Analyzing Credit Risk Using Regression Models

Jay Suresh Singhvi, Benjamin Correia

2024-05-14

1.0 Introduction

1.1 Project Description In this project we used machine learning to analyze a loan default dataset and evaluate the greatest predictors of lenders defaulting on their loans. Using these predictors, we trained a logistic regression model to predict a loan applicant's likelihood to default. We also used a random forest model to analyze the most significant predictors of loan defaults.

1.2 Background When creditors (money lenders like banks) are considering who to lend money to, they must do so cautiously. Haphazardly offering loans to whoever requests them is not great business practice, as many debtors (money borrowers, in this case people) are not fully equipped to pay back money that is lent to them with interest.

In order to maximize the probability of debtors paying them back, creditors must be able to predict the probability of loan applicants to pay them back using financial data.

The goal of this analysis was to use a data set of debtors' financial data and loan statuses to find out which predictors creditors should take into account when deciding on who to lend money to.

2.0 Data Description

The dataset we used is called the Credit Risk Dataset, obtained from Kaggle. It contained 32,581 rows of 12 variables. It's important to note that the dataset was a simulation of credit bureau data made for educational purposes. A quick analysis of these variables will follow:

person_age - numeric: Borrower's Age in years

person_income - numeric: Borrower's income in dollars

person_home_ownership - categorical: Borrower's home ownership status

person_emp_length - numeric: Borrower's length of most recent employment in years

loan_intent - categorical: Borrower's reason for taking out the loan.

loan_grade - categorical: loan grade

loan_amnt - numeric: loan amount in dollars

loan_int_rate - numeric: loan interest rate

loan_status - categorical: levels 1 & 0. 0 - Borrower Did not default 1 - Borrower defaulted

loan_percent_income - numeric - calculated as loan_amnt/person_income

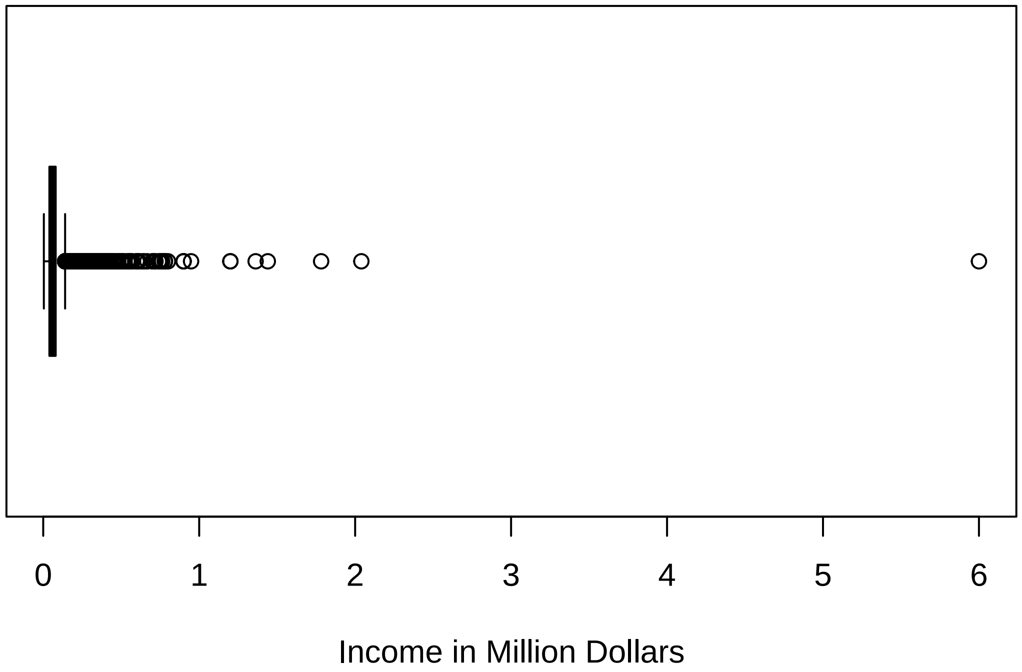
cb_person_default_on_file - categorical: Y if person has a loan default on file, N otherwise

