

## **MICRO CREDIT DEFAULTER PROJECT**

Submitted by:

Abhijit Sarkar

**ACKNOWLEDGMENT**

I would like to express my gratitude for the opportunity and would like to thank **Flip Robo Technologies** for giving me an opportunity to work on this Given project. While going through this project I could find the economy facts and a conceptual thinking’s of the consumer behind and consumptions of the customers. I am very grateful to **DATA TRAINED** team for providing me the adequate Trainings which actually helped me a lot to complete this project in the given time. I took the help from Mr. Mohd Kashif where I faced the problem. Moreover, I took the help of Google, Kaggle, Sklearn library and panda’s library for solving this data set.

**INTRODUCTION**

* **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. 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.

* **CONCEPTUAL BACKGROUND OF THE DOMAIN PROBLEM.**

**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.**

* **REVIEW OF LITERATURE.**

**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).**

* **MOTIVATION FOR THE PROBLEM UNDERTAKEN**

**Here We need to build a model which can be used to predict in terms of a probability for each loan transaction, whether the customer will be paying back the loaned amount within 5 days of insurance of loan. In this case.**

**Label [1] indicates that the loan has been paid i.e., non-defaulter.**

**While:**

**Label [0] indicates that the loan has not been paid i.e., defaulter**.

**ANALYTICAL PROBLEM FRAMING.**

* Mathematical/ Analytical Modelling of the Problem

Firstly, I found Columns with No Missing Values, But While checking Unique contents I saw zeros in many columns, I thought of checking the rows if I have only few rows contains zeros but while checking I saw almost maximum rows with zeros, I replaced those zeros with np.nan and found that almost maximum contains are missing. Even I could see 4 to 5 rows having 90 percentage data missing. The Imputation of those rows can make the prediction model a complete bias. Finally, the columns were dropped. To get better insight on the features I have used plotting like distribution plot, bar plot, reg plot and cat plot, strip plot, count plot as well. With these plotting I was able to understand the relation between the features in better manner. Also, I found outliers and skewness in the dataset so I removed outliers using Zscore and I removed skewness using yeo-Johnson method. I even used Standard Scaler to bring the data under one scale. PCA didn’t work well so I removed that, I have used all the Logistic regression models and other classification models while building model then turned the best model and saved the best model.

* Data Sources and their formats

The data is being collected from my internship company” Flip Robo” and the dataset is in csv format. The Data Set contains 36 columns and 209593 rows.

The size of the data is 7545348.

Columns contains

* Float Values - 21 Columns
* Integer Values - 12 Columns
* Object Values - 3 Columns
* Memory Used: 58 MB
* **Data Pre-processing Don.**

These Steps:

1. Loading Data Set, Locating Null Values

2. Finding the unique values in categorical and numerical columns.

3. Finding Data Missing percentage

4. Finding nan values and replacing nan values.

5. Finding duplicated values in columns rows.

6. Use encoding was not required.

7. Feature Extraction was done here.

* Data Inputs- Logic- Output Relationships

X variables plays a very import role in machine learning for the Prediction of Target variable. Here ‘LABEL ‘is the target variable on which the predictions are being made.

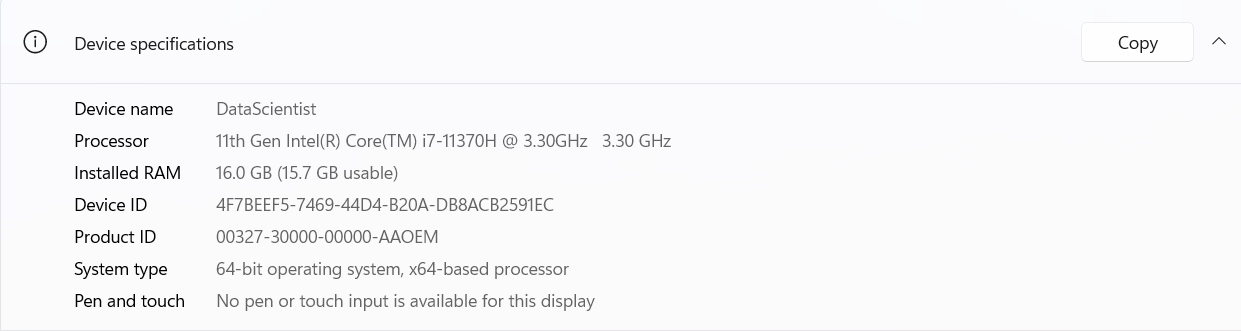
**I used the following to determine the relationship between variable:**

* I have used Catplot for each pair of categorical features that shows the relation with the Target Variable. Used (Box plot and Distplot to determine the relations) And also, for continuous numerical variables I have used lmplot, scatterplot to show the relationship between a continuous numerical variable and target variable.
* Used univariate, Bivariate Graph to check for relations.

**By the Use of these Graph, I Uncovered the is a relationship between continuous numerical variable and The Target Variable.**

* Hardware and Software Requirements and Tools Used

Hardware:

****

Software Used:

* I Used Jupiter Note Book.
* Microsoft Office 2020
* Windows 11 OS

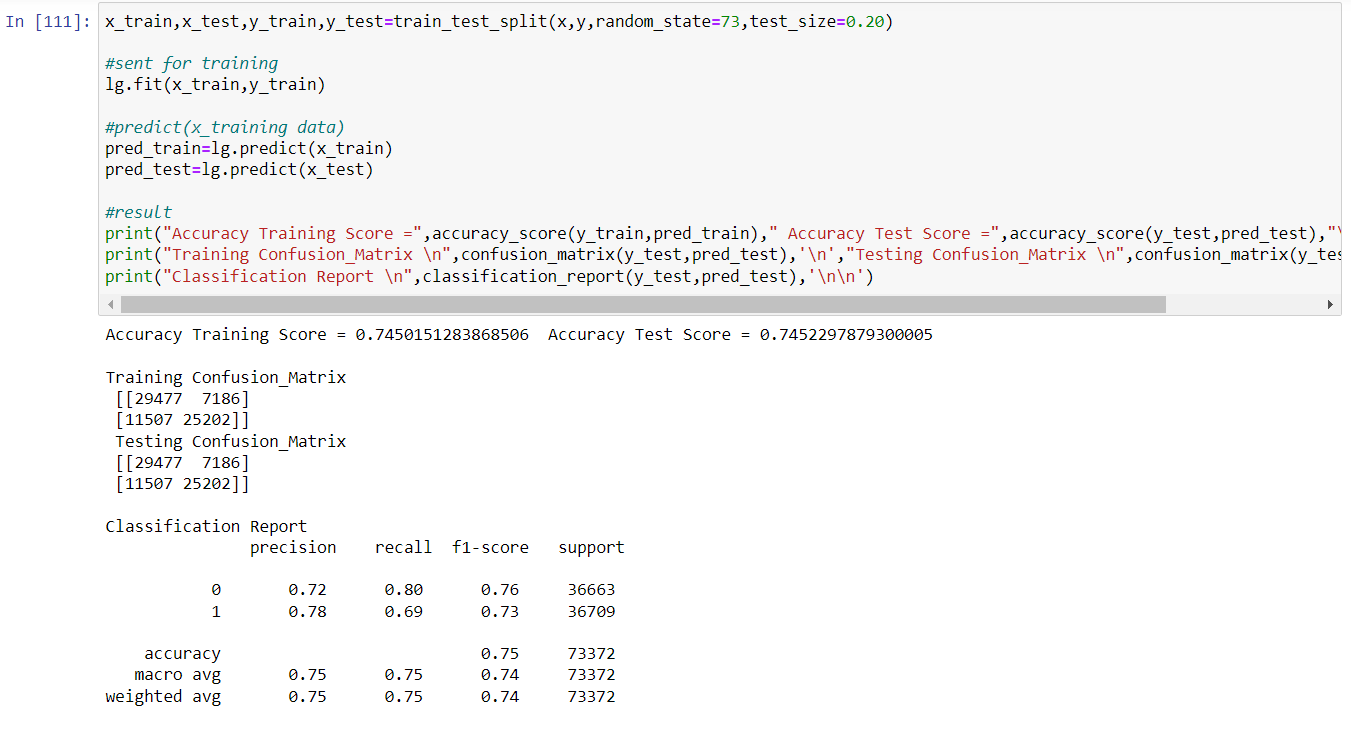
Library used:

1. NumPy
2. Pandas
3. Seaborn
4. Matplotlib
5. SciPy
6. Sklearn
7. Pickle
8. Imbalance Learn

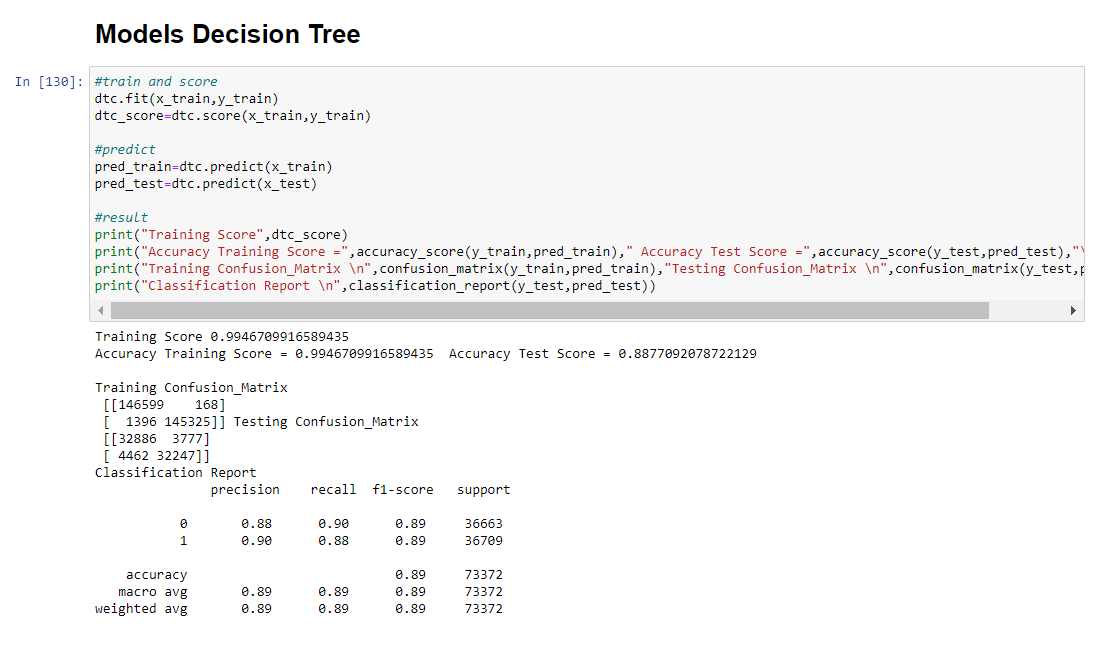
**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)
* I have used the simple imputation method to replace Null values.
* To check outliers, I used boxplot.
* To remove outliers, I have used Zscore.
* To check skewness, I used distplot.
* I remove skewness I have used the yeo-johnson method.
* Use of Pearson’s correlation coefficient to check the correlation between dependent and independent features.
* Also, I have used standardization.
* Then followed by model building with all Classifiers and Logistic regression algorithms.
* Testing of Identified Approaches (Algorithms)
* Logistic Regression
* Decision Tree Classifier
* Extra Trees Classifier
* Random Forest Classifier
* Gradient Boosting Classifier
* Run and evaluate selected models.

