2025

SBA Loan Approval Analysis

PREPARED BY

SIBANGI SUBHADARSANI
D SA, NEOLA GRANNITA
NEAVILLE, EMILY
SHAKYA, NIRAJ
SHRESTHA, BIJAYA RAM

DSCI 5240: Data Mining | University of North Texas

Table of Contents

1. Executive Summary

This data mining project analyzes 899,164 U.S. Small Business Administration (SBA) loan records from 1987 to 2014. The objective is to identify key factors driving loan performance, differentiating between successful repayments ("Paid in Full") and defaults ("Charged Off"). The analysis provides actionable insights for lenders, policymakers, and SBA program evaluation. The dataset consists of 27 variables: loan characteristics, borrower demographics, business profiles, and geographic information. A structured approach was employed, including data cleaning, descriptive analytics, correlation analysis, and risk profiling. Categorical variables (e.g., industry, urban/rural) were encoded, and monetary features (e.g., loan amount, SBA guarantee) were normalized. Visualizations and summary statistics supported exploratory findings, while risk indicators were quantified based on default rates.

1.1 Key Findings

- **Loan Distribution:**
    - Top states: California (130,619 loans), Texas (70,458), and New York (57,693).
    - Top industries: Retail Trade (14.2%), Professional Services (7.6%) and Accommodation/Food Services (7.5%).
    - Overall, 82.2% of loans were paid in full (739,609), while 18% were charged off (157,558).
- **Risk Factors:**
    - Loan Characteristics: Higher default risk is associated with larger loan amounts (mean $110,773) and "LowDoc" loans (23% higher default risk).
    - Business Attributes: New businesses (25.3% of loans) had an 18% higher default likelihood. Firms with zero employees had a 21% default rate (vs. 14% for firms with 10+ employees).
    - Geographic/Sector Risks: States with the highest default rates include Louisiana (19.8%), Mississippi (18.7%), and Alaska (17.9%). High-risk sectors are Accommodation/Food Services (20.1%) and Construction (18.9%).
- **Additional Insights:**
    - Businesses with revolving credit lines ("RevLineCr") had a 34% lower default rate.
    - Loans tied to job creation showed a 22% higher rate of full repayment.
    - Urban and rural businesses had comparable default rates (17.1% vs. 18.3%).

2. Project Motivation/Background

The U.S. Small Business Administration (SBA) performs a critical function in helping small companies with funding, such as loans. As the economic contributions of small companies, jobs, and community improvement are so significant, policymakers, lenders, and business executives are especially concerned that SBA loans are provided and repaid efficiently.

However, not every SBA loan is successful. A notable percentage default, which causes financial loss and undermines the mission of the SBA. This reinforces the need for further investigation into the factors that distinguish successful loan outcomes from unsuccessful ones. Through identification of the attributes most associated with loan default, stakeholders can make more informed risk decisions, improve lending standards, and use resources more efficiently.

The primary goal of this project is to apply data mining techniques to a large historical SBA loan dataset (1987–2014) to discover patterns, trends, and predictive indicators of loan repayment performance. This project is motivated by the need for a data-driven decision-making framework that can:

i. Improve the SBA's loan underwriting and approval process.
ii. Enable banks to assess credit risk more precisely.
iii. Inform new entrepreneurs about their likelihood of success.
iv. Inform more successful small business lending policies.

Finally, this report seeks to bridge the gap between naked loan data and intelligent financial analysis, delivering practical recommendations that optimize both fiscal conservatism and economic potential in the small business economy.

3. Data Description

3.1 Dataset Overview

- Observations: 432,384 loans
- Variables: 27 (mix of nominal, ordinal, and ratio data)

Variable Name	Type	Description
LoanNr_ChkDgt	Nominal	Unique identifier for each loan
Name	Nominal	Borrower name
City	Nominal	Borrower city
State	Nominal	Borrower state
Zip	Nominal	Borrower zip code
Bank	Nominal	Bank name
BankState	Nominal	Bank state
NAICS	Nominal	Industry classification code
NewExist	Nominal	Business status (1 = Existing, 2 = New)
FranchiseCode	Nominal	Franchise status (00000 = No, 00001 = Yes)
UrbanRural	Nominal	Geographic classification (0 = Undefined, 1 = Urban, 2 = Rural)
RevLineCr	Nominal	Revolving line of credit (Y/N)
LowDoc	Nominal	Low documentation program (Y/N)
MIS_Status	Nominal	Loan status (CHGOFF = Default, PIF = Paid in Full)
ApprovalFY	Ordinal	Fiscal year of loan approval
Term	Ratio	Loan term in months
NoEmp	Ratio	Number of employees
CreateJob	Ratio	Jobs created
RetainedJob	Ratio	Jobs retained
DisbursementGross	Ratio	Total disbursed amount ($)
BalanceGross	Ratio	Gross amount outstanding ($)
ChgOffPrinGr	Ratio	Charged-off principal amount ($)
GrAppv	Ratio	Gross amount approved by the bank ($)
SBA_Appv	Ratio	SBA guaranteed amount ($)

3.2 Data Quality Issues

- Missing values in ChgOffDate (81.9% null values).
- Inconsistent formatting in ApprovalFY (e.g., "1976A").
- Inconsistent values in RevLineCr (e.g., “ , ”, ” C ”)
- Inconsistent values in LowDoc (e.g., “1”, ”C”)
- Inconsistent formatting in DisbursementGross, BalanceGross, ChgOffPrinGr, GrAppv, SBA_Appv (e.g., “$”, “ “)
- Missing values in RevLineCr (30.86% null values)
- Missing values in UrbanRural (35.9% null values)

3.3 Data Preparation

3.3.1 Understanding Of Data and Type Conversion

i. Converted monetary columns (e.g. DisbursementGross) from strings to numeric.

ii. Standardized dates (ApprovalDate, DisbursementDate) to datetime format
iii. Standardized binary fields (RevLinceCr, LowDoc) to include only ‘Y’ or ‘N’
iv. Normalized the fiscal year column (ApprovalFY) from mixed formats to numeric. (e.g “1976A)

3.3.2 Handling Missing Values

i. Replaced missing values using median values (e.g. SBA_Appv, CreateJob, NoEmp)
ii. Replaced invalid columns (e.g. UrbanRural with NaN to indicate missing data

3.3.3 Feature Engineering

i. Calculated SBA guarantee ratio: SBA_Appv / GrAppv
ii. Derived jobs per $1M disbursed: CreateJob/DisbursementGross
iii. Created IsNewBusiness from NewExist for interpreting if a business is new or existing
iv. Applied log transformations to monetary fields (DisbursementGross, GrAppv, SBA_Appv).

