**Technical Report**

**Predicting Default Risk of Home Credit Loans**

**Introduction**

**Project Overview**

Credit worthiness is a very interesting domain of lending used to assess an applicant’s probability to repay the credit. It is used by the Home Credit Group to identify those customers who might not repay loans. It is not only important for Home Credit to maximize its returns in the form of interest paid on the borrowed amount, but also to find out if the applicant is a kind of person who will repay the loan and if so what characteristics make a borrower less likely to default.

This is a perfect machine learning application since it helps Home Credit to turn the information they have about all the applicants into something very useful in terms of sound lending decisions. This will enable them to provide more loans to more worthy applicants.

**Problem Statement/Objective**

The objective of this analysis is to use historical loan application data to derive which applicant features better predict whether or not an applicant will or will not repay a loan.

**Dataset Information**

The dataset is from Home Credit Group and retrieved through Kaggle. The dataset contains in depth personal and financial information about the applicants who had been previously been granted loans by Home Credit. It also contains applicants information on the financial background like how have their repayment behaviour been in the past of their loans from Home Credit and other lenders along with their credit card payments. The dataset is composed of seven sets of attributes split into eight files: application\_train/test, bureau, bureau\_balance, previous\_application, POS\_CASH\_balance, instalments\_payments, and credit\_card\_balance.

*Application\_train/test* contains non-identifying demographic information about each borrower, details of their loan and the target variable(repaid or not repaid).It contains information about marital status, education, employment, housing, age, etc.

*Bureau* contains information reported to the credit bureau about previous or concurrent loans held by the borrower at other institutions.

*Bureau\_Balance* contains the reported monthly balance of each loan per borrower.

*Previous\_application* contains information about each of the applicants previously held loans with Home Credit; this file also contains borrower demographic info similar to the application file.

*POS\_CASH\_balance* contains monthly balance data of each of the loans held by the borrower at Home Credit.

*Instalments\_payments* contains monthly balance information of the current loan held by the borrower with Home Credit.

*Credit\_card\_balance* contains monthly balance information for the borrower’s previous credit card loans with Home Credit.

A figure(figure 1.1) of the seven documents and their connection can be found in the Appendix A and a complete list of column descriptions are available in the file, HomeCredit\_columns\_description.csv.

The final merged the dataset contained a total of **307511** observations and **376** variables.

**Methodology**

**Data Preprocessing**

1. Joining/Stitching the 7 data files:

* Recoded the required Categorical variables to Numeric fields by Frequencies for all files except application train and test
* Aggregating: Bureau <-- Bureau Balance using mean value of numerical fields for Bureau Balance
* Aggregating: previous\_application <-- POS\_CASH\_balance using mean value of numerical fields for POS\_CASH\_balance
* Aggregating: previous\_application <-- installments\_payments using mean value of numerical fields for installments\_payments
* Aggregating: previous\_application <-- credit\_card\_balance using mean value of numerical fields for credit\_card\_balance
* Aggregating: application\_train/test<-- bureau\_agg using mean value of numerical fields for bureau\_agg
* Aggregating: application\_train/test<-- previous\_application\_agg using mean value of numerical fields for previous\_application\_agg

1. Handling Missing values:

Features with more than 60% missing values and unique value = 0 were removed. The rest of the missing values were imputed using Median strategy.

1. Handling Outliers:

It is observed that the variable DAYS\_EMPLOYED has very large values like 365243 days (1000 years) which seems to be invalid data and these observations were replaced by Nulls.

1. Label and one hot encoding:

Categorical features with 2 unique values were Label encoded. The label encoded features were: NAME\_CONTRACT\_TYPE, FLAG\_OWN\_CAR and FLAG\_OWN\_REALTY. Categorical features with more than 2 unique values were one hot encoded

1. Negative values in “DAY” fields:

Exploratory analysis of the day fields uncovered that the days were calculated with a reference of the dataset creation date and hence had negative values. Based on the analysis of these fields, the day field were either converted to years or their absolute values were considered in the further analysis

1. The numerical fields were Min/Max scaled for further analysis

**Feature Selection**

1. Random Forest Feature Importance

Random Forests and other ensemble methods are excellent models for classification tasks. They don’t require as much preprocessing as some other methods and can take both categorical and numerical variables as input. This process controls for overfitting and can often produce a very robust, high-performing model. Also, they have a super helpful feature called feature importance. These methods are most often used for prediction, but looking at the feature importance can give a sense of which of your variables have the most effect in these models. Hence, we decided on selecting the top 120 features for model building.

1. Principal Component Analysis (PCA)

Principal Component Analysis is one of the techniques that is used to reduce the dimensions when it comes to the problem statement having a lot of features and variables in question. The approach was to execute PCA on the features before feeding the features to the Logistic Regression. Reducing the dimensionality of features helps in performance of the learning algorithm, specifically the logistic regression because LR assumes all the features are independent. PCA works by determining a lower dimensional features space that can project the entire featurespace of higher dimension. This actually reduces the redundant and correlated features through eliminating the data.

