

**Micro-Credit Defaulter**

**Submitted by:**

Akash Kanjwani

**ACKNOWLEDGMENT**

During the process of completing this project, I have referred following materials for which I owe them great gratitude.

1.Data collection- Using Web scraping tool i.e. Selenium

2. Car24.com for collecting the data *easemytrip.com/*

3. Data trained video tutorials.

4. Scikit-learn <https://scikit-learn.org/stable/>

5. Machine Learning for Dummies by John Mueller and Luca Massaron - Easy to understand for a beginner book.

6. Geeksforgeeks. <https://www.geeksforgeeks.org/>

Besides that all the observation, creations of the models and graphs done by self help.

**Problem Statement:**

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 are very useful in such areas, the implementation of MFS has been uneven with both significant challenges and successes.

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 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 sample data is provided to us from our client database. It is hereby given to you for this exercise. In order to improve the selection of customers for the credit, the client wants some predictions that could help them in further investment and improvement in selection of customers.

**Analytical Problem Framing**

Mathematical/ Analytical Modelling of the Problem

1) The size of table is 209593\* 36 i.e. no. of rows are 209593 and no. of columns are 36(including target).

2) Out of 36 columns 3 columns are object type in nature and rest 33 are numeric type.

3) Null values are not present in the data set as we can see in ths seaborn heatmap so there is no need to adopt imputation technique.

4) In case of object data type, we will apply the encoding technique to convert the values in the numeric format

In our data set all the columns have outliers as we can see, In all the columns the diffrence between mean and standard deviation is very high and the gap between standard deviavtion and the max values is also very high, It shows that the outliers are present in most of the colunms.

In the above dataset most of the columns has skewness :

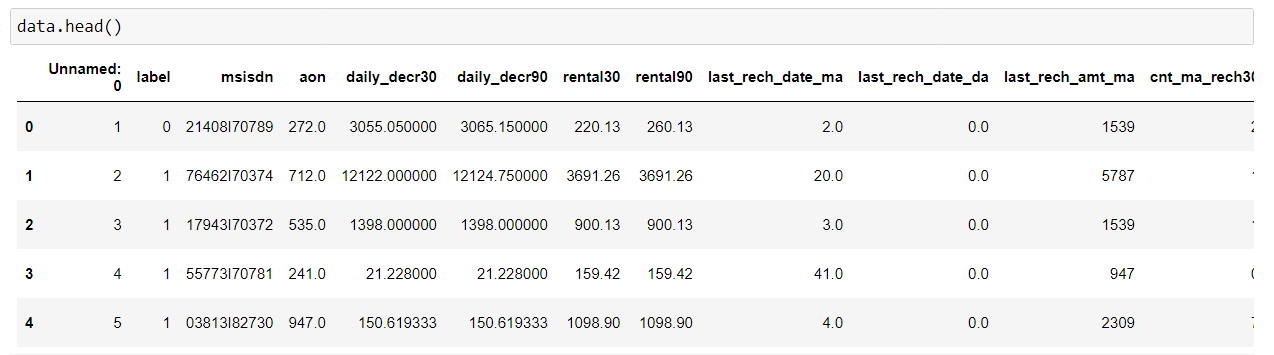
High skewness represents data imbalancy.

