

NAME OF THE PROJECT

Submitted by:-

DUMPALA ADITYA

**ACKNOWLEDGMENT**

I have referred below resources that helps me to completion of this project

🡺stackoverflow

🡺geeksforweeks

🡺google

**INTRODUCTION**

* Business Problem Framing

This project is about providing loans (financial services) to low income populations by Micro-Financial Institution (MFI). MFI provide loan to Group Loans, Agricultural Loans, Individual Business Loans and so on. In Order to achieve this objective, MFI needs to decide criteria for customer selection.

* Conceptual Background of the Domain Problem

This problem is related to banking domain and the banking related knowledge is required for solving this problem, market risk and parameters to decide defaulter, interest charges, benefits, etc.

* Review of Literature

Loan giving capacity will get decided based on below parameters-Daily amount spend & average main account balance in last 30 days, Frequency of recharge for data account & main account in 30/90 days , loan taken in last 90 days & payback time for last 30 days.

* Motivation for the Problem Undertaken

In order to understand to whom loan to be given from lower income earning people and data from telecom industry clearly stats parameters to be taken into consideration to declare borrower as defaulter or not & amount limit also can be decide based on this.

In every country poor population exists to some scale and financial services to be provided to them at affordable level of loan amount to uplift their financial situation, which may reduce the vulnerability factor

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

In this problem there are 209593 records are there, and there are no null values and features are skewed extreamly right side.

* Data Sources and their formats

data provided by client in excel or csv format.

* Data Preprocessing Done
* The given data set has contains 209593 records and 37 columns
* There is no null values are present
* But there are two columns named as ‘unnamed’ and ‘msisdn’ these two are not useful for our prediction because unnamed has only like index and msisdn is customer mobile number and it is unique for every person so we can remove it
* And there is ‘pdate’ is one of the feature and it is in the form of object type and iam converted into datetype and extract day and month not year because year only one year so year is not useful for us after extracting day and month I remove pdate from the data set
* ‘pcircle’ also has only one unique value so we can drop this as well
* there is too much difference in 75% and max, it indicates there are outliers are present
* and some features are having its minimum value is negative but that -ve value for those particular feature does not make any sense like aon (age on cellular network) the value in this feature must be +ve but there are some negative values are there so we have to remove that data and same for remaining features also
* and also there are some features are having 90% of the data is zeros for this particular features we cannot replace zeros by something because nearly 90% and more than 90% of the data is belongs to zero category so it is better to drop those features and I drop those features
* as we discuss previously there is huge difference between 75% and max it indicates there is an outliers are present so iam using box plot to checking outliers and I observe there is huge outliers are present and I am using z score for removing the outliers
* coming to the skewness almost all the features are skewed right side since there are high variance between them iam using powertransformation technique to remove skewness
* and our out has two categories one and zero and also it in highly imbalanced we can use SMOTE for solving this problem

* Data Inputs- Logic- Output Relationships

There is feature cnt\_ma\_rech90 is highly correlated with the target feature and cnt\_ma\_rech30 also means if the data in these features is change there is a high chance to change in the target variable and also there are remaining features also there they also correlates good with the feature

* State the set of assumptions (if any) related to the problem under consideration
* iam considering -ve values in the data set is not makes any sense and I drop them also
* and also there are some fetures has zero values more than 90% so these features also not makes any sense so I drop them as well
* we can observe that there is features having same category but days are changed like cnt\_ma\_rech30 and cnt\_ma\_rech90 so iam testing with scatter plot and I find there is nice relationship between them after I check all the feature like this names and finally Iam using heatmap for checking correlation between them and I find some features having high correlation between them, I remove the features which are having highly correlated with each other

* Hardware and Software Requirements and Tools Used

**Hardware Requirement-**Laptop with below configurations-

Windows Edition-Windows 10

Processor-RYZEN5

Memory-32 GB

System Type-64 bit OS

* **Software Requirement-** Anaconda 3.7 & above , Jupiter Notebook 6.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

**Analytical approach**:- based on EDA and our target variable I am decided to use classification model for this particular problem

**Statistical approach**:- to understand the data easily for the given model the data should be scaled iam using min max scaler for scaling the data this min max scaler is converting the data into max is 1 and minimum is 0 so this type of data will be easily understand to the model