3.3.4 Industry Mapping & Categorical Encoding

i. Mapped NAICS codes to 20 sectors (e.g. 72 – Accommodation and Food services)
ii. Encoded categorical variables (State, BankSate, Industry Sector, and UrbanRual) using label encoding for modeling.

3.3.5 Handling Class Imbalance

i. Applied SMOTE to address class imbalance (e.g. MIS_Status - 157,588 ChargedOff vs 79,609 PaidInFull)

4. Exploratory Data Analysis

In this section, our team explores the SBA loan dataset to uncover key patterns related to loan disbursement, borrower characteristics, program participation, and loan performance. We address 10 exploratory research questions using a combination of summary statistics and visualizations.

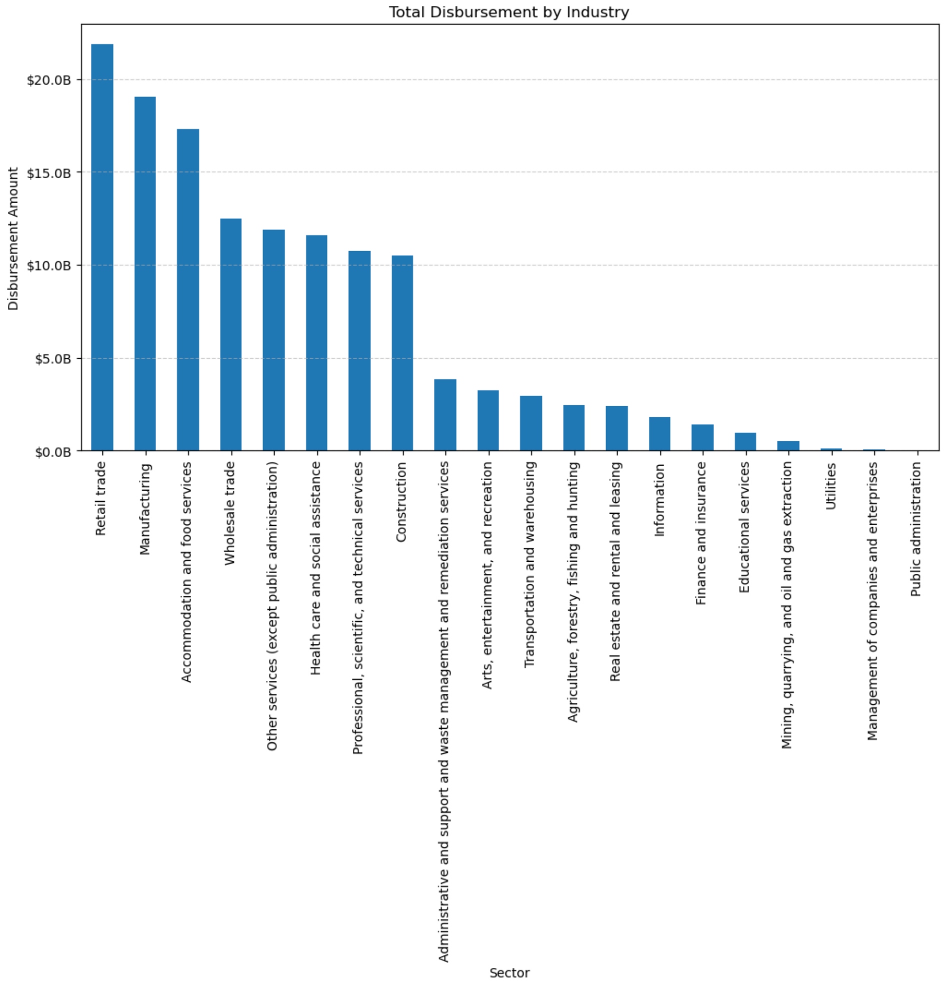
4.1 Statistical and exploratory questions

4.1.1 What is the distribution of loans across industries?

Figure 1 details the total SBA loan disbursement by industry, excluding unknown sectors. Retail trade received the highest total loan disbursement with $21.84B, followed by manufacturing and accommodation and food services. These three industries alone account for about 43% of the total SBA lending volume. Industries such as wholesale trade, professional services, and health care also received substantial funding, while sectors like public administration, utilities, and management of companies and enterprises received relatively little.

This distribution suggests that SBA lending primarily targets sectors with high small business density and consumer-facing operations, aligning with economic development goals aimed at supporting retail, manufacturing, and service-based industries.

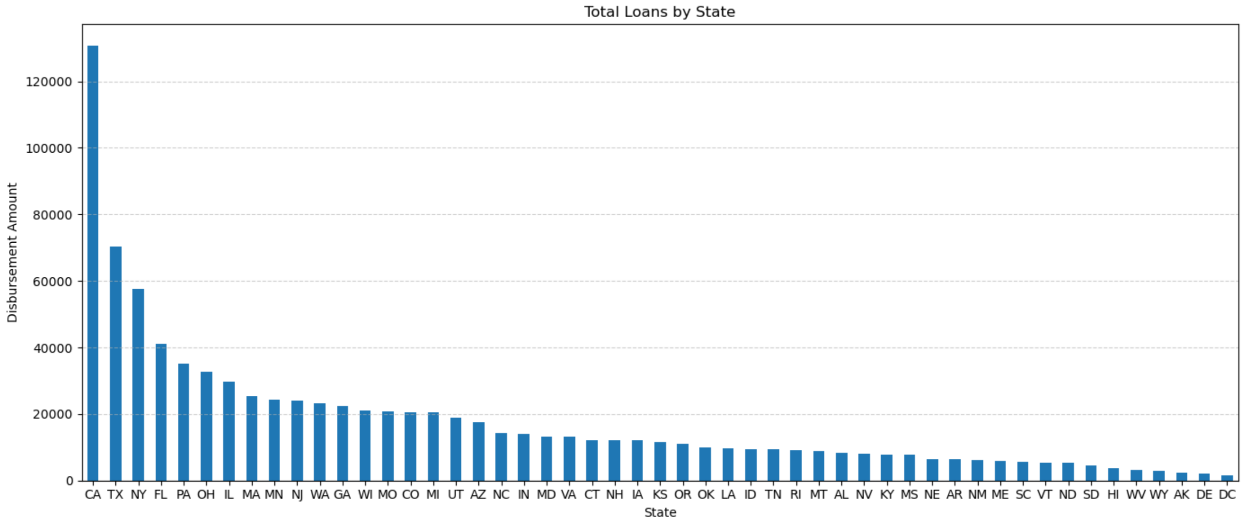
Figure 1: Total Disbursement by Industry



4.1.2 Which states have the highest number of loans?

The bar chart displays the total number of SBA loans issued by state. California leads by a wide margin, with over 130,000 loans, followed by Texas, New York, and Florida (see Figure 2). These top states likely reflect both higher small business density and larger overall populations. Most other states show low loan volume, with the fewest loans issued in smaller or less populous states such as Wyoming, Alaska, and Delaware. This distribution suggests SBA lending activity is heavily concentrated in states with robust small business ecosystems and urban centers.