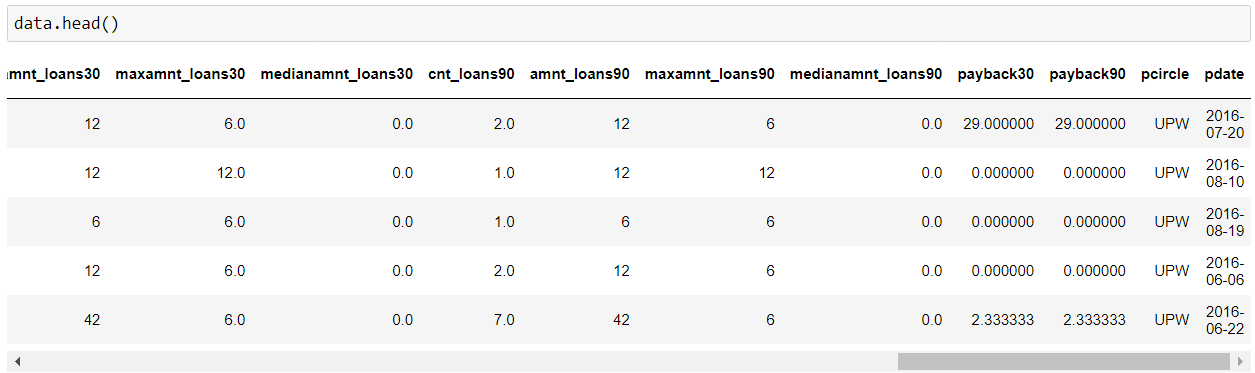
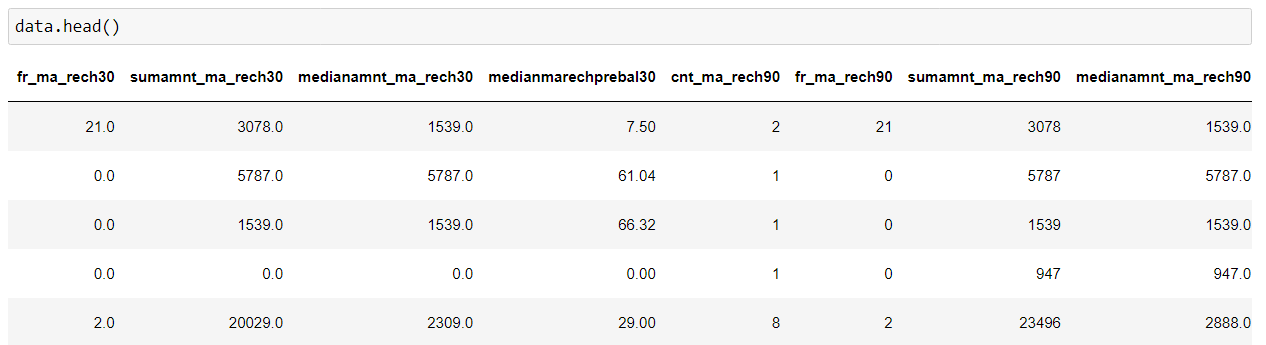
Less skewness represents that our data are equally distributed.

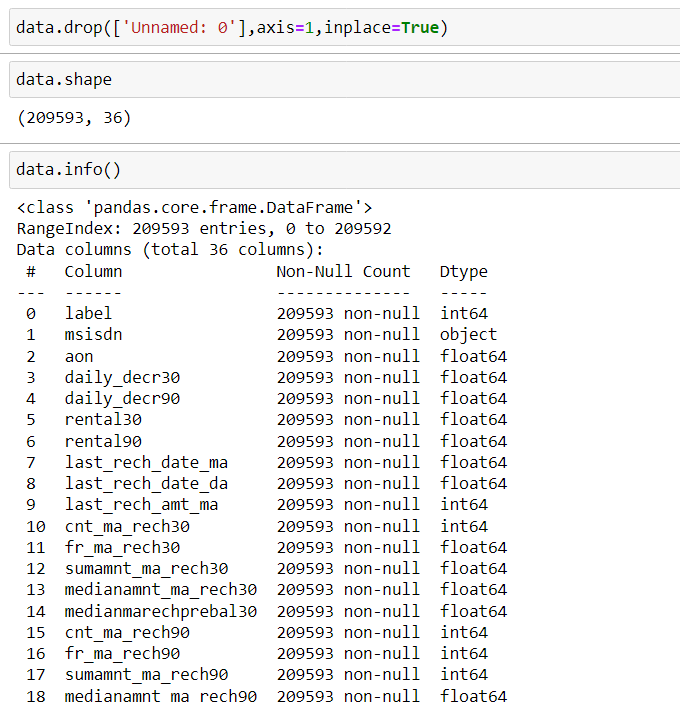
If we see the skewness in our data set then we will find that our most of the colunm's data are skewed.This is because of mean is greater than median in most of the colunms.

