Flip Robo Project

**PROJECT:**

**“Micro-Credit Defaulter Model”**

**Prepared By:**

**Anirudh Bhardwaj**

1.INTRODUCTION

1.1 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.

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.

Conceptual Background of the Domain Problem:

Telecom Industries 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).

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.

We have 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.

2.Analytical Problem Framing

Mathematical/ Analytical Modelling of the Problem

In this particular problem I had label as my target column and it was having two classes 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. So clearly it is a binary classification problem and I have to use all classification algorithms while building the model. There was no null values in the dataset. Also, I observed some unnecessary entries in some of the columns like in some columns I found more than 90% zero values so I decided to drop those columns. If I keep those columns as it is, it will create high skewness in the model. To get better insight on the features I have used plotting like distribution plot, bar plot and count plot. 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 percentile method and I removed skewness using yeo-Johnson method. I have used all the classification algorithms while building model then turned the best model and saved the best model. At last, I have predicted the label using saved model.

2.2 Data Sources and their formats

The data was collected for my internship company – Flip Robo technologies in excel format. The sample data is provided to us from our client database. It is hereby given to us 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.

Also, my dataset was having 209593 rows and 36 columns including target. In this particular dataset I have object, float and integer types of data. The information about features is as follows.

**Features Information:**

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

2. msisdn : mobile number of user

3. aon : age on cellular network in days

4. daily\_decr30 : Daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah)

5. daily\_decr90 : Daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah)

6. rental30 : Average main account balance over last 30 days

7. rental90 : Average main account balance over last 90 days

8. last\_rech\_date\_ma : Number of days till last recharge of main account

9. last\_rech\_date\_da: Number of days till last recharge of data account

10. last\_rech\_amt\_ma : Amount of last recharge of main account (in Indonesian Rupiah)

11. cnt\_ma\_rech30 : Number of times main account got recharged in last 30 days

12. fr\_ma\_rech30 : Frequency of main account recharged in last 30 days

13. sumamnt\_ma\_rech30 : Total amount of recharge in main account over last 30 days (in Indonesian Rupiah)

14. medianamnt\_ma\_rech30 : Median of amount of recharges done in main account over last 30 days at user level (in Indonesian Rupiah)

15. medianmarechprebal30 : Median of main account balance just before recharge in last 30 days at user level (in Indonesian Rupiah)

16. cnt\_ma\_rech90 : Number of times main account got recharged in last 90 days

17. fr\_ma\_rech90 : Frequency of main account recharged in last 90 days

18. sumamnt\_ma\_rech90 : Total amount of recharge in main account over last 90 days (in Indonesian Rupiah)

19. medianamnt\_ma\_rech90 : Median of amount of recharges done in main account over last 90 days at user level (in Indonesian Rupiah)

20. medianmarechprebal90 : Median of main account balance just before recharge in last 90 days at user level (in Indonesian Rupiah)

21. cnt\_da\_rech30 : Number of times data account got recharged in last 30 days

22. fr\_da\_rech30: Frequency of data account recharged in last 30 days

23. cnt\_da\_rech90 : Number of times data account got recharged in last 90 days

24. fr\_da\_rech90 : Frequency of data account recharged in last 90 days

25. cnt\_loans30 : Number of loans taken by user in last 30 days

26. amnt\_loans30: Total amount of loans taken by user in last 30 days

27. maxamnt\_loans30 : maximum amount of loan taken by the user in last 30 days

28. medianamnt\_loans30 : Median of amounts of loan taken by the user in last 30 days

29. cnt\_loans90 : Number of loans taken by user in last 90 days

30. amnt\_loans90 : Total amount of loans taken by user in last 90 days

31. maxamnt\_loans90 : maximum amount of loan taken by the user in last 90 days

32. medianamnt\_loans90 : Median of amounts of loan taken by the user in last 90 days

33. payback30 : Average payback time in days over last 30 days

34. payback90 : Average payback time in days over last 90 days

35. pcircle : telecom circle

36. pdate : date

Data Pre-processing Done

* As a first step I have imported required libraries and I have imported the dataset which was in csv format.
* Then I did all the statistical analysis like checking shape, nun in que, value counts, info etc…...
* Then while looking into the value counts, I found some columns with more than 90% zero values this creates skewness in the model and there are chances of getting model bias so I have dropped those columns with more than 90% zero values.
* While checking for null values I found no null values in the dataset.

I have also droped Unnamed:0, msisdn and pcircle column as I found they are useless.

Next as a part of feature extraction I converted the pdate column to pyear, pmonth and pday. Thinking that this data will help us more than pdate.

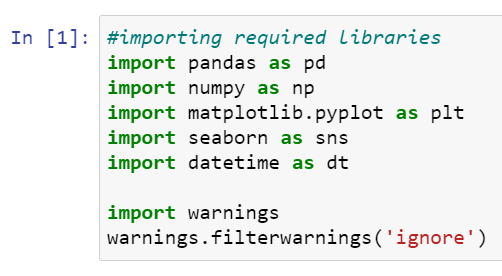
In some columns I found negative values which were unrealistic so I have converted those negative values to positive using abs command.

Also, I have converted all the flaot values in maxamnt\_loans90 to zero as it is specified in the problem statement we can have only 0,6,12 as maximum amount of loan taken by the user in last 30 days. As well I have droped all the data with amnt\_loans90=0 as it gives the persons who have not taken any loans.

2.4 Data Inputs- Logic- Output Relationships

* Since I had all numerical columns I have plotted dist. plot to see the distribution of each column data.
* I have used box plot for each pair of categorical features that shows the relation between label and independent features. Also we can observe whether the person pays back the loan within the date based on features.
* In maximum features relation with target, I observed non-defaulter count is high compared to defaulters.

**Libraries** **required :-**

To run the program and to build the model we need some basic libraries as follows:

import pandas as pd

import numpy as np

import seaborn as sns

Import matplotlib.pyplot as plt

from sklearn.preprocessing import OrdinalEncoder

from sklearn.preprocessing import MinMaxScaler

from sklearn.ensemble import RandomForestClassifier

from sklearn.tree import DecisionTreeClassifier

from xgboost import XGBClassifier

from sklearn.ensemble import GradientBoostingClassifier

from sklearn.ensemble import ExtraTreesClassifier

from sklearn.metrics import classification\_report

from sklearn.metrics import accuracy\_score

from sklearn.model\_selection import cross\_val\_score

With this sufficient libraries we can go ahead with our model building.

