FLIP ROBO

Micro Credit Loan Project

Submitted by:

Apurv Saraf

ACKNOWLEDGMENT

I would like to express my gratitude towards “FlipRobo Technologies” that gave this opportunity to build this project. I would like thank my SME Mr. Nishant Kadian and data trained Mentor for his guidance in building this project. I would also like to thank Data Trained Education for the immense amount of guidance in Machine learning.

References:

- Dataset used is solely property of FlipRobo
- Kaggle
- Medium
- Towards data science
- Analytics Vidhya
- Stack Overflow

INTRODUCTION

Business Problem Framing

- Micro credit Loan Project is a real life project for a telecommunication network provider. The Telecommunication Network provider company collaborated with a Micro finance institution to provide micro-credit on mobile balances to be paid back in 5 days.
- The Consumer is classified to be a defaulter if he/she do not pay back the loan amount within the time duration of 5 days. For the loan amount of 5 (in Indonesian Rupiah), payback amount should be 6 (in Indonesian Rupiah), while, for the loan amount of 10 (in Indonesian Rupiah), the payback amount should be 12 (in Indonesian Rupiah).
- The client has approached for building a ML Algorithm to select the good customers who will pay back the loaned amount

Conceptual Background of the Domain Problem

The domain related concept would be to analyse loan repay tendency of a segment as target group is low-income group

- Review of Literature

To Build the model as explained by client we have used the client's real time database which comprised of details of individual customer with no null values. However the dataset was imbalanced as 87.5% were non-defaulters.

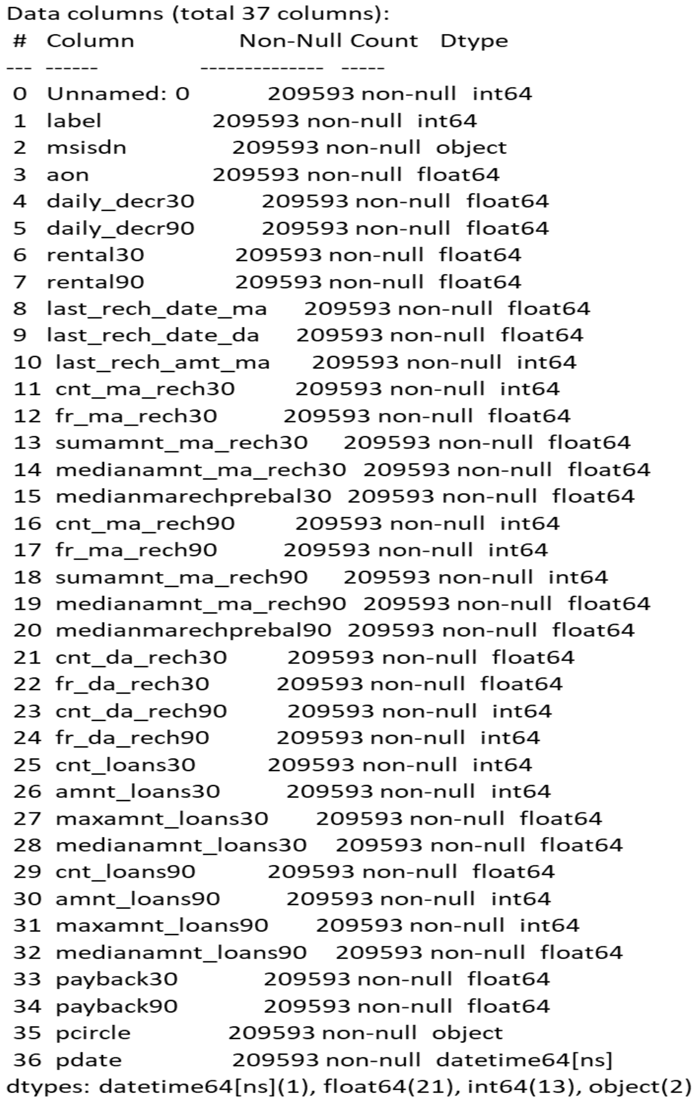
Also as it was a real time dataset comprising of every detail about customer there were 36 features which result to famous curse of dimensionality which we treated using PCA.