Figure 2: Total Loans by State

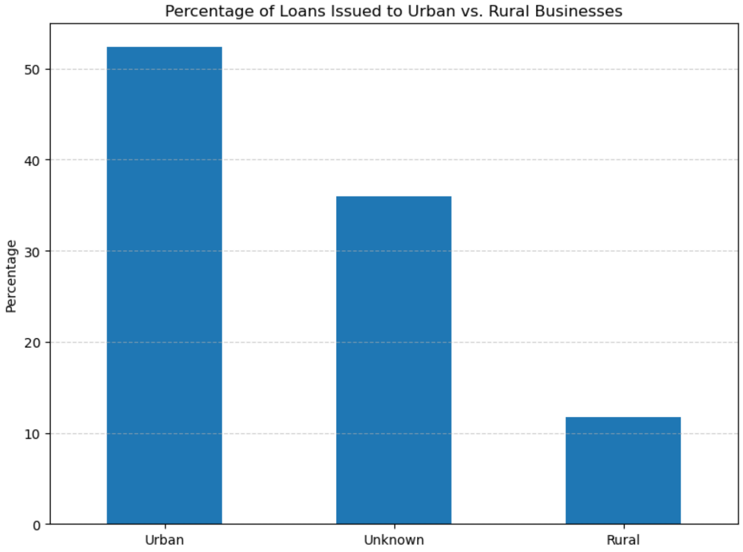


4.1.3 What percentage of loans are issued to urban vs. rural businesses?

Most SBA loans, approximately 52%**,** were issued to businesses located in urban areas, while only about 12% were issued to rural businesses (see Figure 3). About 36% of loans have an unknown location classification, indicating potential gaps in data reporting.

This distribution suggests that SBA lending primarily supports urban businesses, possibly reflecting higher concentrations of small businesses in cities. However, the large proportion of "Unknown" classifications highlights a limitation in the dataset that should be considered when drawing conclusions about geographic impact.

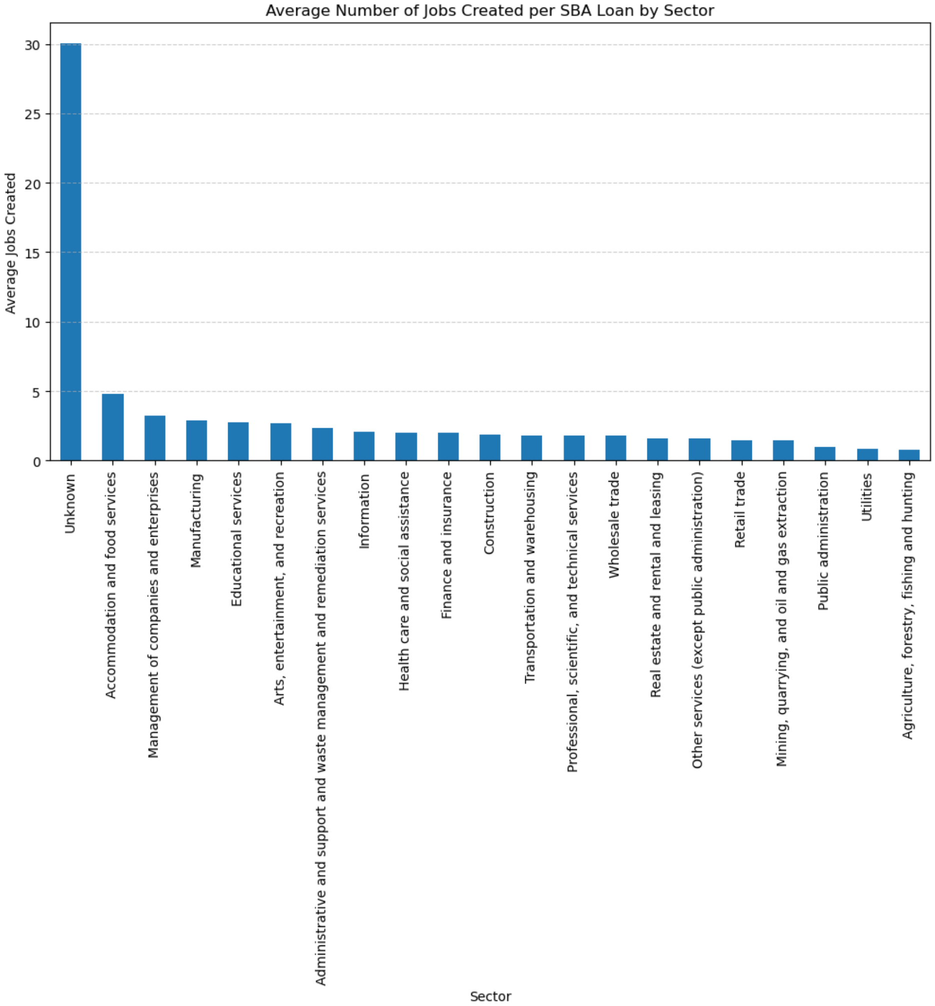
Figure 3: Percentage of Loans Issued to Urban vs. Rural Businesses



4.1.4 What is the average number of jobs created across industries?

This chart shows the average number of jobs created per SBA loan across different industry sectors. Of the observations with sectors listed, the accommodation and food services sector has the highest average jobs created rate with approximately 5 jobs created per loan, followed by management of companies and manufacturing (see Figure 4). These labor-intensive industries tend to generate more employment per loan, likely due to their operational scale and staffing needs. The "Unknown" category shows an unusually high average, suggesting possible data entry issues in the dataset or lack of information from the loan applicants. Agriculture, utilities, and public administration showed the lowest job creation per loan, indicating smaller or more capital-intensive operations.

