

**PROJECT REPORT ON:**

**“Micro-Credit Defaulter Model”**

**SUBMITTED BY**

**Ajit Madame**

**ACKNOWLEDGMENT**

I would like to express my special gratitude to “Flip Robo” team, who has given me this opportunity to deal with a beautiful dataset and it has helped me to improve my analyzation skills. And I want to express my huge gratitude to Ms.Gulshana Chaudhary (SME Flip Robo), she is the person who has helped me to get out of all the difficulties I faced while doing the project.

A huge thanks to my academic team “Data trained” who are the reason behind what I am today. Last but not least my parents who have been my backbone in every step of my life. And also thank you for many other persons who has helped me directly or indirectly to complete the project.

**Contents:**

1. **Introduction**
   * Business Problem Framing
   * Conceptual Background of the Domain Problem
   * Review of literature
   * Motivation for the Problem undertaken

**2.Analytical Problem Framing**

* + Mathematical/ Analytical Modelling of the Problem
  + Data Sources and their formats
  + Data Pre-processing Done
  + Data Input – Logic – Output Relationships
  + Hardware, Software and Tools Used

**3. Data Analysis and Visualization**

* + Univariate Visualization
  + Bivariate Visualizations

**4. Model Developments and Evaluation**

* + Features selections
  + The model algorithms used
  + Interpretation of the result
  + Hyperparameter tuning
  + ROC AUC Curve

**5. Conclusions**

* + Key Finding and conclusions
  + Limitation of this works and scope for future works

**1.INTRODUCTION**

1.1Business 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.

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.

* 1. 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 payed i.e. Non- defaulter, while, Label ‘0’ indicates that the loan has not been payed i.e. defaulter.

* 1. Review of Literature

Microfinance” is often seen as financial services for poor and low-income. In practice, the term is often used more narrowly to refer to loans and other services from providers

that identify themselves as “microfinance institutions” (MFIs)

Microfinance can also be described as a setup of a number of different operators focusing on the financially under-served people with the aim of satisfying their need for poverty alleviation, social promotion, emancipation, and inclusion.

Microfinance institutions reach and serve their target market in very innovative ways. Microfinance operations differ in principle, from the standard disciplines of general and entrepreneurial finance.

An attempt has been made in this report to review the available literature in the area of microfinance. Approaches to microfinance, issues related to measuring social impact versus profitability of MFIs, issue of sustainability, variables impacting sustainability, effect of regulations of profitability and impact assessment of MFIs have been summarized in the below report. We hope that the below report of literature will provide a platform for further research and help the industry to combine theory and practice to take microfinance forward and contribute to alleviating the poor from poverty.

* 1. Motivation for the Problem Undertaken

I have to model the micro credit defaulters with the available independent variables. This model will then be used by the management to understand how the customer is considered as defaulter or non-defaulter based on the independent variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand whether the customer will be paying back the loaned amount within 5 days of insurance of loan. The **relationship between predicting defaulter and the economy** is an important motivating factor for predicting micro credit defaulter model.

**2.Analytical Problem Framing**

* 1. Mathematical/ Analytical Modelling of the Problem

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 payed i.e. Non- defaulter, while, Label ‘0’ indicates that the loan has not been payed i.e. defaulter.

There were 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. Also, I found that, some features are having so many 0 entries so, I replace 0 with suitable values for better functioning. 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 and I removed skewness using yeo-johnson method. I have used 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 dataset is provided by Flip Robo which is in the format csv. This dataset give use for 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 Indonasian Rupiah)

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

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

2.3 Data Pre-processing Done

* 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, nunique, value counts, info etc.
* Then I did value count, in this I found that few features are having more 90% 0 values so avoiding biasness I simply drop this feature.
* I found that, few features are containing 0 value so replace it with mean value.
* I have also dropped 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.

2.4 Data Inputs - Logic - Output Relationships

The Datasets consist mainly of float and int data type variables and a only one feature contain object data type variable. The relationships between the independent variables and dependent variable were analyzed.

All variables are numerical I have plotted dist. plot to see the distribution of each column data.

2.5 Hardware, Software and Tool Used

**Hardware Used:**

Processor – Intel core i3

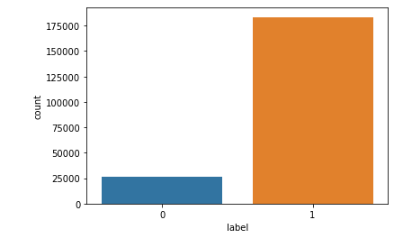
Physical Memory – 8 GB

**Software Used:**

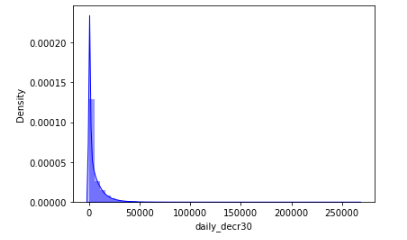
* Windows 10 Operating System
* Anaconda Package and Environment Manager
* Jupyter Notebook
* Python Libraries used: In Which Pandas, Seaborn, Matplotlib, Numpy and Scipy
* sklearn for Modelling Machine learning algorithms, Data Encoding, Evaluation metrics, Data Transformation, Data Scaling, Component analysis, Feature selection etc.

**3.Data Analysis and Visualization**

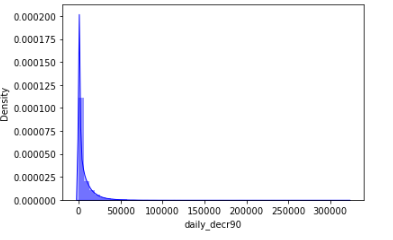
3.1 Univariate Visualization



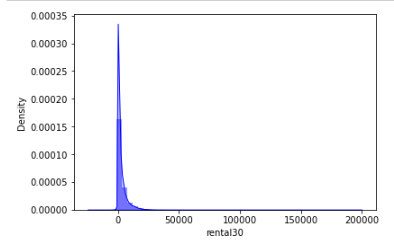
* We can see, label is imbalanced so we need to balanced it.
* So, I will be balanced it in data balancing stage.



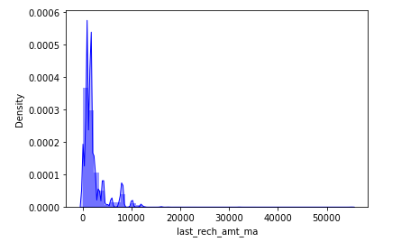
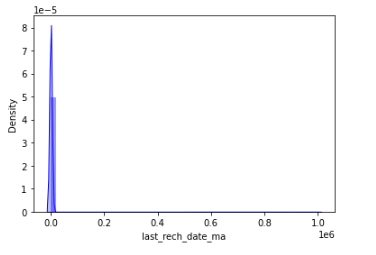
* daily\_decr30 is not normally distributed. It contains some outliers.
* daily\_decr30 is positively or right skewed.



