SI 618 Final Project Report

Yaoqi Liao
School of Information

I. Motivation and Summary

Since I used to work as a credit analyst at a bank, I was in charge of assessing home loans. I used to check customers' background and their financial documents manually. The whole assessment process was quite tedious. So, I want to apply some data analytical skills and machine learning models to improve the assessment process. Suppose we could implement some machine learning models to predict the final loan status accurately. In that case, banks may reduce their bad debt ratio, and customers can have a reduced turnaround time. Hence the general question for this project is how we could use businesses' information to determine whether we should approve their loans?

To find out the answer to the general question, I will focus on three major areas:

1. Explore the relationship between the features of the businesses and the final loan status
   - Geographical plot to show the default rate in each state in the U.S.
   - Regression plot to show the relationship between default rate and median household income by state
   - Bar plot to show the rank of default rate in industries
   - Box plot to show the relationship between loan term and the final loan status
   - Bar plot to show the relationship between company size and the final loan status
   - Bar plot to show the relationship between business operating area and the final loan status
   - Correlation Graph to show the relationships among the features
2. How did the Great Recession influence small businesses?
   - Time Series Analysis: From 2000 to 2014, plot a line chart to show the default trend.
   - The line chart shows the number of small businesses approved for loans from 2000 to 2014.
3. Which machine learning model gives us the best prediction result?
   - Implement different machine learning models: Random Forest, Boosting, Logistic Regression
   - Plot the variable importance

II. Data Sources

I have combined two datasets for this project.

The first dataset is downloaded from Kaggle, which has a CSV format. Here is the link to the dataset: https://www.kaggle.com/mirbektoktogaraev/should-this-loan-be-approved-or-denied. This dataset has been collected from the U.S. Small Business Administration (SBA). To support entrepreneurs and small businesses, the SBA provides a government-backed guarantee on the loan when the business owners borrow money

from banks. So, the SBA has collected information from borrowers to assess whether they would like to provide the guarantee. The dataset contains 899,164 observations and 27 variables in total. Some significant variables include borrowers' states, industry codes, the amount of the money approved by banks, total loan terms, loan types, loan status (note: only two classes exist in the dataset: paid off or charged off). Since this data set has an extensive time range (from 1962 to 2014), I would like to select data from 2004 to 2014 for analysis because I want to focus on detecting the default pattern in the most recent decade. Even though I removed the data from 1966 to 2003 and only kept the data from 2004 to 2014, I found the total number of observations is around 300,000, which is still sufficient for our analysis.

The second dataset is downloaded from the U.S. census website, and it has an Excel Spreadsheet format. Here is the link to the data: https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-income-households.html (Table H-8). The dataset showed the median household income by state in 2020. I will use the income information as an indicator of the wealthiness of the state. Then I will find out whether there is any relationship between the wealthiness of the state and the default rate of the businesses operating in that state.

III. Data Preprocessing

For the loan dataset:

- Step 1: I dropped the loans approved before 2004 because I am only interested in the most recent ten years of data.
- Step 2: Converted a few variables to the correct data types.
- Step 3: Extracted the charged-off year from the charged-off date
- Step 4: Since I am only focusing on the features of the business to figure out their default probabilities, I dropped some variables which contain the banks' information. Specifically, I dropped the loan application number, city, zip, bank state, bank, approval date(kept the approval year), business name (not doing an NLP analysis in this project, hence dropped this variable too), FranchiseCode, charged off date (only interested in the charged-off year)
- Step 5: Dropped all the missing values because the number of missing values is relatively small compared with the total number of records.

For the income dataset:

- Step 1: Converted the Excel Spreadsheet format to CSV format.
- Step 2: A few empty lines are in the CSV file due to file conversion. Hence I dropped the blank lines in pandas to get a cleaned dataset.
- Step 3: Created a new column named "state_abbr," which contains the state abbreviations

IV. Methods and Analysis

1. Explore the relationship between the features of the businesses and the final loan status

State Variable: I calculated the default rate for each state and plotted a U.S. map (see figure 4.1.1 below) to show that certain states have much higher default rates than others. Hence the state variable should be a crucial indicator.

