

MICRO CREDIT DEFAULTER PROJECT

Submitted by:

Asha Vinod Kumar

# ACKNOWLEDGMENT

Firstly, I would like to thank FlipRobo Technologies for giving me the opportunity to work on this project. Also, I would like to thank the DataTrained team, for providing me the knowledge and guidance which helped me a lot to work on this project.

# INTRODUCTION

* Business Problem Framing

The main objective of this project is 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.

* Conceptual Background of the Domain Problem

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.

FlipRobo is 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

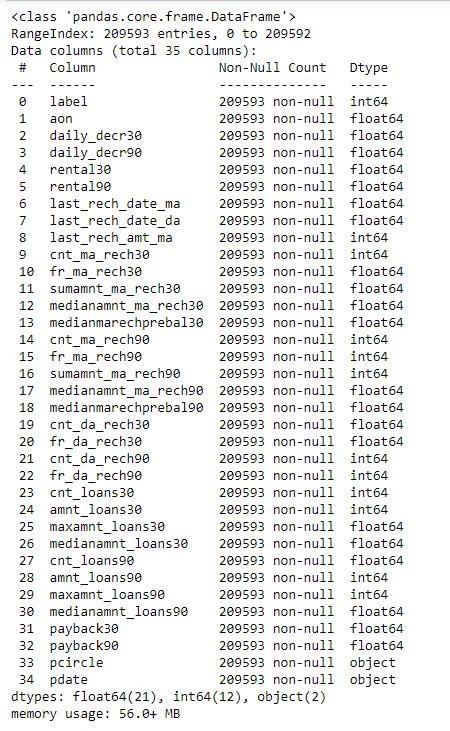
(in Indonesian Rupiah), while, for the loan amount of 10 (in Indonesian Rupiah), the payback amount should be 12 (in Indonesian Rupiah).

* Motivation for the Problem Undertaken
  1. The objective behind to take this project is to harness the required data science skills.
  2. Improve the analytical thinking.
  3. Get into the real-world problem-solving mechanics.

# Analytical Problem Framing

* Data Sources and their formats

The sample data is provided to us from FlipRobo client database. 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. The summary of the dataset are as follows:



* Data Pre-processing Done

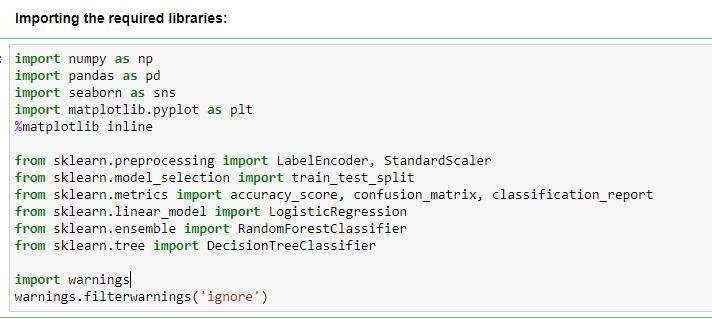
Below are the steps which we have taken in data pre - processing:

* + - Null Values:

We checked for the null values (missing values) and found that there is no null values in the given dataset.

* + - Data Cleaning:
    1. Dropped ‘Unnamed:0’ column as it was not contributing to the dataset.
    2. Dropped 'msisdn' as it’ll not help in the model building.
    3. Split the ‘pdate’ column into day, month, and year and dropped the ‘pdate’ column.
    4. Dropped ‘year’ column as it only contains 2016 as value.
    5. Dropped ‘pcircle’ column as it contains single value (UPW).
* Data Inputs- Logic- Output Relationships

EDA was performed by creating valuable insights using various visualization libraries.



* Hardware and Software Requirements and Tools Used

**Hardware Configuration:**

**Operating System:** Windows 10

**System Type:** 64-bit operating system, x64-based processor

**Processor:** I have used i3 processor with 4GB RAM as hardware.

**Software & Tools:**

* + 1. Jupyter Notebook (used as a notebook to code)
    2. Python (used for scientific computation)
    3. Pandas (used for scientific computation)
    4. Numpy (used for scientific computation)
    5. Matplotlib (used for visualization)
    6. Seaborn (used for visualization)
    7. Scikit-learn (used as algorithmic libraries)

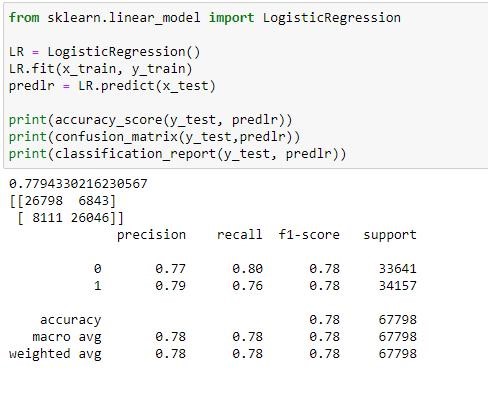
**Model/s Development and Evaluation**

Identification of possible problem-solving approaches (methods)