* Testing of Identified Approaches (Algorithms)

iam testing with 5 below mentioned algorithms they are

**1)Logisticregression**

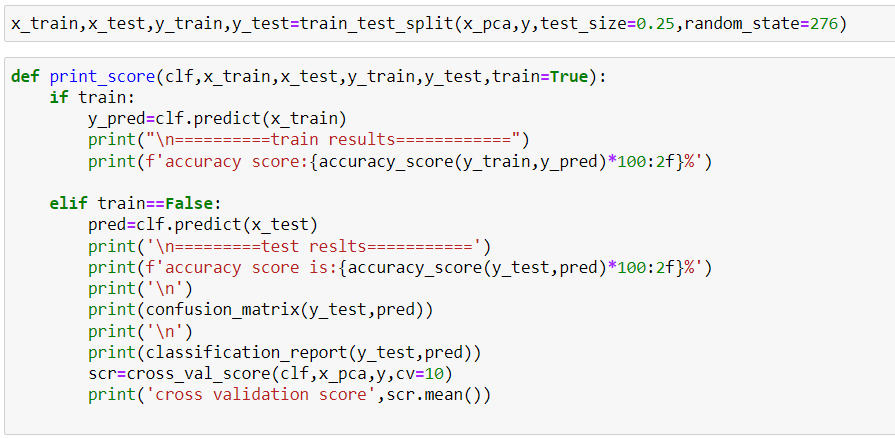
**2)kneighborsclassifier**

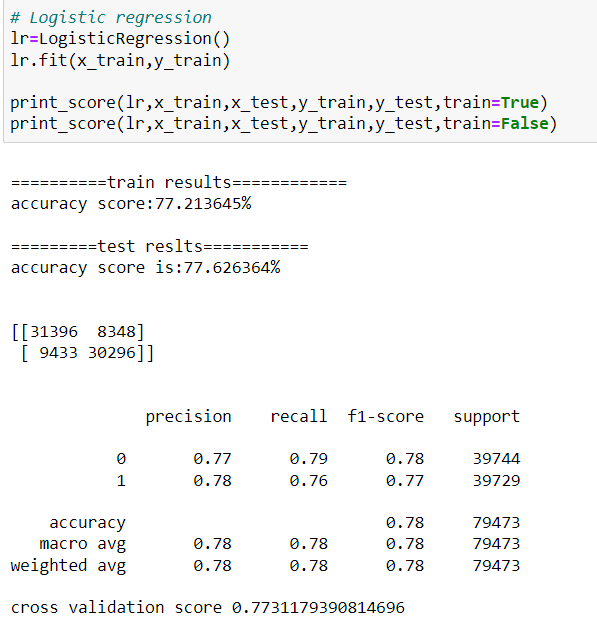
**3)decision tree classifier**

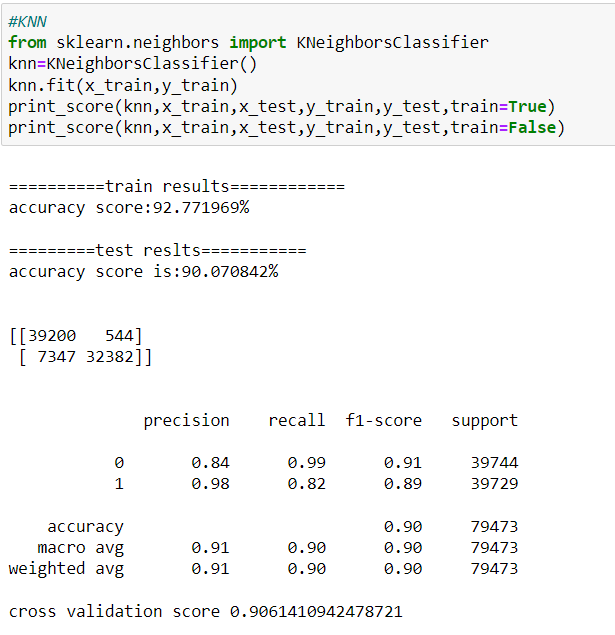
**4)random forest classifier**

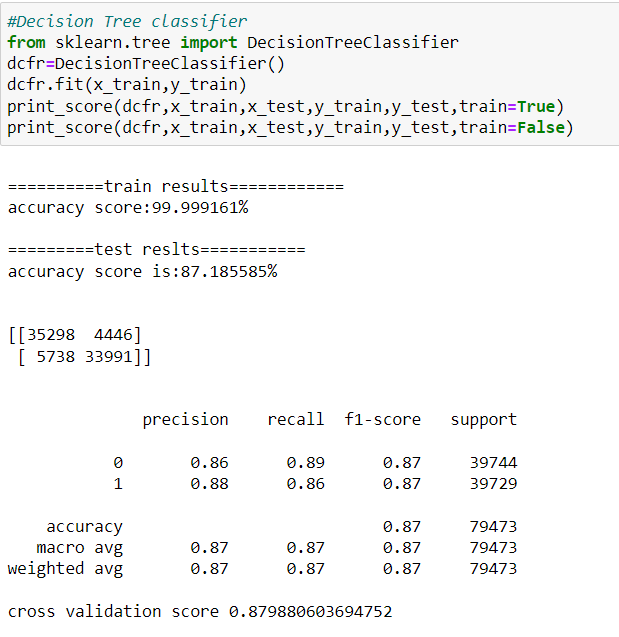
**5)gradient boosting classifier**

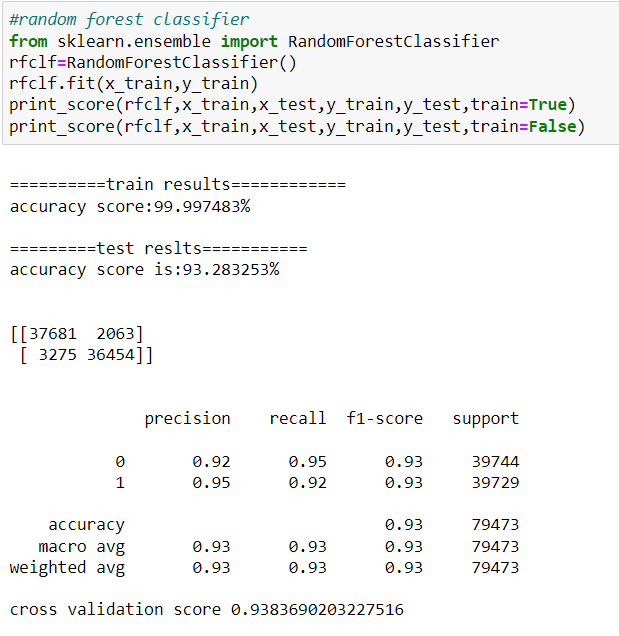
* Run and Evaluate selected models

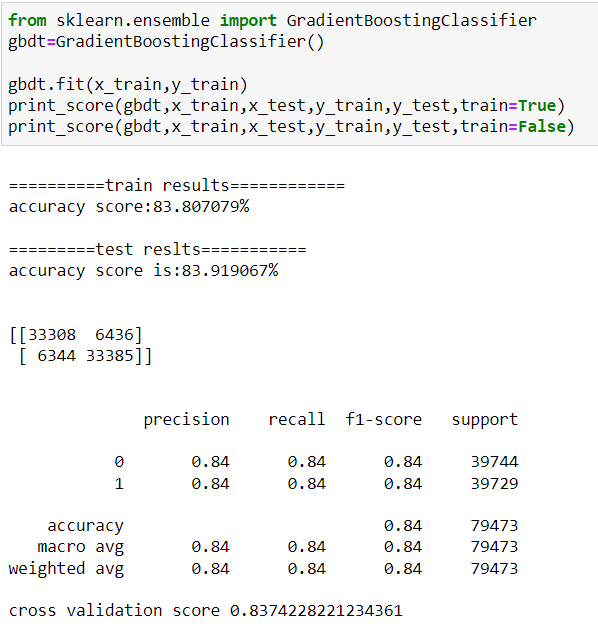




