Predicting Small Business Loan Defaults using Machine Learning Models

Toghrul Rasulov
BUAN 6340

Abstract

The success of small businesses is crucial for a thriving economy, and access to capital through loans plays a significant role in their growth. In this study, we aimed to develop a robust machine learning model to predict loan defaults in the Small Business Administration (SBA) loans dataset obtained from Kaggle. The dataset contains various features such as loan amount, term length, borrower demographics, industry classification, and credit-related information. We employed data cleaning, preprocessing, feature engineering, and feature selection techniques to identify the most relevant features for our analysis. Three different models were tested: logistic regression with grid search cross-validation, LightGBM classifier, and a sequential neural network model with stacked dense layers. The LightGBM classifier achieved the best performance with an AUC score of 0.973 and was selected as the final model. We then used SHAP value analysis to investigate the impact of each feature on the default probability and conducted error analysis for false negatives and false positives. Finally, a scoring function was created that takes raw data, performs cleaning, and generates predictions for new data. The findings from this study can help financial institutions make more informed decisions when lending to small businesses and minimize the risk of defaults.

Introduction

Motivation: Small businesses are a vital part of the economy, and the Small Business Administration (SBA) provides support to these businesses through loans. However, many small businesses face challenges in securing funding due to their size, limited credit history, or other factors. As such, understanding the factors that contribute to loan defaults can help lenders and policymakers provide better support to small businesses.

In this context, we conducted exploratory data analysis on a dataset containing information on SBA loans made to small businesses. The dataset includes various features such as loan amount, industry type, location, and loan terms, as well as an indicator for whether the loan was charged off or paid in full.
Our analysis aimed to identify patterns and relationships between the loan features and loan defaults. We used various visualizations such as bar graphs, heatmaps, and histograms to gain insights into the data. We also performed feature engineering to create new features that could potentially help predict loan defaults.

Based on our analysis, we identified several features that were strongly associated with loan defaults, such as the borrower's credit score, the loan amount, the industry type, and the borrower's location using advanced feature selection techniques. We also found that some features, such as the revolving line of credit and low-doc loan programs, were associated with a higher likelihood of loan default. Additionally, we identified the importance of feature engineering, which allowed us to reduce the number of features while still maintaining good predictive performance.

Overall, our analysis provides insights into the factors that contribute to loan defaults among small businesses and highlights the importance of feature selection and engineering in creating effective predictive models.

Dataset: The dataset used in this project is from the U.S. SBA loan database, which includes historical data from 1987 through 2014 with 23 variables. The dataset has a total of 830 thousand observations. It includes information on whether the loan was paid off in full or if the SBA had to charge off any amount and how much that amount was Each row represents a loan application and includes details about the borrower, loan approval, and loan disbursement. The target variable is 'MIS_Status', which indicates whether the loan was charged off or paid in full. The dataset includes features such as loan amount, loan term, number of employees, industry classification, and loan program. Additionally, the dataset contains information about the bank that approved the loan, the state and city where the borrower is located, and whether the loan is for a new or existing business. The dataset is imbalanced, with a much higher number of approved loans than denied loans. This presents a challenge for building a machine learning model that can accurately predict loan approval or denial.

Data cleaning process: In this project, we performed several cleaning and preprocessing steps on the raw SBA loans dataset to ensure its quality and consistency. The following is a summary of the data cleaning steps:

1. Clean Currency Columns: We first cleaned the currency columns (DisbursementGross, BalanceGross, GrAppv, and SBA_Appv) by removing the '$' and ',' characters and converting the values to the float data type.

2. Clean Binary Columns: The binary columns LowDoc and RevLineCr were cleaned to ensure that they only contain 'Y', 'N', or 'Missing' values. Any value other than 'N' or 'Y', or missing values, were replaced with 'Missing'.
3. Clean Franchise Code Column: The FranchiseCode column was cleaned by converting its data type to 'object' and replacing 0 and 1 with 'Missing'.
4. Clean Urban-Rural Columns: We cleaned the UrbanRural column to reflect three possible values: 'Urban' (1), 'Rural' (2), and 'Missing' (0).
5. Clean NewExist Columns: We also cleaned the NewExist column to represent 'Existing' (1), 'New' (2), and 'Missing' (for any other value).
6. Clean NAICS Column: The NAICS column was cleaned to ensure it contains only 6-digit codes or 'Missing' for undefined values.
7. Clean MIS_Status Column: The MIS_Status column was cleaned by filling missing values with 'P I F' and mapping the values to binary 0 (P I F) and 1 (CHGOFF).
8. Fill Missing Values: Finally, we filled the missing values in the DataFrame with 'Missing' for object columns and 0 for numeric columns.

By performing these data cleaning steps, we ensured that the dataset is in a suitable format and ready for further preprocessing, feature engineering, and modeling. These cleaning steps help improve the quality of the dataset and make it more consistent, which is essential for obtaining reliable results from the analysis and predictive models.

