Analyzing A Credit Score Prediction Model

Evaristus Chibuike Ezekwem & Isaac Heeseok Joo

1. Background

Credit scores and personal financing have been extremely important in shaping our economy as well as the state of our financial well-beings. The banking system highly depends on loaning out the money to businesses and individuals and makes profit based on interest paid on such loans over time. The crucial part in this model is dependent on the abilities of businesses and individuals that borrowed money to repay the bank. If the bank gave out money without paying close attention to whom they are lending out the money, the probability of repayments will likely decrease, which would lead to financial loss to the bank. That is why there are certain metrics in place, such as credit score, for the bank to gauge the likelihood of repayment based on the individual's financial history.

Even with this quantitative measurement, there is still room for improvement to minimize the bank's loss by predicting people who would default with high precision. The current credit scoring algorithm that the banks use is basically predicting the likelihood of individuals paying the money back to the bank so that the bank can decide whether the person qualifies for the loan or not.

The Automated Decision System (ADS) we chose is a competition from Kaggle in which the competitors attempt to improve upon the current system to more accurately predict the probability that an individual will experience financial distress in the next two years.

The finished project that we decided to evaluate was the winning project for this competition by Alec Stephenson, Nathaniel Ramm, and Eu Jin Lok, which had approximately 87\% chance of predicting that an individual will go through financial hardship that would prevent them from paying back the loan, thus declining to lend money to these individuals.

2. Input and Output

There is total of 250,000 data provided from Kaggle, but the method of data collection is not clarified in the competition. Within the data, there are total of 11 input features, which are:

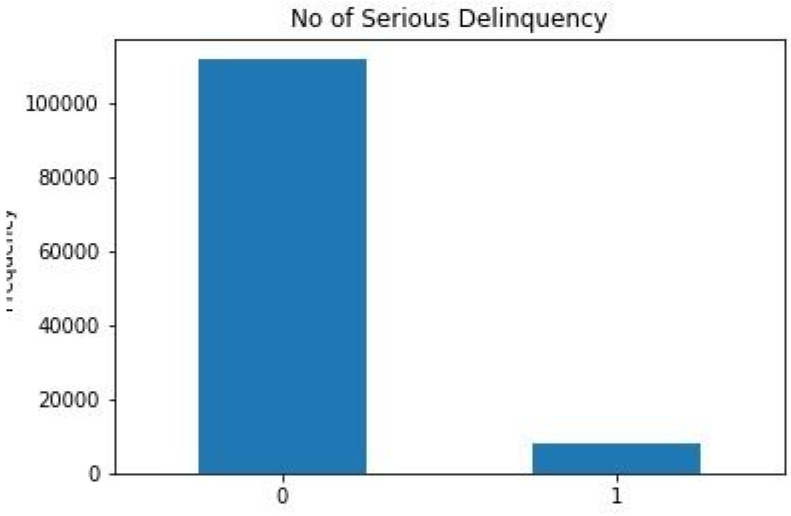
Variable Name	Description	Type
SeriousDlqin2yrs	Person experienced 90 days past due delinquency or worse	Y/N
RevolvingUtilizationOfUnsecuredLines	Total balance on credit cards and personal lines of credit except real estate and no installment	percentage (float)
Age	Age of borrower in years	integer
NumberOfTime30-59DaysPastDueNotWorse	Number of times borrower has been 30-59 days past due but no worse in the last 2 years	integer
DebtRatio	Monthly debt payments, alimony, living costs divided by monthy gross income	percentage (float)
MonthlyIncome	Monthly income	float
NumberOfOpenCreditLinesAndLoans	Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards)	integer
NumberOfTimes90DaysLate	Number of times borrower has been 90 days or more past due	integer
NumberRealEstateLoansOrLines	Number of mortgage and real estate loans including home equity lines of credit	integer
NumberOfTime60-89DaysPastDueNotWorse	Number of times borrower has been 60-89 days past due but no worse in the last 2 years	integer
NumberOfDependents	Number of dependents in family excluding themselves (spouse, children etc.)	integer

As it can be seen, many of the features deal with the basic information of the individuals, such as Age, as well as the financial history or asset of the individual to gauge the possibility of repayment of the debt.

While processing the data, we came across 29731 cases of missing values for Monthly Income, and 3924 cases of missing values for Number of Dependents.

