Small Businesses Loan Default Prediction

Project Description

This project addresses the challenge of predicting loan defaults within the U.S. Small Business Administration's (SBA) 7(a) and 504 loan programs. These programs are extremely important in assisting small businesses and promoting economic growth. The 7(a) program focuses on financing the purchase of a business or getting working capital. These loans are smaller in size, with a maximum amount of $5 million. The 504 program focuses on the purchase of commercial real estate or heavy machinery with a larger maximum loan size of $30 million. Together, these loan programs have allowed small businesses to gain the resources they need, which conventional bank loans may not be willing to provide.

The datasets originate from SBA loan records and are divided into two loan types (refer URL). The 7(a) loan data spans from 2020 to the present, while the 504 data spans further back from 2010 to the present. The datasets contain a wealth of information covering financial attributes such as loan amounts and interest rates, business characteristics like business type and age, and loan terms such as loan length and collateral status. The goal of this project is to create a model that can accurately predict which loans are most likely to default, meaning the borrower fails to repay their obligation. Therefore, for the purposes of this analysis, only loans with definitive outcomes, paid in full or charged off (defaulted), will be considered from the dataset.

Importance of Problem

Loan defaults create a significant challenge for both lenders and the SBA because they cause direct financial loss and can also limit the availability of future funding for other small businesses. Although the programs of the SBA are designed to mitigate lender risk through partial guarantees, defaults still drain needed resources. Lessening the SBA's ability to guarantee future loans can also limit the economic opportunities for business owners and entrepreneurs. Because small businesses are a major driver in job creation, preventing defaults has extensive economic implications.

Correctly anticipating and predicting loan defaults is vital because it opens the opportunity for the SBA and lenders to take appropriate action before a default can occur. The earlier a high-risk loan is identified, the more options the SBA and lenders have, ranging from modifying loan conditions to closer monitoring and financial counselling. It also helps lenders to focus on borrowers with a higher likelihood of success to reduce risk. Predictive modeling directly contributes to achieving long-term sustainability for SBA loan programs as well as towards the larger goal of supporting an adaptive

and growing small business ecosystem through less frequent and severe defaults.

Exploratory Analysis

Before performing exploratory analysis, data preprocessing and cleaning were carried out to ensure the data was fully relevant to the problem statement. To begin, the 7(a) and 504 datasets were merged to achieve a fully inclusive dataset. To maintain relevance, data from before the year 2020 was purged. During this process, the columns names in the dataset were standardized to prevent duplication. A binary response variable, *'LoanSuccess'* was created (from the existing 'LoanStatus' column), which filtered the data to include only loans with paid-in-full and charged-off statuses (PIF=0, ChargedOff=1). Next, the incomplete data in the rows and columns were addressed using a two-stage approach: Removing any features that had more than 70% missing values (which removed five columns) and the replacement of remaining columns by imputing 0 for missing numerical values and “Missing” for categorical values to ensure no empty cells in the dataset. Additionally, some features were manually excluded from the dataset as they were either direct outcomes of loan status or were highly textual data that would increase the complexity of the model predictions.

With the dataset cleaned and structured, an exploratory data analysis (EDA) was performed to better understand the characteristics and trends, and any predictive relationships in the loan data. Initial analysis showed that the dataset exhibited a significant class imbalance. Around 94% of loans were fully repaid compared to about 6% of loans that defaulted (refer Figure 1). While this represents a healthy growth portfolio for the lending institution, it exposes the need for strategies to deal with dataset imbalance in modeling. Program-level comparisons were conducted further (refer Figure 2), it was observed that the gross loan approval amounts were typically higher on SBA 504 loans and lower on SBA 7(a) loans. This aligns with the larger fixed-asset projects accompanied by SBA 504 loans. In Figure 3, the trend of average gross approval amounts over time was visualized. A large spike can be seen in 2021, likely correlating with the COVID-19 pandemic, when many small businesses were requiring loans in greater amounts. This is then followed by a decrease in each year following, as lending activity returned to normal and rising interest rates combined with economic uncertainty led to reduced borrowing. Finally, the number of loans and the proportion of defaults over the five years was analyzed (refer Figure 4). Like the gross approval amounts, it has seen a decrease in loan volume since 2020. The amount of defaulted loans does not appear to change dramatically from 2020 to 2024 indicating that the proportion of defaults has increased over those years, as the number of total loans has decreased.

