MICRO CREDIT DEFAULTER MODEL

Submitted by: Amrutha Dhondale

Acknowledgement

The internship opportunity I have with Flip Robo Technologies is a great chance for learning and professional development.

I perceive this opportunity as a big milestone in my career development. I will strive to use gained skills acknowledge in the best possible way. I would like to extend my appreciation and thanks for the mentors from Data Trained and professionals from FlipRoboTechnologies who had extended their help and support.

References: <https://www.youtube.com/user/krishnaik06>

<https://www.youtube.com/c/CodeWithHarry>

INTRODUCTION

Business Problem: A client in Telecom Industry is collaborating with an MFI (Microfinance Institution) 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. 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.

Background of domain:

• Microfinance is a banking service provided to unemployed or low-income individuals or groups who otherwise would have no other access to financial services.

• Microfinance allows people to take on reasonable small business loans safely, and in a manner that is consistent with ethical lending practices.

• The majority of micro financing operations occur in developing nations, such as Uganda, Indonesia, Serbia, and Honduras.

• Like conventional lenders, micro financiers charge interest on loans and institute specific repayment plans.

• The World Bank estimates that more than 500 million people have benefited from microfinance-related operations. Indonesia is renowned for its large-scale microfinance sector, with a range of commercial banks. More than 56.5 million Micro Small Medium Enterprises1 (MSME), contributed greater than 50% of Gross Domestic Product (GDP) in 2014.

However, many of them do not have adequate access to the bank financing they need to grow their businesses, particularly in rural areas. Some rural communities in Indonesia have no choice but to seek out loans from unregulated moneylenders.

Micro lenders, particularly those operating under Indonesian banks, as well as social enterprise start-ups, are also targeting these communities through their

high mobile penetration rates and are developing the right digital platforms to reach out to them.

Only around 22% of Indonesians are connected to formal financial institutions. Micro-finance is accessible for people in remote areas and on small islands, not just people in the cities. In 2012, there were 143 million unique mobile subscribers, more than double the number of bank account holders (62 million).

Telecommunication operators have more than 300,000 locations at which phone vouchers are sold. Most banks would like to have access to these distribution networks, which would enable them to access the poorest people requiring micro-finance.

MOTIVATION FOR PROBLEM UNDER TAKEN:

Based on data provided from our client database, customer’s repayment of loan is assessed based on different factors. By building the model, we can assess which customers are highly likely to repay the loan, thereby it will be useful for those needy people who will repay the loan and also prevent the loss to the customer by avoiding loans to the defaulters.

ANALYTICAL PROBLEM FRAMING MATHEMATICAL MODELLING OF PROBLEM:

Mathematical modelling is simply the method of implementing statistical analysis to a dataset where a Statistical Model is a mathematical representation of observed data.

While analysing the data, there are an array of statistical models we can choose to utilize. For the given project, we need to predict whether the customer is a defaulter or not. This is a classification problem. There are wide varieties of classification models like decision trees, random forests, nearest neighbour, Logistic Regression.

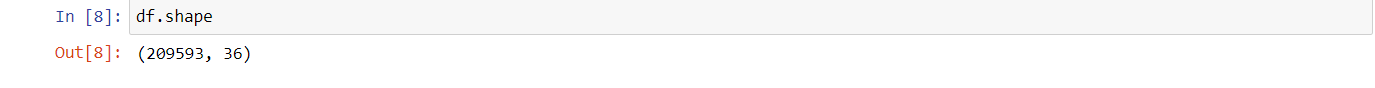
DATA SOURCE AND FORMAT:

The data has been provided by client in a comma separated values(.csv) format.

1. The data will be loaded into pandas dataframe.



1. 2. Checking no. of rows and columns of the data frame and the data type of columns.



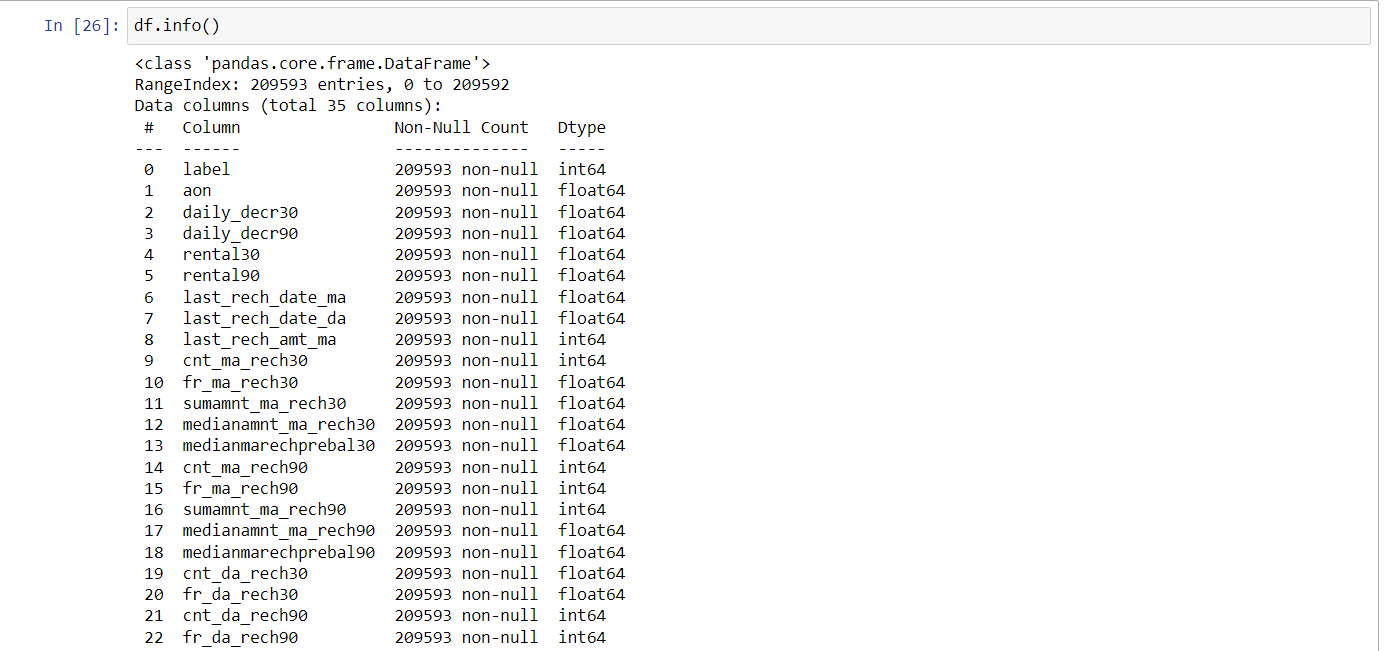
This data set has around 2 lakh rows and 36 columns. There are 3 object columns namely msisdn, pcircle, pdate. Msisdn is the mobile number of customers, Pcircle is the telecom circle and pdate is the date.

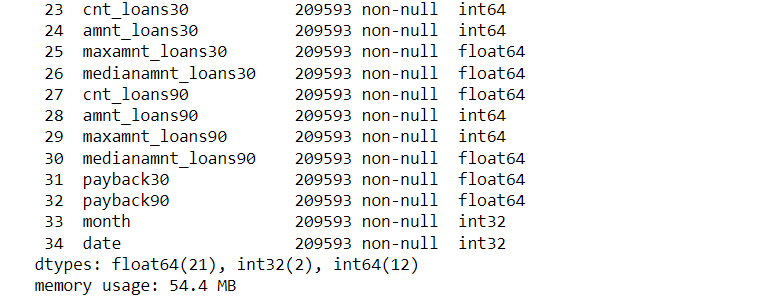
DATA PRE-PROCESSING:

Data pre-processing is a technique of converting raw data into useful format.

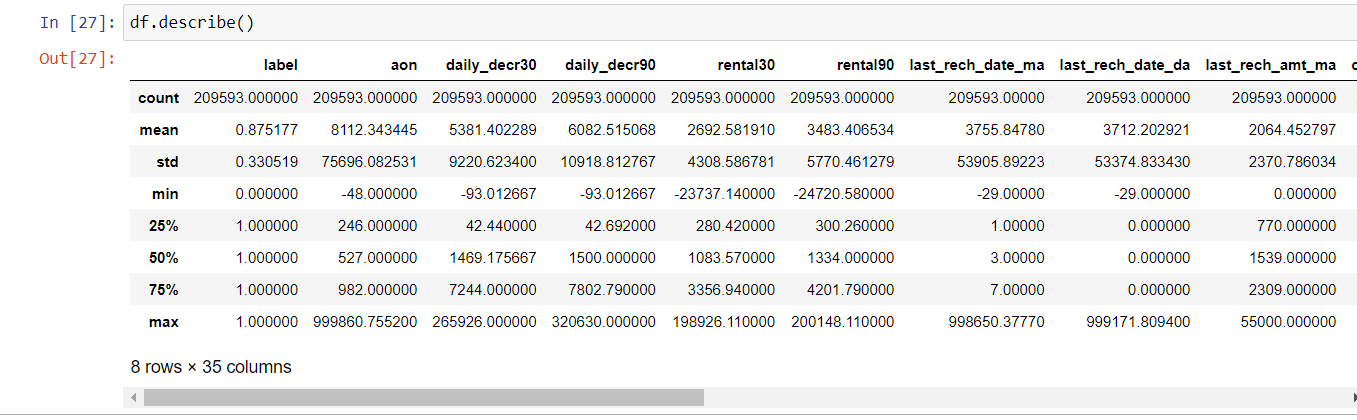
Data cleaning is a part of pre-processing technique which involves filling missing values. For the given dataset it has been mentioned that there are no null values.

Also, the data gathered belongs to 2016 year, hence it won’t be impacting the output due to same entry in all the columns. So dropped the year column. Later converted the month and date columns to integers





We can see that all the columns are of numeric type. Describe method is used to view some basic statistical details like percentile, mean, standard deviation etc. of a data frame or a series of numeric values



We can note that there is a huge difference in 75% value and max value for most of the columns which indicate presence of outliers.

