A Study of Default: Lending Club

Pirro Prifti, Michael Oppong, Nana Yaw Bawiah, Madison Guerinot, Yichu Chen

December 6, 2024

Contents

1 Introduction

1.1 The Data

	Loan Amt.	Funded Amt.	Term	Int. Rate	Installment	Grade	Annual Income
1	3600.00	3600.00	36 months	13.99	123.03	C	55000.00
2	24700.00	24700.00	36 months	11.99	820.28	C	65000.00
3	20000.00	20000.00	60 months	10.78	432.66	B	63000.00
4	10400.00	10400.00	60 months	22.45	289.91	F	104433.00
5	11950.00	11950.00	36 months	13.44	405.18	C	34000.00
6	20000.00	20000.00	36 months	9.17	637.58	B	180000.00

Every loan between a lender and borrower, begins with a question: Will this borrower pay me back? But how can a lender know with confidence which factors to consider when evaluating a given borrower? To help shed light on this question, we analyzed loan information from Lending Club. Lending Club is one of the largest and most prominent peer to peer lending services. The data set we obtained spans from Lending Club's inception in 2007 to the second quarter of 2018. Throughout these 11 years Lending Club accumulated 2.260701×10^6 loans on its books, each with 151 different attributes. The number of loans and attributes allowed us to gain a clear picture of how different factors could contribute to the likelihood of default.

An initial approximation of the most important variables in predicting default, may include annual income and the loan amount. At first glance, these factors could be heavily influential in whether a borrower can pay back their loan. Similarly, a high loan amount could impede a person's ability to pay over time. With a data set with so many attributes, it was important to gain a holistic understanding of the variables and to ensure an unbiased assessment. The data set included continuous and factor variables. Some important attributes in the loan data include the loan amount, funded amount, interest rate, and annual income of the borrower. While these variables are continuous, others were factor variables. Key factor variables include the term of the loan (given in months), loan status, and grade. The loan grades range from A to G and describe the quality of the loan with A being the best. The table above shows a sample of these features as they appear in the data set.

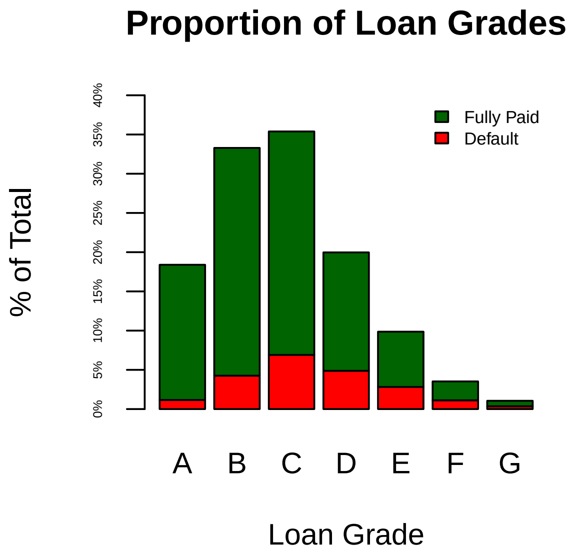
1.2 The Project

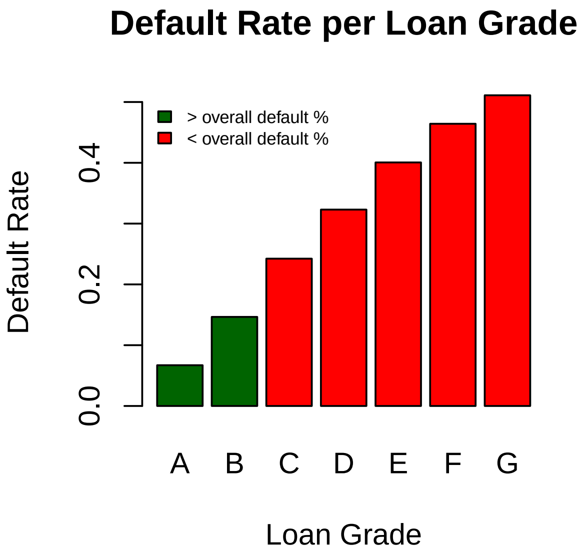
The main insights we will try to draw from this data are the key indicators and attributes leading to default of a given loan. The sheer volume of loans and attributes provided by the Lending Club data set enables us to conduct a robust investigation of default. Further, using classification modeling techniques we can attempt to try and predict this phenomena.

2 Visualizations

The goal of this report is to develop an understanding of defaults for Lending Club. To be able to predict defaults, we must first understand the types of loans given at Lending Club. As a unique peer to peer lending style, we must understand if there are any important differences between the Lending Club borrowers and borrowers in other contexts. This analysis is necessary to contextualize our default predictions and conclusions. The first step in our analysis is to visualize the loans based on key attributes, such as loan grades, interest rates, amounts, and purpose of the loans.