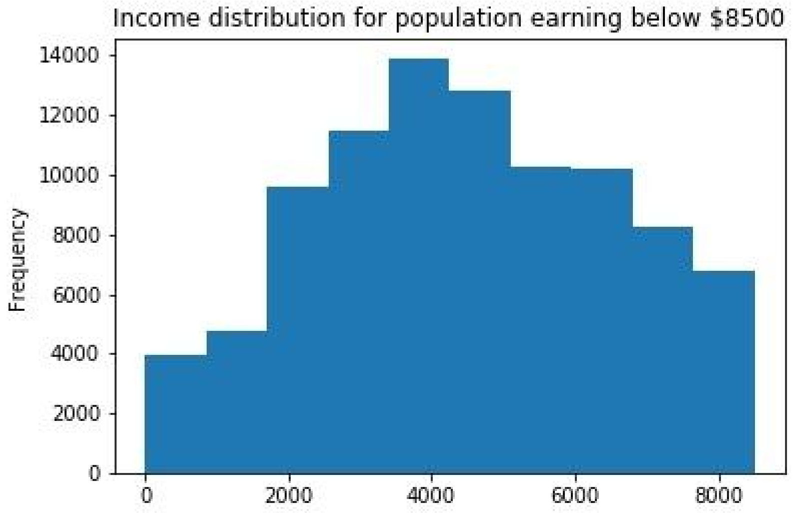
We do not yet have pairwise relationship between all the features, but through examining the features with intuition, we can speculate which features could be related to others. For example, Debt Ratio by definition is the ratio between the debt and the income, so we can be sure that Debt Ratio and Monthly Income will be inversely proportional, and more debt and less income would mean more likelihood of financial hardship, which would increase the chance of missing the payment and experiencing due delinquency. Also, as age is increased, it is more likely that the individuals have families and have dependents, which would mean that Age is proportional to NumberOfDependents.

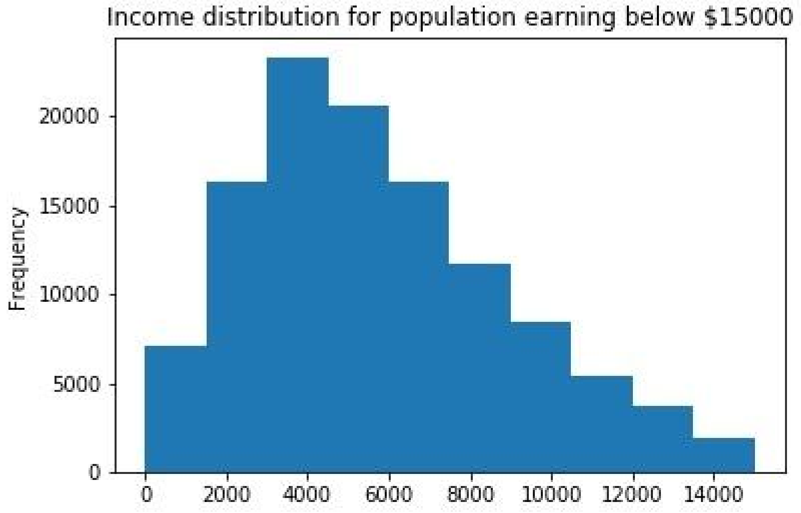
We then examined the distribution of the features to get an overview of what we are dealing with. Firstly, the number of people who have defaulted in paying back the loan by 90 days or more in the past 2 years was very low.



This is heavily skewed which might have the potential to affect the result of the ADS if not properly handled.

We then examined the monthly income of the individuals, which would largely determine the individual’s ability to pay back the debt from a steady stream of income. The vast majority of the population, around 75%, were people who were making less than $8,500 per month, and the overwhelming majority of people who were making less than $15,000 per month were around 95% of the population, so these groups of people in the majority were more closely examined.