* + - Performed EDA (Exploratory Data Analysis).
    - Data Cleaning and dropping the columns which were not contributing to the dataset.
    - Checked for the outliers and tried to remove the outliers of the dataset.
    - Checked for the skewness in the dataset and removed the skewness for better model building.
    - Train- Test the dataset into independent and dependent variables.
    - Model Building.
    - Cross validation score to check if the model is over-fitted.
* Testing of Identified Approaches (Algorithms)

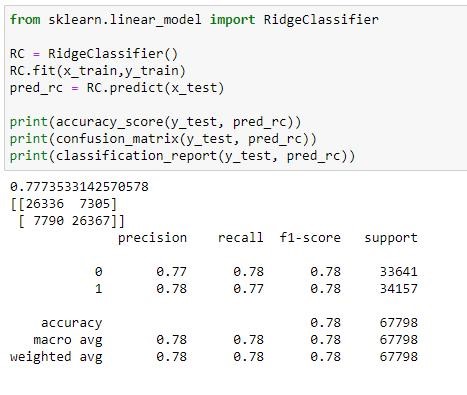
Below are the algorithms used for the training and testing:

* + 1. Logistic Regression.
    2. Ridge Classifier.
    3. Random Forest Classifier.
    4. Decision Tree Classifier.
    5. Gaussian NB.
* Run and Evaluate selected models
  1. **Logistic Regression:**



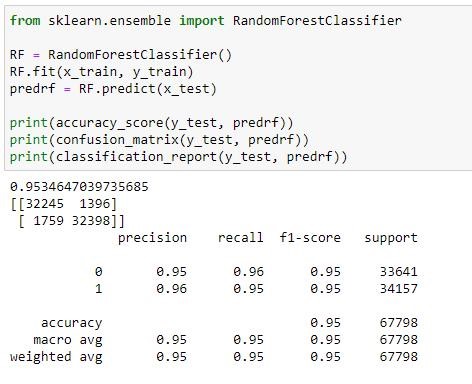
From Logistic Regression we got 78% accuracy score.

**2.Ridge Classifier:**



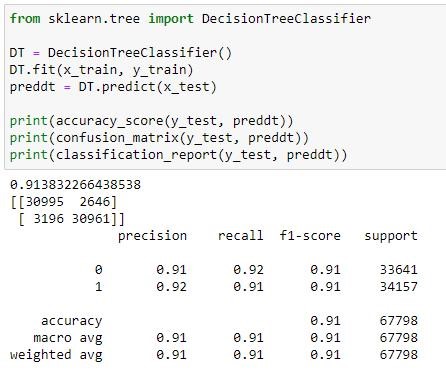
From Ridge Classifier we got 78% accuracy score.

**3.Random Forest Classifier:**



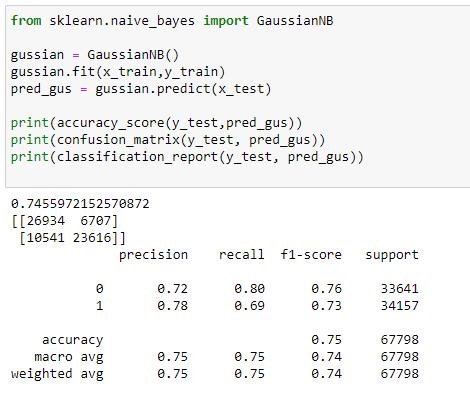
From Random Forest Classifier we got 95% accuracy score.

* 1. **Decision Tree Classifier:**



From Decision Tree Classifier we got 91% accuracy score.

**3.Gaussian NB:**



From Gaussian NB we got 75% accuracy score.

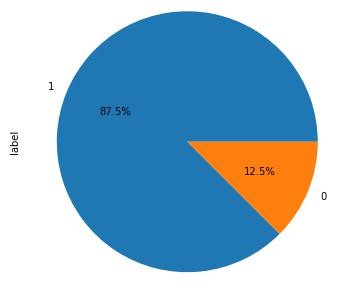
* Key Metrics for success in solving problem under consideration

The key metrics used are as follows:

* + 1. Accuracy Score
    2. Confusion Matrix
    3. Classification Report
    4. F1 Score
    5. Precision & Recall
    6. Cross validation score

Visualizations

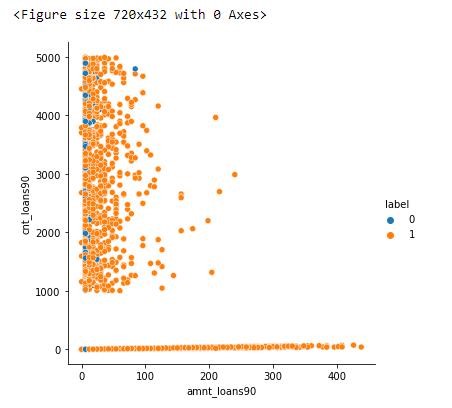
* + Checked if the data is balanced or not.