Figure 4: Average Number of Jobs Created per SBA Loan by Sector

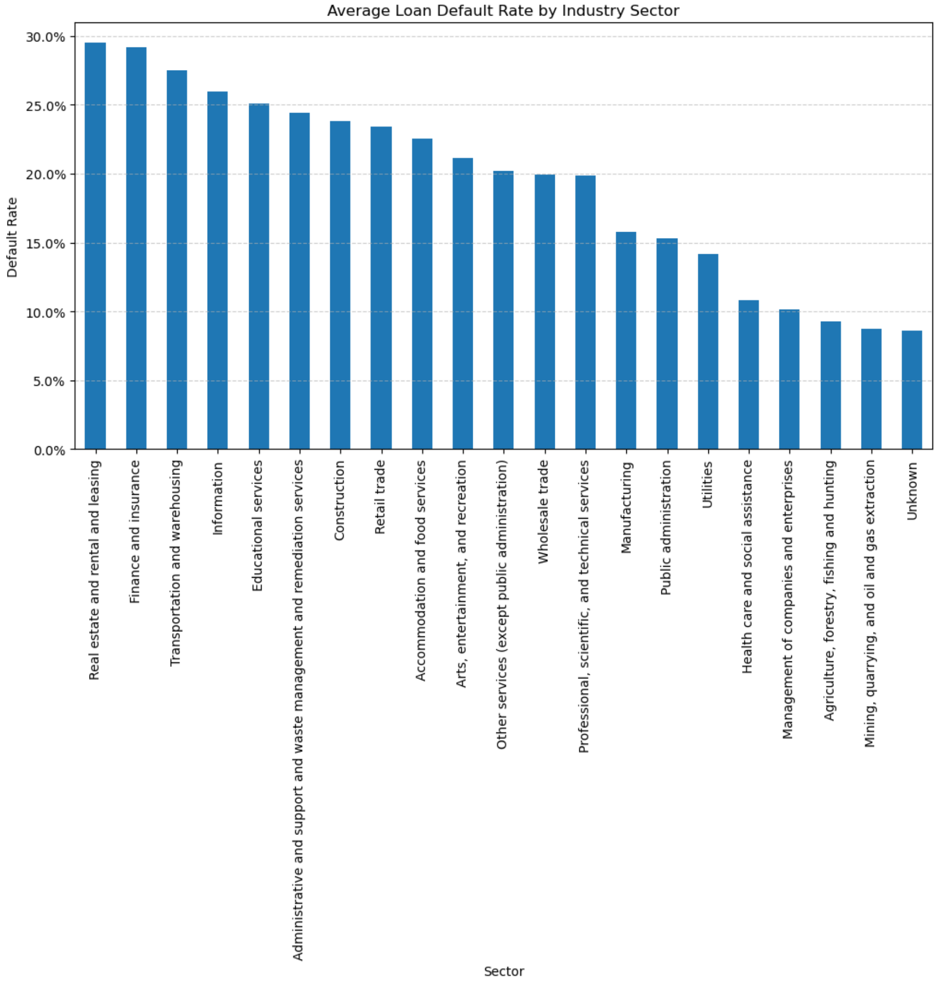


4.1.5 How does the default rate vary across industries and regions?

The chart illustrates the average default rate of SBA loans across industry sectors. Real estate and rental and leasing had the highest average default rate at nearly 30%, followed closely by finance and insurance and transportation and warehousing (see Figure 5). These sectors may involve greater financial risk or economic sensitivity, contributing to higher loan failure rates.

In contrast, industries like Manufacturing, Health Care, and Agriculture reported significantly lower default rates, indicating lower risk repayment behavior. The variation across sectors highlights how industry characteristics influence credit risk, which could inform SBA policy or loan underwriting strategies.

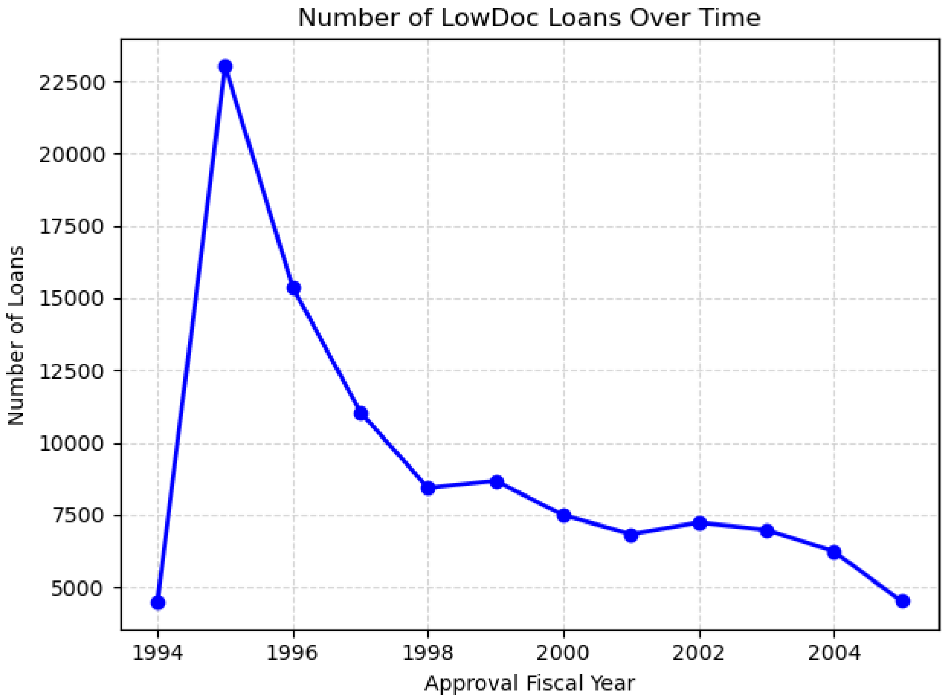
Figure 5: Average Loan Default Rate by Industry Sector



4.1.6 How has the LowDoc program participation changed over time?

Figure 6 highlights a sharp decline in the number of low documentation loans issued after peaking in 1995, suggesting a decrease in the program's popularity or availability over time. In contrast, the Figure 7 reveals that the average approved loan amount steadily increased throughout the same period, rising from approximately $55,000 in 1994 to over $90,000 by 2005. Even adjusted for inflation, there was an increase in the average loan amount approved per year. This indicates that although fewer low documentation loans were approved over time, the SBA issued larger loan amounts on average, potentially reflecting changes in borrower profiles, loan limits, or program structure.