COLUMNS WITH NEGATIVE MINIMUM VALUES:

1)aon

2)daily\_decr30=>Daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah)

3)daily\_decr90=>Daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah)

4)rental30=>Average main account balance over last 30 days

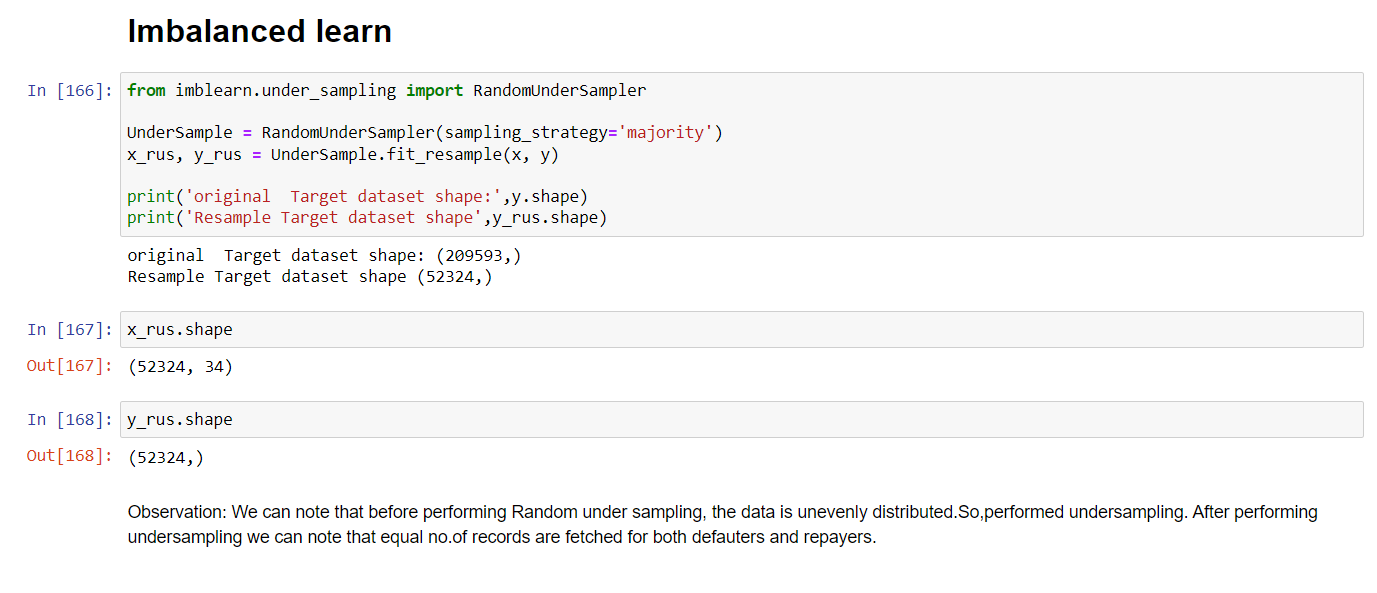
5) rental90=>Average main account balance over last 90 days

6) last\_rech\_date\_ma=>Number of days till last recharge of main account

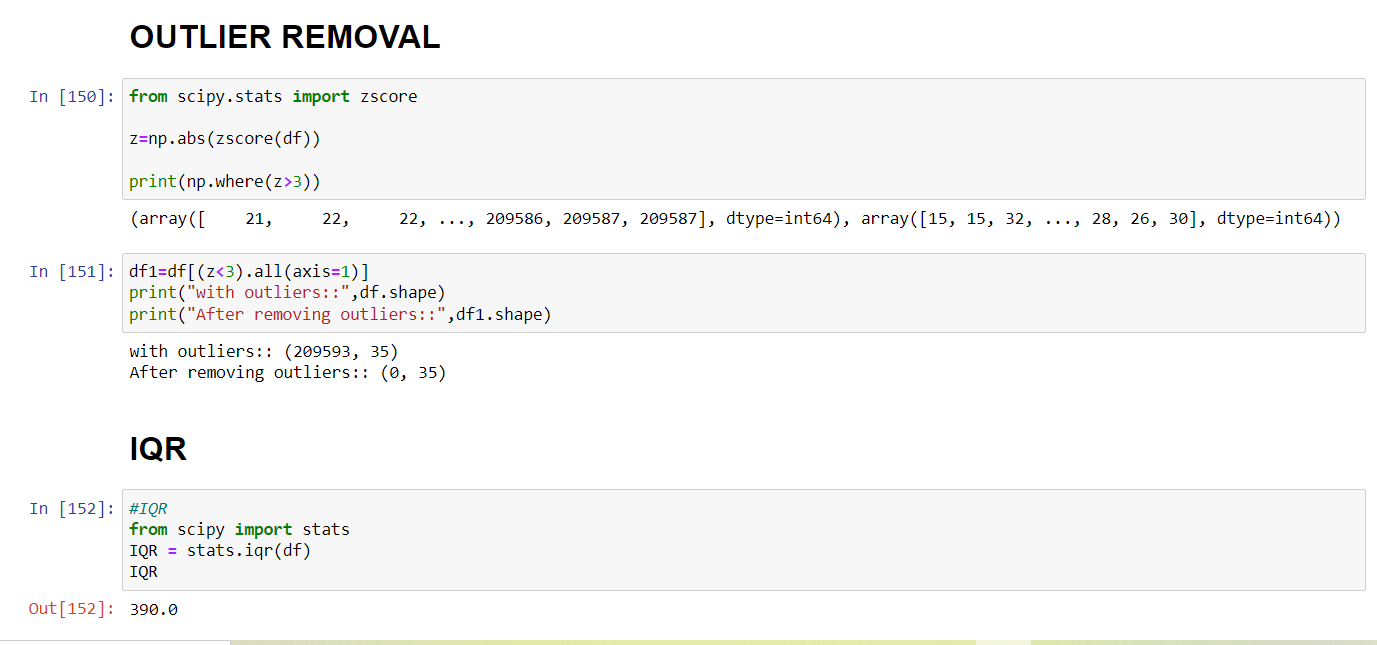
7) last\_rech\_date\_da=>Number of days till last recharge of data account

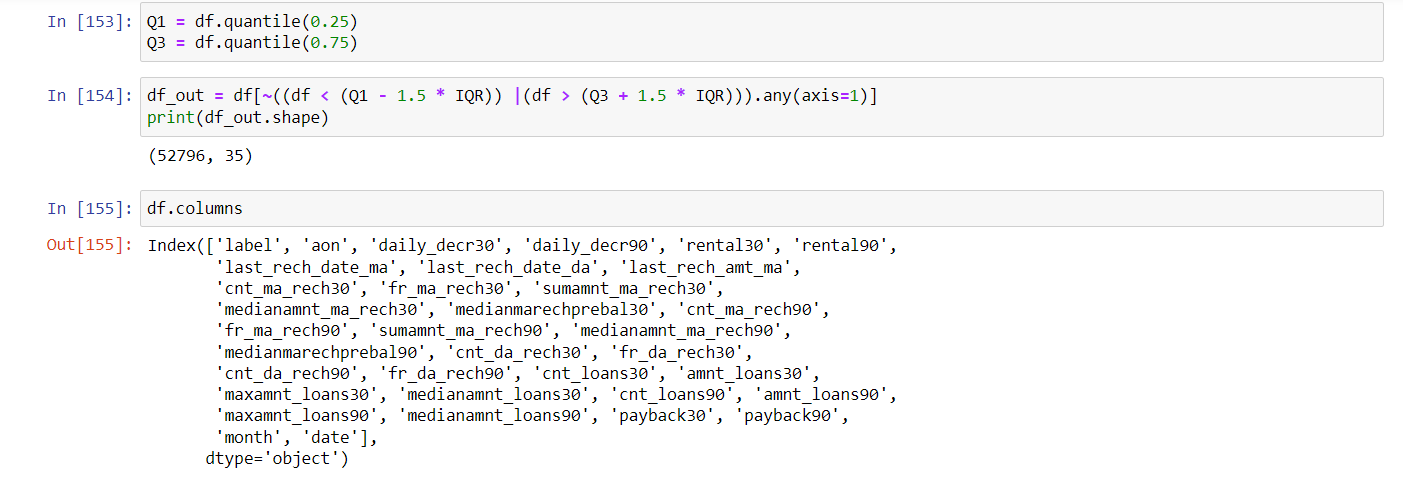
Hardware and Software Used: Software requirement: Anaconda, Jupyter notebook Libraries and packages used: Numpy, Pandas, Sklearn, seaborn, Matplotlib, imblearn, scipy.

Model/s Development and Evaluation Problem-solving approach: The data set is imbalanced since it has large no. of records which contains data about those repaid the loan and less no. of records of those who defaulted loan. This might result in biased predictions. So, used imblearn library to reduce the imbalances. The imblearn library provides different approaches one is Random under sampling. In context of this problem, RandomUnderSampling reduces the no. of records of those who paid the loan. To be precise, random under sampling deletes data from the majority class such that there will be equal no. of samples of both the classes.



Statistical methods used: Outlier removal: Mostly outliers are removed by either z score or IQR(Inter Quartile Range).Tried both these approaches first but, the data loss is high in both these approaches. It has been mentioned in guidelines that the data loss should not exceed 7%. So applied capping technique which is also called as winsorization.