Feature engineering: This is a critical step in our analysis, as it helps transform raw data into more meaningful and informative features that can better represent the underlying patterns in the data. In this project, we performed several feature engineering steps on the cleaned SBA loans dataset to create new features and enhance the existing ones. The following is a summary of the feature engineering steps:

1. Industry: We used the NAICS (industry codes) column to derive the 'Industry' feature. We mapped the NAICS codes to their respective industry sectors to provide a more interpretable representation of the borrowers' industries. This helps us better understand the impact of different industries on loan default rates.
2. SBA_portion: We created a new feature called 'SBA_portion', which represents the proportion of the gross loan amount guaranteed by the SBA. This is calculated by dividing the SBA-backed amount by the total loan amount.This feature helps capture the risk associated with the loan, as higher SBA guarantees may indicate a lower probability of default. However, based on model explainability using SHAP value, we see that most of false negatives have high value of SBA_portion. This showed that the relationship between the feature and the default rate is not always straightforward. One possible explanation for this could be that the SBA guarantees are intended for riskier loans, and that the higher guarantees are being given to businesses that are already at higher risk of default.
3. Same_State: We created a new binary feature 'Same_State', which indicates whether the bank state and the business state are the same. This feature helps us understand if the geographical proximity between the borrower and the lender has an impact on loan default rates.
4. Monthly_amount: We calculated a new feature called 'monthly_amount' by dividing the approved loan amount (GrAppv) by the term length in months (Term). This feature represents the average monthly loan repayment amount and helps us assess the affordability of the loan for the borrower.
5. Binned Numeric Columns (Optional): Optionally, we can create a small number of bins for numerical variables based on quantiles. This step helps simplify the data by grouping similar values together and can potentially improve model performance by reducing the impact of outliers or noise in the data.

By performing these feature engineering steps, we aimed to enhance the dataset with new and informative features that can help improve the performance of our predictive models. These new features help capture essential aspects of the data, such as industry, risk, geographical factors, and loan affordability, which can provide valuable insights into the factors influencing loan default rates.

Findings: The results demonstrated the effectiveness of the LightGBM classifier in achieving an AUC score of 0.973, highlighting its potential for assisting financial institutions in making informed lending decisions to small businesses while minimizing the risk of defaults. Our analysis revealed several key insights into the factors affecting loan default probability:

1. Term length emerged as the most important feature influencing default risk. Longer terms were associated with lower default probabilities, but for very lengthy loans, the default probability is increasing after a certain point. Intuitively, this can be explained by the increased uncertainty associated with longer-term loans, as changes in economic conditions or the business's financial health over a longer period could increase the likelihood of default.
2. Loans where the borrower and lender states were different were more likely to default. This finding suggests that loans involving borrowers and lenders from the same state are less likely to default, possibly due to closer proximity enabling better communication, monitoring, and understanding of local market conditions.

3. Borrowers from California exhibited a different distribution compared to others and were more likely to default. This observation could be influenced by the larger sample size from that state, but it may also suggest regional differences in business operations, industry mix, or economic conditions that could affect loan performance.
4. Loans with missing industry or urban-rural information were more likely to default. This result implies that the lack of specific information about the borrower's industry or location may hinder the lender's ability to accurately assess the risk associated with the loan. Accurate information is crucial for determining the risk profile of a borrower and making well-informed lending decisions.
5. Monthly payment amount (loan amount divided by term length) also contributed to default probability. Higher monthly payments may place more financial strain on small businesses, increasing the likelihood of default as they struggle to meet their debt obligations.
6. Smaller values of the SBA-guaranteed portion were associated with loan defaults. This observation indicates that when the SBA guarantees a larger portion of the loan, the risk of default decreases, likely because the guaranteed portion provides a safety net for lenders and incentivizes them to monitor the loan more closely.

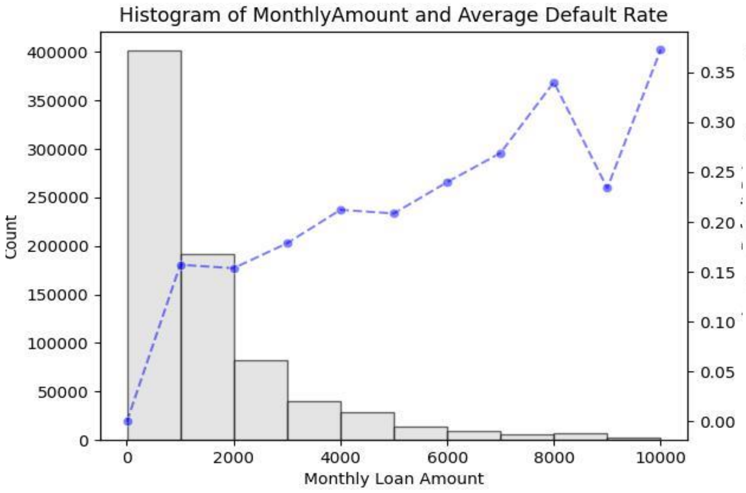
These findings provide valuable insights into the factors that influence loan default probability and have important implications for financial institutions seeking to optimize their risk management strategies. By considering these factors when making lending decisions, financial institutions can minimize the risk of defaults and better support the growth and success of small businesses. Future work could involve refining the model with additional data, exploring other machine learning algorithms, and examining the impact of macroeconomic factors on loan default probabilities.