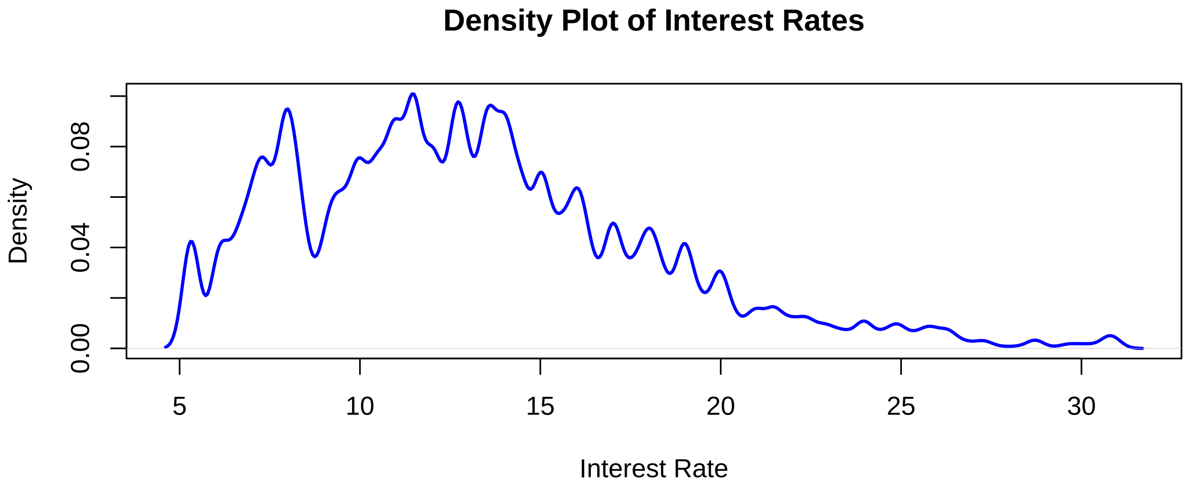
2.1 Bar plot of Loan Grades





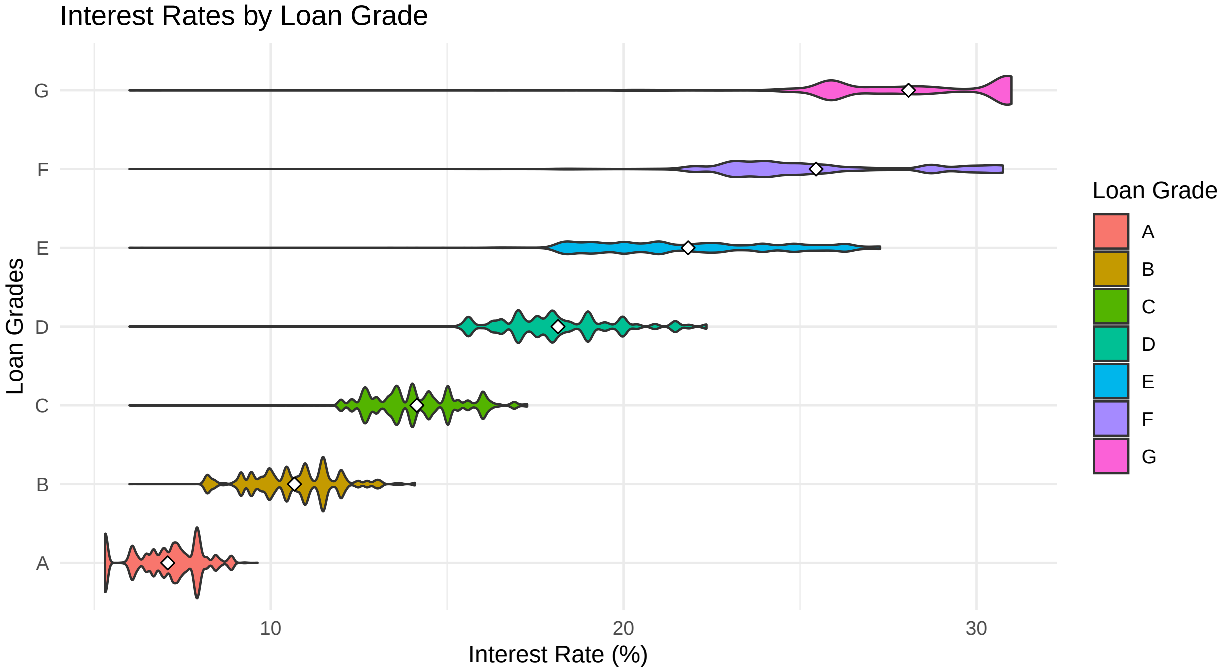
With this bar plot we can gain a better idea of the credit composition of the loans on the books. The Proportion of Loan Grades shows how many loans were within each category of loan grades as a percentage. We can clearly see that the majority of loans (around 55 percent) are ”sub prime” or ”junk” (anything lower than a rating of B). The red portion of each bar shows the proportion of borrowers who defaulted. It is interesting to note that the relative percentage of default is fairly low. The Default Rate per Loan Grade bar plot compares the default rate of each loan grade to the overall default rate of 21.5 percent. The green bars, loan grades A and B, have default rates less than 21.5 percent. The red bars, loan grades C through G, have higher rates of default than the entire population. These graphs clearly show a correlation between the grade of the loan and the likelihood of default, with poorer grades having higher rates of default.