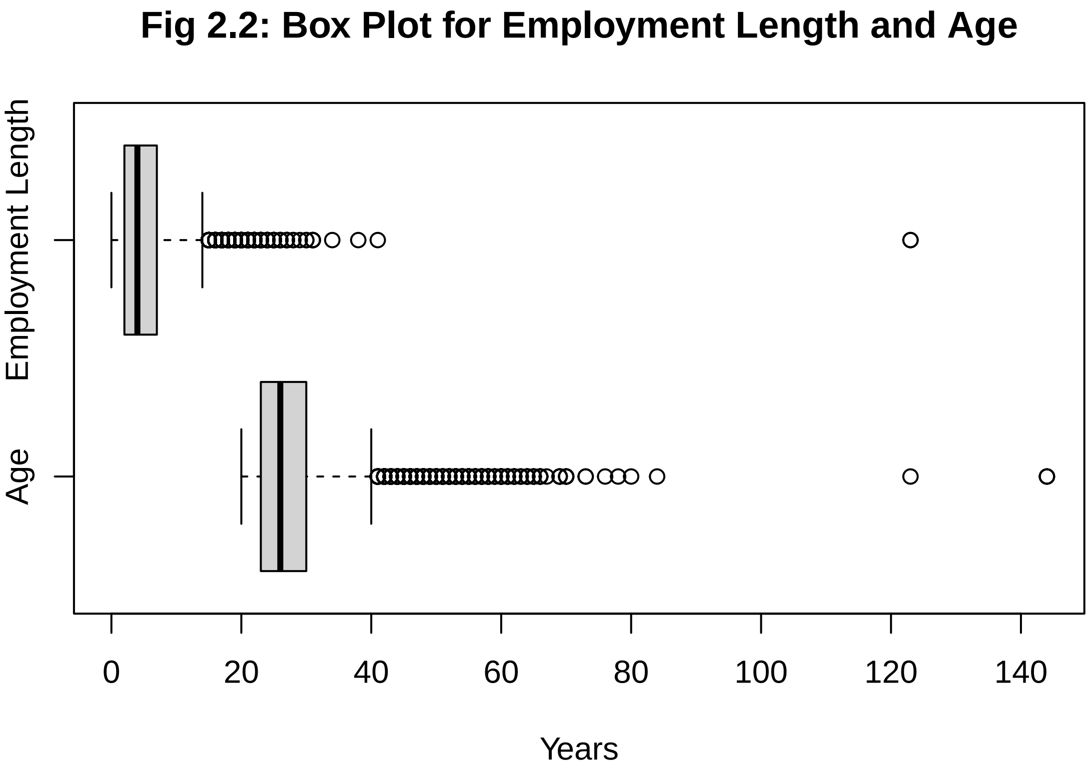
cb_person_cred_hist_length - Length of borrower's credit history in years

We started with 32,581 rows of data before cleaning. First, we removed any rows that contained an empty value. This left us with 28,638 observations.

Just by a glance of the data person_age, person_emp_length, and person_income had some hefty outliers. The box plots below display how much variance existed in the data.

