FLIP ROBO

MICRO-CREDIT DEFAULTER PROJECT

Submitted by:

JASMINE KAUR

INTERNSHIP 18

ACKNOWLEDGMENT

I would like to express my gratitude to ‘Flip Robo Technologies’ that made this project possible. I would like to thank my SME Khushboo Garg for her guidance in building this project. I would also thank DT team to making me capable of handling data science and machine learning projects. Also, I would like to thank my parents for their support.

INTRODUCTION

- ## Business Problem Framing

Financial institution manages credit risk and it is considered to be one of the most important tasks managed by them. If the loan amount is not repaid then there will be no profit. Thus, it is very important to for all the financial institutions to involve in credit risk management; otherwise it will cause huge loss. This is even more applicable to the micro-credit organizations who only deals with one product i.e. loans.

Loan's default is one of the crucial concerns for the banks, so they apply different methods to predict default behaviour of their customers. A Microfinance Institution (MFI) is an organization that offers financial services to low income populations. MFS becomes very useful when targeting especially the unbanked poor families living in remote areas with not much sources of income. The Microfinance services (MFS) provided by MFI are Group Loans, Agricultural Loans, Individual Business Loans and so on.

Many microfinance institutions (MFI), experts and donors are supporting the idea of using mobile financial services (MFS) which they feel are more convenient and efficient, and cost saving, than the traditional high-touch model used since long for the purpose of delivering microfinance services. Though, the MFI industry is primarily focusing on low income families and is very useful in such areas, the implementation of MFS has been uneven with both significant challenges and successes.

Conceptual Background of the Domain Problem

Our aim is to build a model that predicts the probability of repayment for each loan transaction. This means it will determine the probability of repayment of loan within the allotted duration of time. Users are allotted 5 days to repay the loan. Those who pay within 5 days are called Non- defaulters and are labelled as 1. Those who do not pay within 5 days are called Defaulters and are labelled as 0. It is binary classification problem.

Some features that need to be considered for prediction defaulters are:

1. Those who are taking excessive loans.
2. Those who do not have recharged their accounts frequently. As these people are less likely to repay.
3. From how long a person is using the number. As old customers are more likely to be loyal and thus non-defaulter.

Review of Literature

The use of machine learning to create a credit risk management model is not new but it is growing rapidly from recent years. This is because machine learning models are capable of tackling the increased complexity of accurately predicting the behaviour of loan takers. These models can use existing data to train the model and then predicting the probability of repayment of loans. Different scoring models will be used to evaluate certain parameters that could influence the repayment behaviour.

EVALUATION MATRIX: Confusion matrix, Classification Report, ROC_AUC_curve, Cross Validation, Accuracy, F1 score.

MODELS USED: KNeighbors Classifier, Random Forest Classifier, Decision Tree Classifier, Gradient Boosting Classifier.

HYPERPARAMETER TUNING: We have applied hyperparameter tuning using Grid-Search Cross Validation to choose the best parameters.

• Motivation for the Problem Undertaken

Today, microfinance is widely accepted as a poverty-reduction tool, representing $70 billion in outstanding loans and a global outreach of 200 million clients.

We are working with one such client that is in Telecom Industry. They are a fixed wireless telecommunications network provider. They have launched various products and have developed its business and organization based on the budget operator model, offering better products at Lower Prices to all value conscious customers through a strategy of disruptive innovation that focuses on the subscriber.

They understand the importance of communication and how it affects a person’s life, thus, focusing on providing their services and products to low income families and poor customers that can help them in the need of hour.

They are collaborating with an MFI to provide micro-credit on mobile balances to be paid back in 5 days. The Consumer is believed to be defaulter if he deviates from the path of paying back the loaned amount within the time duration of 5 days. For the loan amount of 5 (in Indonesian Rupiah), payback amount should be6(in Indonesian Rupiah), while, for the loan amount of 10(in Indonesian Rupiah), the payback amount should be 12(in Indonesian Rupiah).

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  The target variable is a two class label where label 1 represents non-defaulters and label 0 represents defaulters. Thus, this is a binary classification problem. We have to predict the probability of repayment of loan, that is, whether the customer will repay the loan within 5 days or not. The criteria is that if the loan amount is 5 then the payback amount is 6 and if the loan amount is 10 then the payback amount is 12.

  INSIGHTS FROM THE STATISTICAL SUMMARY:

  1. The columns that have negative values: aon, daily_decr30, daily_decr90, rental30, rental90, last_rech_date_ma, last_rech_date_da, medianmarechprebal30 & medianmarechprebal90.
  2. Some of the features are categorical while some are continuous.
  3. The gap between the 3rd quantile and maximum value is huge for some columns. There are very high chances of outliers in the data.
  4. Some columns are not normally distributed and thus the data is skewed.

DEALING WITH THE NEGATIVE VALUES OF THE DATA: Most of the values that are negative are actually not possible to have negative values. So, we substituted the negative values with the zero.

REMOVING OUTLIERS: We have removed the outliers.

CORRELATION: Some features are highly correlated with each other, so it is better to drop some of these features in the dataset.