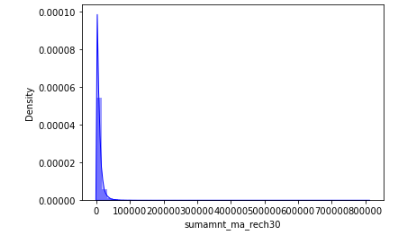
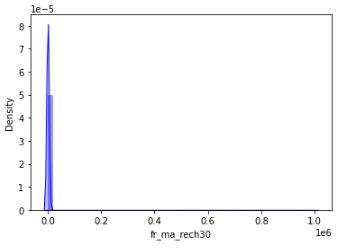
* Daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah) is not normally distributed.
* Maximum daily amount spent in in the range of 0 to 50000.



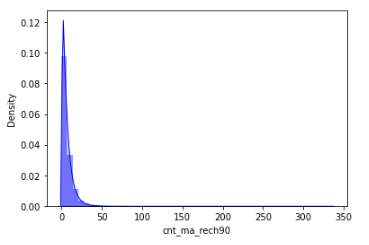
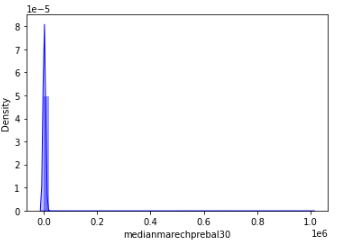
* Average main account balance over last 30 days i.e. rental 30 is in the range of 0 to 50000. maximum rental30 is lies below the 25000.
* rental is not a normally distributed.



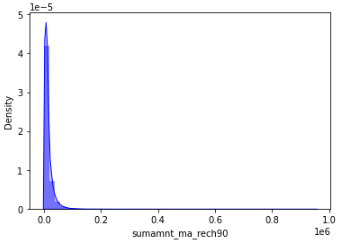
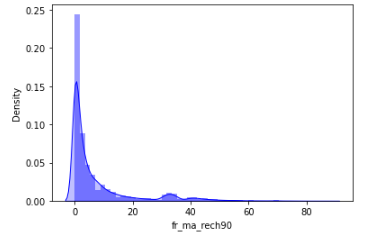
* Maximum Number of days till last recharge of main account lie in the range of 0 to 0.2. But mostly the data are containing 0.0.
* Amount of last recharge of main account is in the range of 0 to 10000. But Maximum data is lies in below 6000. Data is not normally distributed. It is right or positively skewed.



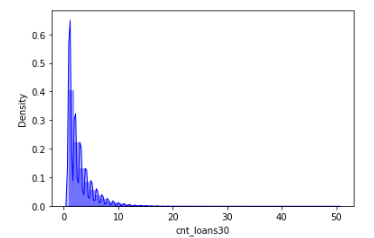
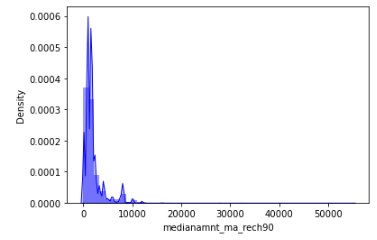
* We can see, almost to time Frequency of main account recharged in last 30 days.
* Less than 50000 time the Total amount of recharge in main account over last 30 days is happened.



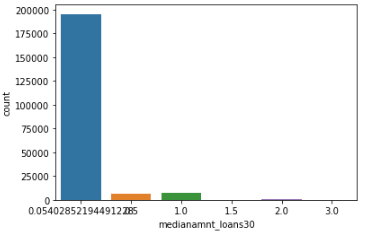
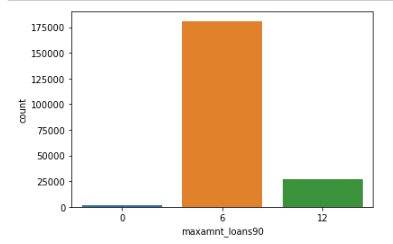
* 0 is a Median of main account balance just before recharge in last 30 days at user level. We can see, 0 is median where the users are doing recharge in last 30 days.
* We can see less 30 Number of times main account got recharged in last 90 days. It has very high count than others. It shows positively skewed.



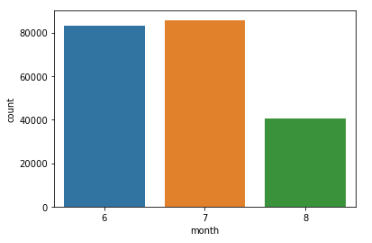
* We can see the users are doing less than 20 time the Frequency of main account recharged in last 90 days. it shows that users are not interested.
* We can see. less than 1.5 Total amount of recharge in main account over last 90 days.



* Less than 10000 Median of number of recharges done in main account over last 90 days at user level.
* We can see, users are taking less than 10 Number of loans taken by user in last 30 days. I show that users are taking less than 10 loans but mostly the users are taking up to 5 loans in last 30 days.

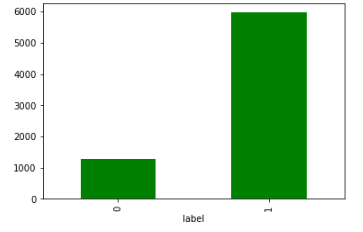
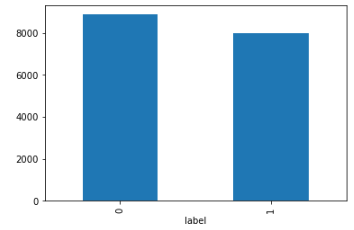


* We can see user are taking loans are less than 6 times followed by 12. 0 It show that, these users are not taking any loan.
* We can see 0 is a Median of amounts of loan taken by the user in last 30 days.

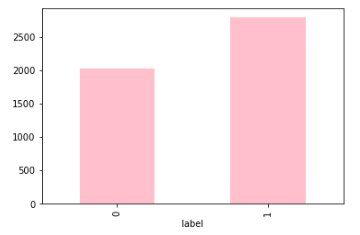
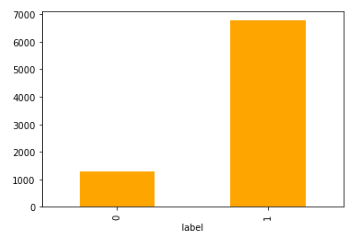


* We can see in 7-month maximum users are taking loan than others followed by 6 months.

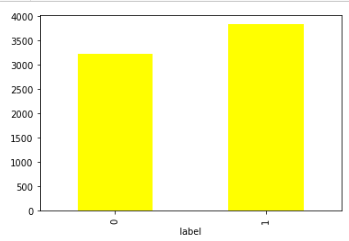
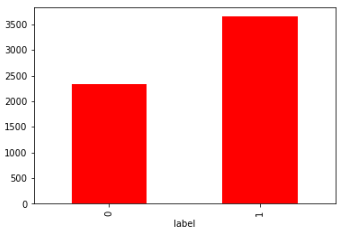
3.2 Bivariate Visualization



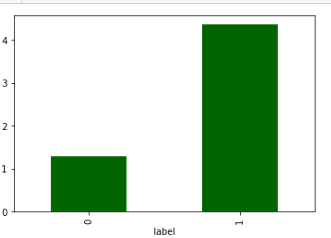
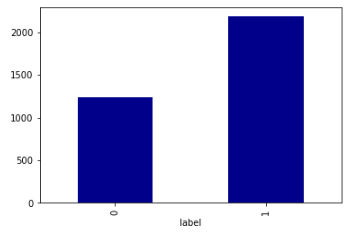
* Customers with high value of Age on cellular network in days (aon) are maximum defaulters (who have not paid their loan amount-0).
* Customers with high value of Daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah) (daily\_decr30) are maximum Non-defaulters (who have paid their loan amount-1).



