PROJECT REPORT

Prithvi Shivashankar

Executive Summary

Increasing number of people struggle to get loans due to a lack of or insufficient credit history. There is therefore a higher need of financial inclusion as this segment of population get taken advantage of by loan sharks and lenders who often are not regulated by the governmental financial organizations. This has resulted in widespread adoption of predictive analytics to be able to analyze people's alternate data such as telco, transactional, previous application data and so on to be able to predict whether such people can repay the loans or not. I have utilized the Home credit Default risk dataset which was part of the Kaggle competition. I used several classification algorithms to predict the repayment capability of Home Credit's clients.

A. Introduction

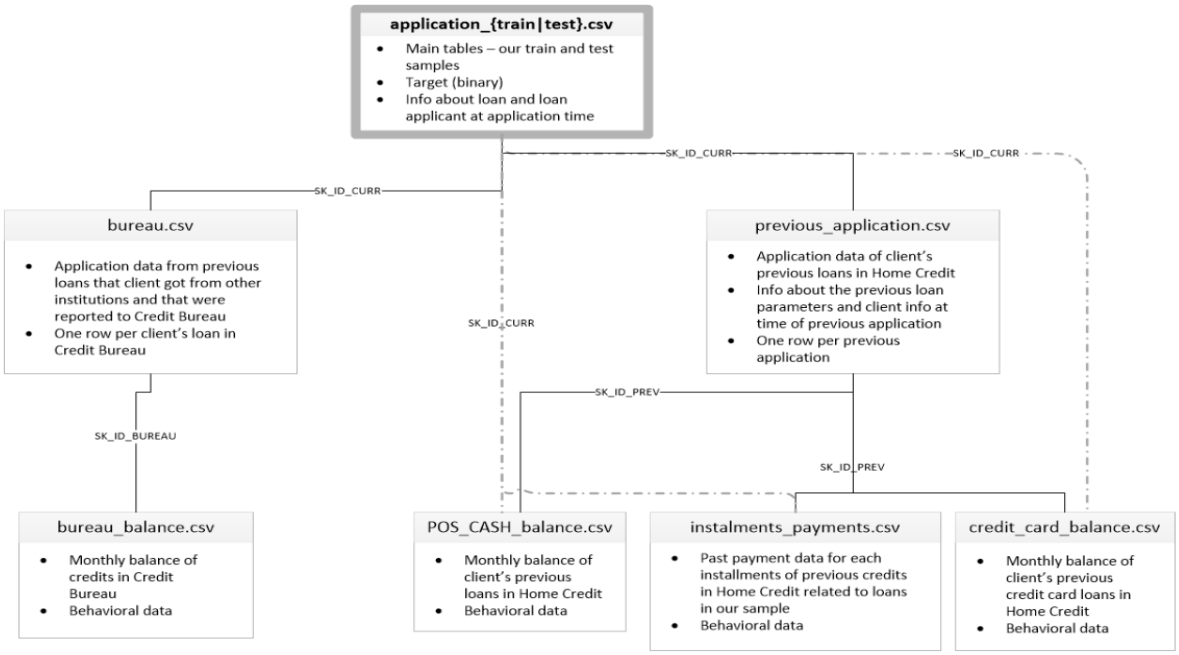
Home Credit is a non-banking financial institution based out of the EU. Their main emphasis in on lending to people with little or no credit history. They use a lot of statistical as well as Machine Learning algorithms to make predictions on a customer's credit worthiness.

There is need for more financial inclusion of people with little or no credit history because: (i) It empowers them to strive for higher financial independence; (ii) Helps them not fall into the debt trap of untrustworthy lenders. For the unbanked populations, there is a need for better algorithms to determine whether they can pay back the loans or not from a variety of alternate data.

