

MICRO CREDIT LOAN CASE

Submitted by:

AKASHDEEP SINGH MANRAL

Batch no.:

1834

**ACKNOWLEDGMENT**

The dataset and all the necessary details of the data was provided to me by company – Flip Robo.

I want to thank my intern mentor miss- Swati Mahaseth for providing assistance in solving my queries, with her help and guidance I was able to complete my project successfully.

**INTRODUCTION**

Data sources and Formats-

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 customer.

Data info-

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Definition | Comment | |
| label | Flag indicating whether the user paid back the credit amount within  5 days of issuing the loan{1:success, 0:failure} | |
| msisdn | mobile number of user |  | |
| aon | age on cellular network in days |  | |
| daily\_decr30 | Daily amount spent from main account  , averaged over last 30 days  (in Indonesian Rupiah) |  | |
| daily\_decr90 | Daily amount spent from main account  , averaged over last 90 days (in Indonesian Rupiah) |  | |
| rental30 | Average main account balance over last 30 days | Unsure of given definition | |
| rental90 | Average main account balance over last 90 days | Unsure of given definition | |
| last\_rech\_date\_ma | Number of days till last recharge of main account |  | |
| last\_rech\_date\_da | Number of days till last recharge of data account |  | |
| last\_rech\_amt\_ma | Amount of last recharge of main account (in Indonesian Rupiah) |  | |
| cnt\_ma\_rech30 | Number of times main account got recharged in last 30 days |  | |
| fr\_ma\_rech30 | Frequency of main account recharged in last 30 days | Unsure of given definition | |
| sumamnt\_ma\_rech30 | Total amount of recharge in main account over last 30 days (in Indonesian Rupiah) |  | |
| medianamnt\_ma\_rech30 | Median of amount of recharges done in main account over last 30 days at user level (in Indonesian Rupiah) |  | |
| medianmarechprebal30 | Median of main account balance just before recharge in last 30 days at user level (in Indonesian Rupiah) |  | |
| cnt\_ma\_rech90 | Number of times main account got recharged in last 90 days |  | |
| fr\_ma\_rech90 | Frequency of main account recharged in last 90 days | Unsure of given definition | |
| sumamnt\_ma\_rech90 | Total amount of recharge in main account over last 90 days (in Indonasian Rupiah) |  | |
| medianamnt\_ma\_rech90 | Median of amount of recharges done in main account over last 90 days at user level (in Indonasian Rupiah) |  | |
| medianmarechprebal90 | Median of main account balance just before recharge in last 90 days at user level (in Indonasian Rupiah) |  | |
| cnt\_da\_rech30 | Number of times data account got recharged in last 30 days |  | |
| fr\_da\_rech30 | Frequency of data account recharged in last 30 days |  | |
| cnt\_da\_rech90 | Number of times data account got recharged in last 90 days |  | |
| fr\_da\_rech90 | Frequency of data account recharged in last 90 days |  | |
| cnt\_loans30 | Number of loans taken by user in last 30 days |  | |
| amnt\_loans30 | Total amount of loans taken by user in last 30 days |  | |
| maxamnt\_loans30 | maximum amount of loan taken by the user in last 30 days | There are only two options: 5 & 10 Rs., for which the user needs to pay back 6 & 12 Rs. respectively | |
| medianamnt\_loans30 | Median of amounts of loan taken by the user in last 30 days |  | |
| cnt\_loans90 | Number of loans taken by user in last 90 days |  | |
| amnt\_loans90 | Total amount of loans taken by user in last 90 days |  | |
| maxamnt\_loans90 | maximum amount of loan taken by the user in last 90 days |  | |
| medianamnt\_loans90 | Median of amounts of loan taken by the user in last 90 days |  | |
| payback30 | Average payback time in days over last 30 days |  | |
| payback90 | Average payback time in days over last 90 days |  | |
| pcircle | telecom circle |  | |
| pdate | date |  | |

Business problem-

1. 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.
2. 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.
3. 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.
4. 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).