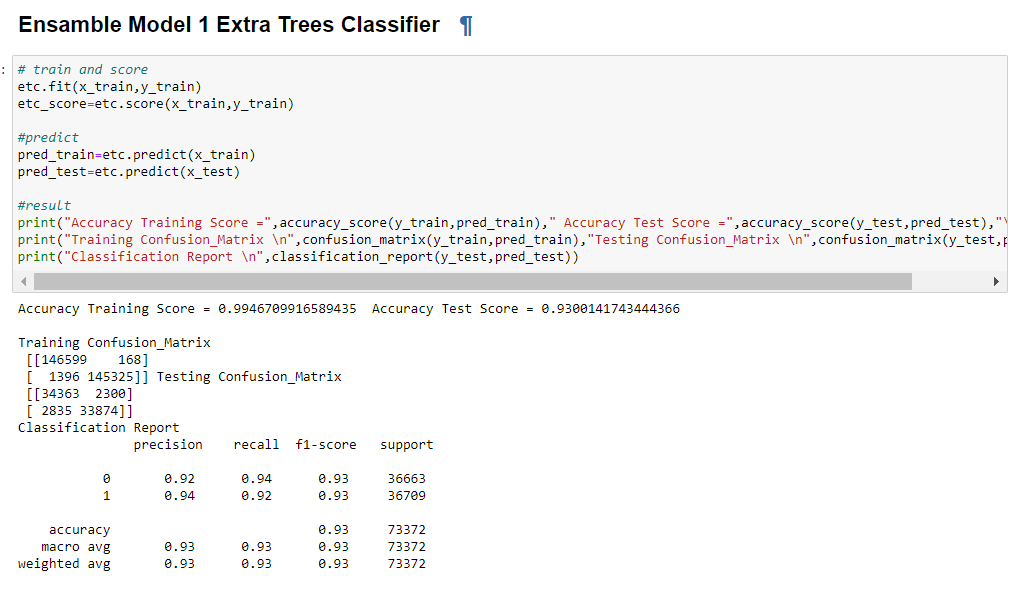
**Model-Logistic Regression**



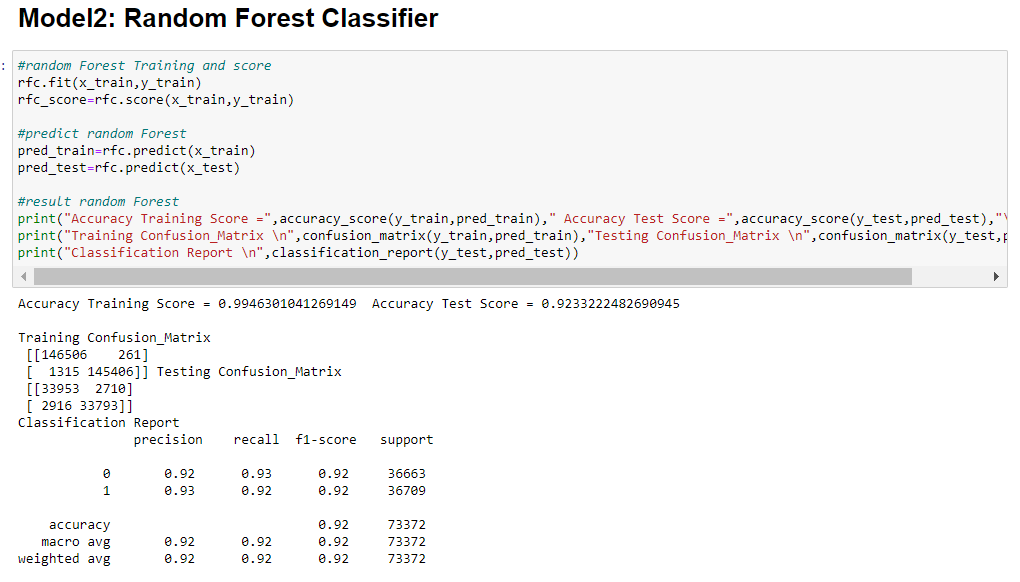
**Model-1**



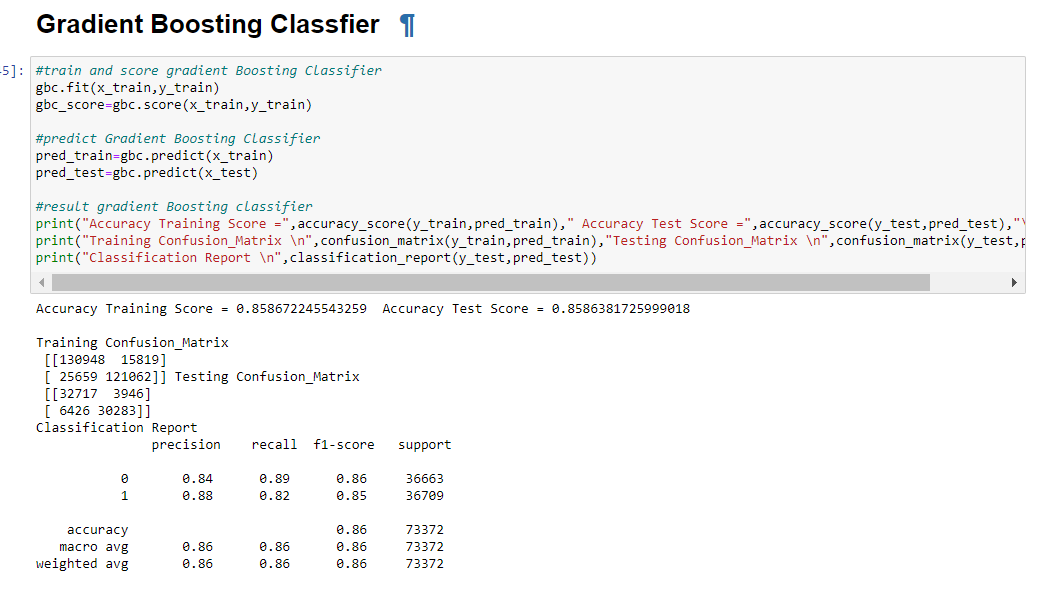
Model-2



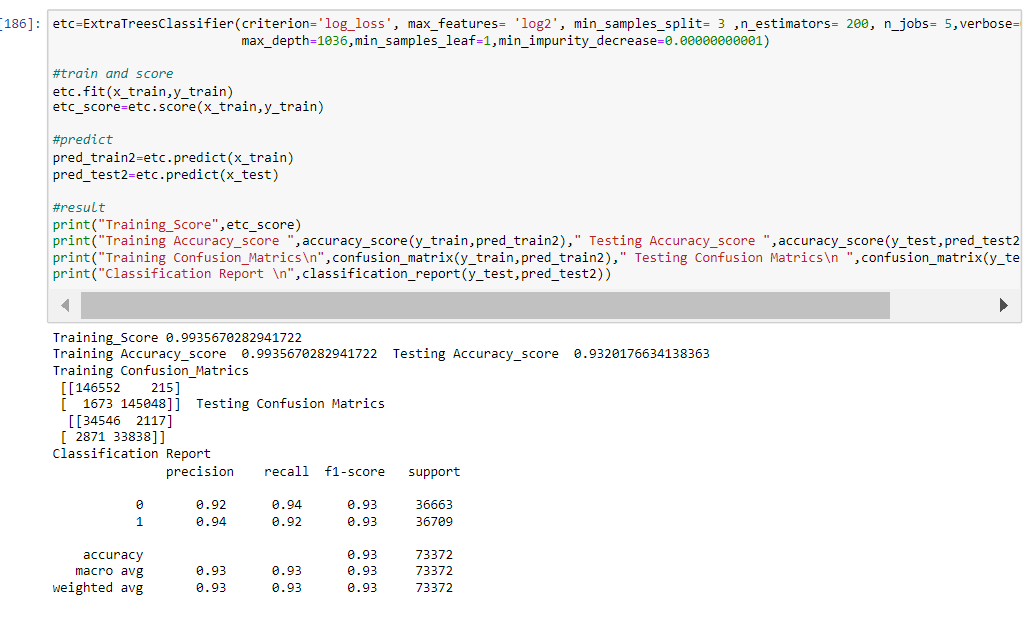
Model-3



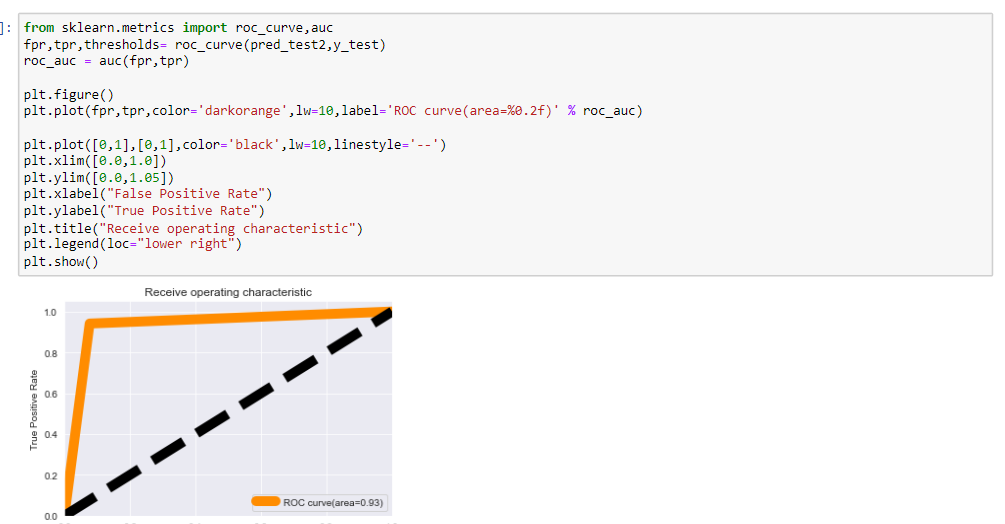
Model-4



Hyper-Parameter on Extra Trees Classifier



Graph:



Saving Model and Loading Model:

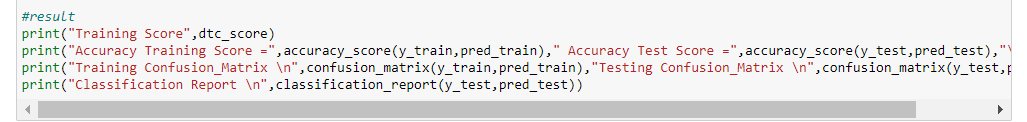


* Key Metrics for success in solving problem under consideration.



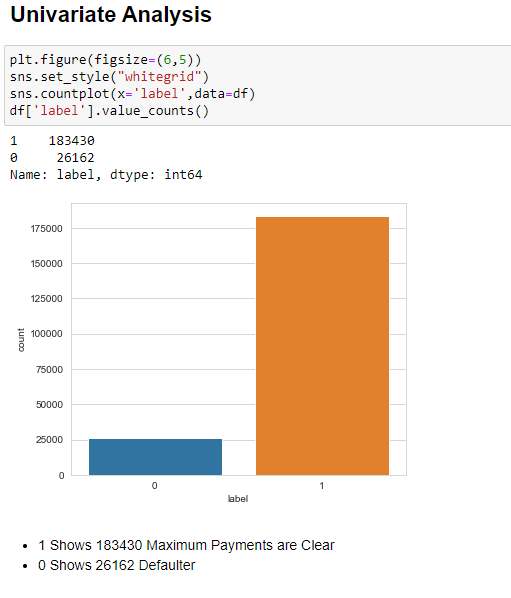
The Following Metrics were Used:

1. Accuracy score
2. Classification report
3. Confusion Matrix
4. F1 score

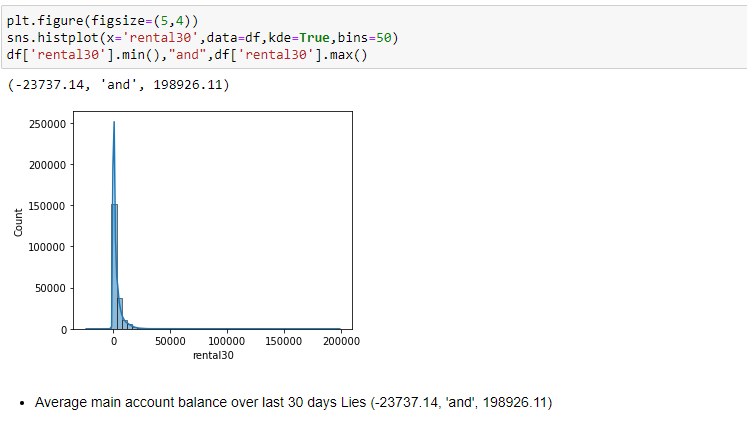


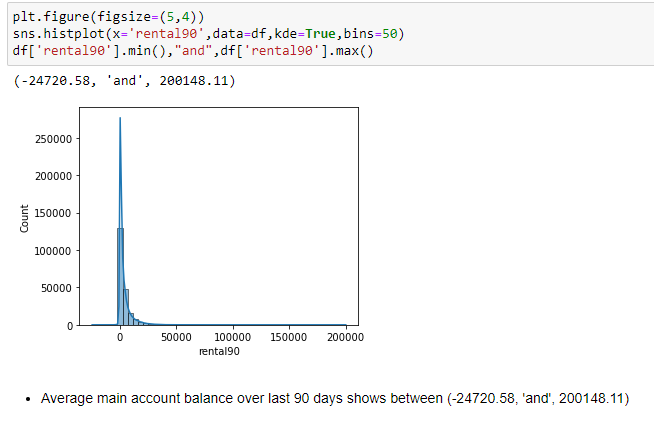
* Visualizations

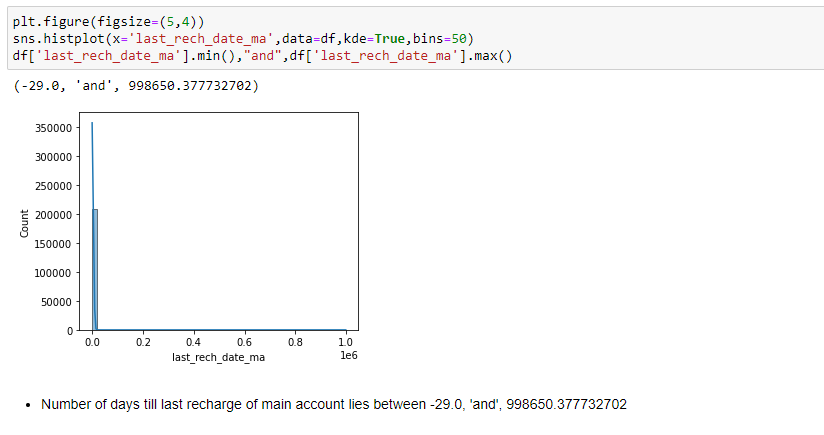
Shows counts of observation for Target Variable.