2.2 Density of Interest Rates



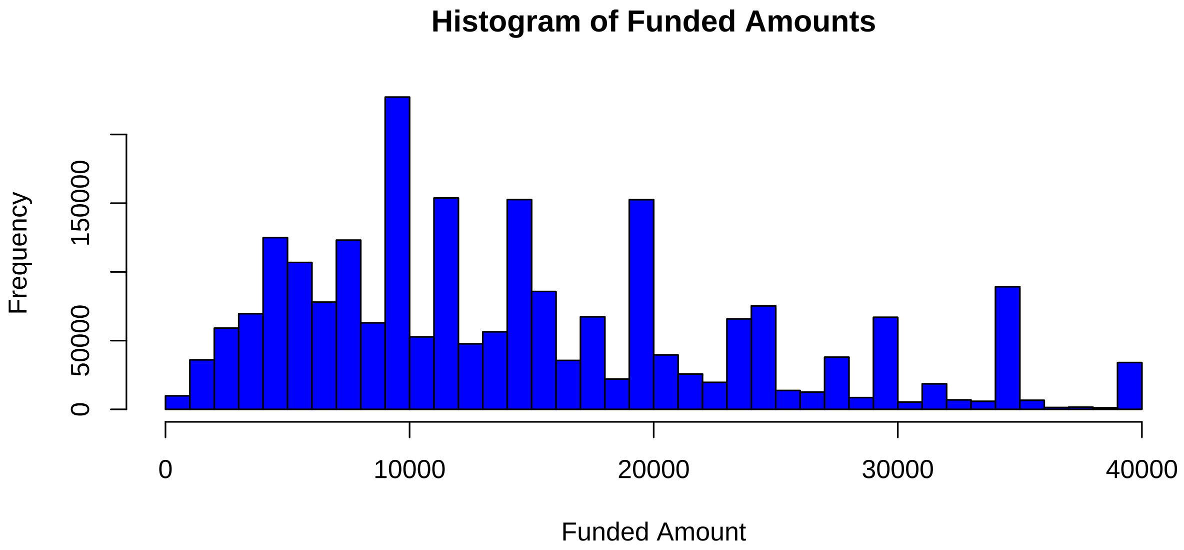
This density plot visualization reveals how common certain interest rates are among the borrowers in the Lending Club. This provides insights into the overall landscape of the lending club's peer lending services. The plot shows that the majority of interest rates fall below fifteen percent. There are relatively few loans given at an interest rate higher than twenty percent with the maximum interest rates reaching into the low thirties. Conversely, no interest rates reach below five percent. While the Lending Club had a non-negligible proportion of loans given at high rates, the majority of loans were given at reasonable rates.

2.3 Interest Rates by Loan Grade



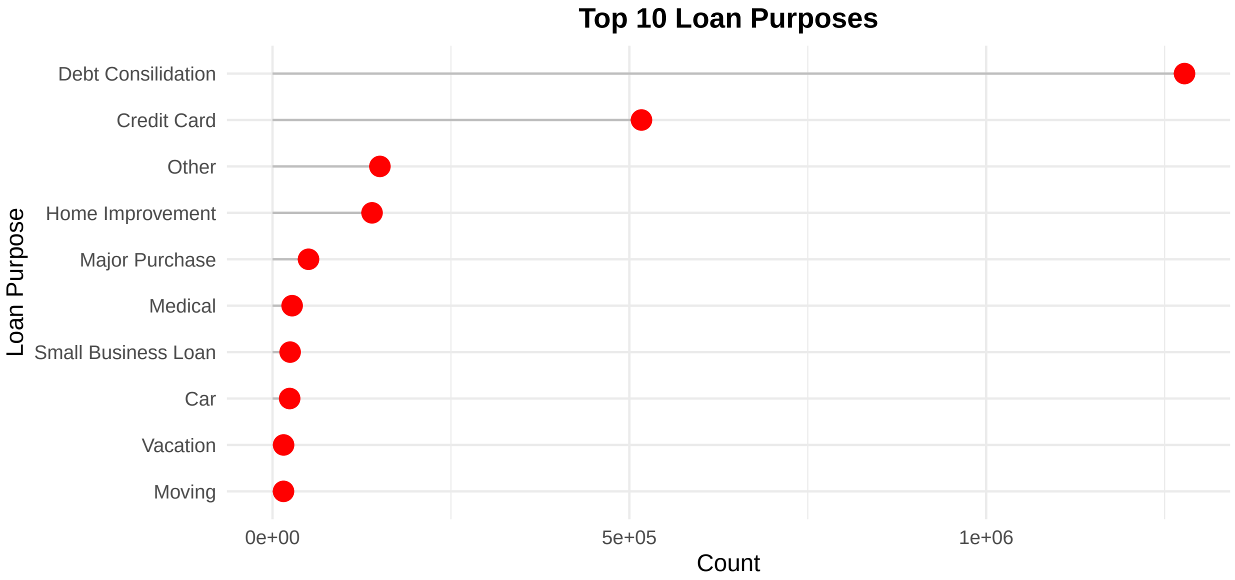
Next, we see how interest rates and loan grades correlate with the above plot. The violin plot shows the interest rates of each grade of loan. The white point in the center represents the mean of the interest rate for each grade. The plot suggests a linear relationship between the grade and the interest rate received. We can see that for Grade A loans, the mean interest rate is roughly seven percent, while for Grade G loans, the average interest rate is approaching thirty percent. The Grades E and F have a large range of interest rates, while the Grade A loans have the tightest distribution of rates. This aligns with our understanding from the first graphs.

2.4 Histogram of Funded Amounts



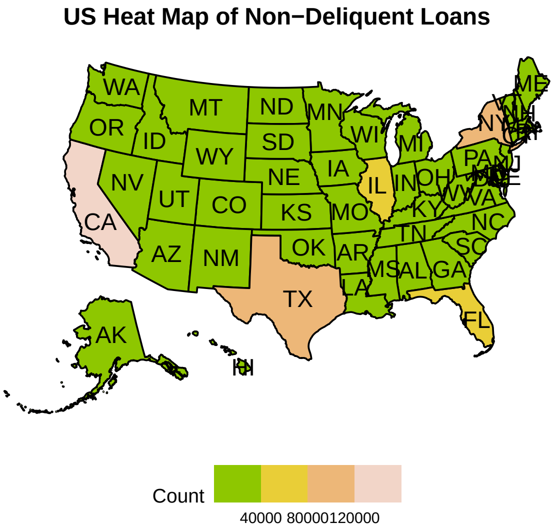
In visualizing the funded amounts using a histogram, one can identify the distribution of the data. The data exhibits a right skew with the majority of loans being less than 20,000 dollars. This visualization and the fact that Lending Club is a peer to peer lender, we can conclude that the bulk of the loans being extended will be of smaller amounts.

