S.B.A Loans Predict Project

Overview

The SBA (Small Business Administration) is an organization established to assist small businesses in obtaining loans. It insures the bank for a certain amount of the loan taken by the company, enabling loans to be provided to small businesses under favorable conditions, thereby facilitating their growth.

If a borrower cannot repay their SBA loan, the lender requests reimbursement from the SBA for the guaranteed portion.

Project objective

The project's objective is to identify loans that are likely to fail and default based on several parameters, such as the loan amount, repayment period, the sector to which the company belongs, and additional characteristics.

Project successes

Success in the project will be defined by the model's ability to identify loans that are likely to default based on their characteristics, thereby preventing their approval from the outset.

Potential Users

Banks, S.B.A Organization: Interested in assessing loans before approval.

Companies: Seeking loans to assess repayment risks

Factors affect

I estimate that among the influencing factors will be the company's sector of activity, the loan amount, and the loan duration.

Additional Factors

I added the Fed interest rate data, which represents the real interest rate for inter-company loans, assuming it also impacts repayment ability

Data Preparation

The dataset contains various loan characteristics, including:

- **Loan year**: The year the loan was approved.
- **Company name**: The name of the company receiving the loan.
- **Loan amount**: The total amount of the loan.
- **SBA-guaranteed amount**: The portion of the loan guaranteed by the SBA.
- **Loan duration**: The term of the loan in years.
- **Sector**: The business sector associated with the loan.
- **Etc**.

Data Preparation Steps:

1. **Sector Grouping**:
   To reduce the number of distinct categories, I grouped business sectors into broader categories based on their characteristics.
2. **Bank Grouping**:
   Similarly, banks were consolidated into groups based on shared patterns identified in their names.
3. **Geographic Information**:
   - **Using Zip Codes**: The company's location was inferred using the zipcode column where available.
   - **Fallback to Street Names**: When zip code information was missing, location data was derived from the street name column.

EDA

Correlation Analysis:

I conducted a correlation analysis between the variables in the dataset to identify potential relationships and patterns.

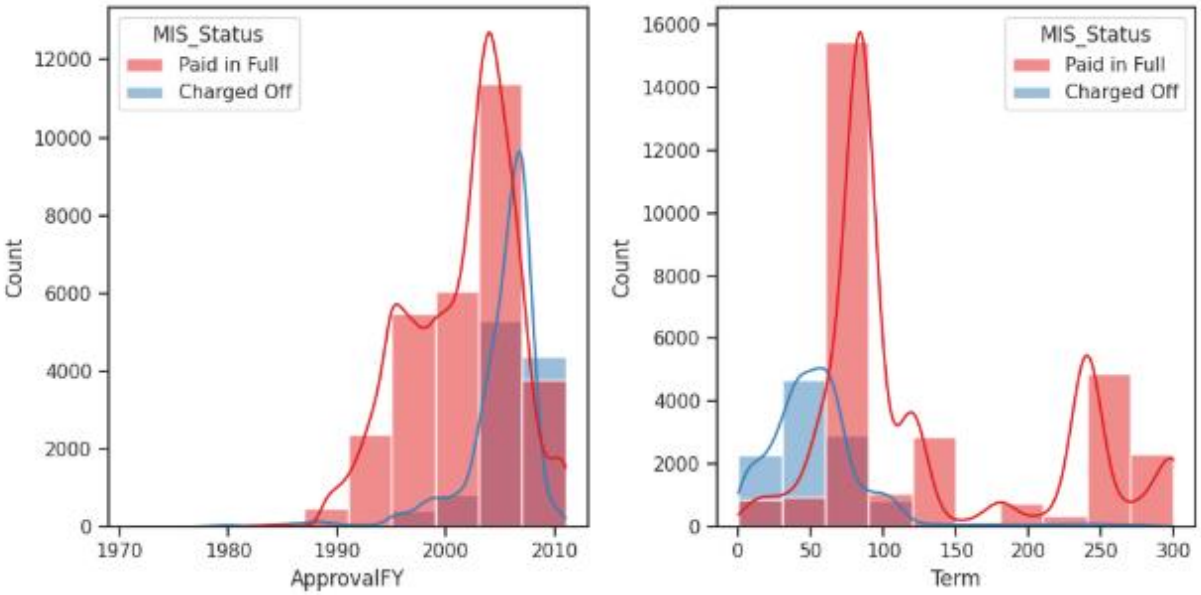
Distribution Analysis by Target Group:

I examined the distribution of variables across the target groups — companies that fully repaid their loans and those that defaulted.

Example Findings:

- Companies that defaulted on their loans tended to have shorter loan durations.
- Defaults were more common among loans issued in later years.

These insights provide an initial understanding of factors that may influence loan repayment behavior, guiding further analysis.



Statistical Tests

To further explore the dataset, I performed the following statistical analyses:

1. **Skewness Test**:
    - Assessed the skewness of each variable to determine the shape of its distribution.
2. **Mann-Whitney U Test**:
    - Conducted this non-parametric test to evaluate whether there are significant differences in the numeric variables between the two target groups (repaid vs. defaulted).
3. **Chi-Square Test**:
    - For categorical variables, I used the Chi-Square test to examine whether there are significant differences in category distributions between the target groups.

Data Cleansing

Outlier Treatment:

- Outliers that did not affect the both correlation and distribution of the variables were replaced with NULL values for consistency.

Missing Values in Categorical Columns:

- Missing values in categorical variables were imputed using the **KNN Imputer**, which estimates the missing values based on the similarity to other observations.

Missing Values in Numeric Columns:

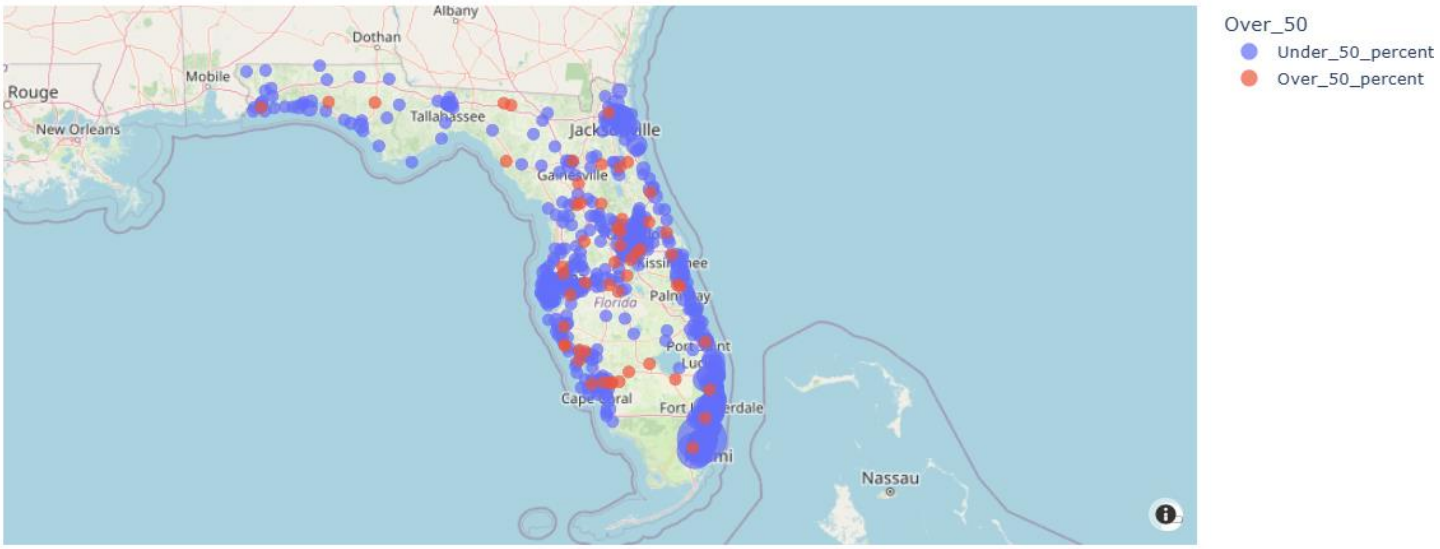
- For numeric variables, missing values were imputed using the **MICE (Multiple Imputation by Chained Equations)** method, which employs regression-based imputation tailored for continuous data.