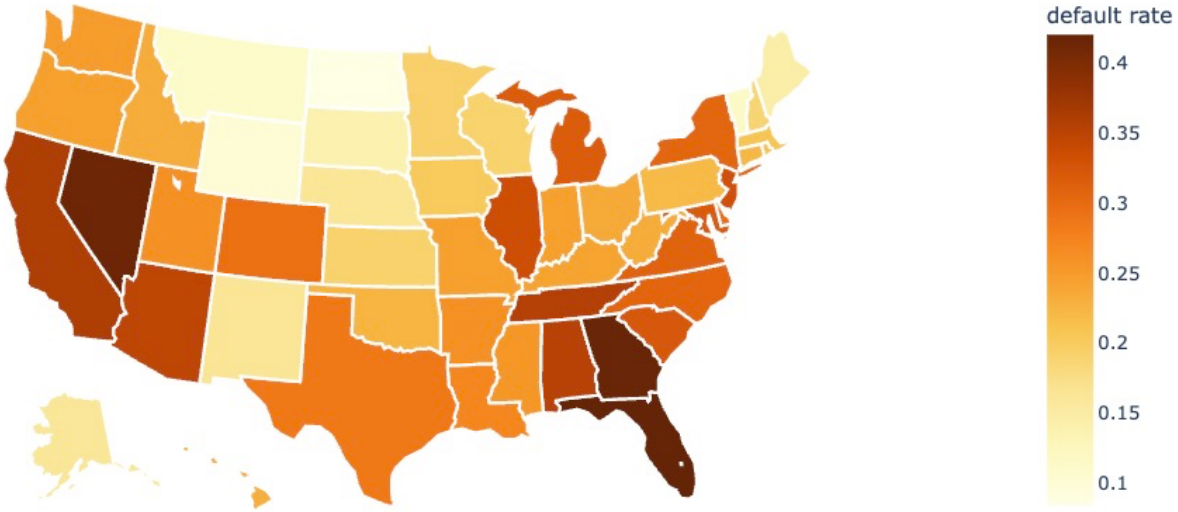

Figure 4.1.1 Default rate by state

Median Household Income by State: After calculating the default rate for each state, I combined the default rate and the median household income by state as a new table. Then I made a regression plot to check the income and the default rate relationship. From figure 4.1.2 below, we cannot see a strong relationship between income and the default rate.

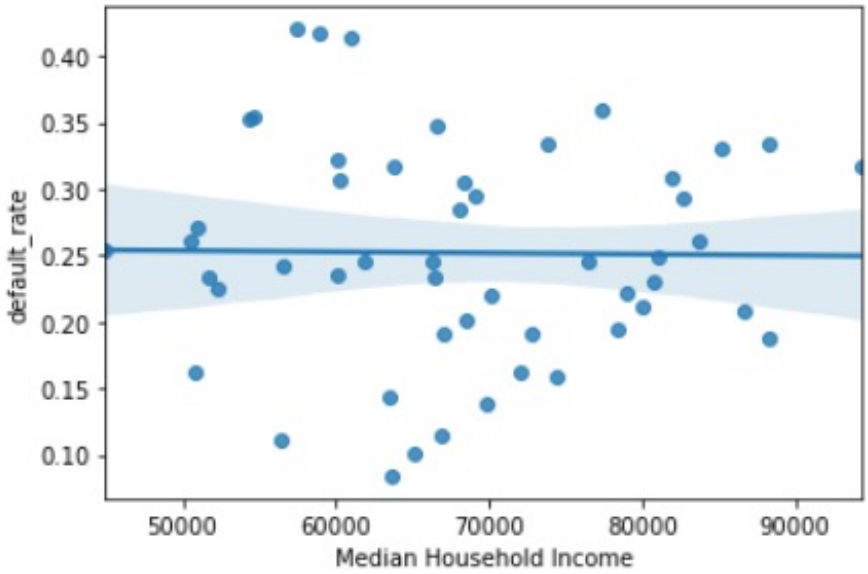

Figure 4.1.2 Default rate by state vs. Median Household Income by state

Industry Variable: I created this variable based on one of the original variables called "NAICS" (i.e., North American Industry Classification System). Since the first two digits of the NAICS classification represent the economic sector, I decided to use the first two digits to create a new "industry" variable and replace the original variable. We draw a bar plot based

on the industry variable (see figure 4.1.3 below) to show that certain industries have higher default rates than others.