Label ‘1’ indicates Non- defaulters & label ‘0’ indicates defaulters.

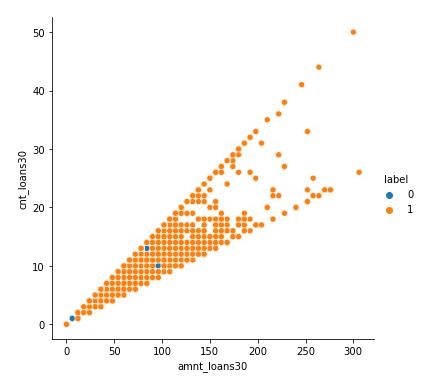
87.5% are non- defaulters and 12.5% are defaulters. This shows that the dataset is imbalance.

* + The number of defaulters are more for 90 days but the loan amount is below 100.

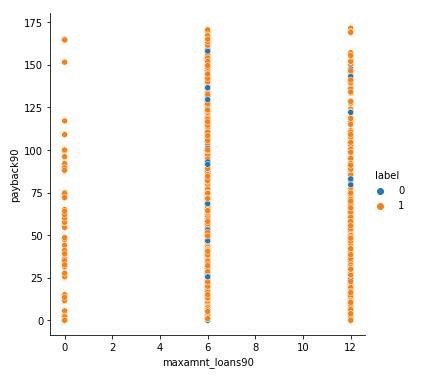
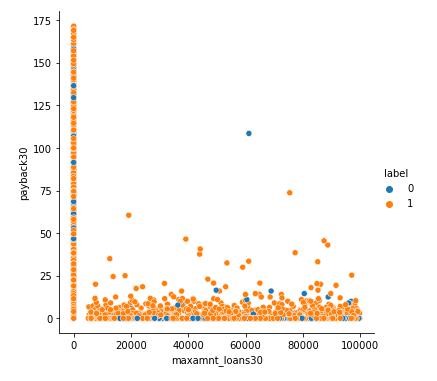


* + The number of loans taken by users in last 30 days is more than

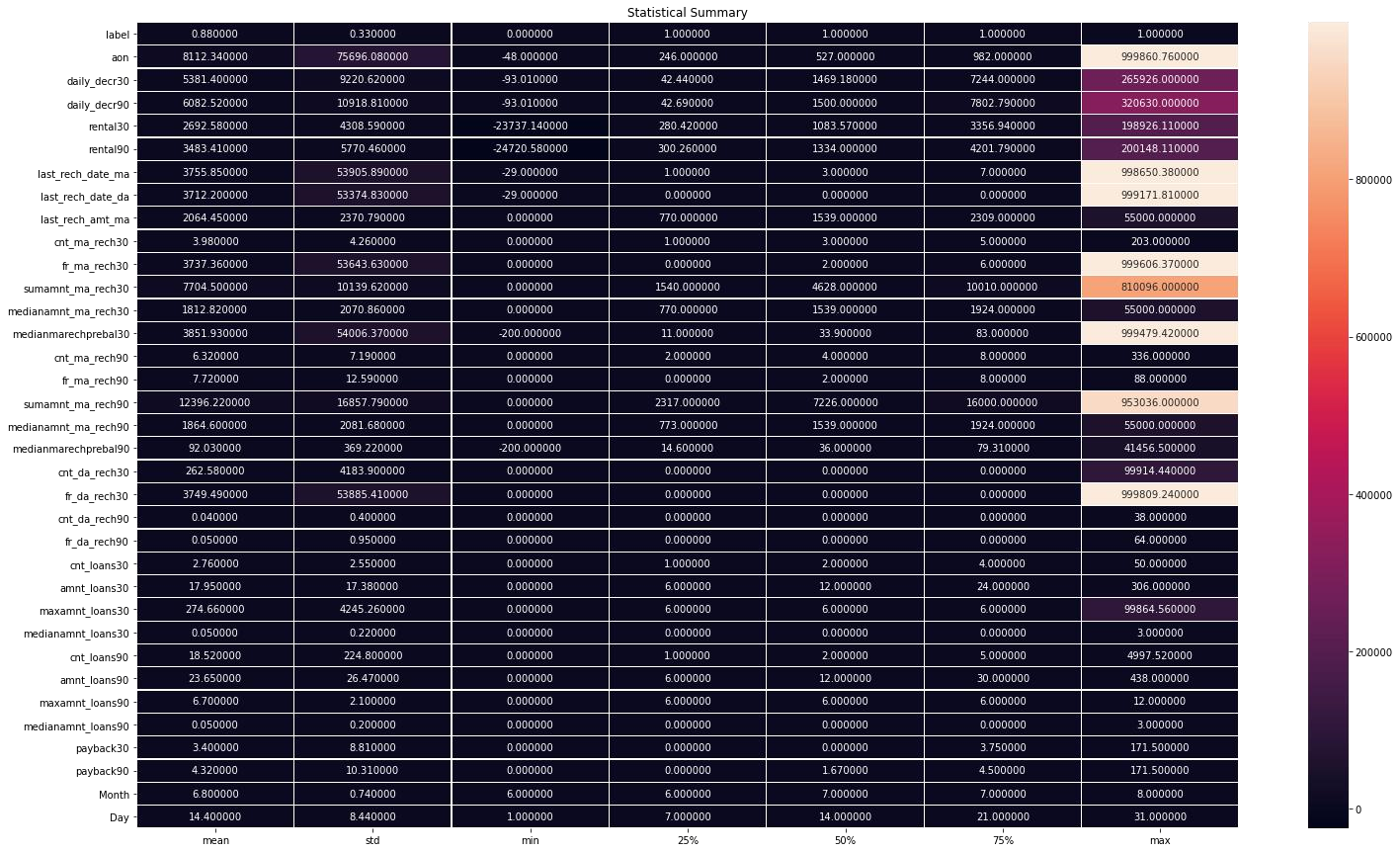
50 but the maximum loan amount taken ranges from 50 to 150.