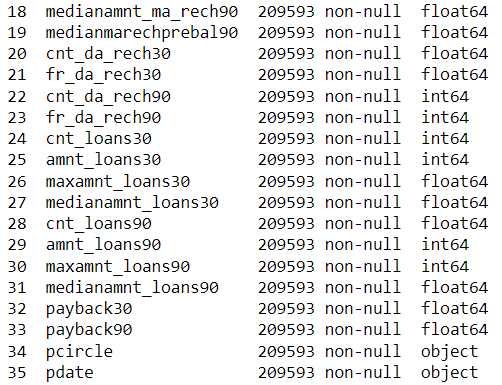
**Data Sources and their formats**

Data has been provided by the Flip Robo technology



****

****

****

**Data Pre-processing Done**

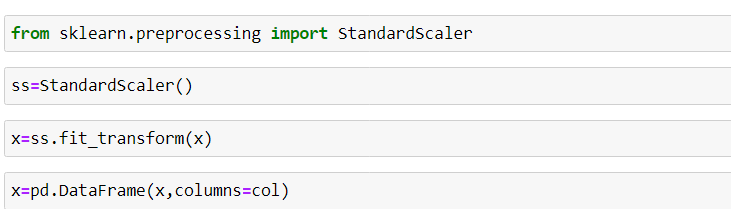
**Outliers, Skewness, Data scaling, Data Cleaning**

Because we had categorical data and absence of numerical data that is why there was no column which can contain outliers and skewness only we had one column i.e. price column, which is numerical in nature but here we can not pre-process the target column.

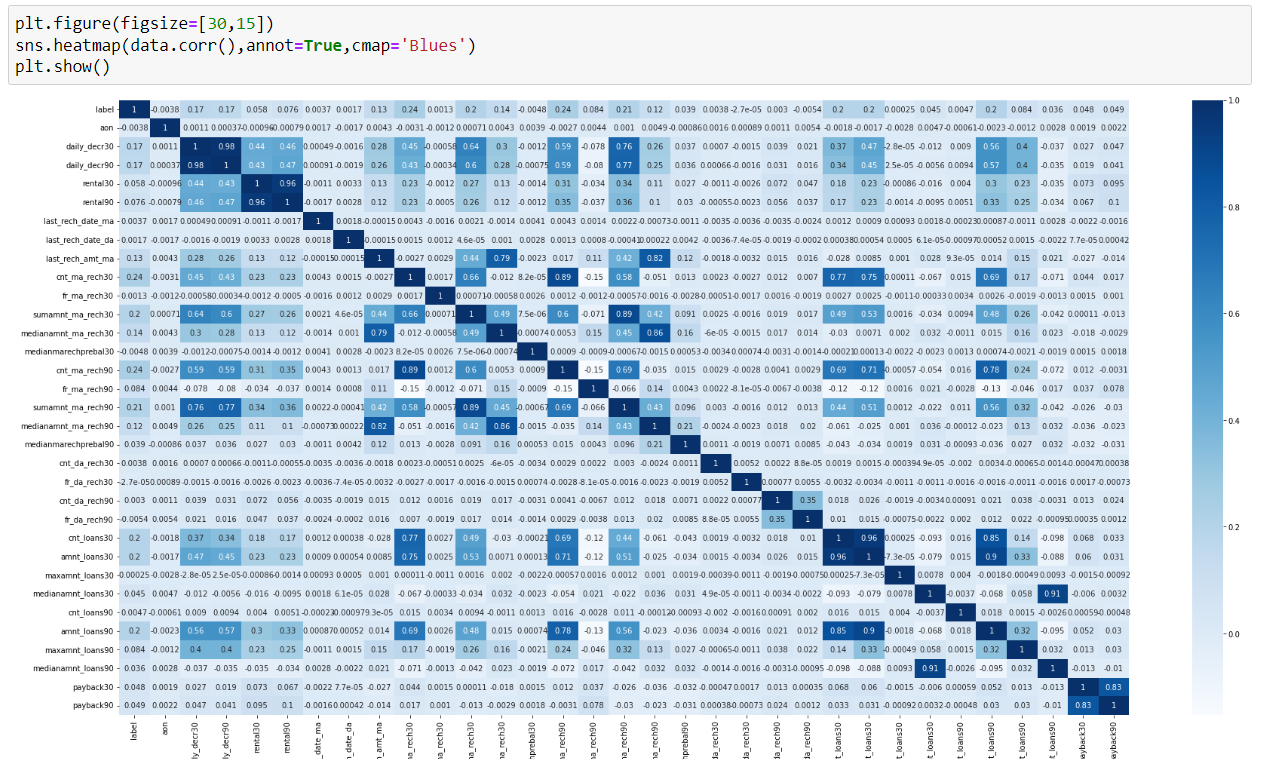
\* For Data scaling we used Standard scaler.