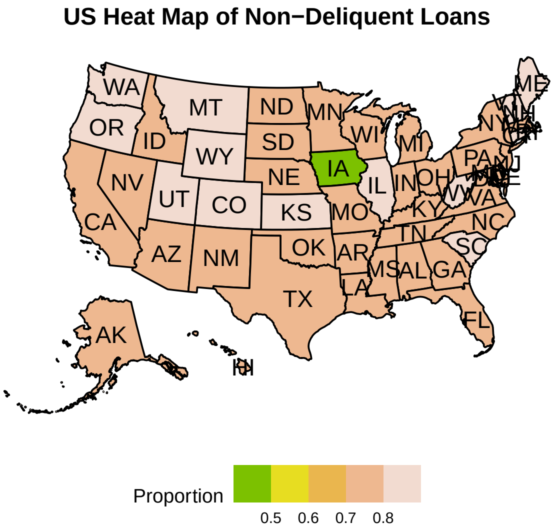
2.5 Analysis of Loan Purposes



The lollipop graph allows us to quickly identify the most common loan purposes. Lending Club was overwhelmingly used for debt consolidation and credit card debt. Relatively few loans were for lifestyle purposes such as a new car, vacation or moving. Understanding the purpose behind the loan helps explain the quality of loans, funding amounts and interest rates. As shown in the other graphs, the median interest rate was in the mid-range and a significant amount of the loans had a rating less than a B. The lower credit ratings are likely correlated to the loan purposes. Debt consolidation and credit card debt are generally correlated with poor credit. We also know that the loans were for smaller amounts. Together, these graphs give a clearer picture of the types of loans given by the Lending Club.