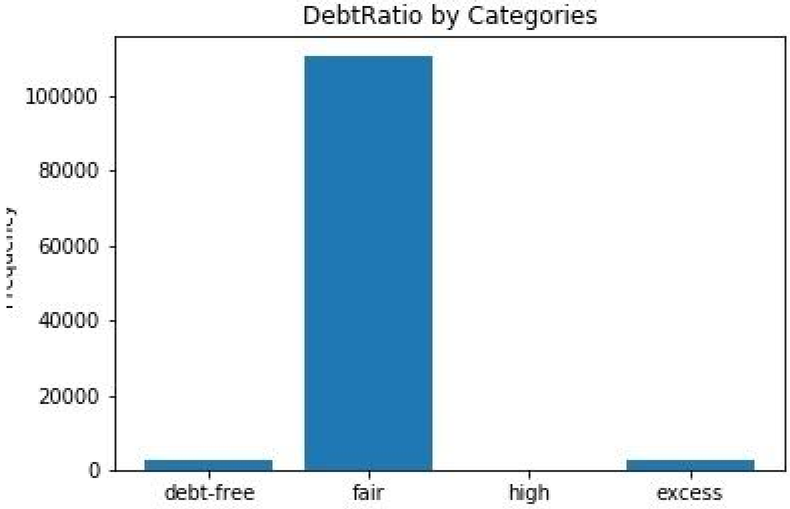


As expected, we see that the mean of both distributions is around $4,000 per month, and it follows an expected distribution of income in our society.

Much more important than the income is the difference in the individual's income versus the debt they have, which is indicated by the variable Debt Ratio. We created four categories from the Debt Ratio to qualitatively understand the distribution.

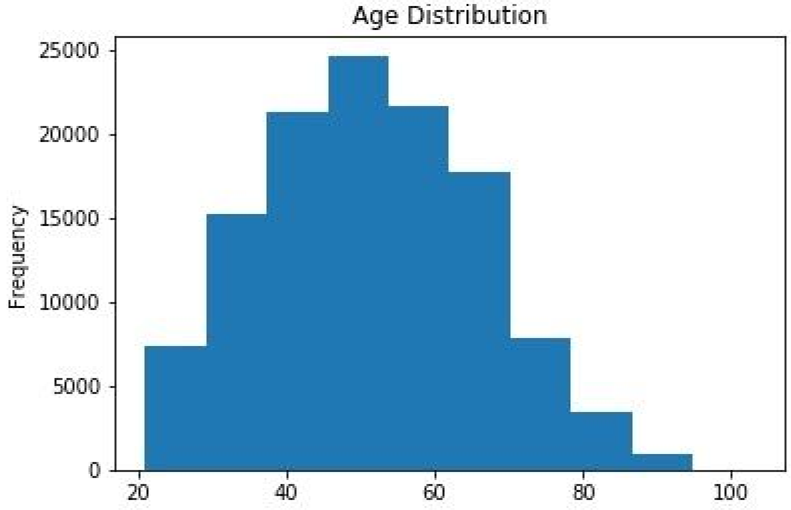
The four categories are:

debt-free	DebtRatio = 0
fair indebtedness	0 < DebtRatio < 1
high indebtedness	DebtRatio = 1
excessive indebtedness	DebtRatio > 1

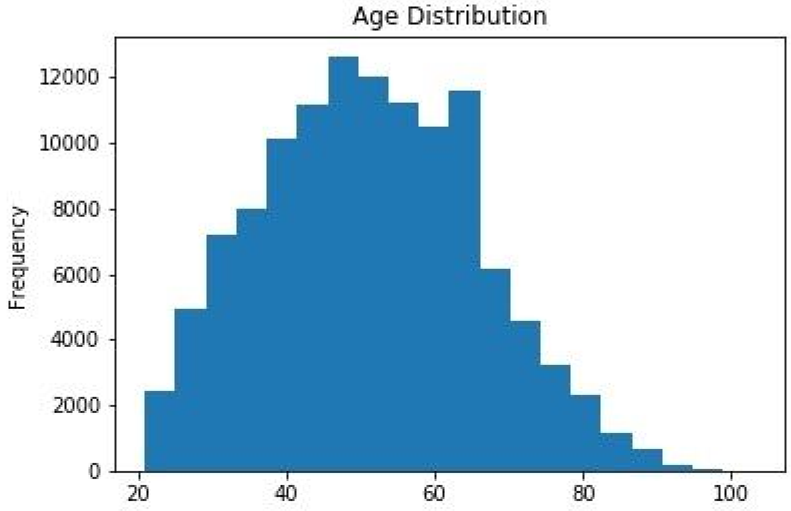


We see that roughly the same number of people are debt-free or excessively indebted, and 24 people fit into the category of high indebtedness since it is very rare that the debt and the income will match exactly equally. Along with the delinquency distribution, we see that the distribution with these four categories are extremely skewed, but note that the purpose of creating these categories were to qualitatively have better understanding of the Debt Ratio.

