PREDICTING THE LIKELIHOOD OF A MORTGAGE LOAN DEFAULT

Tanya Naheta (netID: tsn6109)

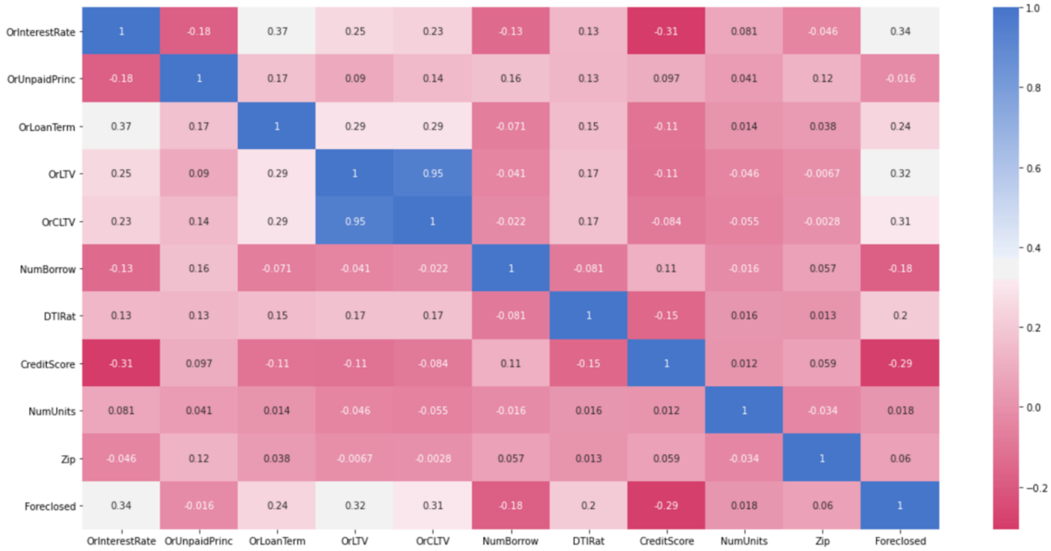
The Federal National Mortgage Association, commonly known as Fannie Mae is a player in the secondary mortgage market. They buy mortgages from lenders and repackage them as mortgage-backed securities (MBS) in the secondary market. Fannie Mac has collected this data and made it public on their website for research purposes; so that the public can analyze the credit performance of their loans. We can use this data not only to assess the performance of Fannie Mae, but also use this data to determine ways to estimate things such as: what are the indicators of a borrower that makes it likely for Fannie Mae to buy and sell, what type of property is in high demand, what are the quality of loans that Fannie Mae are buying and selling etc.

The dataset on their website is available from 2000 and every quarter since then. The data that I will be analyzing is of the second quarter of 2019 – 91 days between 1st April till 30th June. The origination date is a specific date mm/dd/yyyy sometime during this quarter. The “Acquisition” txt file includes loan data from the origination, or delivery date of the loan to Fannie Mae. There are 25 columns and 407389 rows in the Acquisition data file for Q2 2019.

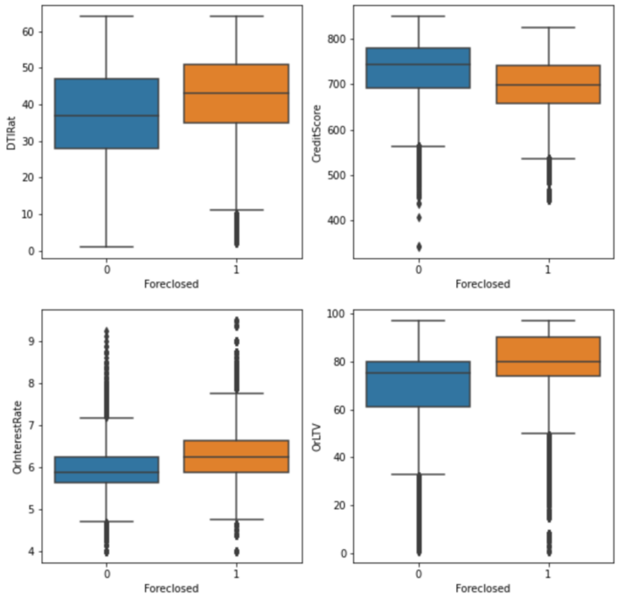
The data describes the mortgage loan (such as interest rate, loan term, origination date, etc.), describes the property (property type, occupancy status, number of units, etc.), and describes the borrowers (credit score, number or borrowers, first-time house buyer indicator, etc.). These three combined gives us a good overall picture about the loan that was acquired by Fannie Mae. This data is for Single-family homes, which is a standalone house as compared to a residential community.