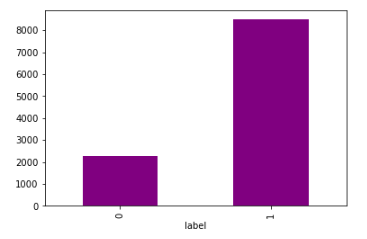
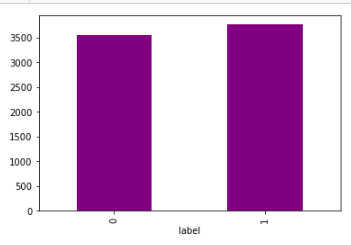
* Customers with high value of Daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah) (daily\_decr90) are maximum Non-defaulters (who have paid their loan amount-1).
* Customers with high value of Average main account balance over last 30 days(rental30) are maximum Non-defaulters (who have paid their loan amount-1).



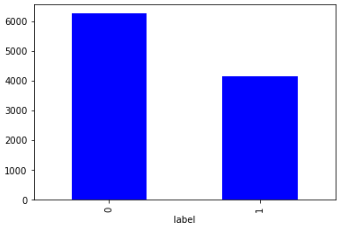
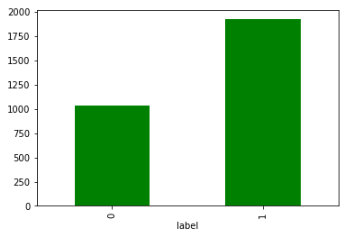
* Customers with high value of Average main account balance over last 90 days(rental90) are maximum Non-defaulters (who have paid their loan amount-1).
* Customers with high Number of days till last recharge of main account (last\_rech\_date\_ma) are maximum Non-defaulters (who have paid their loan amount-1).



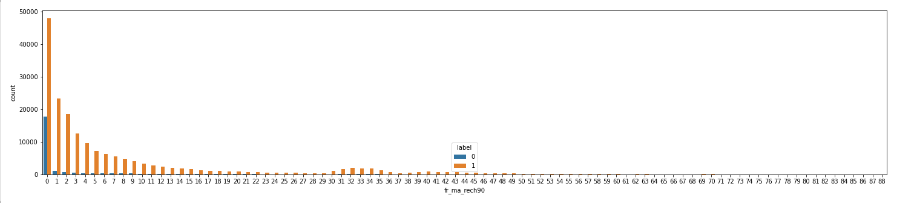
* Customers with high value of Amount of last recharge of main account (in Indonesian Rupiah) (last\_rech\_amt\_ma) are maximum Non defaulters (who have paid there loan amount-1).
* Customers with high value of Number of times main account got recharged in last 30 days(cnt\_ma\_rech30) are maximum value count 1. It means the there is less loan defaulter.



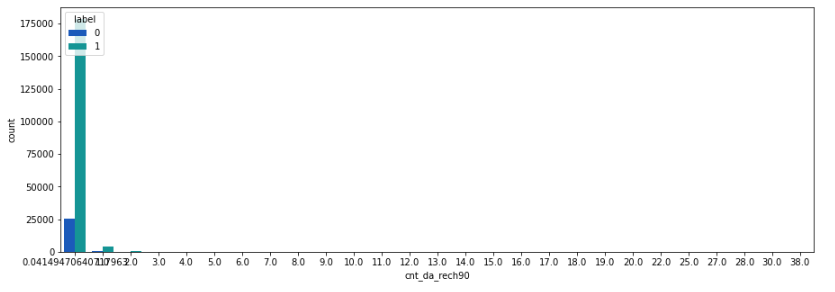
* We can see, Frequency of main account recharged in last 30 days are near to each other’s. In this who's are not defaulting loan are higher than defaulting loan. But who’s are defaulting loan being also high in count but less than not defaulting.
* We can see, maximum customer doing their recharge at time but Total amount of recharge in main account over last 30 days (in Indonesian Rupiah) of defaulting loan is less than not defaulting loan.



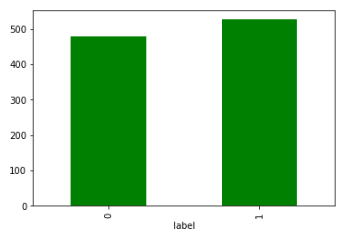
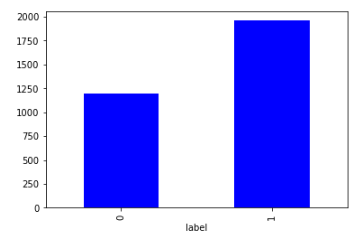
* Customers with high value of Median of number of recharges done in main account over last 30 days at user level (in Indonesian Rupiah) (medianamnt\_ma\_rech30) are maximum Non-defaulters (who have paid their loan amount-1).
* Median of main account balance just before recharge in last 30 days at user level (in Indonesian Rupiah) is high in whose are loan defaulting. It is less in those are not defaulting the loan.



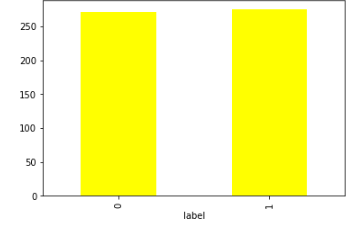
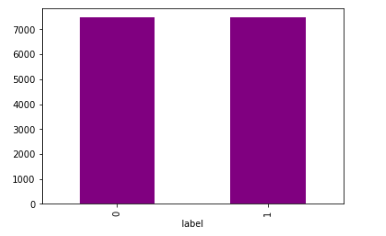
* We can see there are less Frequency of main account recharged in last 90 days who’s defaulting the loan.



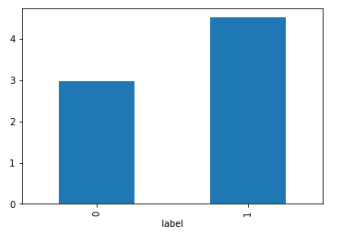
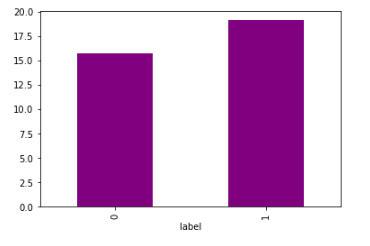
* We can see, there are less defaulter who’s having a smaller Number of times data account got recharged in last 30 days.



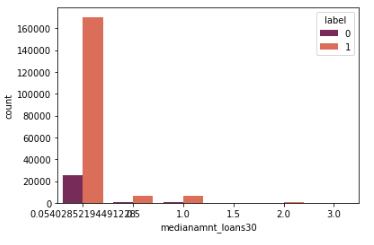
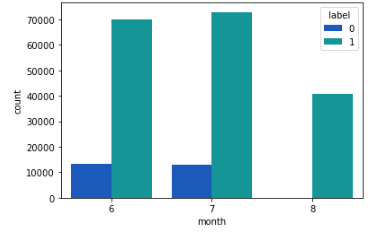
* Median of number of recharges done in main account over last 90 days at user level (in Indonesian Rupiah), in this feature there are less defaulting loan. But up to 1175 the defaulting loan count.
* Number of times data account got recharged in last 30 days(cnt\_dat\_rech30) has high loan not deflater than defaulting loan. But Who's are not defaulting loan it also in high count but little less than not defaulting count.