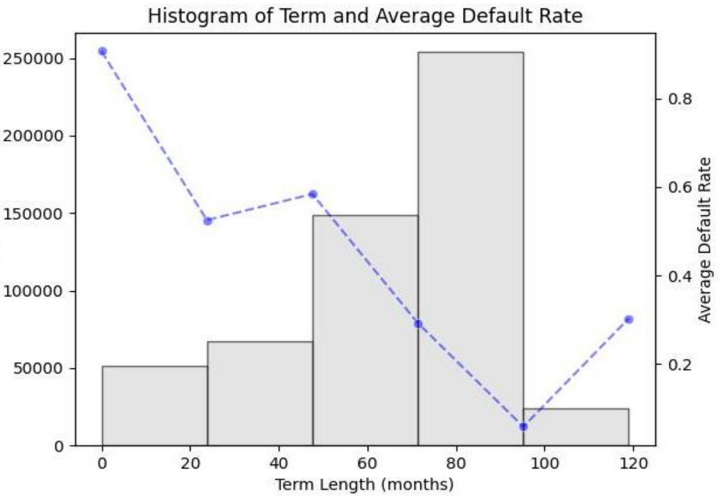
Exploratory Analysis

In this section, we present the findings from our exploratory data analysis, which provides valuable insights into the relationships between various features and loan default rates. The visualizations are grouped into several subsections, each focusing on different aspects of the dataset.

Term Length, Monthly Amount and Default Rates

The above histograms were created to visualize the distribution of Term Length and Monthly Amount, along with their respective average default rates.





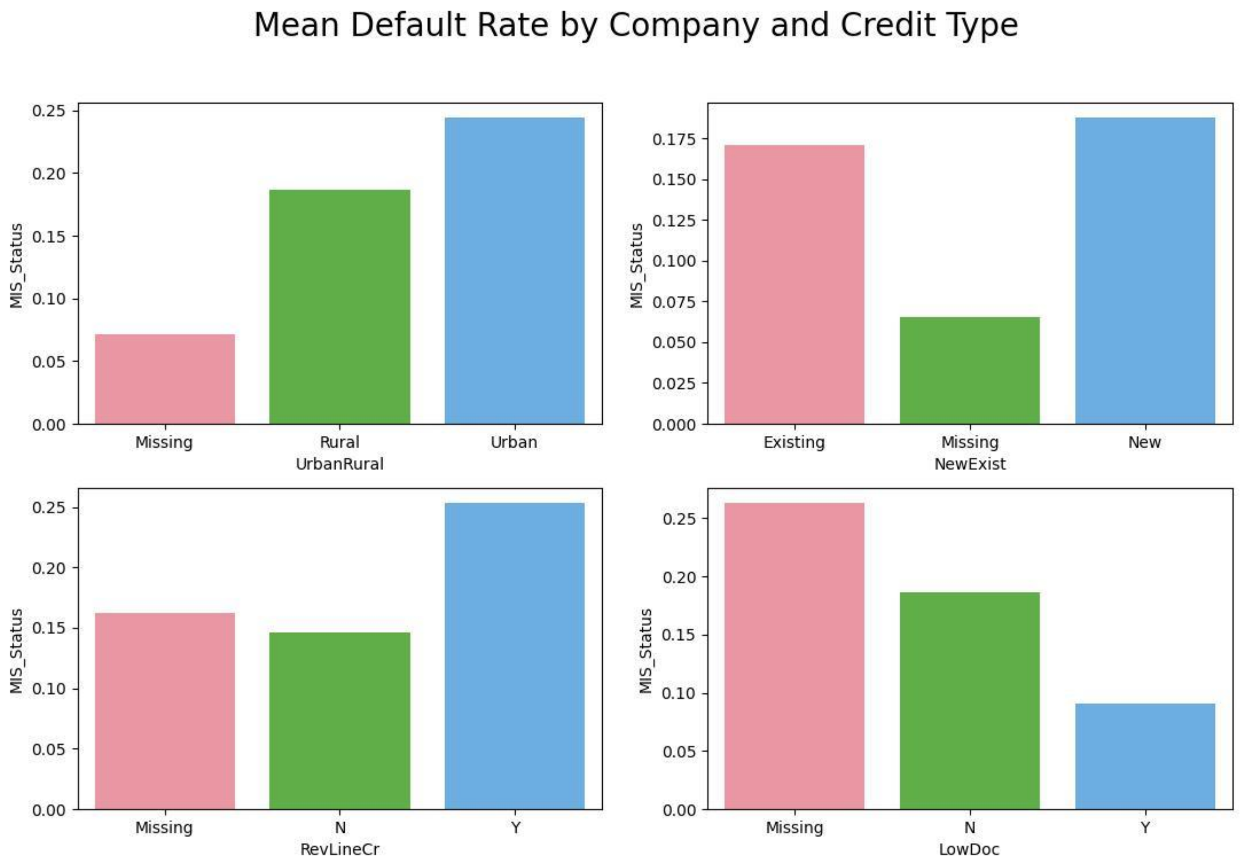
1. Term Length: The histogram reveals a left-skewed distribution, with the highest count of loans (30%) having a term length of 6-8 years. The average default rate exhibits a decreasing trend as the term length increases. However, for loans with a term longer than 8 years, the average default rate starts to increase and oscillate. This could be attributed to the smaller sample size in those bins, which may not provide an accurate representation of the overall trend.
2. The distribution is right-skewed, with the first bin (loan amount up to $1,000) accounting for almost half of the dataset. The average default rate increases as the monthly amount increases, but the relationship appears to be concave, suggesting a diminishing impact after a certain point. Higher loan amounts may be associated with increased financial burden and risk, leading to higher default rates. However, the diminishing impact could be a result of larger, more stable businesses being able to manage higher loan amounts better.

These visualizations suggest that borrowers with longer term lengths and lower monthly amounts may be less likely to default on their loans, although the impact of these factors may be influenced by other variables or specific circumstances.