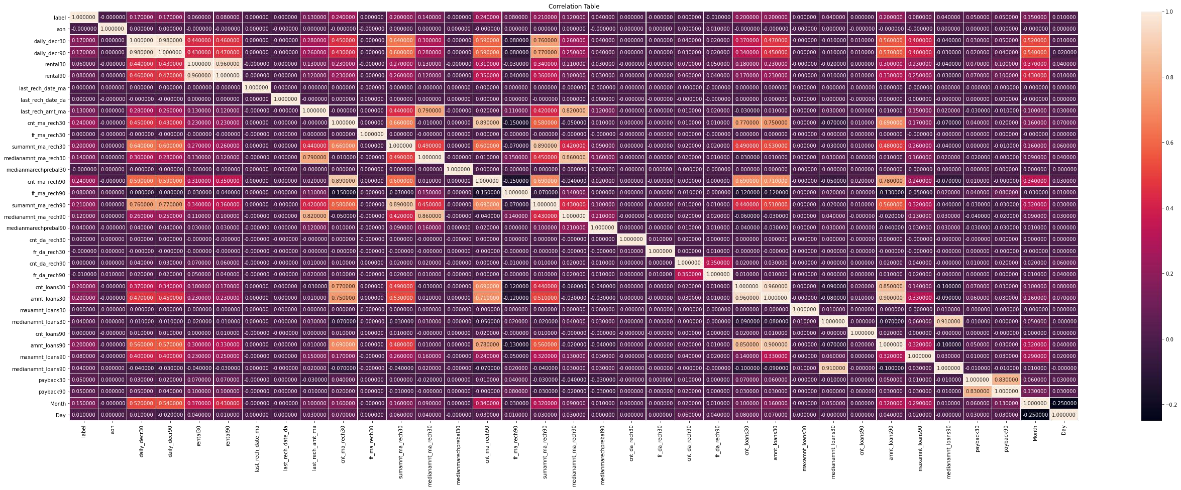
* As the number of days of payback is increasing the number of defaulters are also increasing.



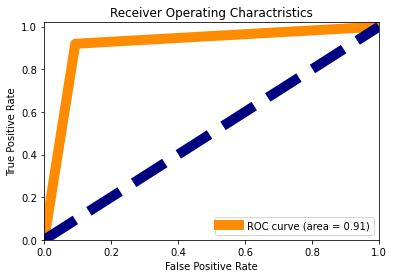
* Statistical Summary using Heat-map



* Heat-map for the correlation table:



* ROC AUC Curve:



Area for the ROC curve is 0.91.

# CONCLUSION

* Key Findings and Conclusions of the Study
  + If the number of days of payback is increasing the chance of defaulters is also increasing. So, we should look for the payback duration.
  + If the loan amount is below 100 and the number of loans taken by users is 90 days, the number of defaulters is increasing.
* Learning Outcomes of the Study in respect of Data

Science

This project helped me to work on the real time industrial data, which helped me to gain the real time experience. In the project I got to work on the different type of algorithms and fitting the best model based on the accuracy score and cross validation score. We achieved accuracy score of 91% using the Decision Tree Classifier.