In summary, the EDA provided key insights on loan amounts, terms, and job support, while also identifying class imbalance, which will be crucial for building effective predictive models.

After understanding the dataset, three new features were engineered to obtain three additional relevant ratios, creating measures that reveal patterns and relationships over raw data. The *'SBA Guarantee Ratio'*, which compares the amount guaranteed by the SBA to the total gross approval amount, when plotted with LoanSuccess, showed that the box for defaulted loans was larger, which means the ratios for the central 50% of these loans were more spread out compared to the successful loans. This suggests lower variation in the levels of guarantee across the pool of unsuccessful loans, proving that the ratio could be a useful predictor (refer Figure 5). The *'Loan Term Burden Ratio'* in Figure 6, which takes the gross approval amount over the loan term in months, had clear separation between the medians, with higher ratios producing better repayment outcomes. Similarly, the *'Jobs Impact Efficiency Ratio'*, which takes the gross approval amount over the number of jobs supported, had a higher median for the successful loans as demonstrated in Figure 7, indicating its value as a predictor.

Solution

Before running models, the dataset was transformed into binary features using 'o*ne hot encoding*'. As the problem requires a binary answer (Yes, No), models that predict such outcomes require binary operators to train the model. The dataset was then split into training and testing sets using an 80/20 partition. The training set was used to fit the models, while the testing set was reserved for evaluating their performance.

Three predictive models were chosen to create and compare for accuracy: *Logistic Regression*, *Random Forest*, and *XGBoost*. The logistic regression model was considered as a baseline because it is well suited for binary classification problems. It allows for simple interpretability and demonstrates how each feature influences the chance of default. The Random Forest Classifier, which uses many decision trees to make predictions, was used as it is a powerful and flexible method for handling complex datasets. Finally, the XGBoost model was developed, which sequentially builds upon decision trees with each tree learning from the errors of the preceding ones. This model learns in a way that accounts for complex patterns and interactions in the data.

On initial evaluation of the models, the predictions were skewed (refer Figure 8) due to the class imbalance of the 'LoanSuccess' predictor variable that was established earlier as visible in Figure 1. To compensate for the imbalance (more loans being paid in full than defaulted on), weight was added to each of the models. On rerunning the models with weightage, the

predictions were balanced such that the number of False Negatives and False Positives increased, which inferred better a prediction count of the correctly predicted loan defaults (True Positive). While the Accuracy may have decreased, it remains in the range that is acceptable (>90%).

Summary

This project predicts loan defaults within the SBA's 7(a) and 504 programs, which are crucial for economic growth. The analysis began with cleaning and merging loan data from 2020 onward, creating a binary 'LoanSuccess' variable. Exploratory analysis revealed a significant class imbalance (94% paid in full vs. 6% defaulted), which required strategic modeling.