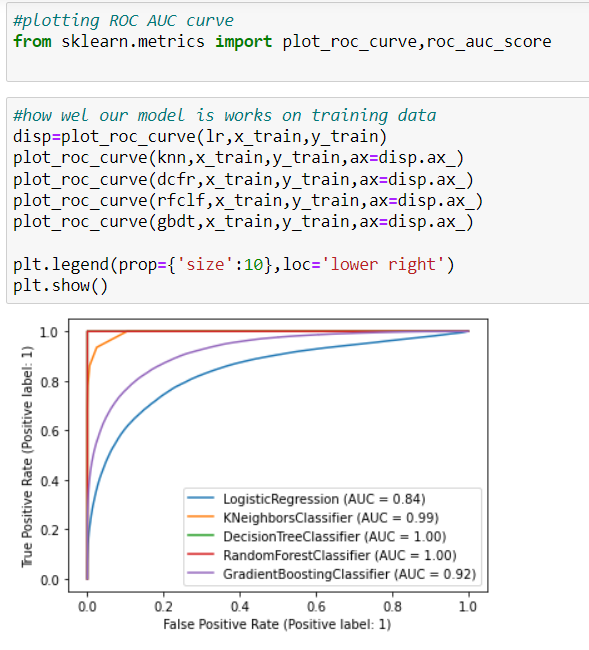




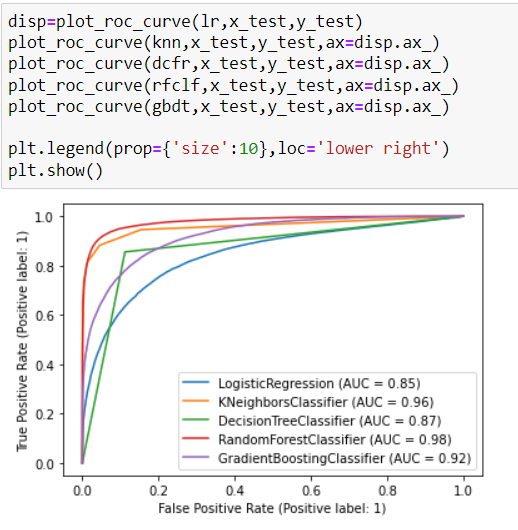




Roc auc for training data:-



for testing data



the above plots shows random forest,gradient boosting clasifier and kneighborsclassifier are given best results but iam choosing kneighbors classifier as final model since our problem is related to bussiness scenerio so we consider false positive rate should be minimum or less so in the knn FP rate is less than all other so this is the best model for choosing as final