Loan Program Characteristics and Default Rates

Next, we visualized the mean default rate across different categorical variables. The following are the descriptions and insights obtained from these visualizations:

1. UrbanRural: The bar graph comparing the mean default rate across UrbanRural categories shows that urban areas have a higher default rate (0.24) than rural areas (0.18), while loans with missing UrbanRural information have the lowest default rate (0.05). This suggests that the location of the business might influence the likelihood of loan default, with urban businesses being more likely to default on their loans.
2. NewExist: The bar graph comparing the mean default rate for new and existing businesses reveals that new businesses have a slightly higher default rate (0.18) than existing businesses (0.16), while loans with missing NewExist information have the lowest default rate (0.07). This insight indicates that newer businesses might be riskier in terms of loan repayment, as they may lack the financial stability and experience of established businesses.
3. RevLineCr: The bar graph comparing the mean default rate across RevLineCr categories (Revolving line of credit) shows that loans with a revolving line of credit (Y) have a higher default rate (0.25) than loans without a revolving line of credit (N), which has a default rate of around 0.15. Loans with missing RevLineCr information also have a similar default rate (around 0.15) as those without a revolving line of credit. This suggests that businesses with revolving lines of credit may have a higher probability of defaulting on their loans, possibly due to the added financial burden or mismanagement of credit.
4. LowDoc: The bar graph comparing the mean default rate across LowDoc categories (LowDoc Loan Program) reveals that loans with a LowDoc status (Y) have the lowest default rate (0.1), followed by loans without a LowDoc status (N) with a default rate of 0.17. Loans with missing LowDoc information have the highest default rate (0.25). This indicates that businesses participating in the LowDoc Loan Program are less likely to default on their loans, perhaps because of the program's streamlined application process and smaller loan sizes.



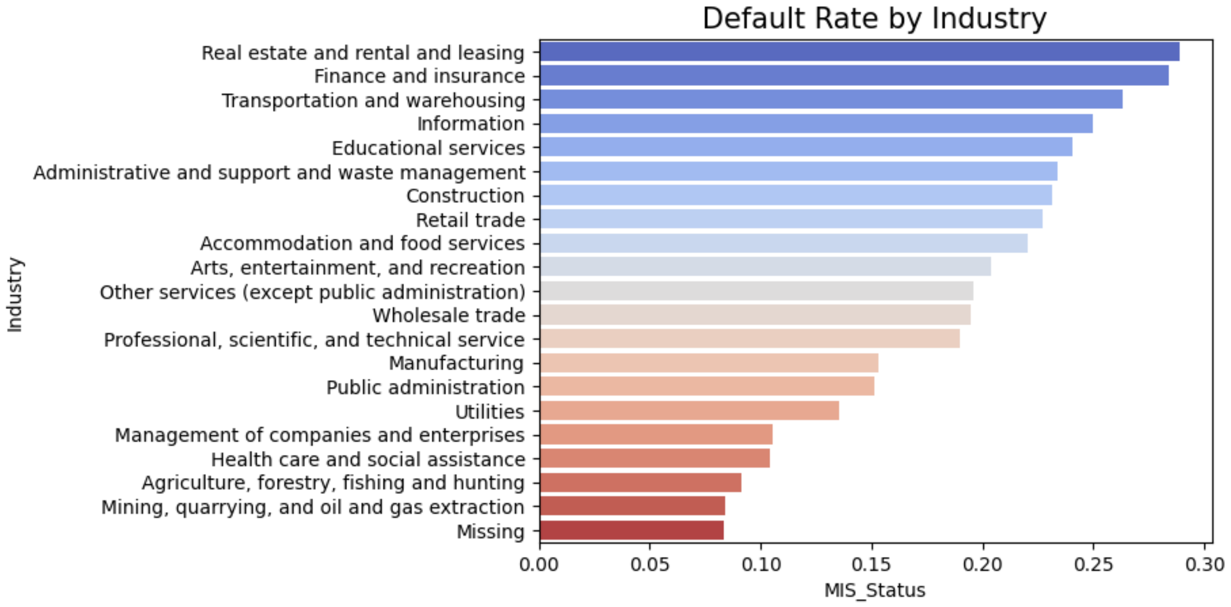
Through these visualizations, we gained valuable insights into the relationships between various categorical features and loan default rates. This exploratory analysis helps inform our feature selection and modeling process by highlighting the potential importance of these variables in predicting loan defaults.

Industry and Default Rates

The bar graph illustrates the average default rate across various industries, revealing deeper insights into the factors that may contribute to the differences in default rates:

1. Administrative and Support and Waste Management and Remediation Services (0.23), Construction (0.23), and Retail Trade (0.23) have the highest average default rates. Businesses in these industries often require significant capital investments and may experience cash flow fluctuations, which could contribute to a higher likelihood of default. Additionally, these sectors are more susceptible to economic downturns and shifts in consumer demand, which can further affect their financial stability.

2. Agriculture, Forestry, Fishing and Hunting (0.09), Mining, Quarrying, and Oil and Gas Extraction (0.08), and the 'Missing' category (0.08) show the lowest average default rates. The lower default rates in these industries might be attributed to several factors, such as government subsidies, long-term contracts, and a lower dependency on consumer demand. These factors can provide a more stable financial environment, making businesses in these sectors less likely to default on their loans.
3. Health Care and Social Assistance (0.1), Management of Companies and Enterprises (0.11), and Utilities (0.14) exhibit moderate default rates. This could be due to the combination of factors such as the necessity of services provided, regulations, and the diverse nature of businesses within these industries. As a result, businesses in these sectors might have different levels of financial stability and risk exposure, leading to a range of default rates.

