

**Micro-Credit Defaulter**

**Project**

**Submitted by:**

Pratiksha Potghan

Data Science Intern at Flip Robo Technologies

**ACKNOWLEDGMENT**

It is my pleasure to present this report. Working on this project was a great experience that gave me very informative knowledge of data analysis.

All the required information and dataset are provided by Flip Robo Technologies (Bangalore) which helped me to complete the project.

I want to thank my SME Mohd Kashif Sir for giving the dataset and instructions to perform the complete case study process.

**INTRODUCTION**

**Problem Statement:**

* **Business Problem Framing**

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

* **Conceptual Background of the Domain Problem**

MFS 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).

* **Review of Literature**

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. An attempt has been made in this report to review the available literature in the area of microfinance. Approaches to microfinance, issues related to measuring social impact versus profitability of MFIs, issue of sustainability, variables impacting sustainability, effect of regulations of profitability and impact assessment of MFIs have been summarized in the below report. We hope that the below report of literature will provide a platform for further research and help the industry to combine theory and practice to take microfinance forward and contribute to alleviating the poor from poverty.

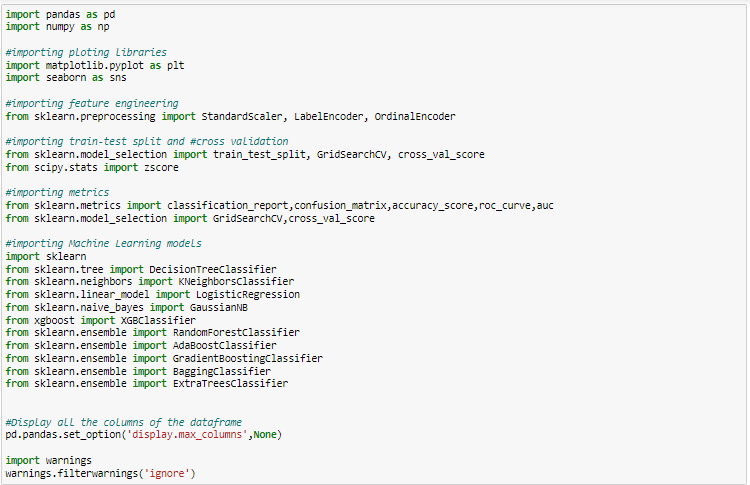
* **Motivation for the Problem Undertaken**

I have to build a model with available independent variable data set by thorough analysis of data. The model will go to management for further research. This micro credit model will help the Finance Company to decide which is defaulter and non-defaulter. Who will return the loan amount within 5 days? So, they can focus on the area which will yield in high return. The relationship between the prediction and economy is important, that will drive a motivation in understand the problem and providing the solution for that problem.