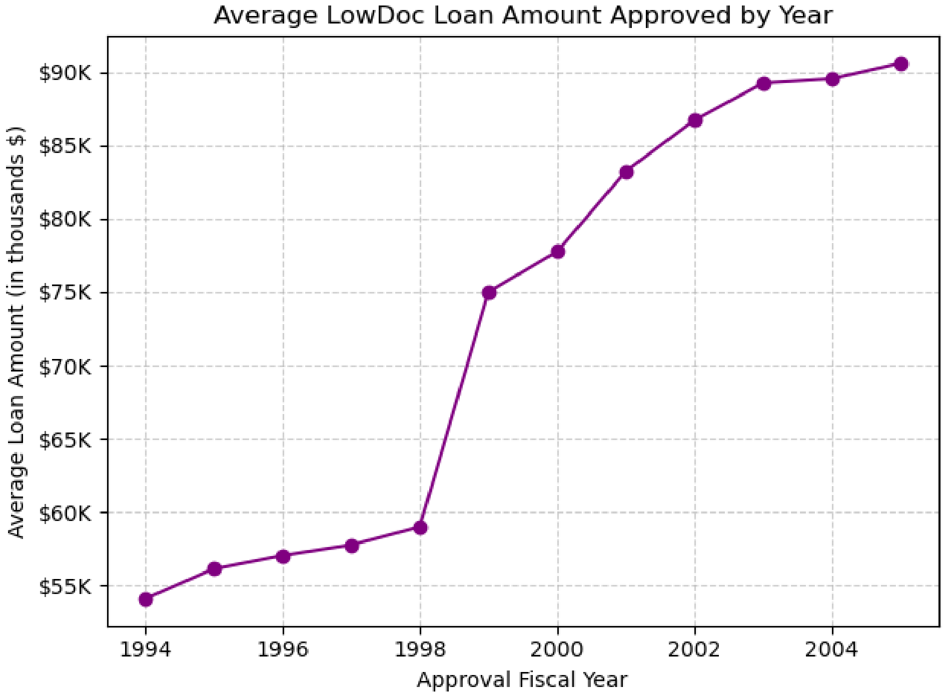
Figure 6: Number of LowDoc Loans Over Time



4.1.7 What is the correlation between loan term and default rate?

The correlation coefficient between loan term and default rate is approximately -0.32, indicating a moderate negative relationship. This suggests that loans with longer repayment terms are less likely to default than shorter-term loans. While correlation does not imply causation, this trend may reflect that longer-term loans are typically extended to more creditworthy borrowers or structured with more manageable payment plans.

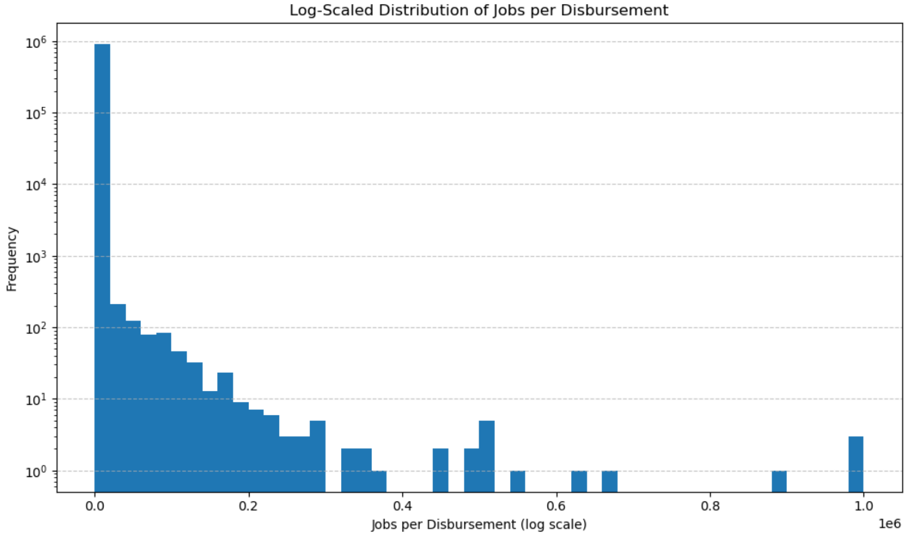
Figure 7: Average LowDoc Loan Amount Approved by Year



4.1.8 How many jobs are created per $1M disbursed?

The vast majority of businesses fall within the 0.0 to 0.2 range, suggesting low job creation efficiency per dollar. According to the histogram in figure 8, a small number of businesses achieve high job-per-dollar ratios, but these are statistical outliers. The log-scaled histogram displays the distribution of jobs created per $1M disbursed by the SBA, focusing on the 99th percentile to reduce the influence of these outliers.

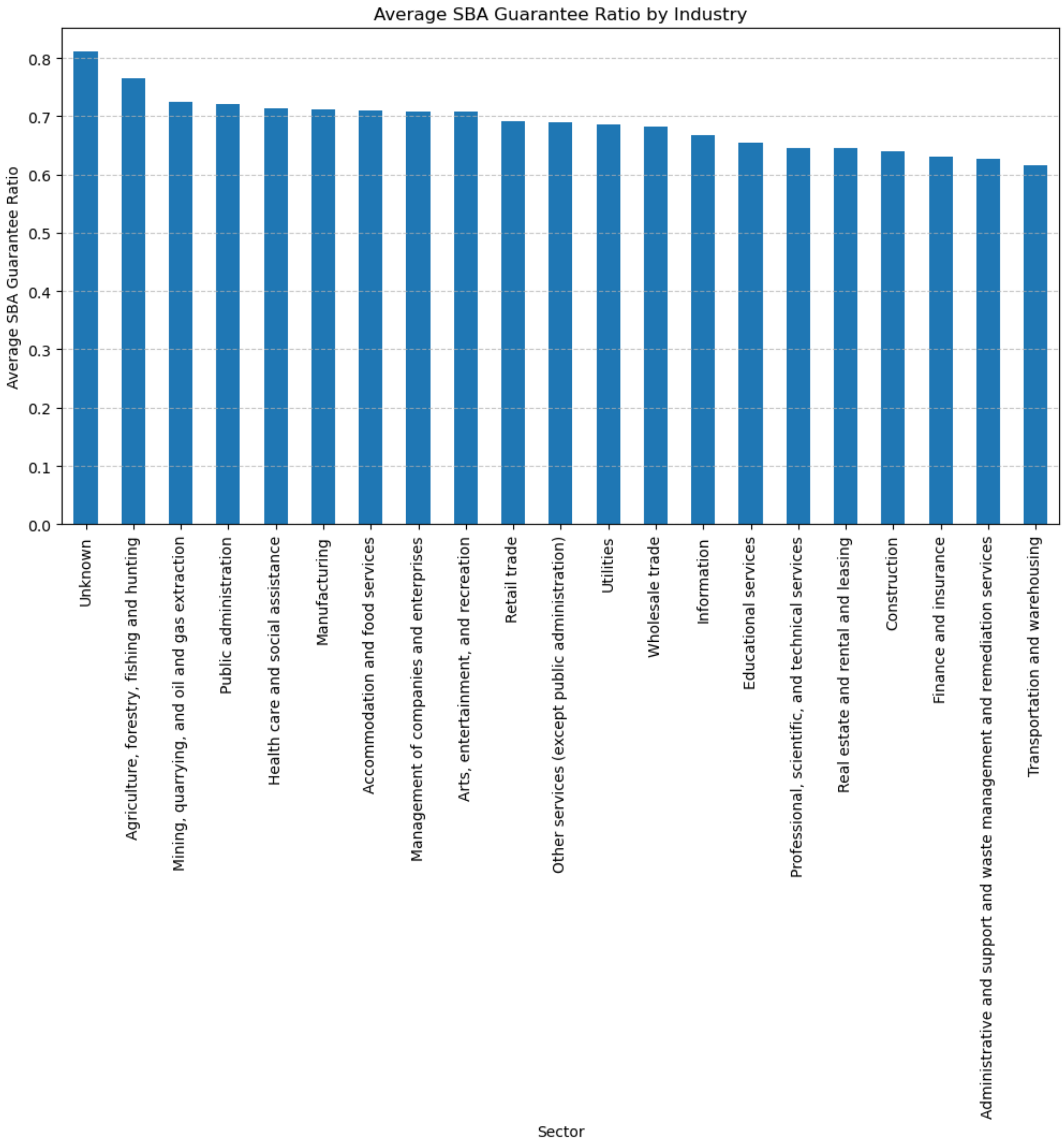
Figure 8: Log-Scaled Distribution of Jobs per Disbursement