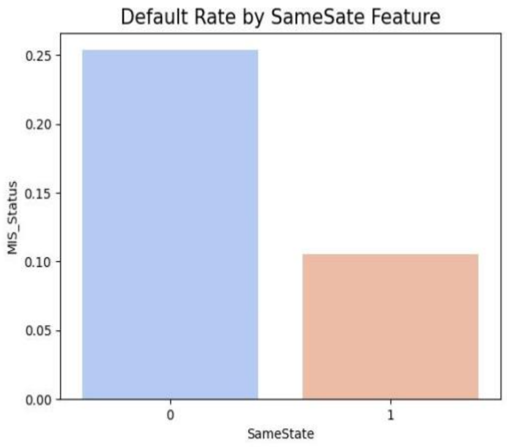


By understanding the industry-specific factors that contribute to loan default rates, lenders can tailor their risk assessments and underwriting processes accordingly. This information can also be incorporated into predictive models to enhance their accuracy and help identify the underlying causes of loan defaults.

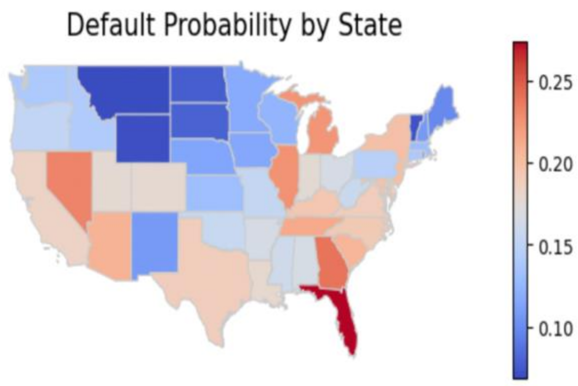
Geographic Factors and Default Rates

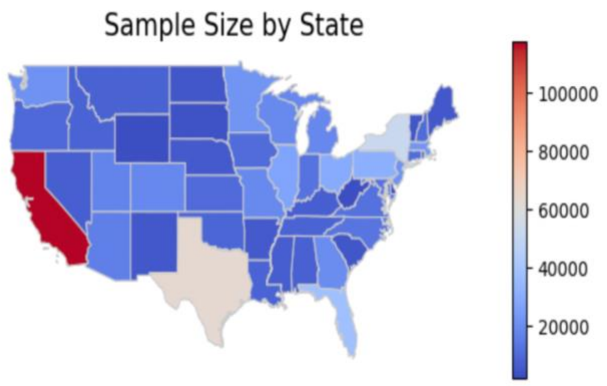
In this section, we analyze the influence of location-related factors on the average default rate, focusing on the engineered feature 'SameState' and a heatmap visualization of US states by average default rate and sample size.

1. SameState: The bar graph shows that when the borrower and lender states are different, the mean default rate is 0.25, while for the same state, it is 0.1. This suggests that loans where the borrower and lender are located in the same state tend to have a lower probability of default compared to loans where the borrower and lender are in different states. This finding suggests that loans involving borrowers and lenders from the same state are less likely to default, possibly due to closer proximity enabling better communication, monitoring, and understanding of local market conditions.



2. Heatmaps of US States: Two side-by-side heatmaps display the average default rate and sample size for each US state.



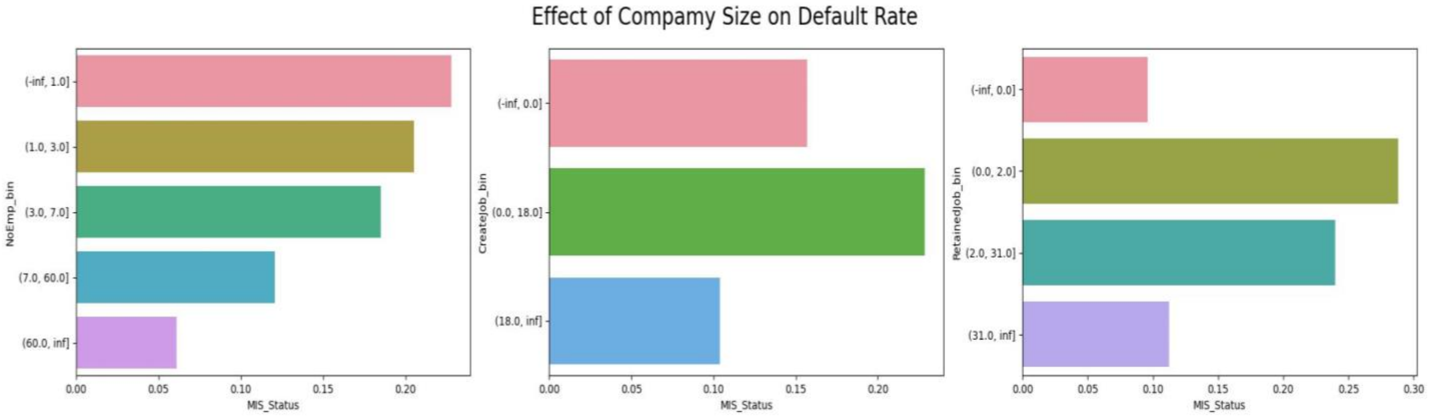


- Average Default Rate: Northern states and some far eastern states (Maine, New Hampshire, Massachusetts) have lower average default rates compared to the southern and central states. Vermont, Montana, and Wyoming have the lowest rates, while Florida, Georgia, and Nevada have the highest rates. The variation in default rates across states may be influenced by factors such as the economic environment, political climate, and the dominant industries in each state. Further research is needed to understand the underlying reasons for these differences.
- Sample Size: California has the largest sample size (around 14% of the dataset), followed by Texas (7% of the total data). The remaining states have similar sample sizes. The distribution of sample sizes across states should be considered when interpreting the findings, as states with smaller sample sizes may not provide an accurate representation of the overall trend.

