Loan Fulfillment Milestone Report 1

Jorge Londono

Problem Statement

Bank loans are an integral part of our economy and are responsible for funding a myriad of different programs such as housing mortgages, business ventures, or short term personal projects. Banks and capital service companies will always face a risk whenever they lend money to a borrower. If the loan in unsecured then the borrower can default on the loan and the lender can end in a very negative situation. Therefore, it is important for lenders to be able to predict whether someone applying for a loan will be able to pay back their loan, or if they will default.

This project will aim to create a model that will classify whether a borrower will default or have the loan charged off. By doing so, the client, in this case being the lender, will be able to better understand whether a risk is worth taking. If a loan is classified as belonging to a borrower who will default or have the loan charged off, this can better influence the decision of the lender.

Data

The data for this project can be found from a Kaggle repository found here: (https://www.kaggle.com/wordsforthewise/lending-club?select=rejected_2007_to_2018Q4.csv.gz)

It contains all the loans by Lending Club from the years 2007 to 2018. The data contains over 150 variables for each loan and its corresponding borrower. There are continuous variables such as annual income of the borrower, and the loan amount that could prove to be important features in our model. Likewise, there are categorical variables such as purpose of the loan, or home ownership of the borrower. Because this data has a vast number of features to test as

well as an extremely large number of observations, our machine learning model will be very strong.

Cleaning and Wrangling

Even though the data came from Kaggle and is therefore pretty clean and organized already, there are a number of procedures needed to have it ready for explanatory data analysis and machine learning.

Firstly, this data set is so massive that loading it normally would take too long and it would be inefficient. Therefore, when we load it *pandas.read_csv()* the optional parameter *chunksize* is set to *10000000.* This way there is not a massive memory requirement needed to load in our dataset.

Next, we need to remove all the data that will not be used for our model. Since the point of our project is to classify whether loan will be delinquent or not, we need to remove all the observations that end up with a loan status that is not either: Fully paid, default, or charged off. Once those have been removed, there will be a new column created that tells us whether the result of the loan was fully paid or whether it was 'delinquent'. In this case, delinquent means that the loan defaulted or was charged off.

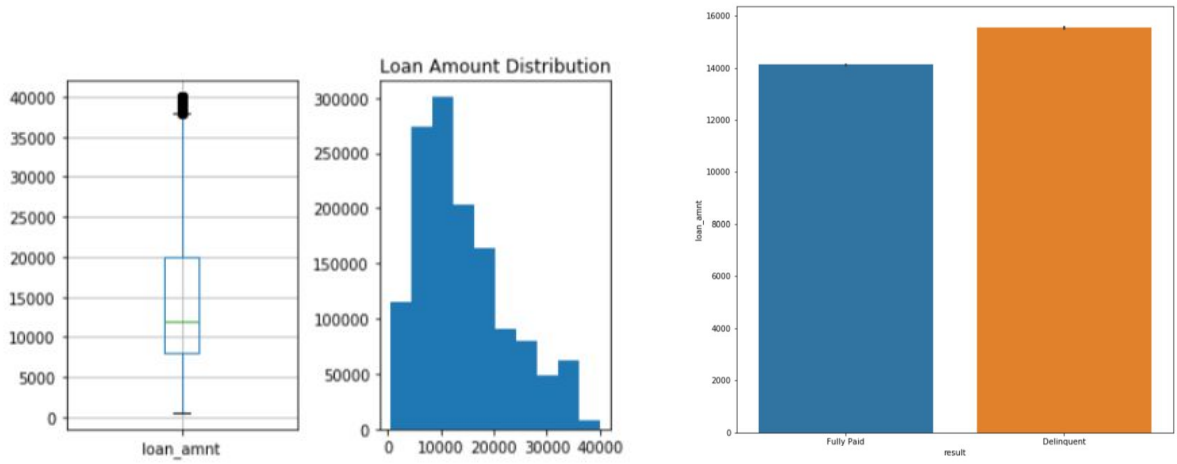
Finally, we were left with a Data Frame of 1,346,111 observations for us to experiment, explore and model. Below is a table that shows the total amount of delinquent, versus fully paid loans.

