

**“A project report on**

**Micro credit –**

**Loan Defaulter Model”**

**SUBMITTED BY**

**HIMAJA IJJADA**

**ACKNOWLEDGMENT**

I express my sincere gratitude to FlipRobo Technologies for giving me the opportunity to work on the Micro credit loan defaulter project using machine learning algorithms. I would also like to thank FlipRobo Technologies for providing me with the requisite datasets to work with. And I would like to express my gratitude to Mr. Mohd Kashif (SME FlipRobo) and Ms. Sapna Verma (SME FlipRobo) for being of a great help in completion of the project.

Most of the concepts used to predict the Micro-Credit loandefaultersarelearned from Data Trained Institute and below documentations.

* https://scikit-learn.org/stable/
* https://seaborn.pydata.org/
* https://www.scipy.org/
* https://imbalanced-learn.org/stable/

**CONTENTS**

* **Introduction**
* Business Problem Framing
* Conceptual Background of the Domain Problem
* Review of Literature
* Motivation for the Problem Undertaken
* **Analytical problem framing**
* Mathematical/ Analytical Modeling of the Problem
* Data Sources and their formats
* Data Preprocessing Done
* Data Inputs- Logic- Output Relationships
* Assumptions
* Hardware and Software Requirements and Tools Used
* **Model/s Development and evaluation**
* Visualizations
* Identification of possible problem-solving approaches (methods)
* Testing of Identified Approaches (Algorithms)
* Run and Evaluate selected models
* Key Metrics for success in solving problem under consideration
* Interpretation of the Results
* **Conclusion**
* Key Findings and Conclusions of the Study
* Learning Outcomes of the Study in respect of Data Science
* Limitations of this work and Scope for Future Work

**INTRODUCTION**

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

**Review of Literature**

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.

**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.

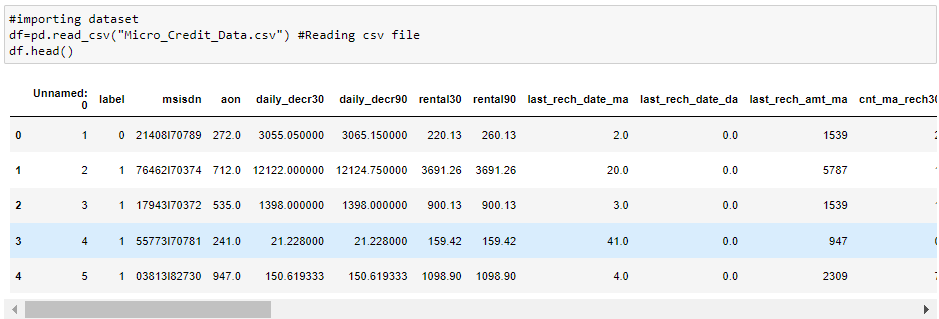
**ANALYTICAL PROBLEM FRAMING**

**Mathematical/ Analytical Modeling 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 payed i.e. Non- defaulter, while, Label ‘0’ indicates that the loan has not been payed 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 tuned the best model and saved the best model. At last I have predicted the label using saved model.

**Data Sources and their formats**

The data was collected for my internship company – FlipRobo 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.



**Dataset Description**

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

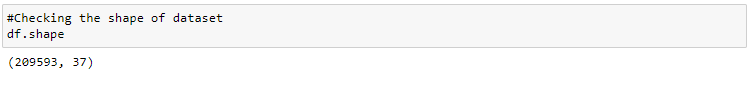
34. payback90 : Average payback time in days over last 90 days

35. pcircle : telecom circle

36. pdate : date

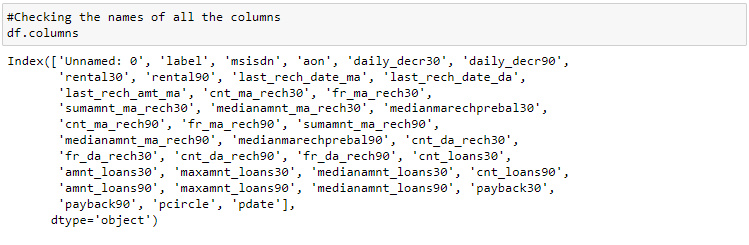
**Data Preprocessing Done**

**Checking the shape of dataset**

****

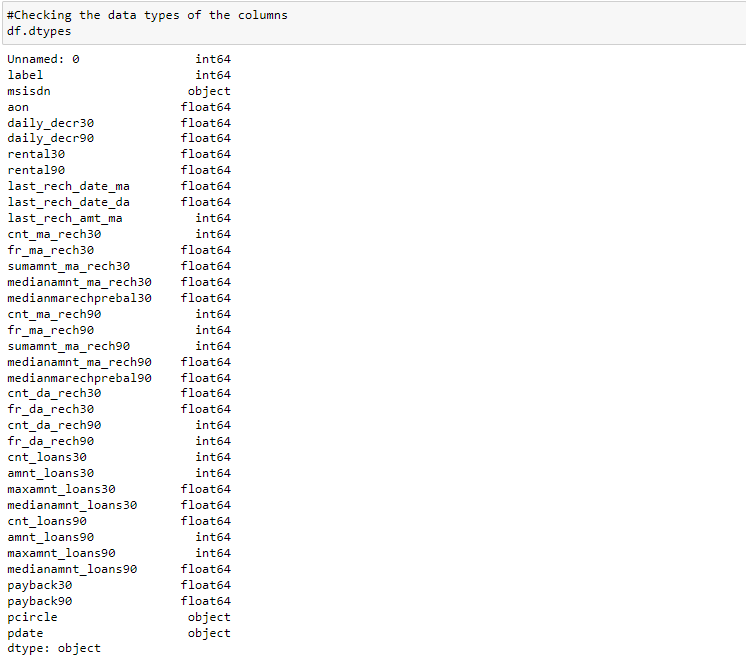
The dataset has 209593 rows and 37 columns

**Checking the names of all the columns**



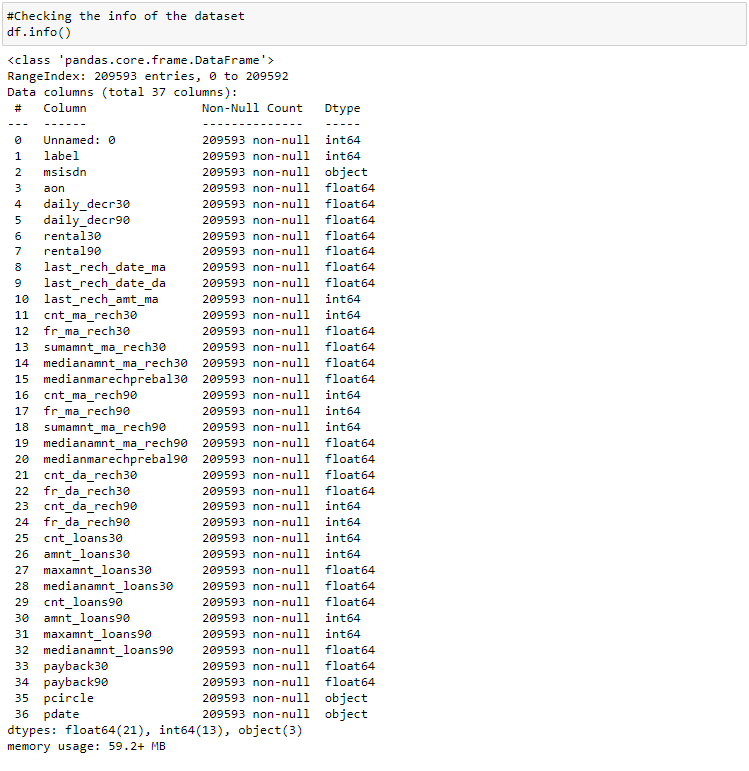
Above is the list of the column names in the dataset

**Checking the data types of the columns**

****

We can observe that the data has 3 types of data - integer, float and object.

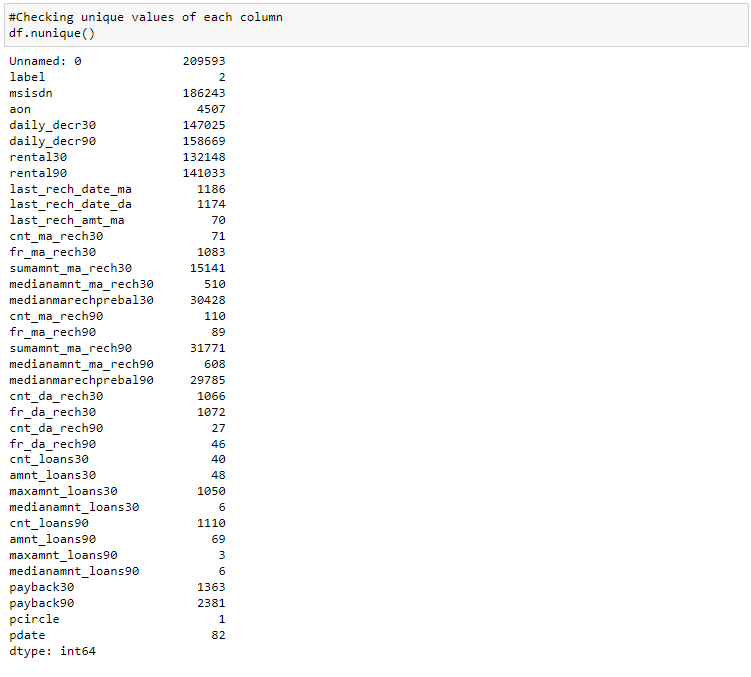
**Checking the info of the dataset**

****

## Observations

* There are no null values in the dataset
* We can see the datatypes of each column
* The column 'pdate' should be a datetime datatype so we need to change it from object to datetime datatype.