- <u>Data Sources and their formats</u>

1. The dataset is provided in csv format.
2. There is no null value in the data. But there are unrealistic values.
3. The label column is imbalanced. Label 1 i.e. non-defaulter has 87.5% records whereas label 0 i.e. defaulters has 12.5% records.
4. Description of the features in the dataset:

i. **Label** : Flag indicating whether the user paid back the credit amount within 5 days of issuing the loan {1: success, 0: failure}

ii. **msisdn**:mobile number of users

iii. **aon** : age on cellular network in days

iv. **daily_decr30**: Daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah)

v. **daily_decr90**: Daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah)

vi. **rental30**: Average main account balance over last 30 days

vii. **rental90**: Average main account balance over last 90 days

viii. **last_rech_date_ma**: Number of days till last recharge of main account

ix. **last_rech_date_da**: Number of days till last recharge of data account

x. **last_rech_amt_ma**: Amount of last recharge of main account (in Indonesian Rupiah)

xi. **cnt_ma_rech30**: Number of times main account got recharged in last 30 days

xii. **fr_ma_rech30**: Frequency of main account recharged in last 30 days

xiii. **sumamnt_ma_rech30**: Total amount of recharge in main account over last 30 days (in Indonesian Rupiah)

xiv. **medianamnt_ma_rech30:** Median of number of recharges done in main account over last 30 days at user level (in Indonesian Rupiah)

xv. **medianmarechprebal30**: Median of main account balance just before recharge in last 30 days at user level (in Indonesian Rupiah)

xvi. **cnt_ma_rech90**: Number of times main account got recharged in last 90 days

xvii. **fr_ma_rech90**: Frequency of main account recharged in last 90 days

xviii. **sumamnt_ma_rech90**: Total amount of recharge in main account over last 90 days (in Indian Rupee)

xix. **medianamnt_ma_rech90**: Median of number of recharges done in main account over last 90 days at user level (in Indian Rupee)

xx. **medianmarechprebal90**: Median of main account balance just before recharge in last 90 days at user level (in Indian Rupee)

xxi. **cnt_da_rech30**: Number of times data account got recharged in last 30 days

xxii. **fr_da_rech30**: Frequency of data account recharged in last 30 days

xxiii. **cnt_da_rech90**: Number of times data account got recharged in last 90 days

xxiv. **fr_da_rech90**: Frequency of data account recharged in last 90 days

xxv. **cnt_loans30**: Number of loans taken by user in last 30 days

xxvi. **amnt_loans30**: Total amount of loans taken by user in last 30 days

xxvii. **maxamnt_loans30:** maximum amount of loan taken by the user in last 30 days

xxviii.**medianamnt_loans30**: Median of amounts of loan taken by the user in last 30 days

xxix. **cnt_loans90**: Number of loans taken by user in last 90 days

xxx. **amnt_loans90**: Total amount of loans taken by user in last 90 days

xxxi. **maxamnt_loans90**: maximum amount of loan taken by the user in last 90 days

xxxii. **medianamnt_loans90**: Median of amounts of loan taken by the user in last 90 days
xxxiii.**payback30**: Average payback time in days over last 30 days
xxxiv.**payback90**: Average payback time in days over last 90 days
xxxv. **pcircle**: telecom circle
xxxvi.**pdate**: date

- <u>Data Pre-processing Done</u>

It starts with collecting the data and ends with building the entire model that gives the prediction with very high accuracy. The steps involved in data pre-processing is very important to transform the raw data into a meaningful and efficient format. Without pre-processing, we might end up considering wrong information, thus it will in turn results in model that does not predict correctly.

The process of data pre-processing has 4 steps:

i. CLEANING: Column wise empty cells analysis was performed and it was found that there is no missing value. But there are some columns with unrealistic values, so we have dealt with that. There were some columns with only one unique value, so we dropped those columns.
ii. FORMATTING: We divided the date column into three different columns namely date, month and year.There were outliers in the dataset, so we have removed them. Skewness was present in the data, so we have dealt with it.
iii. SCALING: We have scaled the data using standard scaler.
iv. LABEL ENCODING AND NORMALIZATION: We converted the non numerical columns into numerical columns using label encoder.

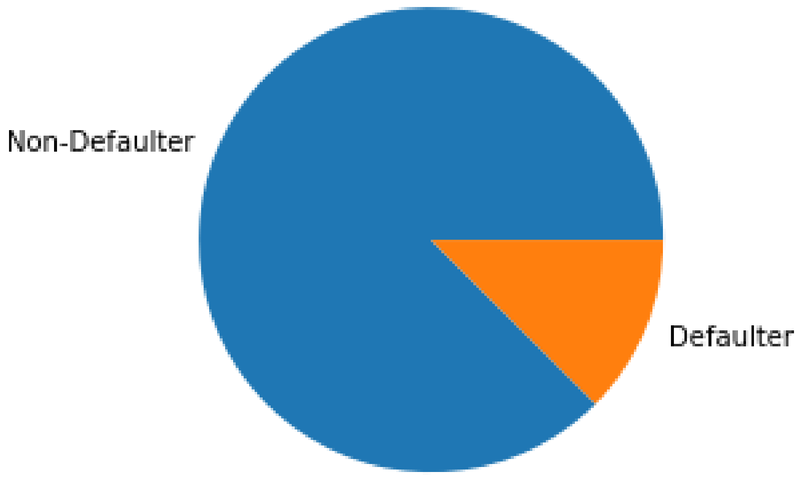
• Data Inputs- Logic- Output Relationships

The target variable in the project is a binary class with 0 and 1 as their values. The dependent variable is of int64 type and there are 203386 rows. The distribution of the class is:

1: Non-defaulters - 183431

0: Defaulters – 26162

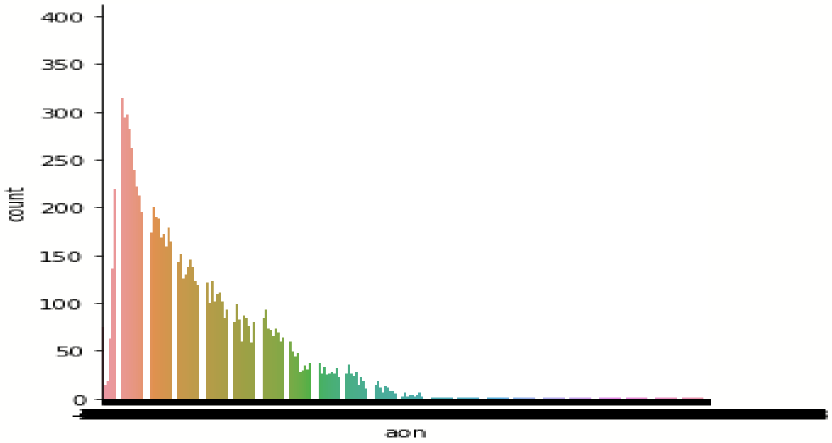
The representation of the target variable distribution:



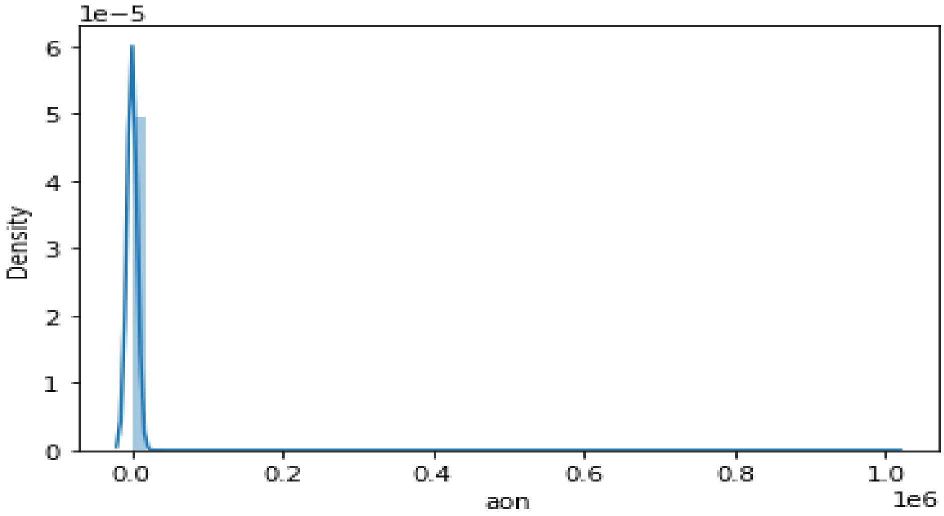
There are 19 independent variables with 203386 rows each.

Representation of some of the feature variables are:

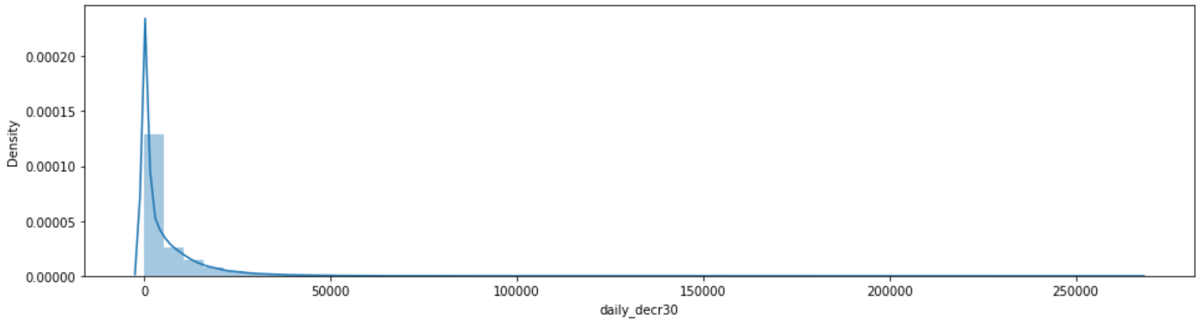
1. Age on cellular network



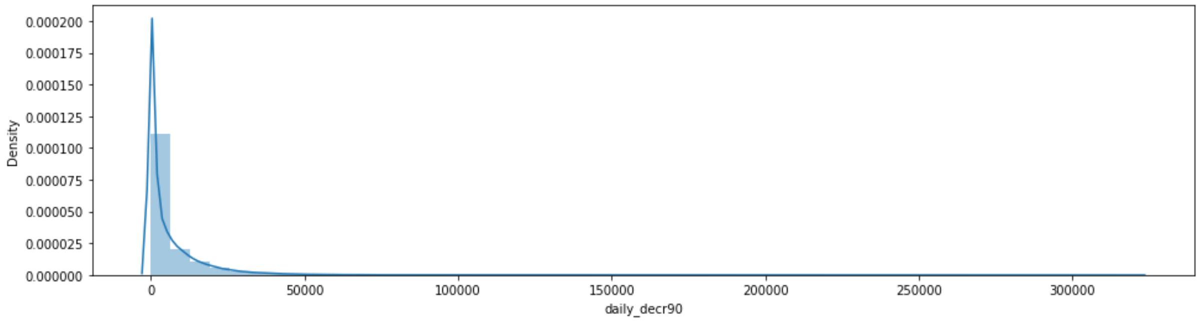
2. Distribution of age on cellular network