Testing of Identified Approaches (Algorithms):

List of algorithms used:

• Logistic Regression

• Decision Tree Classifier

• KNeighborsClassifier

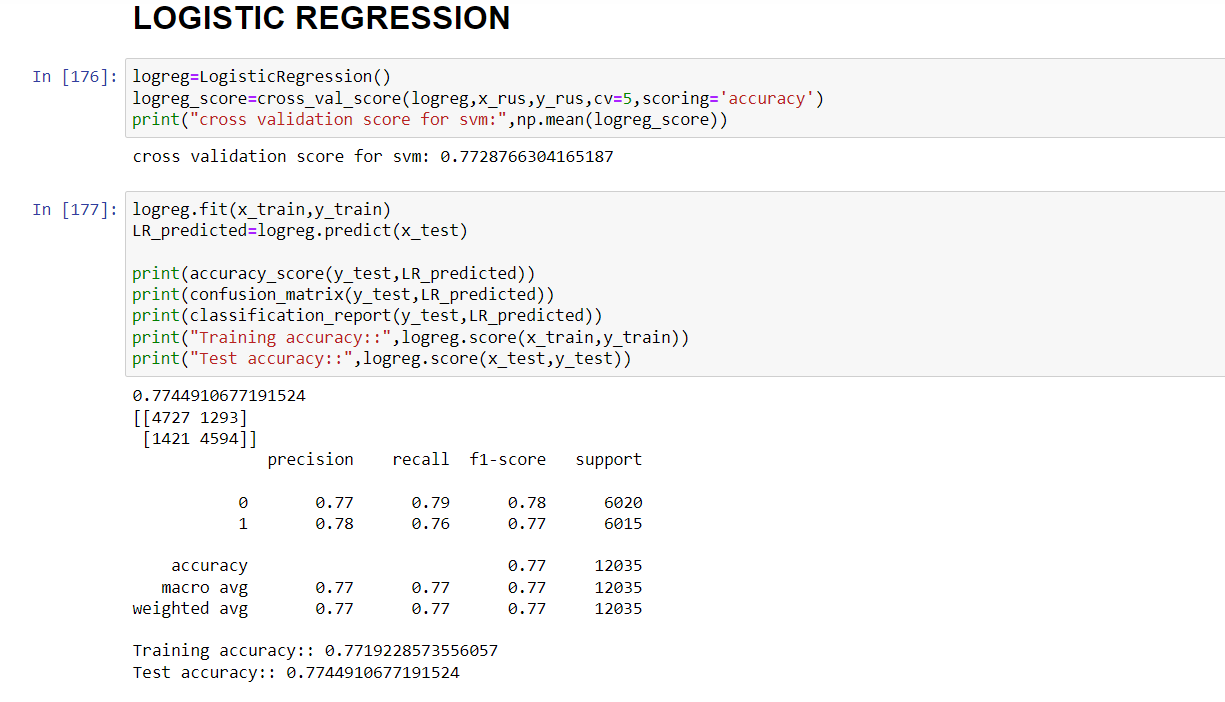
• RandomForestClassifier

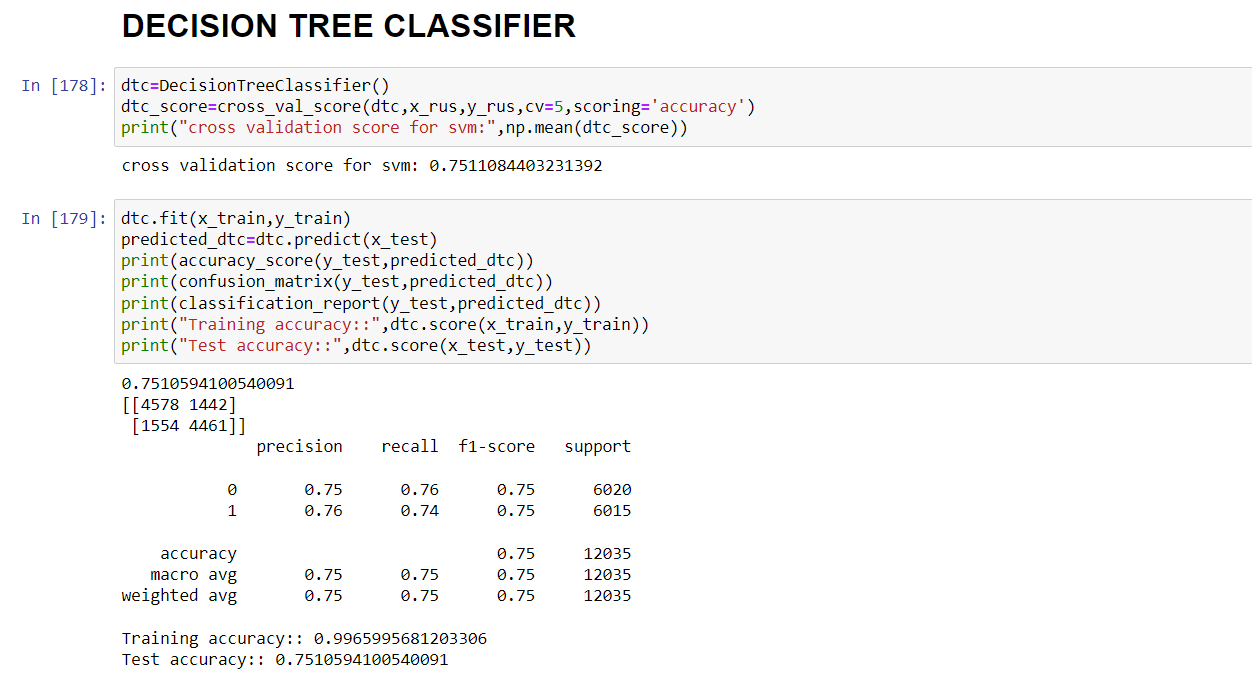
• AdaboostClassifier

• BaggingC;assifier

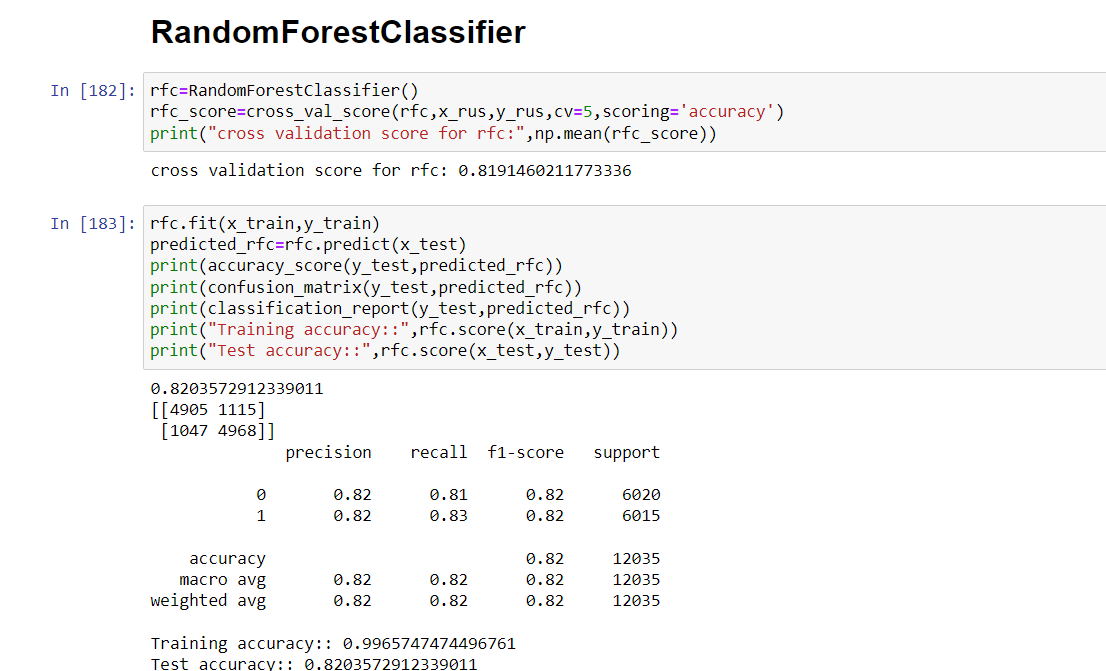
• GradientBoostingClassifier

Run and evaluate selected models: Cross-validation is used to test the model's ability to predict new data that was not used in estimating it. Cross validation used in scenarios where we need to avoid over fitting