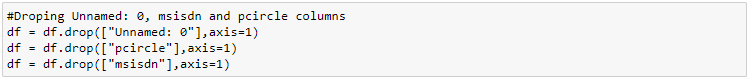
**Checking unique values of each column**

****

**Observations**

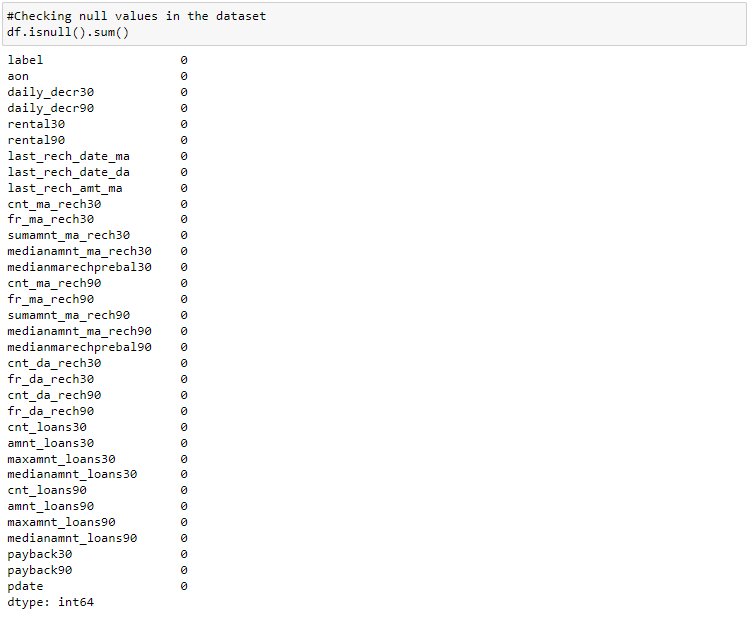
* 'Unnamed:0' is an index column in raw dataset which has all the entries as unique numbers so we can drop this column.
* The column 'pcircle' has only one entry in the column, which has no contribution with our model training. So we can drop this column also.
* The column 'msisdn' is a column with phone numbers of users, which has no contribution to the model building. So this column can be dropped.

**Dropping Unnamed: 0, msisdn and pcircle columns**

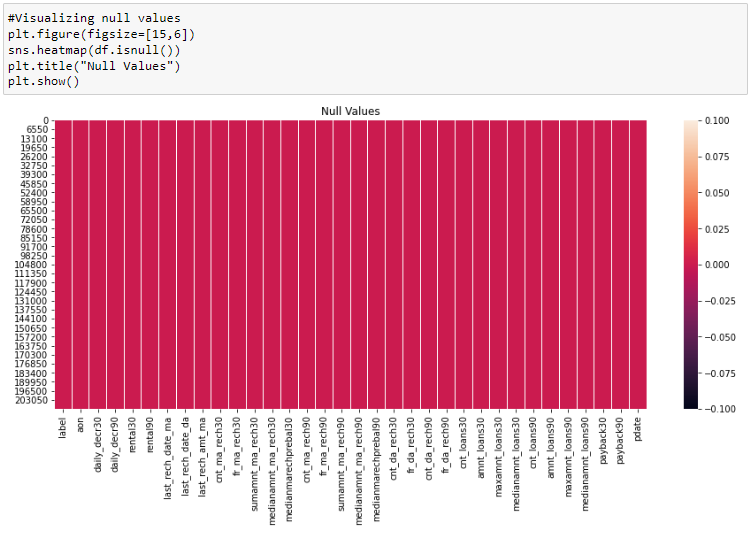
****

We have dropped the above columns from our dataset

**Checking null values in the dataset**

****

**Visualizing null values**

****

## Observations

There are no null values in our dataset.

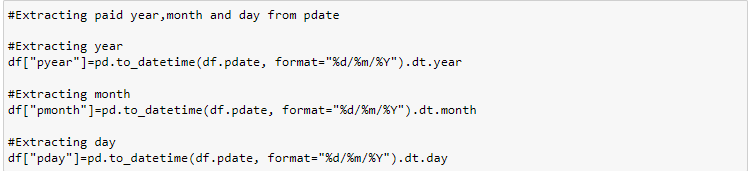
**Feature Engineering:**

**Converting object data type to datetime**

****

We have converted the pdate column from object to datetime datatype

**Extracting paid year,month and day from pdate**

****

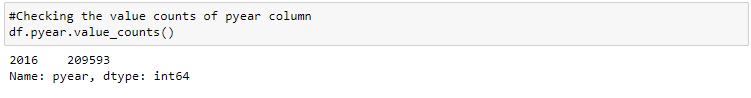
We have extracted the date month and year from the pdate column

**Dropping pdate column after extraction**

****

We have dropped the pdate column as it is redundant in the dataset

**Checking the value counts of pyear column**

****

## Observations

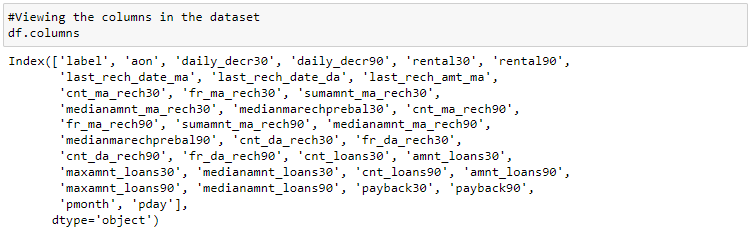
The column pyear has only one entry in all the rows. So this can be dropped as it has no insights for model training

**Dropping pyear column**

****

We have dropped the 'pyear' column from the dataset

**Viewing the columns in the dataset**

****

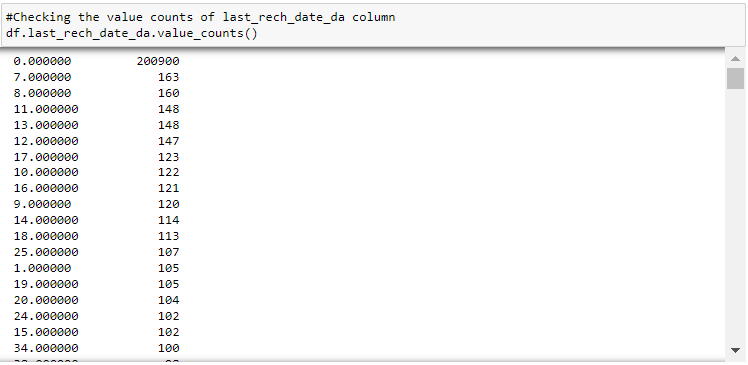
**Dropping the people who haven't taken any loan**

****

## Observations

Dropping people who haven't taken any loans as we don't have any insights from this column for model training.

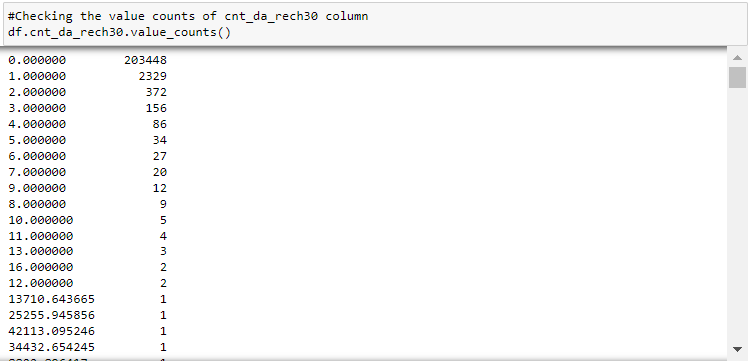
**Checking the value counts of last\_rech\_date\_da column**

****

## Observations

We have 97% zeros in this column.

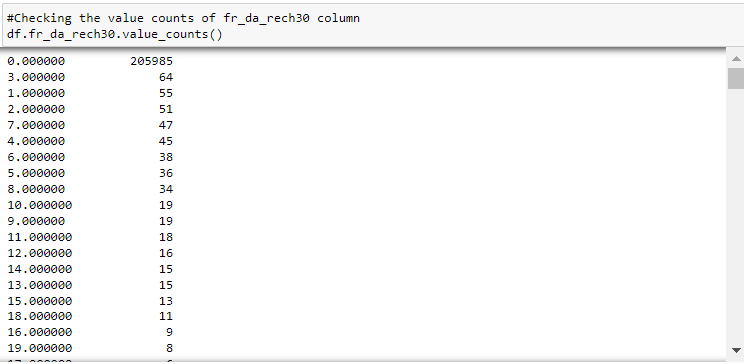
**Checking the value counts of cnt\_da\_rech30 column**

****

## Observations

We have 98% zeros in this column.