In this analysis, we selected 46 principal components which explained 95% of the variance for the model.

**Model Selection -**

Using the 120 selected features the data was fitted with KNN, Naive Bayes, LDA, Decision Tree, and Logistic Regression.

The following metrics have been used to measure the implementation and results of the classifiers and algorithms used:

1. **ROC/AUC**

The area under the ROC (receiver operating characteristic) curve is the perfect measure of the classifier’s performance. The size of the area determines how well the classifier was able to predict the behavior of the applicant. The AUC can range from 0 to 1. The area of 1 indicates that the classifier perfectly identified each and every applicant weather they would repay the loan or not and the area of 0 means that the classifier is not able to identify and label any of the true positives. An area of 0.5 means that the classifier randomly labels a point. As the objective of our classifier is to classify the segment of the applicants, AUC/ROC is the perfect metric to measure the performance

1. **Classification Accuracy**

It is measured by calculating the ratio of correct predictions to the total number of input points to the algorithm. This is also one of the metrics that is suitable to this problem. Classification accuracy is sometimes misleading, specifically in the case of imbalanced data. In our problem statement this was the case and hence we tried to take care of the imbalance in the data by using the hyper parameter that balanced the input samples.

1. **F1 Score (Precision & Recall)**

F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances). High precision but lower recall, gives you an extremely accurate, but it then misses a large number of instances that are difficult to classify. The greater the F1 Score, the better is the performance of our model. F1 Score tries to find the balance between precision and recall.

1. **Precision**

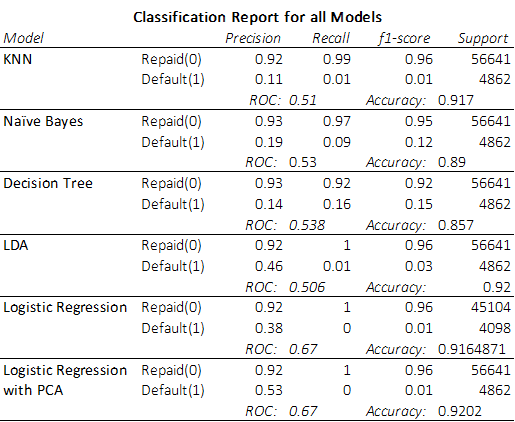
It is the number of correct positive results divided by the number of positive results predicted by the classifier.

1. **Recall**

It is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive).

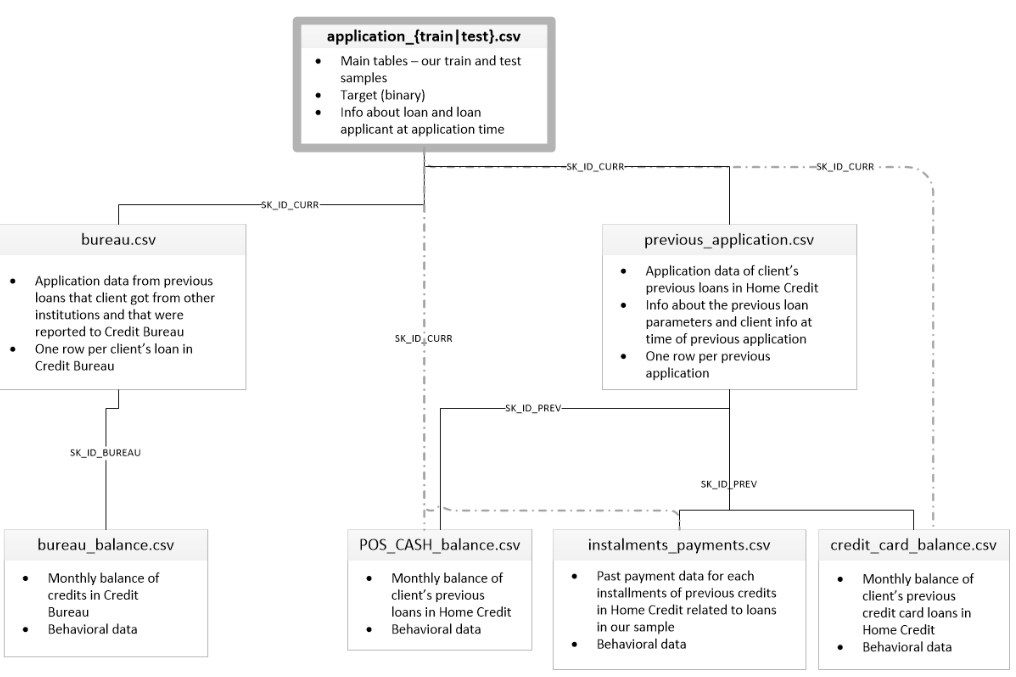
**Discussion / Conclusion:**

Logistic regression is the most suitable algorithm that is used to classify binary classification problems. This was confirmed by our analysis. As we can see the distance similarity classification algorithms have a low ROC and precision for class - 1 as compared to Logistic Regression, therefore there is more support for a Logistic Regression model.