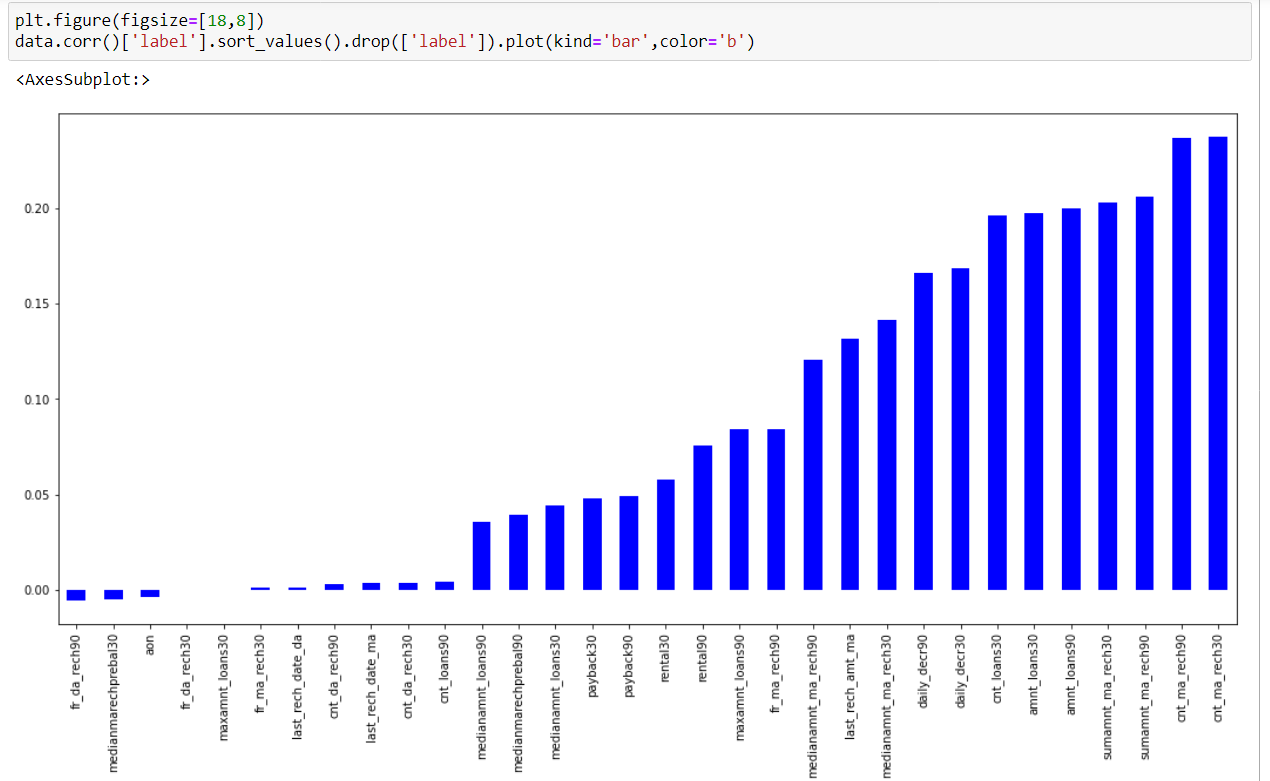
\* Data cleaning using pandas function

**Scaling the data**

****

**Data Inputs- Logic- Output Relationships**



****

**Hardware and Software Requirements and Tools Used**

**Anaconda Navigator**

**Jupyter Notebook**

**Language-Python**

Selenium

**Many lib.-------**

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import numpy as np

import warnings

import selenium

warnings.filterwarnings('ignore')

from sklearn.preprocessing import power\_transform

from scipy.stats import zscore

from sklearn.preprocessing import MinMaxScaler

import statsmodels.api as si

from scipy import stats from statsmodels.stats.outliers\_influence

import sklearn

from sklearn.linear\_model import LinearRegression,Lasso,Ridge,ElasticNet

from sklearn.model\_selection import train\_test\_split,GridSearchCV,cross\_val\_score

from sklearn.tree import DecisionTreeRegressor

from sklearn.svm import SVR

from sklearn.neighbors import KNeighborsRegressor

from sklearn.ensemble import RandomForestRegressor,AdaBoostRegressor,GradientBoostingRegressor