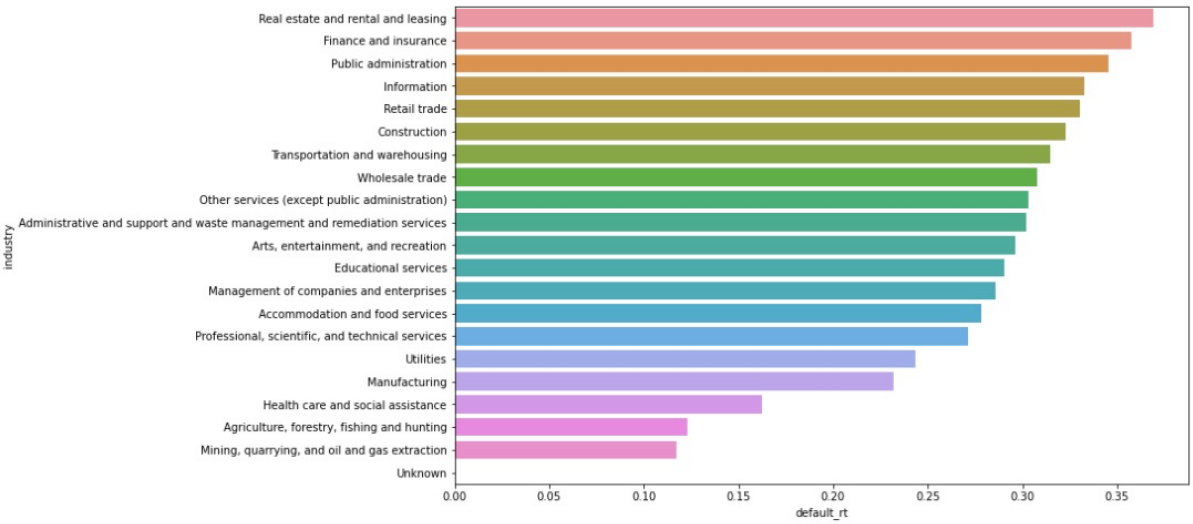

Figure 4.1.3 Default rate by industry

Term_Length Variable: Similar to the industry variable, the term_length variable is created based on the original term variable to indicate the loan term length. If the loan term is less than or equal to 36 months, it is a short-term loan. If the loan term is longer than 36 months but less than or equal to 120 months, it is a medium-term loan. For the loan term that is longer than 120 months, it is a long-term loan. From figure 4.1.4, we can see that the short-term loan has the highest default rate. From figure 4.1.5, we can see most of the charged-off loans have a loan term of fewer than 100 months.