B. Data Description

There are 8 different datasets with the main one being the application dataset that contains 300000 records, all the applicant information and each record corresponds to the one loan application. In addition to this, there is the bureau dataset, which consists of all the previous credits provided by other financial institutions to the Credit Bureau and there can be multiple credits for each client (multiple rows for one applicant for the current loan). There is the bureau balance dataset which contains information about the monthly balances of previous credits reported to the Credit Bureau. The POS (Point of Sales) cash balance dataset has details relating to the pervious point of sales and cash loans the current loan applicant had taken. The Credit Card Balance dataset has details of the previous credit cards the current loan applicants had. Additionally, we have the previous application dataset that has details of the previous Home Credit loans for current Home Credit loan applicants. Finally, there is the installment payments dataset which contains information regarding previous repayment history of clients for the previous Home Credit loans.

The relation for the dataset can be found below:



C. Preprocessing Data

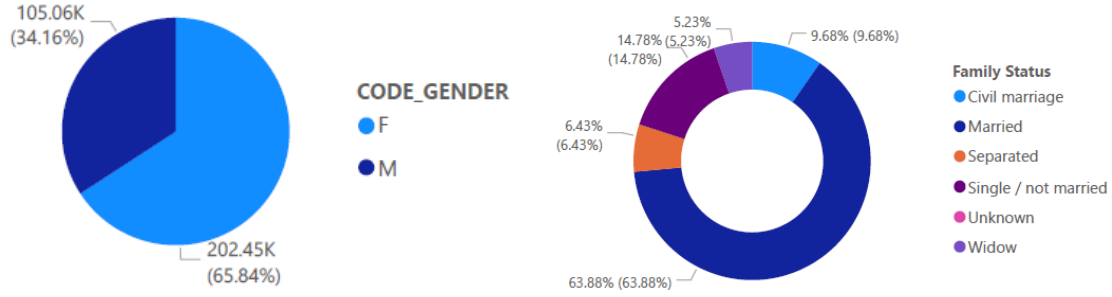
A lot of variables in each of the datasets had many missing values some even ranging to 50-60%. I used the VIM library in R to check for the patterns in missing data within the applicant dataset and mostly found them to be missing completely at random. Based on this finding, I used the MICE imputation (PMM) technique to impute over 32 variables with missing values in the Application dataset.

(MICE imputation in the BUAN6357_Shivashankar_Feature_project.R file)

The next challenge was that for each applicant value, there were multiple rows in other datasets such as the previous application, bureau dataset and so on. While joining each table based on the unique applicant Id's, for each current Home Credit applicant, there were multiple columns and hence there was a need to perform aggregation techniques on each row for the 7 other datasets. Since, this can be a very time consuming process, I have utilized the FeatureTools package in Python to automate the feature Engineering process and retrieve the final dataset which contains the sum, median, average and mode for each of the applicant ID's relating to the previous credit information and so on. In addition to the FeatureTools package, I utilized the Dask library in Python to parallelize the Feature Engineering process and reduce the overall computation time. I had retrieved over 1100 features after performing this operation. I had then applied the Boruta feature selection algorithm on a sample of 5000 data points to obtain the most important features contributing to the variation of the Target variable and was able to overall achieve 35 important features to build the predictive models. **(FeatureTools and Dask in the Feature_Engineering.ipynb file)**