* Key Metrics for success in solving problem under consideration

**Zscore:-**  it is one of the metric I used to remove outliers

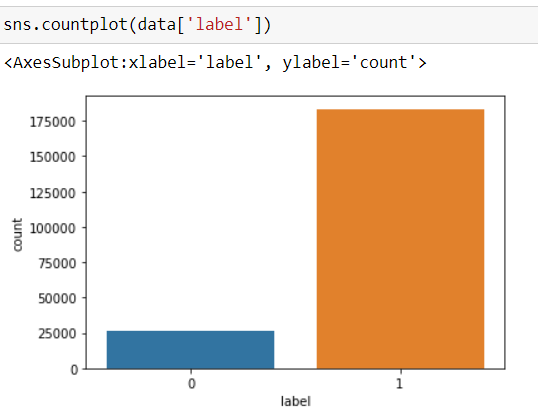
**Correlation:-**  it is nothing but finding the correlation between the features not with the target variable

**SMOTE:-**  it is used for balancing the target variable since our target variable is imbalanced

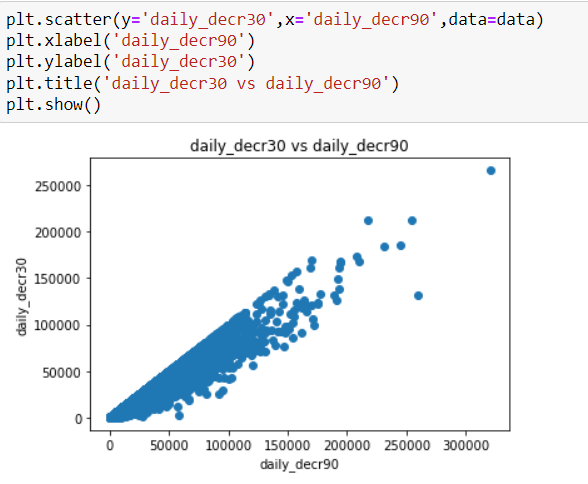
**Min max scaler:**- for scaling the data to avoid high variance between the data points