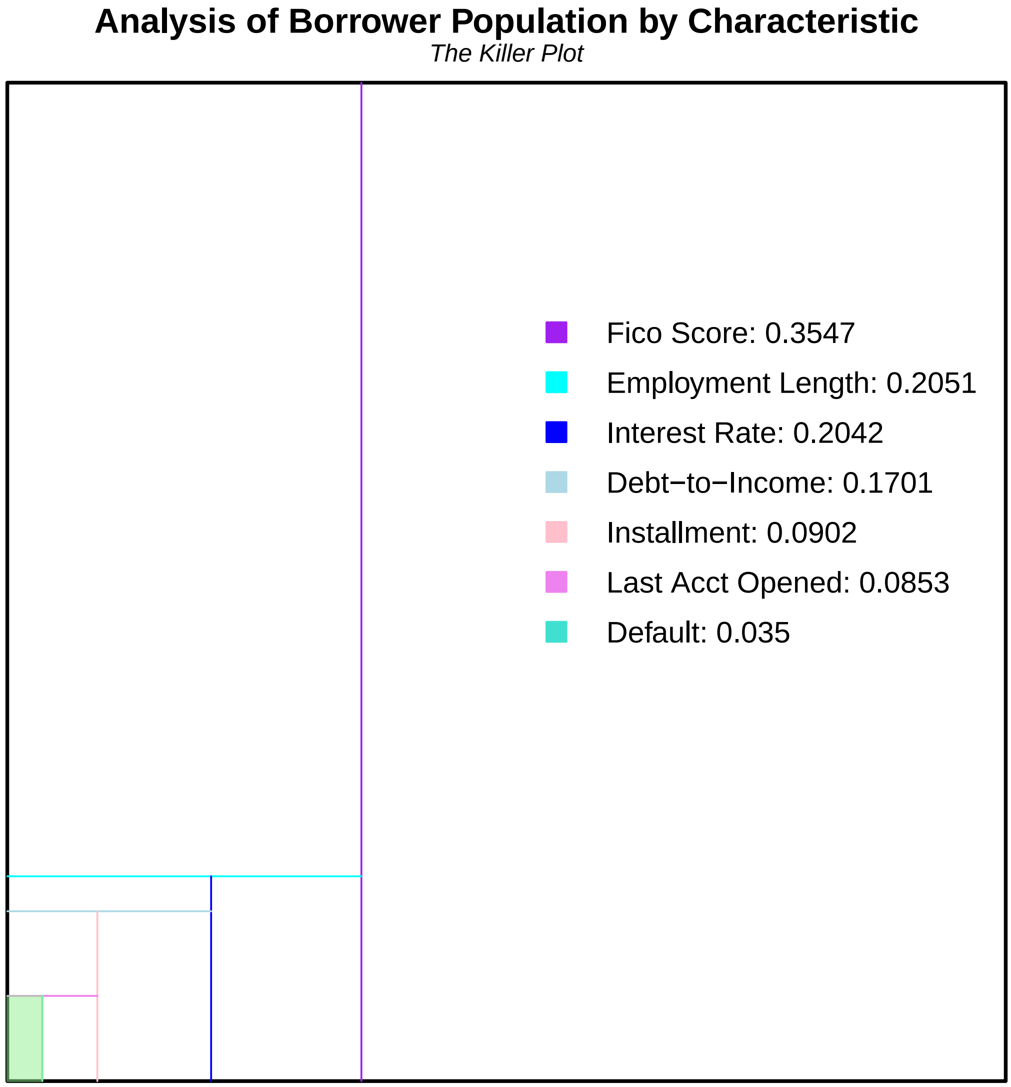
2.6 Examining Geographical Trends





Here we can see two heat maps of the US showing the count and proportion of non-delinquent loans. This will give us a picture if there are geographical patterns, tendencies, or trends regarding loan default. Taking a look at the count graph we can see that larger states like Texas, New York, and California naturally have higher counts (of non-delinquent loans) likely due to their larger population size overall. Looking at this graph alone can be misleading, but examining the count of non-delinquent loans over total loans in the state (proportion) will give us a better picture if some states have a tendency for default. Once we look at the proportion, most states have anywhere from 70 to 80 percent of good loans on the books. A handful of states like Washington, Oregon, and Utah show proportions as high as 80 plus percent. However, there is one outlier, Iowa. Examining this further, we can see that this is likely due to chance as there is a low amount of loans from the state in general. With six non-delinquent loans and fourteen total loans from the state, it is easy to see that our sample size is not large enough to warrant drawing the conclusion that loans from Iowa are more likely than others to default.

2.7 Killer Plot



The Killer Plot allows us to quickly ascertain whether a given borrower is likely to default. The plot divides the population by characteristics to show how common those characteristics are. For example, the above graph depicts a borrower with the following characteristics: a Fico Score around 600 (we used a range of plus or minus seventy-five the input score), an employment length of eight years or more, a debt to income ratio of ten percent or more, an installment payment of at least three hundred dollars, and it has been at least six months since they last opened a credit account. The first input, Fico score, is shown by the largest vertical line in the plot. Each sequential line further divides the population, until at last, the plot shades the defaulting population with the input characteristics.

The input characteristics were chosen because our later analysis showed they were the most important in predicting default. In addition to seeing the likelihood of default, we can see how similar a borrower is to the rest of the Lending Club. This plot is useful for lenders and borrowers alike. It allows a lender to assess the borrower's propensity towards default. It also allows a borrower to use the tool and see how changing aspects of their profile could change their prospects as a borrower. For example, the borrower could quickly see if changing the installment or the interest rate would change how a lender perceives the borrower's risk or profile. This gives the borrower some insight on how to make themselves more attractive to lenders. The plot has additional functionality by allowing a borrower or lender to isolate fewer variables. For example, they could input just Fico score and installment if these were the two variables of interest.

The visual assessment of the attributes of Lending Club loans revealed no surprising or outlandish results that would limit our conclusions to Lending Club alone. For example, the interest rates were largely between ten to fifteen percent. These numbers are fairly typical for loans of different types. The Fico scores of the borrower's were slighly higher than expected, however, the scores were not so high as to limit our conclusions.

3 Model Inputs

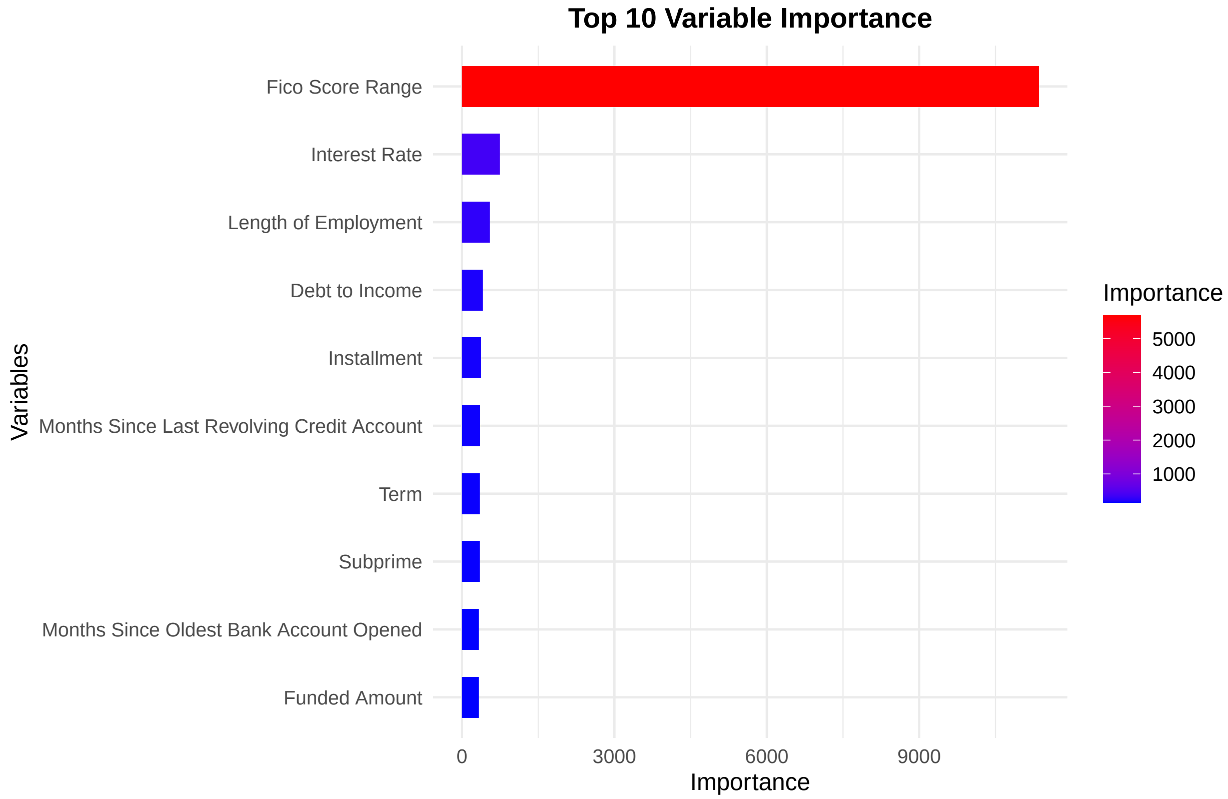
3.1 Data Cleaning

The size of the data set posed challenges for computation and modeling. To reduce the dimensionality, it was necessary to clean the data. In addition to improving the computational efficiency, cleaning the data allowed us to gain a clearer picture of the most important attributes of the loans.

