Analysis of Home Credit Group's Credit Default Risk

Nirav P. Sheth

I. Introduction:

Mortgages or home loans are a 5-30 year loans given to consumers to typically purchase property. Before the mortgage lender provides the fund, they must predict if the applicant will be able to pay back their loans with a variety of credit and related information. On the other side, the applicant purchasing the home wants to find a loan that suits their needs. Traditionally, loans are provided to applicants with adequate credit history. However, the unbanked population, consumers without adequate credit history, will go to an untrustworthy lender whom could take unfair advantage of their situation because reputable banks will not provide them with loans.

However, there could be other ways to determine the credit worthiness of applicants. For this project, I will be testing if non-traditional sources of credit history can determine if the applicants will repay the loan?

A. Client:

Home Credit Group goal is to broaden financial inclusion for the unbanked population. Their vision is to provide better services to the unbanked population to ensure fair and equal opportunities to purchasing property. Overall, if they can better predict the likelihood of loan repayment, then they can expand their services to a larger population and reduce their cost to service the loans.

B. Data:

The primary data provided in CSV format by Home Credit Group are the current applicants financial information, previous applicants, available and relevant information by the Credit Bureau, remaining balances on existing loans, and credit card information. The current applicants has about 308K applications. In addition, they also provided ~2M to 13M lines of supporting credit history information.

A dataset has a mix of behavioral, descriptive, and credit history. The main difficulty is that about half of the clients do not have historical information, so the use of behavioral and descriptive data will be an important source to judge the clients.

C : Methodology:

I will need to develop the dataset for the machine learning by using data wrangling techniques, exploratory data analysis, statistical analysis, and machine learning. A common connector of data will be the applicant ID generated per client. The next steps will be to gain some deeper understanding of the data through statistical analysis and generate some baseline knowledge. Then I will use supervised machine learning to generate a probability of loan repayment. The probability for each applicant will help determine the likelihood of repayment, but the actual decision is binary (yes or no). Therefore, I will plot the probability results onto a histogram and determine the cutoff value or the value that determines full repayment of loan. The optimal cutoff value will be determined based on the value of

the ROC to the application test dataset. In the end, a binary classifier will be given to each applicant to determine if they will repay the loan.

D : Deliverables:

The deliverables for this project are finalized proposal, repository on Github, explanation of methodology with code, and finalized report.

Capstone Link

https://github.com/nervster/CapstoneProject1
https://www.kaggle.com/c/home-credit-default-risk

II. Data Cleaning

My datasets consists of seven fairly clean because it came from Kaggle. The seven datasets are one current applicant dataset with six historical informational datasets. An important cleaning step was to aggregate the historical datasets to convert from the '1:many' to '1:1' relationships, so I will can merge it into the current applicant dataset. For example, one current applicant could have multiple previous loans. Thus, I used Pandas' pivot_table function to help with the aggregation. I focused one 3 main data sets: Previous_Application, Bureau, and Installment Payments because it had a largest amount of informative data such as previous loans, type of loans, and loan payment habits.

For the Previous_Application dataset, it has 37 columns with 15 columns with NaN values. To aggregate the 37 columns, I used the sum or Numpy's mean methods if the column was numerical. If the columns were a category or string based, then I used 'lambda x: len(x.unique())' to count the unique categories that each applicant was part of. I repeated the same thought process for Bureau and Installment Payments datasets. As a result, I ended up with a dataframe with 185 columns with a mix of Float, Integer and Object data types.

Missing Values/Null Values

This dataset has an extremely large amount of missing values (ranges from 99.6% missing values to 0%) due to the aggregation step and lack of historical information for current applicants. To reduce the missing values or null values, I used the fillna method with the 'inplace=True' parameter to replace all null values to '0'. For simplicity, I decided to use '0' as the dataframe's null value. This step is extremely important because it will help allow the feature elimination model to find the best columns that provides the best insight to this problem.

The current threshold is 1E-8, however, the threshold should increase as I improve this model. The low threshold is most likely due to amount of null values being used across the dataframe. This is also an indication that using more efficient null value techniques could be a method to help improve my model.