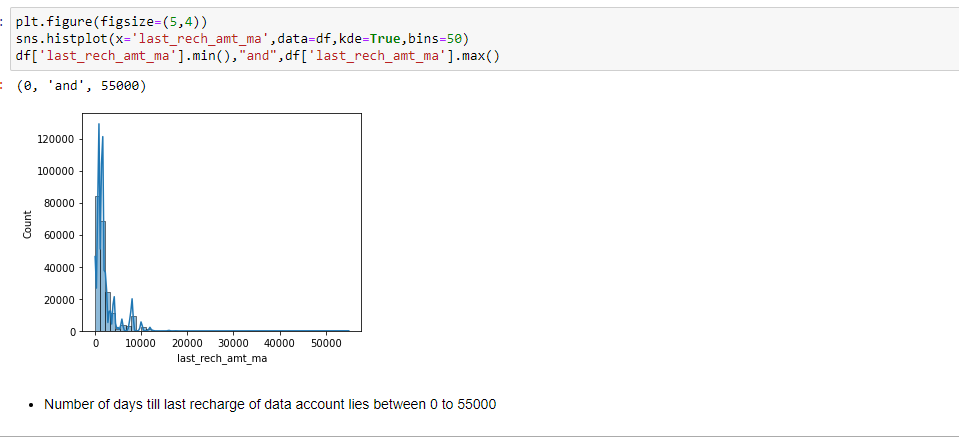


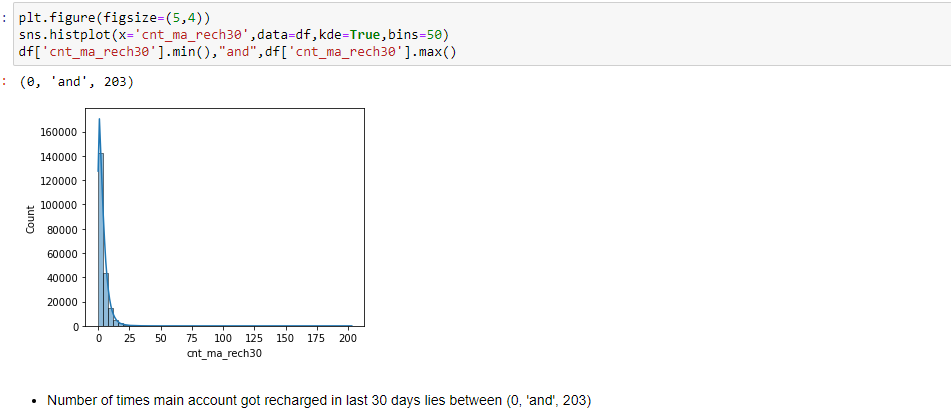
Distribution:

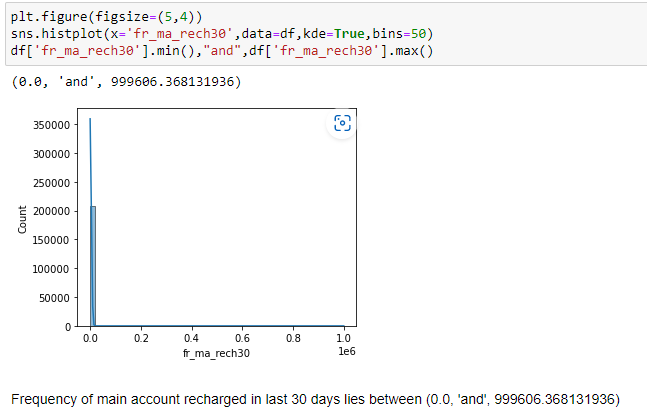


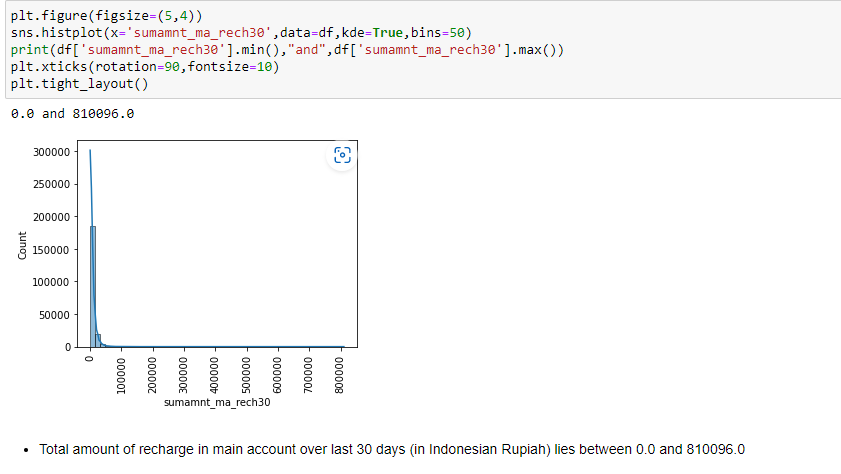


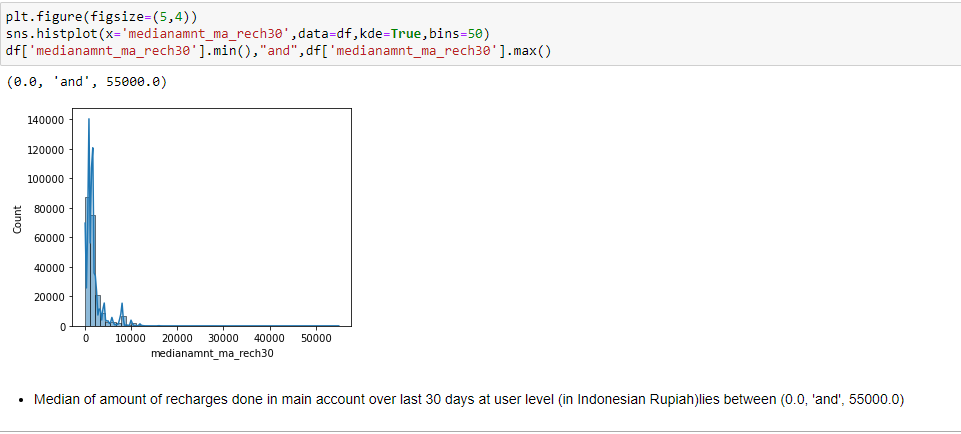


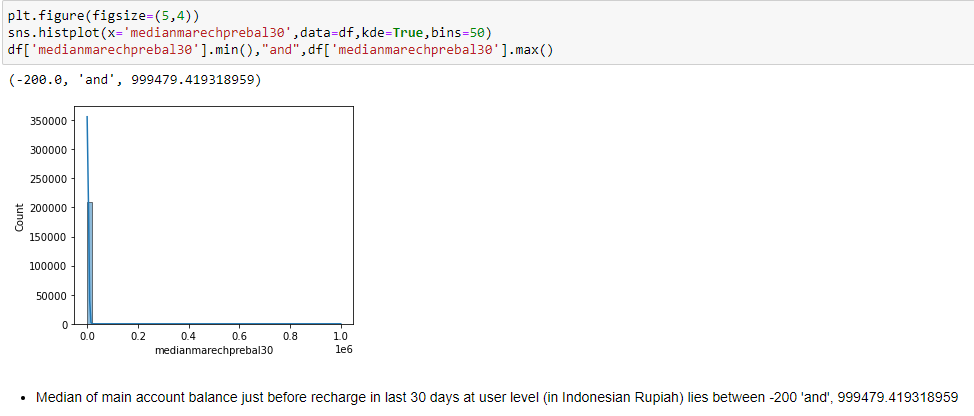


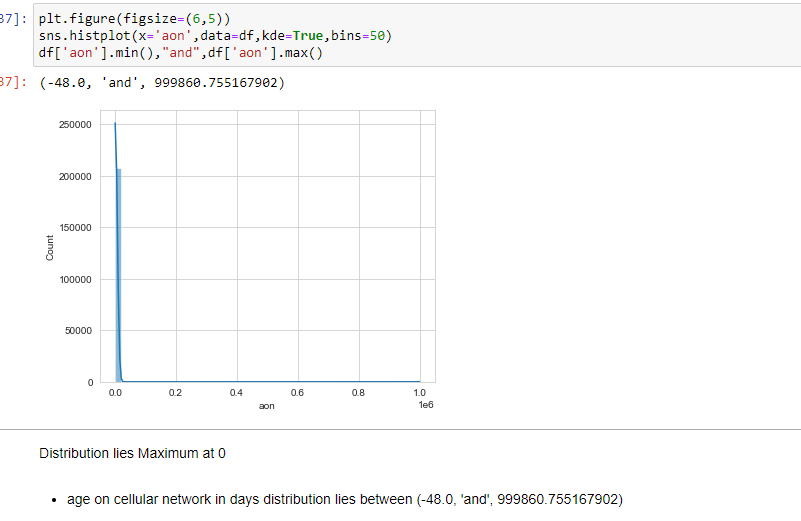




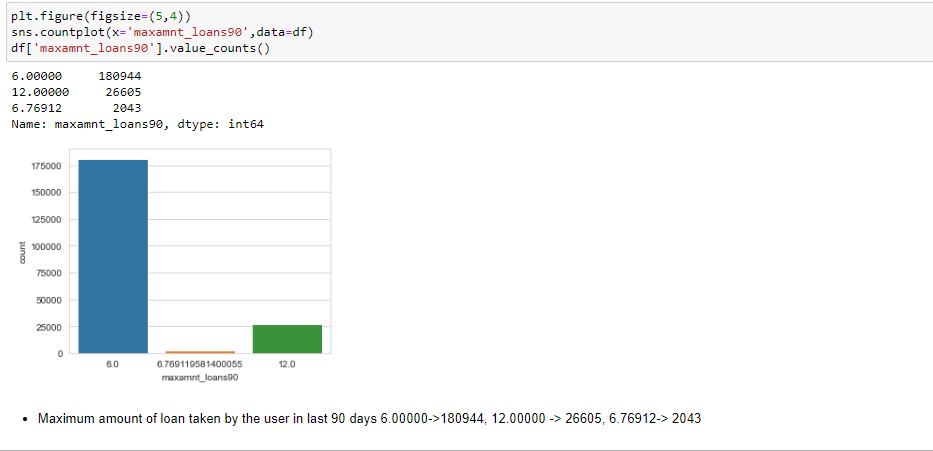


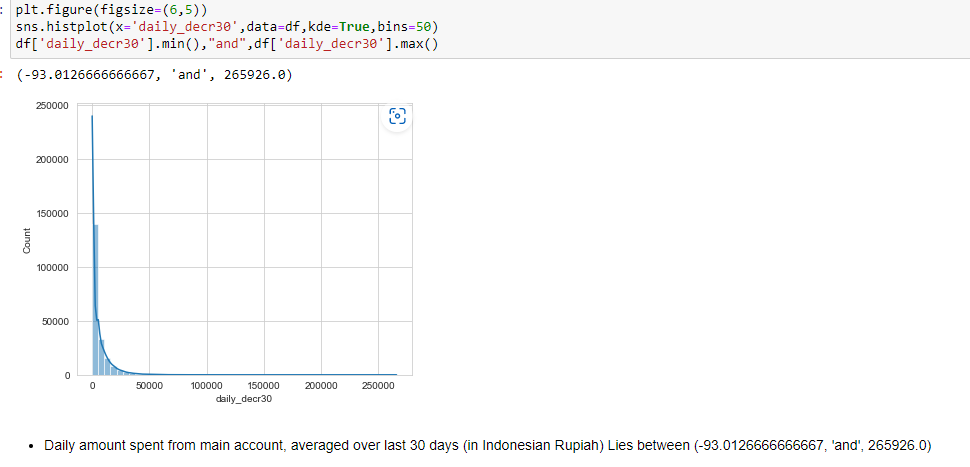


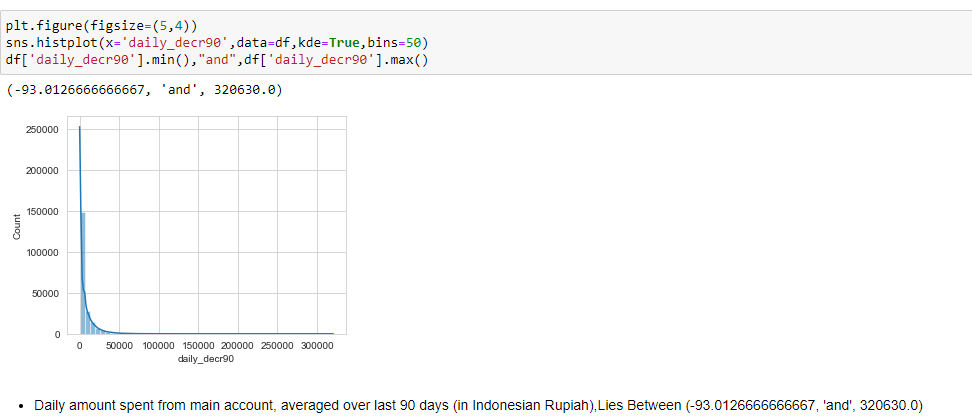


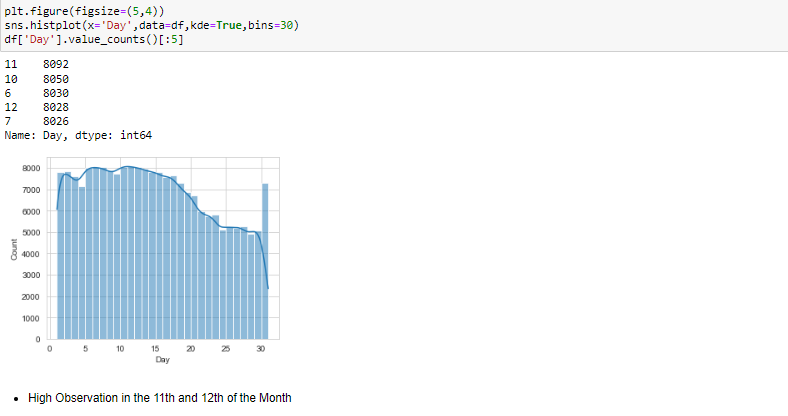


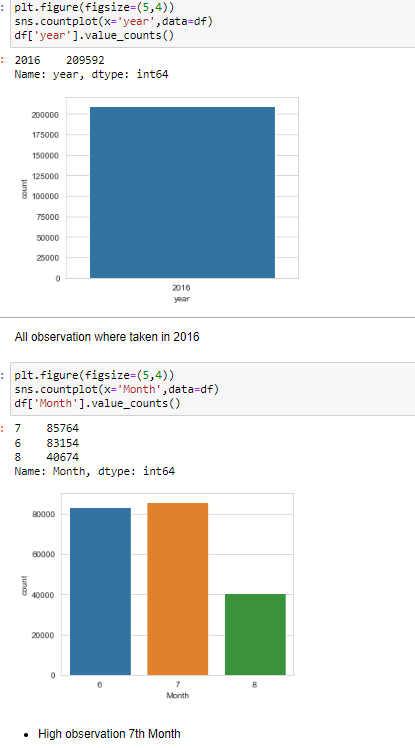


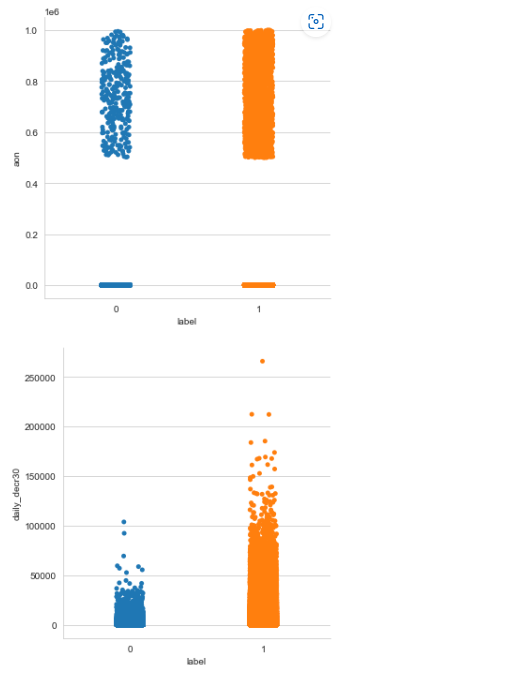


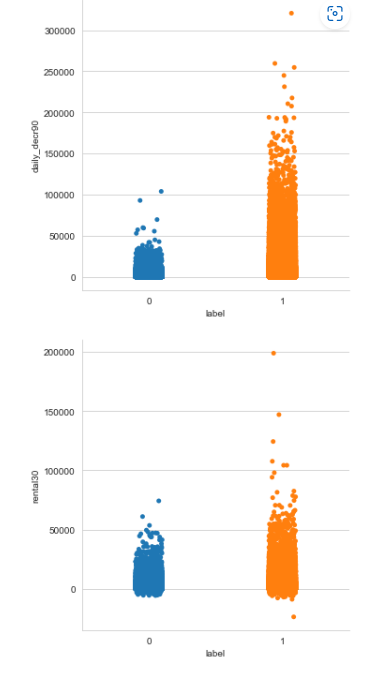


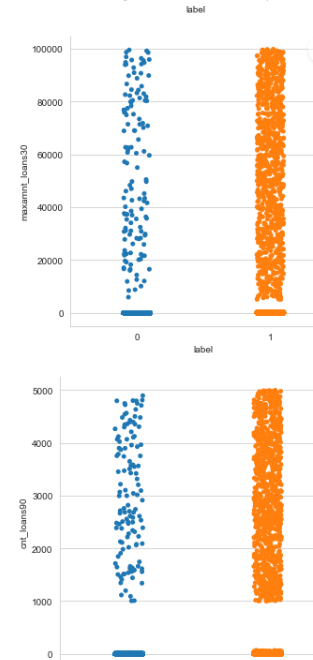


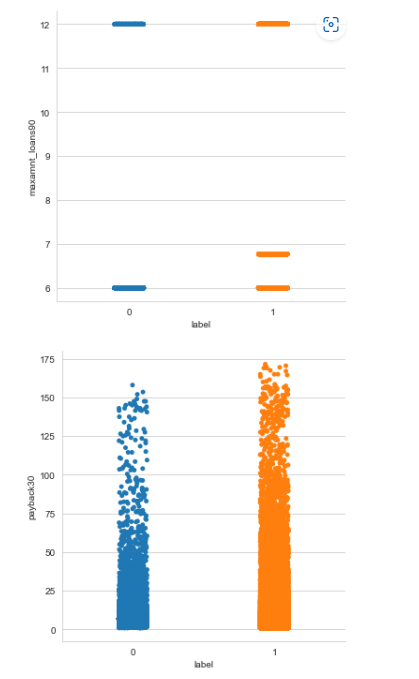
****

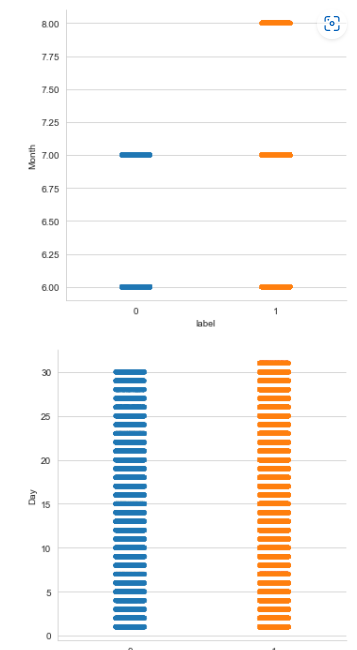