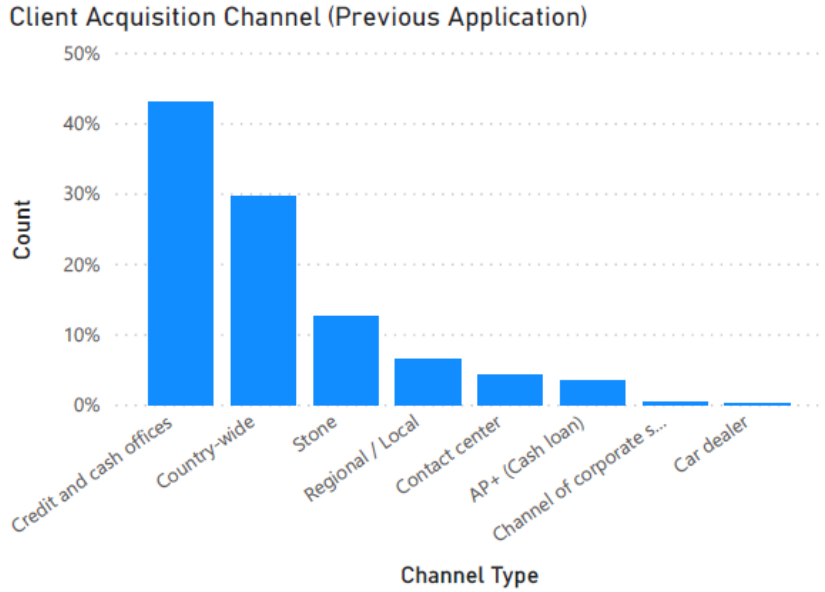
D. Exploratory Data Analysis

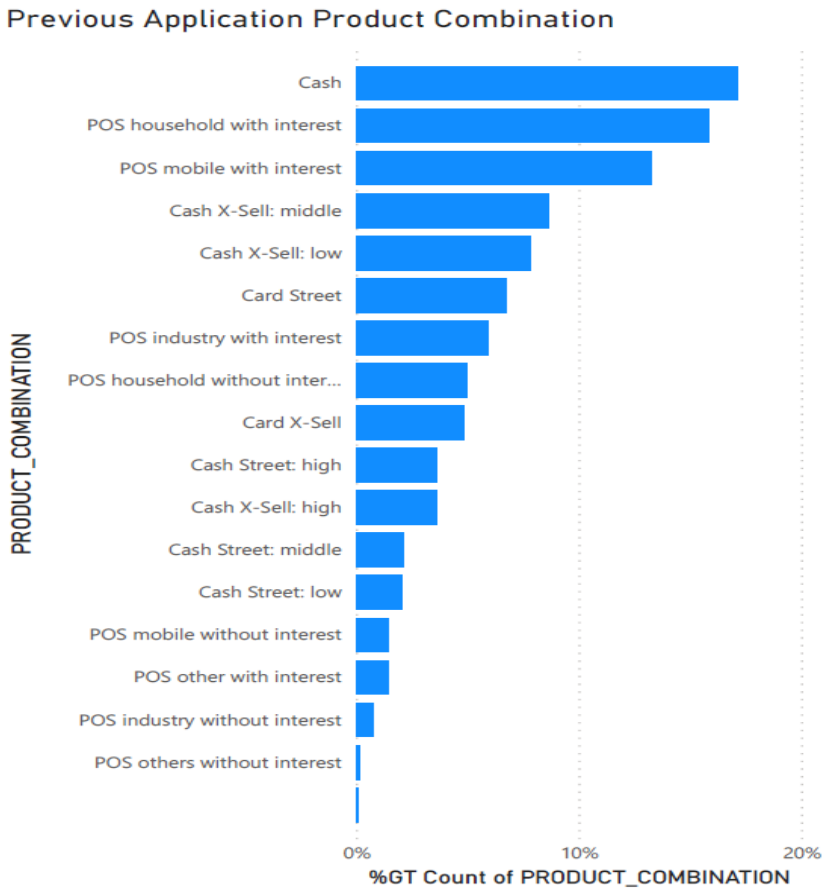
The below pie chart shows the distribution of male and female applicants as well as the family status in this dataset.