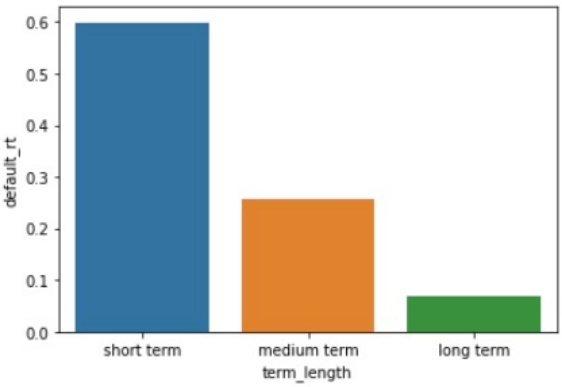

Figure 4.1.4 Term length vs default rate

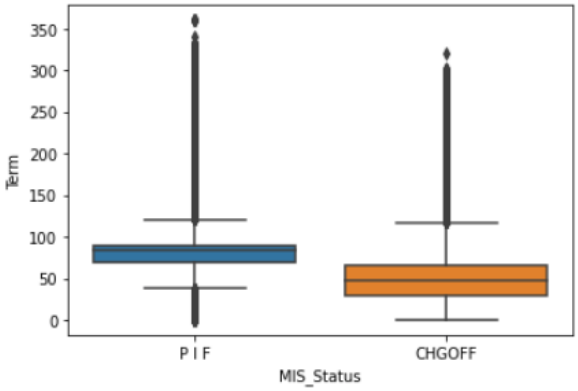

Figure 4.1.5 Loan term vs loan status

Business Size: Based on the original variable "NoEmp," which indicates the number of employees in the company, I classified the business size into two categories: Small('S') and extra small('XS'). If the number of employees is less than or equal to 10, it is an 'XS' company; otherwise, it is an 'S' company. Comparing the two types of businesses, we can see that the default rate in the 'XS' company is almost twice as high as that of the 'S' company. This result indicates that the company size (i.e., the 'NoEmp' variable in the original dataset) might be a good variable to predict the loan status.

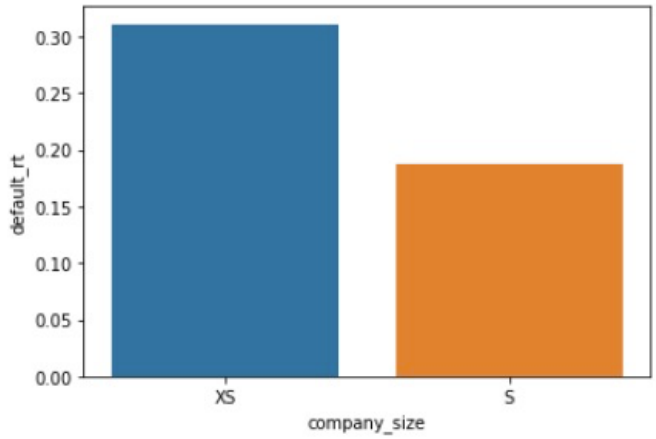

Figure 4.1.6 Business size vs default rate

UrbanRural Variable: The "UrbanRural" variable indicates whether the business operates in an urban or rural area. If the variable's value is 1, the company operates in a metropolitan area. If the variable's value is 2, it operates in a rural area. If the variable's value is 0, it indicates that the business location is undefined. After doing the exploratory analysis (see figure 4.1.7 below), I found that the default rate is higher in urban areas.

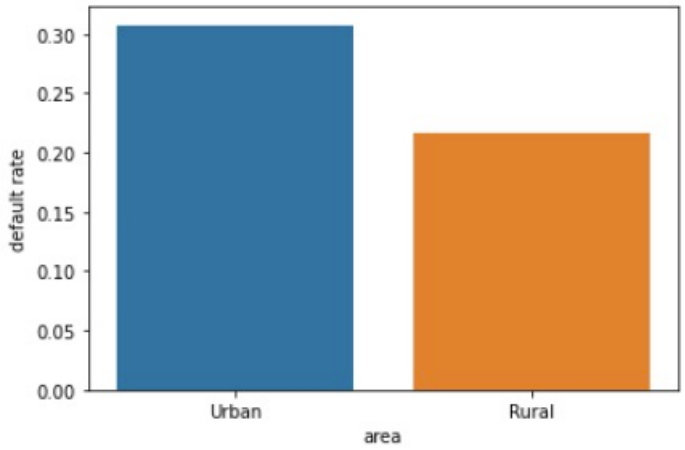

Figure 4.1.7.Business Operating Area vs Default Rate

Revolving line of credit: A revolving line of credit is a type of loan offered by banks. Based on my domain knowledge in the banking industry, the revolving line of credit usually has a higher credit risk than traditional loans. With a revolving line of credit, as soon as the debt is repaid, the user can borrow up to her credit limit again without going through another loan approval process. So the credit assessment for a revolving line of credit usually takes more information from the borrowers and will have deeper analysis in the business. The result shown in the figure 4.1.8 is surprising because the default rate of the revolving line of credit is lower than the traditional loans.