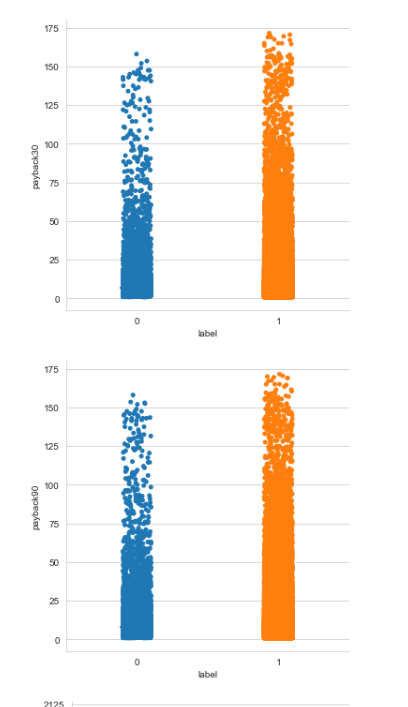




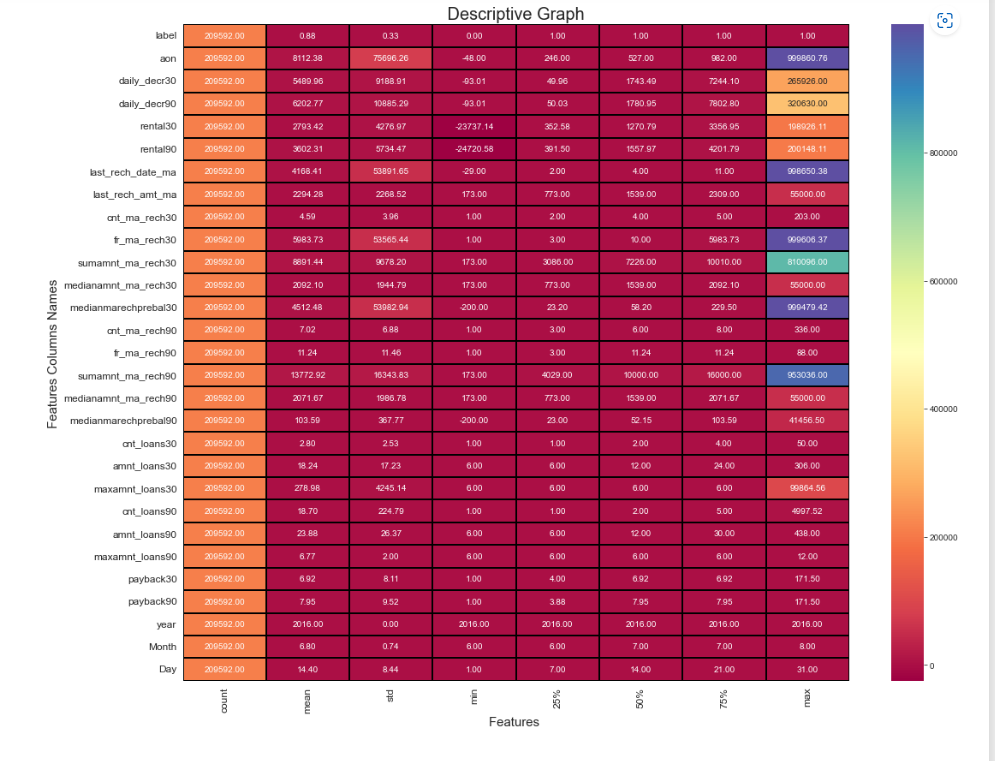








Descriptive Graph:



# Few Observations:

1. Null Values- All values are intact (No Null Values)

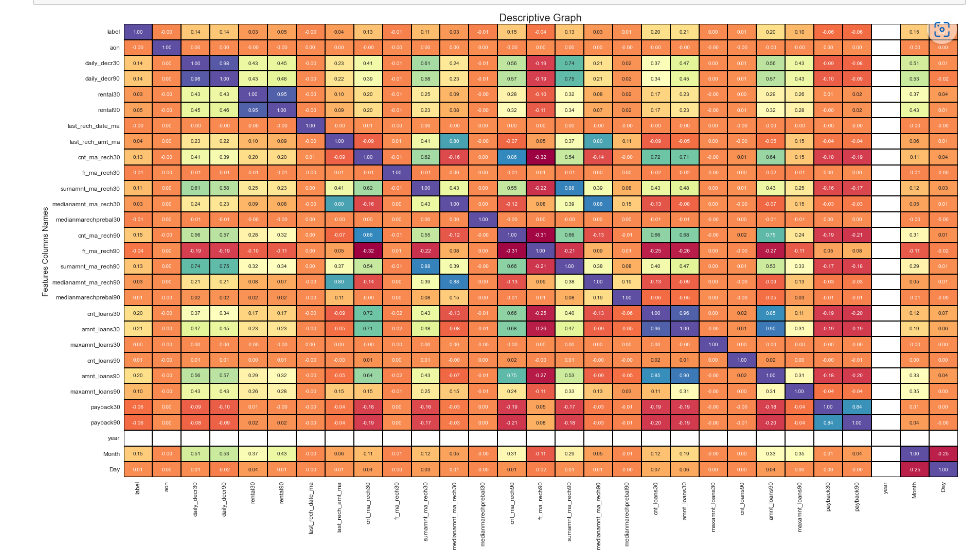
2. Right Skew- aon, daily\_decr30, daily\_decr90, rental30, rental90, last\_rech\_date\_ma, last\_rech\_amt\_ma, cnt\_ma\_rech30 fr\_ma\_rech30 sumamnt\_ma\_rech30 medianamnt\_ma\_rech30 medianmarechprebal30 cnt\_ma\_rech90 fr\_ma\_rech90 sumamnt\_ma\_rech90 medianamnt\_ma\_rech90 medianmarechprebal90 cnt\_loans30 amnt\_loans30 maxamnt\_loans30 cnt\_loans90 amnt\_loans90 maxamnt\_loans90 payback30 payback90 year Month Day