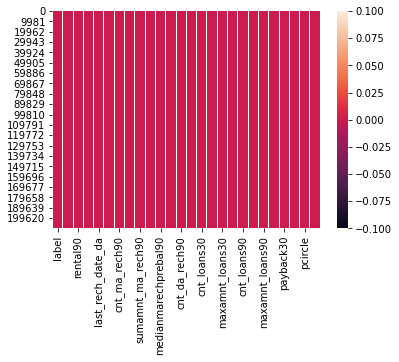
Problem Undertaken-

1. 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 first thing we did was the Exploratory Data Analysis (EDA)of our dataset. EDA was done using various mathematical/ analytical tools. We made use of heat maps , count plots , histograms , distplot, relplots, to carve out the valuable information of the users.
2. We used the heatmap to find out if there was any missing data in the dataset. This is done to eliminate or handle the missing data because such missing values may affect the data prediction score.



- This the heat map , showing a single colour, reflects there are no null values in the dataset.

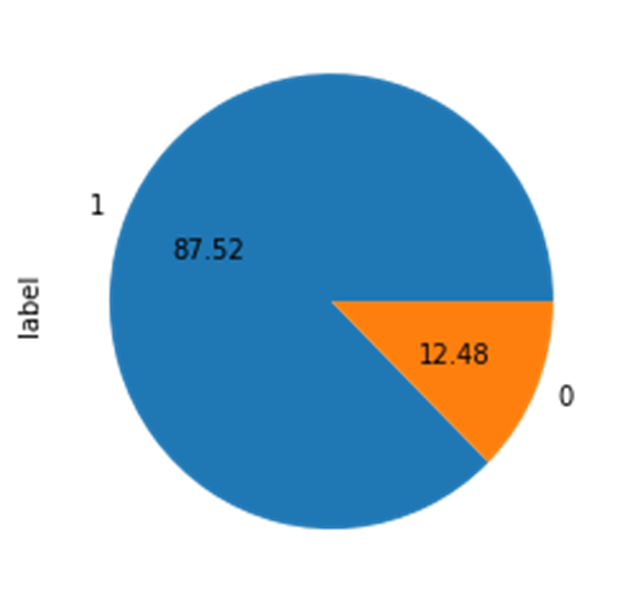
1. We then made use of statistical tools to get the valuable information from data.
2. Countplots –

Label =Flag indicating whether the user paid back the credit amount within 5 days of issuing the loan{1:success, 0:failure}