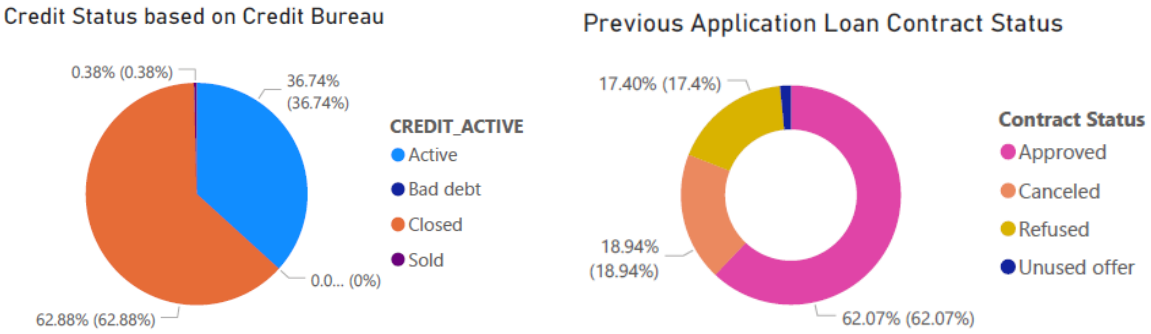
I then investigated the loan contract type as well as the proportion of applicants who repaid the loan and it is evident that the cash loans are the most sought after and around 90% of the applicant repaid the loans.

I have further looked into the previous loan applicants' product combinations and the main client acquisition channels. We can see that the credit and cash offices is the main channel for customer acquisition and that cash loan was the most preferred product for the previous applicants.





Finally, I analyzed the Credit Bureau status and the previous loan contract status to get an idea of how many loans were approved and whether some of the previous applications were bad debt or not.



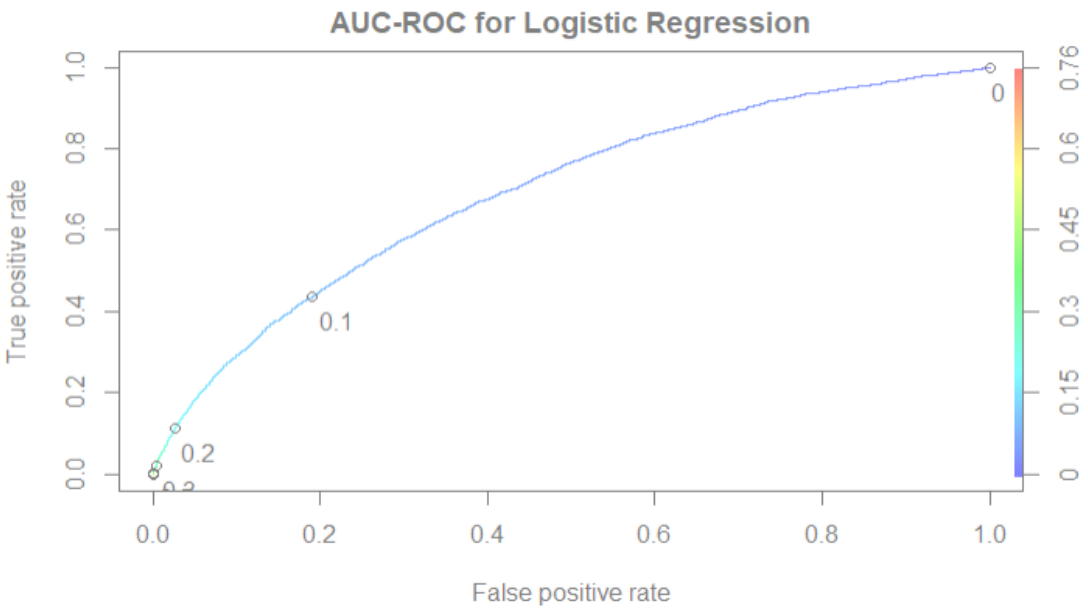
From the above 2 visualizations, we can see that not a high amount of previous applications was refused, and negligent amount of Credit Bureau credit status was a bad debt.
Hence, after a preliminary analysis, we can see that most people within the unbanked populations are able to repay the loans and using sophisticated machine learning algorithms, we would be able to better gauge and reduce of the risk of bad loans.