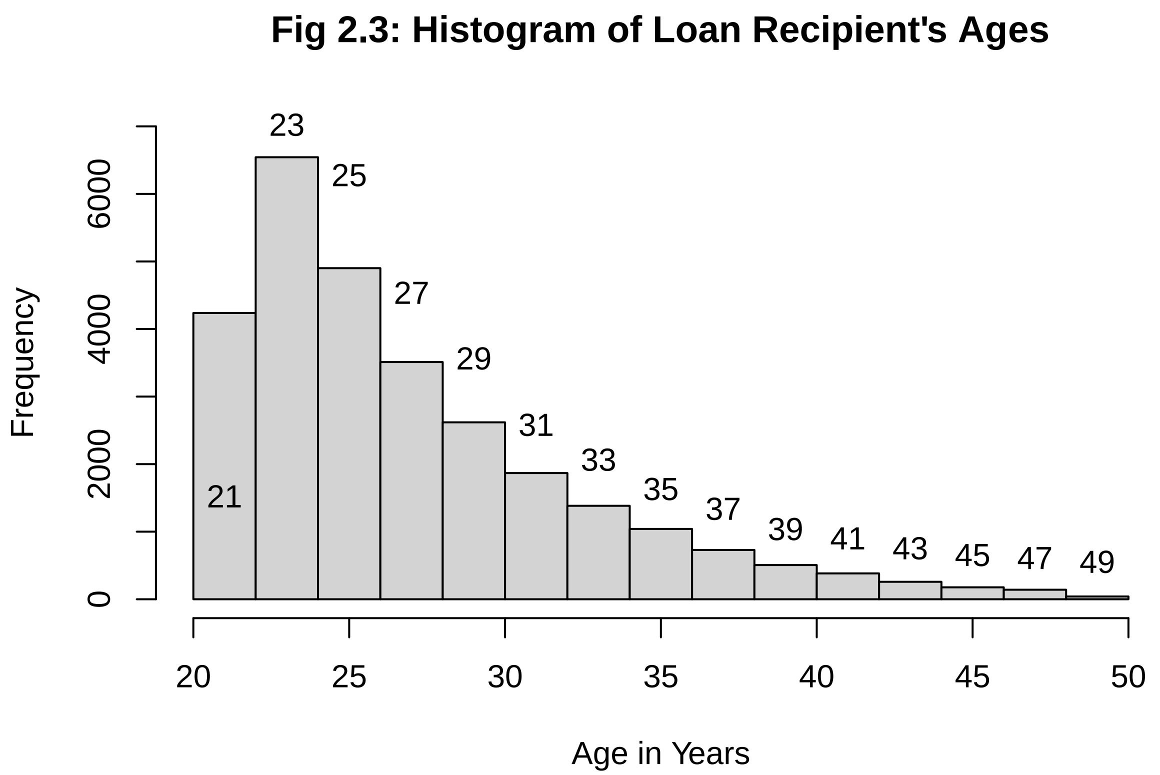
Fig 2.1: Box Plot for Income





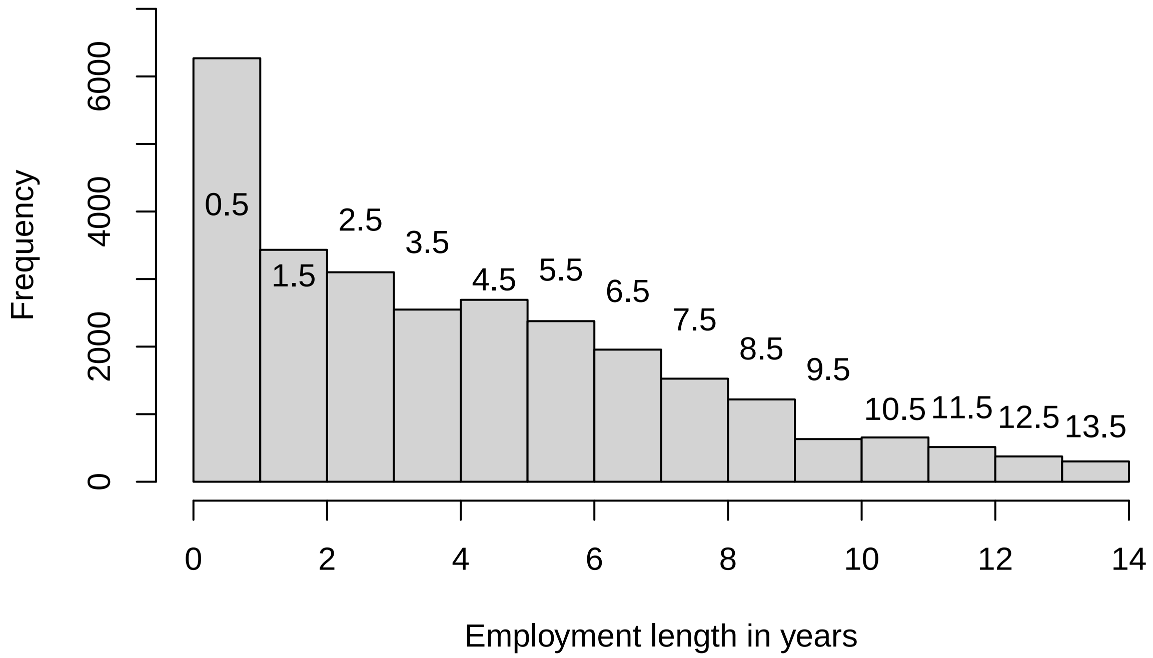
A typical standard for removing outliers is removing any values over 1.5 interquartile ranges above Q3 or below Q1.

As for age, we wanted to remove the incredibly large outliers because we simply didn't have enough data on elderly applicants to make sufficient conclusions about them, and some of the ages were outright unrealistic in Fig 2.2. Some of these outliers, however, were representative of a true variance in the ages of the population. In fact, over 1000 observations exist within the 40-50 y/o range. We used a cutoff of fifty years old as a maximum age to remove these outliers present in Fig 2.2. 28340 remain.