* We can see, Frequency of data account recharged in last 30 days(fr\_da\_rech30) are almost same for both defaulting loan and not defaulting loan.
* We can see maximum amount of loan taken by the user in last 30 days are almost same for both defaulting the loan and not defaulting the loan. It means that users are taking maximum loan and not paid the loan amount or defaulting the loan.

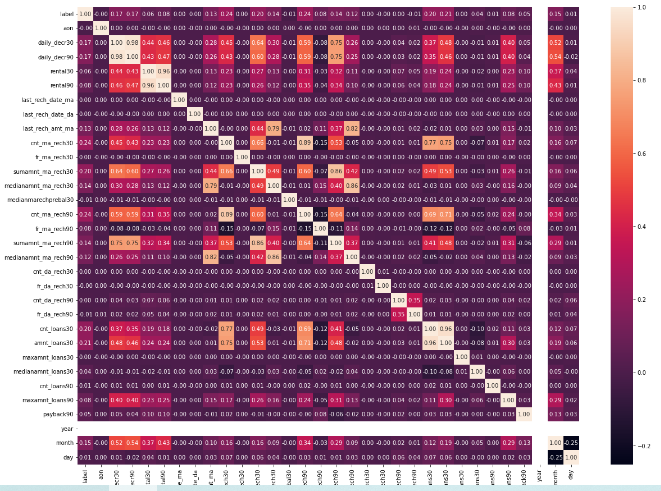


* Number of loans taken by user in last 90 days is high and it also high in loan are not defaulting. But loan defaulting little lower than non-defaulting users.
* Customers with high value of Average payback time in days over last 90 days(payback90) are maximum Non-defaulters (who have paid their loan amount-1).



* We can see, in 7- month customers are taking loan is high than 6 month and it loan defaulting also high in 7-month than 6 - month. But in 8 Month non defaulting loan are not present. it means in this month customer are not defaulting the loan.
* Median of amounts of loan taken by the user in last 30 days is high in non-defaulting loan. Here are only 6 categories are given in which 0.05 is having high non defaulting customers.

**Correlation of the features with target column**



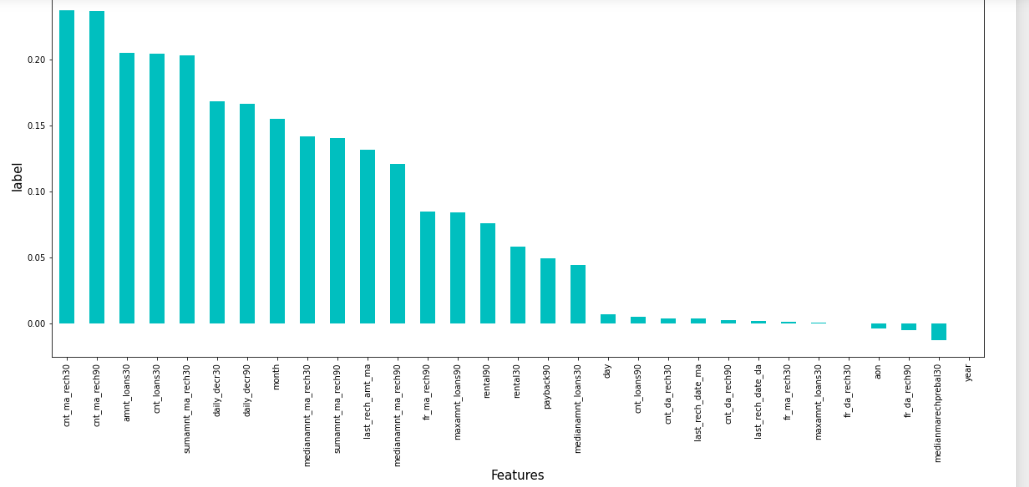
**Outcome of Correlation**

1. daily\_decr30 has 17% positive correlation with target column.
2. cnt\_ma\_rech30 has 24% positive correlation with target column.
3. cnt\_ma\_rech90 has 24% positive correlation with target column.
4. cnt\_loans30 has 20% positive correlation with target column.
5. amt\_loan30 has 21% positive correlation with target columns.
6. month has 15% positive correlation with target column.

**Features correlation to each other**

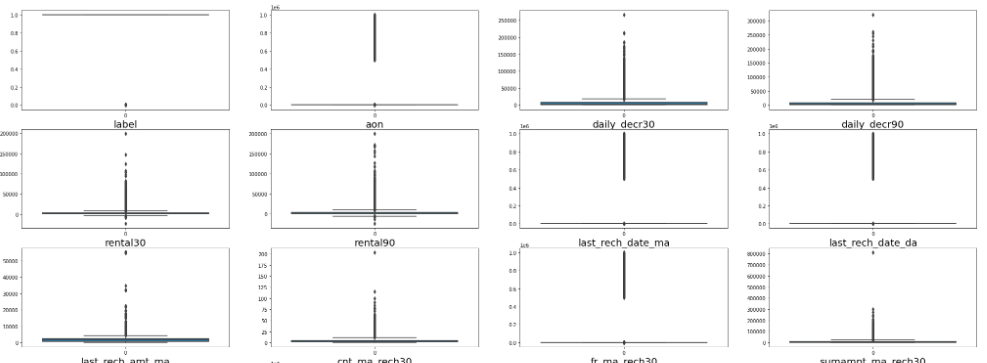
1. daily\_decr30 have 98%, 75% positive correlation with daily\_decr90 and sumamnat\_ma\_rech90.
2. rental90 has 98% positive correlation with rental30.
3. last\_rech\_amt\_ma has 82% and 79% positive correlation with medianamnt\_ma\_rech90 and medianamnt\_ma\_rech30.
4. cnt\_ma\_rech30 has 89% positive correlation with cnt\_ma\_rech90.
5. cnt\_loan30 has 77% positive correlation with cnt\_ma\_rech30.
6. sumamnt\_ma\_rech30 has 86% positive correlation with sumamnt\_ma\_rech90.
7. medianamnt\_ma\_rech90 has 86% positive correlation with medianamnt\_ma\_rech30.