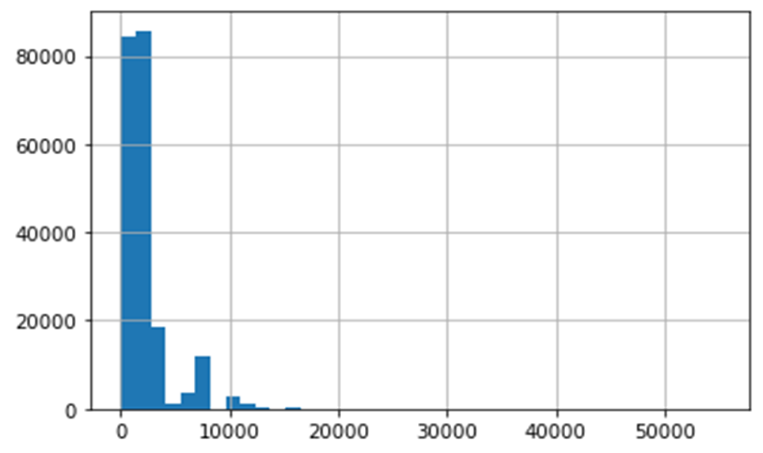


1. Pie charts-

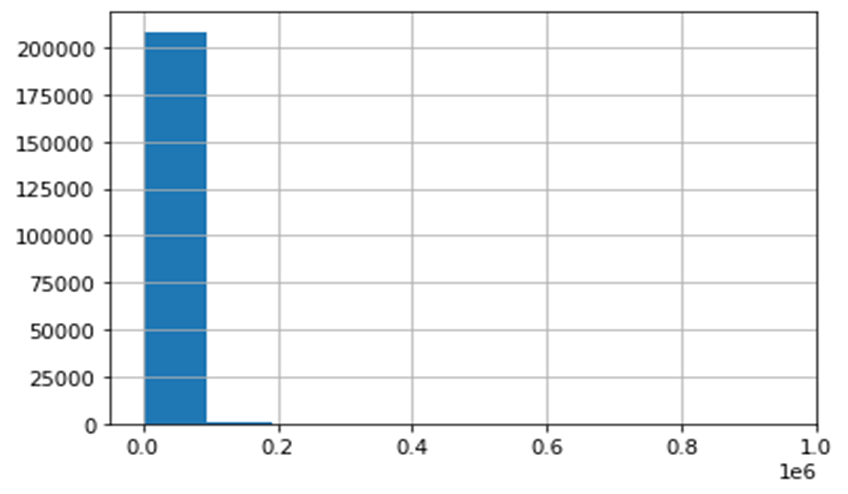
Percent of people-  
 1- who were able to payback within 5 days of issuing loan  
 0- who were not who were able to payback within 5 days of issuing loan



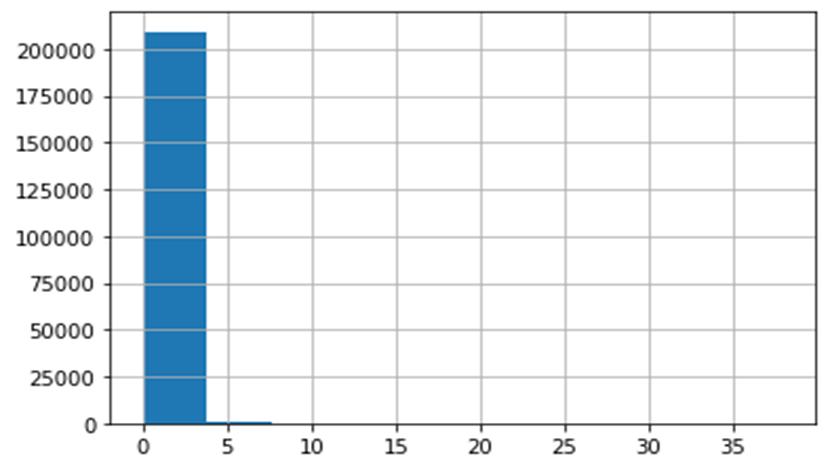
1. Histograms-
2. last\_rech\_amt\_ma=Amount of last recharge of main account (in Indonesian Rupiah)  
     
   observation-  
   1- max people have amount of last recharge of main account (in Indonesian Rupiah) between 0-1000 days approx.  
   2- min people have amount of last recharge of main account (in Indonesian Rupiah) beyond 10,000 days



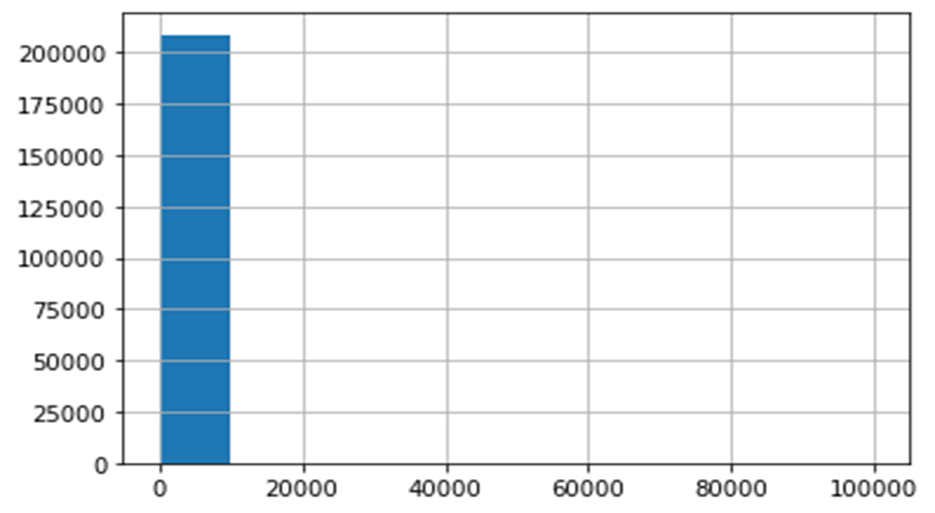
1. sumamnt\_ma\_rech90=Total amount of recharge in main account over last 90 days (in Indonesian Rupiah)  
     
   observation-  
   1-most of the people have done recharge in main account ,over last 90 days (in Indonesian Rupiah), of amount between (0.0-0.1)(le6) (0-100000 in Indonesian Rupiah).  
   2-very few people have done recharge of amount between (100000-19000 Indonesian Rupiah )



