FLIP ROBO

Report on

Micro Credit Loan

Submitted by:

PRANKUL JAIN

ACKNOWLEDGMENT

Foremost, I would like to express my deepest and sincere gratitude towards the team at “Flip Robo Technologies” for their continuous support towards the project and other facets of the assignment along with offering a position as an Intern.

Besides them, extending my gratitude towards the academic team at “DataTrained” for their continuous mentoring sessions and their support in the knowledge transfer sessions.

Finally, from the very onset on my assignment I’ve extensively used online platforms like stackoverflow, w3school, kaggle and many others. I’m obligated to pay my gratitude to every personage involved in making content on these platforms.

INTRODUCTION

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. 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. 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). Using the historical data of the customer on their recharges, we will be predicting the defaulters with the help of Machine Learning models.

Analytical Problem Framing

The given dataset has 209593 rows and 35 columns. Using this dataset we will be training the Machine Learning models on 77% of the data and the models will be tested on 33% data. Although the given dataset doesn't have any null value, we can expect outliers and un-realistic values for certain variables. This data was collected for the UPW telecom circle in the year 2016. Below are the definition for each variable available on the dataset

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
rental90	Average main account balance over last 90 days
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
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
sumamnt_ma_rech90	Total amount of recharge in main account over last 90 days (in Indonesian Rupiah)
medianamnt_ma_rech90	Median of amount of recharges done in main account over last 90 days at user level (in Indonesian Rupiah)
medianmarechprebal90	Median of main account balance just before recharge in last 90 days at user level (in Indonesian 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 rechargedin 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
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

• Data Preprocessing Done

- “Unnamed:0” – This row is a dummy row just like an indexing starting from 1
- “msisdn” – The data definition column above clearly states that this is a subscriber mobile number and they are randomly generated and will not have any meaning in the prediction of credit defaulters.
- “pcircle” – This feature is same throughout the rows (UPW) and will not have any effect on the target variable

Post removal of the columns Unnamed:0 and msisdn, We can see the datatypes of the remaining columns.
most of the columns are of numerical data type except pcircle and pdate (Telecom circle and Date respectively) and I have the confirmation that the dataset has no null values.

Since we have dropped the pcircle, we will be extracting the features from the date, here we’ll be extracting date and month ignoring year because it’s the same for every row (i.e., 2016).

Post data extraction from the date column, we deleted the pdate and currently have 35 features.

Now that we have all the columns in numerical type. We can explore thedata and its relationship with the target variable.

I have used ".describe" to understand the shape of the data. You can view the snippets of the same below.

Data Inputs- Logic- Output Relationships

Exploratory data analysis (EDA) is conducted across the dataset. The object here is to establish a relationship between the target column and the input data. To achieve this I've made use of boxplot,kdeplot/distplot, scatter plot, barplot and strip plot. Most numeric continuous data have a linear relationship with the taget column (house price).

Hardware and Software Requirements and Tools Used

Hardware necessary:

- RAM- 8GB or above
- Processor – Core i5 or above
- SSD- 250GB or above

Software necessary:

- Anaconda

Libraries:

- import pandas as pd: pandas is a popular Python-based data analysis toolkit which can be imported using import pandas as pd. It presents a diverse range of utilities, ranging from parsing multiple file formats to converting an entire data table into a numpy matrix array. This makes pandas a trusted ally in data science and machine learning.
- import numpy as np: NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional

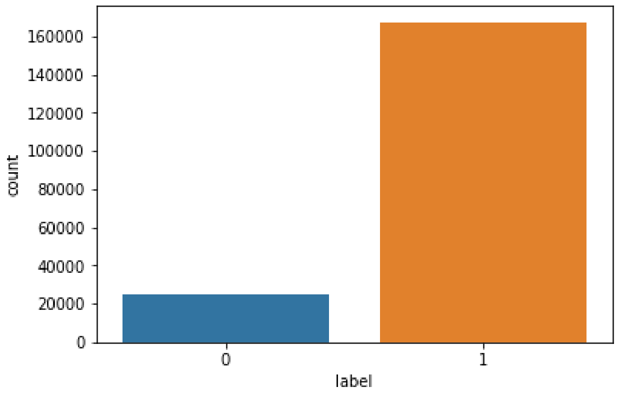
array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.