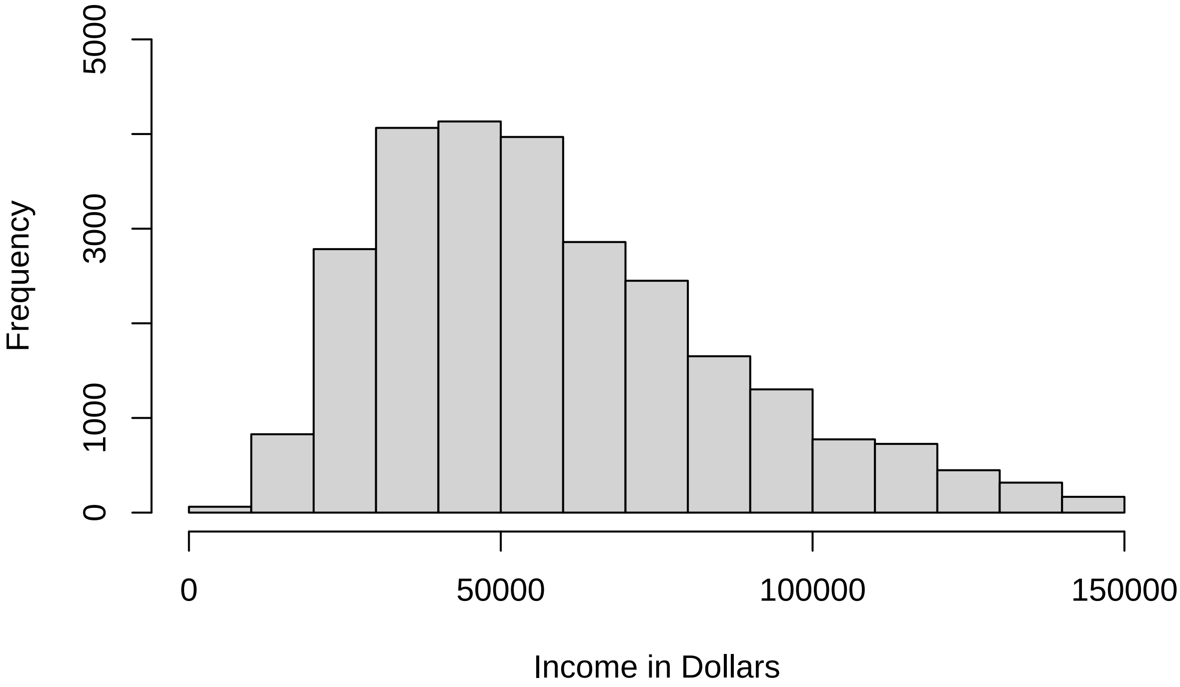
Employment length is another variable for which it probably would not be great practice to indiscriminately remove all of the outliers, as there can be a great degree of natural variance. By making a boxplot of the data the massive outliers are clear in Fig 2.2. We elected to ultimately remove those observations above fifteen. 27599 remain

Fig 2.4: Histogram of Loan Recipient's Length of Current Employme



Income is, of course, the most likely of our variables to vary greatly. First we removed the greatest few outliers that we could see from Fig 2.1. Afterwards, making a labeled histogram showed us that income is relatively normally distributed, skewed to the right, and has a mean around the 50,000 mark. There is a sharp dropoff in the number of observations with income above 150,000, which should be a reasonable point upon which we stopped considering measurements to base our model upon and that is shown in Fig 2.5. 26536 remain

Fig 2.5: Histogram of Loan Recipient's Income



Like most of our numeric data, Loan Amount is normally distributed and skewed to the left. There is a hard dropoff in the number of loans above 25,000, so we naturally removed those from the dataset and plot a histogram shown in Fig 2.6. 25701 remain.