Looking at more personal information of individuals instead of their financial history or assets, we examined the ages of the individuals.

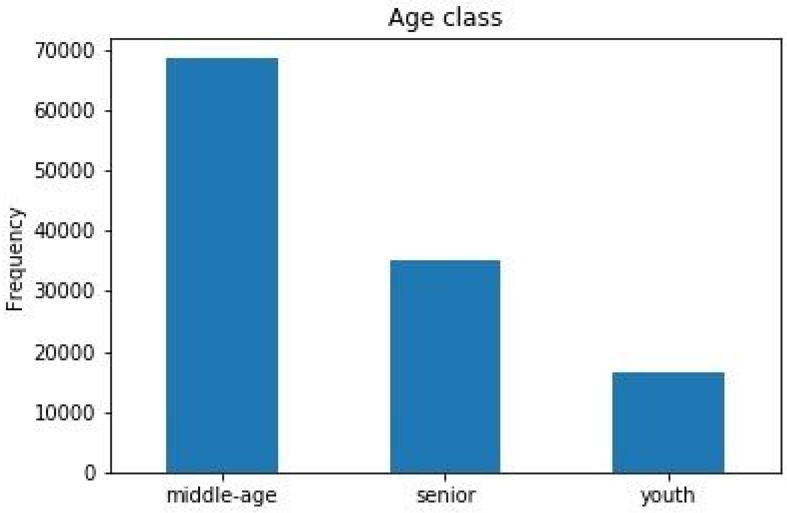


From this distribution, the age, which is a protected attribute in our analysis, seems to follow a normal distribution. However, notice that finer resolution on the age shows an anomaly in the distribution.

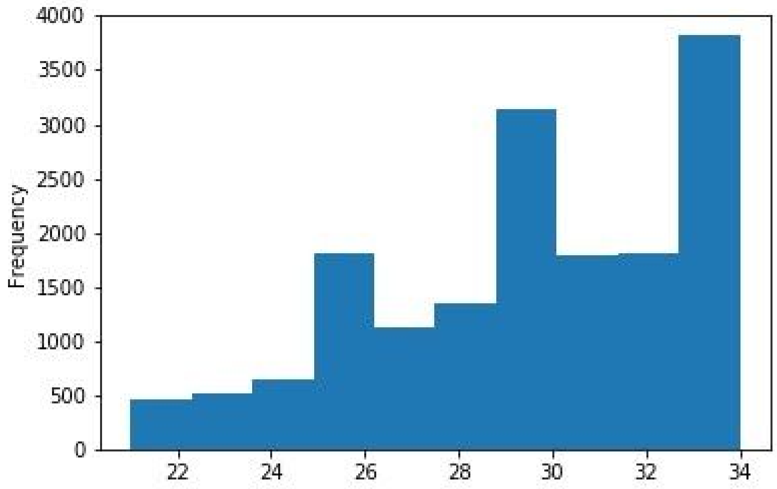


This distribution reveals a sharp decline in the number of population who are over 65 years. This could be caused by fact that most individuals in this age-range are retired and a good number of them could be living off of their retirement savings, which reduces the need for loans.

Again with the age, just like the Debt Ratio, there is a feature to better understand and categorize the age group.



While the middle-age population appeared to be most represented, while the youths were least represented in the distribution. Looking more closely at the youth population,



We see that there are abnormal spikes when they are 25, 29, and 33. This could be due to certain dramatic changes in their lives, such as refinancing their student debt after graduating, buying a car or house, etc, which all are periods in youths’ lives where they need large sums of money to take out.

3. Plan for the future

As mentioned before, having table of pairwise relationship between all the features would provide us with better picture for this project.

Our plan is to take the winning model from the competition to get the prediction score, and examine the misclassified individuals or classes in order to develop a hypothesis on why the model provided misclassification or skewed classification on those individuals. We then plan on running a test on the hypothesis by doing a more in-depth analysis using LIME and Shap, and evaluate the model on fairness metrics using age and number of dependents as protected attributes.

Finally, if time allows, we plan on examining the results of the model after being trained with privacy preserving data that was generated from the original data, but with different privacy methodology.