Feature engineering

Geographic Analysis

Using zip code data, I mapped the coordinates of companies and analyzed loan default patterns.

- Regions with **over 50% defaults** were highlighted and compared to regions with **less than 50% defaults**.
  This visualization illustrates geographic trends in loan repayment behavior.



Word Cloud Analysis

For each group (repaid and defaulted), I created word clouds to visualize the most frequent words appearing in company names. This provided insights into potential patterns or trends associated with company characteristics.

Pay Full

CHGOFF

Among companies that repaid loans, words like "medical" and "care" were common, while in those that defaulted, words like "home," "investment," "finance," and "trucking" were more frequent.

Model Selection and finetuning

In this section, I evaluated several models and their performance on the **validation** set after being trained on the **training** set to determine the best-performing model.

The models tested included:

SVC, Logistic Regression, Random Forest, Gradient Boosting and XGBoost.

The **XGBoost model** demonstrated the best performance, achieving:

The results are summarized in the table below.

	AUC	F1	Accuracy
SVC	0.732018	0.623804	0.828239
Logistic	0.748014	0.649068	0.835517
RandomForest	0.889853	0.854495	0.923823
GradientBoost	0.896680	0.860274	0.925764
XGB	0.906839	0.872119	0.931344

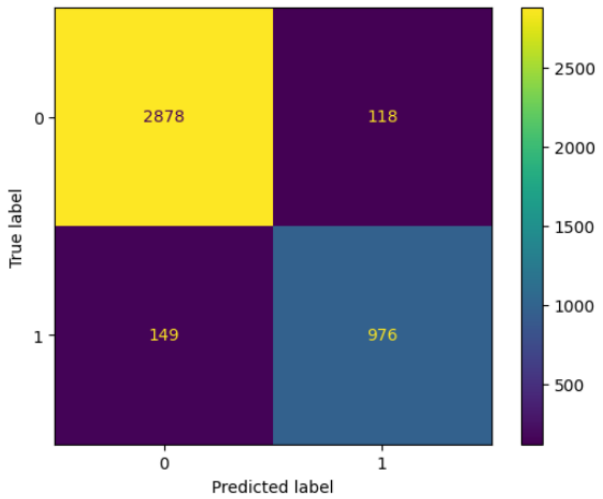
Finally, I performed fine-tuning of the **XGBoost model** using RandomizedSearchCV.
To ensure a more robust evaluation, I implemented a custom scoring function that assessed the model's performance on the **validation set** rather than the data used for training.

The final model's performance metrics on the test population are as follows:

AUC: 0.91

F1: 0.87

Accuracy: 0.93



Confession Matrix

Conclusions

The graph below highlights the most important features and their impact on the likelihood of loan repayment:

1. **Loan Duration:** Longer loan durations are associated with a higher likelihood of repayment.
2. **Loan Year:** Loans issued in more recent years are more likely to be classified as non-repaid.
3. **Loan Amounts:**
    - Higher loan amounts approved by the bank increase the likelihood of repayment.
    - Conversely, higher amounts guaranteed by the SBA correlate with a higher likelihood of non-repayment.
4. **Other Factors:**
    - The **Fed interest rate** and the **sector category** of the borrowing company have relatively minor impacts on loan repayment likelihood.