- Motivation for the Problem Undertaken

The main motivation behind this project is to understand and analyse the micro finance business , Also to understand if any telecommunication network provider tomorrow comes with such solution of extending its services to micro-lending how well that project will perform and how diversified would be the customers response towards it.

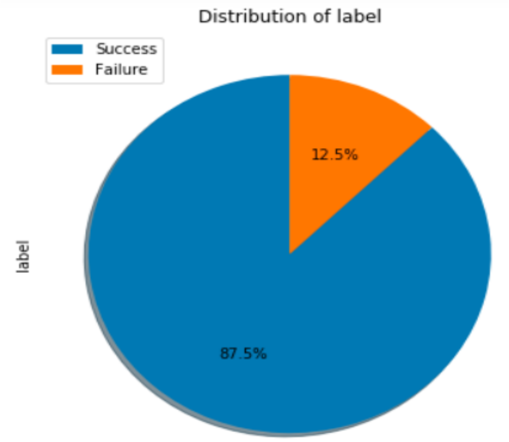
Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem
    1. This is a classic classification problem where target variable is “Label Feature”.
    2. The data set comprised of below features with no null values.



```
Data columns (total 37 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Unnamed: 0            209593 non-null  int64
 1   label                 209593 non-null  int64
 2   msisdn                209593 non-null  object
 3   aon                   209593 non-null  float64
 4   daily_decr30          209593 non-null  float64
 5   daily_decr90          209593 non-null  float64
 6   rental30              209593 non-null  float64
 7   rental90              209593 non-null  float64
 8   last_rech_date_ma     209593 non-null  float64
 9   last_rech_date_da     209593 non-null  float64
 10  last_rech_amt_ma      209593 non-null  int64
 11  cnt_ma_rech30         209593 non-null  int64
 12  fr_ma_rech30          209593 non-null  float64
 13  sumamnt_ma_rech30     209593 non-null  float64
 14  medianamnt_ma_rech30  209593 non-null  float64
 15  medianmarechprebal30  209593 non-null  float64
 16  cnt_ma_rech90         209593 non-null  int64
 17  fr_ma_rech90          209593 non-null  int64
 18  sumamnt_ma_rech90     209593 non-null  int64
 19  medianamnt_ma_rech90  209593 non-null  float64
 20  medianmarechprebal90  209593 non-null  float64
 21  cnt_da_rech30         209593 non-null  float64
 22  fr_da_rech30          209593 non-null  float64
 23  cnt_da_rech90         209593 non-null  int64
 24  fr_da_rech90          209593 non-null  int64
 25  cnt_loans30           209593 non-null  int64
 26  amnt_loans30          209593 non-null  int64
 27  maxamnt_loans30       209593 non-null  float64
 28  medianamnt_loans30    209593 non-null  float64
 29  cnt_loans90           209593 non-null  float64
 30  amnt_loans90          209593 non-null  int64
 31  maxamnt_loans90       209593 non-null  int64
 32  medianamnt_loans90    209593 non-null  float64
 33  payback30             209593 non-null  float64
 34  payback90             209593 non-null  float64
 35  pcircle               209593 non-null  object
 36  pdate                 209593 non-null  datetime64[ns]
dtypes: datetime64[ns](1), float64(21), int64(13), object(2)
```