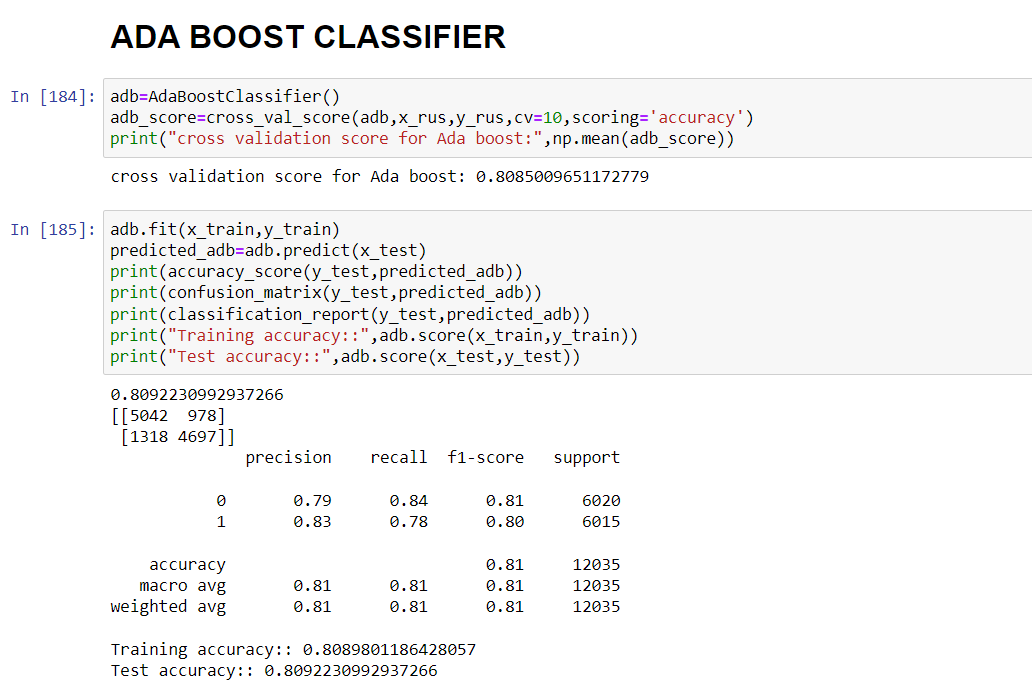


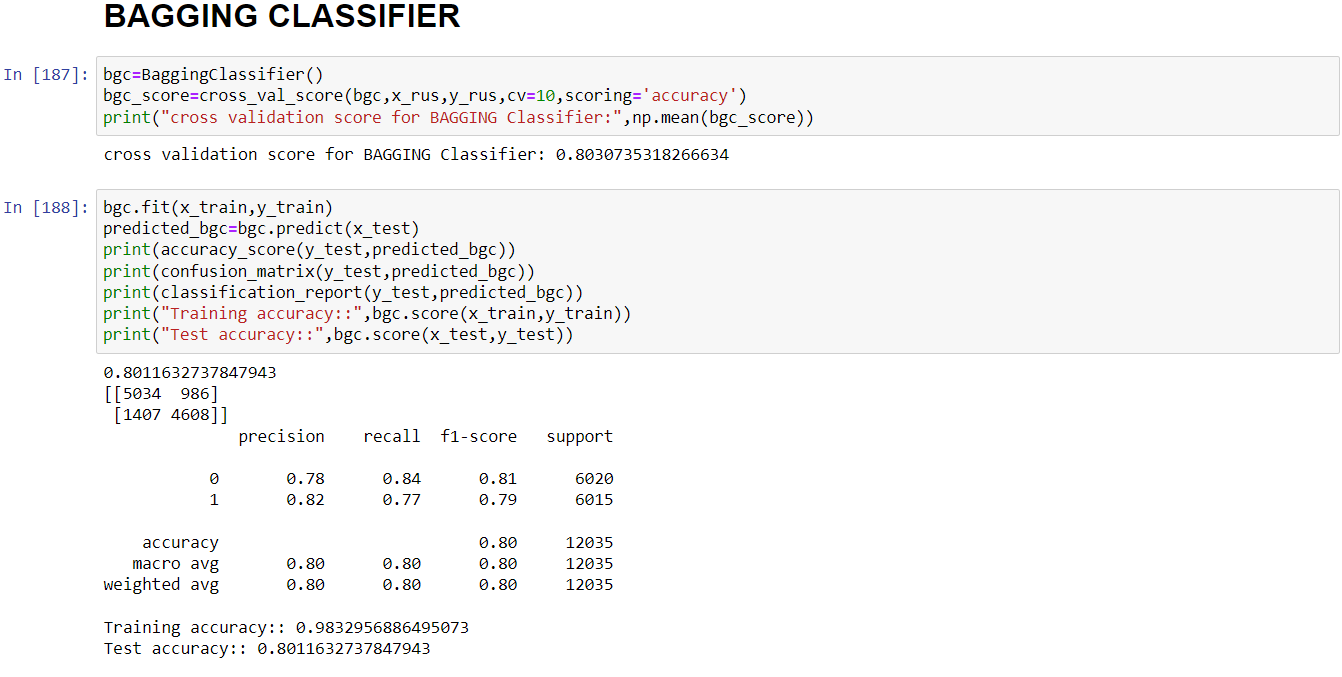


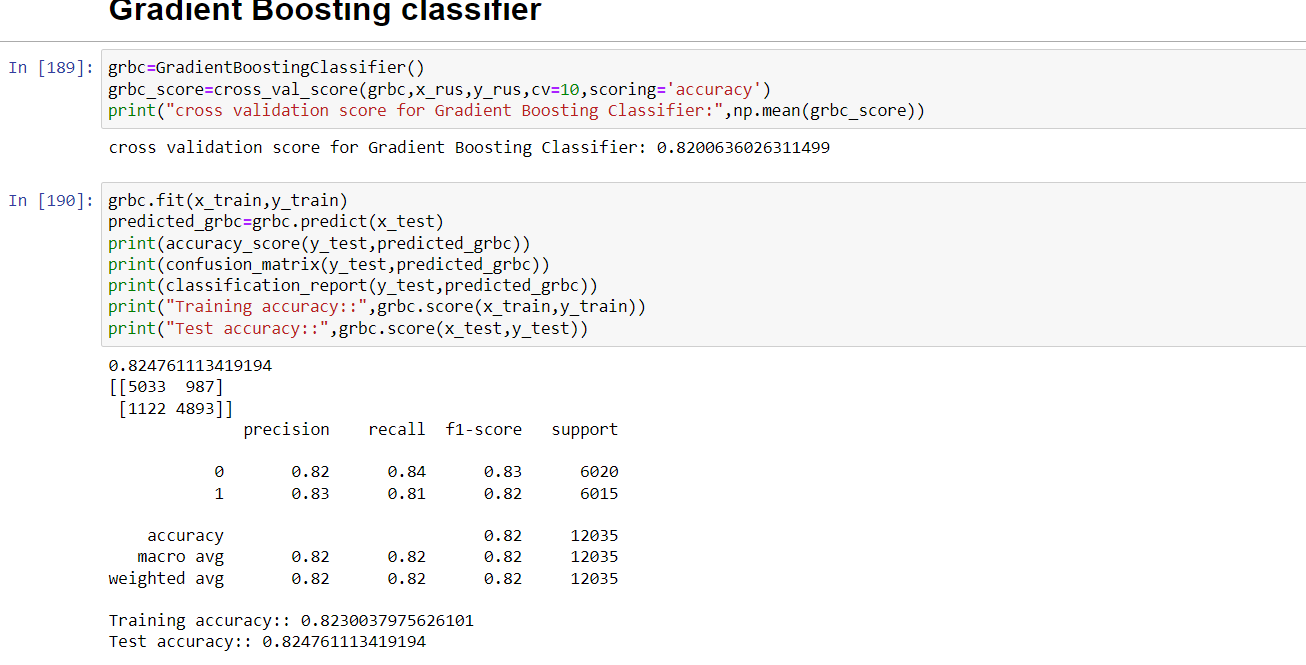




Ensemble models in machine learning operate on a similar idea. They combine the decisions from multiple models to improve the overall performance. The idea behind bagging is combining the results of multiple models to get a generalized result. Here I have used the following ensemble techniques.







Key Metrics for success in solving problem under consideration:

A confusion matrix helps us gain an insight into how correct our predictions were and how they hold up against the actual values.

The following metrics are used:

1)Accuracy: Accuracy is the ratio of the total number of correct predictions and the total number of predictions.

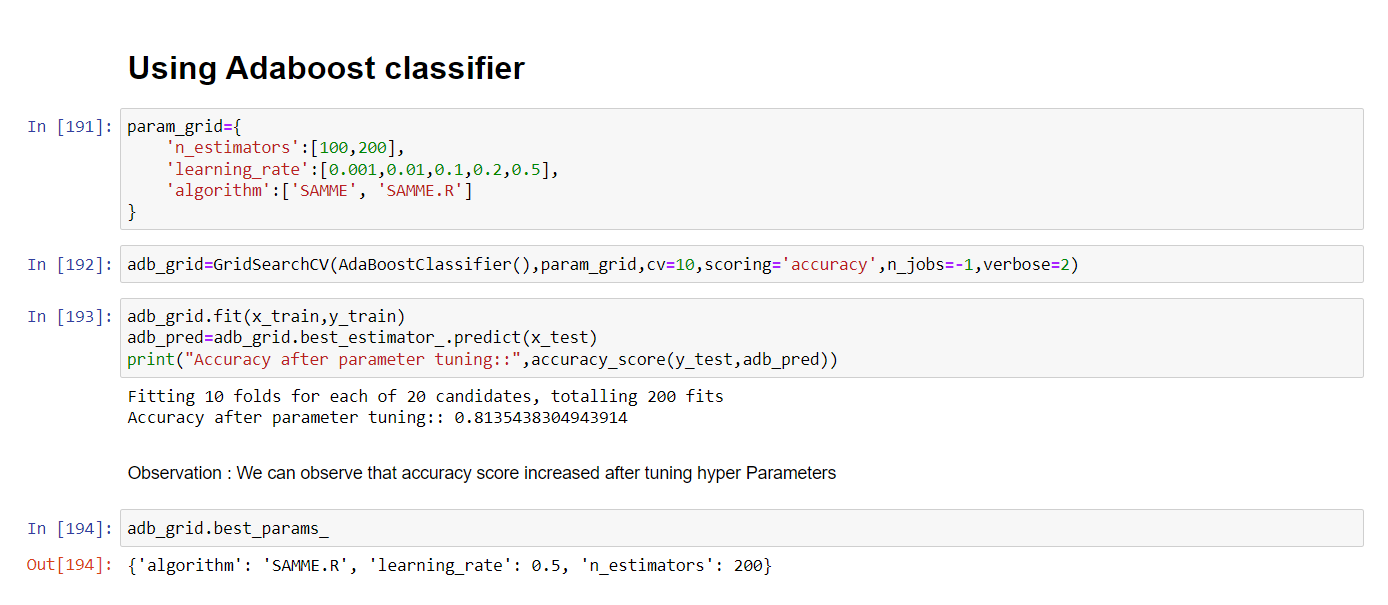
2)Precision: Precision is the ratio between the True Positives and all the Positives

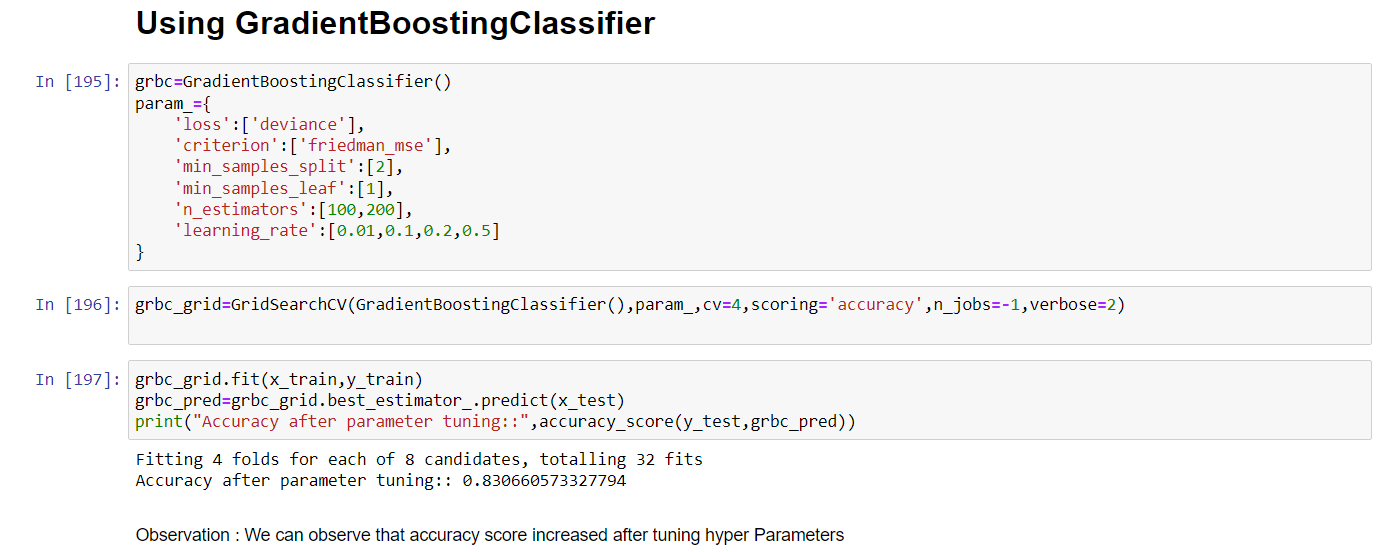
3)Recall: The recall is the measure of our model correctly identifying True Positives