Fig 2.6: Histogram of Loan Amount

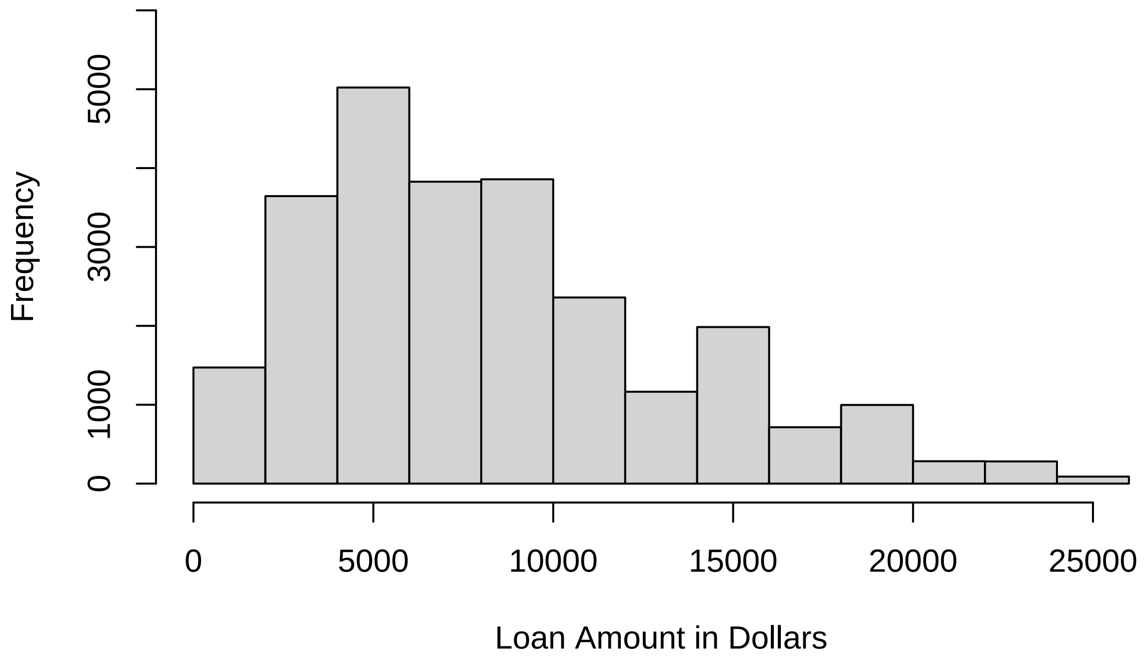


Table 1 displays the levels of the predictor variable, Loan_Status, along with the frequencies. From this, we computed the imbalance in data and compared it to this table:

Degree of Imbalance Proportion of Minority Class None >40% Mild 20-40% Moderate 1-20% Extreme <1%

After computation, the imbalance in the data was 21% which, according to the table is categorized as mild.

Table 1: Loan_Status levels in Dataset

Loan_Status	Frequency
non-default	20147
default	5554

After cleaning our data we were left with 25,701 observations. 6,880 observations was been removed from our data set. Now that we cleaned our data, we were ready to begin modeling it.

3.0 Analysis

First, we partitioned our data into training an testing sets. 75% of our data was used for training which equals to 19,277 observations and remaining was testing set of 6424 observations.

We trained a logistic regression model using our training data. To start, we used the formula 'loan_status ~ . -loan_grade' This formula indicates that we used all of our variables except for loan grade as independent variables to predict loan_status, our dependent variable. We elected not to use loan_grade as a predictor, as it is a grading system lenders often use to categorize loans by risk. Although it is a great predictor of risk, it is one that was not standardized and was the outcome of statistical analysis on credit risk that has already been carried out. Consider this model as LR1.

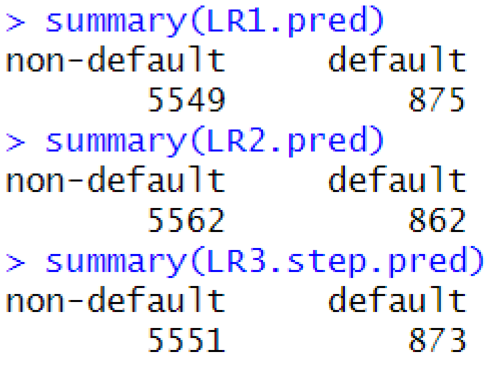
Using predict() function, we generated predictions using our linear model. We passed in the testing data to create predictions of applicant's default status using our logistic regression model. Afterwards, we created a confusion matrix to compare the predictions to the actual outcomes.

LR1 model performed with 84.92% accuracy. Unfortunately, our sensitivity was 46.61% which is very low. Sensitivity is a measure of how accurately our model predicts true positives. In order to decrease risk, our main goal for our model's sensitivity was to be very high, as to prevent lending to borrowers who were expected not to default, but do.

The first thing we did to improve our model was remove insignificant predictors which was found by calling summary() on model LR1. Age, Income, employment length, credit history length, and default history had p-values above 0.05, so we removed these and then trained another multiple linear regression model on the same data. Consider this model LR2

Model LR2 had accuracy of 84.74%, and a slightly lower sensitivity of 45.75%. For our purposes, this was not an acceptable trade-off. Then we used step wise regression for selecting our predictors and then predict the loan's applicant likelihood to default.