3.Data Analysis and Visualization

3.1 Identification of possible problem-solving approaches (methods)

To remove outliers I have used percentile method. And to remove skewness I have used yeo-johnson method. We have dropped all the unnecessary columns in the dataset according to our understanding. Use of Pearson’s correlation coefficient to check the correlation between dependent and independent features. Also I have used Normalization to scale the data. After scaling we have to balance the target column using oversampling. Then followed by model building with all Classification algorithms. I have used oversampling to get rid of data misbalancing. The balanced output looks like this.



Testing of Identified Approaches (Algorithms)

Since label was my target and it was a classification column with 0-defaulter and 1-Non-defaulter, so this perticular problem was Classification problem. And I have used all Classification algorithms to build my model. By looking into the difference of accuracy score and cross validation score I found RandomForestClassifier as a best model with least difference. Also to get the best model we have to run through multiple models and to avoid the confusion of overfitting we have go through cross validation. Below are the list of classification algorithms I have used in my project.

* XGBClassifier
* DecisionTreeClassifier
* BaggingClassifier
* AdaBoostClassifier

Key Metrics for success in solving problem under consideration

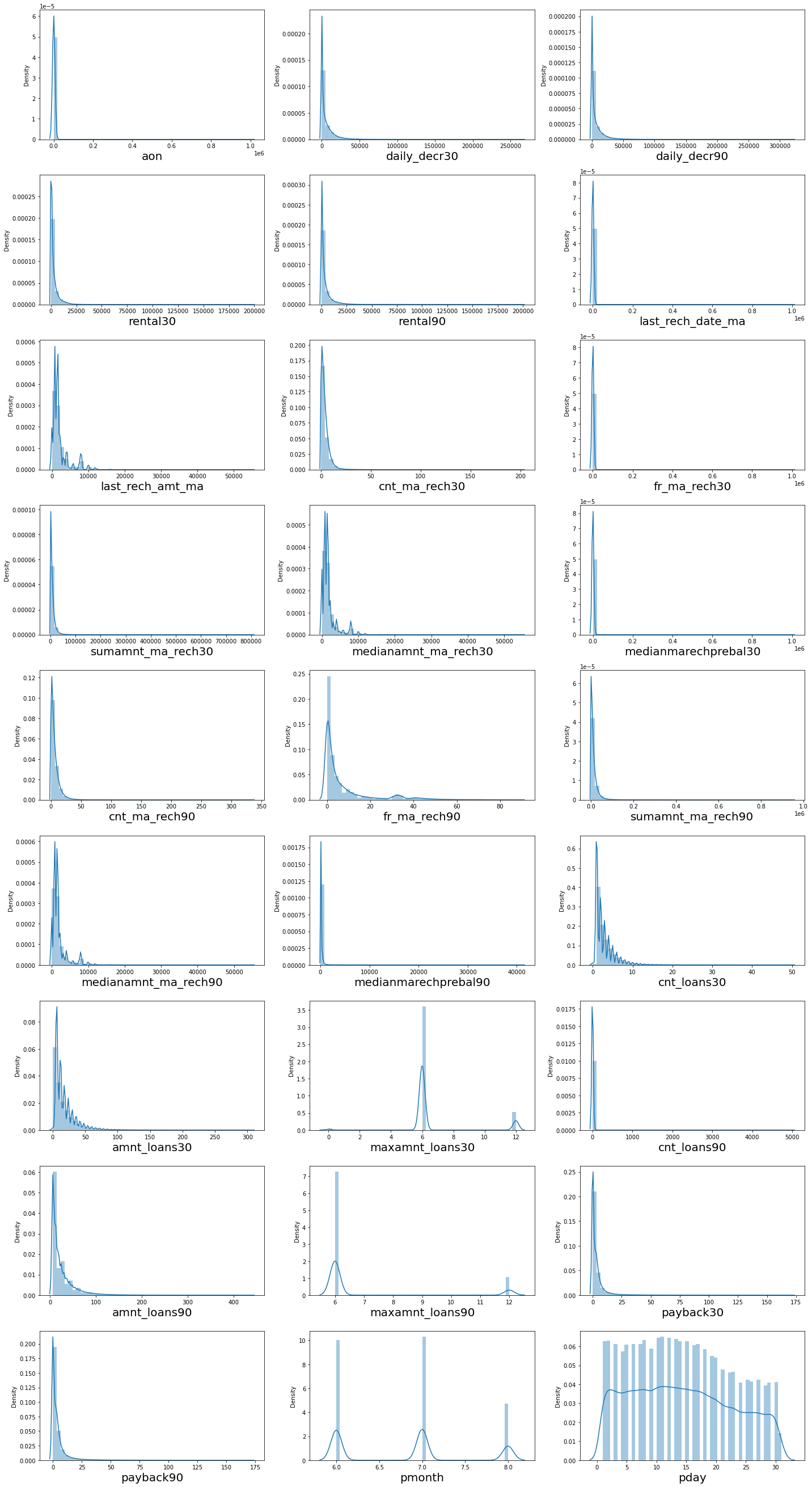
I have used the following metrics for evaluation:

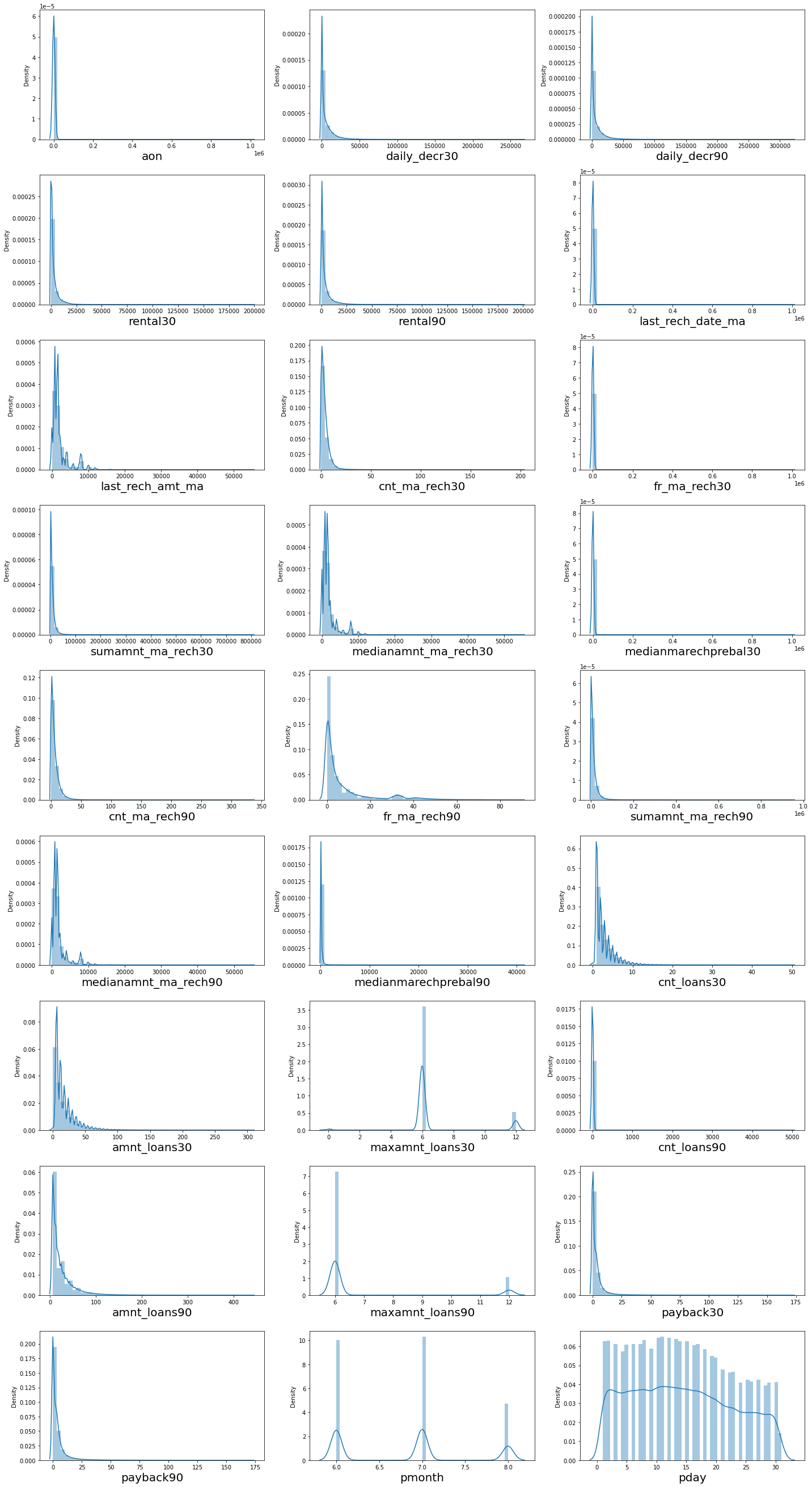
* Precision
* Recall
* Accuracy score
* F1-score:
* Cross\_val\_score
* AUC\_ROC \_score

3.4 Visualizations

I have used bar plots to see the relation of numerical feature with target and I have used 2 types of plots for numerical columns one is disp plot for univariate and bar plot for bivariate analysis.

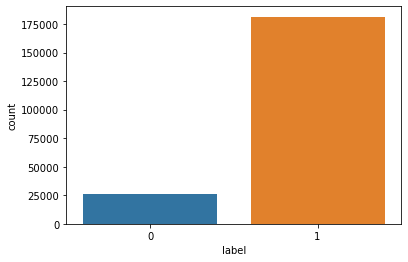
1. **Univariate Analysis for numerical columns:**