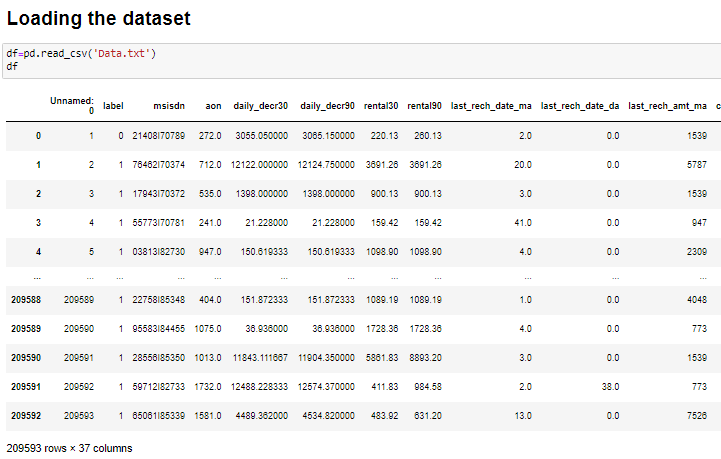
**Analytical Problem Framing**

* **Mathematical/ Analytical Modeling of the Problem**

Firstly I imported all the required libraries

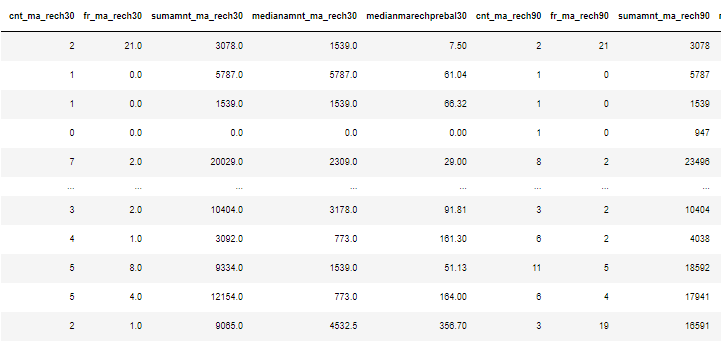
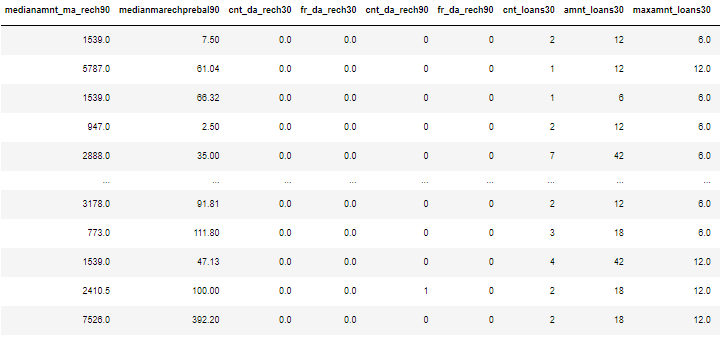


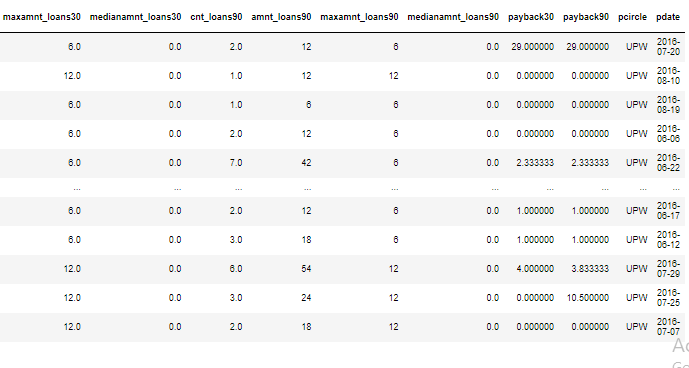
After importing libraries, I have loaded the dataset using read\_csv as below.



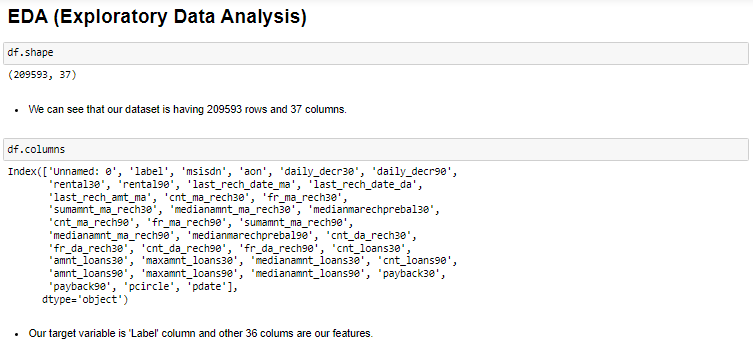
We can see that there are 37 columns, from which column ‘label’ is our target variable and other 36 columns are our independent variables.

In coming steps, we have converted the column ‘pdate’ to the day, month, year format.





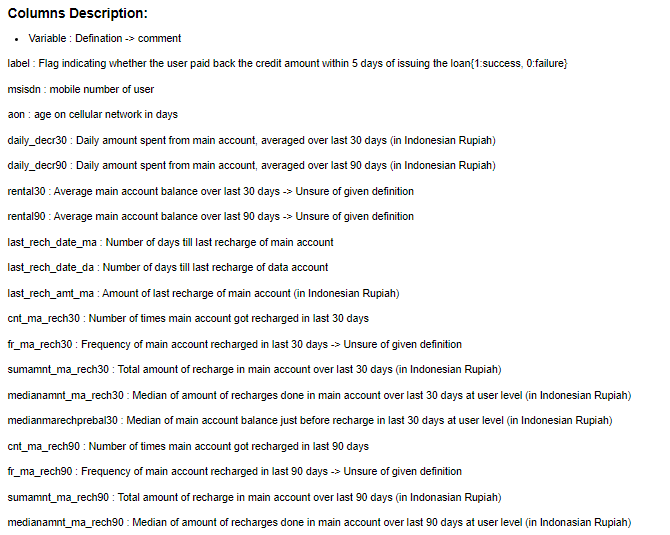
I have performed Exploratory data analysis (EDA), Data Cleaning, Data Visualizations, checking Statistical Data, checking correlation, Feature Selection and then finally building model.