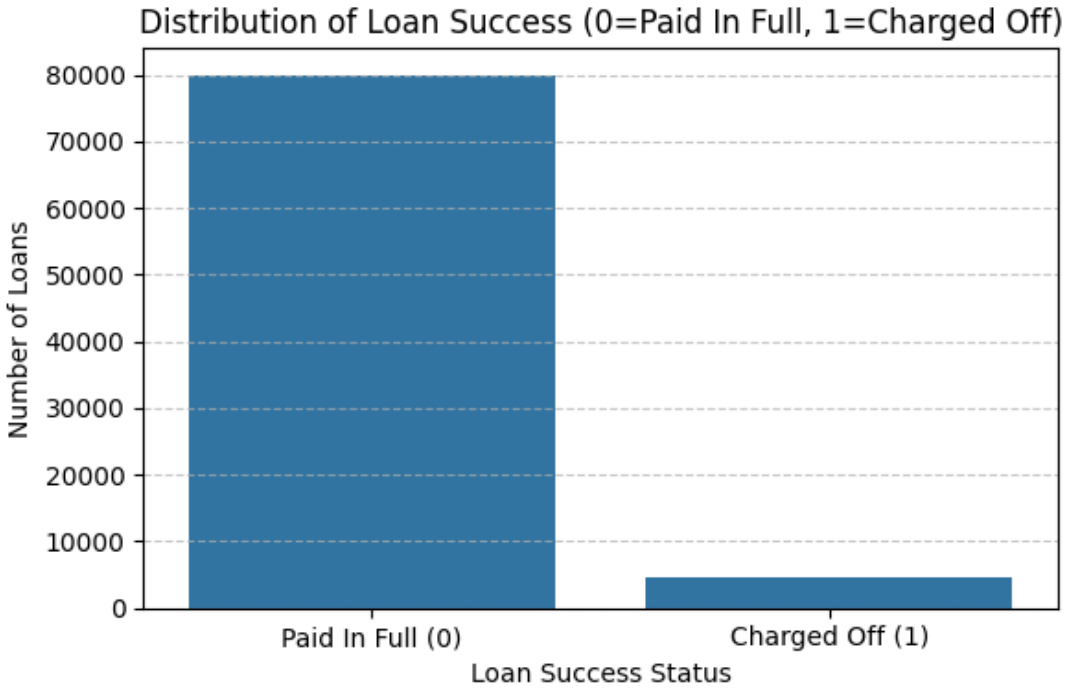
Three new features were engineered to improve predictive power. The models chosen – Logistic Regression, Random Forest, and XGBoost – were initially skewed due to the data imbalance. To counteract this, a weighting system was implemented, which improved the models' ability to correctly predict the minority class (defaults) and provided a more balanced and reliable prediction framework.

Practical Implementations

- *Enhanced Risk Assessment:* The models provide data-driven risk scores to supplement traditional loan review processes. This leads to more informed decisions, reduces human bias, and improves loan portfolio quality.
- *Targeted Lending:* Insights from the models can identify applicant segments with a higher likelihood of success. This allows for more targeted marketing and specialized loan products, optimizing customer acquisition.
- *Policy Refinement:* Model insights offer evidence to refine lending policies. Understanding which factors predict success allows institutions to adjust criteria for fairer and more effective lending.
- *Early Warning System:* The framework can be adapted to monitor ongoing loans and predict early signs of distress. This enables proactive interventions like financial counseling to prevent defaults.