	id
result	
Delinquent	269360
Fully Paid	1076751

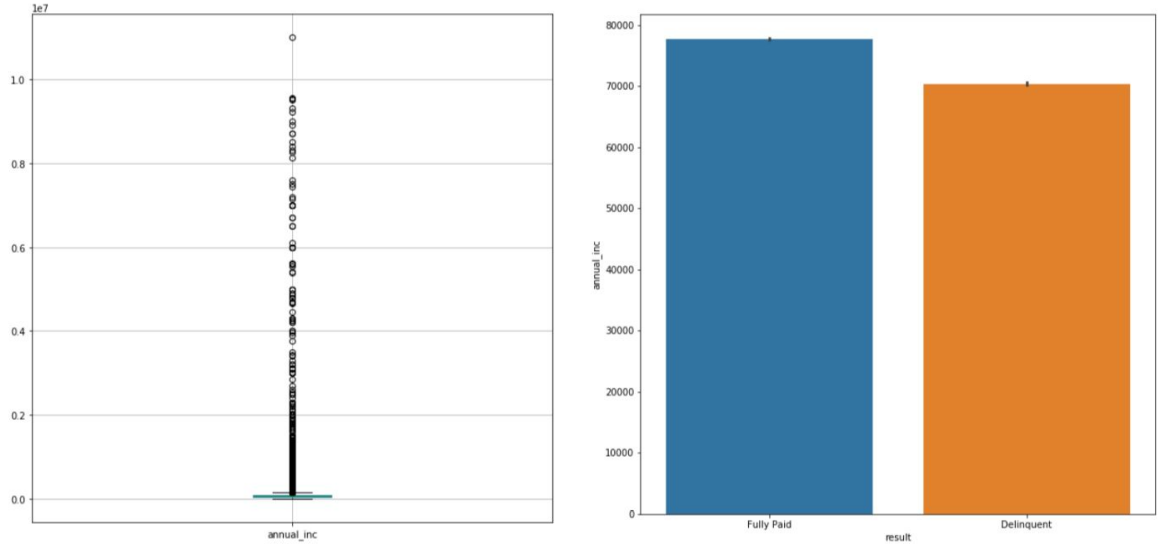
Exploratory Data Analysis

In order to select the best features to test out in our model, we need to see if we can find any trends or visual hints that will lead us to a hypothesis about the data we possess.

Starting with the continuous variables loan amount, which represents the total amount borrowed, and annual income, which reflects the total income the borrower earns per year.



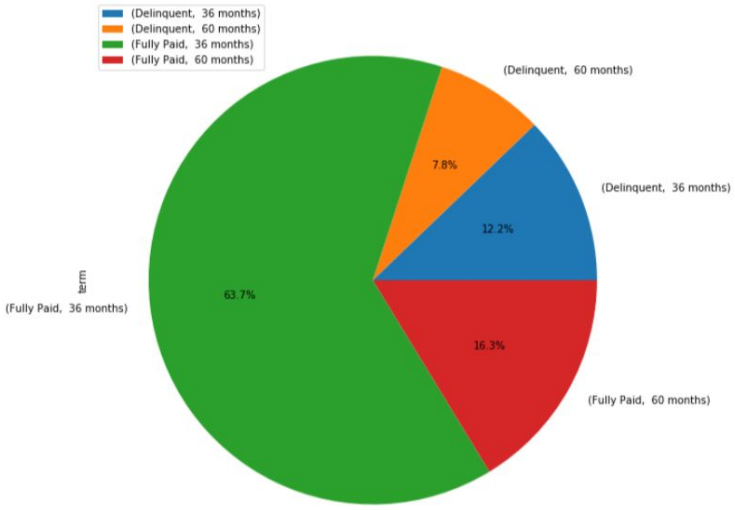
Looking at the spread of the loan amount data we see that it is not normal and skewed to the right. Therefore, this variable will probably have to be normalized in order for it to work better in our model. The right graph compares the mean loan amount of the 2 groups, fully paid and delinquent. It is clear from the graph that delinquent amounts on average have a larger loan amount.



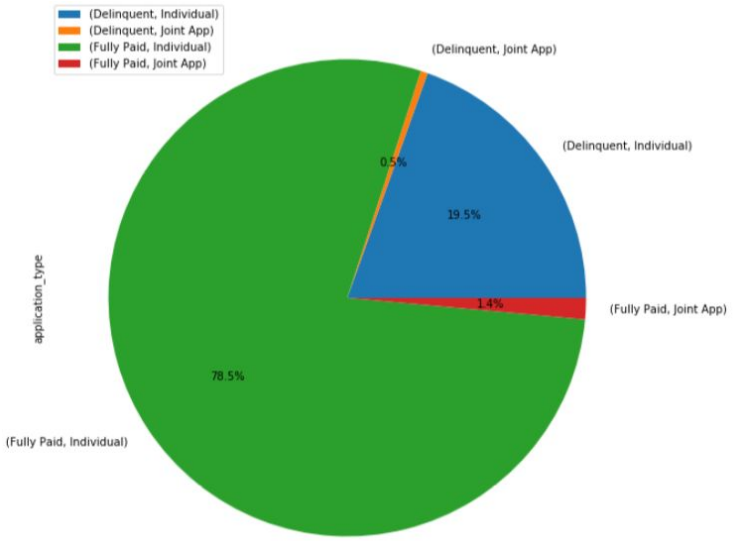
Just like the spread of the loan amount, the variable of annual income definitely has a problem with skewness towards the right. Again, this will probably have to be normalized in order to make our model work better. Additionally, the graph on the left shows us that delinquent loans correspond on average to lower annual income from the borrower than fully paid loans.

One of the categorical variables that are included in this data set is term amount. This in fact, is no continuous variable, as there are only 2 options: 36 months, and 60 months. Another

categorical variable is application type which just like the previous variable only has 2 options: individual, or joint.