***What can we tell about the relationships between a loan default and some of the variables?***

Firstly, we declare that the Loan Default (binary value), also known as 'Foreclosed' is the target variable for this analysis. The most important question when it comes to analyzing mortgage loans is to understand what variables affect the probability of default and to what degree. First, to gain an overall idea of all the different relationships, we have plotted a correlation matrix which measures the correlation between the different attributes and the binary variable 'Foreclosed'.

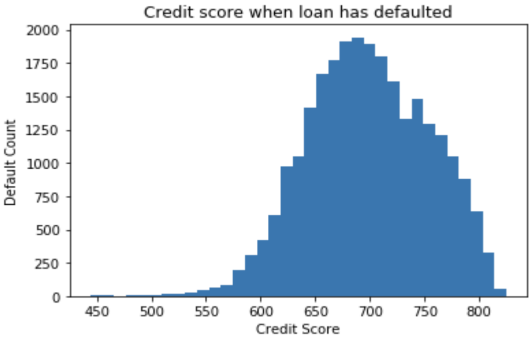


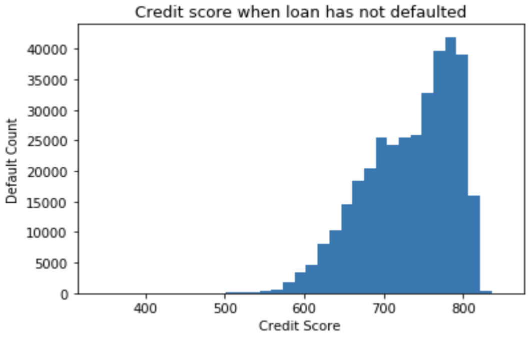
From the matrix above, we can see that some of the highest correlation to the target variable exists in the columns: OrInterestRate, OrLTV, OrCLTV, OrLoanTerm. Also, as expected, CreditScore seems to have a negative correlation to the target variable of Foreclosed. The correlation matrix above provides a high-level overview of our attributes, we can dive deeper into the data and compare the different attributes we have.

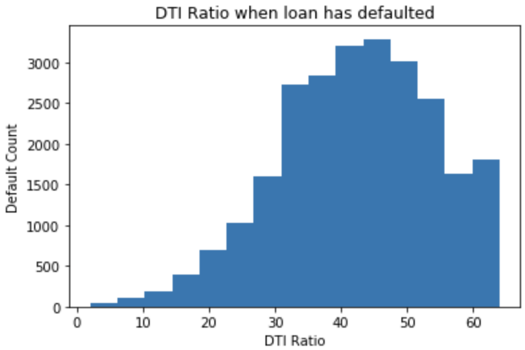


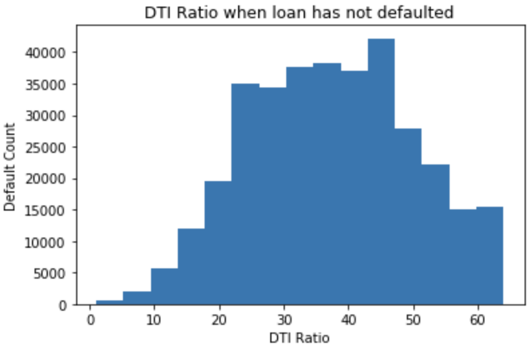
Furthermore, the boxplot below shows us the difference between the distributions of the attributes when we separate the population into loans that have foreclosed and ones that have not. This strengthens our understanding of the relationships as we can see a difference in the means.

Furthermore, the following plots show the distributions of the attributes when the loan has been defaulted and when it has not. We can start to hypothesize the reasons for differences in the distributions. We can see that the credit score is definitely right skewed when the loan was not defaulted. This makes sense because it is less likely for people with high credit scores to default on a loan. The same is the case for the debt to income ratio, however it is not as exaggerated. The graph is centralized on a higher debt to income ratio when the loan has defaulted..