****

****

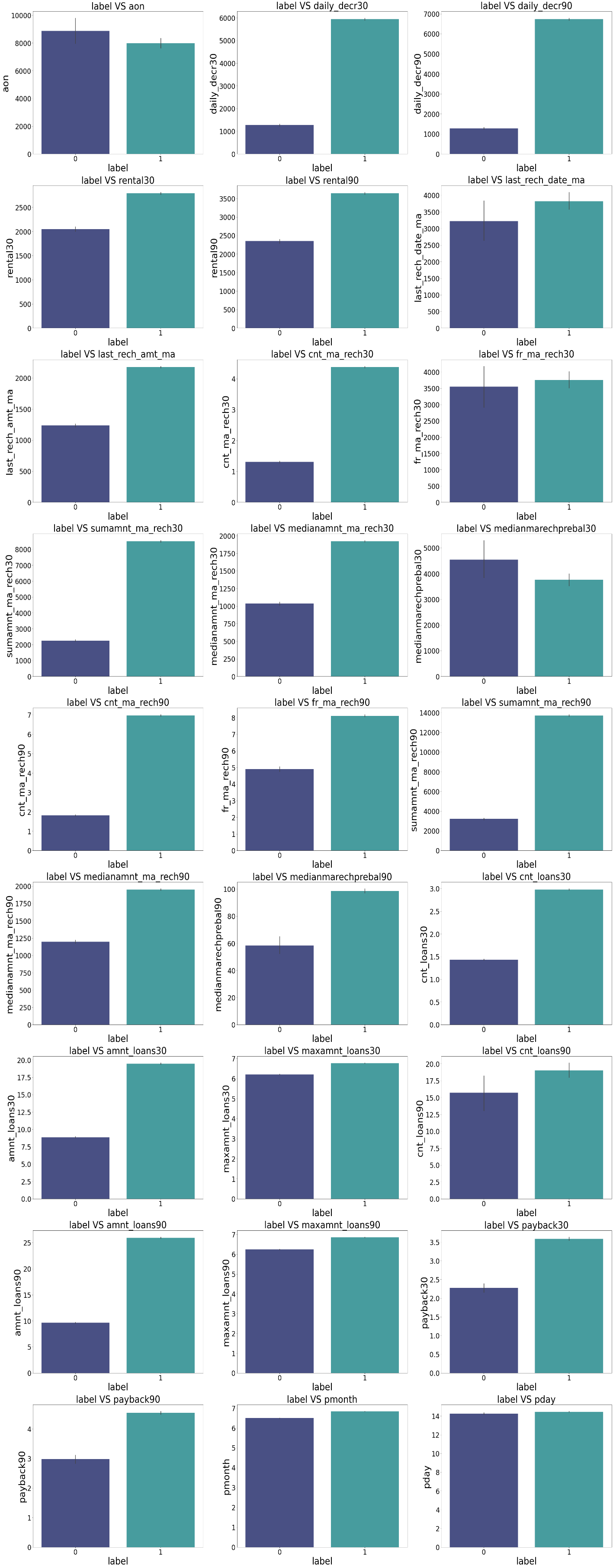
**Observations:**

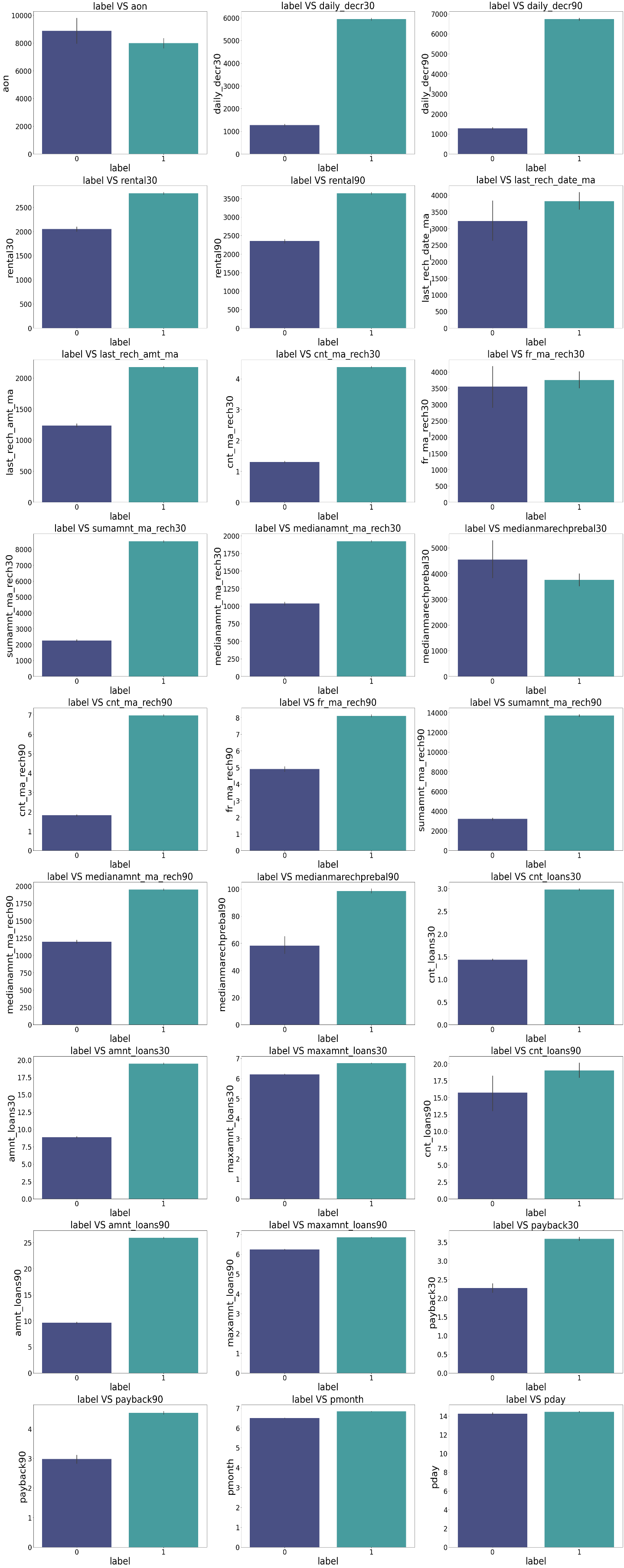
We can clearly see that there is skewness in most of the columns so we have to treat them using suitable methods.



There is a data misbalancing issue so we have to treat this by using oversampling or under sampling.

Bivariate analysis for numerical columns:

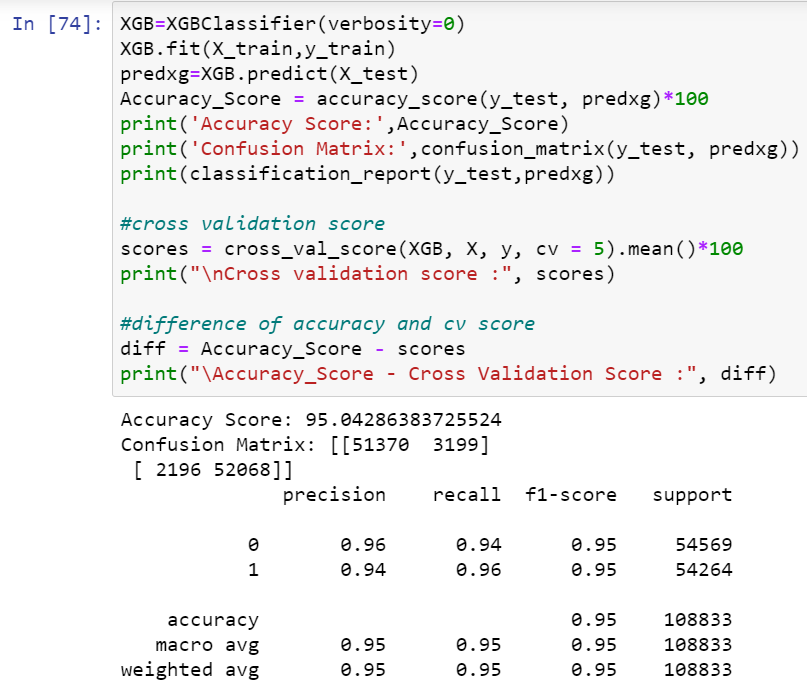
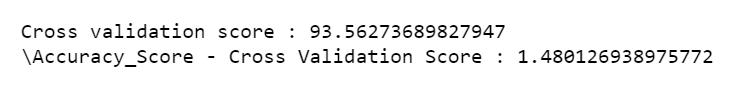
****