3. The Distribution of our target variable is as below:



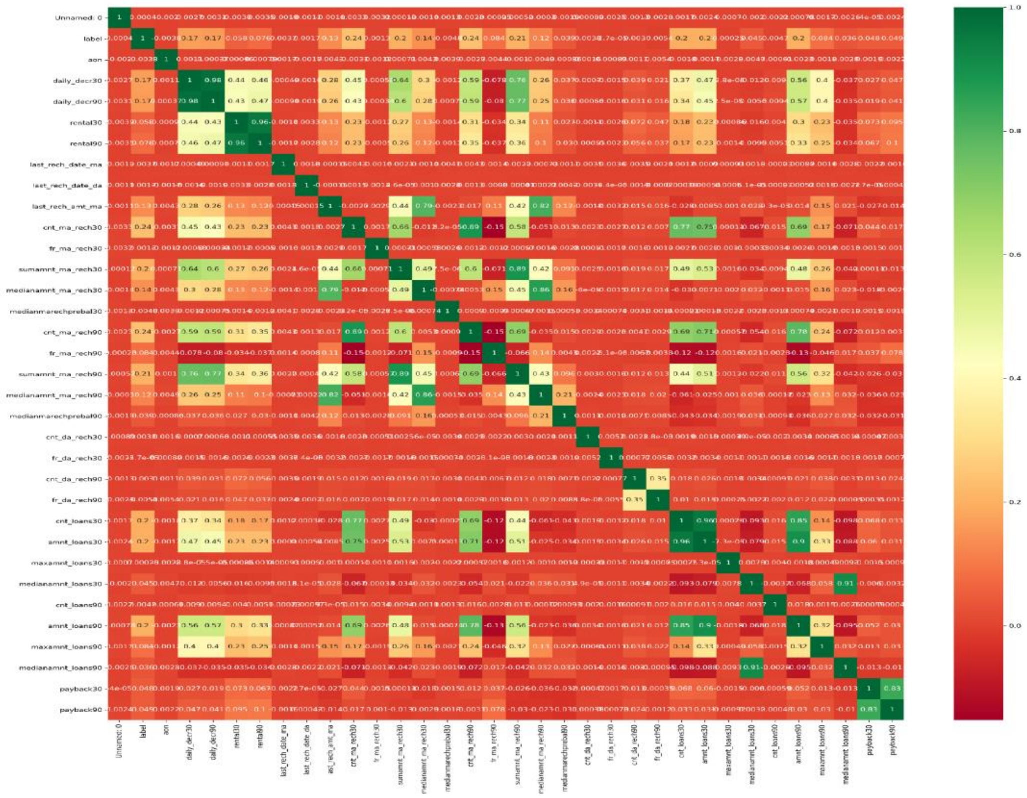
- Data Sources and their formats

Dataset is a real time dataset shared by FlipRobo Technologies in CSV format which was imported in python using pandas.

- Data Preprocessing Done:
  1) In data preprocessing various analysis being done like using describe we checked he statistical behaviour of data
  2) In statistical analysis we came across conclusion that there are outliers present in the dataset, data is skewed.
  3) Then in univariate and bi-variate analysis we saw distribution of the dataset and correlation with target variable.
  4) Then during feature selection we dropped highly correlated features
  5) Also using Z-Score we cleaned outliers

• Data Inputs- Logic- Output Relationships

So basically using SNS.heatmap saw how correlated the i/p and target variables are



• State the set of assumptions (if any) related to the problem under consideration

So I have taken two assumptions one during outlier removal using Zscore large amount of data dropped precisely 22 % which assuming wont affect the predictions much.Secondly data is imbalanced which using smote technique could be done.

• Hardware and Software Requirements and Tools Used

- Hardware Requirement:
  A computer with a processor i3 or above with 4GB Ram or above

- Software Required:
  1) 1) Python 3.6 or above
  2) Jupyter Notebook using Annaconda

- Tools/Libraries Used:

1) For computing and data input/ output Numpy,Pandas,sklearn,scipy
2) For Visualising mostly Seaborn and Matplotlib
3) For saving Model Pickle