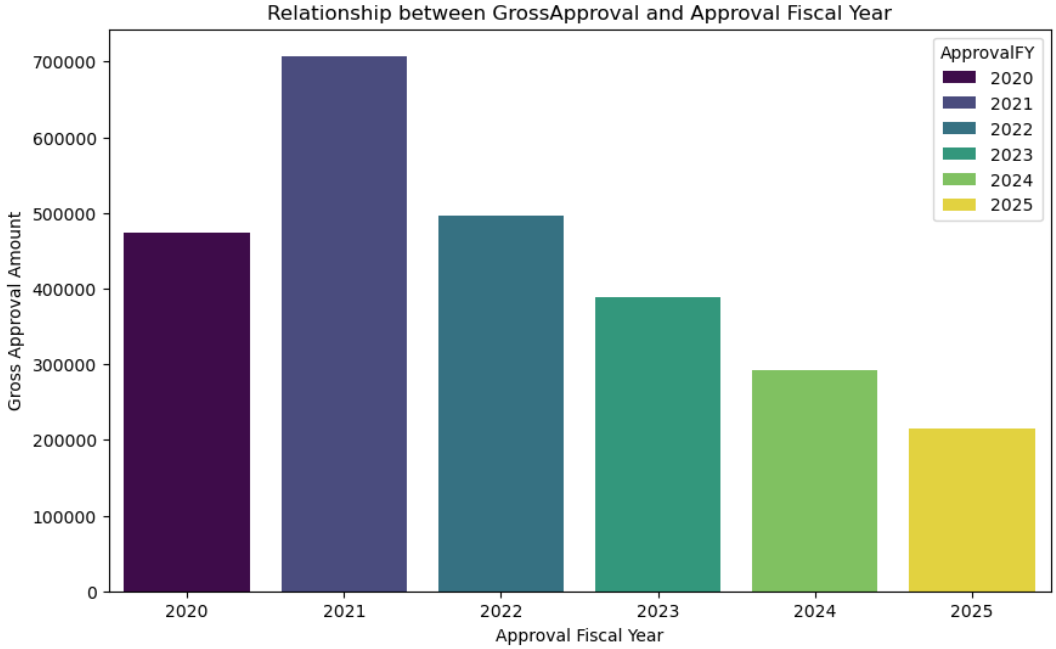
Appendix

1. URL - https://data.sba.gov/en/dataset/7-a-504-foia
2. Figure 1.

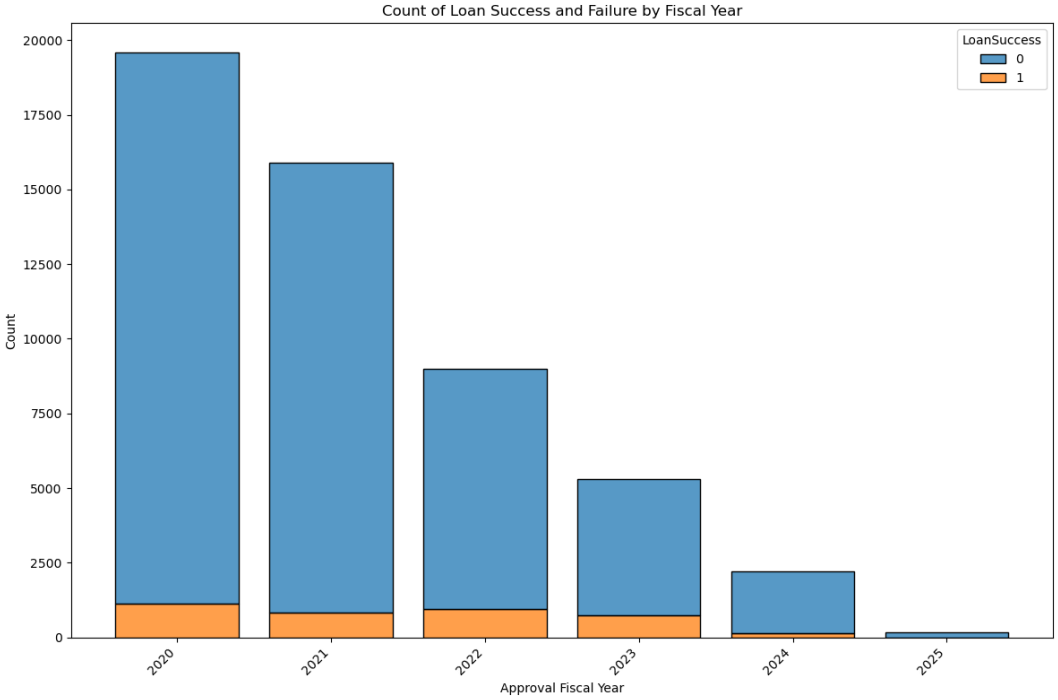