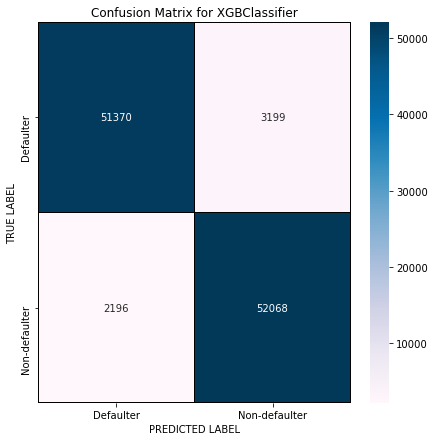
****

3.5 Run and Evaluate selected models

1. Model Building:

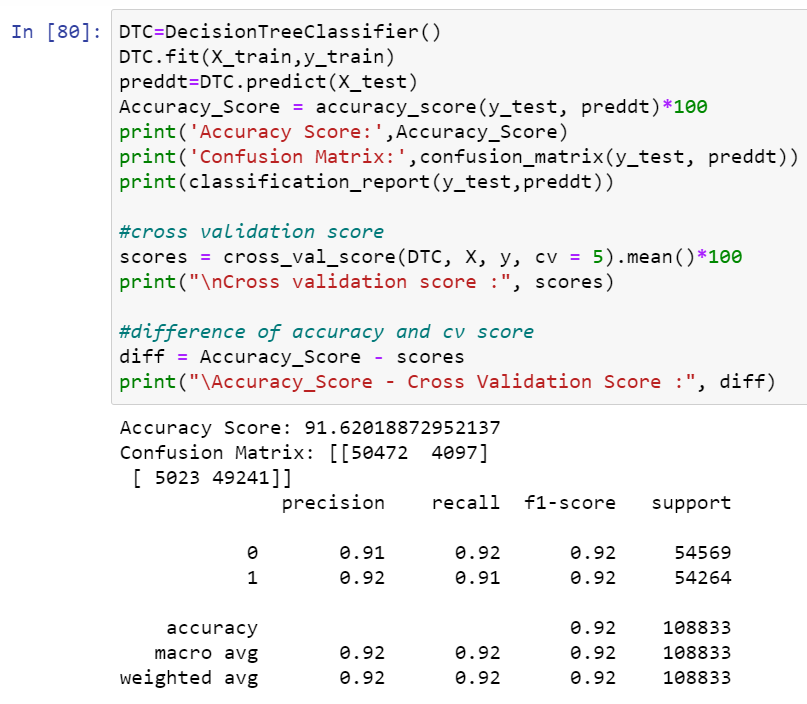
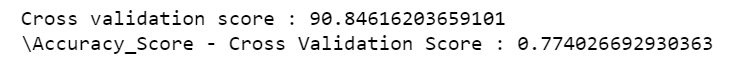
1. **XGBClassifier:**

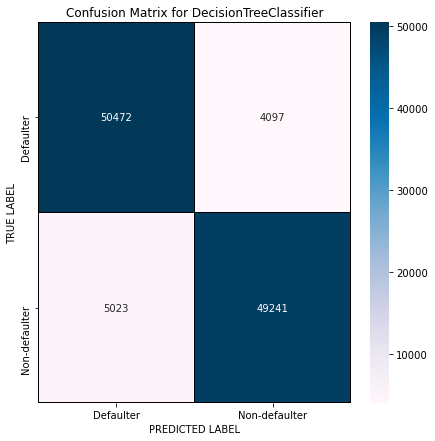
****

* **XGBClassifier** has given me 95% accuracy and the difference between model accuracy and cross validation score is 1.48%, but still we have to look into multiple models.

We can see the true values and predicted values in XGB Classifier model using confusion matrix.

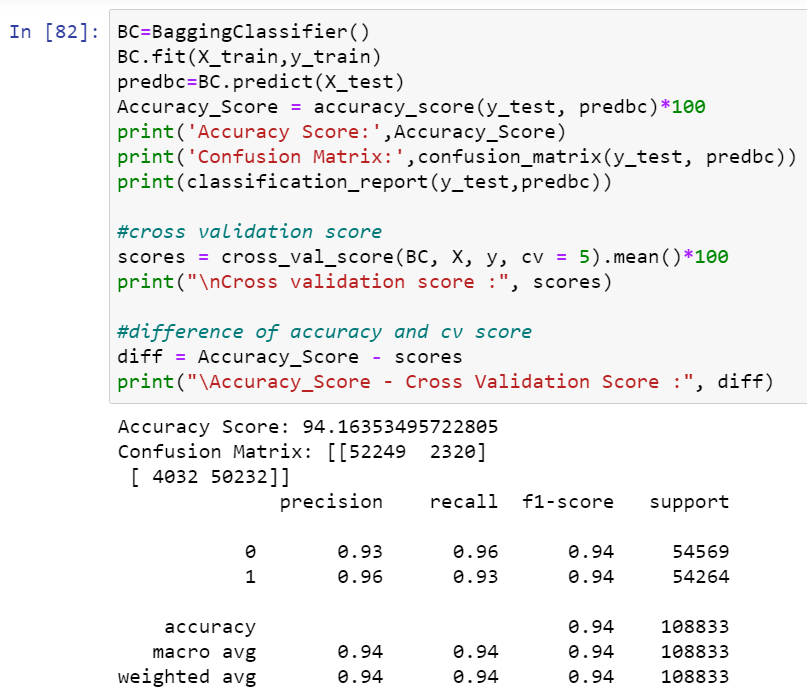
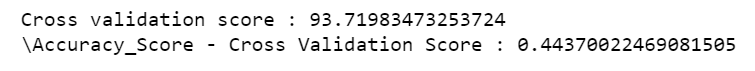
1. **DecisionTreeClassifier:**

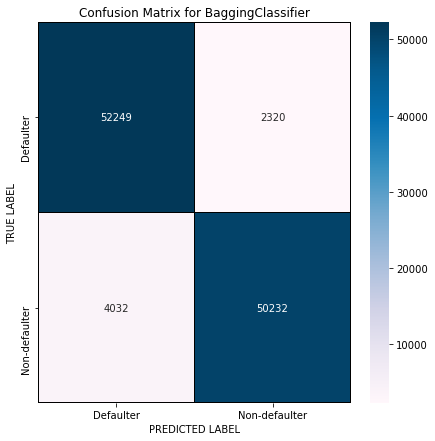


* **Decision Tree Classifier** is giving me 92% accuracy and the difference between model accuracy and cross validation score is 0.77%.