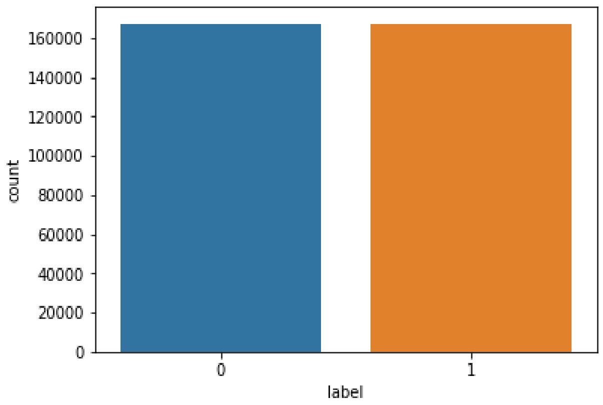
- import seaborn as sns: Seaborn is a data visualization library built on top of matplotlib and closely integrated with pandas data structures in Python. Visualization is the central part of Seaborn which helps in exploration and understanding of data.
- Import matplotlib.pyplot as plt: matplotlib.pyplot is a collection of functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.
- from sklearn.ensemble import RandomForestRegressor
- from sklearn.model_selection import cross_val_score

Model/s Development and Evaluation

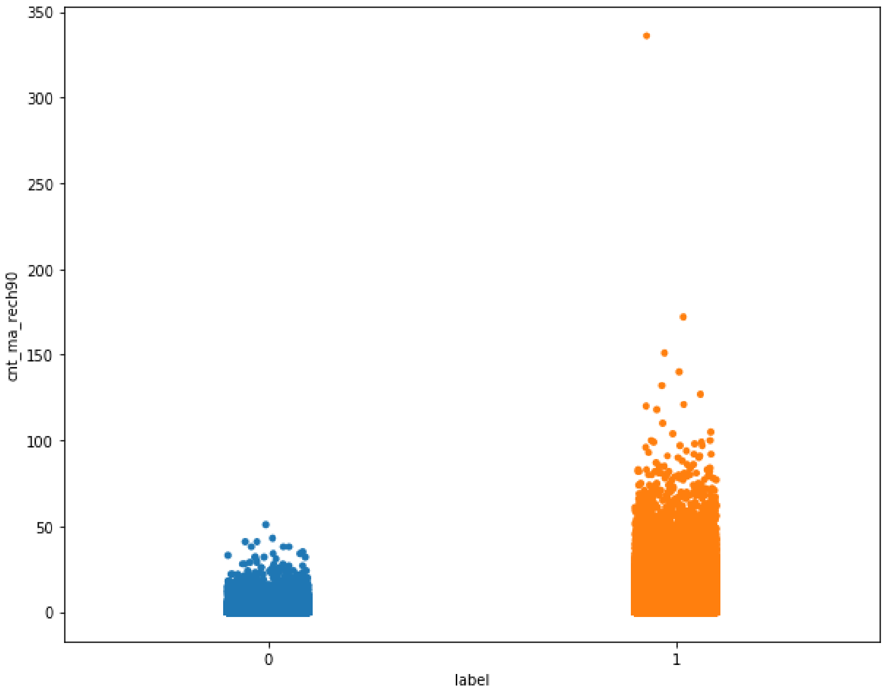
We identified that there was class imbalance in the dataset, before proceeding with the model building, I’m balancing the class using SMOTE over-sampling technique. This is necessary because, if the balancing is not done the model tends to predict the majority class better and the minority class will not be accurately predicted.

Below plots compare the class count before and after performing SMOTE.