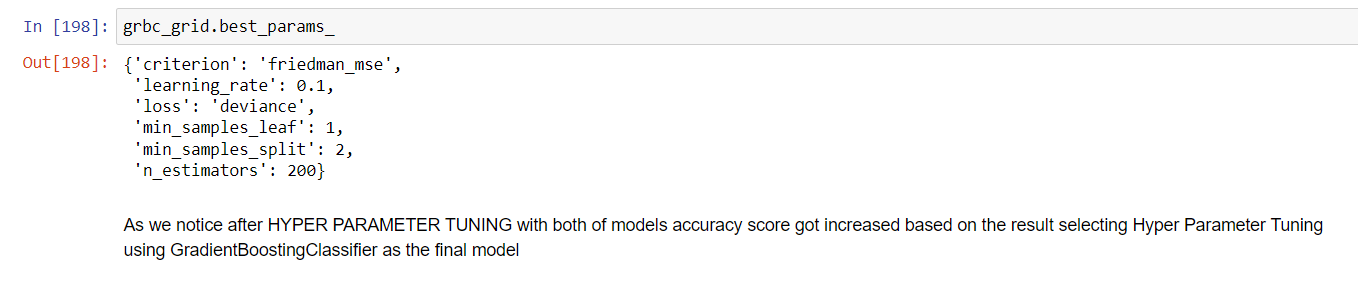
4)F1 score: F1 Score is needed when you want to seek a balance between Precision and Recall.

HYPER PARAMETER TUNING:

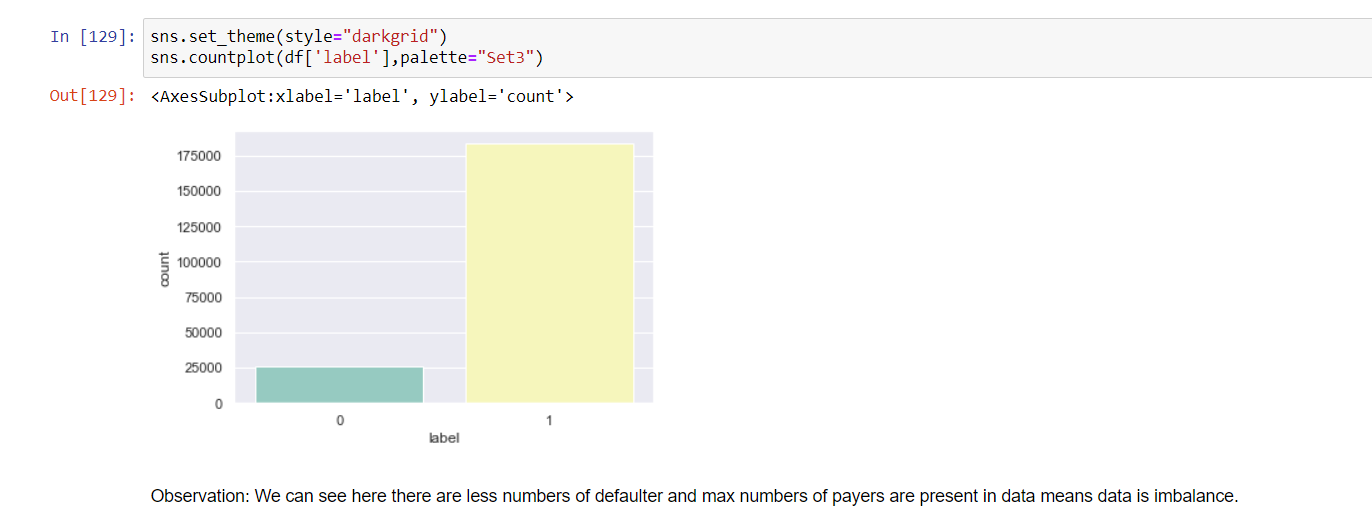
Hyper parameter tuning is used to increase the performance of the algorithm

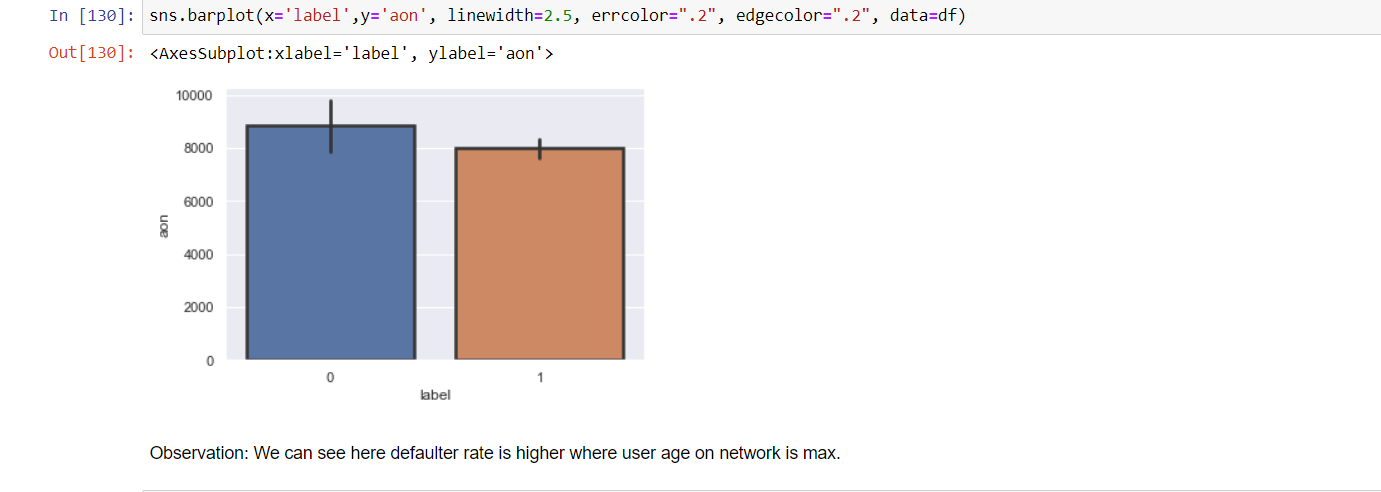






Visualizations:



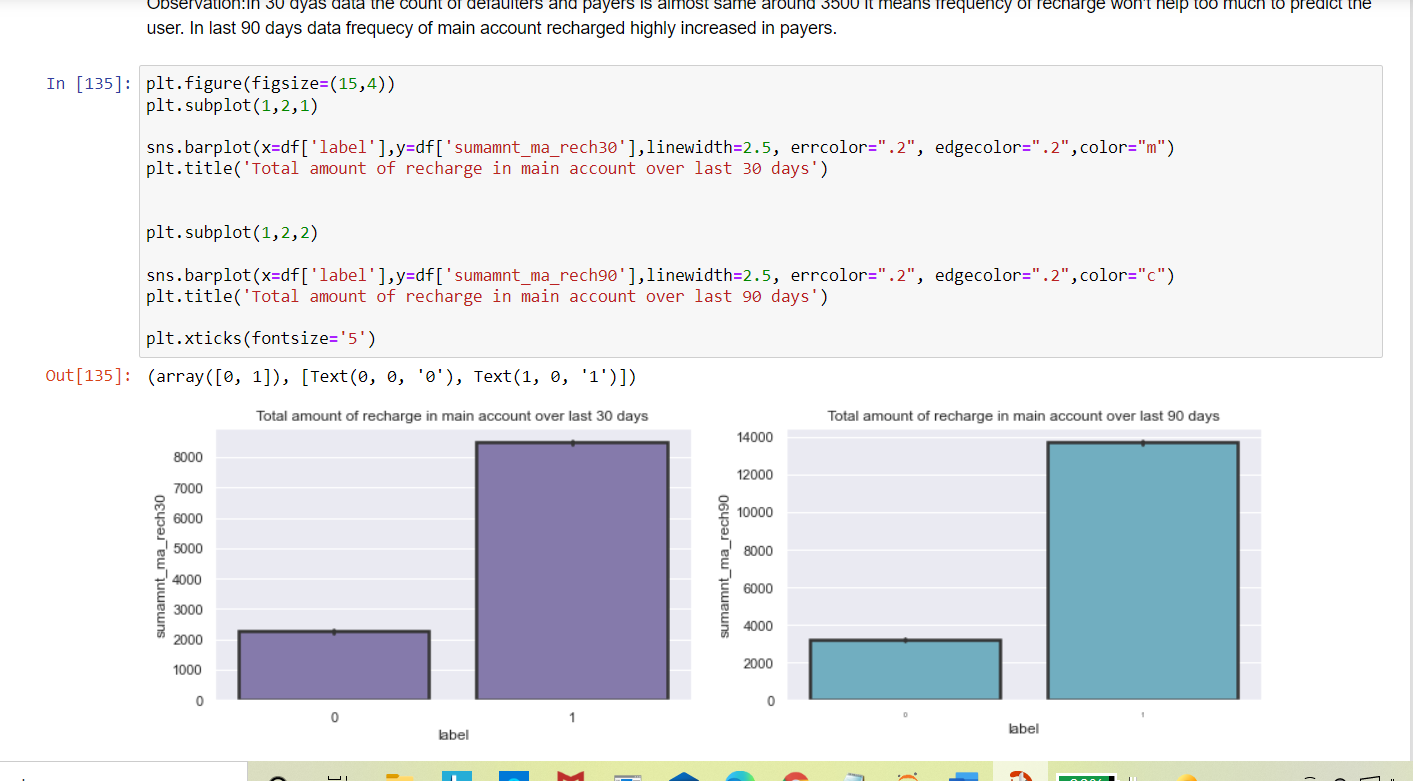






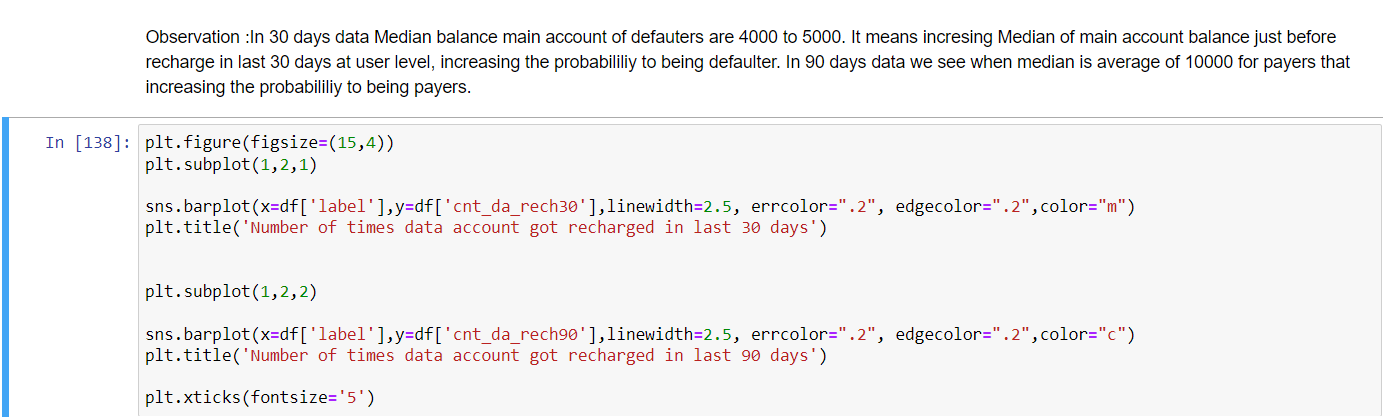


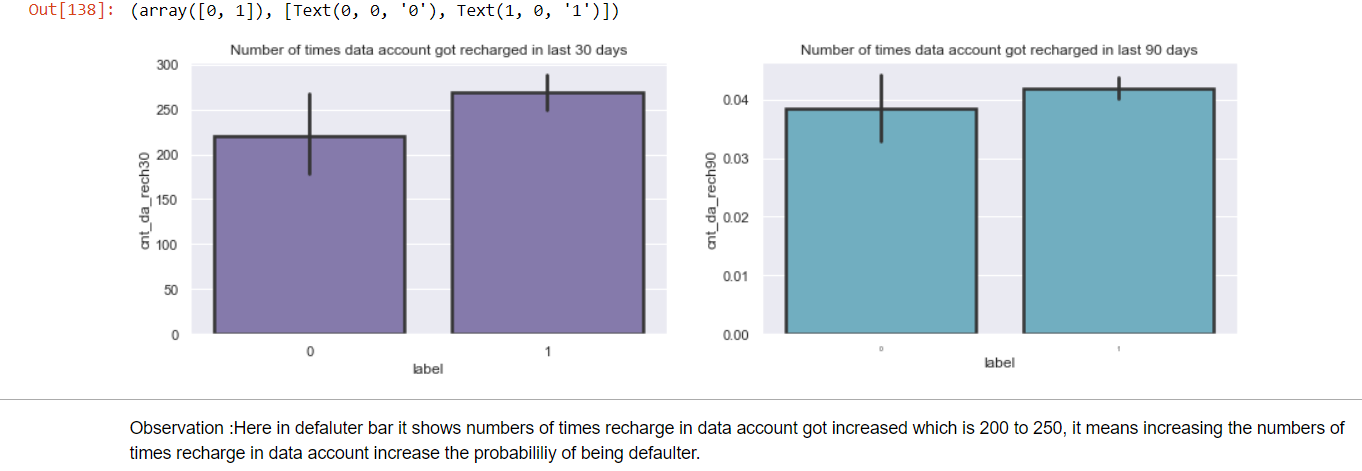




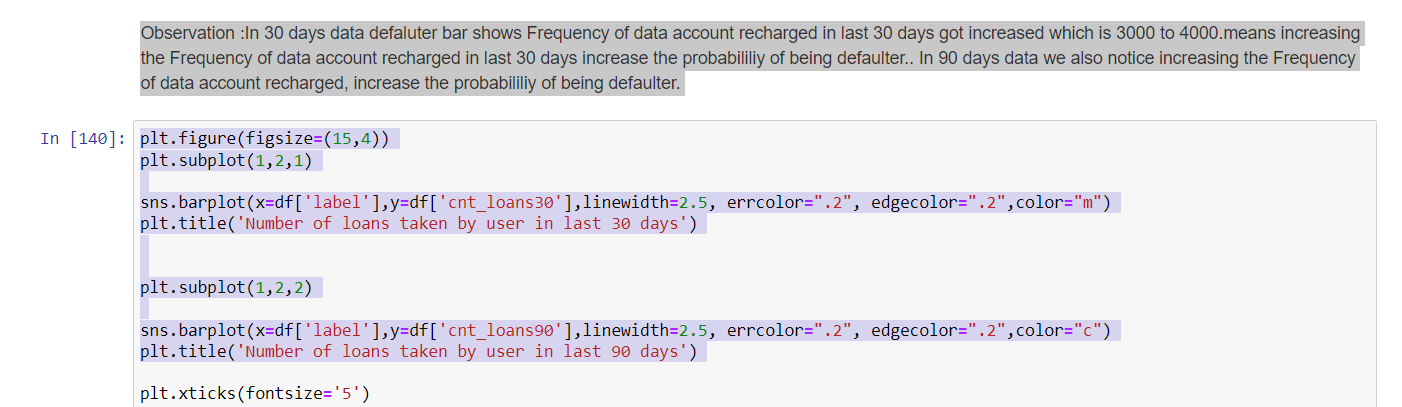


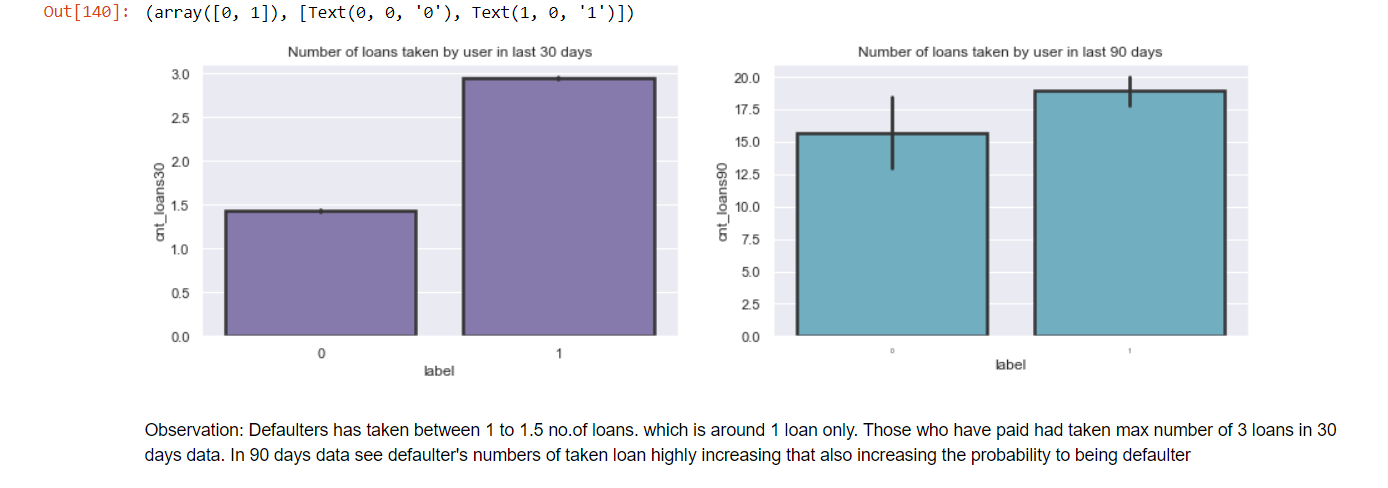




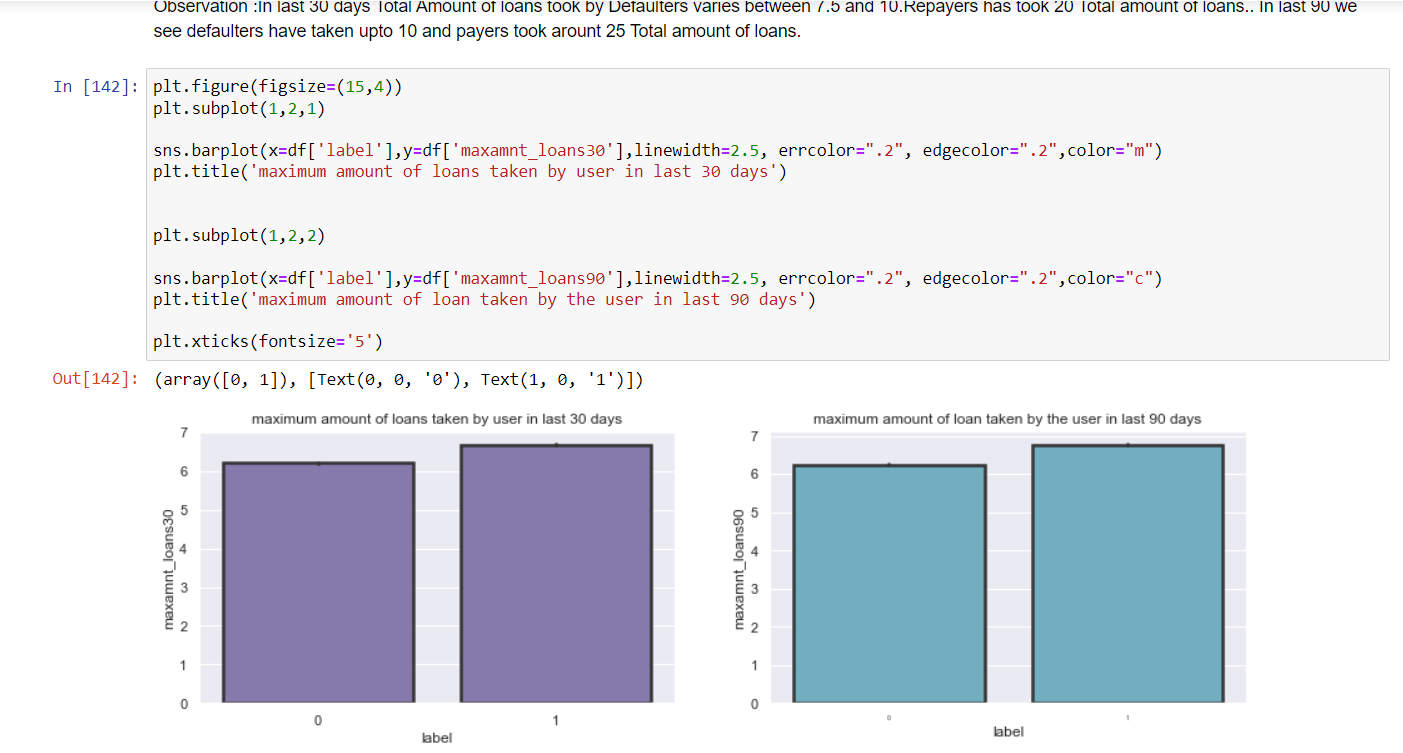


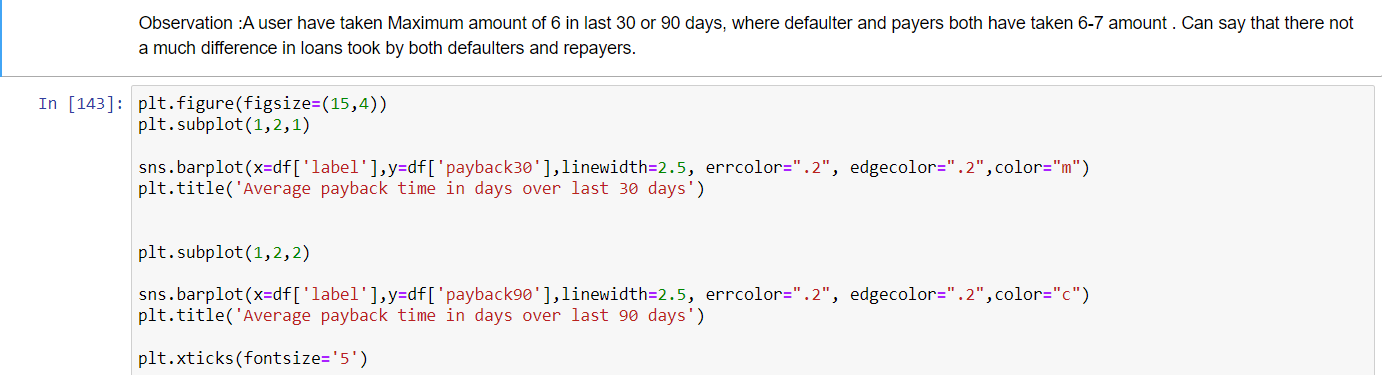


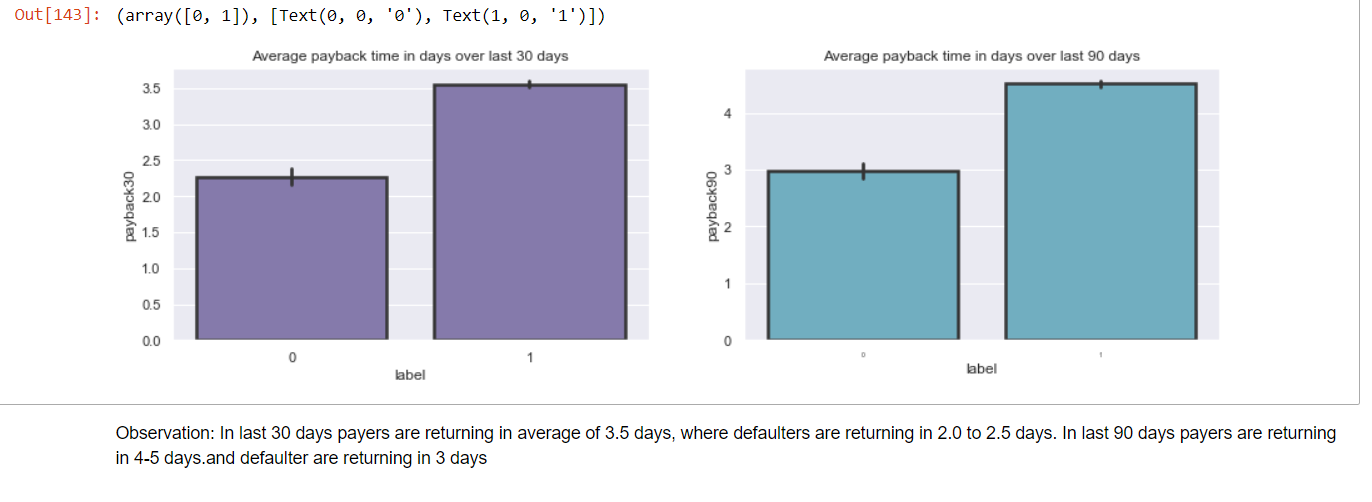


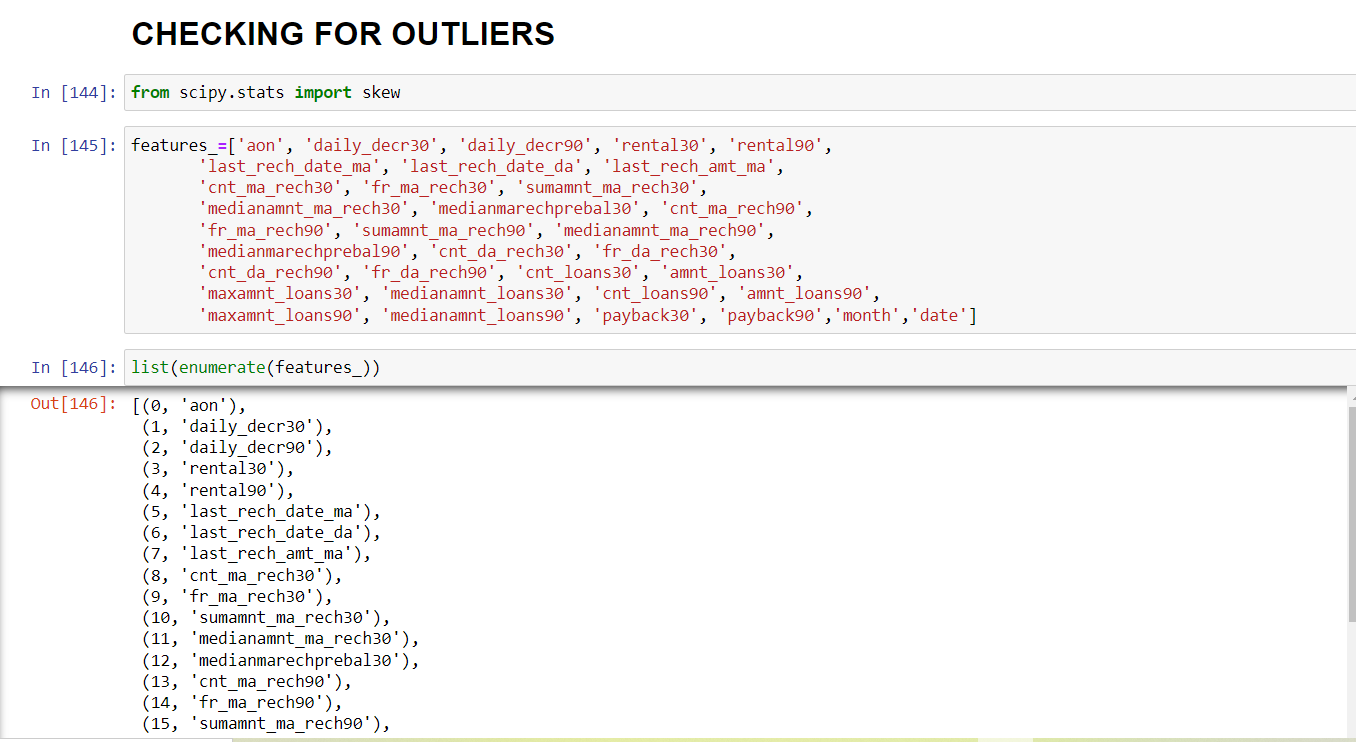


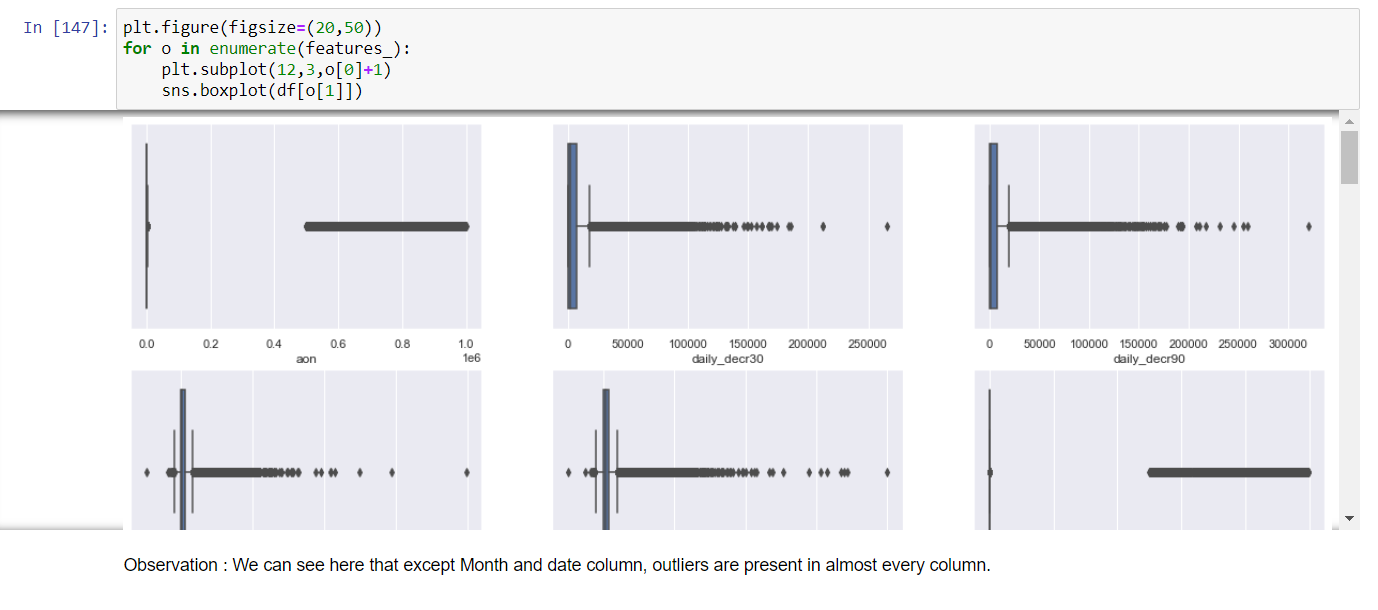


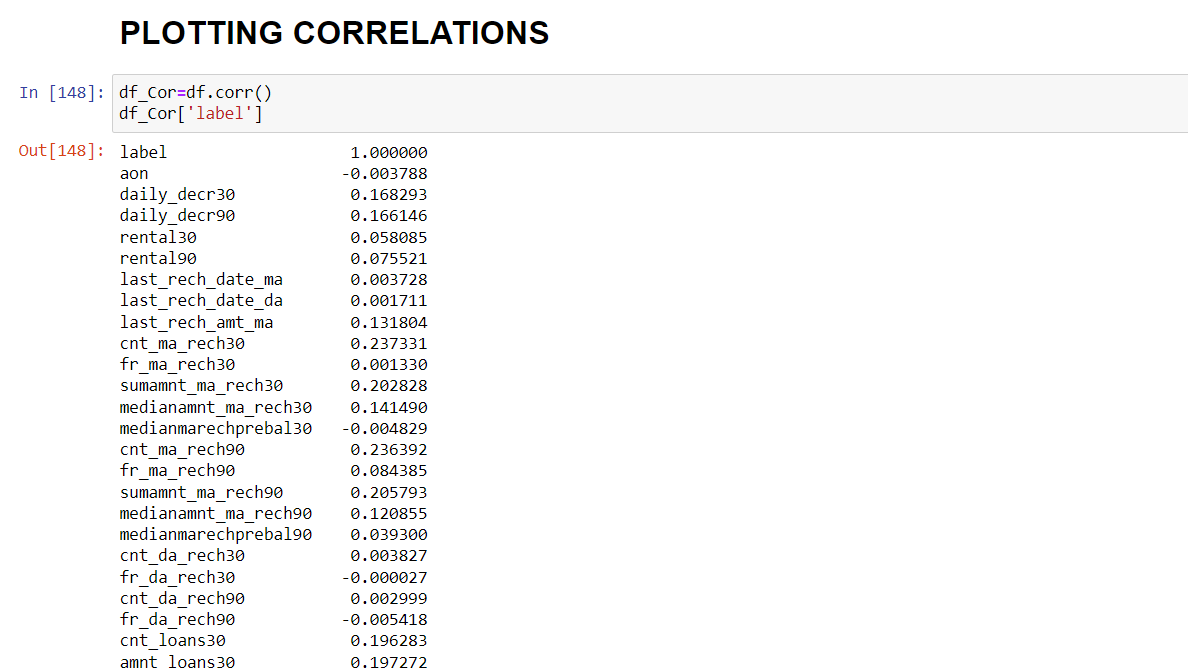


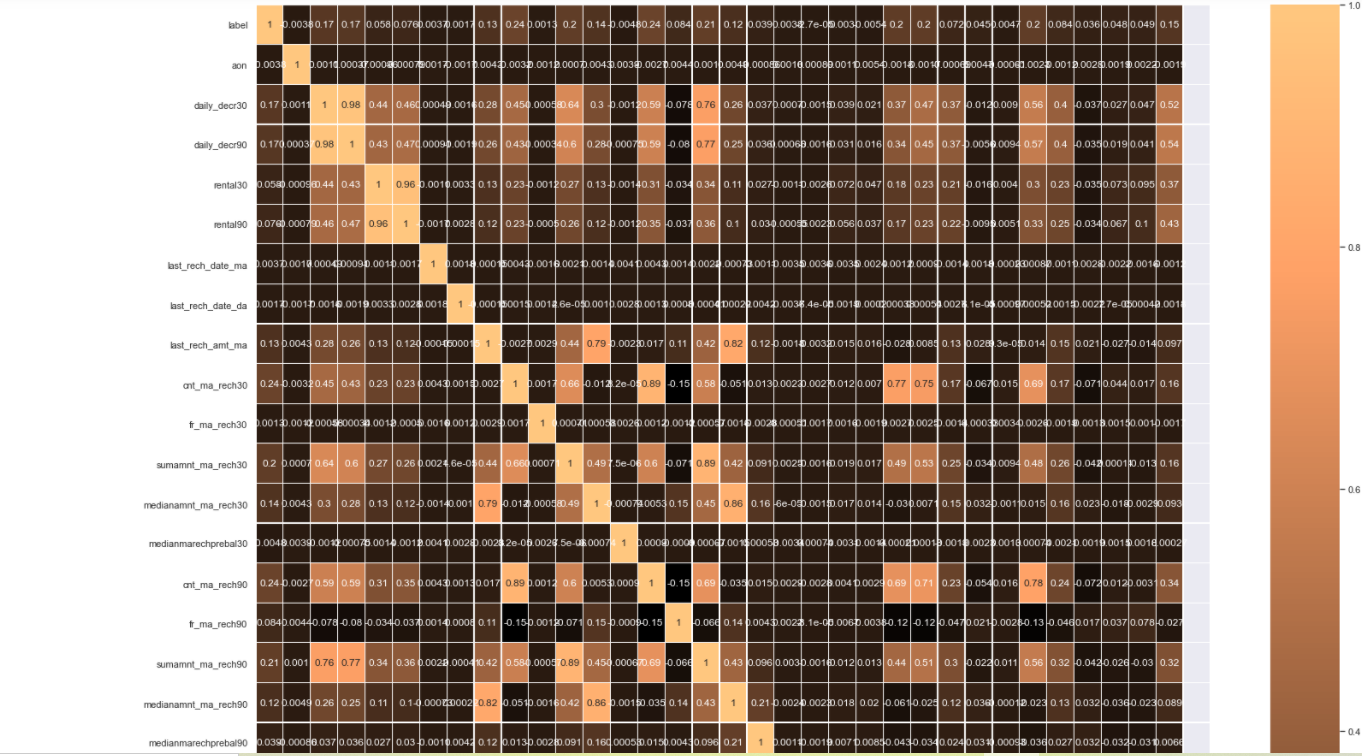












Interpretation of the Results:

• We can note that there is less data about defaulters and more about those who did repay their loan. Hence can say that the data is imbalanced

• With increase in Age on Network, defaulting rate is higher.

• The data is collected based on different parameters for two time periods. One observation is for 30 days and other is for 90 days.

Analysing the parameters separately.

• For 30 days:

1) With the increase in Average Main balance, there is a probability of defaulting.

2) Defaulters recharged Main account max number between 1 and 2 times. Whereas re-payers recharged for 4 plus times.

3) On an average the defaulters have recharged for a max of 1000 Indonesian Rupiah for Main Balance.