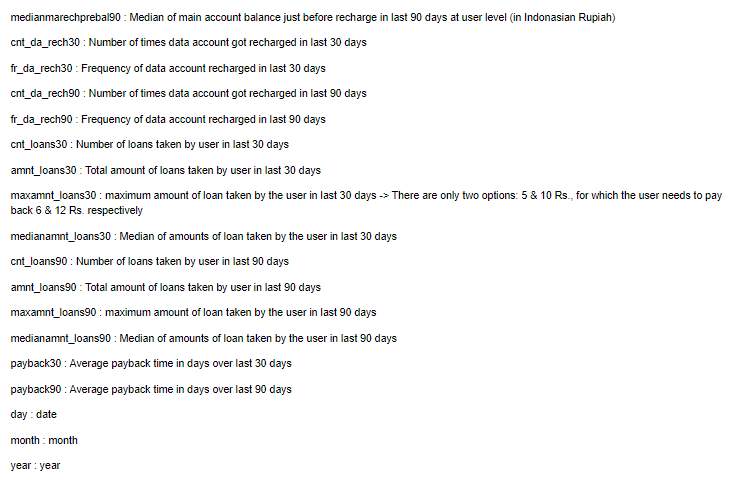


In EDA, I checked for count of dataset using df.shape, after that checked all 37 columns of dataset.

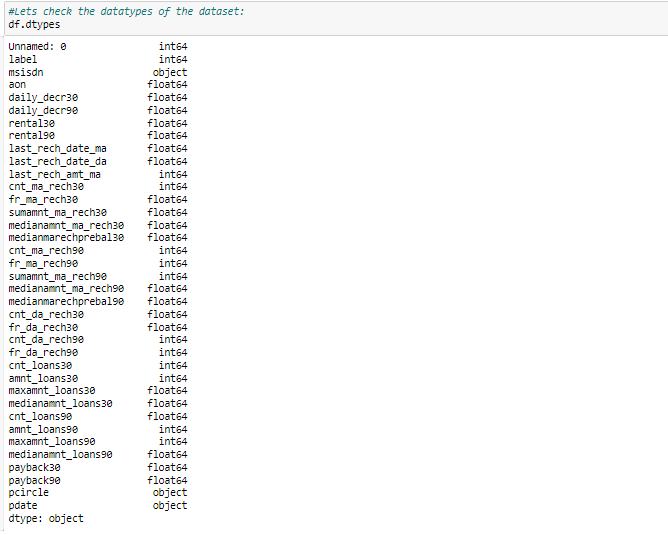
* **Data Sources and their formats**

The columns of the data are as below:

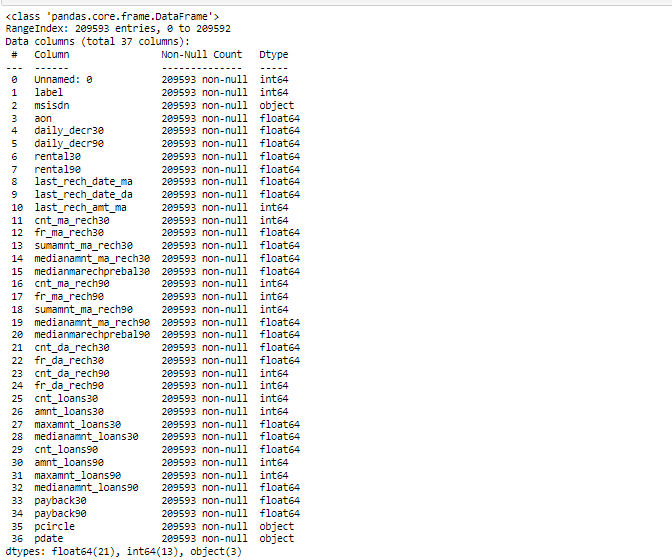




I have checked the dtypes of these columns as below:



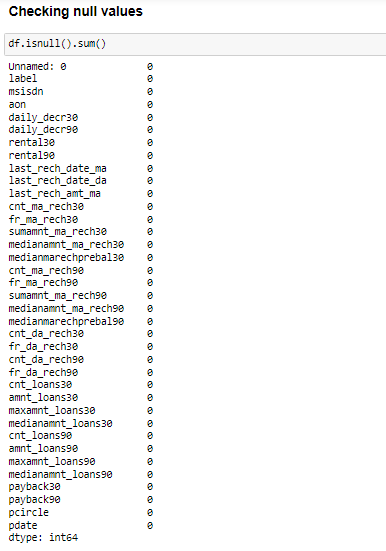
Now we can see dtypes of each column. For more detail information regarding dtypes and non-null values, I have used df.info() as below.



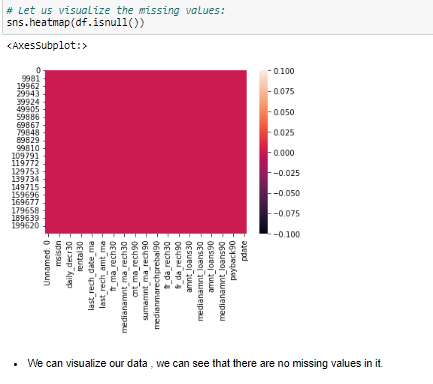
I can see that there are only 3 object dtypes and 34 numerical dtypes. Also I observed that there were no null values in the data.