Outliers/Feature Elimination

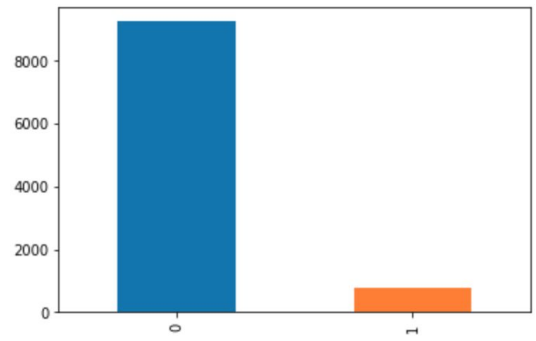
Before this step, I had 185 columns with a mix of Float, Integer, and Object data types. To prevent overfitting the future machine learning model, I need to decrease the number of columns in my dataframe. Therefore, I decided to use the 'SelectFromModel' meta-transformer with the 'LassoCV' classifier.

The LassoCV works with only numerical data types, so I used the Pandas' get_dummied method to convert all of my 'Object' based columns into a numerical. I was then able to use the SelectFromModel with a 1E-8 threshold limit to reduce my dataframe to 15 columns. From here, I will be continue my exploration and statistical analysis to find additional insights and improve my model.

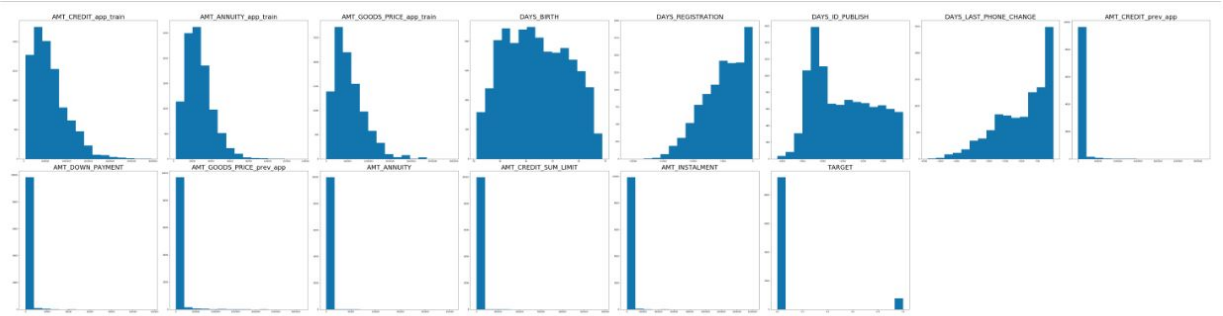
Data Storytelling

My data has many incomplete values but I believe that is important to the overall story because we are trying to gain a better understanding of the group of people without credit information. I started by creating a correlation graph to the bar chart. I also wanted to see how many delinquent loans the dataset had. Upon further research, I noticed that Home Credit Group had twice the amount of delinquent loans than the United States' national average. (Source: Delinquency Rate for US)

```
Percent of Deliquent:  7.75 %
Percent of Non-Delinquent:  92.25 %
```



Other characters about my dataset is that the average age of my population is about 44 years old. I compared the delinquent loans age to non-delinquent and noticed that younger generation are more likely to be delinquent. The lower income population tend to be less delinquent, but the price of the loan did not seem to have that large of an impact. Lastly, I generated a histogram plot of all loans. I will be able to use this plot to refer back to during the later steps.



Data Storytelling

Are there variables that are particularly significant in terms of explaining the answer to your project question?