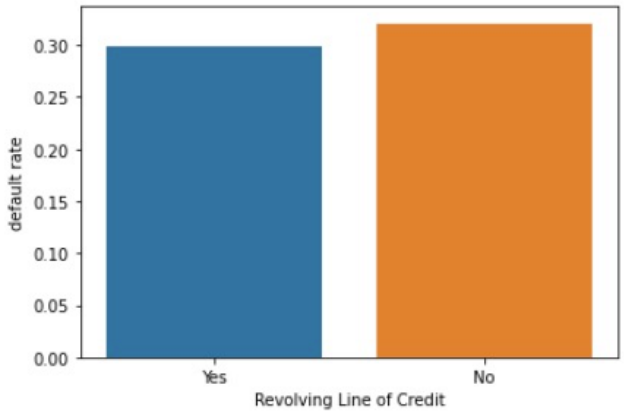

Figure 4.1.8 RLC vs. default rate

Correlation Heatmap: To see the relationships among different variables, I draw a correlation heatmap. From the heatmap, we can see that a few variables are highly correlated with each other. Specifically, the variables “GrAppv”(the gross amount of loan approved by the bank) and “SBA_Appv” (SBA’s guaranteed amount of approved loan) are positively correlated. “GrAppv” and “DisbursementGross” also have a strong positive correlation of 0.95. “SBA_Appv” and “DisbursementGross” has a correlation of 0.93. The three variables mentioned above are about the loan amount. Hence it makes sense that they have a very high correlation. Therefore, in the next step of fitting models, I will use “GrAppv” as one of my independent variables and drop the other two.

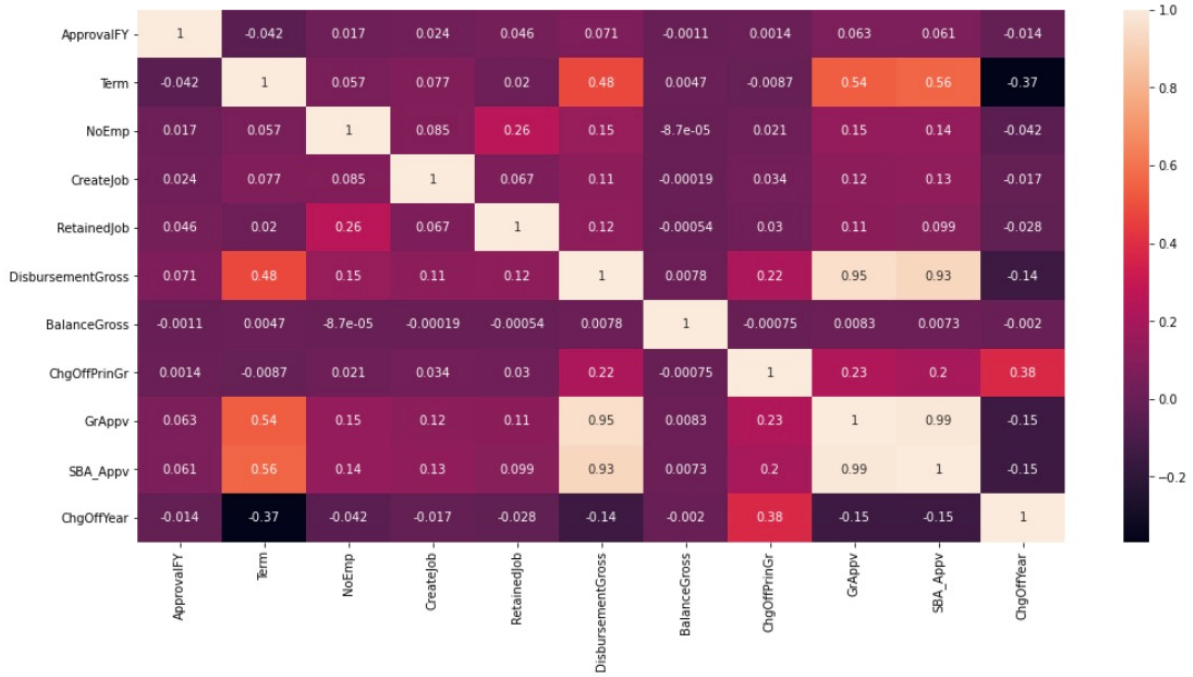

Figure 4.1.9 Correlation Heatmap

2. How did the Great Recession influence small businesses?

To see the great recession's impact, I first plot the trend of the default rate from 2004 to 2014. And then, I plot the trend line of the number of applications from 2004 to 2014. As shown in the line plots below, the default rate dramatically increased from 2004 to

2008, but the number of approved applications (figure 4.2.2) decreased significantly from 2007 to 2010. The results here reflect the tough time during the great recession.