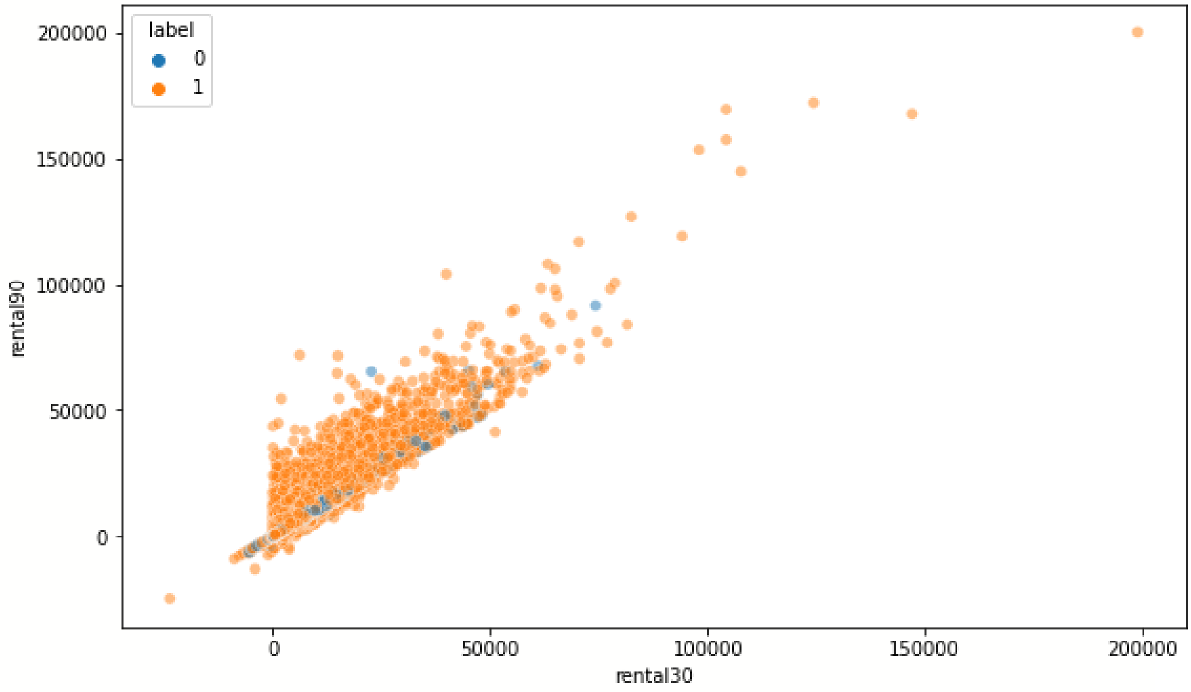
3. Distribution of daily_decr30



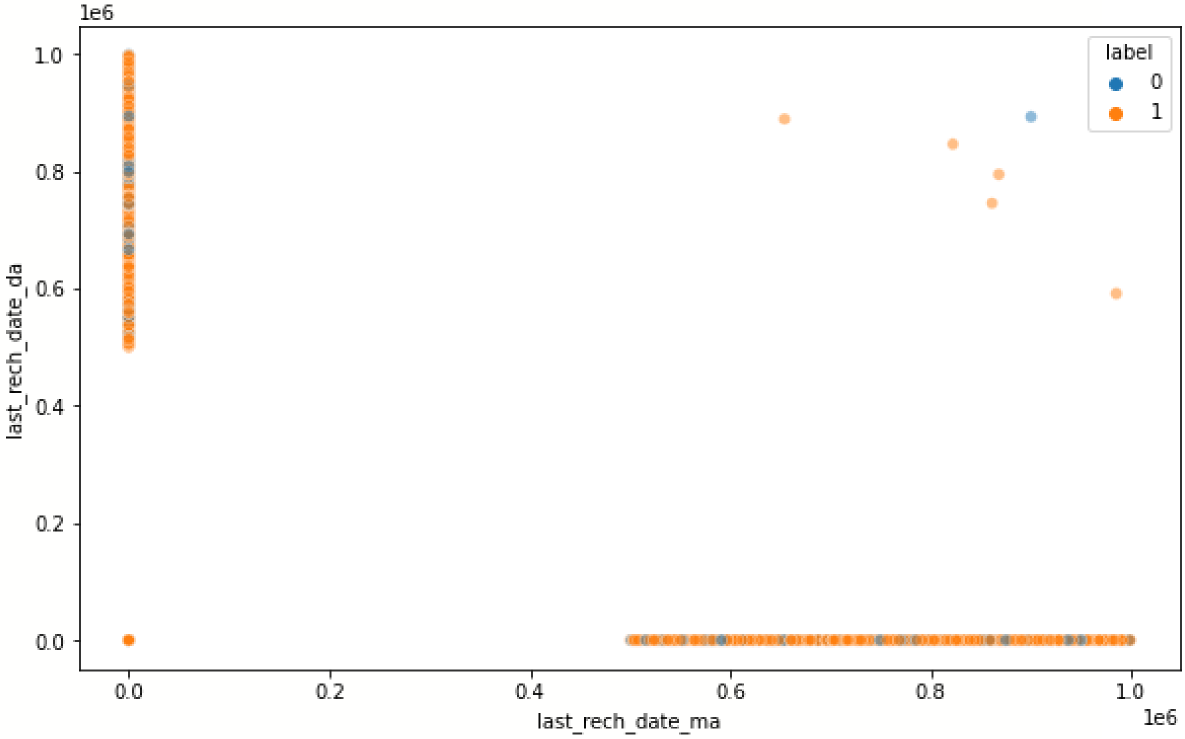
4. Distribution of daily_decr90



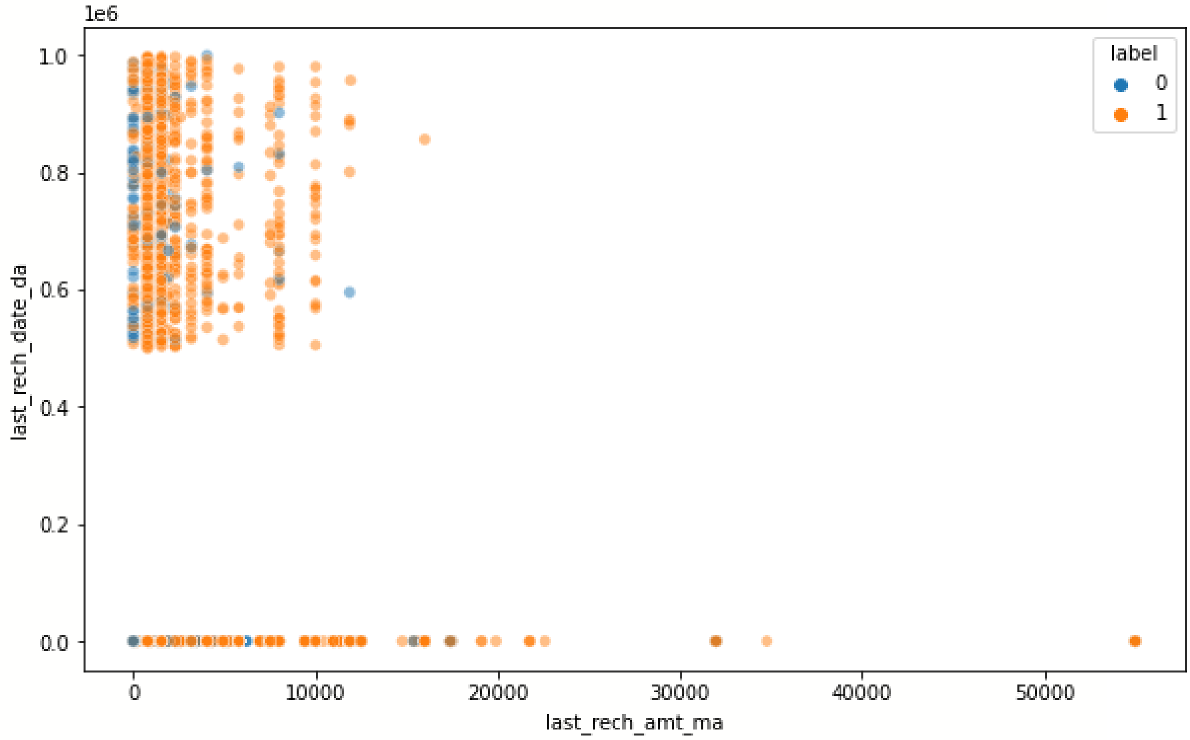
5. Scatterplot of rental30,data.rental90