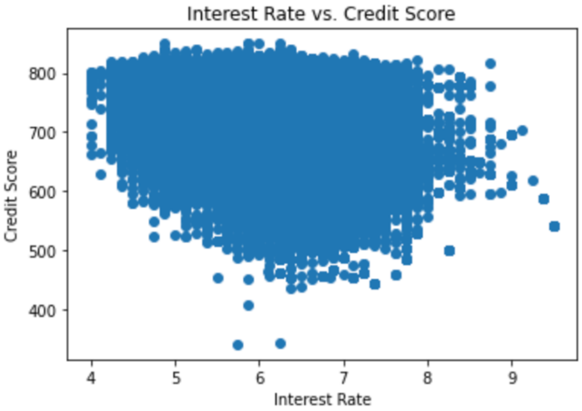








It was also worth exploring the possibility of covariance between the credit score and interest rate variables as it is likely that someone with a low credit score will be charged a high interest rate on their loan. Also, this would be important to understand as we proceed further with the regression model. However, after plotting these two variables on a scatter plot, there doesn't seem to be much correlation between the data



PREDICTING THE LIKELIHOOD OF A MORTGAGE LOAN DEFAULT

Tanya Naheta (netID: tsn6109)

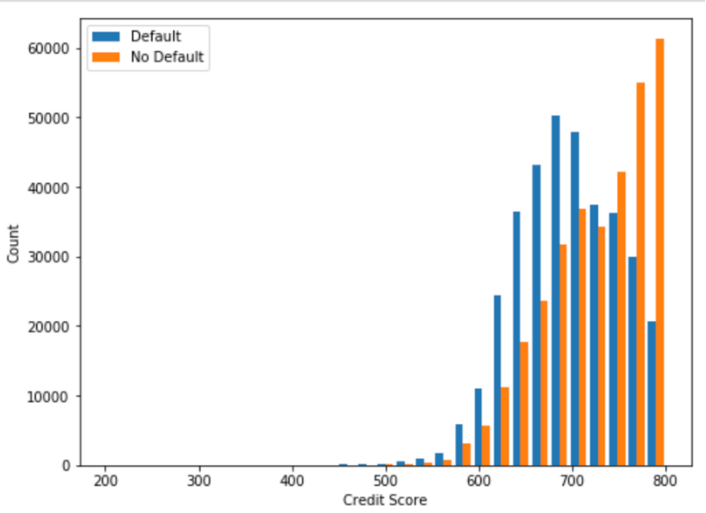
### *Is it acceptable to say that credit score and interest rate has an effect on the likelihood of a default?*

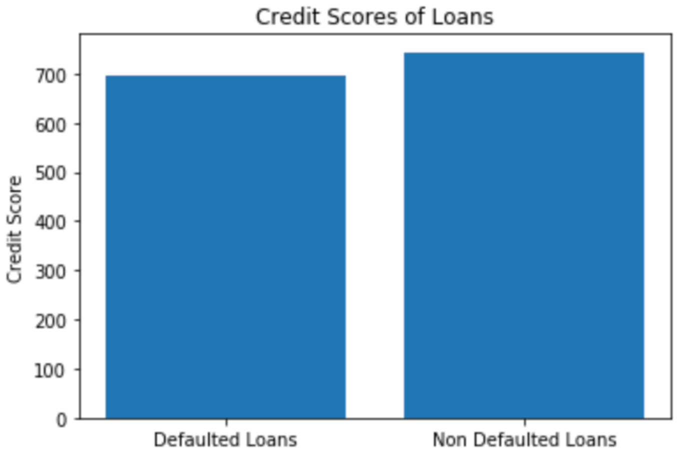
One of the hypotheses regarding this dataset is that the credit score will have an inverse correlation to the likelihood of a loan default. This is because people with a higher credit score should be in a more financially stable position and less likely to default. Another hypothesis is that the original interest rate on the loan will have a strong correlation to the likelihood of a loan default. This is because a higher interest rate would mean that borrowers must pay off more money, making it harder for them to do so. We will be testing both hypotheses by comparing the attributes with our target variable.

We will be implementing the Mann Whitney test approach to answer this question. This approach fits well as we have data from two samples (credits scores/interest rates from defaulted loans and non-defaulted loans) that come from the same underlying population. The null hypothesis is that there is no difference between the credit scores/interest rates of the loans from defaulted loans vs. from non-defaulted loans, and we will consider alpha at 0.05.