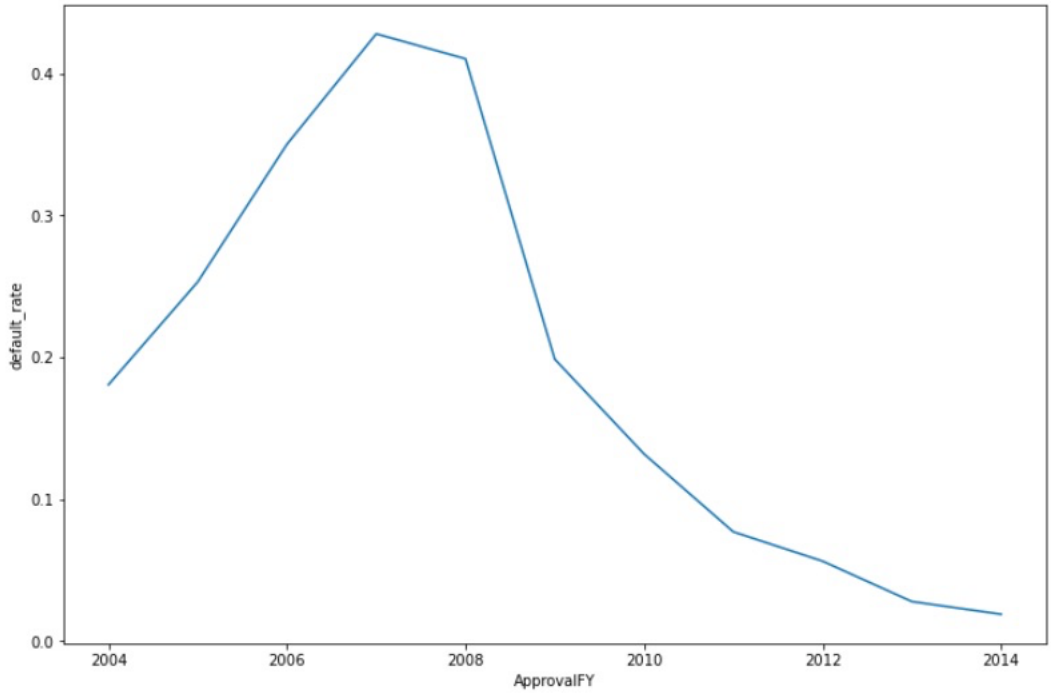

Figure 4.2.1 Default rate through year 2004 to 2014

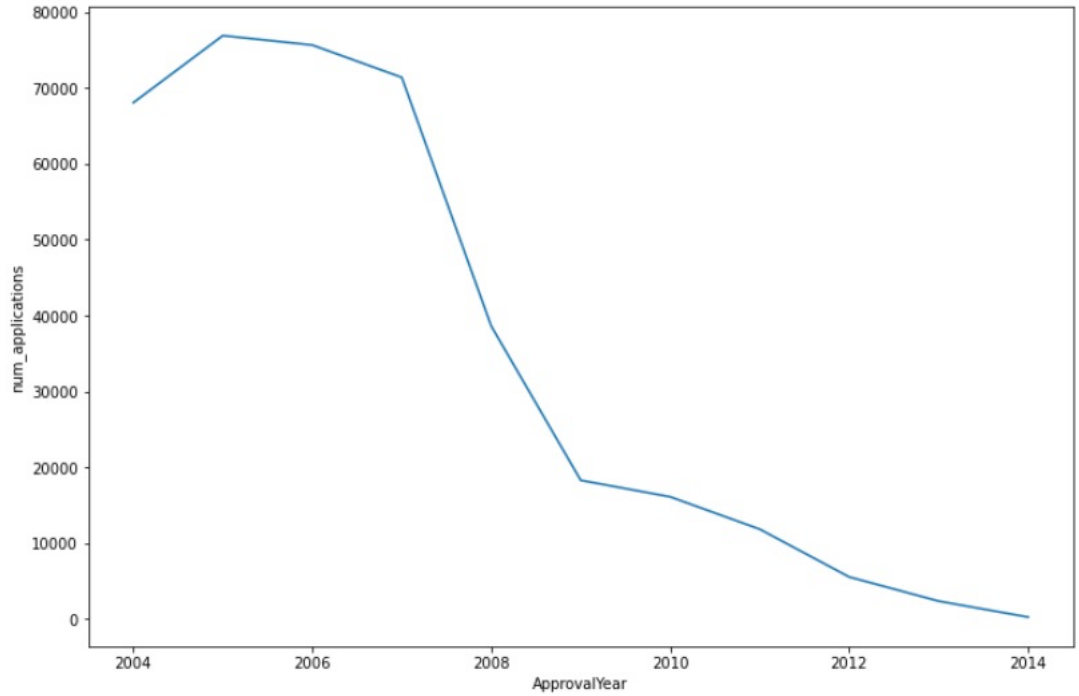

Figure 4.2.2 Number of applications through year 2004 to 2014

3. Which machine learning model gives us the best prediction result?

Firstly, I picked the features based on my domain knowledge and the correlations among the variables. Then I split the dataset into 70% training and 30% testing. I tried three different models(i.e., random forest, logistic regression, and AdaBoost) for this dataset. Since the dataset is highly imbalanced, I used the AUC score and the balanced accuracy

score as measuring metrics for my models. The prediction results are shown in the table below. The random forest model is the best in terms of prediction.

Models	AUC score	Balanced accuracy score
Random Forest	0.97	0.91
Logistic Regression	0.82	0.65
Adaboost	0.95	0.88

Table 4.3.1 Prediction results for different models