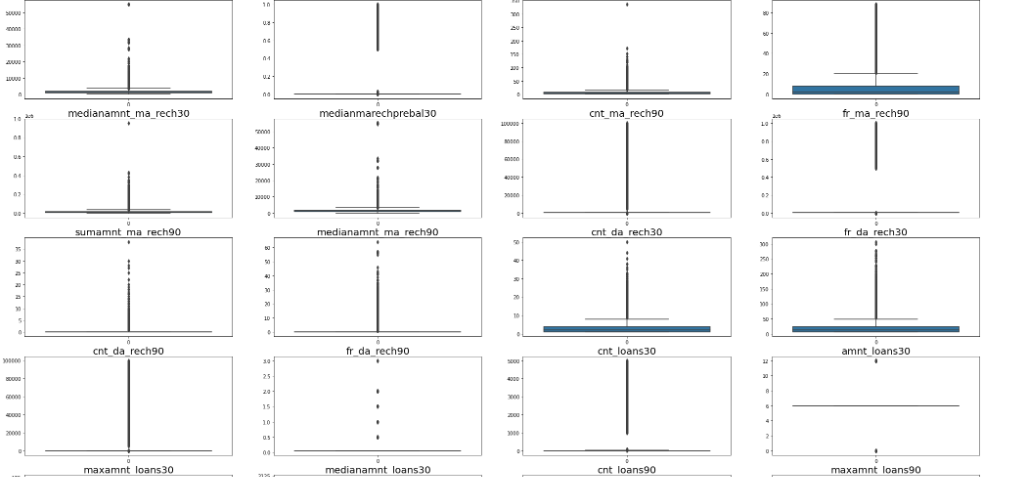
**Visualizing correlation of feature columns with label column**

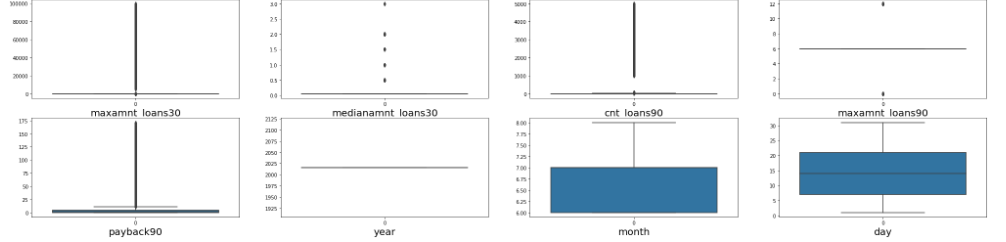


* It is observed that cnt\_ma\_rech30, cnt\_ma\_rech90, amnt\_loans30 and sumamnt\_ma\_rech30 have the highest positive correlation with label.
* year, medianmarechprebal30 and fr\_do\_rech90 have the highest negative correlation with label.

**Checking Outliers**

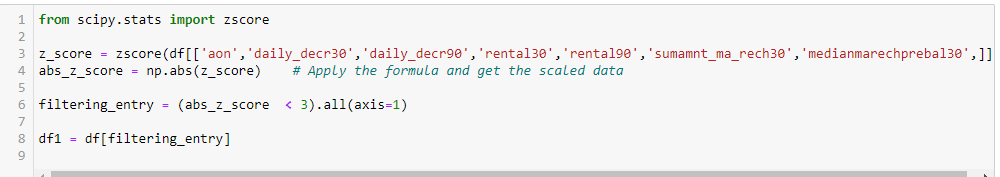
****

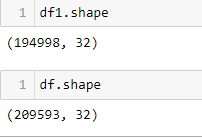
****

****

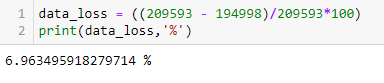
* We can see almost all features are having outliers except year, month, day.
* But I am only removing outliers only those features are having continuous data.

**Removing Outliers**

****



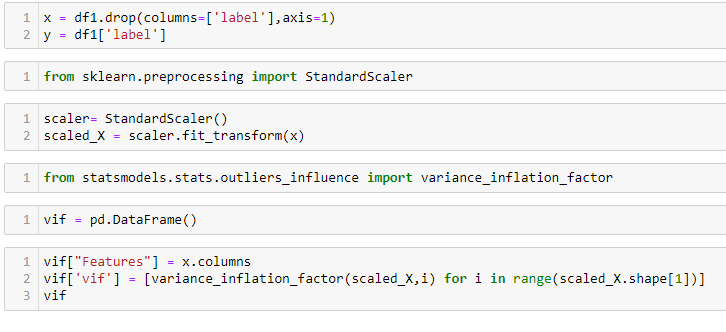
**Percentage of data loss**

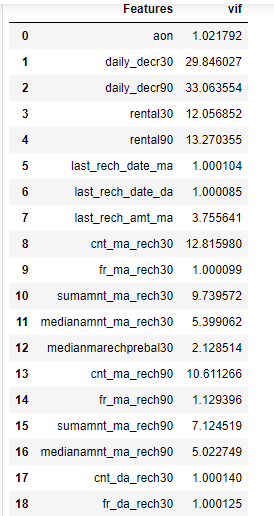
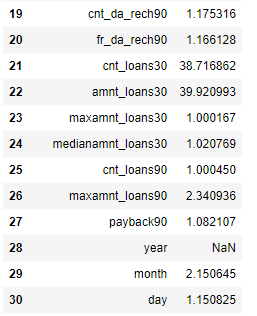


4. Models Development and Evaluation

4.1 Features Selection

Features were first checked for presence of multicollinearity and based on the respective ANOVA f-score values, the feature columns were selected that would best predict the Target variable, to train and test machine learning models.



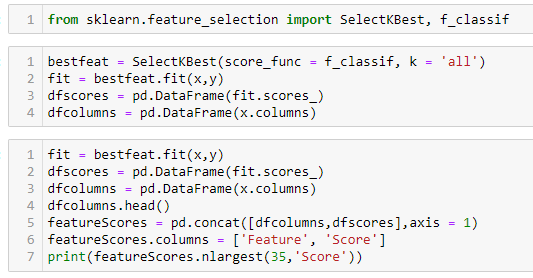
 

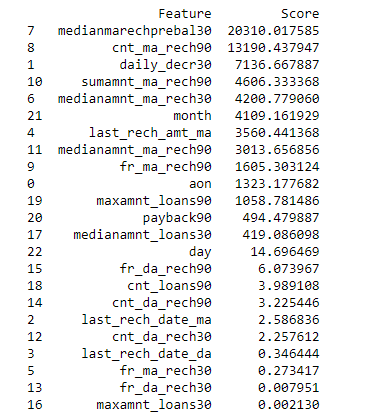
* If VIF > 10, It means multicollinearity is present.
* Multicollinearity are existing in amnt\_loans30, cnt\_loans30, cnt\_ma\_rech90, daily\_decr90, daily\_decr30, sumamnt\_ma\_rech30, nt\_ma\_rech30, rental90, rental30, daily\_decr90 and daily\_decr30, Based on ANOVA F scores.