3. Figure 2.

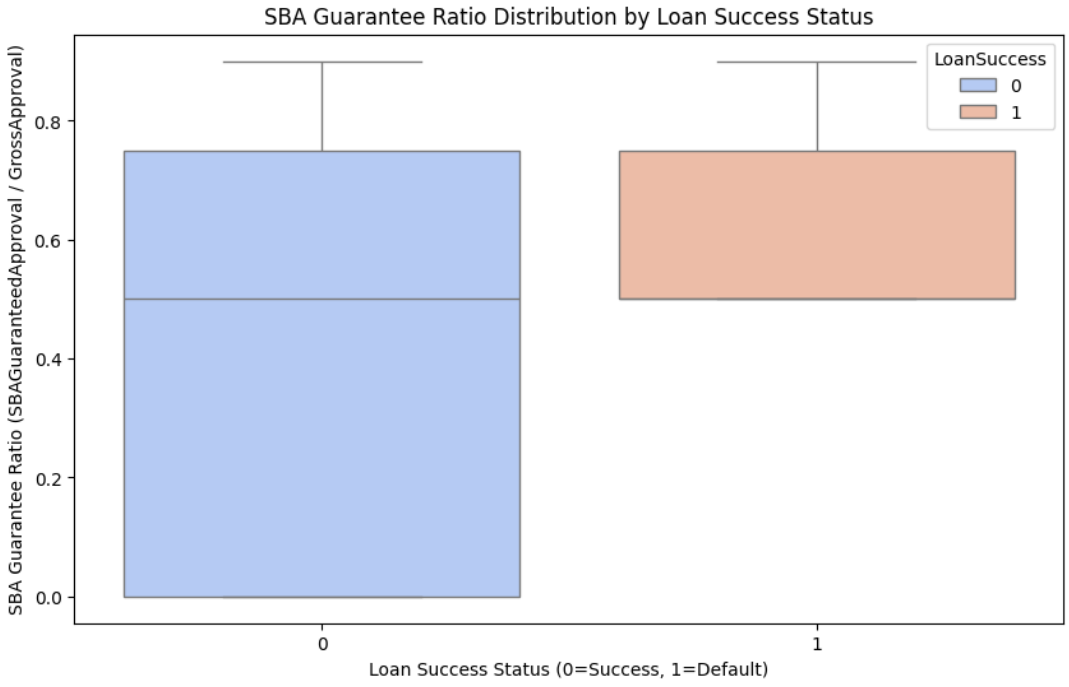
4. Figure 3.



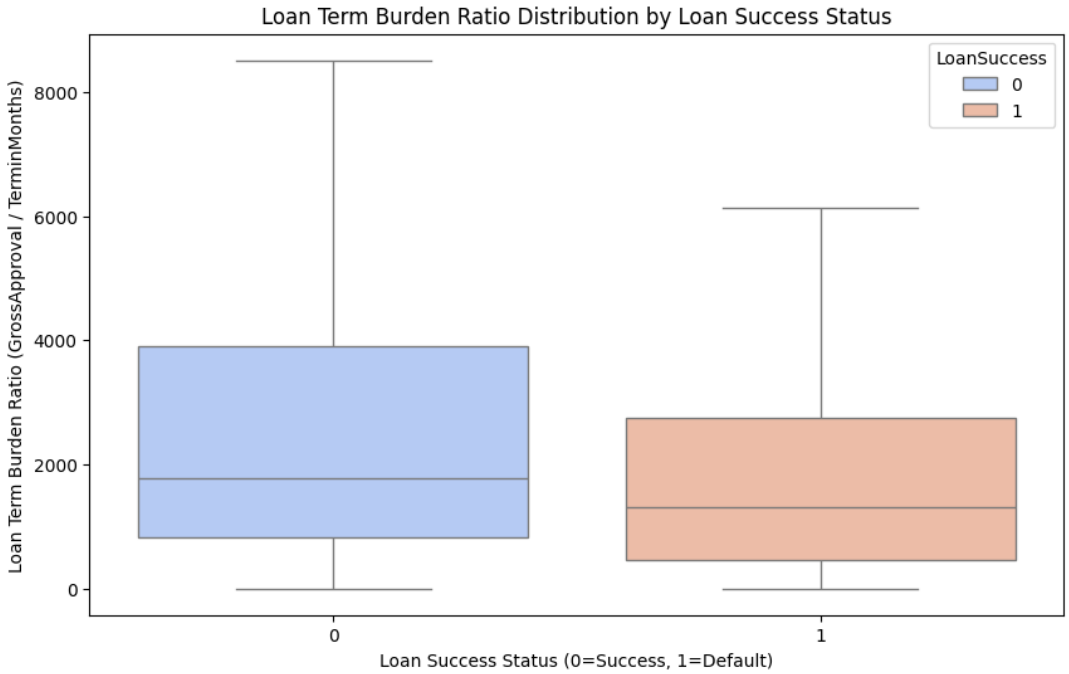
5. Figure 4.