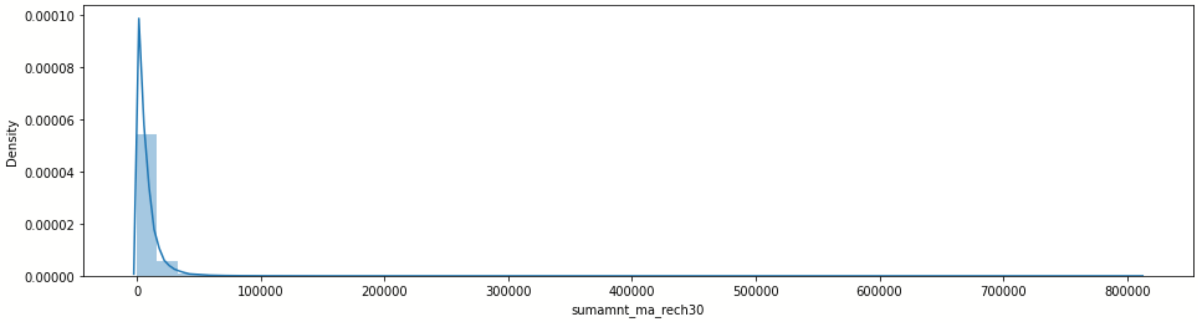
6. Scatterplot of last_rech_amt_ma and last_rech_date_da



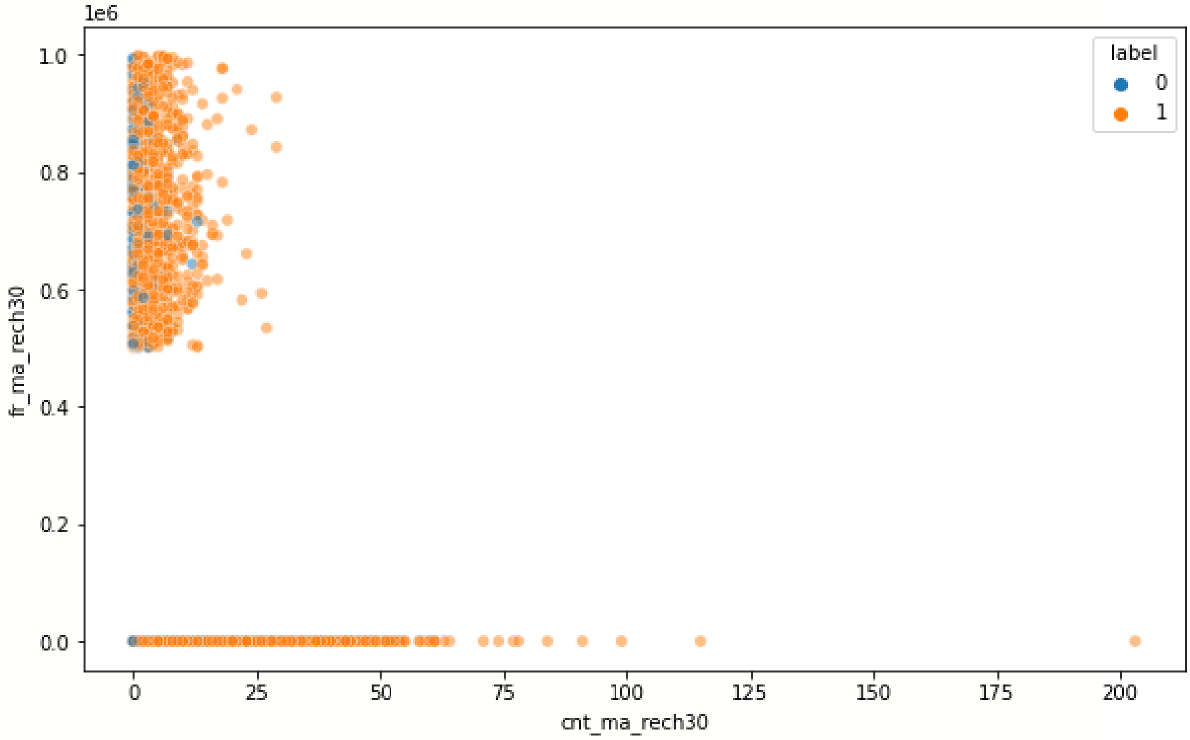
7. Scatterplot of last_rech_amt_ma and last_rech_date_da