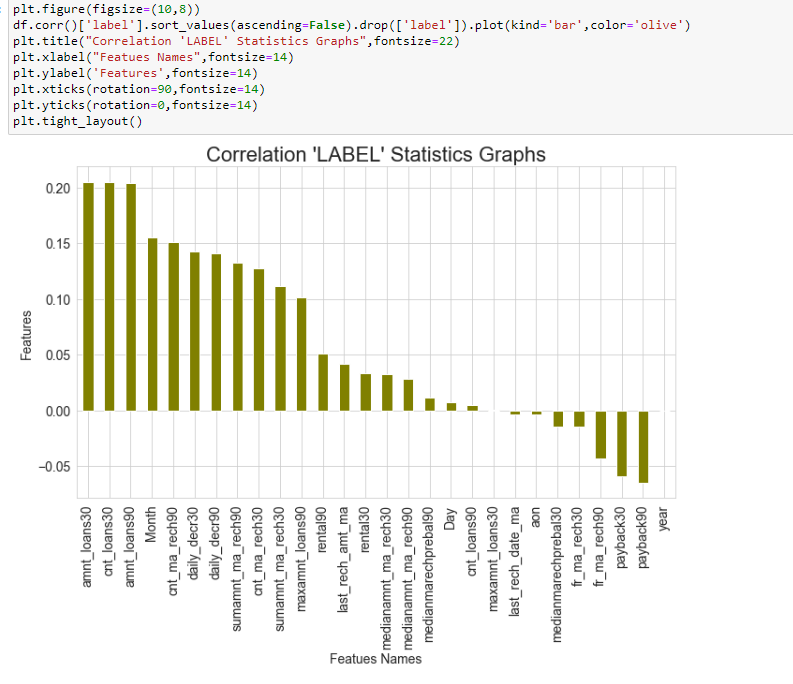
3. Left Skew- Null

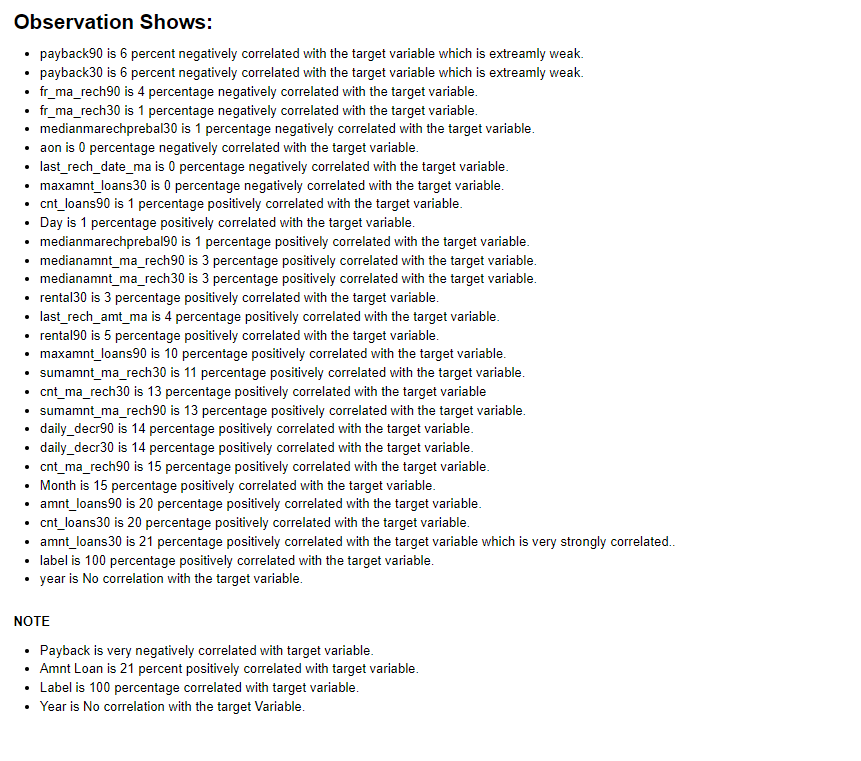
4. Standard Deviation-aon daily\_decr30 daily\_decr90 rental30 rental90 last\_rech\_date\_ma last\_rech\_amt\_ma cnt\_ma\_rech30 fr\_ma\_rech30 sumamnt\_ma\_rech30 medianamnt\_ma\_rech30 medianmarechprebal30

5. Outliers- aon, cnt\_ma\_rech30 fr\_ma\_rech30 sumamnt\_ma\_rech30 medianamnt\_ma\_rech30 medianmarechprebal30 cnt\_ma\_rech90 fr\_ma\_rech90 sumamnt\_ma\_rech90 medianamnt\_ma\_rech90 medianmarechprebal90 cnt\_loans30 amnt\_loans30 maxamnt\_loans30 cnt\_loans90 amnt\_loans90 maxamnt\_loans90 payback30 payback90 year Month Day

Correlation

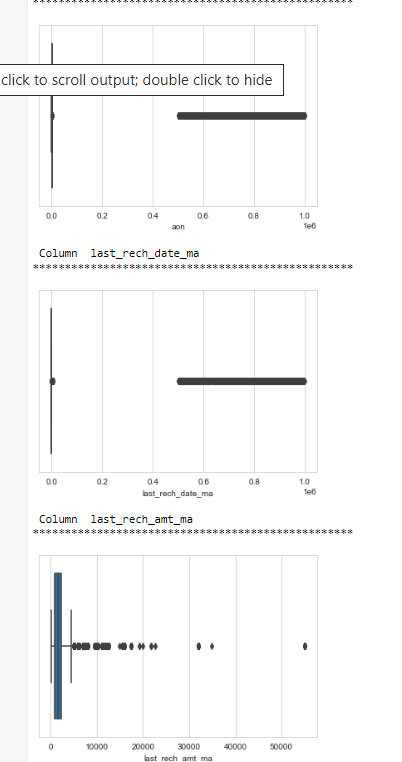


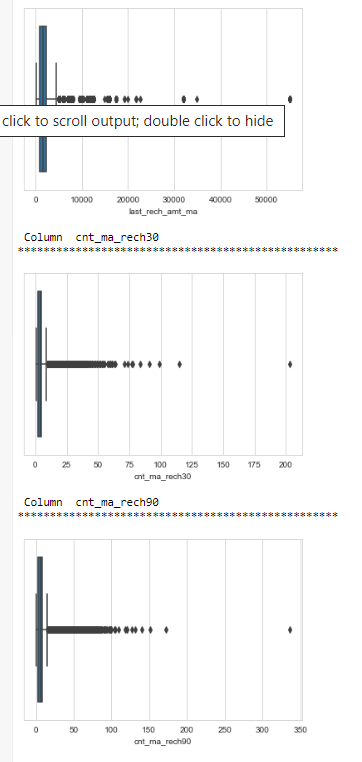


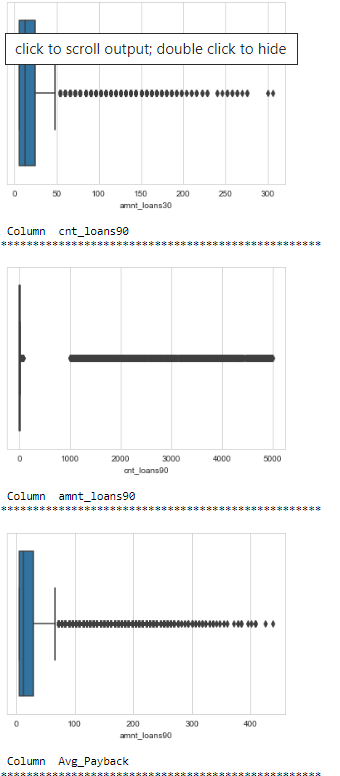


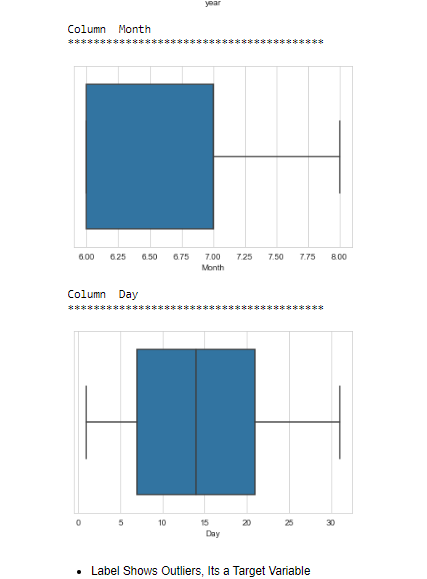
Outliers:

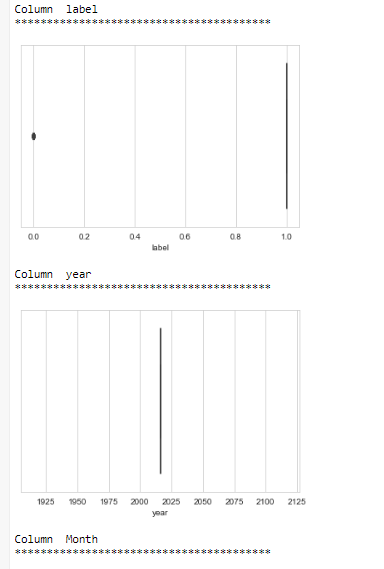
ALL float columns contains outliers:



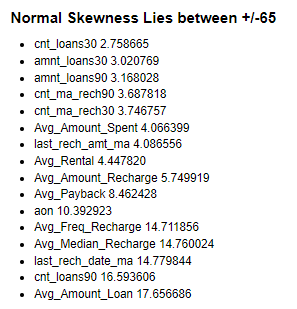


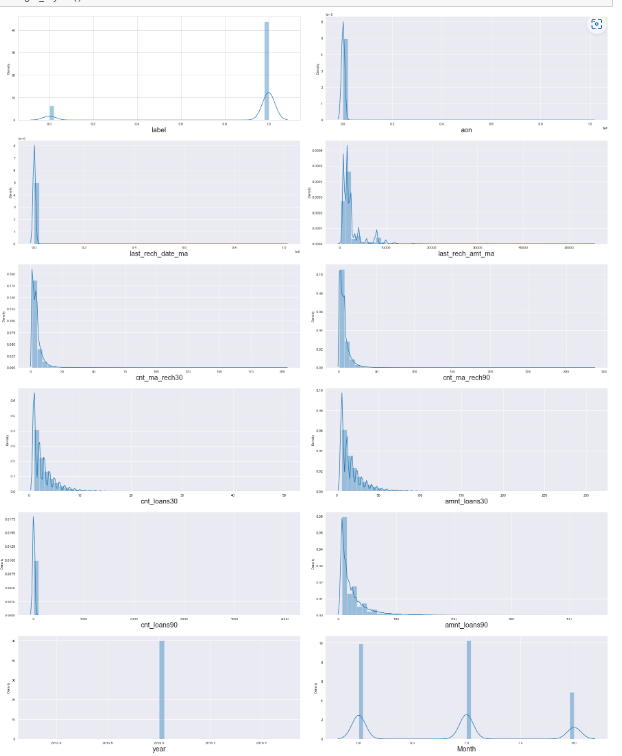


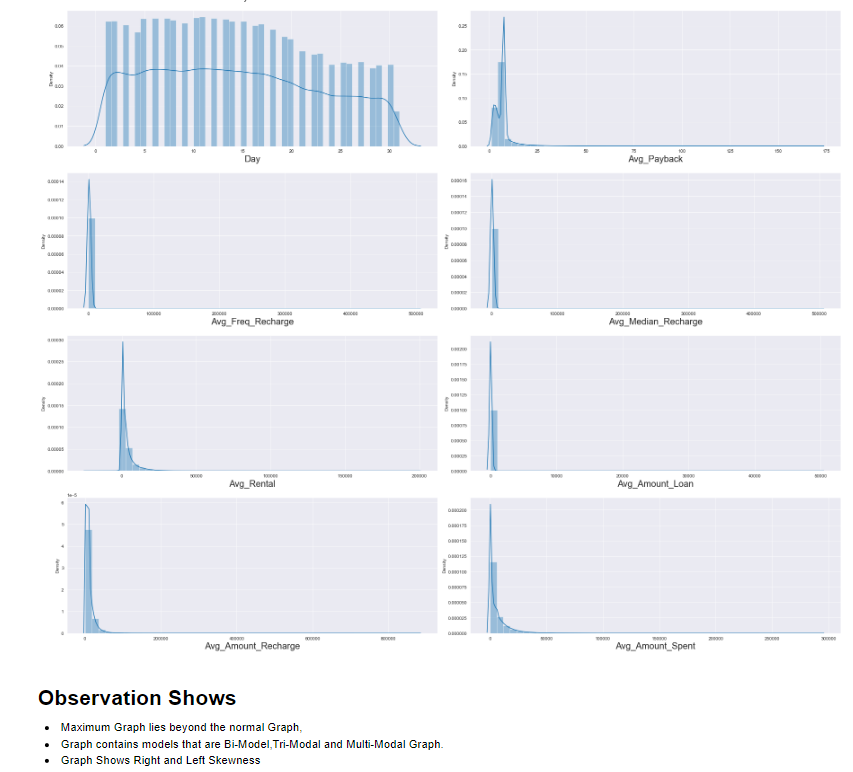




Skewness in these columns







* **Interpretation of the Results**
* This dataset was very special as it had a separate train and test datasets.
* Firstly, the datasets were having no null values but full of null zeros values which is firstly converted to null values. Then Imputation is done of that dataset. entries maximum columns so we have to be careful while going through the statistical analysis of the datasets.
* I found maximum numerical continuous columns were in relationship with target column.
* I notice a huge number of outliers and skewness in the data so we have chosen proper methods to deal with the outliers and skewness. If we ignore this outlier and skewness, we may end up with a bad model which has less accuracy.
* Then scaling both train and test dataset has a good impact like it will help the model not to get biased.
* We have to use multiple models while building model using train dataset as to get the best model out of it.
* Extra Trees and Random Forest were the best among all the models. Result received with both model is approx. **above 92 percentage.**
* Finally selected Extra Trees Classifier as CV\_Score has better result and result was quite noticeable.

**CONCLUSION**

* Key Findings and Conclusions of the Study

In this project report, we have used machine learning algorithms to predict in terms of a probability for each loan transaction, whether the customer will be paying back the loaned amount within 5 days of insurance of loan. We have mentioned the step-by-step procedure to analyse the dataset and finding the correlation Between the features. Thus, we can select the features which are not correlated to each other and are independent in nature. These feature set were then given as an input to five algorithms to predict output. Hence, we calculated the performance of each model using different performance metrics and compared them based on these metrics.

And saved the model as filename="micro\_fin.pkl"

* Learning Outcomes of the Study in respect of Data Science

I found that the dataset was quite interesting to handle as it contains all types of data in it. Improvement in computing technology has made it possible to examine social information that cannot previously be captured, processed and analysed. New analytical techniques of machine learning can be used in property research. The power of visualization has helped us in understanding the data by graphical representation it has made me to understand what data is trying to say. Data cleaning is one of the most important steps to remove missing value and to replace null value and zero values with their respective mean, median or mode. This study is an exploratory attempt to use machine learning algorithms in finding the probability for each loan transaction against each customer. To conclude the application of Machine Learning in prediction is still at an early stage. I hope this study has moved a small step ahead in providing solution to the companies.

The changes I faced was when I saw a lot of zero and I was thinking to impute those zero but as the missing rows percentage were too high, so finally I had to drop those columns. Another Issue was while using binning process taking the median values helped me to achieve a good accuracy. I was actually thinking to get best out of those. However, I finally achieved a good model out of this.

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

This model doesn't predict future probability. The future will be unpredictable at all times due to this, the risk in investment in an import factor. This can predict for a time period and needs to be updated on a basis. Machine can predict when loan can be given but it can’t predict the intensity of the person who is grabbing the loan.