First, we removed columns that were mostly filled with NA and empty string values as these variables provide little insights. Similarly, we removed any columns which were single valued as this does not give us any new information. We next analyzed the remaining variables to identify those which were clearly unconnected to default and removed those variables accordingly. For example, the columns url and title were removed because they do not provide any information on the likelihood of default. Lastly, we removed variables that would lead to data leakage, like recoveries, which only occurs after someone has already defaulted.

Next, we engineered a binary variable, default, to better categorize the dependent variable. This variable combined the loan statuses that are conventionally considered delinquent. This left one category of loans that did not fit this binary mold, current. Current means that the loans are still in term and are not delinquent. They cannot give us any insight on the characteristics of a good or bad loan since they have not come to term, so we omitted them from the analysis. We also created a dummy variable using the loan grade column which puts loans into either prime or sub-prime buckets which is characterized by having a loan grade less than a rating of B. See Appendix A for code on how this was accomplished.

3.2 Variable Selection

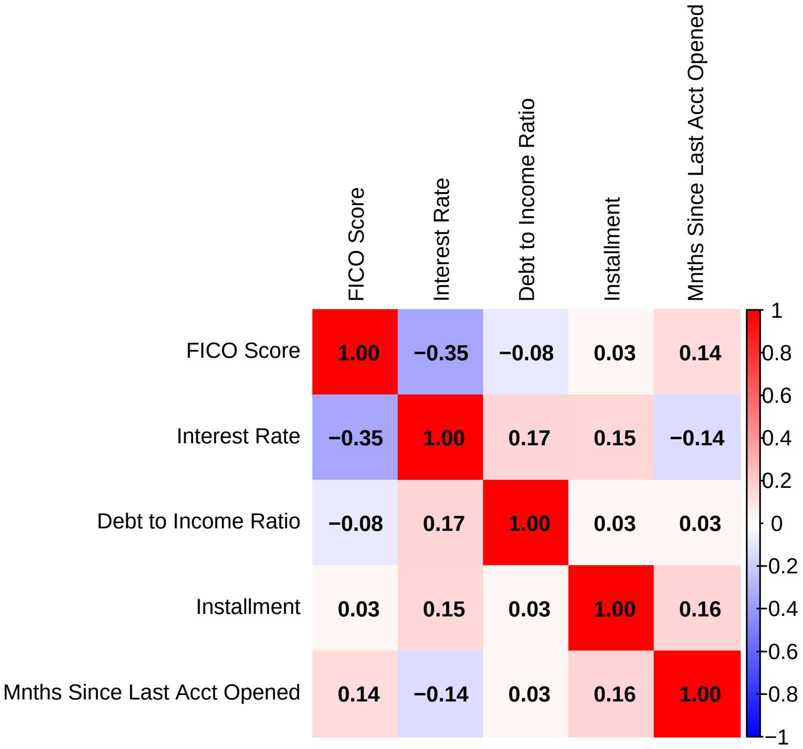


With the cleaned data, the next step was to split the data into training and test sets. This split was still too large for the random forest model to run, so we took a stratified random sample from the training set to still give the model an accurate representation of the data, but small enough so that it will run. After omitting the NA values, as the model cannot handle these values, the sample was small enough to preform the analysis.

We ran the random forests model on all variables except those that we deemed redundant (such as loan grade and sub grade) and used default model parameters for classification problems (ntree of 500, a node size of 1, and mtry of 10). Using this model we were able to extract each variable's importance. We used this data to create a data frame and then selected the top ten variables according to the mean decrease in the gini coefficient as calculated by the model. Fico score (high value of the last fico range) is overwhelmingly the most important variable for predicting default with features like interest rate, employment length, and debt-income ratio falling greatly behind. See Appendix B for code on how this was accomplished.