8. Distribution of sumamnt_ma_rech30



9. Scatterplot of cnt_ma_rech90 and fr_ma_rech90



- <u>State the set of assumptions (if any) related to the problem under consideration</u>

1. We have assumed that it is not possible to negative values in various features. Those features must have minimum values as zero. So, we have applied it in the project.
2. We have assumed that outliers distort our analysis, so we have applied tools to remove them from the data.
3. We have assumed that if any column has only one unique value then it will not help us in analysing, so we have removed it.
4. We have assumed that the number of the customer cannot tell anything about the behaviour of the customer related to the repayment of the loan, so we have removed it.
5. We have assumed that imbalanced class distribution of the target variable will give biased results, so we have balanced the class using SMOTE.

Hardware and Software Requirements and Tools Used

Hardware: 4GB RAM, 64 BIT 0/S, I 3 Processor

Software tools: Google colab, python, notepad

Libraries and packages used:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.preprocessing import StandardScaler
from statsmodels.stats.outliers_influence import variance_inflation_factor
from sklearn.metrics import f1_score,roc_auc_score
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import train_test_split, cross_validate

from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import KFold
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score,classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import roc_curve, roc_auc_score

from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier,GradientBoostingClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.neighbors import KNeighborsClassifier
```

Model/s Development and Evaluation

- ## Identification of possible problem-solving approaches (methods)
  1. Boxplot for summarizing variations and checking outliers
  2. Histograms are used to check distribution
  3. Correlation matrix was used to analyse correlation between the features.
  4. Standard scaler was used to standardize the data.
  5. We have split the data using train test split.
- ## Testing of Identified Approaches (Algorithms)
  1. KNearestNeighbors
  2. Random Forest Classifier
  3. Decision Tree Classifier
  4. Gradient Boosting Classifier
- ## Run and Evaluate selected models

KNEARESTNEIBHORS

```
Accuracy: 0.8660866152034541

Precision: 0.94480433981145
Recall: 0.7788728725251823
Classification report:
               precision    recall  f1-score   support

           0       0.81      0.95      0.88     45652
           1       0.94      0.78      0.85     46064

    accuracy                           0.87     91716
   macro avg       0.88      0.87      0.87     91716
weighted avg       0.88      0.87      0.87     91716
```

CONFUSION MATRIX

```
array([[43585,  2129],
       [10033, 35969]])
```

RANDOM FOREST CLASSIFIER

```
Accuracy: 0.907377120676872
Precision: 0.9171459661122338
Recall: 0.8965786731503994
Classification report:
               precision    recall  f1-score   support

           0       0.90      0.92      0.91     45652
           1       0.92      0.90      0.91     46064

    accuracy                           0.91     91716
   macro avg       0.91      0.91      0.91     91716
weighted avg       0.91      0.91      0.91     91716
```

```
CONFUSION MATRIX
array([[42064,  3650],
       [ 4688, 41314]])
```

DECISION TREE CLASSIFIER

```
Accuracy: 0.8422194600723975
Precision: 0.8561700976302676
Recall: 0.8243313650573115
Classification report:
               precision    recall  f1-score   support

           0       0.83      0.86      0.84     45652
           1       0.86      0.82      0.84     46064

    accuracy                           0.84     91716
   macro avg       0.84      0.84      0.84     91716
weighted avg       0.84      0.84      0.84     91716
```