* **Data Pre-processing Done**

For cleaning data firstly I checked null or missing values in data using df.isnull().sum() and I got results as below.

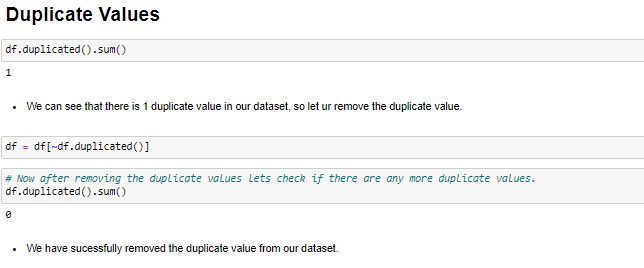


I also visualized the null values data using heatmap as below.

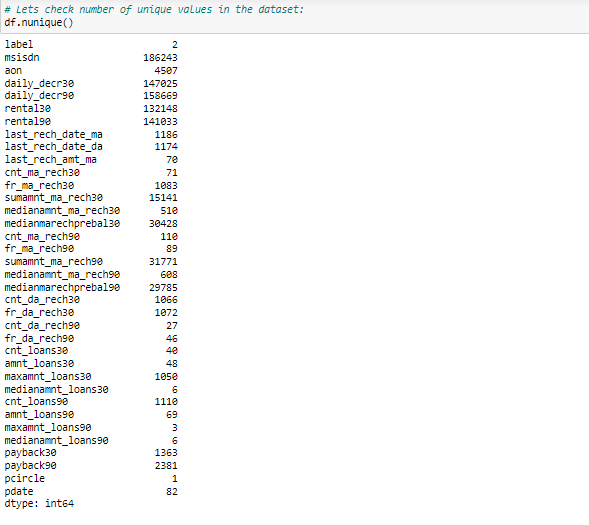


As there were no null values in the dataset, so there was no need of filling any missing or NaN data.

After checking for null values, I checked number of unique values and duplicate values in data if any.



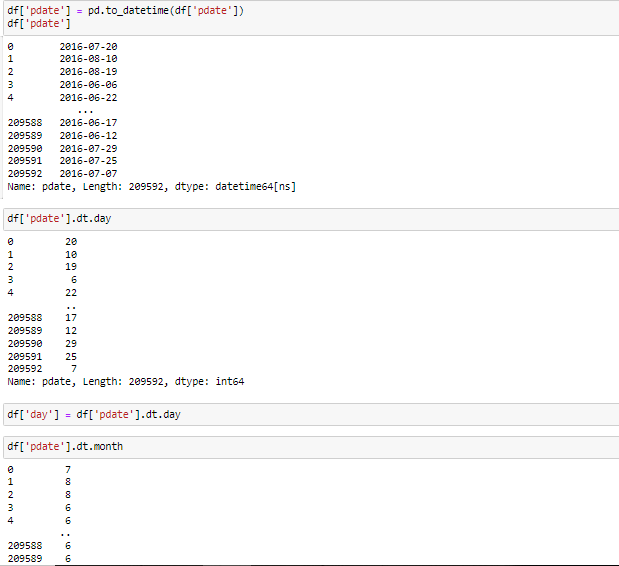
There was 1 duplicate value, which we removed using above function.



By using unique function we came to know that pcircle column had only 1 unique value, so I dropped it. I also dropped column ‘unnamed: 0’ as this was not giving any relevant input to the data.

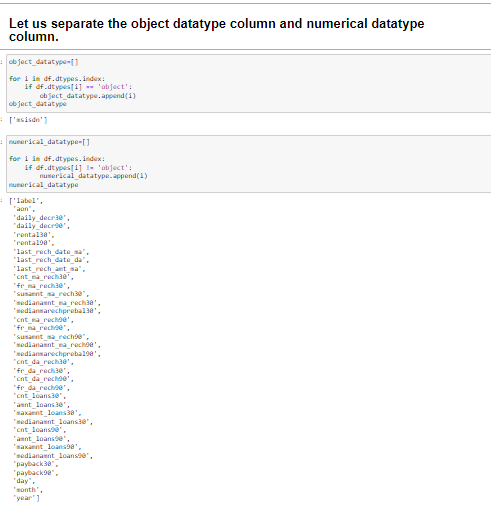
Then I converted the pdate column to the day, month and year using feature extraction.

As per below figure, First I extracted the day and month column from ‘pdate’, year was not extracted as there is only one unique value of year present in the dataset.