E. Empirical Analysis

Logistic Regression:

I have initially used a logistic regression model to predict the TARGET variable – TRUE/FALSE. TRUE, if the customer has payment difficulties and FALSE, if the customer can repay the loan. Logistic regression results below:

```
[1] "Train Confusion matrix:"

        FALSE  TRUE
  False 97422    11
  True   7640     6
[1] "Overall train results:"

Point estimates and 95 % CIs:
-------------------------------------------------------
Apparent prevalence                  0.93 (0.93, 0.93)
True prevalence                      1.00 (1.00, 1.00)
Sensitivity                          0.93 (0.93, 0.93)
Specificity                          0.35 (0.14, 0.62)
Positive predictive value            1.00 (1.00, 1.00)
Negative predictive value            0.00 (0.00, 0.00)
Positive likelihood ratio            1.43 (1.01, 2.04)
Negative likelihood ratio            0.21 (0.11, 0.39)
-------------------------------------------------------
[1] "Train accuracy rate: 0.927188115608257"
[1] "Train error rate: 0.0728118843917434"


[1] "Test Confusion matrix:"

        FALSE  TRUE
  False 33825     3
  True   2717     3
[1] "Overall test results:"

Point estimates and 95 % CIs:
-------------------------------------------------------
Apparent prevalence                  0.93 (0.92, 0.93)
True prevalence                      1.00 (1.00, 1.00)
Sensitivity                          0.93 (0.92, 0.93)
Specificity                          0.50 (0.12, 0.88)
Positive predictive value            1.00 (1.00, 1.00)
Negative predictive value            0.00 (0.00, 0.00)
Positive likelihood ratio            1.85 (0.83, 4.12)
Negative likelihood ratio            0.15 (0.07, 0.33)
-------------------------------------------------------
[1] "Test accuracy rate: 0.925577322972529"
[1] "Test error rate: 0.0744226770274707"
```



Based on the above results, it is observed that logistic regression does provide a high level of accuracy. The test accuracy is 92.5% and based on the AUC-ROC curve, the model is doing a good job in correctly classifying clients who do not have payment difficulties and are less likely to default. On the other hand, the specificity for the test dataset is 50%, which means that only 50% of the people who are likely to have payment difficulties are classified correctly.

Gradient Boosting Model:

After running a few other models, most of which did not yield very favorable results, I ran the Gradient boosting with a 5-fold Cross Validation. Model results:

```
Confusion Matrix and Statistics

          Reference
Prediction False  True
     False   273    37
     True  33555  2683

               Accuracy : 0.0809
                 95% CI : (0.0781, 0.0837)
    No Information Rate : 0.9256
    P-Value [Acc > NIR] : 1

                  Kappa : -8e-04

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.008070
            Specificity : 0.986397
         Pos Pred Value : 0.880645
         Neg Pred Value : 0.074038
             Prevalence : 0.925577
         Detection Rate : 0.007470
   Detection Prevalence : 0.008482
      Balanced Accuracy : 0.497234

       'Positive' Class : False
```

Based on the above results, it is observed that even though the accuracy (8%) is very low, the specificity (98%) is very high, which means that 98% of the time, the model is correctly classifying the group of people who have payment difficulties in this case.

(Machine Learning code in the BUAN6357_Shivashankar_Final_project.rmd file)

F. Conclusions

Using a large dataset consisting of customer's data relating to where they live, building information, previous loan as well as credit card applications, I analyzed and derived insights regarding what causes a customer to default or be able to pay back the loans. I have concluded that the score of clients from external data source, previous credit information as well as information from previous loan installment payments have a major role to play in deciding whether they are able to pay the loan or not. Through such data, banks and other lending institutions should be able to extend loans to the unbanked populations.

Sources:

Home Credit Default Risk Kaggle competition:
https://www.kaggle.com/c/home-credit-default-risk

Magazine Article
Alternative Data: The Great Equalizer to Lending Inequalities?
https://www.forbes.com/sites/forbestechcouncil/2019/08/14/alternative-data-the-great-equalizer-to-lending-inequalities/#1c30e27d2449

Dask Documentation:
https://docs.dask.org/en/latest/

Boruta Feature Selection process in R:
https://www.datacamp.com/community/tutorials/feature-selection-R-boruta

Deep Feature Synthesis(FeatureTools):
http://www.jmaxkanter.com/static/papers/DSAA_DSM_2015.pdf