We can see the true values and predicted values in Decision Tree Classifier model using confusion matrix.

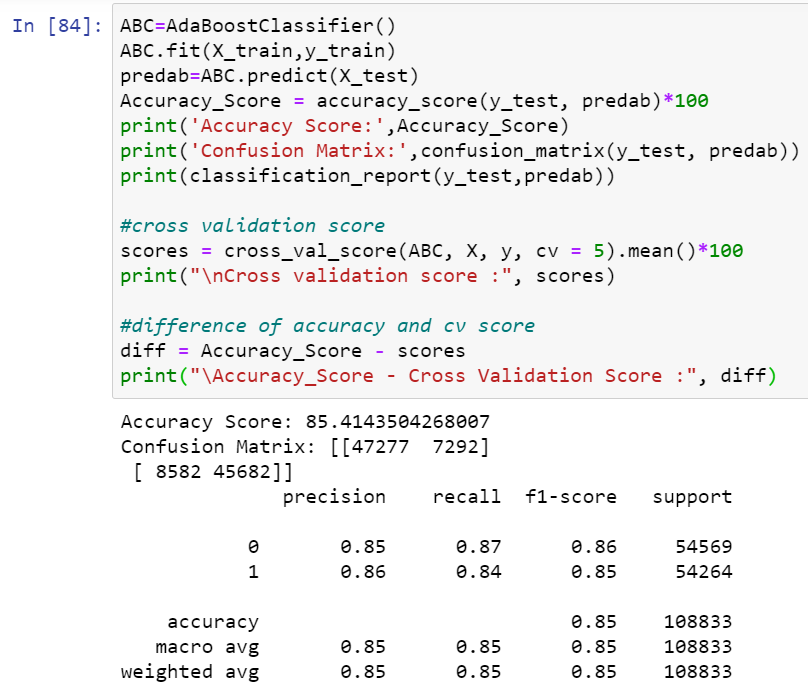
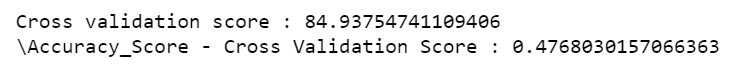
1. **Bagging Classifier:**

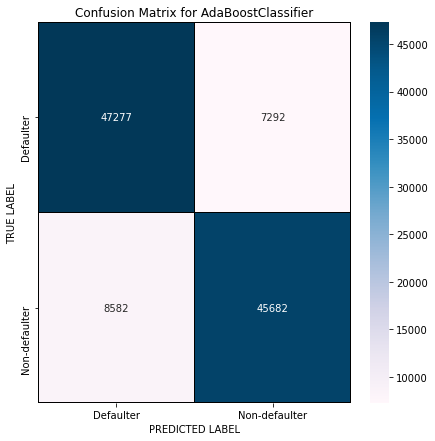
****

* **Bagging Classifier** is giving me 94% accuracy and the difference between model accuracy and cross validation score is 0.44%.

We can see the true values and predicted values in Bagging Classifier model using confusion matrix

1. **AdaBoost Classifier:**

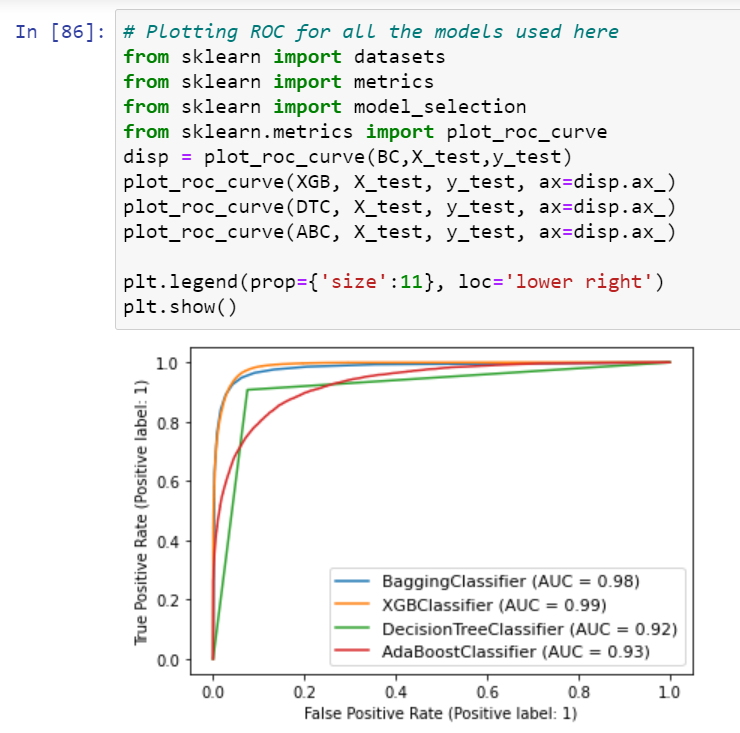
****

* **AdaBoost Classifier** is giving me 85% accuracy and the difference between model accuracy and cross validation score is 0.48%.

We can see the true values and predicted values in Adaboost Classifier model using confusion matrix.