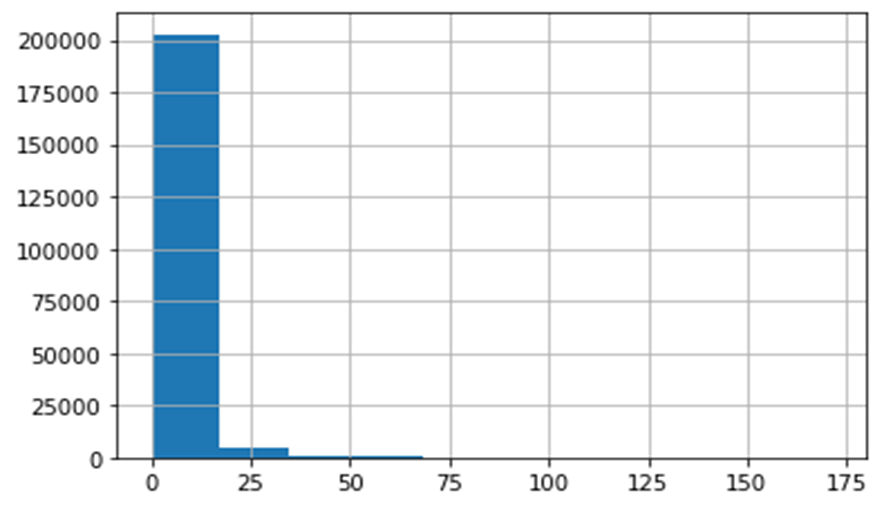
1. cnt\_da\_rech90=Number of times data account got recharged in last 90 days  
     
   observation-  
   Number of times data account got recharged in last 90 days  
   a-0-3 times mostly  
   b-beyond 3 times very less



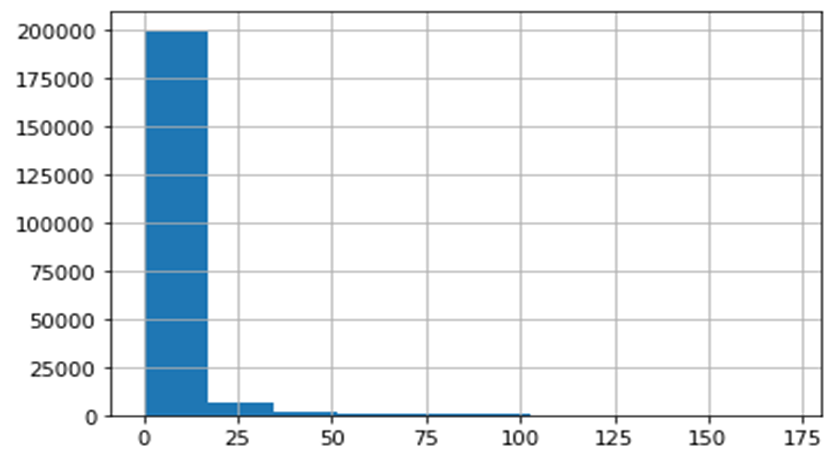
1. cnt\_da\_rech30=Number of times data account got recharged in last 30 days  
     
   observation-  
   Every data account got recharged between 0-10000 times in last 30 days