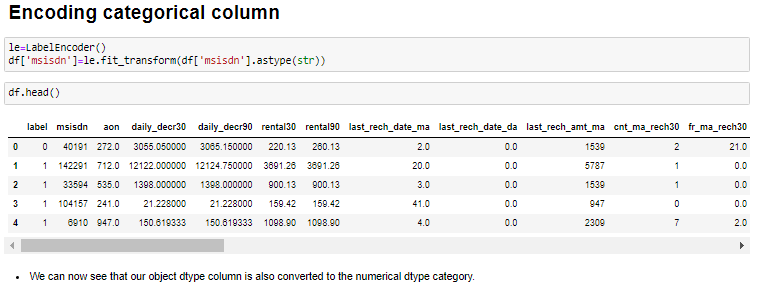


* **Data Inputs- Logic- Output Relationships**

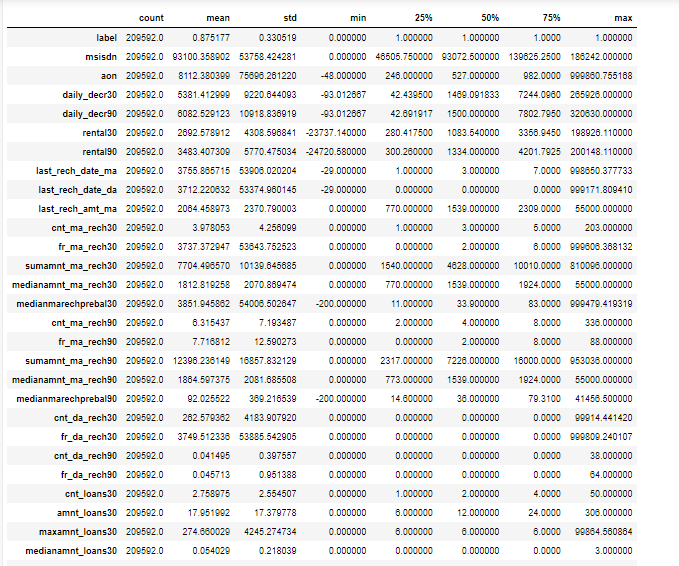
I checked and separated numerical and object dtypes.



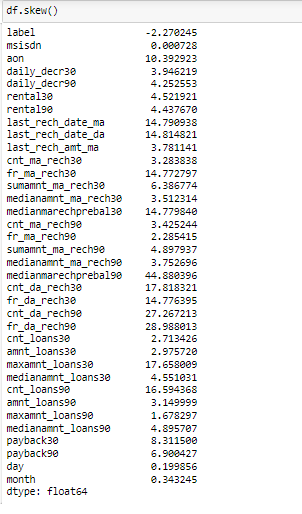
Here only 1 column ‘msisdn’ is in categorical dtype, so I used Encoding technique to encode the data of column ‘msisdn’.



After encoding I checked for the statistical data which gave me count, mean, minimum, standard deviation, 25%, 50%, 75%, maximum values.

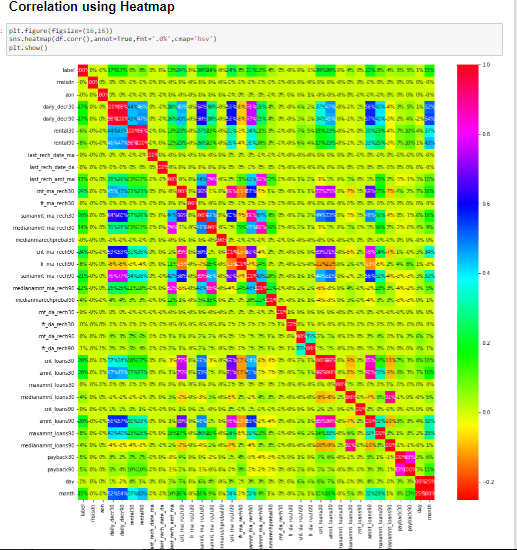


After checking this data, I checked for skewness, outliers in the data.

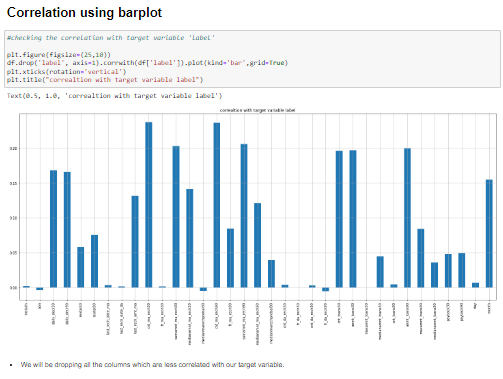




After this I check for the correlation using heatmap as below.



After that I checked the correlation of ‘label’ with other features using barplot.