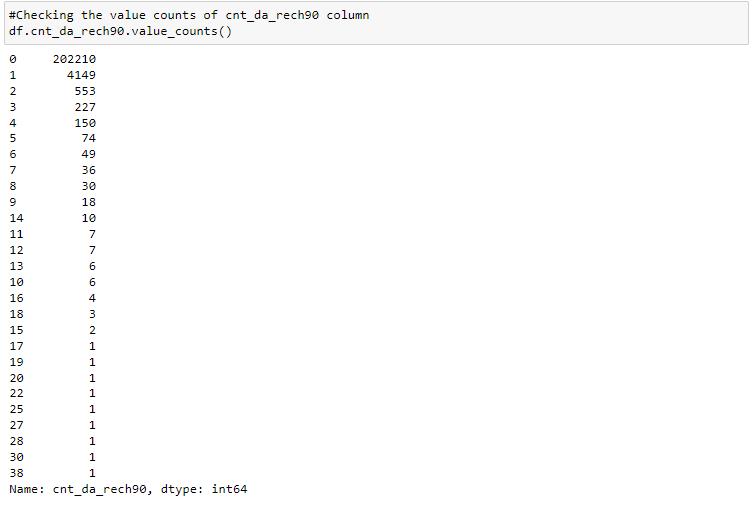
**Checking the value counts of fr\_da\_rech30 column**

****

## Observations

We have 99% zeros in this column

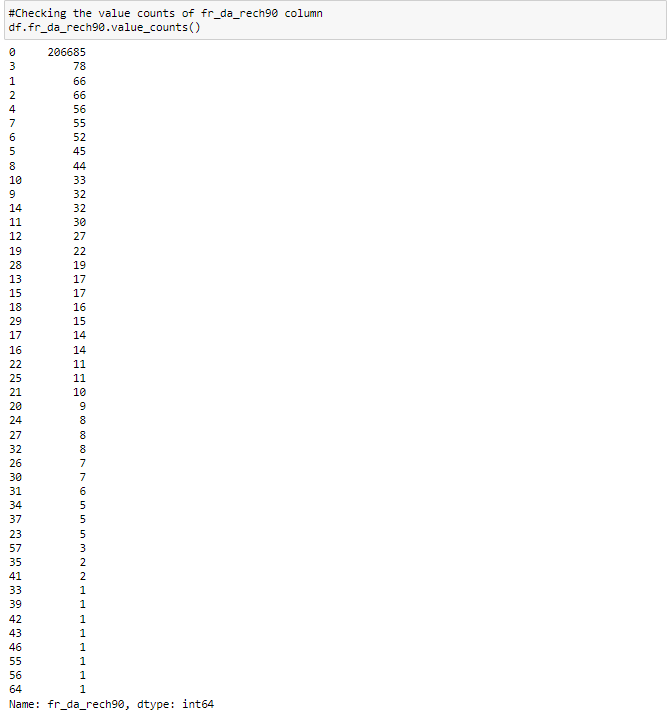
**Checking the value counts of cnt\_da\_rech90 column**

****

## Observations

We have 97% zeros in this column.

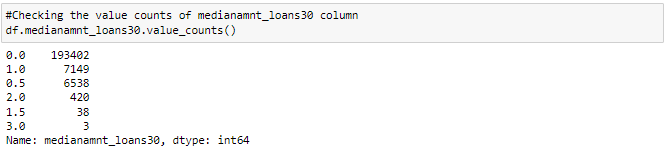
**Checking the value counts of fr\_da\_rech90 column**

****

## Observations

We have 99% zeros in this column.

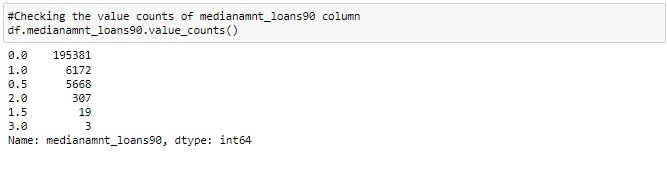
**Checking the value counts of medianamnt\_loans30 column**

****

## Observations

We have 93% zeros in this column.

**Checking the value counts of medianamnt\_loans90 column**

****

## Observations

We have 94% zeros in this column.

In all the above columns we can observe there are more than 90% of the entries as zeroes which leads to skewness in the dataset. So we have to drop these columns.

**Dropping columns with more than 90% zeros**

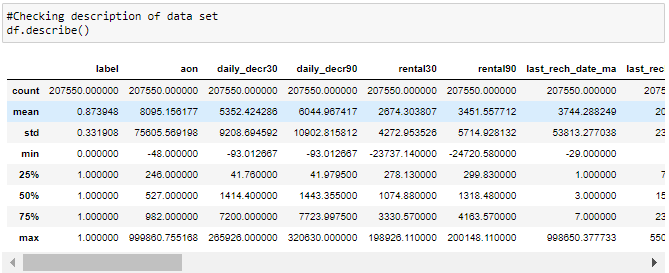
****

We have dropped those columns with more than 90% of entries as zeroes to avoid skewness in the dataset

Now let us check the statistical summary of the dataset

**Data Inputs- Logic- Output Relationships**

**Checking description of data set**

****

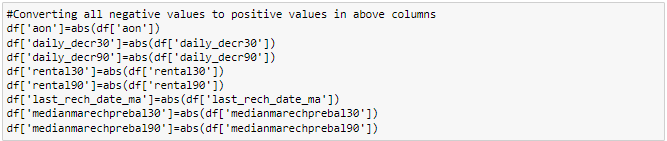
**Observations**

We can observe there are negative values in the following columns

* aon
* daily\_decr30
* daily\_decr90
* rental30
* rental90
* last\_rech\_date\_ma
* medianmarechprebal30
* medianmarechprebal90

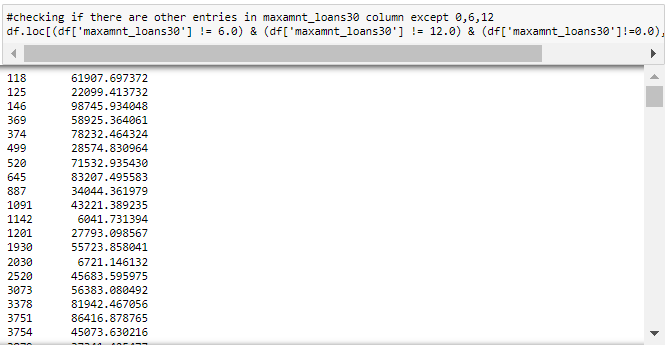
The entries in the above columns cannot be negative as they are age, account balance and number of days. So we shall change them to positive.

**Converting all negative values to positive values in above columns**

****

We have converted the negative values in the entries to positive

**Checking if there are other entries in maxamnt\_loans30 column except 0,6,12**

****

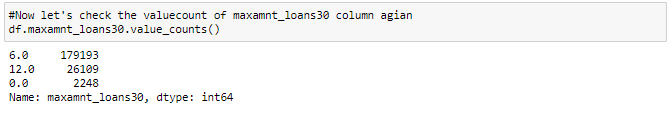
In the problem statement it is specified that we can have only 0,6,12 as maximum amount of loan taken by the user in last 30 days. So converting all the above values to zero

**Converting the values into zero which are not 0, 6 and 12**

****

We have converted those values into zeroes which are not 0, 6 and 12.

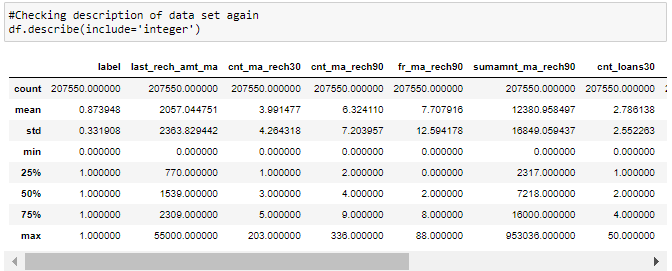
**Checking the value count of maxamnt\_loans30 column again**

****

## Observations

Here we have a class imbalance in the above column, i.e., the entries are not evenly distributed. But this is not the target column so we need not fix this

**Checking description of data set again**

****

**Assumptions**

* The count of all the columns are same which says there are no null values in the dataset
* We need not consider the statistical summary of the target as the entries in it are classes rather than values
* The following columns have Low standard deviation, which means data are clustered around the mean
  + cnt\_loans30
  + amnt\_loans30
  + maxamnt\_loans90
  + pmonth
  + pday
* The other columns have High standard deviation, which indicates data in those columns is more spread out.
* The following columns have mean almost equal to median, which says the distribution of curve is normal
  + maxamnt\_loans90
  + pmonth
  + pday
* The other columns have mean greater than the median (50th percentile) , which says the distribution is skewed to right
* The 'pmonth' column has no much difference between the 75% (3rd quantile) and the max values, which shows there are less chance for presence of outliers
* All the other columns have huge difference between the 75% (3rd quantile) and the max values, which shows the chance for presence of outliers

**Checking unique values of target column**

****

There are only two unique values in target column. Therefore it is a binary classification problem

**Checking for empty observations**

****

We can observe that there are no empty observations in the target column.

**Hardware and Software Requirements and Tools Used**

**Hardware Used:**

* Processor AMD Ryzen 9 5900HX(8 Cores 16 Logical Processors)
* Physical Memory: 16.0GB (3200MHz)
* GPU: Nvidia RTX 3060 (192 bits), 6GB DDR6 VRAM, 3840 CUDA cores.