**Dropping columns whose having multicollinearity**

****

**Selecting K Best Features**

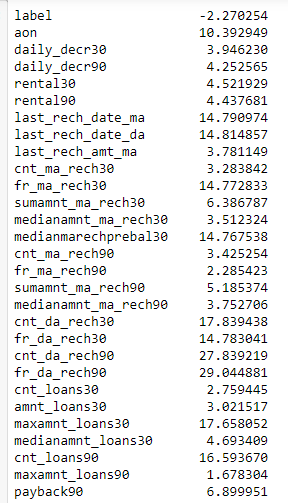
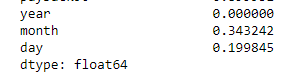




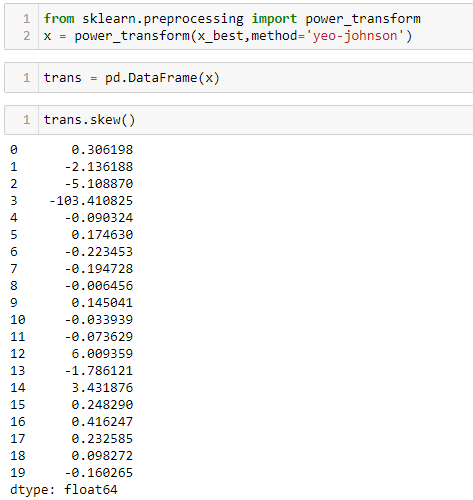
**Selecting best features based on their scores**

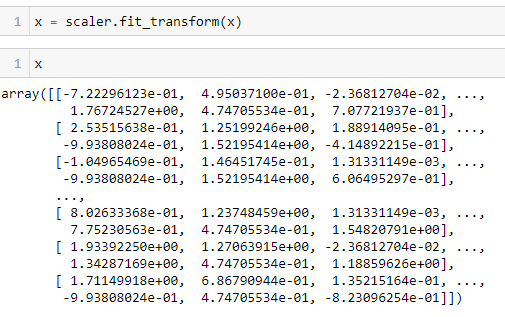
****

**Data Skewness Before removing**

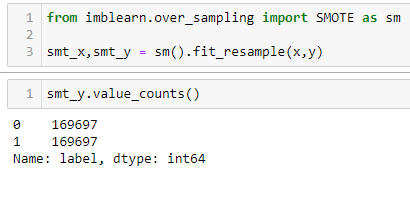
 

**Data After Removing Skewness**





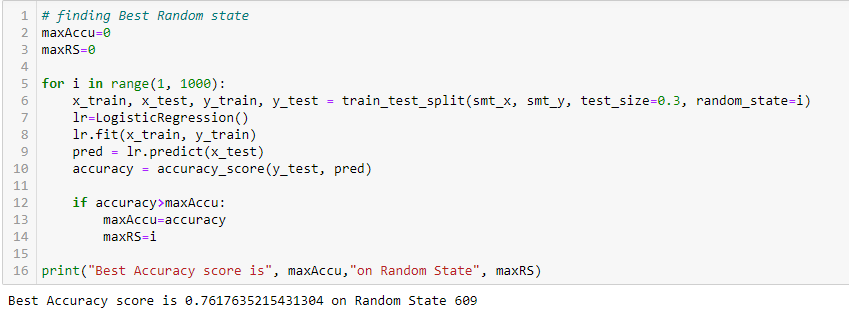
**Balancing Dataset**



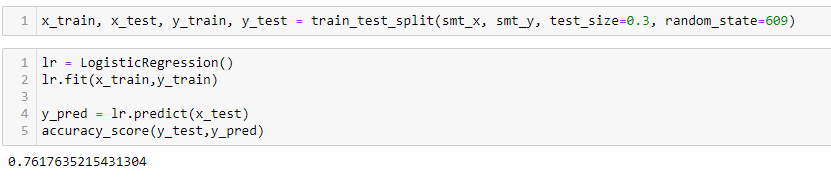
4.2 The model algorithms used

**Logistic Regression:**

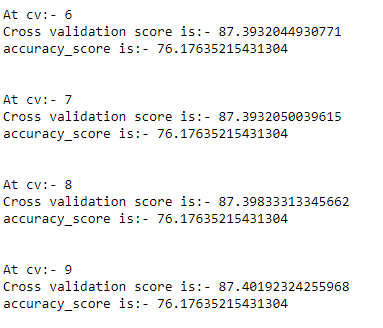
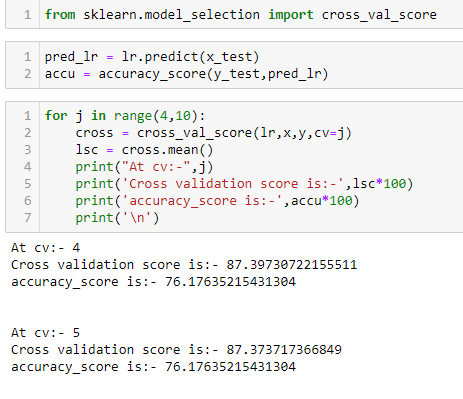
**Finding Best Random State**