4) Defaulters has recharged the data account for a maximum of 200 to 250 times. With increase in No. of times data accounts recharge, probability of defaulting is high

5) A defaulter may default after 2 days, re payers took average of 3.5 days.

6) Defaulters took 1 loan, re payers took 3 loans.

• FOR 90 DAYS:

1) the defaulters have spent a max of 1000 from main account, Repayors has spent 7000 rupiah.

2) Defaulters average main account balance =2000 to 2500 Repayors average main account balance = 3500

3) Defaulters recharged main account for 2 times. Re-payers recharged main account for 7 times.

4) Defaulters’ frequency of main account recharge is 5, Repayors frequency of main account recharge is 8.

CONCLUSION

• Key Findings and Conclusions of the Study:

The defaulting rate is higher in old customers. Defaulters recharge for the main account less no. of times but does recharge for data account more no. of times. Re payers recharge the main account more no. of times when compared to defaulters.

• Learning Outcomes of the Study in respect of Data Science One of the challenges I faced while data cleaning is outlier removal, in most of the scenario’s Z-score will be used as outlier removal technique since it performs quite well with less data loss. In our data set, Z-score has caused 22% data loss.

Then I tried another famous technique called InterQuartileRange it caused around 80% data loss. Another technique is replacing the outlier data with mean or median.

But when we observe this data set there is a huge difference between minimum and maximum values. If we calculate mean or median it won’t give appropriate values as it includes the outlier value (maximum ones).So not using this approach. As we are not dropping the outliers, another approach is capping or winsorization of outliers.

Using percentile capping. Values that are less than the value at 10th percentile are replaced by 10th percentile value, and the value greater than 90th percentile are replaced by 90th percentile value.

The other challenge is when I used the imbalanced data, the accuracy was very high but there was bias in predictions.

So, I used imblearn to reduce the imbalances in the target variable.

Limitations of this work and Scope for Future Work:

This data set contains data of the year 2016 belonging to PSW telecom circle. If we get data of other years along with other telecom companies we can predict on varied scenarios.

Thank You