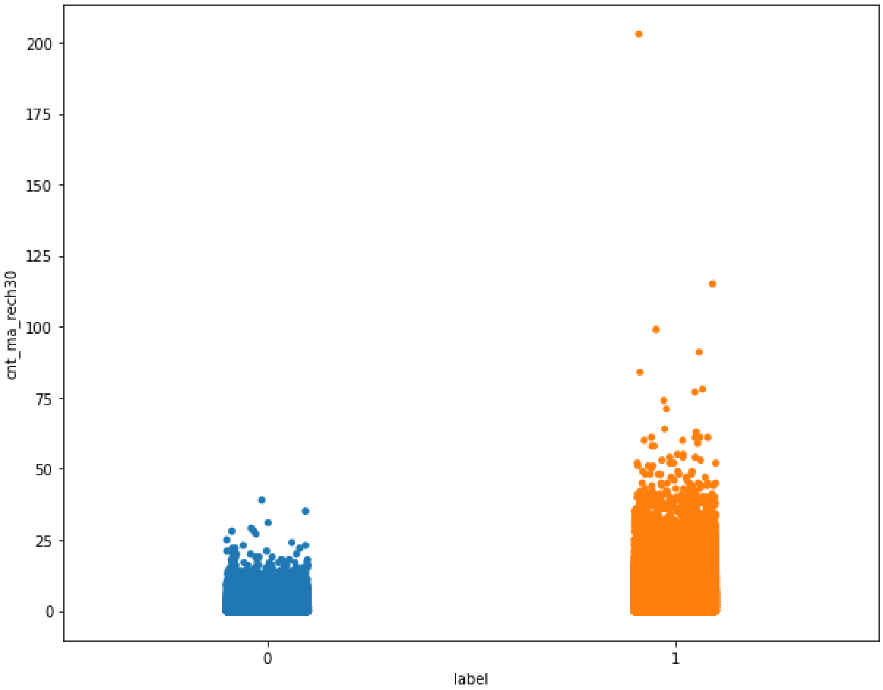




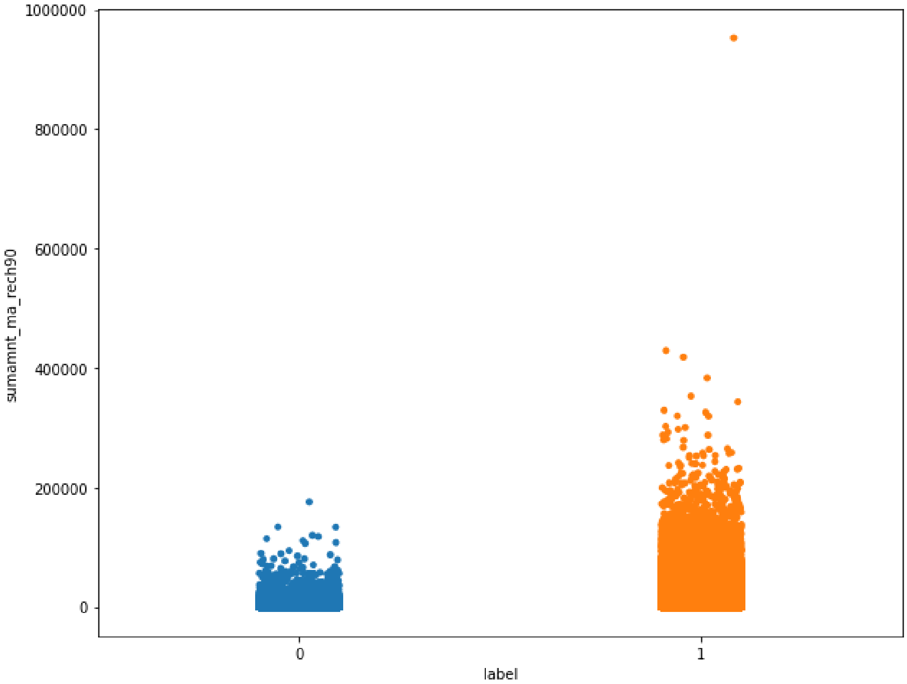
- Visualizations



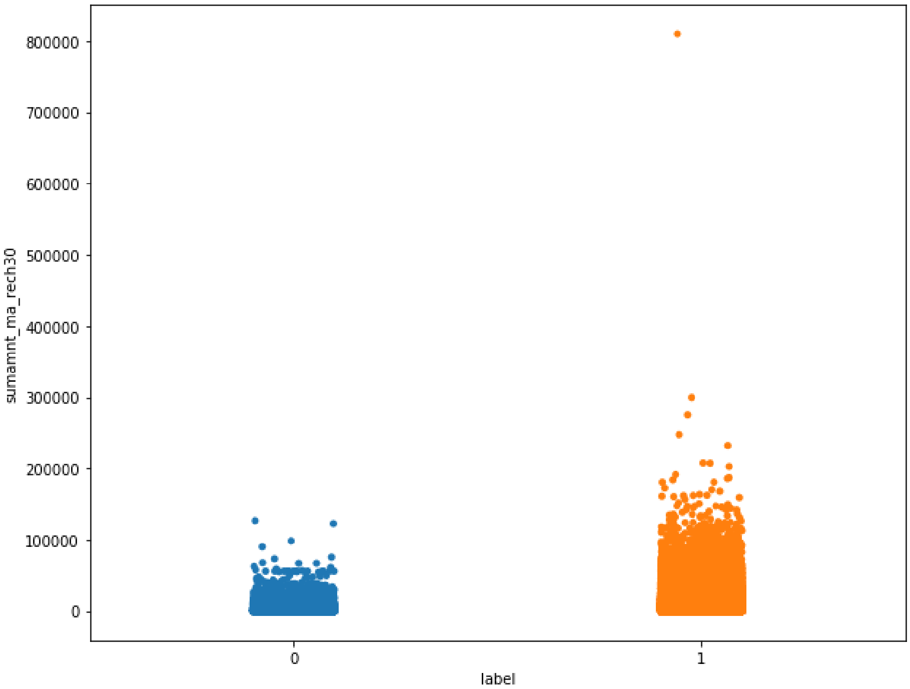
- We can see that the customers who repaid the loan made recharges more than 100 times and all the customers who didn't repay the loan recharged less than 50 times in 90 days