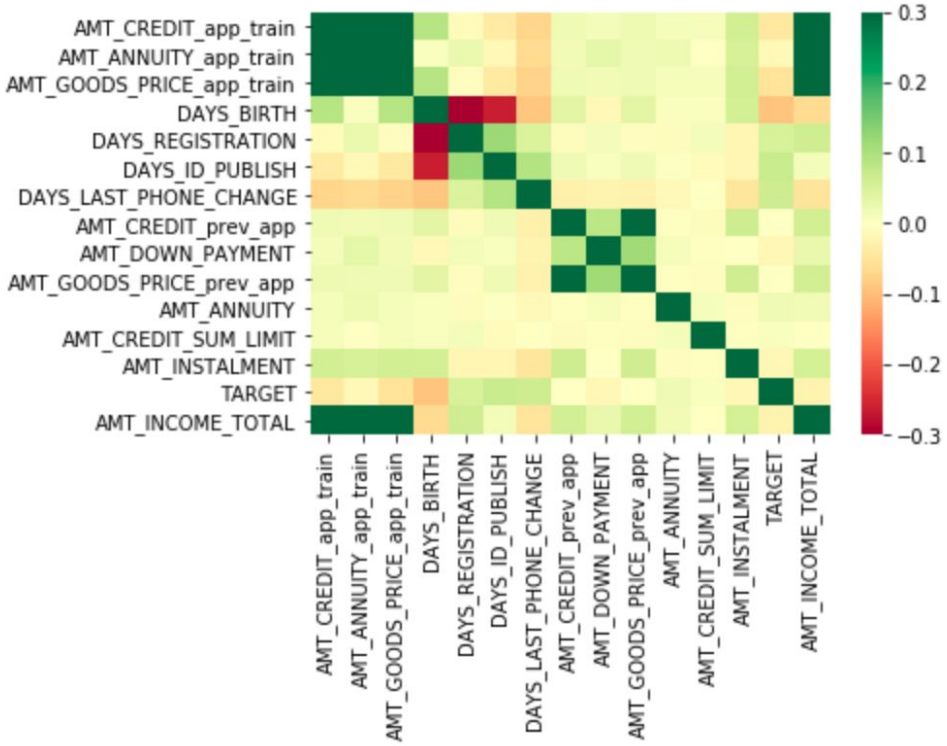
Using the Pearson R Correlation method, I have identified 4 columns that are extremely important to the TARGET data.

1. DAYS_ID_PUBLISH: This columns identifies the number of days the client changed the identity on the application. This column has a Pearson R correlation of .071. This dataset shows that clients with payment difficulties tend to have a larger or changed their application closer to the days published.
2. DAYS_LAST_PHONE_CHANGE: This columns identifies the number of days the client changed their phone number. This columns has a Pearson R correlation of .067. This dataset shows that clients with payment difficulties tend to have a larger or changed their phone closer to the application date.
3. DAYS_REGISTRATION: This columns identifies the number of days the client changed their application. This columns has a Pearson R correlation of .053.

4. DAYS_BIRTH: This column identifies the client's age. This columns has a Pearson R correlation of .093. We tend to see younger clients that are more delinquent on their loans.

This columns are quite interesting for the fact that columns 1-3 are behavioral based columns while column 4 is descriptive based. While the correlation is low, it is the highest of the features.

<u>Are there strong correlations between pairs of independent variables or between an independent and a dependent variable?</u>



Using the Seaborn library and heatmap function, I plotted the Pearson R correlations onto a heat map and saw a few observations. I decided to use .3 to -.3 as the range because this dataset has very low correlations, so it allows us to see the correlations in further detail.

1. Positive Correlation:
    a. AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE, and AMT_INCOME
    b. AMT_INSTALMENT, AMT_GOODS_PRICE, AMT_CREDIT, AMT_ANNUITY,
2. Negative Correlation:

a. TARGET against AMT_CREDIT, AMT_ANNUITY, and AMT_GOODS_PRICE.
b. DAYS_BIRTH against AMT_CREDIT, AMT_GOODS_PRICE, DAYS_REGISTRATION, DAYS_ID_PUBLISH

While there are strong correlations, I do not see many low correlations. With that said, I believe that identifying how the TARGET doesn't correlate with the size of the loan is helpful to narrow the scope of where to help the business to focus their efforts. Overall, it seems like the delinquencies seems to be a behavioral factor. I will identify further testing going forward.

What are the most appropriate tests to use to analyse these relationships?

I would like to test how behavioral factors affect delinquency. To design this test, I would use columns 1-3 (identified above) against TARGET. I would Z-Test to test the difference of means between TARGET with 0 and 1 per column is significant. In other words, I will split each columns into 2 group: first column with TARGET identified with 0 and second column with TARGET identified with 1. The null hypothesis would test if there is no significant difference and the alternative hypothesis would test if there is a significant difference.