4 Modelling

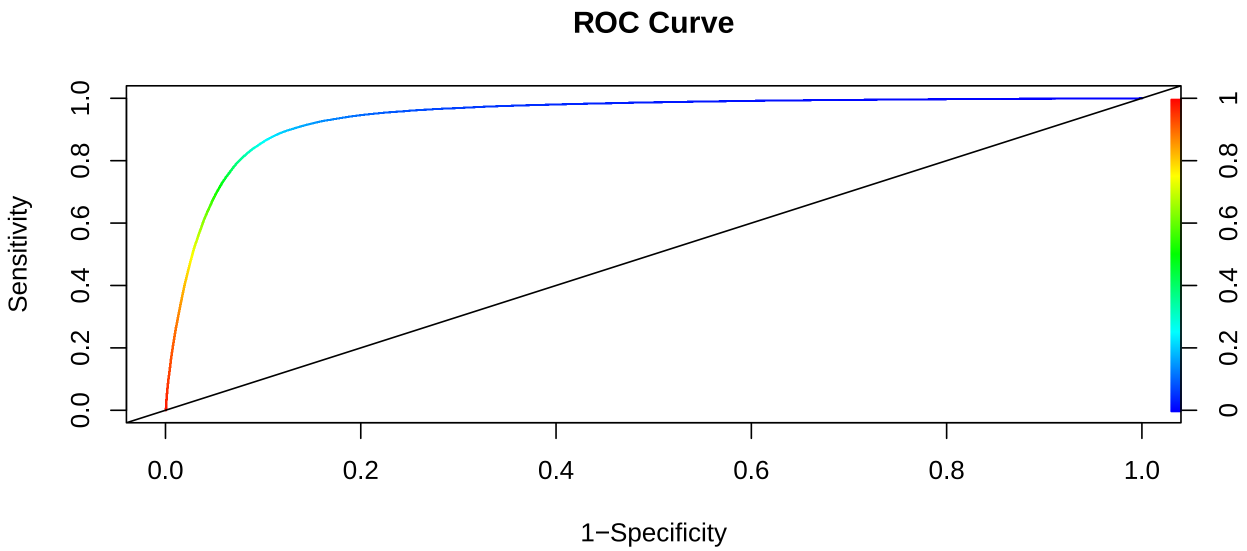
4.1 Logistic Regression



Now that we have the key variables that help predict default, we can use these as inputs to a model. Logistic regression is a powerful classification tool that we can develop to help us do just that. For model simplicity, we will select the top six variables. These are last Fico score, interest rate, employment length, debt-to-income ratio, installment, and month since oldest revolving account opened. Now that we have our variables of interest we can divide our data into a training set which will be used to develop the model, and a test set which will be used to evaluate our model. This split was done at 70 percent so we can have enough data for the model to pick up the underlying patterns.

With our variables selected and data ready, we can implement the model. After building the logistic regression now we can run a summary to check the significance of our variables and check some assumptions of the model to make sure this is statistically valid. All our variables are significant with a p value of essentially zero. Next, we created a correlation matrix to see if any of our variables are highly correlated with each other which would suggest multicollinearity which is an important assumption of logistic regression. As we can see no high correlation exists, with the largest correlation being between interest rate and last Fico score at a correlation of .35.

4.2 Model Evaluation



Accuracy	Sensitivity	Specificity
0.90	0.71	0.94

Table 1: Results (%)

After building our model, we must evaluate its performance to truly see how powerful it and the predictors are. Our logistic regression model outputs probabilities or values between zero and one, so to get our response we need a threshold where if a given output it is above it, we classify it as a defaulted loan and fully paid otherwise. As a starting point and benchmark, we will use 50 percent.

Now that we have our predicted response, we can compare this with the actual values from the data itself. It must be noted that since we are now in the evaluation stage, we will be using unseen data, the test set, to do our prediction and comparison. The comparison between our predicted and actual values of the response will manifest itself in a confusion matrix. Using the confusion matrix we will calculate the key metrics which will help us to evaluate the models performance. These are sensitivity, which measures how well we predict default or ones, specificity, which tells us how well we predict a fully paid loan or zeros and accuracy, which measures the strength of the model in regards to predicting both defaulted and fully paid loans. As seen in the table above, our accuracy and specificity are exceptionally high; however, with there being a noticeable difference in sensitivity this could suggest that there needs to be some adjustment in the model. Since sensitivity measures how well we classify a loan as defaulted, which is the most costly and key thing we are trying to predict, with this metric being noticeably smaller, this suggests an area we can try to optimize.

Another way we can try to evaluate our model is by plotting out an ROC curve and finding the area underneath it. The ROC curve, as shown above, plots sensitivity against one minus the specificity. A perfect model will be able to predict both with full accuracy and so the plot would form a ninety degree angle in the top left corner. Thus, the closer we can get to forming such a plot, the better our performance. For this reason, we take the area under the curve to signify how well the model performs with the value one being a perfect evaluation. Our area under the curve is .93 which suggests a very well performing model.

4.3 Model Optimization

	Actual	
Predicted	Positive	Negative
Positive	0.00	1.00
Negative	1.50	0.00

Table 2: Loss Matrix

Accuracy	Sensitivity	Specificity
0.90	0.78	0.93

Table 3: Optimized Results (%)

As mentioned in Section 4.2, the low sensitivity gives us reason to try and optimize our model. We decided the most appropriate way to optimize was through the implementation of a loss matrix. A loss matrix assigns a penalty to making certain predictions. In this case, predicting a loan will be fully paid when it defaults, or a false negative, is much more costly than inaccurately classifying a loan as defaulted when it actually is fully paid, or a false positive. The former results in a loss of capital for the investor whereas the latter simply means an investor will forgo that lending opportunity which only results in an opportunity cost and not a material loss. That being said, as can be seen in the first table, we assigned false negatives a penalty of 1.5 as opposed to the false positive penalty of just 1.

Unlike the machine learning model classification tree, we cannot directly input the loss matrix into the model and let it do the calculation. Instead a way we can use this to optimize our results is by using the proportion of our false negative penalty over the total penalties as a new threshold. Given our loss matrix, our new threshold is 40 percent. We can now incorporate this threshold in the computation of our predictions of default and fully paid classes and recalculate our performance metrics and see how they have changed. As seen in table two, our accuracy has stayed the same, but sensitivity has increased by seven percentage points while specificity has only decreased by one percentage point! This is a great result as we have essentially maintained all of our metrics while showing a significant increase in our desired metric, sensitivity.