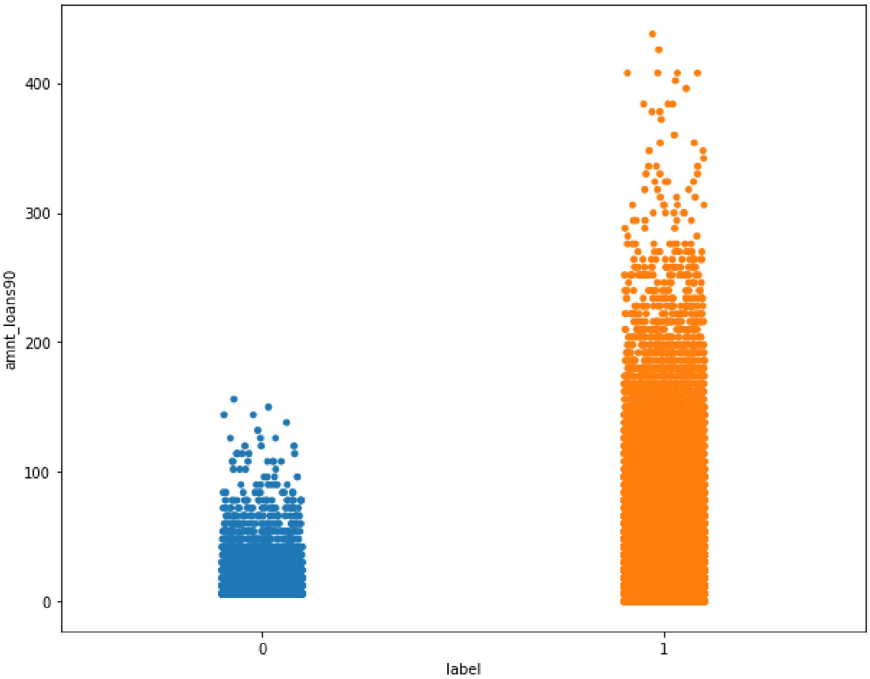
- We can say that the customers who didn't repay the loan, recharged the main account less than 25 times.



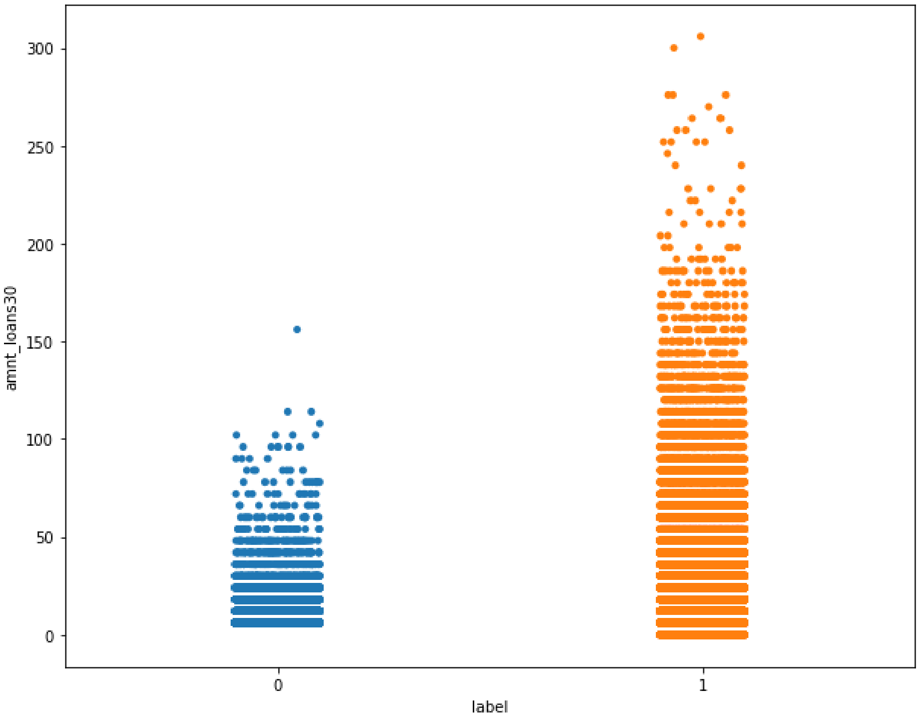
- From the above observation, I can say that the customers who didn't repay the loan recharged their main account less than 100000 Indonesian over the past 90 days