**Software Used:**

* Windows 10 Operating System
* Anaconda Package and Environment Manager: Anaconda is a distribution of the Python and R programming languages for scientific computing, that aims to simplify package management and deployment. The distribution includes data- science packages suitable for Windows and provides a host of tools and environment for conducting Data Analytical and Scientific works. Anaconda provides all the necessary Python packages and libraries for Machine learning projects.
* Jupyter Notebook: The Jupyter Notebook is an open-source web application that allows data scientists to create and share documents that integrate live code, equations, computational output, visualizations, and other multimedia resources, along with explanatory text in a single document.
* Python3: It is open source, interpreted, high level language and provides great approach for object-oriented programming. It is one of the best languages used for Data Analytics and Data science projects/application. Python provides numerous libraries to deal with mathematics, statistics and scientific function.
* Python Libraries used:
* **Pandas:** For carrying out Data Analysis, Data Manipulation, and Data Cleaning etc.
* **Numpy:** For performing a variety of operations on the datasets.
* **matplotlib.pyplot, Seaborn**: For visualizing Data and various relationships between Feature and Label Columns
* **Scipy:** For performing operations on the datasets
* **Statsmodels:** For performing statistical analysis

**sklearn** for Modelling Machine learning algorithms, Data Encoding, Evaluation metrics, Data Transformation, Data Scaling, Component analysis, Feature selection etc.

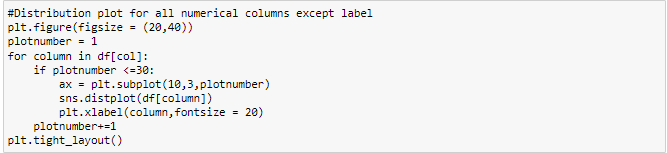
**MODEL/S DEVELOPMENT AND EVALUATION**

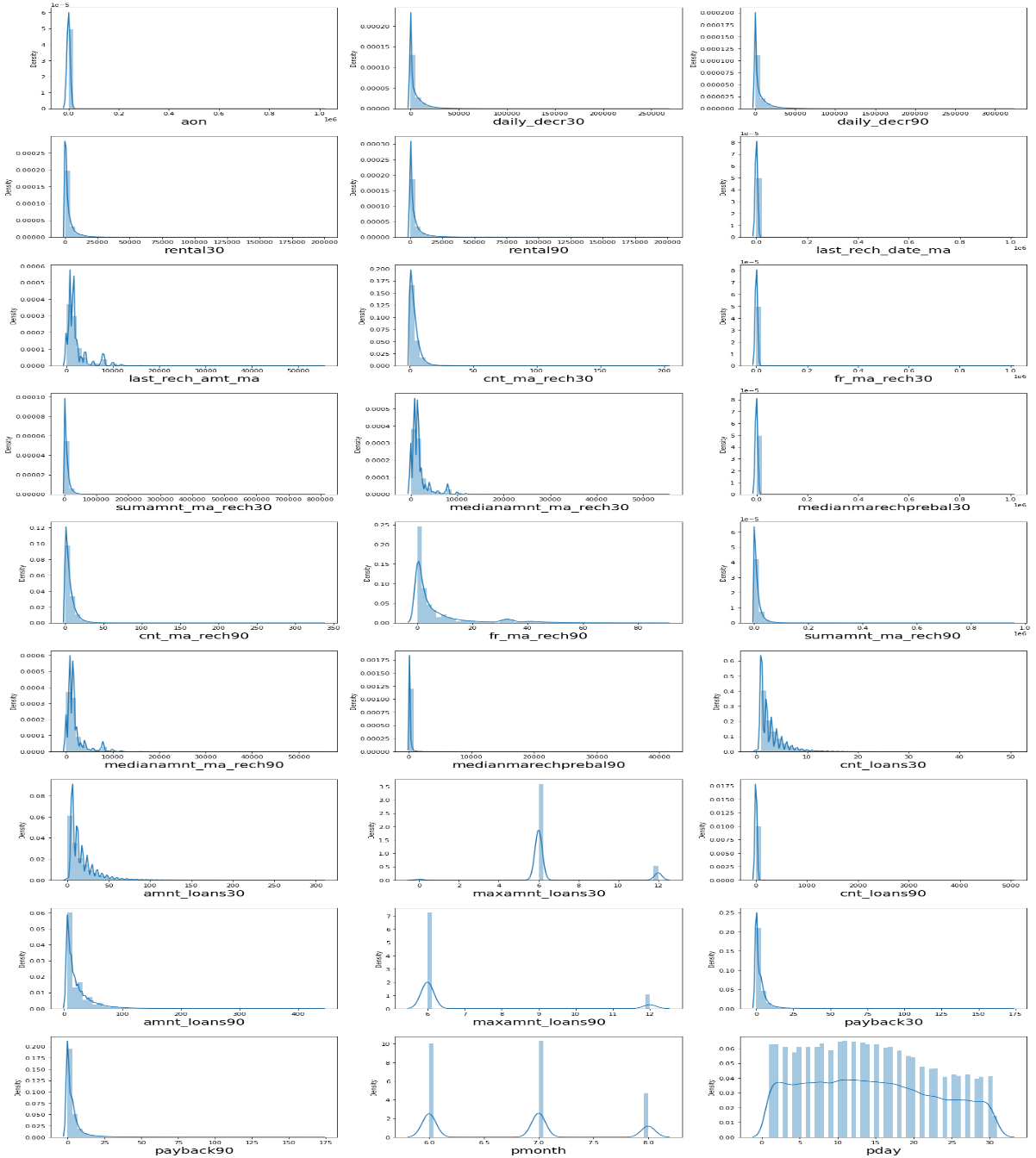
**Visualizations**

## Univariate Analysis:

Let us separate the columns with numerical data fro performing the univariate analysis

****

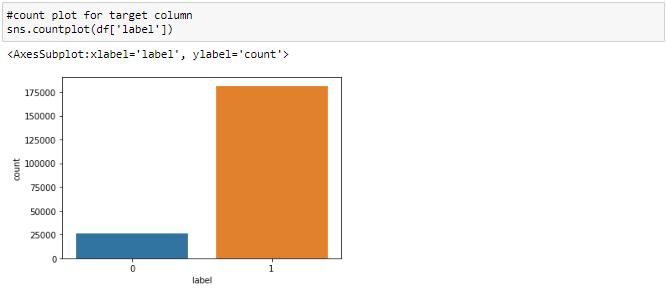
****

****

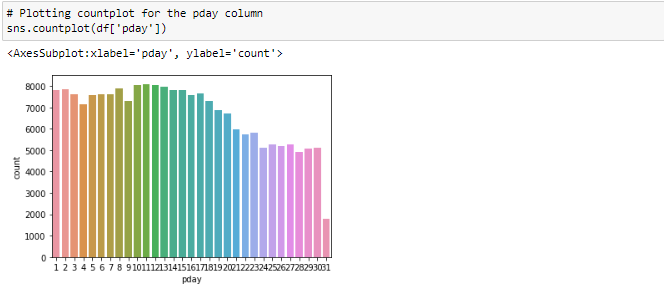
## Observations

* We can clearly see that there is skewness in most of the columns so we need to treat them.
* The following columns are skewed to the right
  + 'daily\_decr30'
  + 'daily\_decr90'
  + 'rental30'
  + 'rental90'
  + 'last\_rech\_amt\_ma'
  + 'sumamnt\_ma\_rech30'
  + 'medianamnt\_ma\_rech30'
  + 'payback30'
  + 'payback90'
  + 'amnt\_loans90'
  + 'medianamnt\_ma\_rech90'
  + 'fr\_ma\_rech90'
  + 'cnt\_ma\_rech30'
  + 'cnt\_ma\_rech90'
  + 'sumamnt\_ma\_rech90'
* The following columns are normally distributed
  + 'aon'
  + 'last\_rech\_date\_ma'
  + 'fr\_ma\_rech30'
  + 'cnt\_loans30'
  + 'amnt\_loans30'
  + 'medianmarechprebal90'
  + 'cnt\_loans90'
  + 'medianmarechprebal30'
* Though some of the columns have integers as entries the data is considered categorical, so we need not take them into consideration.

**Count plot for target column**

****

## Observations: There is a data imbalance issue so we have to treat this by using oversampling or undersampling techniques.

****

**Observations**

* All the days till third week of the month have similar no of entries
* The last week of every month has lesser entrie sthan the first 3 weeks.

## Bivariate Analysis

## 

## 

**OBSERVATIONS:**

1. Customers with high value of Age on cellular network in days (aon) are maximum defaulters (who have not paid there loan amount-0).

2. 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 there loan amount-1).

3. 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 there loan amount-1).

4. Customers with high value of Average main account balance over last 30 days (rental30) are maximum Non-defaulters (who have paid there loan amount-1).

5. Customers with high value of Average main account balance over last 90 days (rental90) are maximum Non- defaulters (who have paid there loan amount-1).

6. Customers with high Number of days till last recharge of main account (last\_rech\_date\_ma) are maximum Non- defaulters (who have paid there loan amount-1).

7. 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).

8. Customers with high value of Number of times main account got recharged in last 30 days(cnt\_ma\_rech30) are maximum Non-defaulters(who have paid there loan amount-1).

9. Customers with high value of Frequency of main account recharged in last 30 days (fr\_ma\_rech30) are maximum Non-defaulters (who have paid there loan amount-1) and also the count is high for defaulters comparatively Non-defaulters are more in number.