We observe that the columns cnt\_ma\_rech30 and cnt\_ma\_rech90 are highly positively correlated with label this means as the cnt\_ma\_rech30 and cnt\_ma\_rech90 are increasing the probability of cutomer being non-fraudulent is also increasing.

We also observe that the columns aon, medianmarechprebal30 and fr\_da\_rech90 are negatively correlated with label this means as the aon, medianmarechprebal30 and fr\_da\_rech90 are increasing the probability of customer being fraudulent is also increasing.

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

By looking into the target variable label we assumed that it was a classification type of problem. We observed multicollinearity in between columns so we assumed that we will be using Principal Component Analysis (PCA).

* **Hardware and Software Requirements and Tools Used**

We used Jupyter notebook for this project.

The tools, libraries and packages we used for accomplishing this project are pandas, numpy, matplotlib, seaborn, scipy stats, sklearn.decomposition pca, sklearn standardscaler, collections counter, imblearn SmoteTomek, GridSearchCV, joblib.

**Model/s Development and Evaluation**

* **Identification of possible problem-solving approaches (methods)**

We first converted all our categorical variables to numeric variables with the help of label encoder to checkout the correlation between them and dropped the columns which we felt were unnecessary.

The data was imbalanced so through imblearn’s SmoteTomek package we were able to handle the imbalanced data by increasing the number of fraudulent transactions on relevant data points.

The data was improper scaled so we scaled the feature variables on a single scale using sklearn’s StandardScaler package.

There were too many (37) feature variables in the data so we reduced it to 7 with the help of Principal Component Analysis(PCA) by plotting Eigenvalues and taking the number of nodes as our number of feature variables.

* **Testing of Identified Approaches (Algorithms)**

The algorithms we used for the training and testing are as follows: -

• Extreme gradient boosting classifier

• Decision tree classifier

• KNeighbors classifier

• Logistic Regression

• GaussianNB

• Random forest classifier

• Ada boost classifier

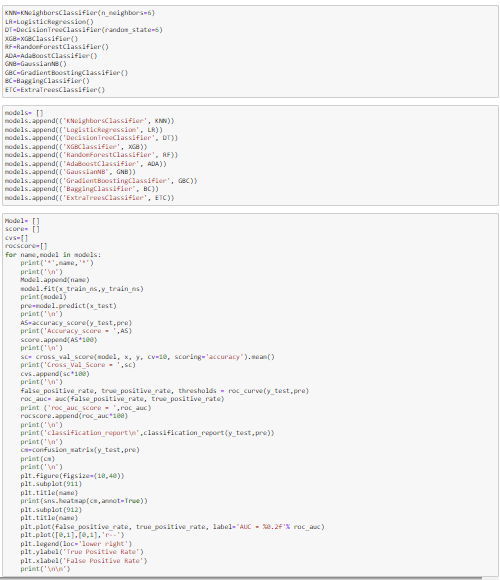
• GradientBoostingClassifier

• Bagging classifier

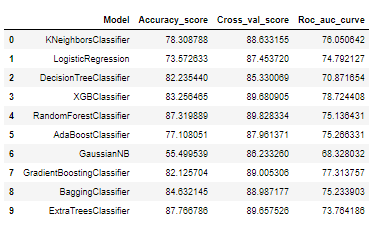
• Extra trees classifier

* **Run and Evaluate selected models**

We used below function for checking accuracy score, cv score and auc-roc score of all the classification models.

****

After performing the above function. I got below results.