Company Size and Default Rates

In this set of visualizations, we analyzed the relationship between the mean default rate and various company size features, such as the number of business employees, the number of jobs created, and the number of jobs retained. The following are the descriptions and insights obtained from these bar graphs:

1. Number of Business Employees: The bar graph shows that as the number of business employees increases, the average default rate decreases. This suggests that larger companies may have a lower risk of default, possibly due to factors such as economies of scale, better access to resources, and more stable cash flows.
2. Number of Jobs Created: The mean default rates for the three bins (0 jobs, 0-18 jobs, and more than 18 jobs) are 0.15, 0.23, and 0.1, respectively. While the relationship between the number of jobs created and the default rate is not entirely clear, it is possible that companies creating more jobs may have lower default rates due to better business performance, increased economic activity, or other factors that contribute to financial stability.
3. Number of Jobs Retained: As the number of retained jobs increases (from the second to the fourth bin), the average default rate decreases (0.27, 0.23, 0.11). The first bin, representing missing values, has the lowest rate (0.09). This trend suggests that companies retaining more jobs may have a lower risk of default, possibly due to factors such as employee loyalty, stability, and company performance.



In conclusion, our exploratory data analysis provides insights into the relationships between various features and loan default rates. These findings can help inform the development of predictive models for loan defaults and support decision-making processes in the lending industry.

Machine Learning Models

Feature Selection: We performed feature selection using two rounds of feature importance techniques to select the most important features for our model. The initial dataset had 137 features after one-hot encoding. We fit the LGBMClassifier to the whole dataset and then used feature importance scores and permutation feature importance techniques to shortlist the features. In the first round of feature selection, we started with the entire dataset and obtained feature importance scores based on the LGBMClassifier. We also calculated permutation feature importance scores for each feature. We then selected the most important features based on a combination of both importance measures. In the second round of feature selection, we repeated the process on the shortlisted features from the first round. This helped us further reduce the number of features and we ended up with a final set of 16 features that did not significantly affect model performance.

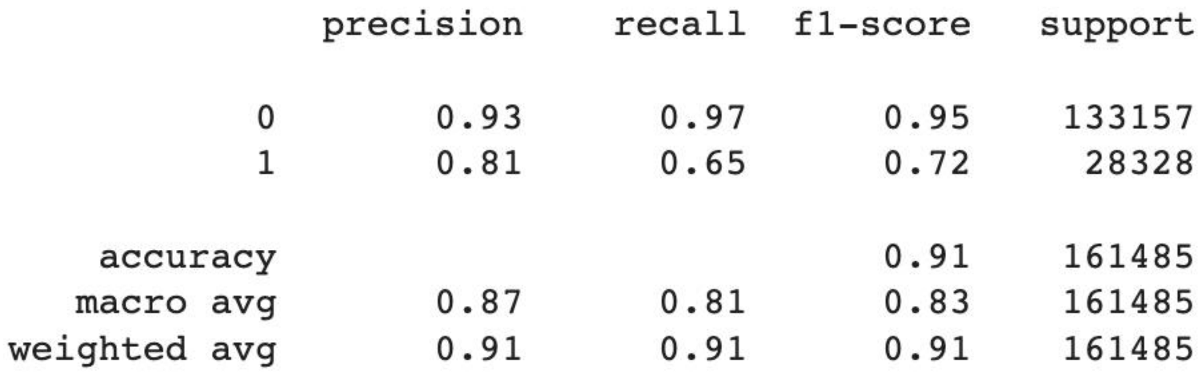
The shortlisted features were: *NoEmp, Industry_Missing. MonthlyAmount, UrbanRural_Missing, RevLineCr_N, State_CA, RetainedJob, RevLineCr_Y, SBA_portion, GrAppv, SameState, LowDoc_Y,* CreateJob, Term, SBA_Appv DisbursementGross

These features were selected based on their high importance scores and their relevance to the problem statement. The features were also consistent with our exploratory data analysis, where we found that loan-term, credit, company, industry, and location features were significant predictors of loan default. Since some of the features were correlated with each other, we excluded some of the variables found during exploratory analysis.

Model Selection and Results: We attempted to train and evaluate three different models on our preprocessed dataset using selected features.

1. The first was logistic regression, which performed poorly with an ROC-AUC score of 0.66 on the test data. The poor performance of logistic regression may be attributed to the complexity of the dataset and the categorical nature of the data. Logistic regression assumes a linear relationship between the input features and the target variable, which does hold in our dataset based on our exploratory analysis and visualizations.

2. The second model was a deep learning model, specifically a feedforward neural network. Despite adjusting the architecture and parameters, the deep learning model performed moderate, achieving an ROC-AUC score of 0.93. Deep learning models are not very effective for our problem, since they require large amounts of data to train effectively, and our dataset may not have been sufficient to support the complexity of the model. However, there are techniques like embeddings that can be used to help deep learning models learn hidden characteristics of categorical features such as borrower states or cities which may explain economic environment, political climate, and demographic characteristics that may affect the success of the small businesses. Below, you can see the classification report based on the deep learning model predictions of the test data.