Credit Scores

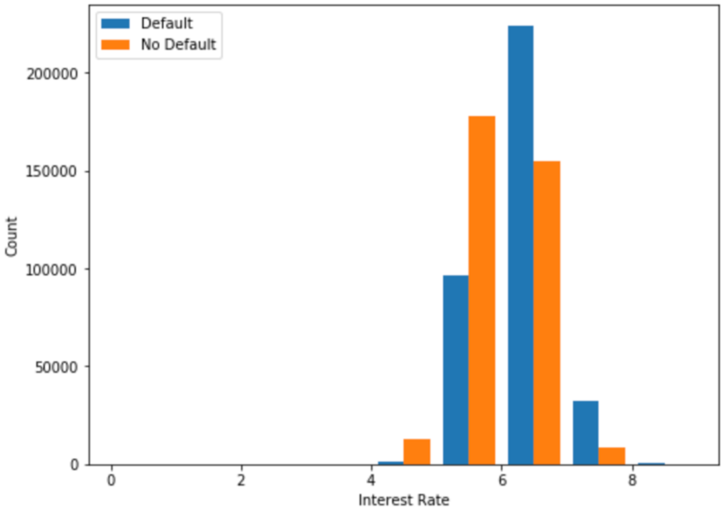
Based on our calculations, the p value is 4.65e-19 which is much smaller than the alpha, so we reject the null hypothesis. This means that we reject that there is no difference in credit scores between defaulted and non-defaulted loans and say that there is a strong indication that there is a difference.

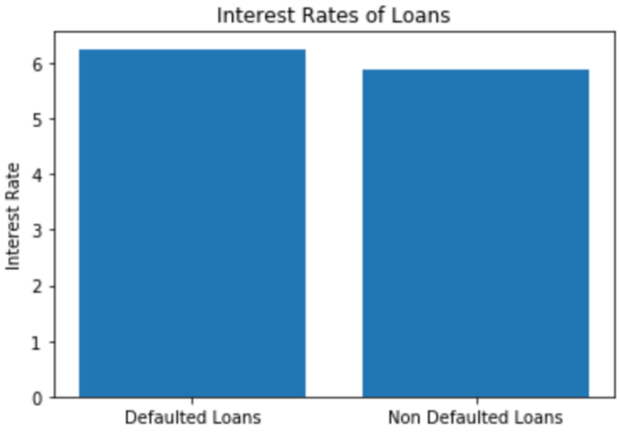




## Interest Rates

Based on our calculations, the p value is 6.5e-27 which is much smaller than the alpha, so we reject the null hypothesis. This means that we reject that there is no difference in interest rates between defaulted and non-defaulted loans and say that there is a strong indication that there is a difference.





* Note that in our calculations, we limited the sample size to 500. This is because with the extremely large sample size (over 354,000 for each sample), even the most minuscule difference between the groups turn out to be very significant and leads to a p value of 0.0.

PREDICTING THE LIKELIHOOD OF A MORTGAGE LOAN DEFAULT

Tanya Naheta (netID: tsn6109)

***Can we create a regression model to predict whether a loan will default or not based on certain attributes?***

We will build a logistic regression algorithm to answer this question our question is a binary one. We will use several attributes from our data set as the dependent variables in this predictive modelling technique which are listed clearly in the code.

*Pre-Processing:*

A large portion of the analysis was spent cleaning and pre-processing the data. To summarize: The data was separated into two files – Acquisition file and performance file which needed to be merged. We used the attributes from the acquisition dataset to train our model and merged it with the performance data set which contains the target variable – loan default. We convert this target variable to a binary 0 or 1 based on the 'Foreclosure date' column and drop the rest of the columns.

However, while analyzing the data we found some duplicate loan ID's in the performance data set. To handle this, we chose the latest entry so that we get the most up to date information. Also, the amount of data of non-defaulted loans was much greater than defaulted loans, which is an issue because it can incorrectly train our model. To correct this, we ised up-sampling on the minority class to ensure the model does not skew. Furthermore, to clean the data we performed steps such as dropping the attribute columns with a lot of null values and converting dates into separate columns for year and date so that the data contain numeric values. We also normalized the data so that they have values between 0 to 1 so that the attributes can be compared more efficiently.

*Building the regression model:*

We saved 30% of the data for testing purposes, and 70% of the data for training/tuning purposes. To find the parameters that would best fit the data we used gridsearchCV as a method of hyper parameter tuning.

*Results:*

This logistic regression model had an accuracy score of 0.75 and F1 score of 0.76. Further backed up by area under the curve metric of 0.83.

In conclusion, we were able to justify our initial hypothesis that the credit score is inversely correlated with our target variable, and the original interest rate is correlated with our target variable.

In the future, to improve these results, the categorical columns could be handled slightly differently by analyzing the frequencies of each of the occurrences combined with their correlation to the target variable. Also, we could compare the regression model another algorithm such as K-nearest neighbors, NBC, or Decision tree to see which model suits this dataset and problem type.

It would also be very interesting to analyze the mortgage data from the housing crisis of 2008 and compare it with future years to see what qualities of the loans have changed.