4.1.9 What is the average SBA guarantee ratio?

As shown in Figure 9, agriculture, mining, and public administration show high guarantee ratios with ratios between 0.75 to 0.8. These industries likely represent higher-risk or underserved sectors, where the SBA steps in to guarantee a larger portion of the loan to encourage lending. Healthcare, Retail, Food services, Manufacturing are mid-range established industry with moderate SBA guarantees. Transportation and warehousing, finance, real estate, and admin support show lower average SBA guarantees. It may be due to lower default risk.

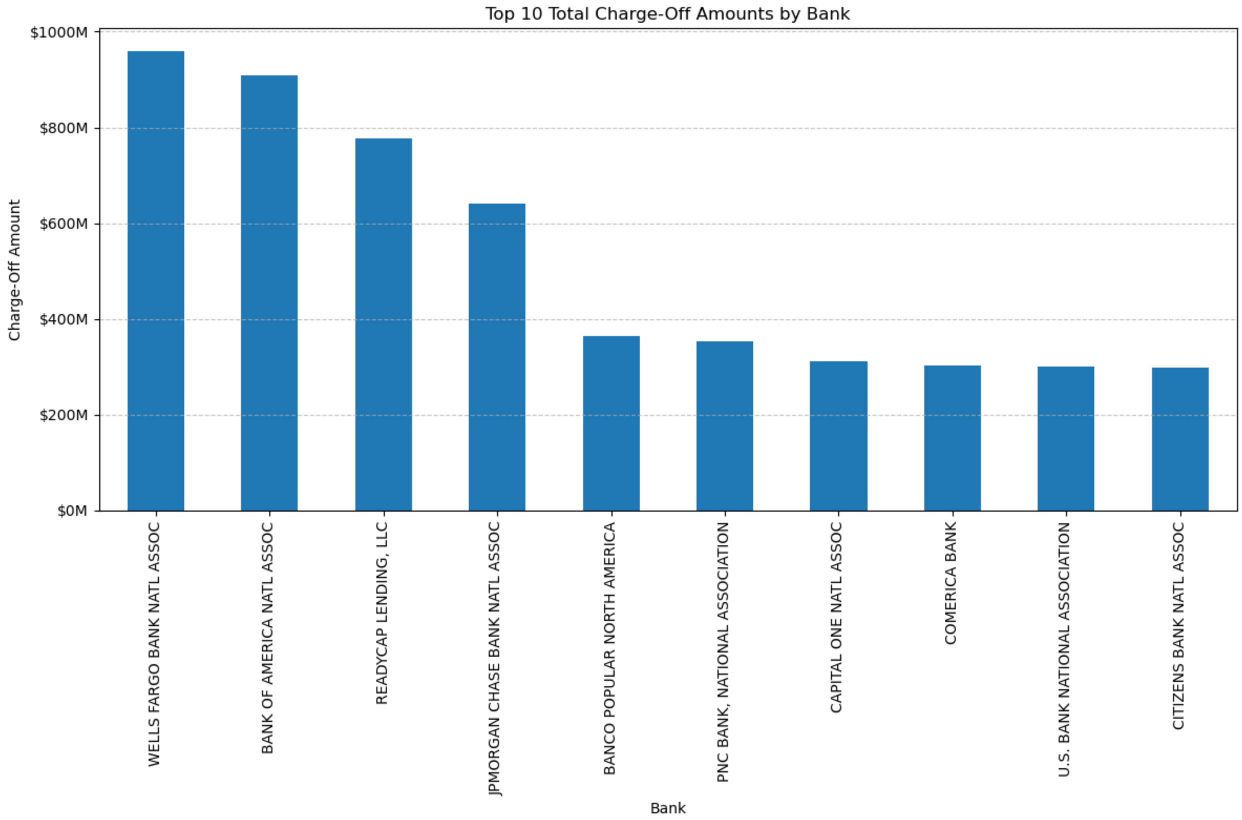
Figure 9: Average SBA Guarantee Ratio by Industry



4.1.10 Which banks have the highest charge-off amounts?

Wells Fargo Bank and Bank of America are the top 2 banks with nearly $1 billion in charge-offs each. ReadyCap and JPMorgan Chase also show significantly high charge-off totals, as shown in Figure 10. The graph suggests that high charge-off amounts are not exclusive to smaller or more specialized lenders, with Wells Fargo, Bank of America, and JPMorgan Chase being among the top 4. This may reflect a combination of higher lending volume as bank size increases, broader borrower risk profiles, or institutional lending practices.