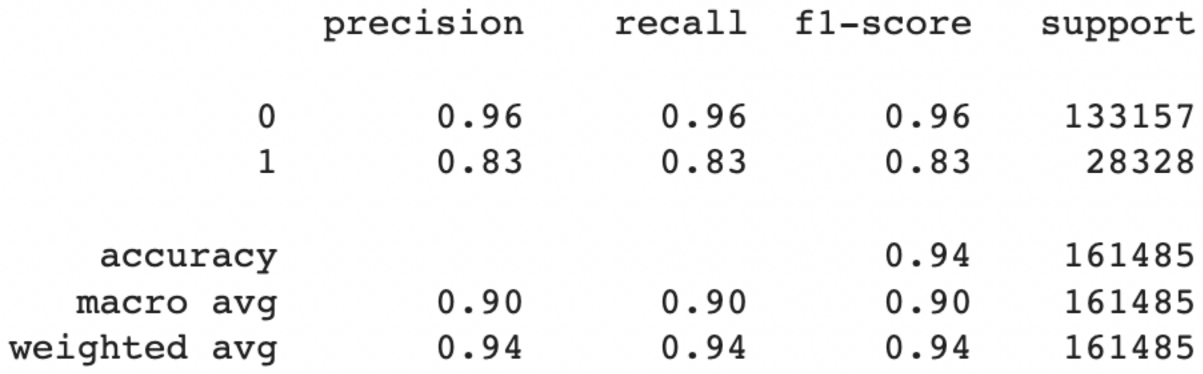


```
              precision    recall  f1-score   support

           0       0.93      0.97      0.95    133157
           1       0.81      0.65      0.72     28328

    accuracy                           0.91    161485
   macro avg       0.87      0.81      0.83    161485
weighted avg       0.91      0.91      0.91    161485
```

3. Finally, we used LightGBM, which performed the best with an ROC-AUC score of 0.97. LightGBM, a gradient boosting decision tree model, was able to perform well due to its ability to handle complex nonlinear relationships between features and the target variable, as well as its efficient computation and scalability. Additionally, LightGBM is capable of handling imbalanced datasets, which is important in our case where the default class is significantly smaller than the non-default class. The below classification report displays the performance of the model on the test data using the best threshold value calculated based on macro average f-1 score.



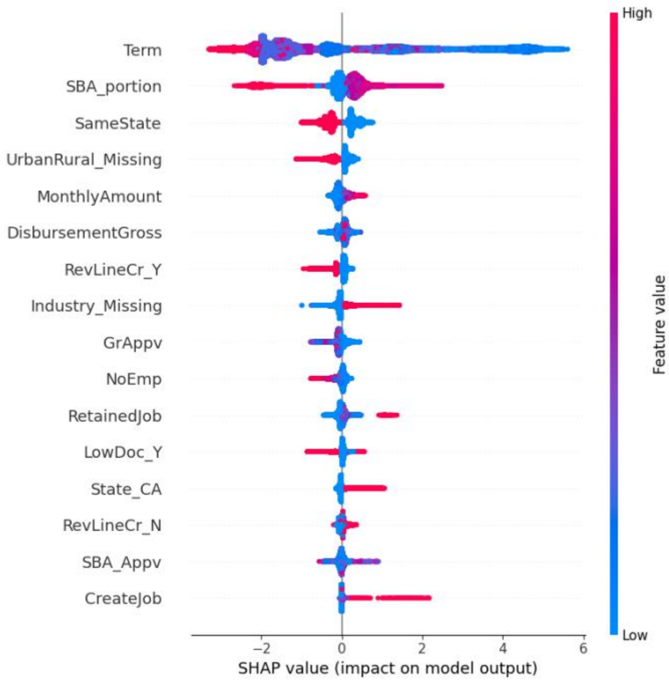
```
              precision    recall  f1-score   support

           0       0.96      0.96      0.96    133157
           1       0.83      0.83      0.83     28328

    accuracy                           0.94    161485
   macro avg       0.90      0.90      0.90    161485
weighted avg       0.94      0.94      0.94    161485
```

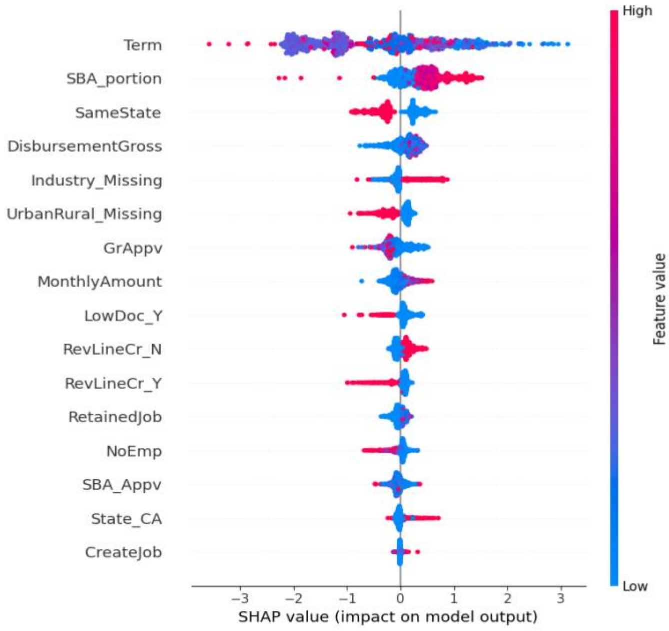
Model Explainability and Error Analysis: After training our final model, we performed feature importance analysis using SHAP (SHapley Additive exPlanations) values. SHAP values provide local explanations for each sample in the dataset, by calculating the contribution of each feature to the model output.