**Appendix:**

1. Descriptions of each of the data files.
2. **Application:** This is the main training and testing data with information about each loan application at Home Credit. Every loan has its own row and is identified by the feature SK\_ID\_CURR. The training application data comes with the TARGET indicating 0: the loan was repaid or 1: the loan was not repaid
3. **Bureau:** Data concerning client's previous credits from other financial institutions. Each previous credit has its own row in bureau, but one loan in the application data can have multiple previous credits.
4. **Bureau\_balance:** Monthly data about the previous credits in bureau. Each row is one month of a previous credit, and a single previous credit can have multiple rows, one for each month of the credit length.
5. **Previous\_application:** Previous applications for loans at Home Credit of clients who have loans in the application data. Each current loan in the application data can have multiple previous loans. Each previous application has one row and is identified by the feature SK\_ID\_PREV.
6. **POS\_CASH\_BALANCE:** Monthly data about previous point of sale or cash loans clients have had with Home Credit. Each row is one month of a previous point of sale or cash loan, and a single previous loan can have many rows.
7. **credit\_card\_balance:** Monthly data about previous credit cards clients have had with Home Credit. Each row is one month of a credit card balance, and a single credit card can have many rows.
8. **installments\_payment:** Payment history for previous loans at Home Credit. There is one row for every made payment and one row for every missed payment.



1. List of features removed for the analysis:
2. Dropping columns with more than 60% missing values
3. Prev\_CCB\_CNT\_DRAWINGS\_ATM\_CURRENT
4. Prev\_FLAG\_LAST\_APPL\_PER\_CONTRACT
5. Prev\_RATE\_INTEREST\_PRIVILEGED
6. Prev\_RATE\_INTEREST\_PRIMARY
7. Prev\_CCB\_AMT\_PAYMENT\_CURRENT
8. Prev\_CCB\_AMT\_DRAWINGS\_ATM\_CURRENT
9. Prev\_CCB\_CNT\_DRAWINGS\_POS\_CURRENT
10. Prev\_CCB\_AMT\_DRAWINGS\_OTHER\_CURRENT
11. Prev\_CCB\_AMT\_DRAWINGS\_POS\_CURRENT
12. COMMONAREA\_MODE
13. COMMONAREA\_AVG
14. COMMONAREA\_MEDI
15. NONLIVINGAPARTMENTS\_MODE
16. NONLIVINGAPARTMENTS\_AVG
17. NONLIVINGAPARTMENTS\_MEDI
18. FONDKAPREMONT\_MODE
19. LIVINGAPARTMENTS\_AVG
20. LIVINGAPARTMENTS\_MEDI
21. LIVINGAPARTMENTS\_MODE
22. FLOORSMIN\_MODE
23. FLOORSMIN\_MEDI
24. FLOORSMIN\_AVG
25. YEARS\_BUILD\_AVG
26. YEARS\_BUILD\_MODE
27. YEARS\_BUILD\_MEDI
28. OWN\_CAR\_AGE
29. Prev\_CCB\_MONTHS\_BALANCE
30. Prev\_CCB\_AMT\_CREDIT\_LIMIT\_ACTUAL
31. Prev\_CCB\_AMT\_DRAWINGS\_CURRENT
32. Prev\_CCB\_AMT\_INST\_MIN\_REGULARITY
33. Prev\_CCB\_AMT\_PAYMENT\_TOTAL\_CURRENT
34. Prev\_CCB\_CNT\_DRAWINGS\_CURRENT
35. Prev\_CCB\_AMT\_RECEIVABLE\_PRINCIPAL
36. Prev\_CCB\_AMT\_RECIVABLE
37. Prev\_CCB\_SK\_DPD\_DEF
38. Prev\_CCB\_SK\_DPD
39. Prev\_CCB\_CNT\_INSTALMENT\_MATURE\_CUM
40. Prev\_CCB\_AMT\_TOTAL\_RECEIVABLE
41. Prev\_CCB\_NAME\_CONTRACT\_STATUS\_Approved
42. Prev\_CCB\_AMT\_BALANCE
43. Prev\_CCB\_NAME\_CONTRACT\_STATUS\_Active
44. Prev\_CCB\_NAME\_CONTRACT\_STATUS\_Demand
45. Prev\_CCB\_NAME\_CONTRACT\_STATUS\_Completed
46. Prev\_CCB\_NAME\_CONTRACT\_STATUS\_Refused
47. Prev\_CCB\_NAME\_CONTRACT\_STATUS\_Sentproposal
48. Prev\_CCB\_NAME\_CONTRACT\_STATUS\_Signed
49. Dropping columns with unique value = 0

* Bureau\_CREDIT\_TYPE\_Interbank credit
* Bureau\_CREDIT\_TYPE\_Loan for purchase of shares (margin lending)
* Prev\_NAME\_TYPE\_SUITE\_nan
* Prev\_PRODUCT\_COMBINATION\_nan

Decision Tree Output