To understand which variable is most predictive, I draw a feature importance plot. (See figure 4.3.2) As a result, the term variable is the most important one, followed by 'GrAppv.'

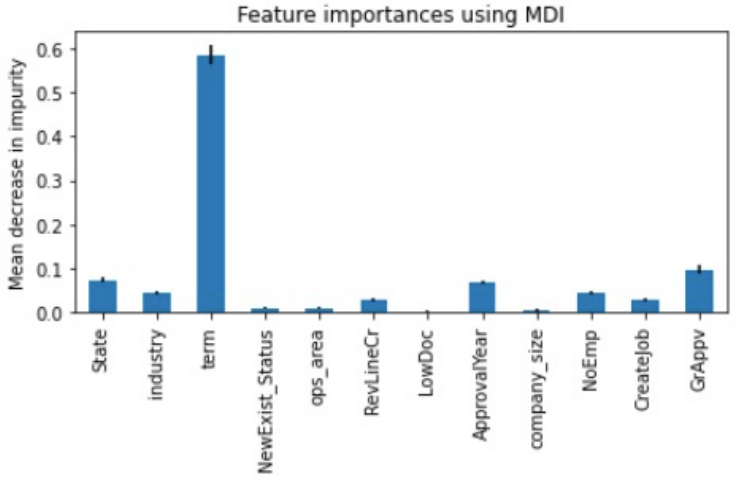

Figure 4.3.2 Feature importance

V. Challenges and Discussion

Overall, this project is interesting, but I encountered challenges while I was doing the data exploration. First, the dataset has a high dimension (more than 20 variables), and I need to determine which variables might be helpful in the models. So I did a lot of exploratory data analysis. In addition to the EDA, I also used my domain knowledge in the banking industry to interpret the variables. Secondly, the dataset is highly imbalanced; the regular accuracy score is not a good metric to measure how well the model fits the data. Hence, I used the AUC and the balanced accuracy scores to measure the results. But if I have more time, I would like to explore the oversampling method to fit different models and see how the models perform under the oversampling condition.