import xgboost as xg from sklearn.metrics

import mean\_squared\_error,mean\_absolute\_error,r2\_score

**Pandas**- For making data frame

**Matplotlib and seaborn-** For data visualization

**Numpy-** For numerical python

**MInMAxScaler-** For data scaling

**Power transform**-For removing skewness

**From metrice** - mean\_squared\_error,mean\_absolute\_error,r2\_score -For checking the model accuracy.

**Regression-** For regression modeling

**Ensamble-** For boosting and bagging

**Grid search cv-** For hyperparameter tuning

**Cross\_Val\_Score**- For cross validation

**Model/s Development and Evaluation**

Approaches Firstly it is import to know about which type of modelling we are going to construct, For this problem we used regression models because our target variable is numerical type and we had to predict the flight price. When we go for regression we have to use some metrics like mean\_squared\_error, mean\_absolute\_error, r2\_score In order to do this work we have to find out the best random state by which we can achieve good accuracy. Then we split our data set into the train part and test part using train test split. When we done with the modelling we have to use cross validation for real accuracy(without underfitting and overfitting). Hyperpameter is must for increasing the model accuracy in order to build good model for this we use Grid search cv. We followed all the above approaches to build our Machine learning model.

**Model learning phase**

So basically it will become important to know which type of machine learning model we are going to construct.It depends on the target. Here we are going to use classification model because our target column is catagorical in nature.

For classification model, Some matrix we are going to find like:

* Confusion matrix
* Accuracu Score
* Classification report

**Models:**

Decision Tree classifier

KNeighbors classifier

Support vectod machine classifier

Naive bayes classifier

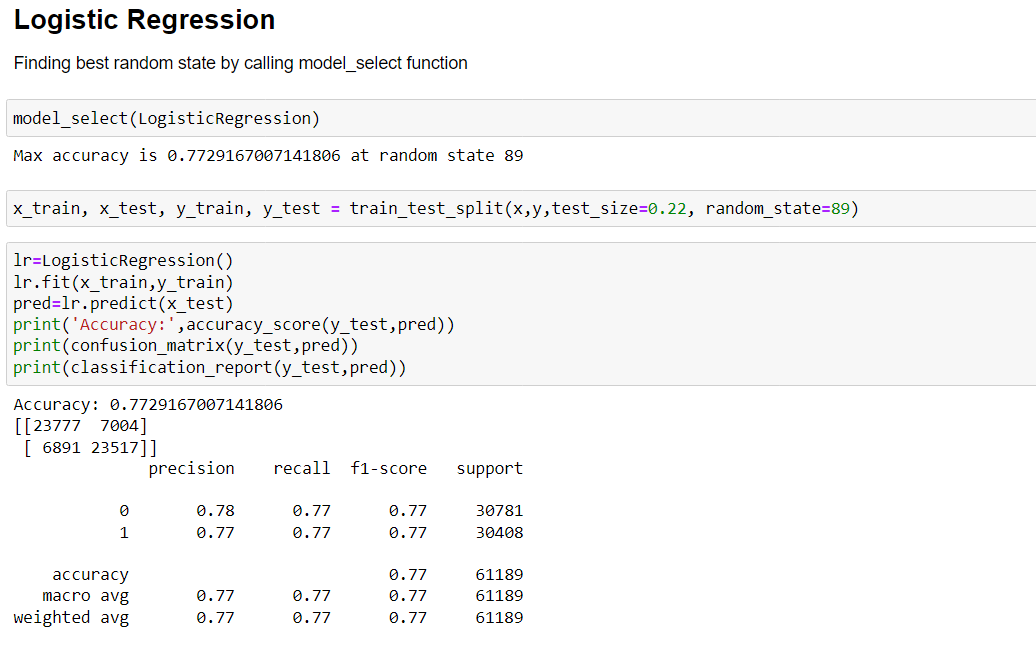
For bagging and boosting :