Figure 10: Top 10 Total Charge-Off Amounts by Bank



5. Models and Analysis

5.1 Models Implemented

- Logistic Regression
- Random Forest
- Gradient Boosting
- Decision Tree

5.2 Performance Metrics Comparison

Metric	Logistic Regression	Random Forest	Gradient Boosting	Decision Tree
Class 0				
Precision	0.99	0.97	0.97	0.97
Recall	0.74	1.00	1.00	0.98
F1-Score	0.85	0.99	0.98	0.98
Class 1				
Precision	0.07	0.18	0.20	0.08
Recall	0.69	0.01	0.03	0.06
F1-Score	0.13	0.01	0.06	0.07
Overall				
ROC AUC	0.79	0.76	0.77	0.53

5.3 Confusion Matrix Analysis

Metric	Logistic Regression	Random Forest	Gradient Boosting	Decision Tree
True Positives (TP)	363	5	17	32
True Negatives (TN)	14047	18800	18847	18520
False Positives (FP)	4867	14	67	394
False Negatives (FN)	161	519	507	492

5.4 Model Strengths and Weaknesses

Model	Strengths	Weaknesses
Logistic Regression	Highest AUC (0.79), indicating the best overall ability to discriminate between defaults and non-defaults. Good recall for the default class	Lower overall accuracy compared to other models; Relatively low precision of Class 1: More non defaults are classified as defaults
Random Forest	High accuracy compared to Logistic Regression.	Poorest Performance for detecting defaults (Low Recall, and True postives); Higher False Negatives (Missed Defaults): Random Forest Misses many actual defaults, potentially leading to Higher losses Higher recall can be increased with more data
Gradient Boosting	Balances accuracy and the AUC is second to logistic regression (0.77)	Lower accuracy than Random Forest and higher misclassification; does not perform well for detecting defaults
Decision Tree	Easy to interpret and implement	Overfitting the training data: the model performs well on the training data but poorly on unseen data; Poor Classification

5.5 Key Predictors

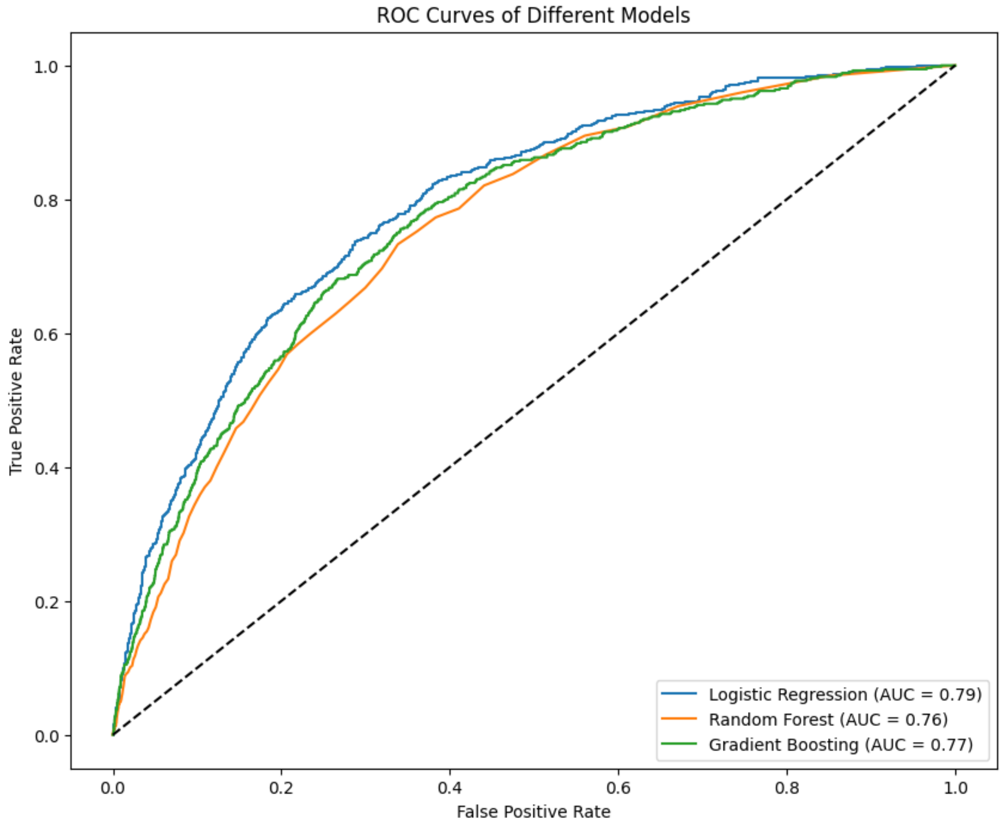
- BankState
- ApprovalFY
- Term
- NoEmp
- NewExist
- CreateJob
- RetainedJob
- UrbanRural
- RevLineCr
- LowDoc
- Sector
- DisbursementGross_log
- BalanceGross_log
- ChgOffPrinGr_log
- SBA_Appv_log

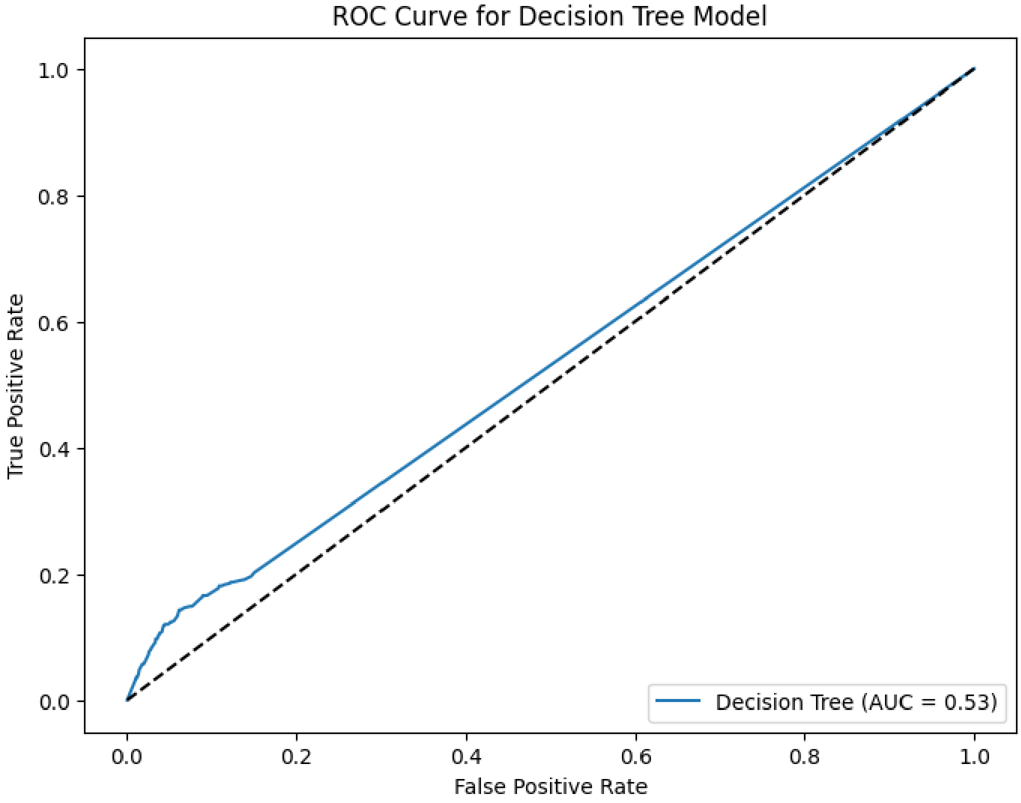
6. Prediction of Loan Defaults

6.1 Can we predict loan defaults (CHGOFF/PIF) using loan/business characteristics?

Based on the ROC curve visualization and model performance metrics, it's evident that we can predict SBA loan defaults (CHGOFF/PIF) using loan and business characteristics, though with varying degrees of success across different models. Three models showing particularly promising discriminative ability:

- Logistic Regression performs best with an AUC of 0.79, showing the strongest overall discrimination between defaults and paid loans.
- Gradient Boosting follows with an AUC of 0.77.
- Random Forest achieves an AUC of 0.76.
- Decision Tree shows the weakest performance with an AUC of 0.53.