```
CONFUSION MATRIX
array([[39277,  6437],
       [ 8174, 37828]])
```

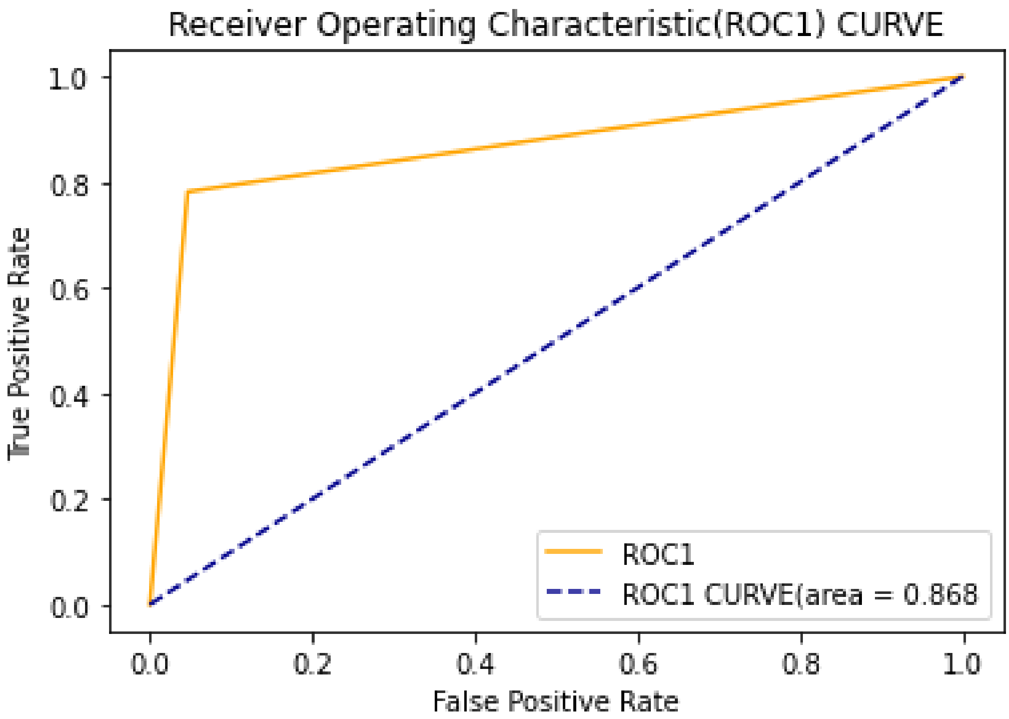
GRADIENT BOOSTING CLASSIFIER

```
Accuracy: 0.7918138601770683
Precision: 0.8044842846820809
Recall: 0.7734673497742272
Classification report:
               precision    recall  f1-score   support

           0       0.78      0.81      0.79     45652
           1       0.80      0.77      0.79     46064