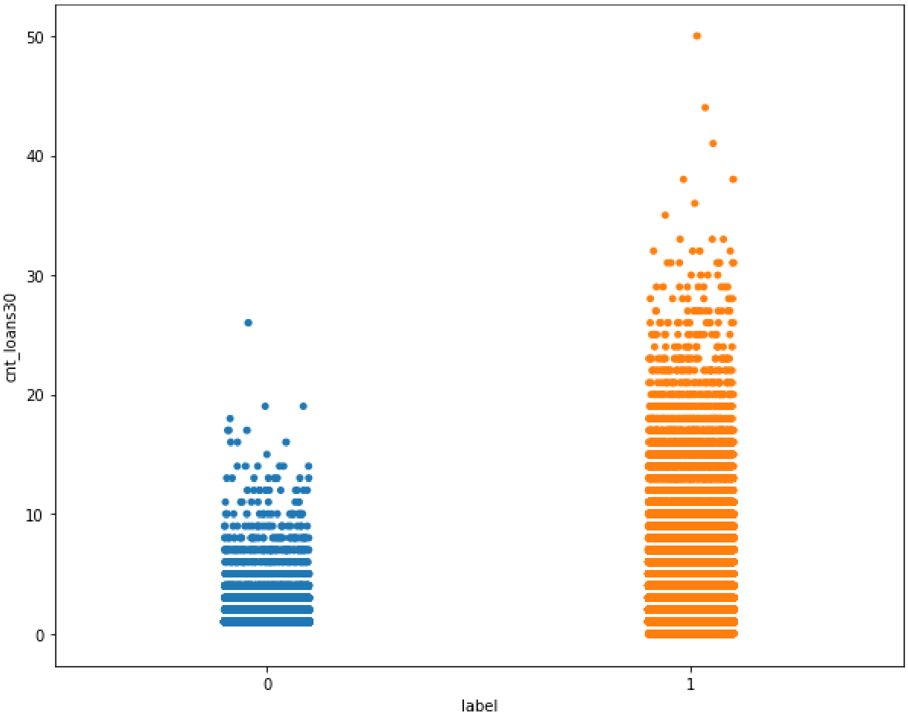
- From the above observation, I can say that the customers who didn't repay the loan recharged their main account less than 50000 Indonesian over the past 30 days



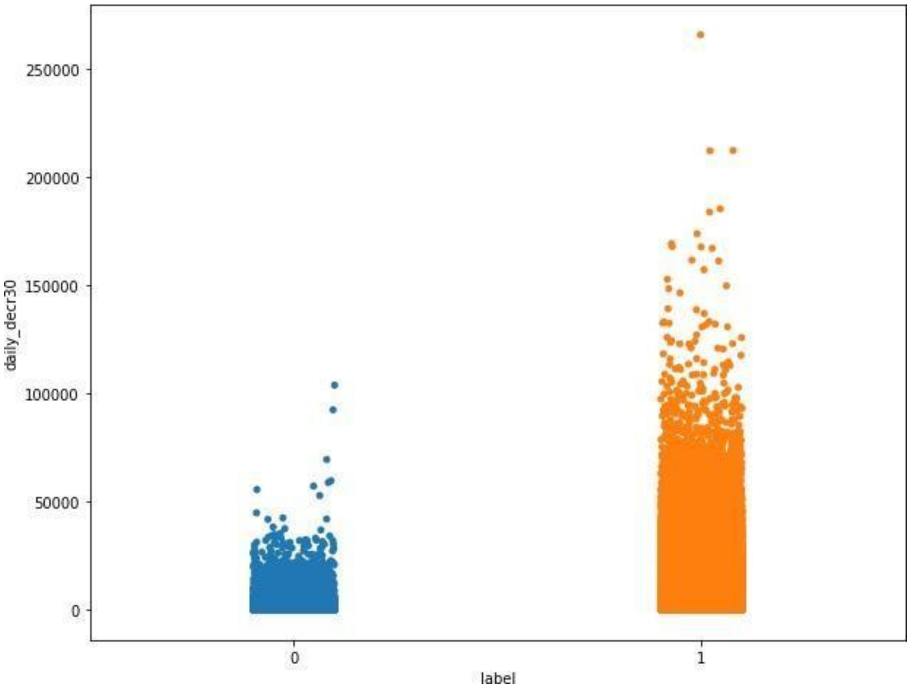
- Upon reviewing, we can say that the customers didn’t repay the loan, took less than 100 Indonesian rupiah as loans over the 90 day period.