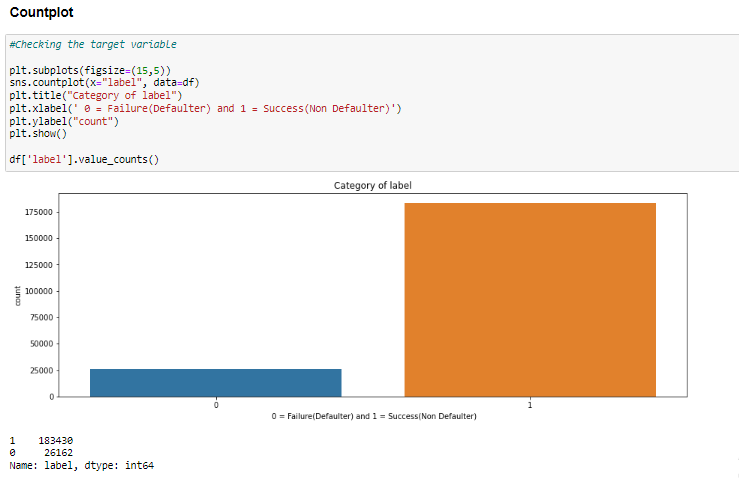


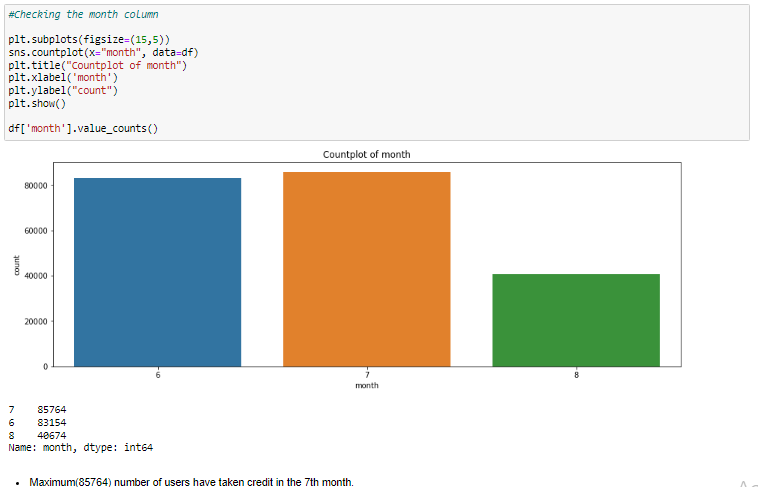
* **Key Metrics for success in solving problem under consideration**

Accuracy is not a appropriate measure of model performance here and we used the metric **AREA UNDER ROC CURVE** to evaluate models performance because high roc score will mean high recall which means the model does well by not classifying legit transactions as fraudulent.

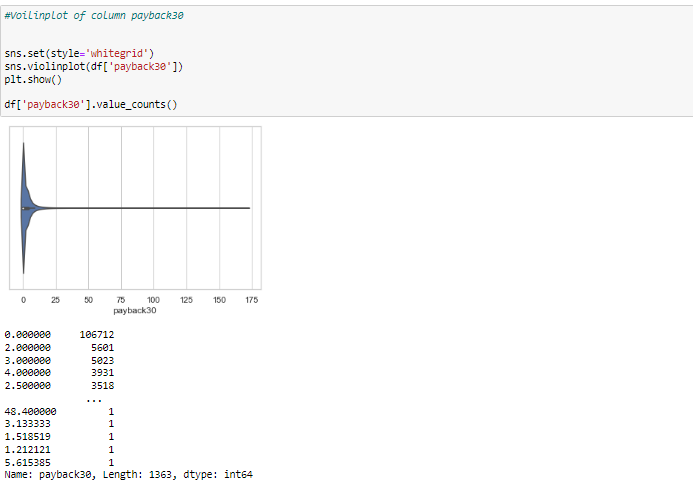
* **Visualizations**

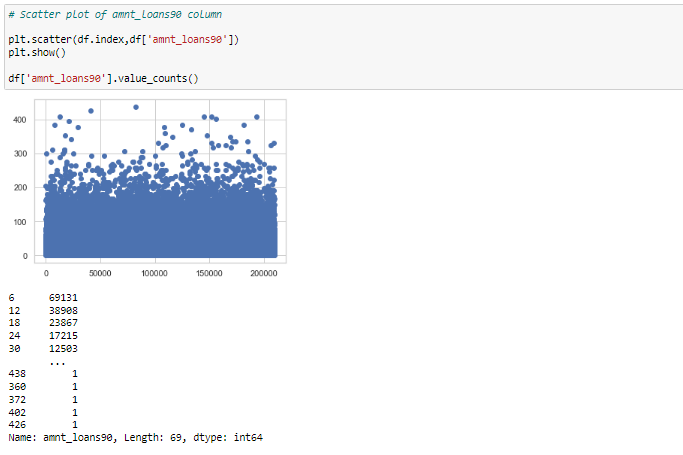
Following are some visualizations done.