10. Customers with high value of Total amount of recharge in main account over last 30 days (in Indonesian Rupiah) (sumamnt\_ma\_rech30) are maximum Non-defaulters (who have paid there loan amount-1).

11. Customers with high value of Median of amount 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 there loan amount-1).

12. Customers with high value of Median of main account balance just before recharge in last 30 days at user level (in Indonesian Rupiah) (medianmarechprebal30) are maximum defaulters (who have not paid there loan amount-0).

13. Customers with high value of Number of times main account got recharged in last 90 days(cnt\_ma\_rech90) are maximum Non-defaulters(who have paid there loan amount-1).

14. Customers with high value of Frequency of main account recharged in last 90 days (fr\_ma\_rech90) are maximum Non-defaulters (who have paid there loan amount-1).

15. Customers with high value of Total amount of recharge in main account over last 90 days (in Indonesian Rupiah) (sumamnt\_ma\_rech90) are maximum Non-defaulters (who have paid there loan amount-1).

16. Customers with high value of Median of amount of recharges done in main account over last 90 days at user level (in Indonesian Rupiah) (medianamnt\_ma\_rech90) are maximum Non-defaulters (who have paid there loan amount-1).

17. Customers with high value of Median of main account balance just before recharge in last 90 days at user level (in Indonesian Rupiah) (medianmarechprebal90) are maximum Non-defaulters (who have paid there loan amount-1).

18. Customers with high value of Number of loans taken by user in last 30 days (cnt\_loans30) are maximum Non-defaulters (who have paid there loan amount-1).

19. Customers with high value of Total amount of loans taken by user in last 30 days (amnt\_loans30) are maximum Non-defaulters (who have paid there loan amount-1).

20. Customers with high value of maximum amount of loan taken by the user in last 30 days (maxamnt\_loans30) are maximum Non-defaulters (who have paid there loan amount-1).

21. Customers with high value of Number of loans taken by user in last 90 days (cnt\_loans90) are maximum Non-defaulters (who have paid there loan amount-1).

22. Customers with high value of Total amount of loans taken by user in last 90 days (amnt\_loans90) are maximum Non-defaulters (who have paid there loan amount-1).

23. Customers with high value of maximum amount of loan taken by the user in last 90 days (maxamnt\_loans90) are maximum Non-defaulters (who have paid there loan amount-1).

24. Customers with high value of Average payback time in days over last 30 days (payback30) are maximum Non-defaulters (who have paid there loan amount-1).

25. Customers with high value of Average payback time in days over last 90 days (payback90) are maximum Non-defaulters (who have paid there loan amount-1).

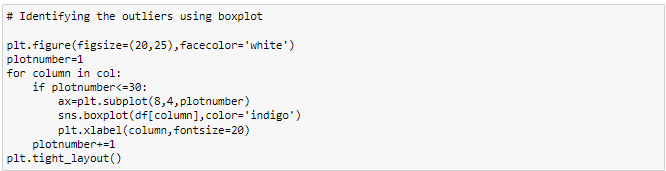
26. In between 6th and 7th month maximum customers both defaulters and Non-defaulters have paid there loan amount.

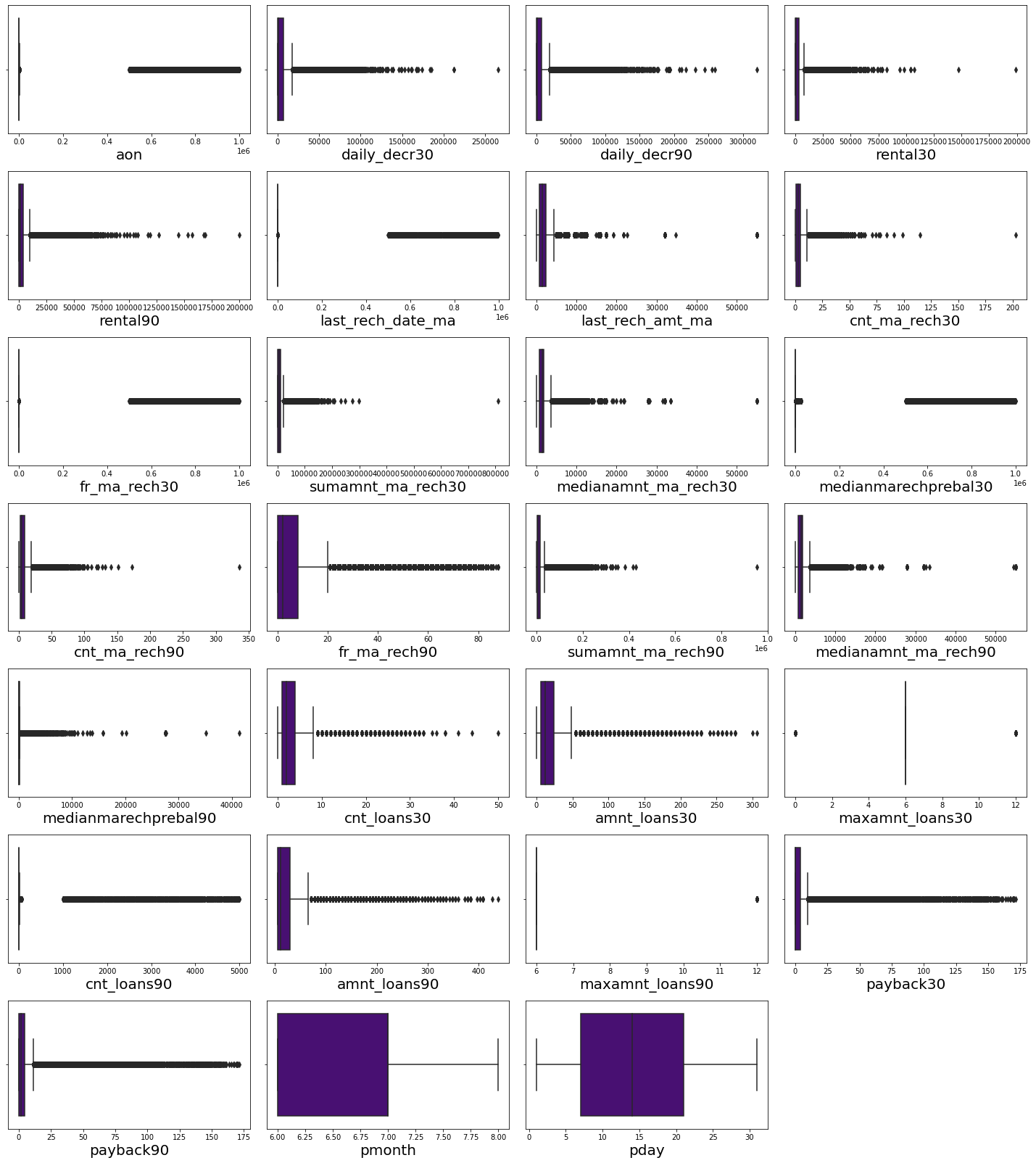
27. Below 14th of each month all the customers have paid there loan amount.

**Identification of possible problem-solving approaches (methods)**

### **Checking for outliers:**

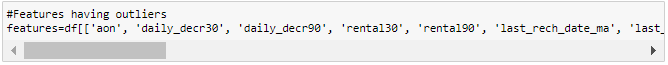
**Identifying the outliers using boxplot**

****

****

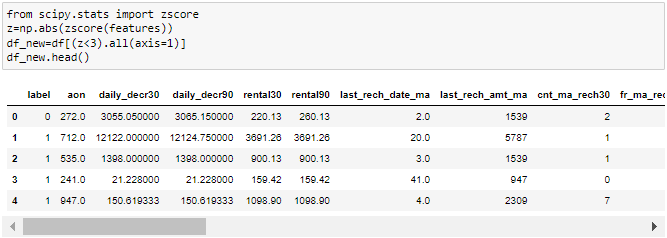
## Removing Outliers

### i) Zscore method

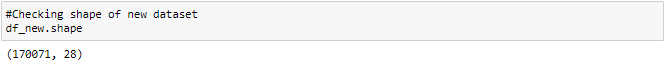
****

Above are the list of columns with outliers in the dataset.

## Removing the outliers using zscore method and viewing the dataset

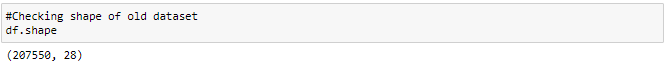
****

## Checking shape of new dataset

****

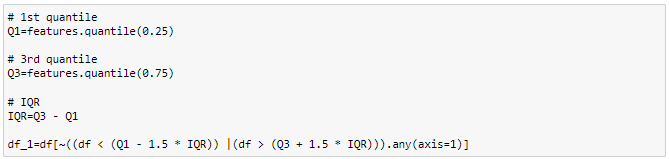
The new dataset has 170071 rows and 28 columns.

## Checking shape of old dataset

****

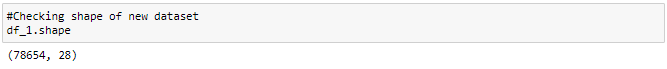
In Z-score method the data loss is more than 10% so let us have a look into IQR method to remove outliers.

### ii) IQR method:

****

We have removed the skewness of the dataset using IQR method.

**Checking shape of new dataset**

****

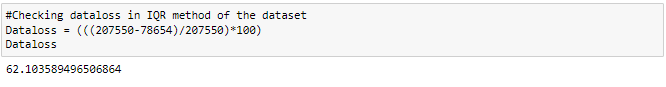
In the new dataset we have 78654 rows and 28 columns.