**By looking into the difference of model accuracy and cross validation score i found Bagging Classifier as the best model with 95.16% accuracy and the difference between model accuracy and cross validation score is 0.44.**

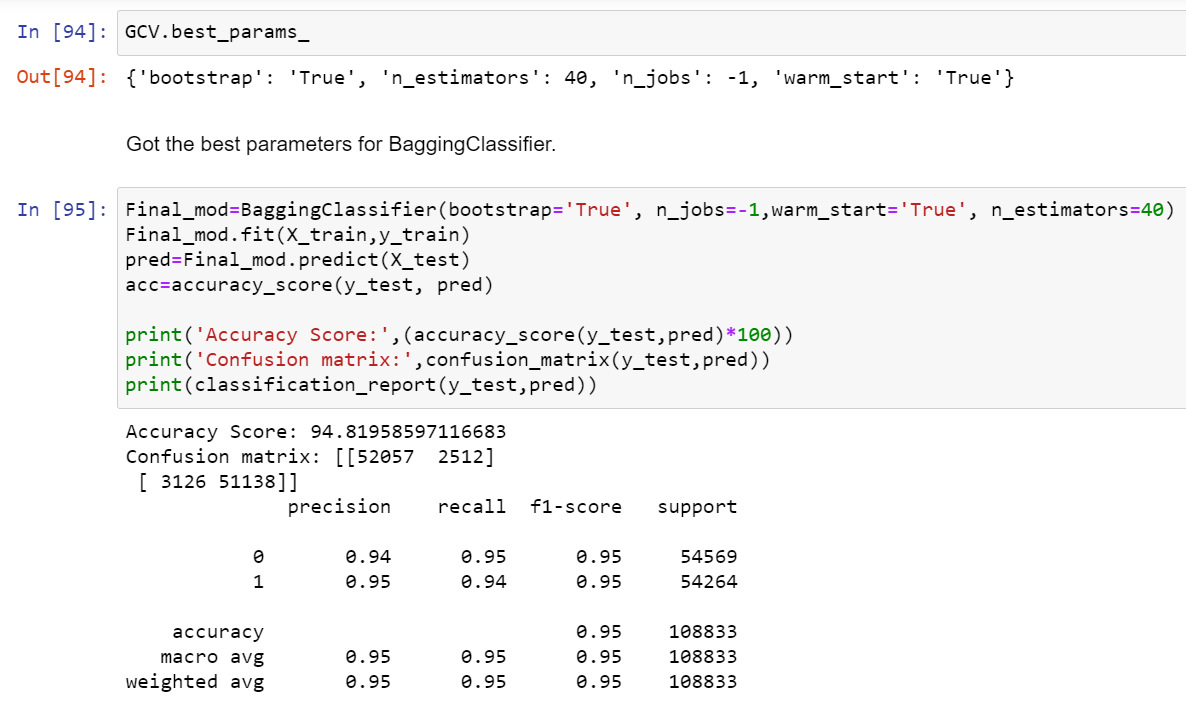
**2. ROC-AUC Curve:**

****

AUC value is high for XGB Classifier and Bagging Classifier.I got least difference in model accuracy and cross validation score for Bagging Classifier so BC is my best model.

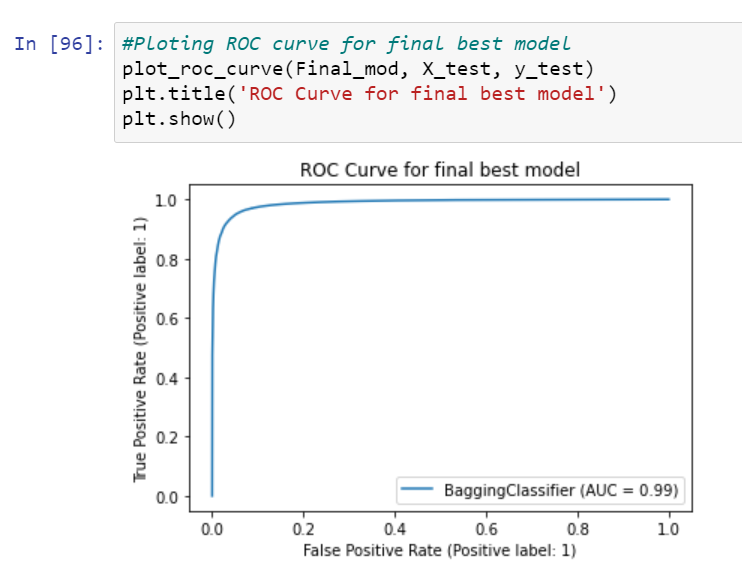
**3. Hyper Parameter Tunning:**

****

****

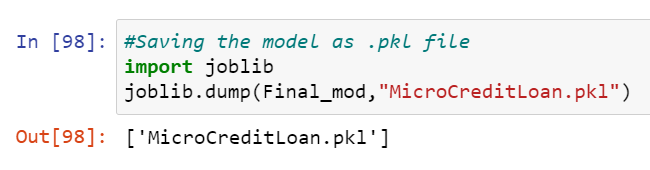
* I have choosed all parameters of BaggingClassifier, after tunning the model with best parameters I have incresed my model accuracy from 94.16% to 94.82%.