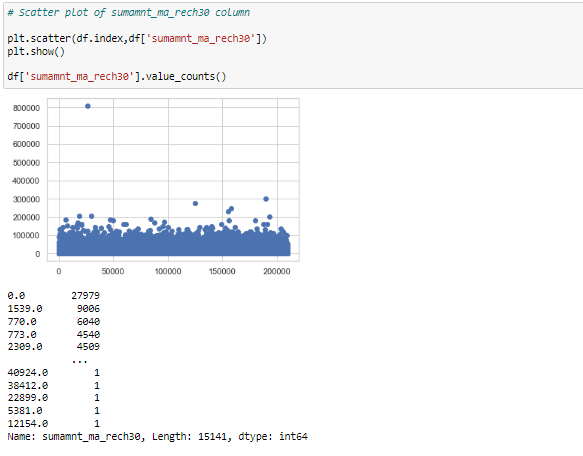


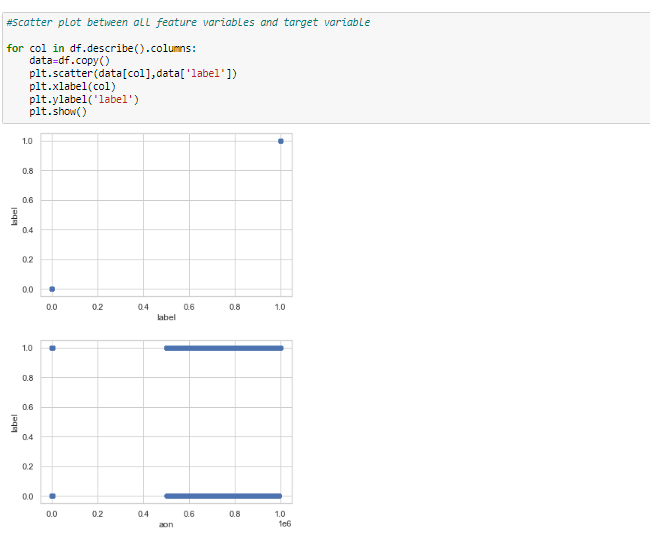


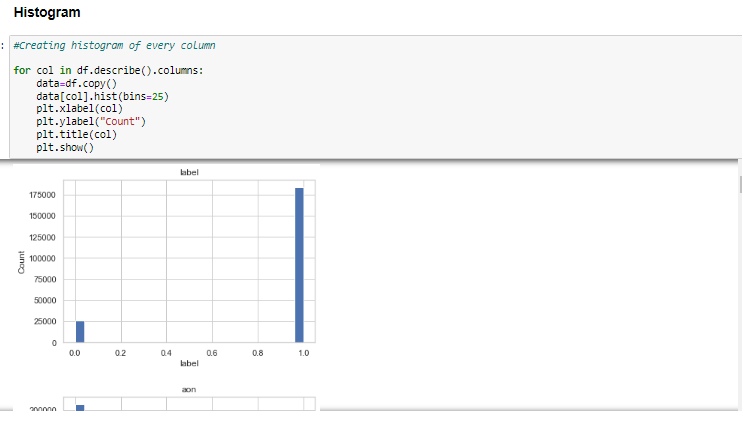












**Observation:**

1. We can observe that there are 183431 number of Non defaulters whereas 26162 number of defaulters.
2. We can observe that this is a very imbalanced data set.
3. Maximum (85764) number of users have taken credit in the 7th month.
4. Maximum (8092) number of users have taken credit on 11th day of the month.
5. Above count plot shows the count of number of loans taken by user in last 30 days
6. Sumamnt\_rech90 and amnt\_loans30 are increasing the number of non-defaulters are also increasing.
7. As cnt\_ma\_rech30 and cnt\_ma\_rech90 are increasing the number non defaulters are also increasing.

* **Interpretation of the Results**

From the visualization we interpreted that the data was very imbalanced and the target variable was highly positively correlated with the columns cnt\_ma\_rech30 and cnt\_ma\_ma\_rech90.

From the preprocessing we interpreted that data was improper scaled, there were hidden features present in the data which needed to be extracted.

From the modeling we interpreted that XGBClassifier works best with respect to our model with rocscore 0.90.

**CONCLUSION**

* **Key Findings and Conclusions of the Study**

Key Findings and Conclusions of the Study In this project we have tried to show how to deal with unbalanced datasets like the MicroCreditDefaulter where the instances of fraudulent cases is few compared to the instances of non-fraudulent cases. We have argued why accuracy is not a appropriate measure of model performance here and used the metric AREA UNDER ROC CURVE to evaluate how method of SmoteTomek technique can lead to better model training. The best score of 0.90 was achieved using the best parameters of XGBClassifier through GridSearchCV though both random forest and gradient boosting models performed well too.

* **Learning Outcomes of the Study in respect of Data Science**

This project has demonstrated the importance of sampling effectively, modelling and predicting data with an imbalanced dataset.

Through different powerful tools of visualization we were able to analyse and interpret different hidden insights about the data.

Through data cleaning we were able to remove unnecessary columns and outliers from our dataset due to which our model would have suffered from overfitting or underfitting.

The few challenges while working on this project were:-

• Improper scaling

• Too many features

• Hidden features

• Imbalanced data

• Skewed data due to outliers

The data was improper scaled so we scaled it to a single scale using sklearns’s package StandardScaler.

* **Limitations of this work and Scope for Future Work**

1. In this data set, first drawback is the data is huge and it is difficult to handle. Because of huge data set, it takes lots of time for Visualization, training the model and in hyper parameter tuning.
2. The data set contains lot of outliers and skewness present in the dataset 0.
3. There are lot of classification algorithm, we have chosen few machines learning algorithm to make the prediction.
4. This is an early stage for making the prediction for Micro Credit Defaulter. The Finance company will make the use of this prediction and yield the high return.