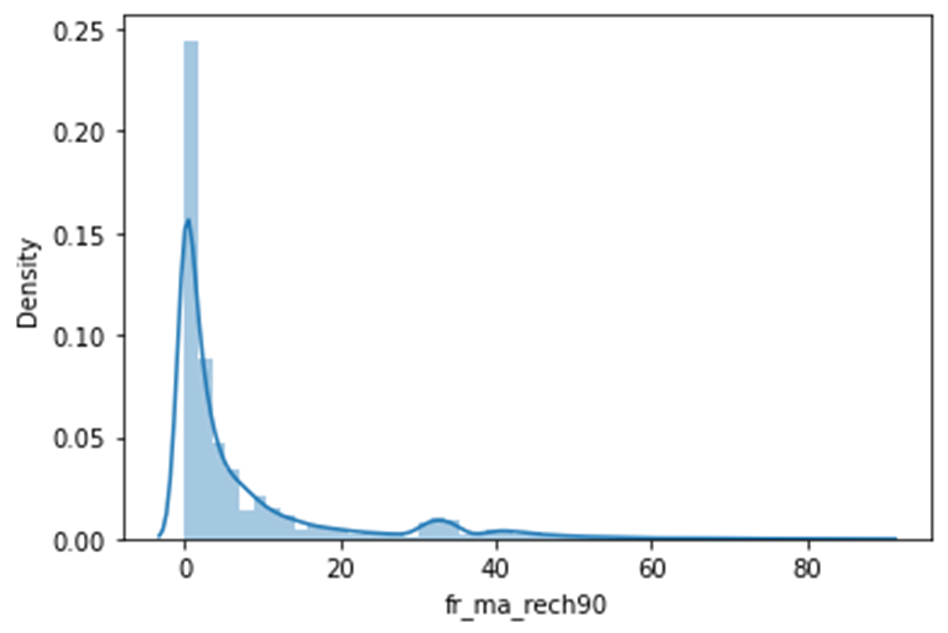
1. payback30=Average payback time in days over last 30 days  
     
   observation-  
   most of the people had avg payback time of 0-15days



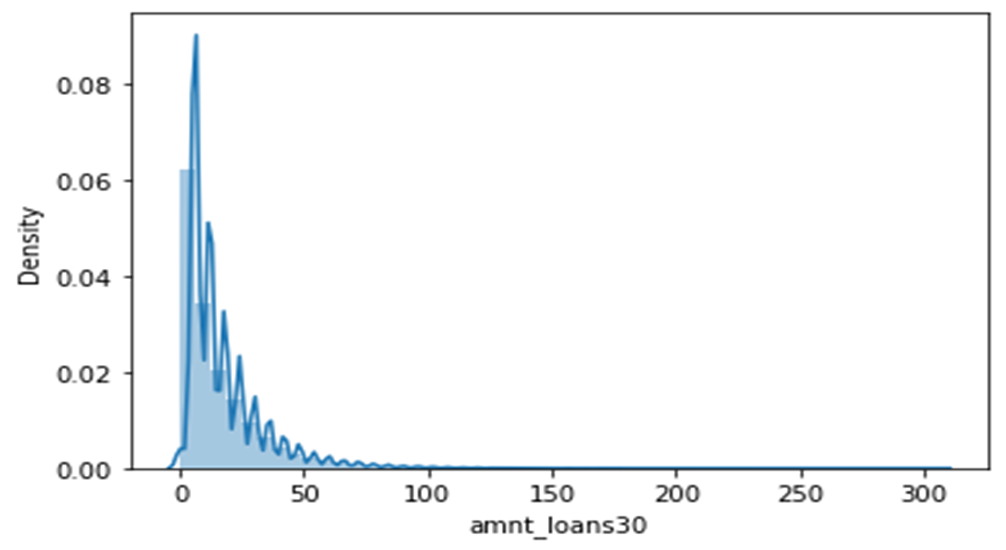
1. payback90=Average payback time in days over last 90 days  
     