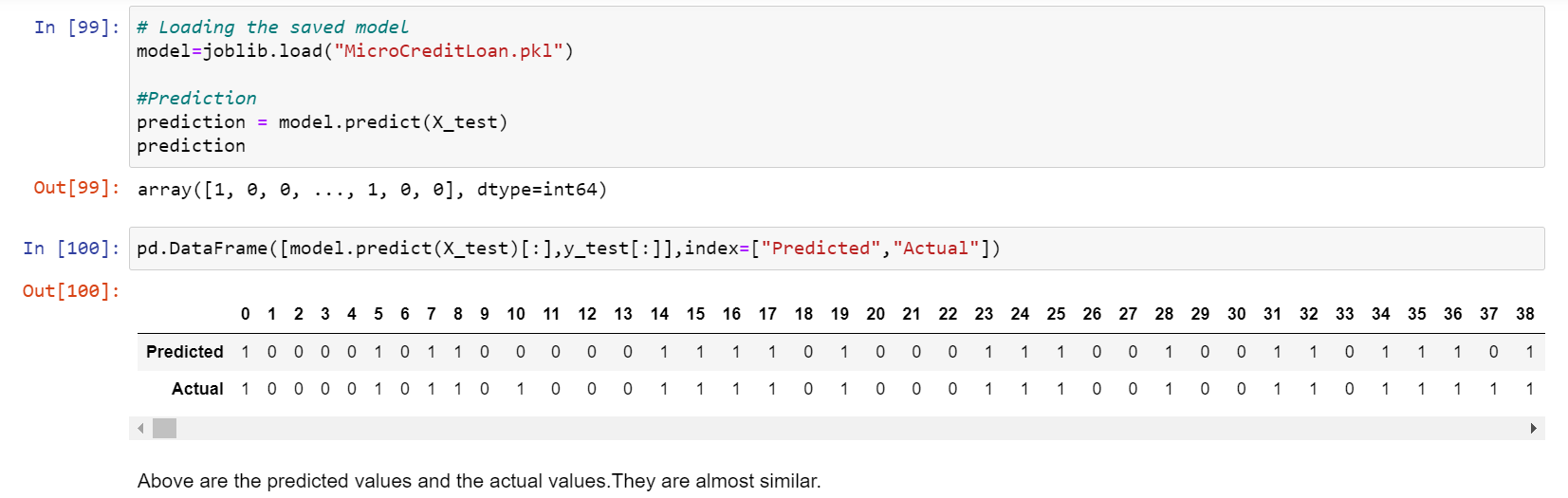
**ROC Curve for final model:**

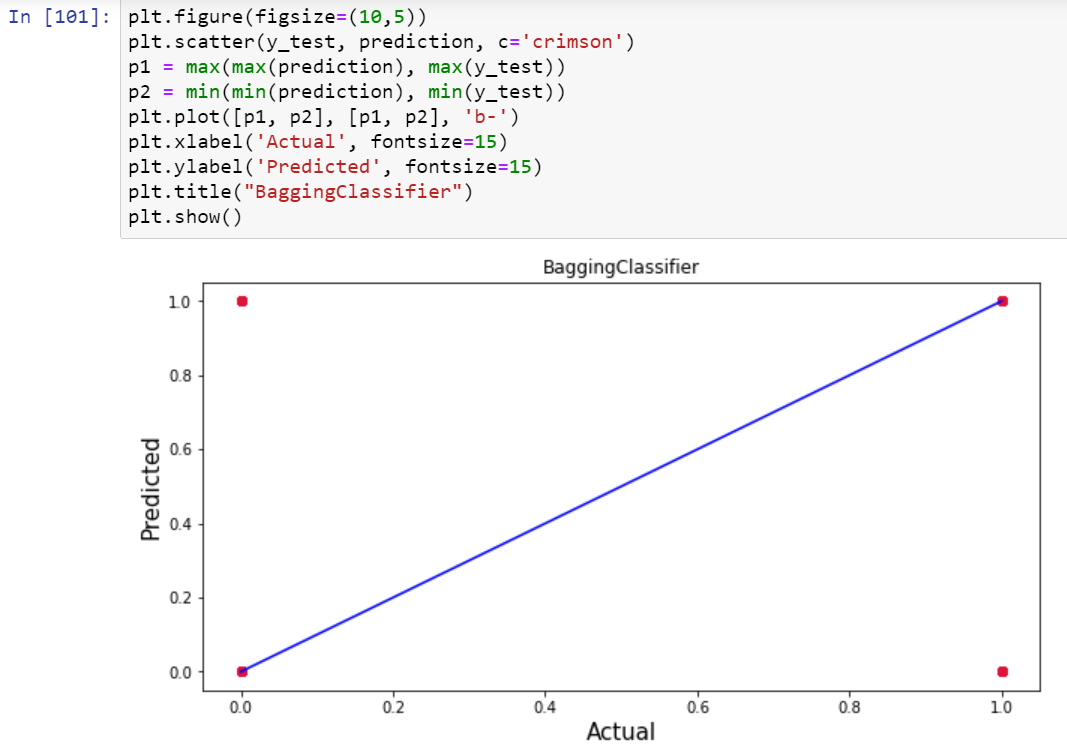
****

Great after hyperparameter tuning we got improvement in roc curve and AUC also.

1. **Saving the model and Predictions:**

* I have saved my best model using .pkl as follows**.**
* Now loading my saved model and predicting the test values.





Plotting Actual vs Predicted, To get better insight. Bule line is the actual line and red dots are the predicted values.

Interpretation of the Results:

* The dataset was very challenging to handle it had 37 features with 30days and 90days information of customers.
* Firstly, the datasets were not having any null values.
* But there was huge number of zero entries in maximum columns so we have to be careful while going through the statistical analysis of the datasets.
* And proper plotting for proper type of features will help us to get better insight on the data. I found maximum numerical columns in the dataset so I have chosen bar plot to see the relation between target and features.
* 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 dataset has a good impact like it will help the model not to get biased. Since we have not removed outliers and skewness completely from the dataset so we have to choose Normalization.
* We have to use multiple models while building model using dataset as to get the best model out of it.
* And we have to use multiple metrics like F1\_score, precision, recall and accuracy score which will help us to decide the best model.
* I found Bagging Classifier as the best model with 94.16% accuracy score. Also, I have improved the accuracy of the best model by running hyper parameter tunning.
* At last, I have predicted whether the loan is paid back or not using saved model. It was good!! that I was able to get the predictions near to actual values.

**4.CONCLUSION**

In this project report, we have used machine learning algorithms to predict the micro credit defaulters. 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 correlated to each other and are independent in nature. These feature set were then given as an input to four algorithms and a hyper parameter tunning was done to the best model and the accuracy has been improved. Hence, we calculated the performance of each model using different performance metrics and compared them based on these metrics. Then we have also saved the best model and predicted the label. It was good that the predicted and actual values were almost same.

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 unrealistic values and zero values. This study is an exploratory attempt to use four machine learning algorithms in estimating micro credit defaulter, and then compare their results.

To conclude, the application of machine learning in micro credit is still at an early stage. We hope this study has moved a small step ahead in providing some methodological and empirical contributions to crediting institutes, and presenting an alternative approach to the valuation of defaulters. Future direction of research may consider incorporating additional micro credit transaction data from a larger economical background with more features.