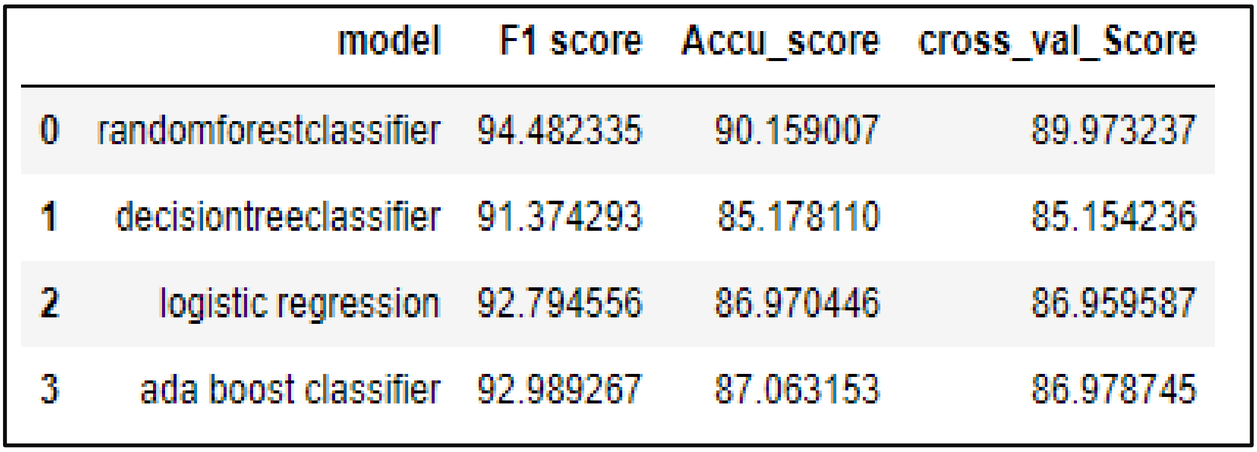
Model/s Development and Evaluation

As stated earlier it’s a classification problem and below models were used:

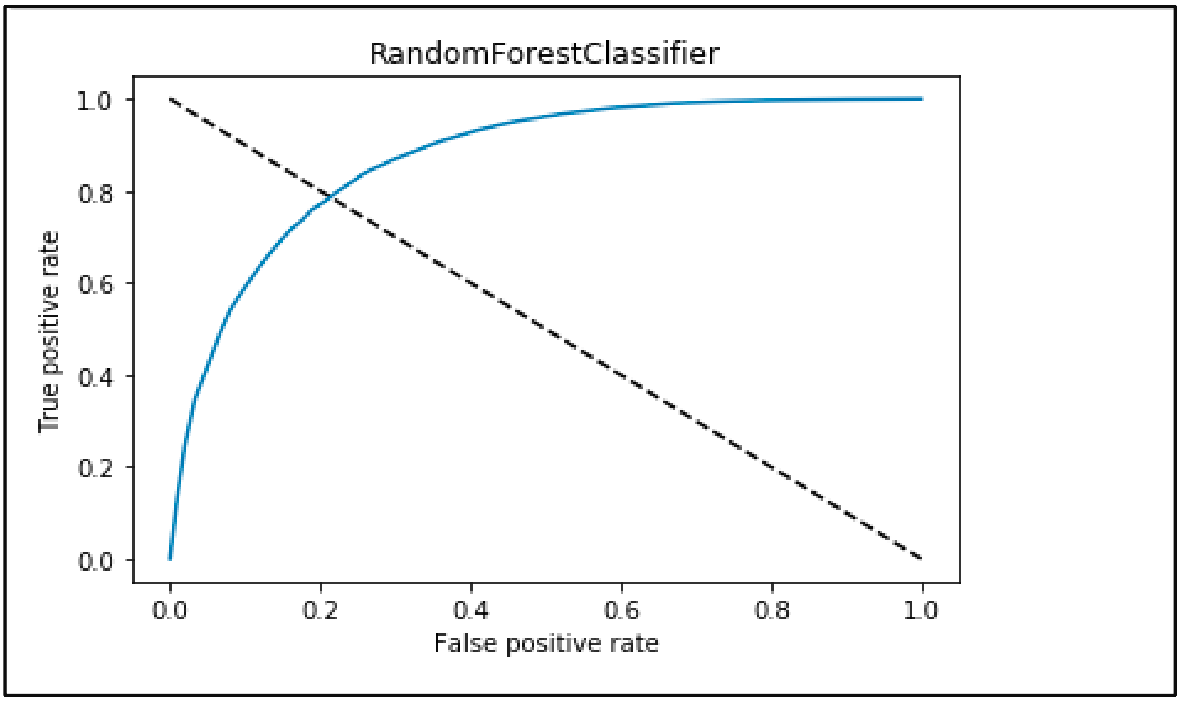
- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- AdaBoost Classifier

Evaluation being done using cross validation and various AUC and ROC Curve, accuracy score and classification metrices.

CONCLUSION



	model	F1 score	Accu_score	cross_val_Score
0	randomforestclassifier	94.482335	90.159007	89.973237
1	decisiontreeclassifier	91.374293	85.178110	85.154236
2	logistic regression	92.794556	86.970446	86.959587
3	ada boost classifier	92.989267	87.063153	86.978745



While evaluating the models found that Random Forest classifier is best fit model for the problem statement. Also would like to mention we can more fine tune the predictions by working more on feature tuning, and outlier management and also using smote for imbalanced data