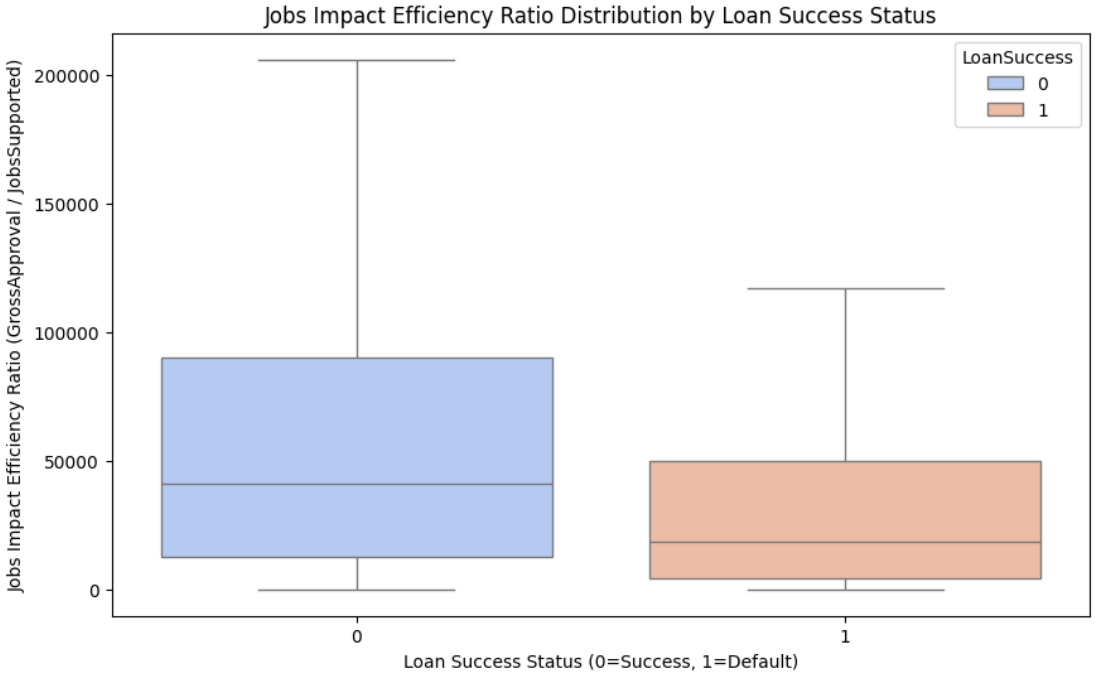
6. Figure 5.



7. Figure 6.



8. Figure 7.



9. Figure 8.

Logistic Regression	Without Weights		Predicted	
			Negative	Positive
	Actual	Negative	1548	0
		Positive	122	0
Random Forest	Without Weights		Predicted	
			Negative	Positive
	Actual	Negative	9635	41
		Positive	323	437
XGBoost	Without Weights		Predicted	
			Negative	Positive
	Actual	Negative	9602	74
		Positive	172	588

10. Figure 9.

Logistic Regression	With Weights		Predicted	
			Negative	Positive
	Actual	Negative	701	847
		Positive	24	98
Random Forest	With Weights		Predicted	
			Negative	Positive
	Actual	Negative	9638	38
		Positive	312	448
XGBoost	With Weights		Predicted	
			Negative	Positive
	Actual	Negative	9541	225
		Positive	78	682

11. Figure 10

Weightage	Without Weights			With Weights		
Model	Logistic Regression	Random Forest	XGBoost	Logistic Regression	Random Forest	XGBoost
Accuracy	92.69%	96.51%	97.64%	47.84%	96.65%	97.10%
Precision	0.00%	91.42%	97.55%	10.37%	92.18%	97.43%