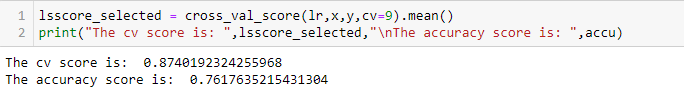


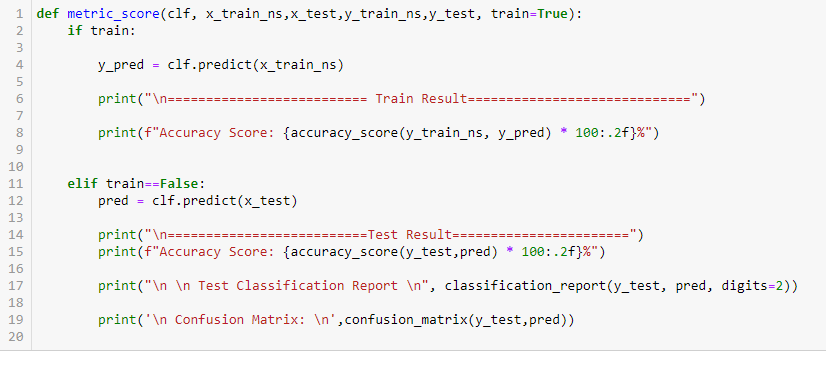
**Train and Test**

****

**Cross Validation of the model**

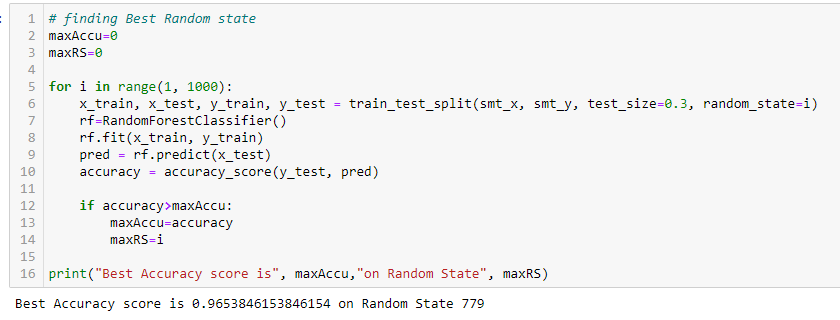
****

****

****

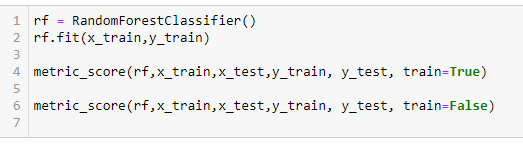
**Random Forest Classifier:**

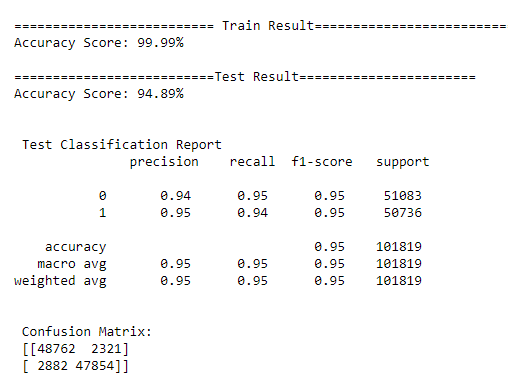
**Finding Best Random State**

****

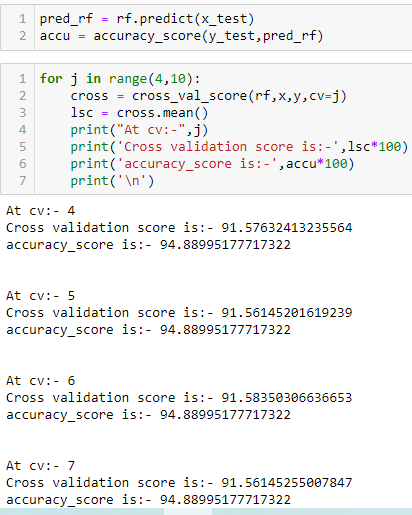
**Train and Test**

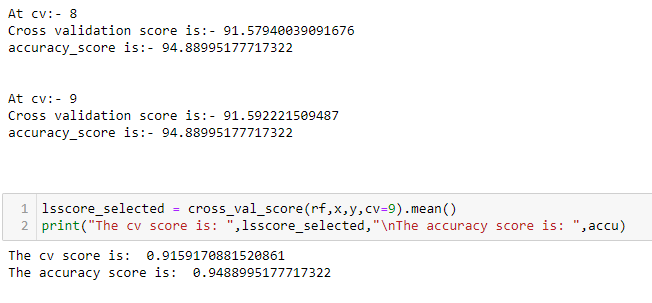
****

****

****

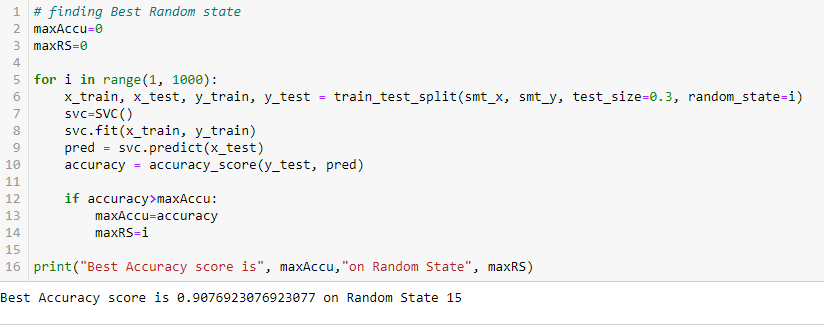
**Cross Validation of the model**

****

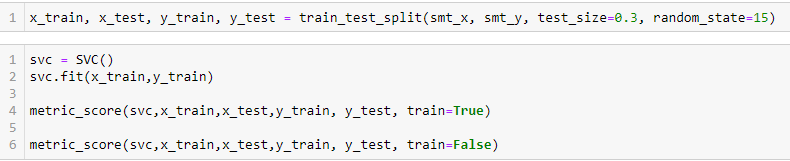
****

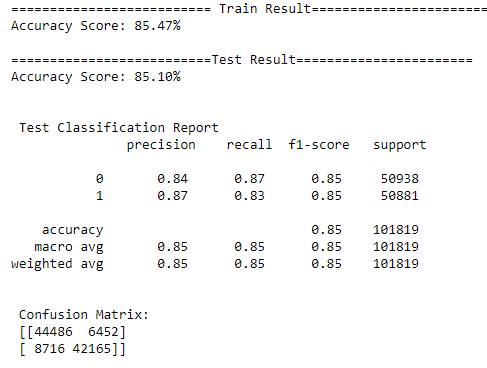
**SVC**

**Finding the Best Random State**

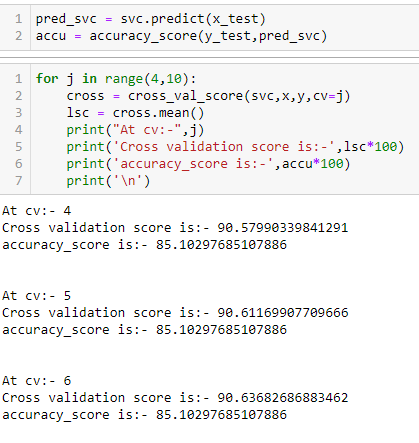
****

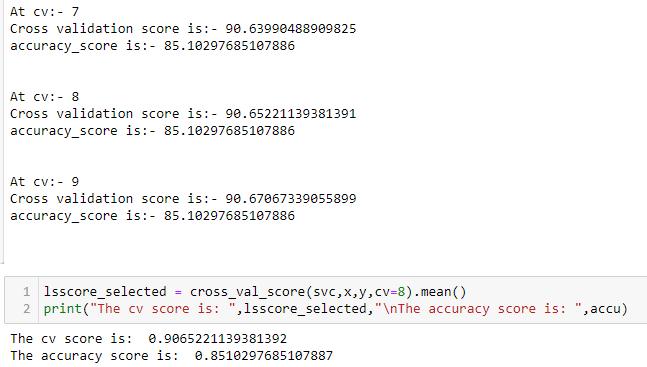
**Train and Test**

****

****

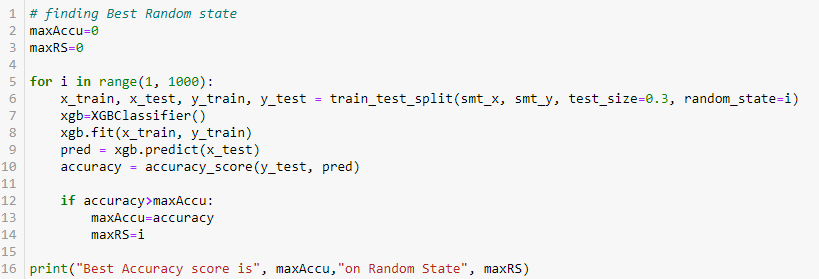
**Cross Validation of the model**

****

****

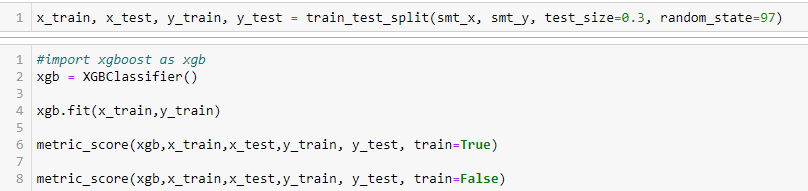
**XGBoost Classifier:**

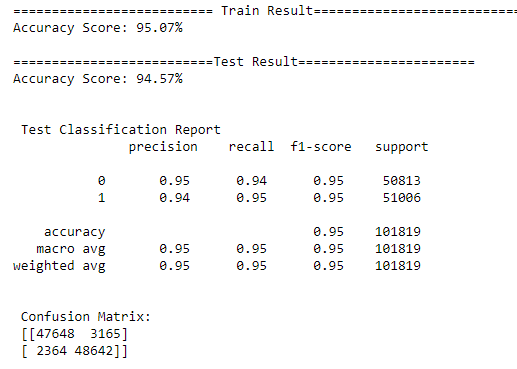
**Finding the Best Random State**

****

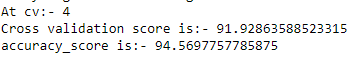
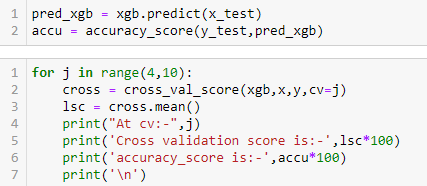
****