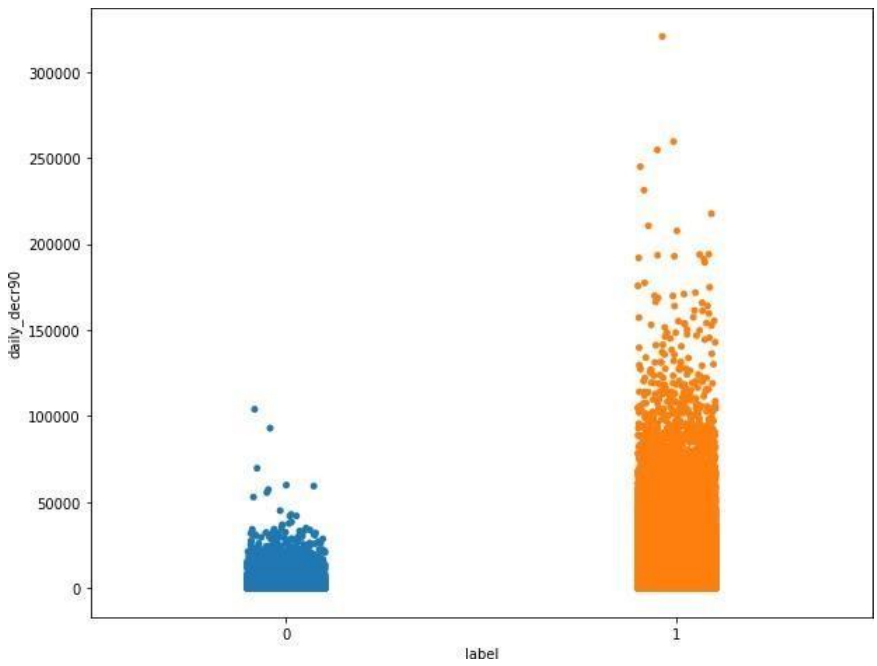
- When we look at the 30 day period, customers who took less than 100 Indonesian rupiah as loans didn’t repay the loan.
- In both of the above cases, we can also say that the lower number of loans were recorded as a result of customers not repaying the amount in time. Hence lesser amount



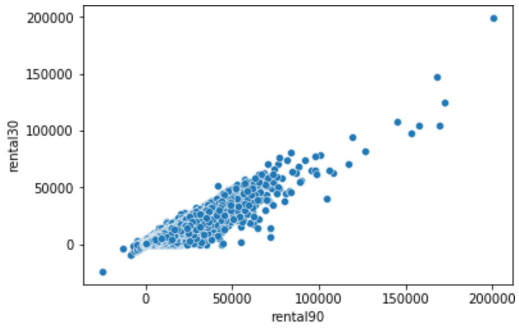
- Here, we can also say that, the lesser the number of loans taken, higher the chance of defaulting the loan. The above figure suggests that the customers didn't repay when they took less than 15 loans over the period of 30 days.

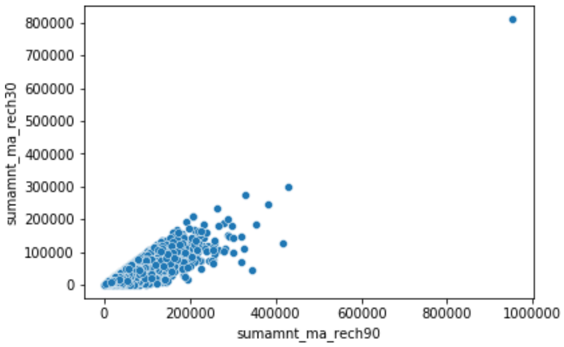


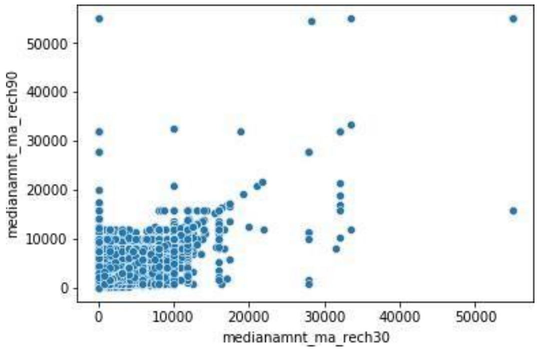
- The above figure suggests that the customers who didn't repay the loan spent less than 50000 Indonesian rupiah on an average over the past 30 days.