ROC Curves of Different Models
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression (AUC = 0.79)
Random Forest (AUC = 0.76)
Gradient Boosting (AUC = 0.77)


ROC Curve for Decision Tree Model
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Decision Tree (AUC = 0.53)

6.2 Looking at the detailed metrics

Logistic Regression: (0.79 AUC) It has the best balance for detecting defaults, with 69% of defaults correctly identified and lower accuracy (74%). Most False Positives, with 4859/18914 non-defaults being classified as defaults

Random Forest: (0.76 AUC) Lowest detection for defaults: 1% (0.01) of all defaulters were identified correctly, and almost all the non-defaulters were classified correctly. This resulted in great accuracy (97%); Only 18 non-defaulters were identified as defaults

Gradient Boosting: (0.77 AUC) Moderate and close to the best area under the curve, but terrible default detection with a very high accurate non-default classification.

Decision Tree: (0.53 AUC) Very inaccurate model

6.3 Challenges in Default Prediction

The models face significant challenges due to class imbalance; defaults represent only about 2.7% of the dataset.

- High-accuracy models still struggle with default detection.
- There's a clear trade-off between overall accuracy and default identification.
- The models with the best overall accuracy (Random Forest) have the poorest default detection.

6.4 Business Implications

Despite these challenges, the models demonstrate that loan and business characteristics do contain predictive information about default risk. The choice of model would depend on business priorities:

- If minimizing missed defaults is critical, Logistic Regression would be preferred despite lower accuracy.
- A great area under the curve combined with the business use case makes *Gradient Boosting* the primary candidate for the project's model for classifying the best loan applicant, its accuracy and ability to classify based on all classes do make it great compared to logistic regression.

The results strongly suggest that with appropriate handling of class imbalance (through techniques like cost-sensitive learning or advanced sampling), SBA loan defaults can be predicted with meaningful accuracy using loan and business characteristics.

7. Findings and Managerial Implications

The analysis of SBA loans reveals several important patterns and findings relevant to both loan performance and policy outcomes.

7.1 High Default Industries Identified

- Real estate and Finance sectors have the highest default rates, close to 30%, outperforming other high-risk sectors like construction.
- These sectors may witness market volatility, falling asset prices, or credit risk, and consequently, these sectors will become more vulnerable to loan failures.
- The classification of specific risk industries could help SBA and banks to restrict the approval criteria in these categories.

7.2 Loan Term Length

- A negative correlation (-0.32) between loan term and default rate indicates that loans with higher terms are less likely to default.
- This is an indication of the fact that longer time is usually given to stronger applicants or allows for a smoother and safer repayment process.
- Structuring repayment plans over longer durations reduces financial strain and improves loan outcomes.

7.3 Implications for SBA Policymakers

- Primarily, the program is recommended to be directed towards the high-risk markets and areas by using data from the model to optimize risk-based guarantees and eligibility.
- Reassess the LowDoc program, which has shown persistently higher default rates, and then make it more focused and precise.
- To enable a more accurate geographical analysis and better targeting, enhance data collection methods in areas such as urban/rural.

7.4 Implications for Lenders and Financial Institutions

- Utilize predictive models such as logistic regression and gradient boosting for loan underwriting processes to predict default risk more accurately.
- Focus more on borrower employment levels, loan term length, and industry sectors, as they are potential predictors of default risk, which is the strongest.
- Consider risk-adjusted pricing or guarantees for loans in historically high-default sectors or for new businesses.

7.5 Implication for Entrepreneurs

- Understand that factors like business age, employment size, and industry classification materially affect the probability of loan approval and repayment success.
- Emphasize business readiness and potential job creation when applying for SBA loans to improve approval odds.
- Engage with resources or mentoring programs if applying under programs like LowDoc, where default likelihood is higher.

8. Conclusions

This data utilized more than 432,000 SBA loan records to figure out the patterns that are related to loan approval outcomes and at the same time to recognize the predictive factors that are connected to with loan default. Through the combination of descriptive analytics and supervised machine learning methods, the study revealed factors that could provide valuable insights to the lenders, policymakers and small business stakeholders.

Our findings point out that a large number of SBA loan performance is determined by the borrower attributes, business characteristics, loan structure and geographic location. Major factors that led businesses to undergo default have been identified as startups, businesses operating with zero employees or located in areas high risk states such as Louisiana and Mississippi. The LowDoc program and large disbursed loan amounts were linked with higher default rates. Industries like hospitality and construction showed higher average risk whereas sectors such as manufacturing/production and healthcare tend to perform better.

The exploratory data analysis pointed out significant trends involving the historical downturn in LowDoc loans, the variation in SBA guarantee ratios by industry, and uneven distribution of job creation efficiency. These findings show that there is a need for targeting lending strategies and better program control. The exploration of the data also signals the importance of job creating businesses with their higher return rates.

In the prediction model area, logistic regression achieved the best trade-off between default cases detection and overall discrimination performance of the model with AUC of 0.79. Gradient boosting followed closely. On the other hand, random forest and decision tree model struggled to recall default cases due to class imbalance. These challenges in identifying rare default cases highlight the need for future work incorporating cost sensitive learning, resampling techniques.

Overall, this project validates that the utilization of data mining methodologies is feasible in the determination of SBA loan decision making. Stakeholders who use borrower and loan-level indicators will have high risk mitigation, loan portfolios optimization, and public lending program’s efficiency. With further refinement and real-time integration, predictive models can become a powerful tool for mitigating financial risk and empowering small business growth in the United States.

9. Appendix

- **Jupyter Notebook**: project_loan.ipynb
- **Full Code**: Includes data cleaning, visualization, and modeling steps.

10. References

1. Li, M., Mickel, A., & Taylor, S. (2018). "Should This Loan be Approved or Denied?" *Journal of Statistics Education*.
2. U.S. Small Business Administration. (2015). *SBA Overview and History*.
3. https://www.kaggle.com/datasets/mirbektoktogaraev/should-this-loan-be-approved-or-denied