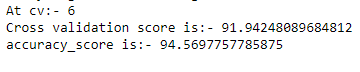
**Train and Test**

****

****

**Cross Validation of the Model**

****

****

****

****

****

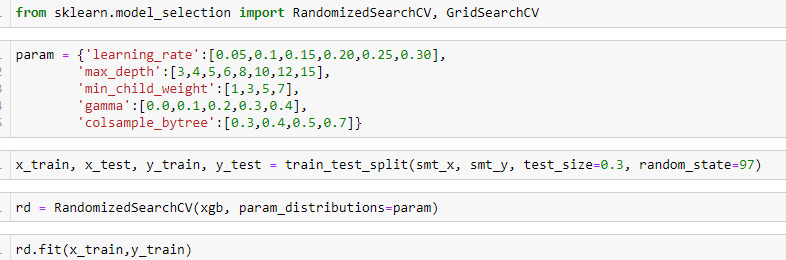
****

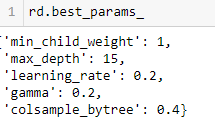
****

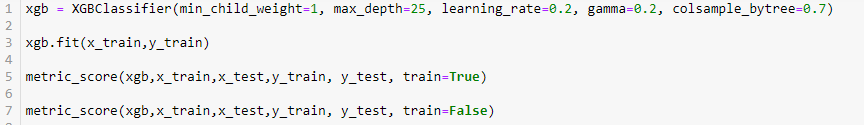
4.3 Interpretation of the results

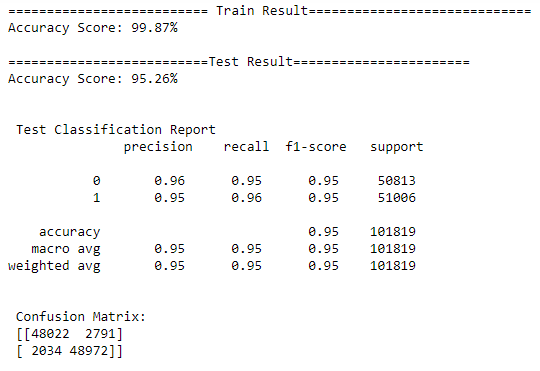
Based on comparing Accuracy Score results with Cross Validation results, it is determined that XGBoost Classifier is the best model.

4.4 Hyperparameter Tuning

****

****

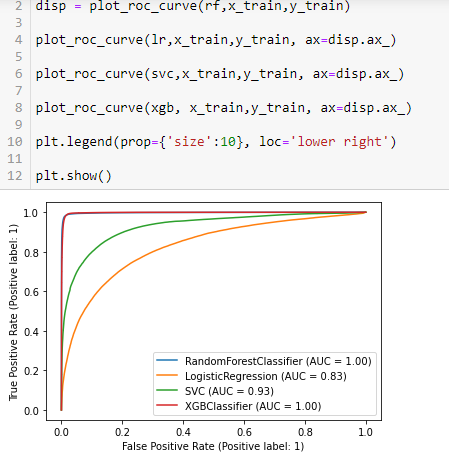
****

****

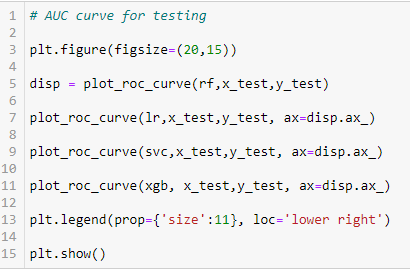
Based on the input parameter values and after fitting the train datasets The XGBoost Classifier model was further tuned based on the parameter values yielded from RandomizedSearchCV. The XGBoost Classifier model displayed an accuracy of 95.26%.

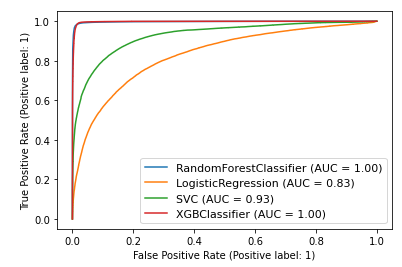
4.5 ROC AUC Curve

**AUC Curve for training**

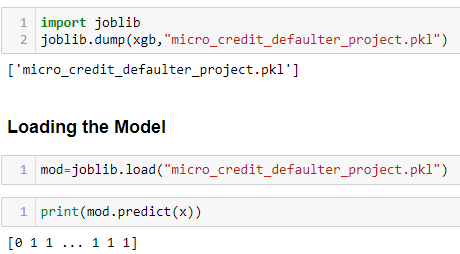
****

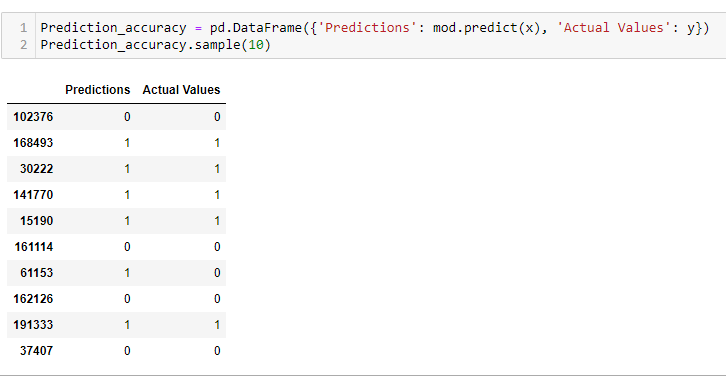
**AUC Curve for Testing**

****

****

**The Model Save and Testing**

****

****

**5. Conclusions**

5.1 Key Finding and Conclusions

Once a loan is written off, it is no longer on the books of the lender. However, the MFI’s decision to write-off the loan is often driven by prudential accounting or regulatory requirements and is not necessarily a signal that the debt has been forgiven. Therefore, the lender may continue to attempt to recover it, and so the borrower is not free from the prospect of a collections call or visit. The borrower may be barred from future borrowing from the same MFI, or (if reported to a credit bureau) other MFIs. At the microfinance level, bankruptcy is not an option. And, in most developing countries debt counseling or rehabilitation services are not available to assist defaulters.

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 the predicted and actual values were almost same.

5.2 Limitation of this works and scope for future works

* First drawback is the length of the dataset it is very huge and hard to handle.
* Followed by more number of outliers and skewness these two will reduce our model accuracy.
* Also, we have tried best to deal with outliers, skewness and zero values. So, it looks quite good that we have achieved an accuracy of 95.26% even after dealing all these drawbacks.
* Also, this study will not cover all Classification algorithms instead, it is focused on the chosen algorithm, starting from the basic ensemble techniques to the advanced ones.