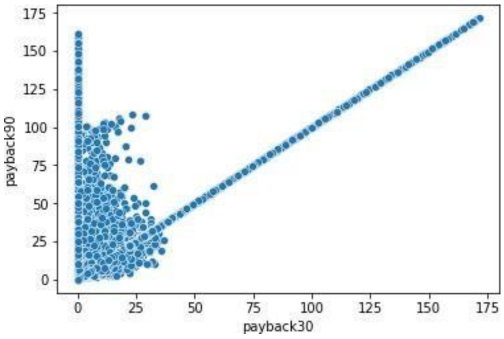


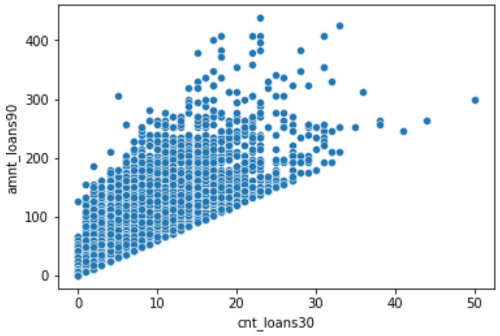
- The above figure suggests that the customers who didn’t repay the loan spent less than 35000 Indonesian rupiah on an average over the past 90 days.

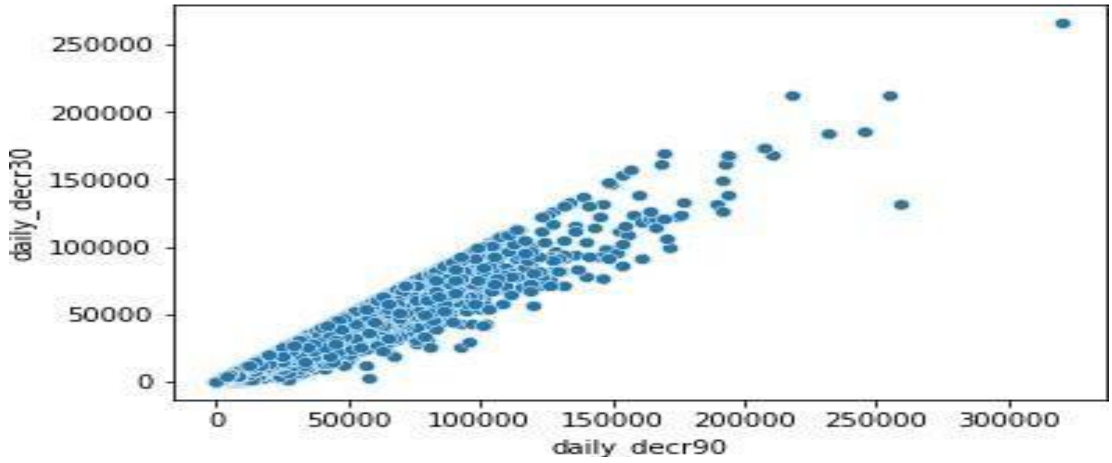


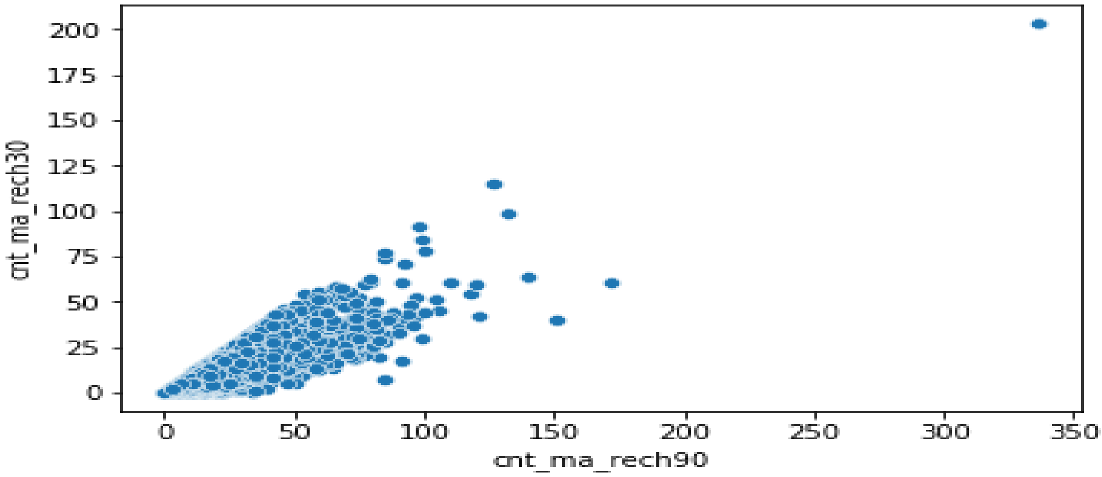


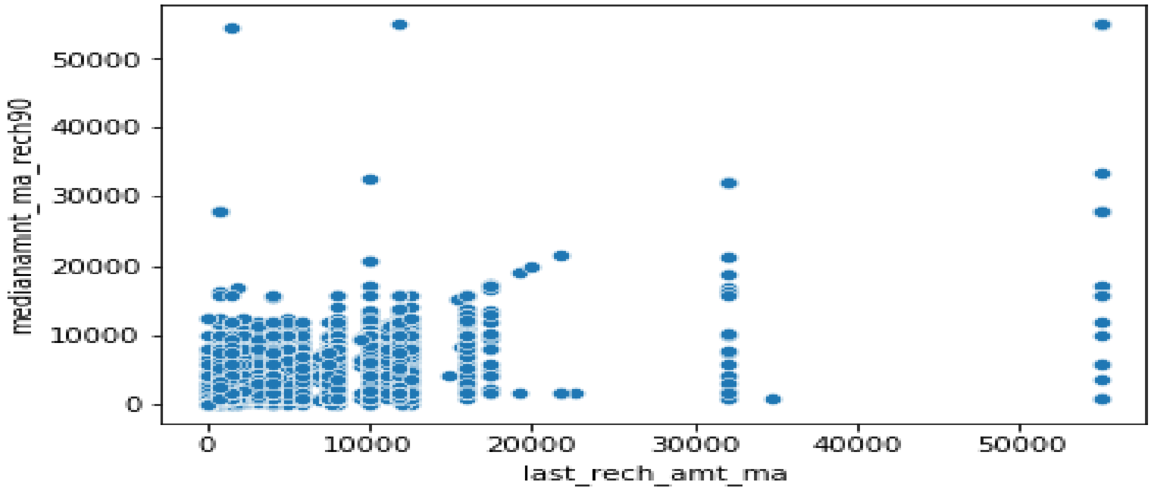


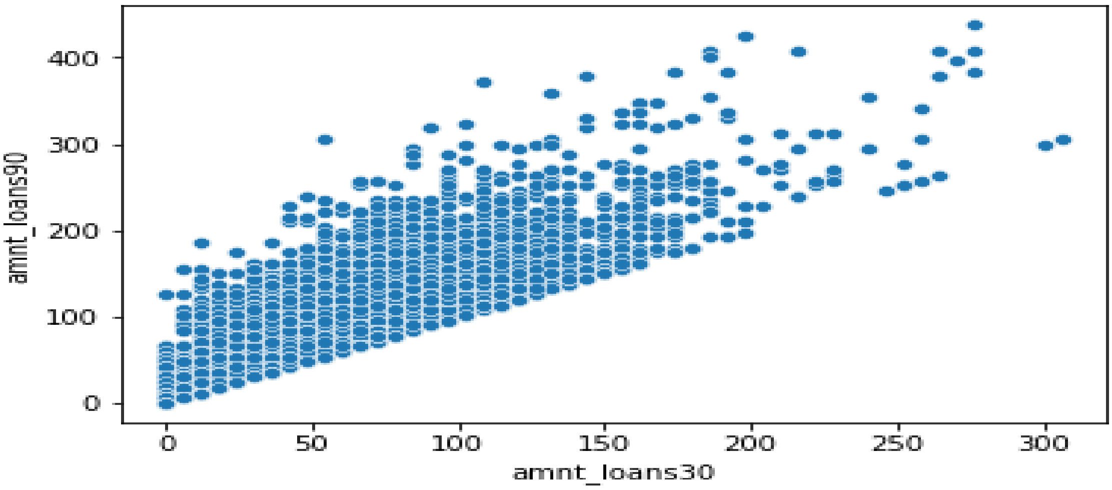





daily_decr30
250000
200000
150000
100000
50000
0
0
50000
100000
150000
200000
250000
300000
daily_decr90


cnt_ma_rech30
200
175
150
125
100
75
50
25
0
0
50
100
150
200
250
300
350
cnt_ma_rech90


medianamnt_ma_rech90
50000
40000
30000
20000
10000
0
0
10000
20000
30000
40000
50000
last_rech_amt_ma


amnt_loans90
400
300
200
100
0
0
50
100
150
200
250
300
amnt_loans30

CONCLUSION

- Key Findings and Conclusions of the Study

We have successfully built a model using multiple models and found that the Random forest Classifier model.

Below are the details of the model's metrics predicting the dataset

1. Average precision of 0.93
2. Average recall of 0.93
3. F1 Score is 0.93

The ability of a classifier to distinguish between classes (AUC) is also 0.93