**Checking shape of old dataset**

****

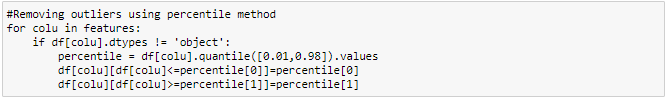
The dataset previously had 207550 rows and 28 columns.

**Checking data loss in IQR method of the dataset**

****

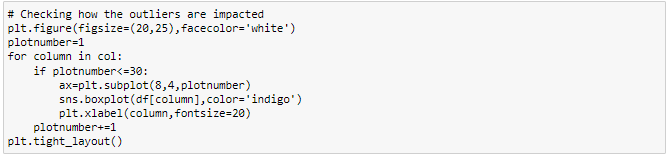
In IQR method the data loss is more than 50%, which is not acceptable. So let us have a look into percentile method to remove outliers.

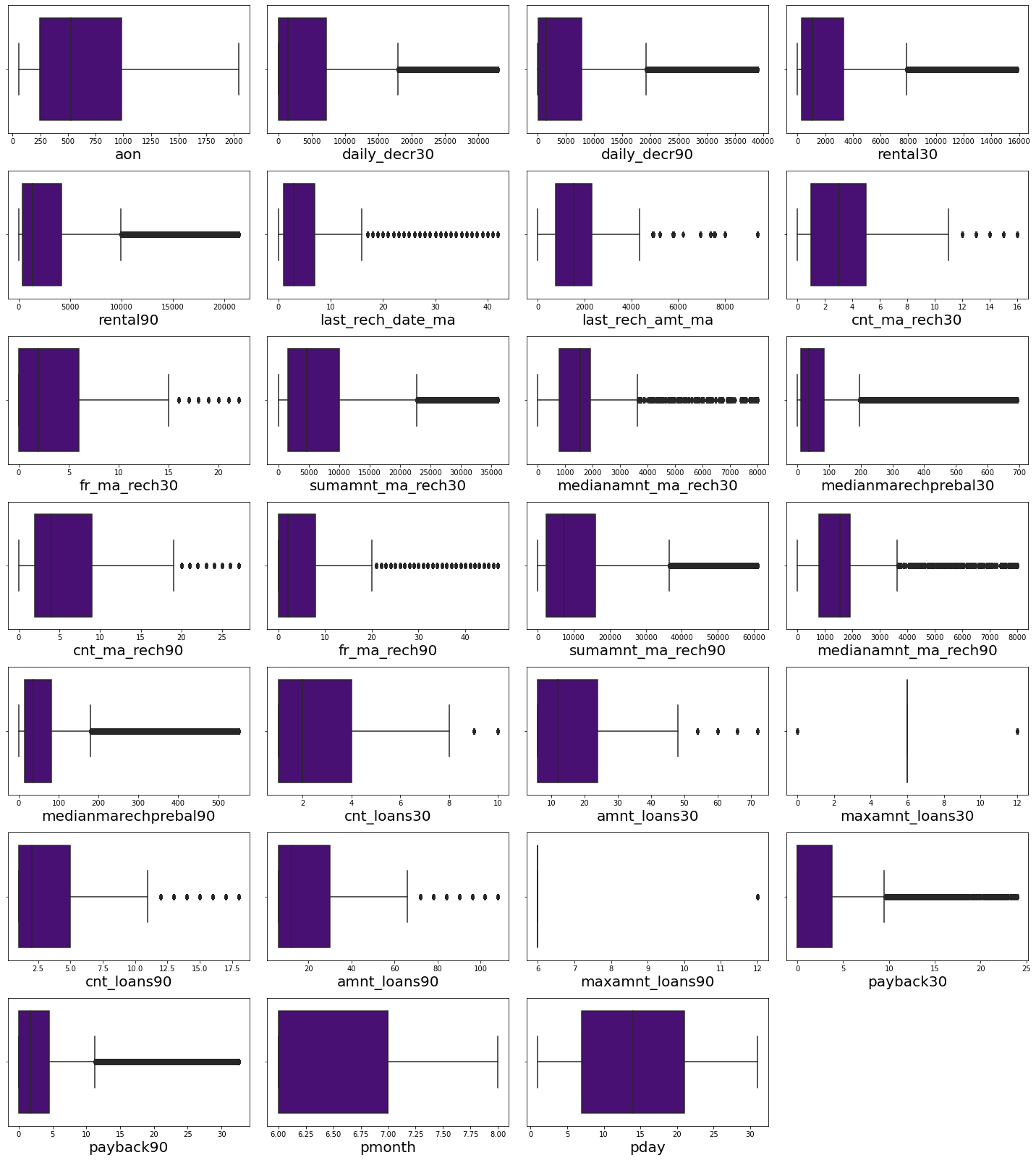
### iii) Percentile Method:

****

We have successfully removed outliers in the dataset using percentile method.

**Checking how the outliers are impacted**

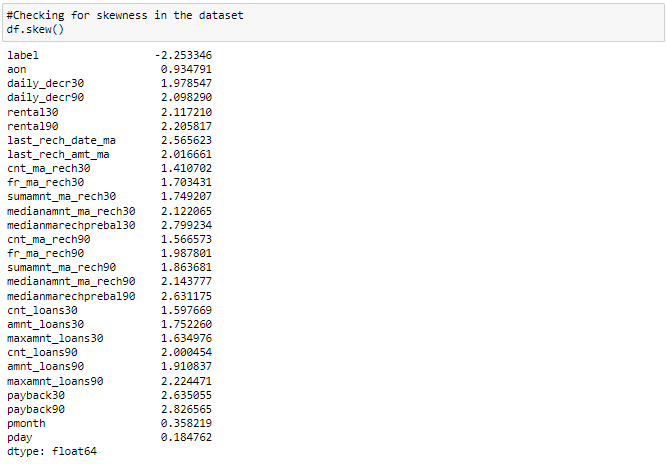
****

****

## Observations

We can observe that the Outliers has been significantly reduced in all the columns.

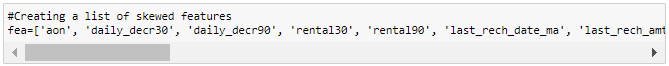
## Checking for skewness:

****

## Observations

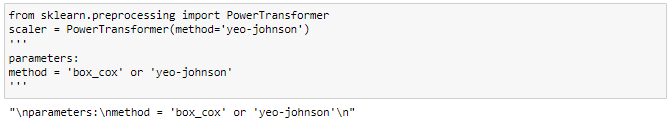
There is skewness in almost all columns except pmonth,pday and as label is my target i should not remove skewness from this column.

## Removing skewness

****

Creating a list as fea with all the columns having skewness.

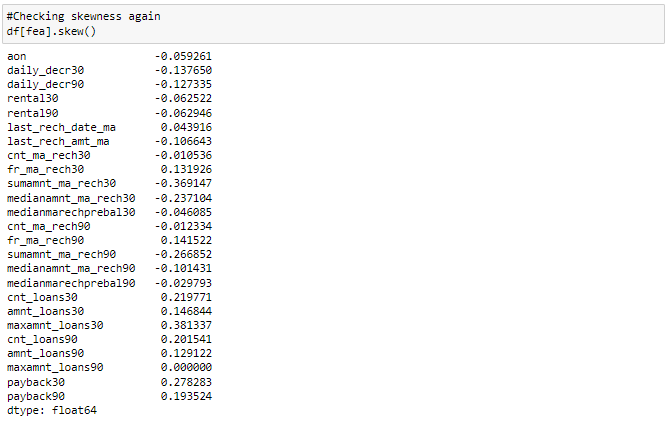
## Removing skewness using yeo-johnson method:

****

Using yeo\_johnson method for removing the skewness

****

The skewness has been removed from the dataset.

****

Skewness in all the columns has been reduced.

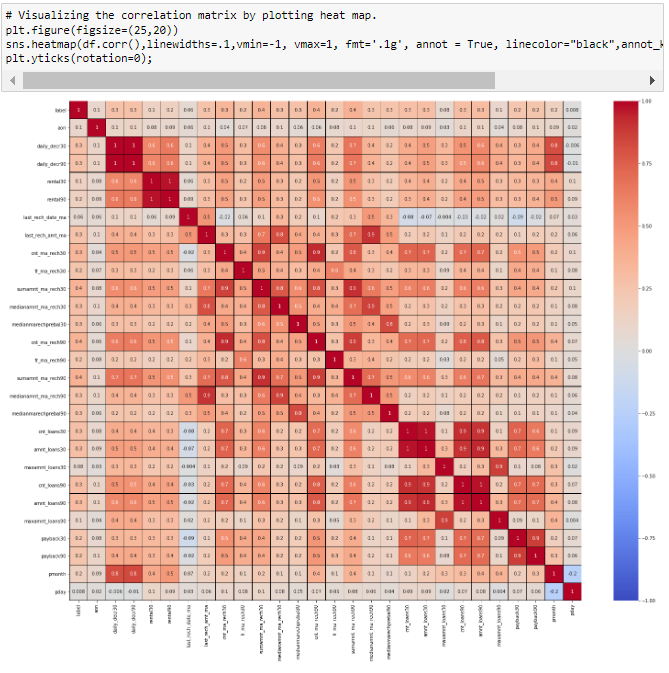
## Correlation

****

## Observations

Above are the correlations of all of the features. To get better visualization on the correlation of features,let us plot it using a heat map.