    accuracy                           0.79     91716
   macro avg       0.79      0.79      0.79     91716
weighted avg       0.79      0.79      0.79     91716
```

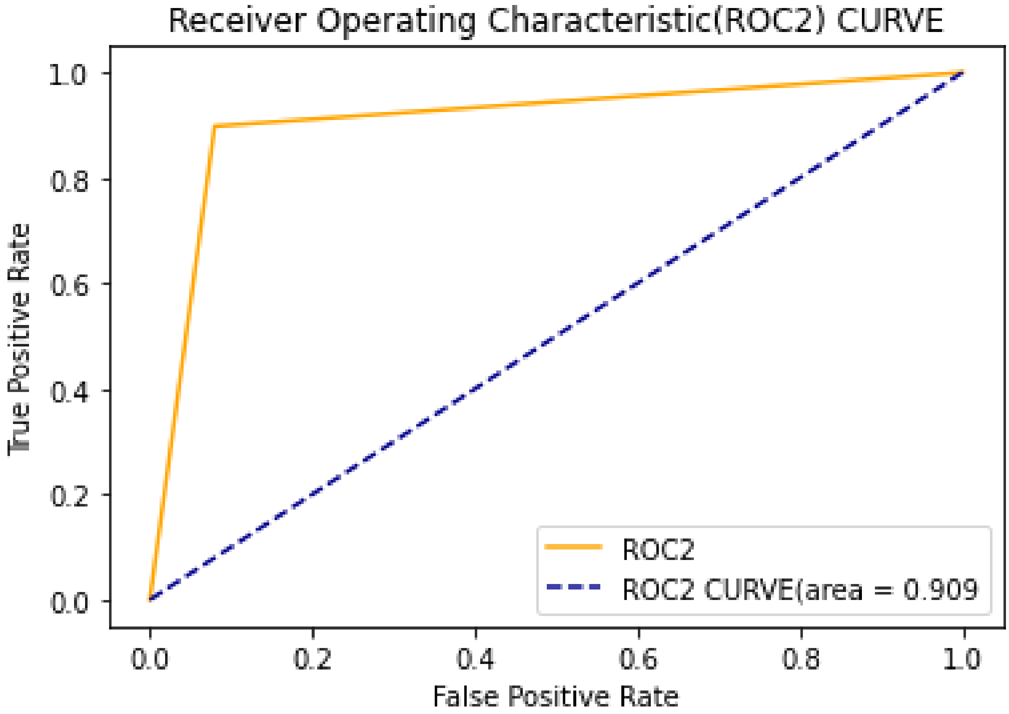
```
CONFUSION MATRIX
array([[37108,  8606],
       [10433, 35569]])
```

- Visualizations

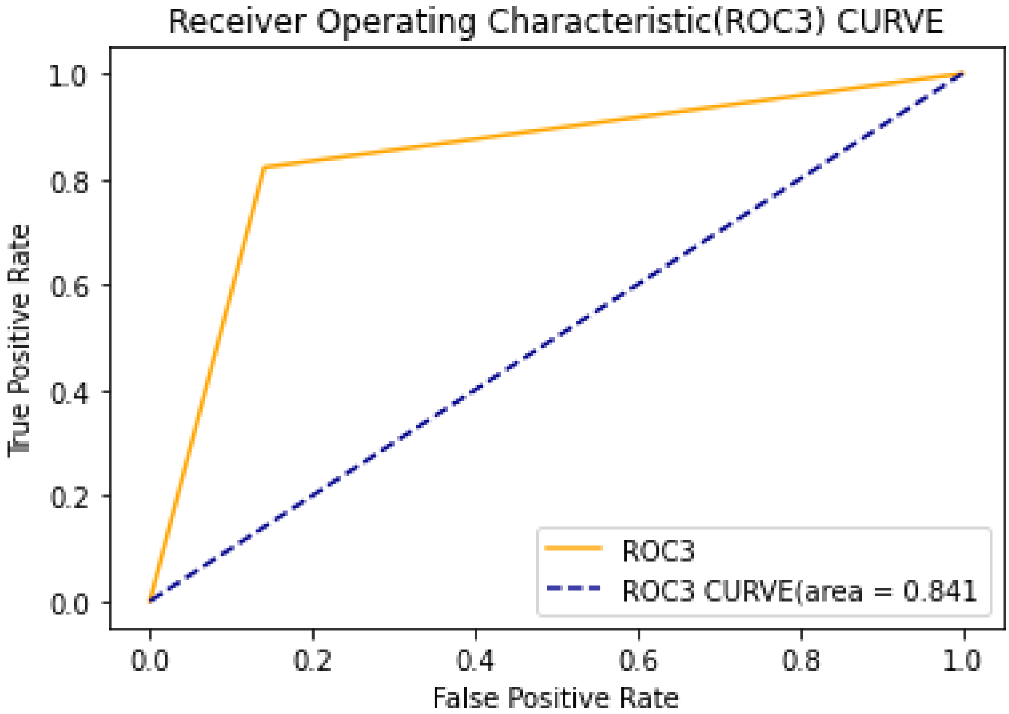
1. KNearestNeighbors:



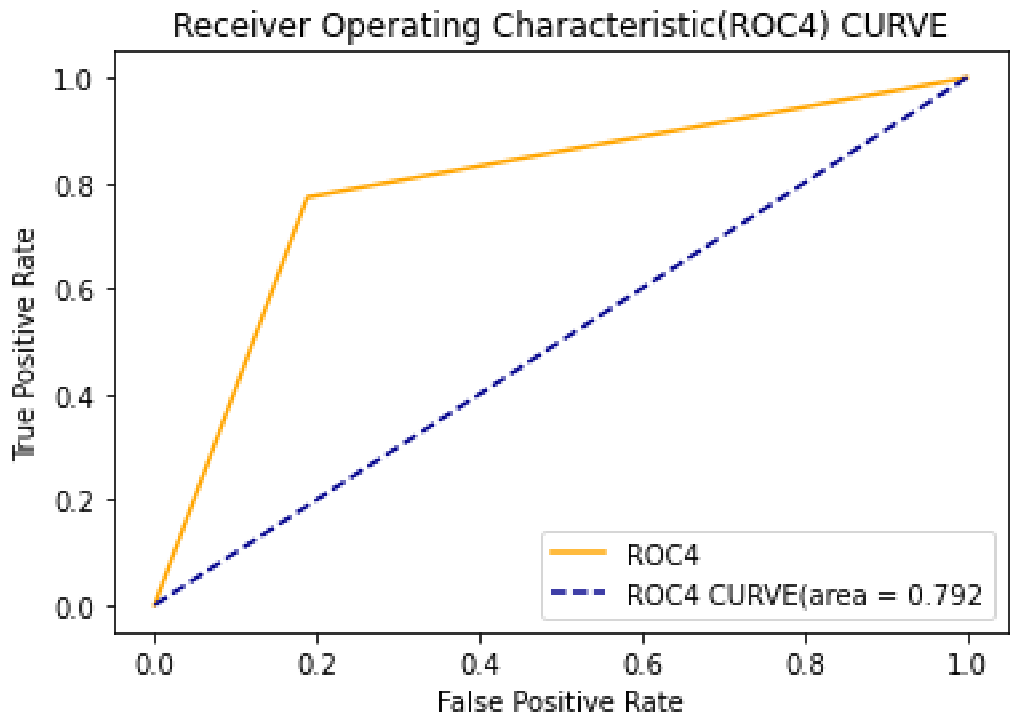
2. Random Forest classifier



3. Decision Tree Classifier



4. Gradient Boosting Classifier



Interpretation of the Results

The difference between the accuracy score and their cross validation score is minimum for gradient boosting classifier, so we will apply hyper-parameter tuning using

The PCA shows that 10 features provide more than 95% of the information about the data. So, we have selection top 10 features for the model.

Applied train test split and got the following results:

Modeling Algorithm	Accuracy
KNearestNeighbors	0.866087
RandomForestClassifier	0.908609
DecisionTree	0.841969
GradientBoostingClassifier	0.791814

Modeling Algorithm	**Accuracy**
KNearestNeighbors	0.866087
RandomForestClassifier	0.908609
DecisionTree	0.841969
GradientBoostingClassifier	0.791814

Applied cross validation:

Modelling Algorithm	**CVSCORE**
KNearestNeighbors	0.8713791
RandomForestClassifier	0.9119832
DecisionTree	0.8465723
GradientBoostingClassifier	0.790586

The difference between the two is minimum for gradient boosting classifier. After applying hyper-parameter tuning, we got the accuracy of the model equal to 80.22%.

CONCLUSION

5. Key Findings and Conclusions of the Study

The project demonstrated the use of machine learning on a challenging dataset to analyse the ability of repaying the loan. To achieve the best result, data pre-processing is performed very carefully. We have applied PCA and selected top 10 features as they were giving more than 95% of the information. Random forest was also performing very well but we took gradient boosting classifier for applying hyper-parameter tuning.

6. Learning Outcomes of the Study in respect of Data Science

- Data cleaning is quite tedious task and also very time consuming. We encountered many issues but research, study and guidance.
- Data visualization helped in analyzing the data
- All the libraries and packages helped us achieving our goal.

7. Limitations of this work and Scope for Future Work

- We could have used randomized search CV to apply hyper parameter tuning.
- Size of the dataset is huge, so it was difficult to handle but we got the opportunity to deal such datasets.
- There were many unrealistic values and we tried our best to deal with it.
- In future, we would apply more algorithms to have better results.

||THANK YOU||