We first calculated the SHAP value for each observation and created summary plots to explore the overall relationship between the selected features and the model output. This allowed us to see which features had the highest impact on the model's decision-making process, how higher and lower values affect the predictions. The top features were *Term, SBA_portion, and SameState, MontlyAmount* which is consistent with our earlier exploratory analysis. Negative SHAP values in the x axis mean that the feature has a negative impact on the model output (decreases the predicted probability of the target class), while positive SHAP values mean that the feature has a positive impact on the model output (increases the predicted probability of the target class).

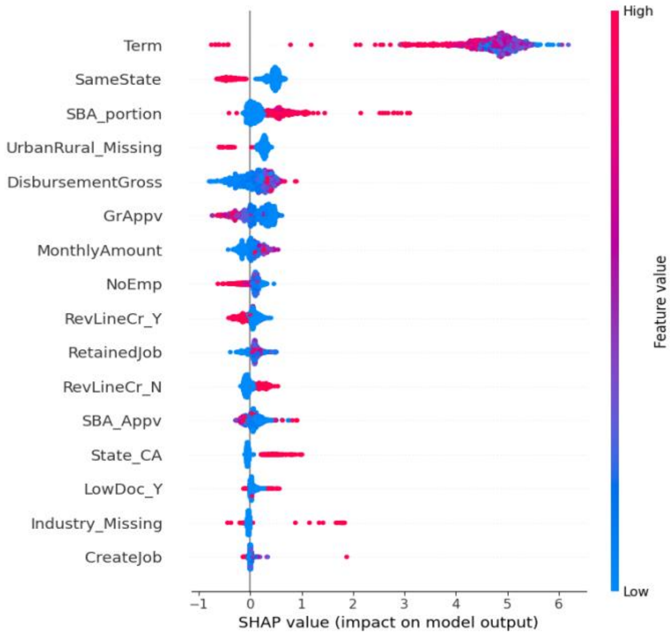


Additionally, we used a SHAP dependence plot to explore the relationship between features and the different type of errors. These plots showed how changes in features impacted false negatives and false positives. To do this, we first calculated the prediction error for each sample and rank them in descending error. The predictions with very high error (residual) values represent the false negatives, and the ones with smaller values represent False Positives.

False Negative (class 1, predicted 0): We first see that false negatives generally have shorter *Term* value, which therefore is not very helpful by itself to explain the error. We further found that the relationship between the SBA_*portion* feature and the default rate is not as straightforward as we initially thought. Specifically, SHAP values for the false negatives showed that higher values of *SBA_portion* (see the plot on the right) are contributing positively to probabilities of default, which is the opposite of what we initially hypothesized. One possible explanation for this could be that the SBA guarantees are intended for riskier loans, and that the higher guarantees are being given to businesses that are already at higher risk of default.



False Positives (class 0, predicted 1): From the plot given on the right, we can see that there are many “paid off” loans with shorter term length predicted as ‘default’ class.One possible explanation for why false positives have low term values is that the short term loans may be less risky, so they are more likely to be approved and have a lower probability of default. However, in our model, the presence of other risk factors may have caused some of these short-term loans to be classified as high risk and incorrectly labeled as false positives. Additionally, it's possible that some of the false positives were due to errors or inaccuracies in the data. To address this issue, we may need to explore other features that could help distinguish between low-risk and high-risk loans with short term lengths, and/or improve the accuracy of the data to reduce errors and inconsistencies. We could also consider modifying our model or adjusting the classification threshold to better balance the trade-off between false positives and false negatives.



In summary, using SHAP values allowed us to gain insights into which features had the highest impact on the model output, and how changes in those features affected the model's predictions, and further understand root causes of errors. This information can be used to further finetune the model by addressing issues causing the most errors, improve model interpretability and inform decision making in the lending process.

Conclusion

The main objective of this project was to build an accurate model to predict loan default. We began by conducting an exploratory analysis of the loan dataset and identified key features that influenced default rates such as credit score, industry type, location, loan amount, and SBA portion. We found that SBA portion did not have a clear relationship with default rates. After selecting relevant features through a two-stage feature selection process, we trained several models including logistic regression, LightGBM, and deep learning models. LightGBM proved to be the best performing model, achieving an AUC score of 0.974. Model is able to "paid of" loans with 96% and "defaulted" loans with 83% accuracy. We then used SHAP values to explain the predictions made by the LightGBM model. We found that loan portion guaranteed by SBA, term length, and state had the most significant impact on the model's output. Furthermore, we also observed that the effects of SBA portion values term lengths contributed are not very straightforward and need more detailed hyphothesis tests.

In conclusion, our selected model was highly accurate in predicting loan defaults, and our analysis helped to identify key features that influence loan defaults. Our findings could be useful for lenders in identifying risky loans and potentially mitigating losses.

References:

1. Li, Min, Amy Mickel, and Stanley Taylor. ""Should This Loan be Approved or Denied?": A Large Dataset with Class Assignment Guidelines." *Journal of Statistics Education* 26.1 (2018): 55-66.

2. Kaggle dataset by MIRBEK TOKTOGARAEV. https://www.kaggle.com/datasets/mirbektoktogaraev/should-this-loan-be-approved-or-denied, 2019.

3. ChatGPT, 2023.