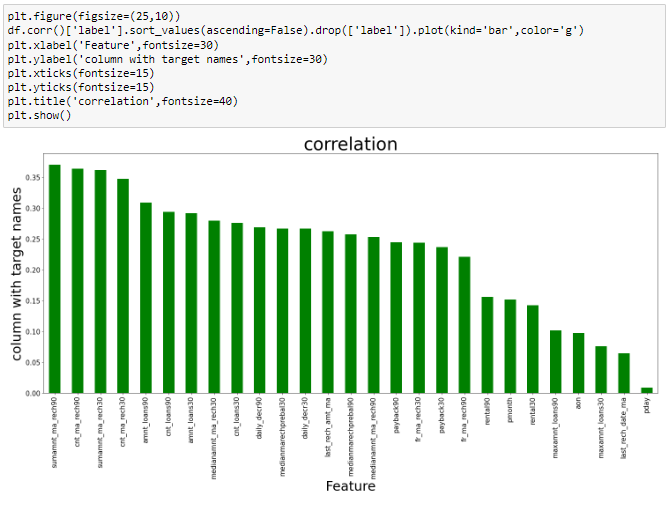
**Visualizing the correlation matrix by plotting heat map.**

****

**Observations**

* We can observe that there are some fetures which are positively correlated with other features which says there is multicollinearity in the dataset. We need to terat the multicollinearity before building a model.

Let's visualize the correlation of all the features with target to get better insights.

****

## Observations

* All the features are positively correlated with the target. We don't have any negatively correlated features with the target.
* Even though pday feature is least correlated with the target, we can keep it and proceed.
* 'sumnamt\_ma\_rech90', 'cnt\_ma\_rech90', 'sumnamt\_ma\_rech30' are most correlated with the target.

## Separating Features and Target:

****

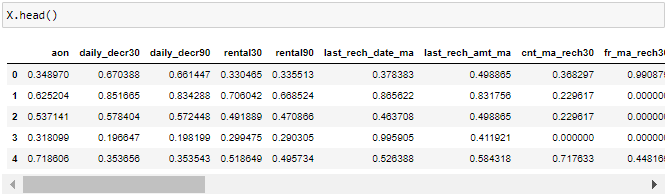
We have separated the target and independent columns.

# Scaling

****

We have scaled the data using MinMax scaler

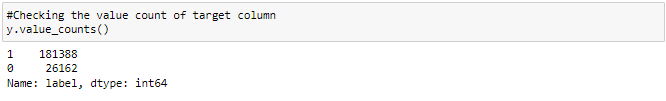
**Viewing the data after scaling**

****

This is the data of independent variables after scaling.

## Balancing

**Checking the value count of target column**

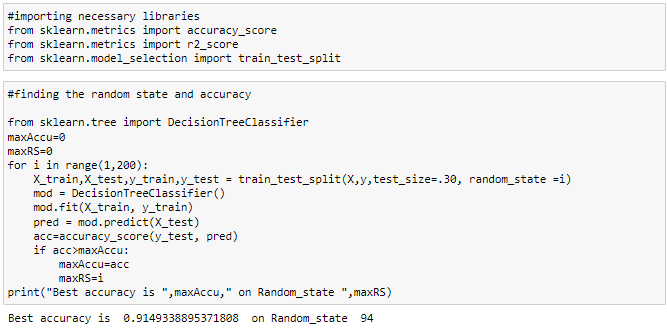
****

**Balancing the target variable using oversampling**

****

Now the data looks balanced.

## Finding Best Random State and Accuracy:

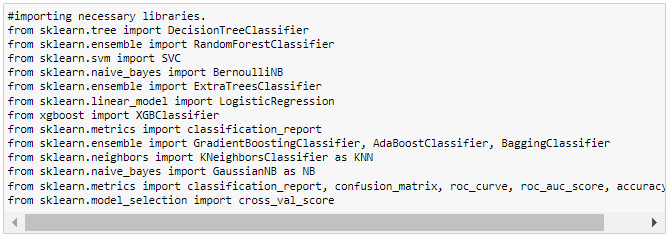
****The above is the best accuracy and random state.

**Creating the train and test split**

****

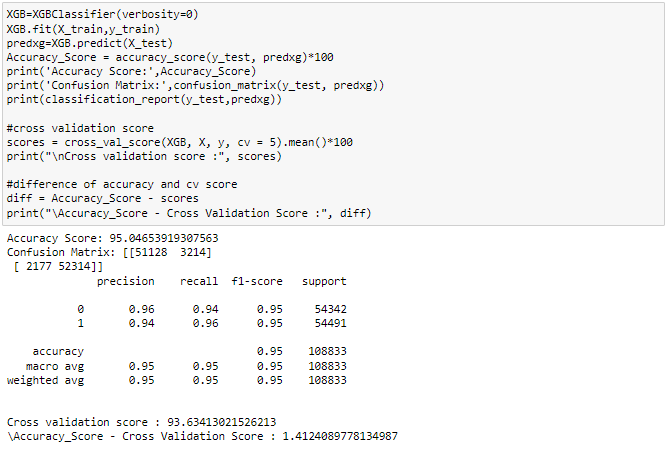
We havecreated the train and test split.

## Classification Algorithms:

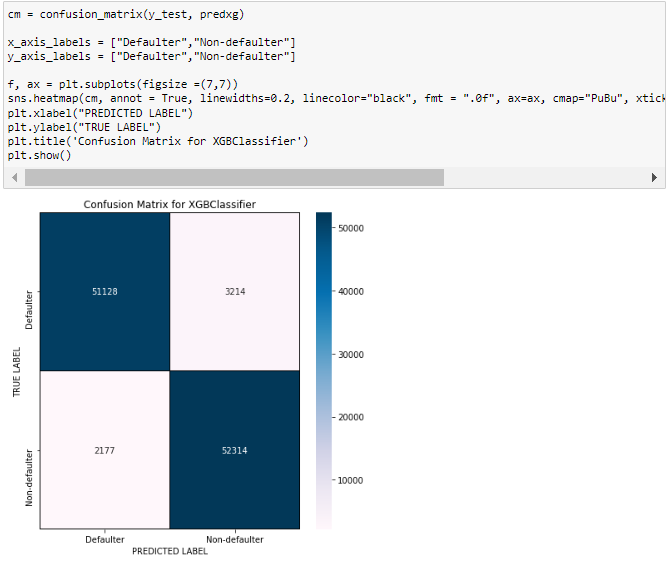
****

We have imported all the necessary libraries for the classification algorithms to be evaluated

## i) XGB Classifier:

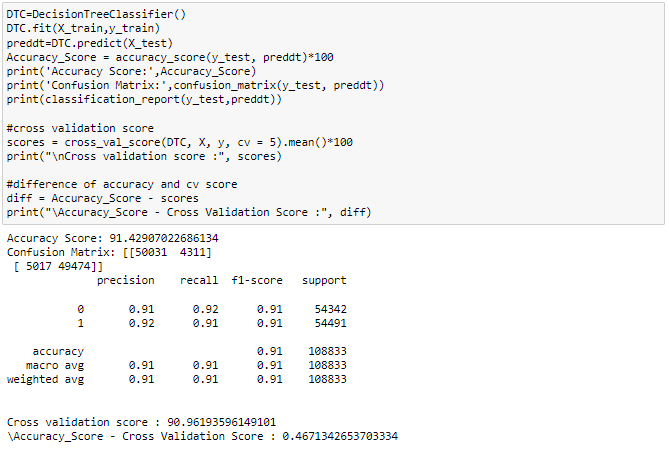
****

XGBClassifier is giving 95% accuracy.

****

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

## ii) DecisionTreeClassifier:

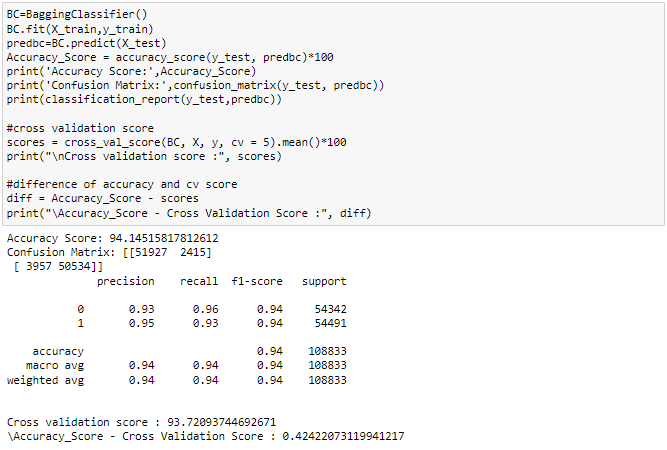
****

DecisionTreeClassifier is giving 91.42% accuracy.