5 Conclusion

Our analysis of the Lending Data led to several important conclusions. The most important borrower characteristics for predicting default was overwhelmingly the Fico score range. While Fico scores do not provide a complete picture of a given borrower's profile, they are exceedingly helpful in predicting default on a loan. Other important factors in predicting default are the interest rate, employment length, and debt to income ratio. Lending Club itself is a unique institution by distributing peer-to-peer loans. In this way, it has several unique aspects. Overall, the borrowers had high Fico scores and relatively small loan amounts. Interestingly, the loans were primarily for debt consolidation. We tend to assume that borrowers seeking debt consolidation are inherently riskier than those looking to finance home projects. However, the Lending Club as a whole had a small default ratio. In our visualizations, we did not see any outlying variables that would limit our results to Lending Club. In future work, we can test our predictions to see if the analysis can be repeated in other lending scenarios.

It is essential for both borrowers and lenders to have an accurate assessment of the likelihood of default. For lenders, this enables them to give competitive loans while maintaining profitability. For borrowers, they can gain an understanding of how to become an attractive borrower. This may have important implications for small businesses and individuals alike.

6 Appendix A

The below code was used to clean the data set to allow for modeling as discussed in the section Data Cleaning.

```r
# Data Cleaning

# Loading in database connection
dcon <- dbConnect(SQLite(), dbname =
  "/Users/pirroprifti/Desktop/R for Data Science/Final Project/accepted_db.db")

# Query data for loans that have come to term
res <- dbSendQuery(conn = dcon, "
SELECT *
FROM accepted
WHERE loan_status NOT IN ('Current', 'In Grace Period') ;")
clean <- dbFetch(res, -1)
dbClearResult(res)

# Drop Unnecessary Columns According to NAs and ""
index = c()
for (i in 1:ncol(clean)){
  perc_na = mean(is.na(clean[,i]))
  if (perc_na > .5){index = append(index, i)}
}
clean = clean[, -index]

# Dropping unimportant columns
# head(acc[, 1:35], 1)
drop = c("id", "issue_d", "url", "earliest_cr_line", "title", "emp_title",
         "zip_code")
clean = clean[, !names(clean) %in% drop]

# head(acc[, 36:70], 1)
drop = c("last_pymnt_d", "last_credit_pull_d" )
clean = clean[, !names(clean) %in% drop]

# head(acc[71:103, ], 1)
drop = c("disbursement_method", "initial_list_status")
clean = clean[, !names(clean) %in% drop]

# Removing data leakage features
drop = c("recoveries", "collection_recovery_fee", "total_rec_prncp",
         "last_pymnt_amnt", "total_pymnt_inv", "total_pymnt",
         "funded_amnt_inv", "debt_settlement_flag", "out_prncp_inv", "out_prncp",
         "total_rec_int", "total_rec_late_fee")
clean = clean[, !names(clean) %in% drop]

# Dropping Single Value Columns
index = c()
for (i in 1:ncol(clean)){
  colm = clean[!is.na(clean[,i]), i]
  if (length(unique(colm)) < 2){index = append(index, i)}
}
```

```r
clean = clean[, -index]

# Designing Dummy and Factor Variables

# Dependent Variable
indicator = ifelse(clean$loan_status == "Fully Paid" |
       clean$loan_status ==
       "Does not meet the credit policy. Status:Fully Paid", 0, 1)
clean$default = as.factor(indicator)

# Creating Prime/Sub prime Variable
indicator = ifelse(clean$grade == "A" |
       clean$grade == "B", 0, 1)
clean$subprime = as.factor(indicator)

# Converting to numeric
clean$mths_since_recent_inq = as.numeric(clean$mths_since_recent_inq)

# Uploading clean data into database for easy retrieval
dbWriteTable(conn = dcon, name = "acc_clean", clean,
             append = TRUE, row.names = FALSE)
```

7 Appendix B

See below for the code on variable selection.

```r
# Loading data
dcon <- dbConnect(SQLite(), dbname =
        "/Users/pirroprifti/Desktop/R for Data Science/Final Project/accepted_db.db")

res <- dbSendQuery(conn = dcon, "
SELECT *
FROM acc_clean;")
mydf <- dbFetch(res, -1)
dbClearResult(res)

# Creating factor variables
mydf <- mydf %>%
  mutate_if(is.character, as.factor)

# Train test split at 70%
set.seed(123, sample.kind = "Rejection")
spl = sample.split(mydf$default, .7)
train_set = mydf[spl, ]
test_set = mydf[!spl, ]

# Perform stratified sampling to get a small enough sample
stratified_sample = train_set %>%
  group_by(default) %>%
  sample_n(size = floor(.05 * n()), replace = FALSE)

# Omitting NAs so Model Runs
```

```
clean = na.omit(stratified_sample)

# Finding key features

# Removing Irrelevant and Redundant Columns
# ("last_fico_range_low", "grade", "sub_grade", "addr_state", "loan_status")

# Running Random Forest Model
rf_mod = randomForest(default ~ . -grade -sub_grade -addr_state -loan_status
                      -last_fico_range_low, data = clean, ntree = 500,
                      nodesize = 1, mtry = 10)

# Checking For Important Variables with Random Forests and Getting Top 10
importance_data = importance(rf_mod)
var_imp_df = data.frame(Variables = rownames(importance_data),
                        Importance = importance_data[, 1])
top_vars <- var_imp_df[order(-var_imp_df$Importance), ][1:10, ]

# Uploading variable importance data frame to database for quick plotting
dbWriteTable(conn = dcon, name = "var_imp", top_vars,
             append = TRUE, row.names = FALSE)
```