**Cross validation:**- it is used for to overcome the overfitting/ underfitting problem

* Visualizations

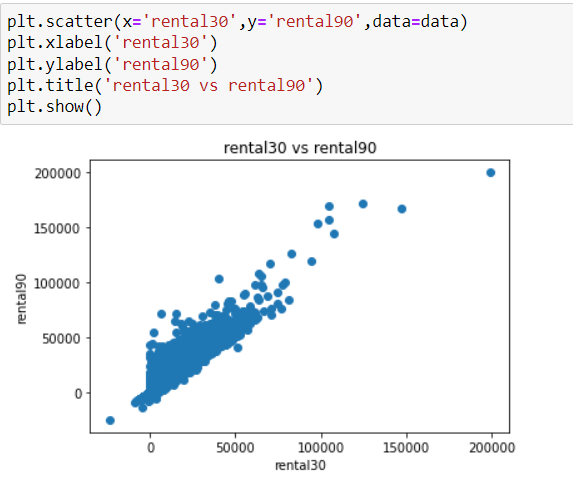


Observation:- the target variable is highly imbalanced



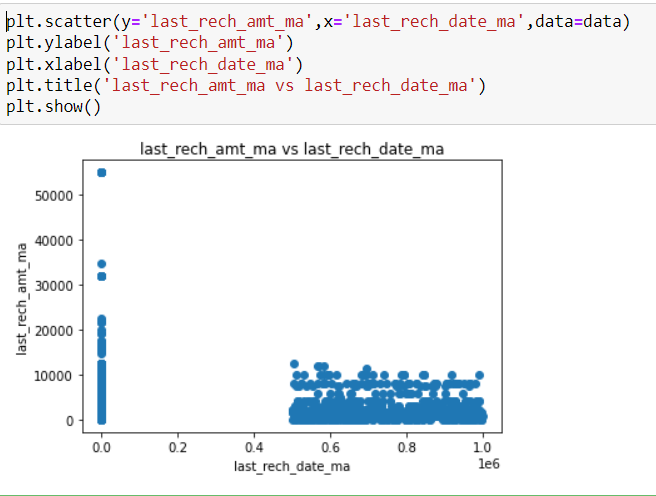
#observation

#we can see that Daily amount spent from main account, averaged over last 30 days and Daily amount spent from main account, averaged over last 90 days are highly correlated with each other.



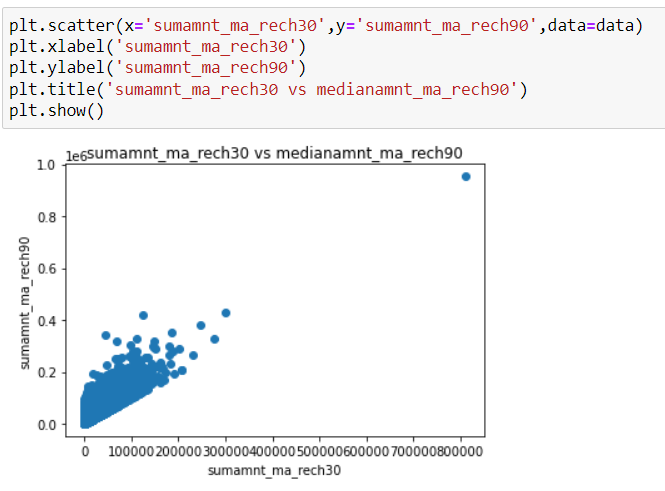
#observation

# rental30 and rental90 are correlating each other



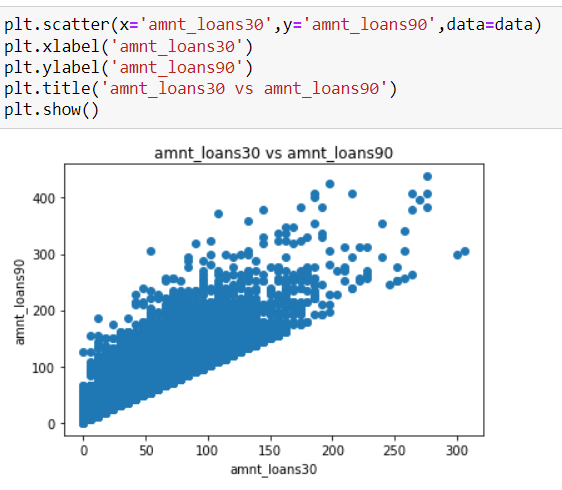
Observation

There is no relationship exists



Observation:-

There is a relationship exists between sumamt\_ma\_rech90 and sumamt\_ma\_rech30



There is relationship exits between amnt\_loans30 and amntloans90

* Interpretation of the Results

From all the visualizations I observe that there is multicollinearity exists between the features but when I plot them by using heatmap less no of features are correlated with each other

**CONCLUSION**

* Key Findings and Conclusions of the Study
* I am observe that the data is highly skewed and
* multicollinearity between the features
* outliers in the continuous features
* Learning Outcomes of the Study in respect of Data Science

**The following are the learning outcomes in the process of solving this problem**

* Dealing with large data like data set having more no of records (more than 1 lakh)
* Handling highly skewed data (right side)
* Removing outliers by using zscore with the help of box plot
* Balancing the imbalaned target variable using SMOTE technique
* Scaling the data by using min max scaler
* Using PCA for reducing high dimentionality problem and multicollinearity problem
* Using different algorithms for train and test the data set
* Using auc roc curves for selecting final model/ algorithm
* Hyperparameter tunning by using different parameters for improving model accuracy
* Saving the model by using pickle
* Limitations of this work and Scope for Future Work

Limitation of this work is we must have minimum knowledge related to banking domain if we don’t have it becomes difficult to assume something in the process of EDA and etc