****

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

## iii) BaggingClassifier:

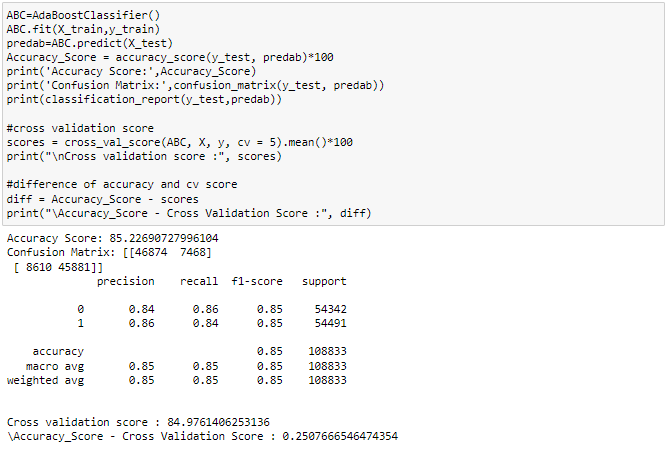
****

BaggingClassifier is giving 94.14% accuracy.

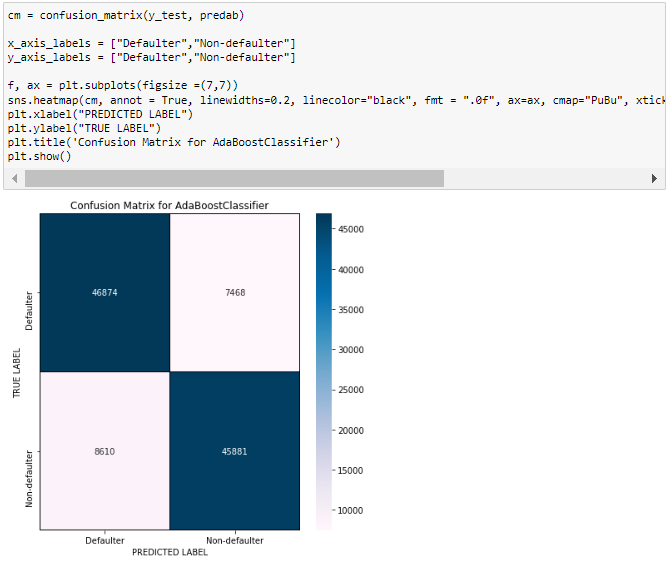
****

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

## iv) AdaBoostClassifier:

****

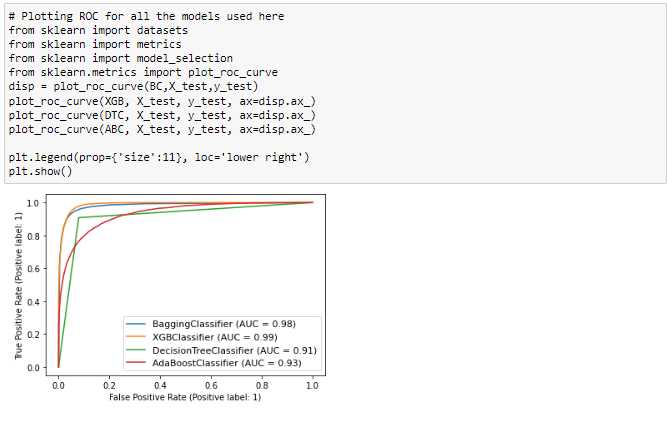
AdaBoost Classifier is giving 85% accuracy.

****

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

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

## ROC-AUC curve:

****

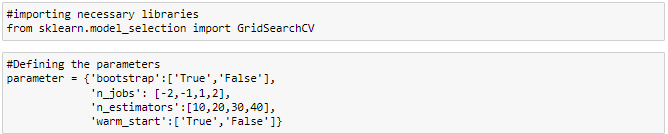
Above is the ROC curves for all the models that we have predicted. The AUC values can also be seen in the plot.

AUC values for XGBClassifier and BaggingClassifier are higher compared to other models. We have least difference in model accuracy and cross validation score for BaggingClassifier so BaggingClassifier can be considered the best model to proceed with hyper parameter tuning.

**Testing of Identified Approaches (Algorithms)**

## Hyper Parameter tuning:

**Importing necessary libraries**

****

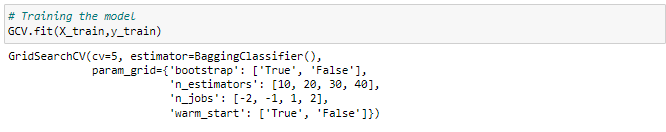
We have given the list of parameters for BaggingClassifier model.

**Defining grid search CV for bagging classifier**

****

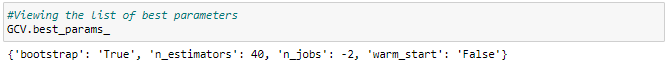
Running grid search CV for BaggingClassifier

**Training the model**

****

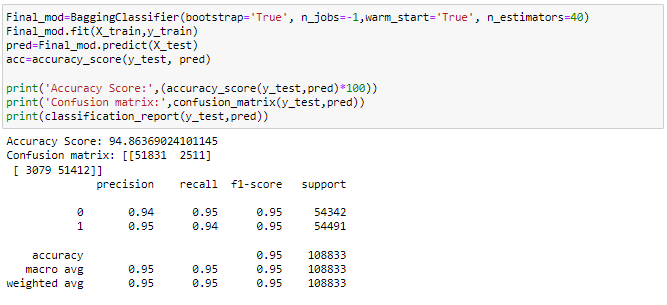
We have trained the model with the above defined GCV.

**Viewing the list of best parameters**



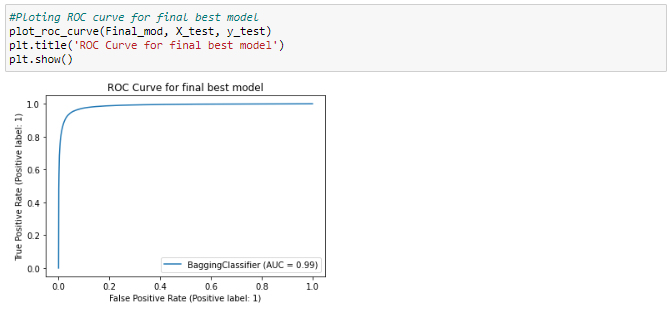
We have got the best parameters for BaggingClassifier

**Run and Evaluate selected model**



The accuracy of our final model has improved from 94.14% to 94.86%

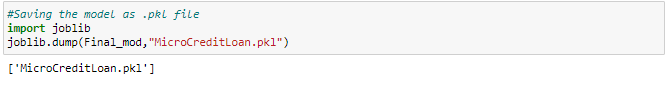
# AUC ROC CURVE for final model

**Plotting ROC curve for final best model**

After hyper parameter tuning we can notice the improvement in roc curve and AUC too. This is near ideal.

**Key Metrics for success in solving problem under consideration**

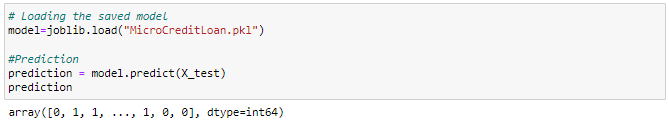
# Saving the model

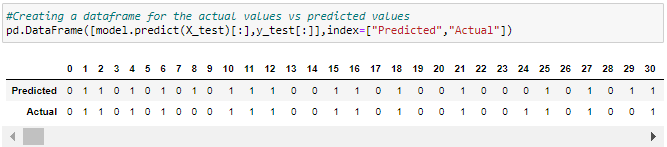


We have saved the final model as MicroCredit.pkl

# Predictions

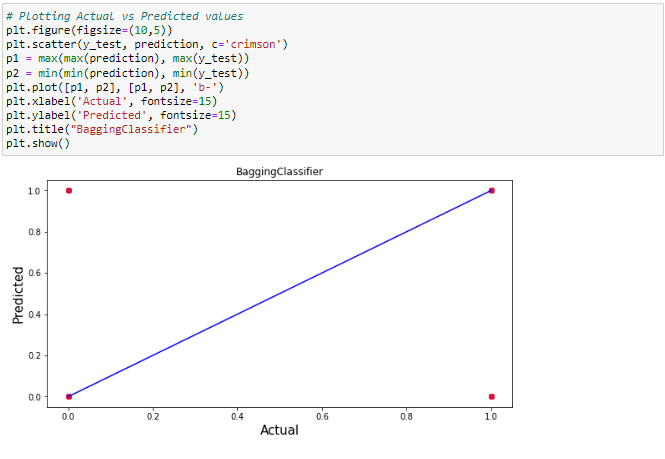
**Loading the saved model**

****

****

Above are the predicted values vs the actual values. They look similar with minimal exceptions

**Plotting Actual vs Predicted values**

****

**Observations**

* We have plotted the Actual vs Predicted values to get better insights.
* Here we can observe that the Blue line is the actual values and red dots are the predicted values.
* This says there are minimal to no exceptions, most of the predicted values are in sync with the actual 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 have noticed that most of the columns are skewed, so we have chosen proper methods to deal with the outliers and skewness. If we ignore the outliers and skewness we may end up with a model which has lesser accuracy at predicting the output.
* 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 BaggingClassifier as the best model with 94.14% accuracy\_score. Also I have improved the accuracy of the best model by running hyper parameter tuning.
* At last I have predicted whether the loan is paid back using saved model. The predictions looked similar to the actual values.

**CONCLUSION**

**Key Findings and Conclusions of the Study**

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 analyze 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 tuning 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.

**Learning Outcomes of the Study in respect of Data Science**

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 analyzed. 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.

**Limitations of this work and Scope for Future Work**

* First drawback is the shape of the dataset it is such a huge dataset that makes it 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 94.86% 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 ensembling techniques to the advanced ones.