Using stepwise regression, we got the function loan_status ~ person_home_ownership + loan_intent + loan_amnt + loan_int_rate + loan_percent_income + cb_person_cred_hist_length. This means that home ownership, loan intent, loan amount, interest rate, credit/income ratio, and credit history length are considered the most significant predictors of loan defaults according to our method. This model had a slightly lower sensitivity than the initial model, but the decrease in performance was so minute that we considered this one more trustworthy, as it worked on less parameters. Consider this model LR3

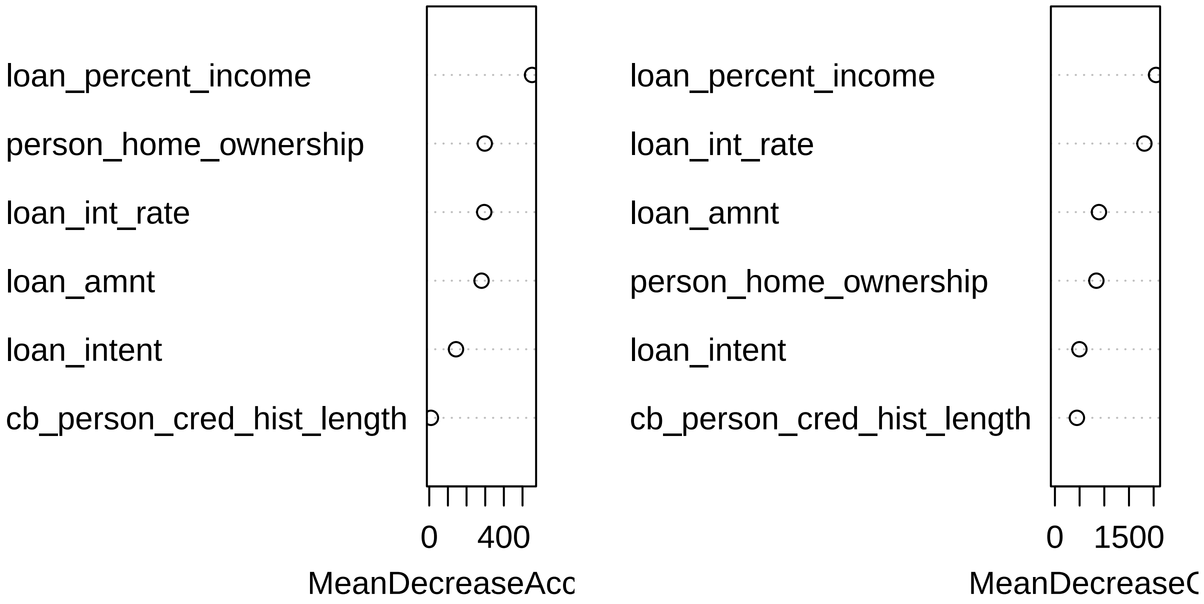


```
> summary(LR1.pred)
non-default     default
       5549         875
> summary(LR2.pred)
non-default     default
       5562         862
> summary(LR3.step.pred)
non-default     default
       5551         873
```

This shows the summary of classification of test data after applying all 3 models through which we clearly saw the split in the dataset and likelihood of person getting defaulted.

After analyzing different logistic regression models and selecting the model generated by stepwise regression (LR3), we used random forest on LR3 model to see the most significant predictors of loan defaults.

Fig 2.7: Random Forest on LR3



The plot from random forest suggested that loan_percent_income, loan_int rate and person_home_ownership were the most significant predictors of loan defaults among the predictors used in the formula returned by model LR3 which were person_home_ownership, loan_intent, loan_amnt, loan_int_rate, loan_percent_income, cb_person_cred_hist_length

4.0 Conclusions

The results of our testing continuosly showed that home ownership, loan intent, loan amount, interest rate, loan__percent__income, and credit history length are the most significant predictors of loan defaults.

The most surprising result to us was that having a default in one's credit history is a very weak predictor, with a p-value of over 0.8.

In the future, we intend to build more precise regression models, use more statistical methods to compare our models, and possibly use clustering to find out if there are trends in the types of debtors who default on their loans.

Because we measure loan__percent__income as a function of loan__amnt and person__income, it may be necessary to build our models taking that into account.

5.0 References

https://www.kaggle.com/datasets/laotse/credit-risk-dataset/discussion