   observation-  
   most of the people had avg payback time of 0-15days



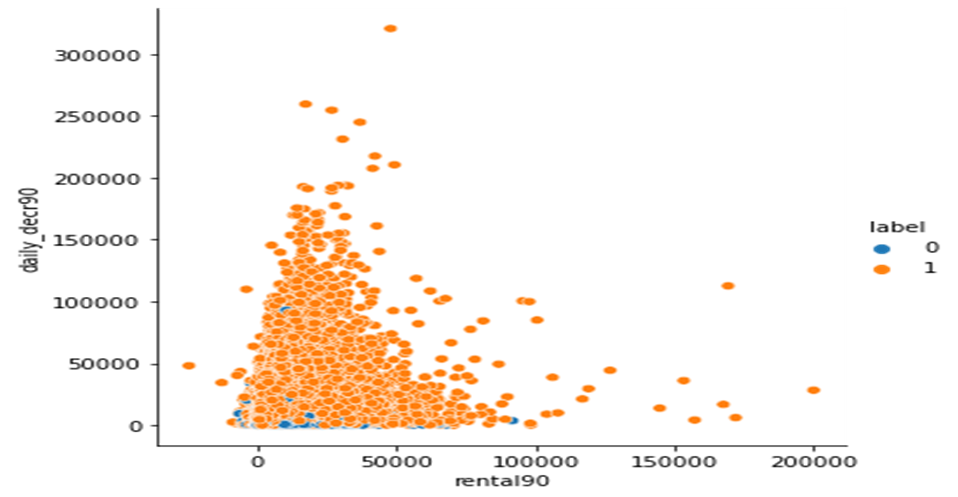
1. Distplots-
2. fr\_ma\_rech90=Frequency of main account recharged in last 90 days  
     
   observation-  
   highest Frequency of main account recharged in last 90 days - is between 0-5 days



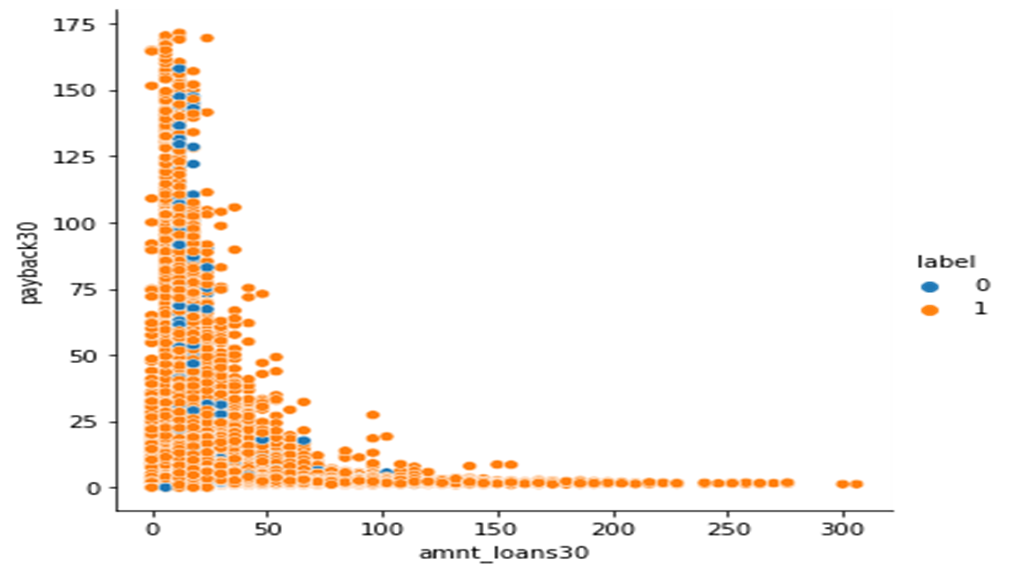
1. amnt\_loans30-Total amount of loans taken by user in last 30 days  
     
   observation-  
   Total amount of loans taken by user in last 30 days-  
   a-max- between 0-10 times  
   b-min- beyond 100



1. relplots-
2. rental90= average main account balance over last 90 days  
   daily\_decr90= daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah)  
     
   observation-  
   1-most of the people have their avg main account balance over last 90 days between 0-50,000 and their daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah) was also between 0-50000  
   2-very few peoplwe had their avg main account balance between 1lakh-2lakh and their spending was around 50000  
   3-a few cases were also there whose rantal 90 was around 60,000 and daily\_decr90 was more than 3lakh



1. payback30=Average payback time in days over last 30 days  
     
   observation-  
     
   most of the people low total amount of loan(0-50loans) were having average payback time of(0-50 days)



**Model/s Development and Evaluation**

After getting the analytical information from the dataset,Outliers were detected and removed from the dataset

1. Outliers- An outlier is a data point that is noticeably different from the rest. They represent errors in measurement, bad data collection, or simply show variables not considered when collecting the data.
2. Outlier detection- it is done through zscore method

Zscore- Z score is an important concept in statistics. Z score is also called standard score. This score helps to understand if a data value is greater or smaller than mean and how far away it is from the mean. More specifically, Z score tells how many standard deviations away a data point is from the mean.