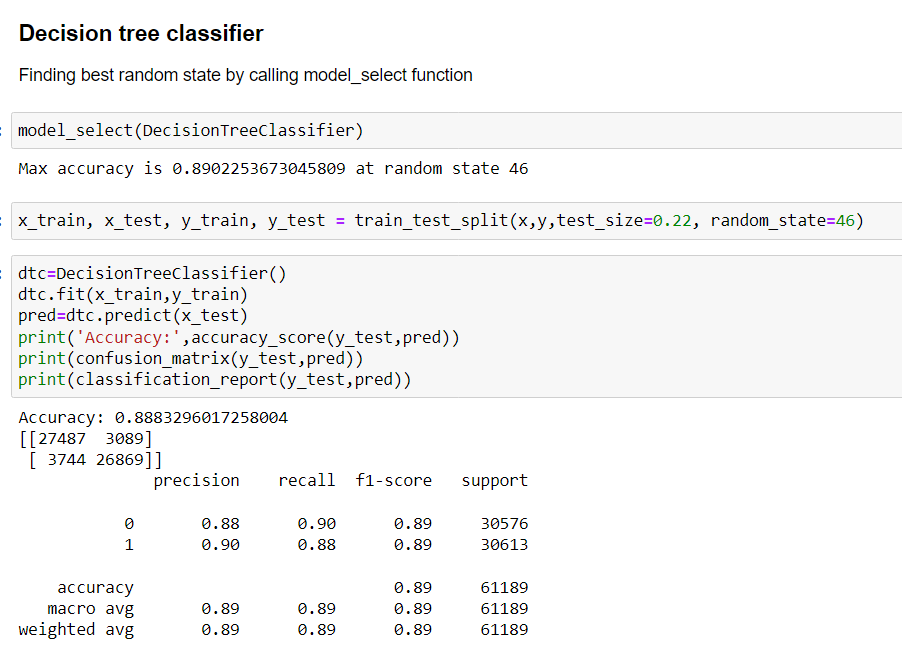
Random Forest classifier

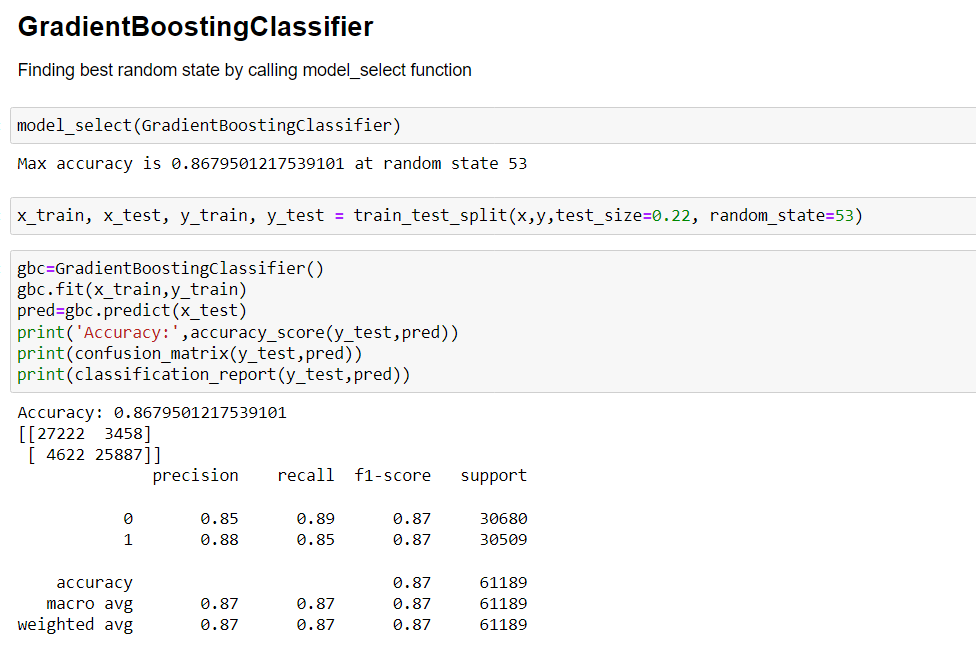
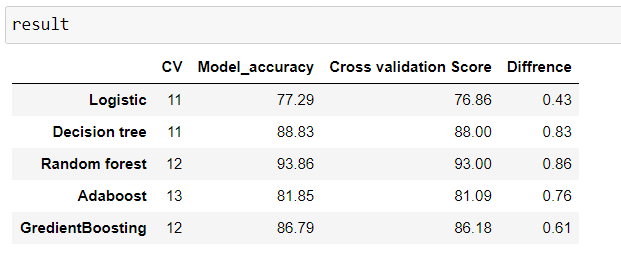
AdaBoost classifier

Gradient Boosting classifier

XGboost classifier

****

****

****

### 

### 

### Best Model

### We are chosing Random forest classifier as a best model, Because this model has least diffrence between model acuracy score and cross validation score and also its accuracy is highest as compare to others.That is why we are choosing RFC

### Model accuracy is : 93.86

### Cross\_validation score : 93

### Diffrence : 0.86

### 

### 