From this pie chart we can see that the proportion of delinquent loans versus fully paid loans is ls larger when there are 60 month payment terms than 36 month payment terms.



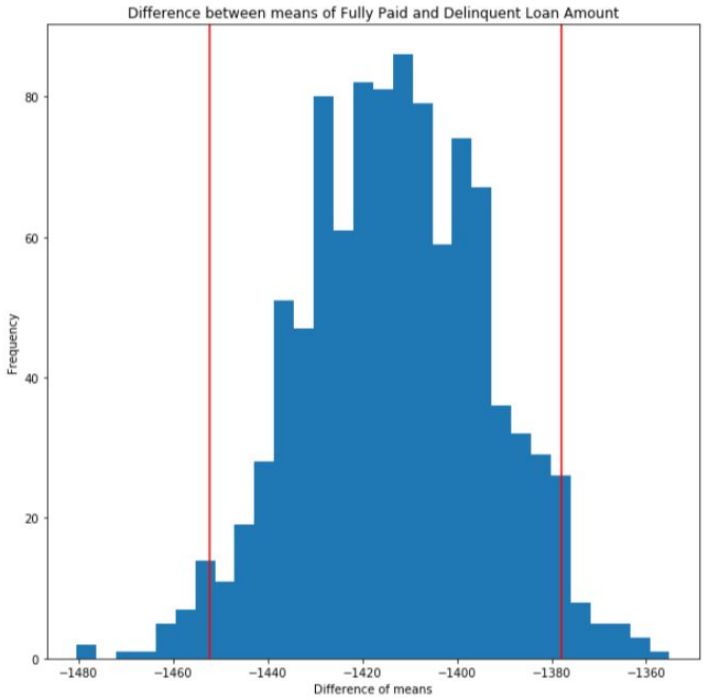
This pie chart shows us that around ⅓ joint applicants end up with delinquent loans while about ¼ individual applicants end up fully paying off the loans. This shows us that joint applicants are more likely to not fully pay their loan.

Statistical Analysis

From the initial analysis above, there are two main points that should be tested with a statistical

analysis. There is a difference in the mean loan amounts between delinquent and fully paid group, and there is also a difference in the mean annual incomes between the delinquent and fully paid groups. We will be conducting hypothesis testing via bootstrapping to test whether these observations are statistically significant. Bootstrapping allows us to replicate our data multiple times, and we can use the measurements of that data to figure out if our assumptions are not simply derived by chance up to a specific significance level.

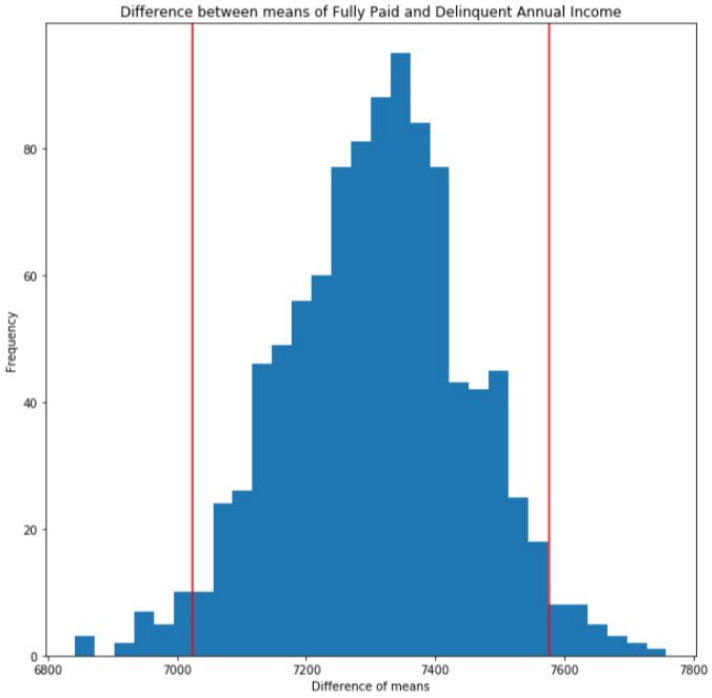
Our first null hypothesis is that there is no difference in the mean loan amount of fully paid and delinquent groups. The alternate hypothesis would be that there is a difference in the 2 means.



A two tailed test with a 95% confidence interval shows that the difference between the means of the two groups is between -1452 and -1378.

Because this interval does not include 0, at 95% significance we can reject the null hypothesis that there is no difference in the mean loan amount of fully paid and delinquent groups, and we can support our alternate hypothesis. This means that this can be a feature in our model that could be very helpful.

The 2nd null hypothesis is that there is no difference in the mean annual income of fully paid and delinquent groups. The alternate hypothesis would be that there is a difference in the 2 means.



A two tailed test with a 95% confidence interval shows that the difference between the means of the two groups is between 7023 and 7576.

Because this interval does not include 0, at 95% significance we can reject the null hypothesis that there is no difference in the mean annual income of fully paid and delinquent groups, and we can support our alternate hypothesis. This also shows us that this can be a useful feature for our model.