***Z score = (x -mean) / std. deviation***

1. If the z score of a data point is more than 3, it indicates that the data point is quite different from the other data points. Such a data point can be an outlier.
2. The outliers are then removed from the datset.
3. After the completion of EDA , the data was divided into two variables

X= All the independent variable

Y= The dependent variable

1. Both x and y were trained and tested on various machine learning models,such as- Logistic Regression, DecisionTreeClassifier, RandomForestClassifier,
2. Logistic Regression- Logistic Regression is used when the dependent variable(target) is categorical

For ex-

When we have to decide, true or false, yes or no, 0 or 1, types of solutions.

1. Logistic Regression assumptions-
2. The dependent variable must be categorical in nature
3. .The independent variable should not have multi-collinearity
4. Logistic Regression Equation-
5. We know the equation of the straight line can be written as:

Logistic Regression in Machine Learning

1. In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

Logistic Regression in Machine Learning

1. But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

Logistic Regression in Machine Learning

1. Decision Tree Classifier-
2. Decision Tree is a Supervised learning techniquethat can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where**internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcome.**
3. In a Decision tree, there are two nodes, which are the **Decision Node** and**Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
4. The decisions or the test are performed on the basis of features of the given dataset.
5. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
6. A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.



1. Random Forest Classifier-
2. Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase
3. The Working process can be explained in the below steps and diagram:
4. **Step-1:** Select random K data points from the training set.
5. **Step-2:** Build the decision trees associated with the selected data points (Subsets)
6. **Step-3:** Choose the number N for decision trees that you want to build.
7. **Step-4:** Repeat Step 1 & 2.
8. **Step-5:** For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.



1. Metrices used for evaluation-
2. Confusion matrix- A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making.

****

**TP- True** Positive Value**-** The predicted value matches the actual value and actual value is positive and machine also predicted a positive value.

**TN-** True Negative value- The predicted value matches the actual value and the actual value is negative and machine also predicted the negative value.

**FP-** False Positive- The predicted value was false. The actual value was negative but model predicted a positive value.

**FN-** False Negative- The predicted value was false. The actual value was positive but model predicted negative value.

1. Accuracy score-

https://cdn.analyticsvidhya.com/wp-content/uploads/2020/04/Equation_Accuracy.png

1. Classification Report-

It is one of the performance evaluation metrics of a classification-based machine learning model. It displays your model’s precision, recall, F1 score and support. It provides a better understanding of the overall performance of our trained model.

**Precision**- It is the ratio of true positives to the sum of true and false positives.

**Recall**- It is the ratio of true positives to the sum of true positivs and false negatives

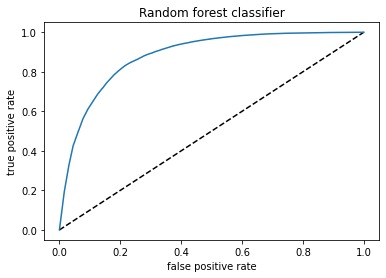
**F1 score**- It is the weighted harmonic mean of Precision and Recall. The closer the value of F1 score to 1.0 , the better the result.

**Support-** It is the number of actual occurences of class in the datasets.

**CONCLUSION**

1. Results- Based on the accuracy score ,**RandomForestClasifier** was the best model for prediction.
2. AUC ROC curve- The Receiver Operator Characteristic **(ROC)** curve is an evaluation metric for binary classification problems. It is a probability curve that plots the **TPR**against **FPR**at various threshold values and essentially separates the ‘signal’ from the ‘noise’. The **Area Under the Curve (AUC)**is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.

So we made a AUC ROC curve of our model which is based on the classification report and it looked like-



1. We also made a data frame comparing the predicted and actual values and the model was working very efficiently.

**The model was able to predict, whether the user would be able to pay back the loan within 5 days of issuing the loan or not, provided all the given details of the user**.

1. Limitations- The model was working on the 90% accuracy.so the predicted values does not give 100% guarantee. In some cases, the predicted result would not be matching with the actual results.