Micro-Credit Defaulter/Non Defaulter Model

**OBJECTIVE**

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.

**PROCEDURE**

1. The very first step to start the project is to import the data set. we have successfully imported the dataset using pandas library .
2. This dataset is about a company giving loans to the customers in form of recharge amount in their cell phone talktime of amount of 5 and 10 Indonesian Rupiah. This data also shows that customer repay loans with a interest, for example for 5 and 10 ,

6 and 12 Indonesian Rupiah is collected , interest of 1 and 2 respectively.

1. The customer who is unable to repay the loans within 5 days is considered as defaulter and who clears the loan within the time is considered as non defaulters which has been represented by 0 and 1 in the dataset respectively.
2. In this dataset there are 36 attributes for a particular msisdn (customer) which shows the records such as how old is the customer , what is the avg recharge in last 30 , 90 days , ,maxm loan taken in 30 , 90 days etc, the details of which will be provided below.
3. After Importing the datset we found that this dataset has 209593 rows and 36 columns out of which 183431 corresponds to label 1 , that is non defaulters and 26162 belongs to label 0 that is defaulters .
4. Clearly from the above distribution of defaulters and non defaulters we conclude that data is highly unbalanced , as percentage of defaulters and non defaulters is 12 and 88 respectively as per the dataset.
5. In some technical point of view , this dataset has 21 , 12 , 3 features float , int and object feature respectively.

**EXPLORATORY DATA ANALYSIS**

We have performed some data visualization on this dataset in order to be more familiar with the dataset which in turn helps to get the hidden pattern of the dataset.

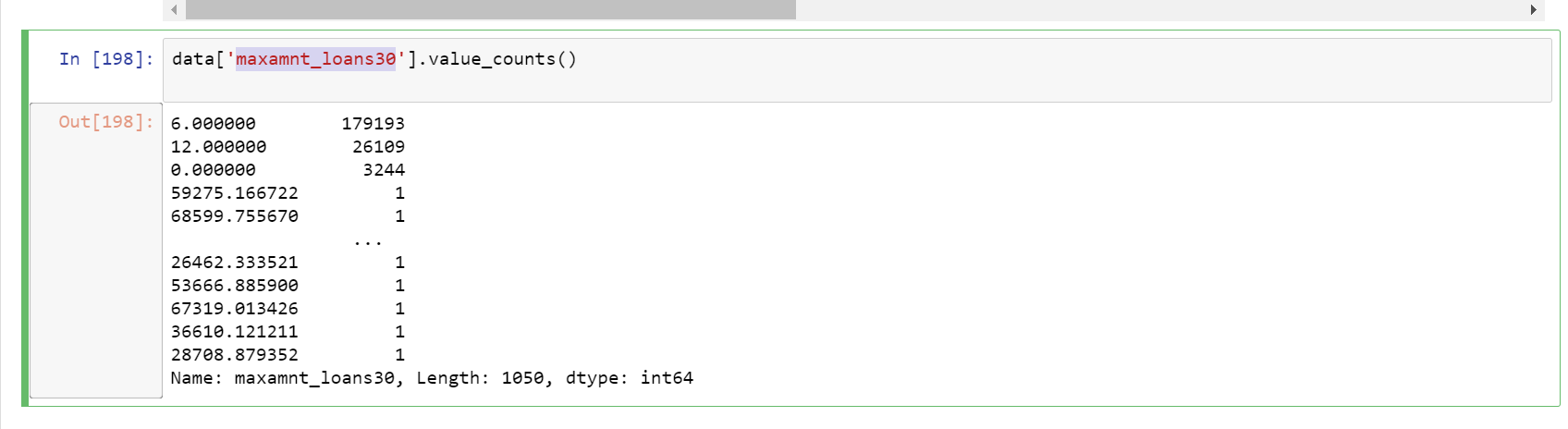
1] Null Value Analysis –

This dataset does not have any feature with null value.

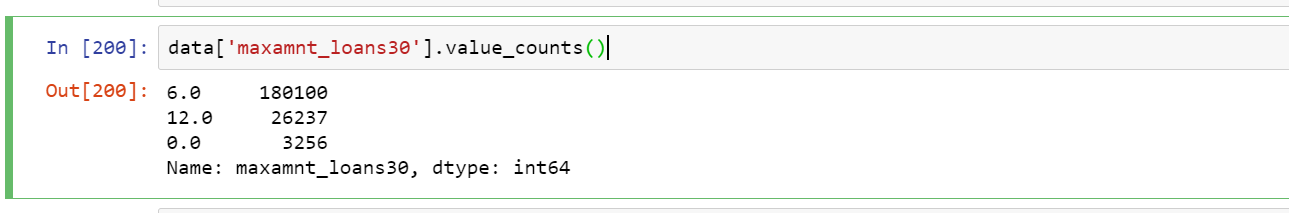
2]Univariate Analysis-

We have performed techniques like count plot, histogram plots to check the trend and pattern :-

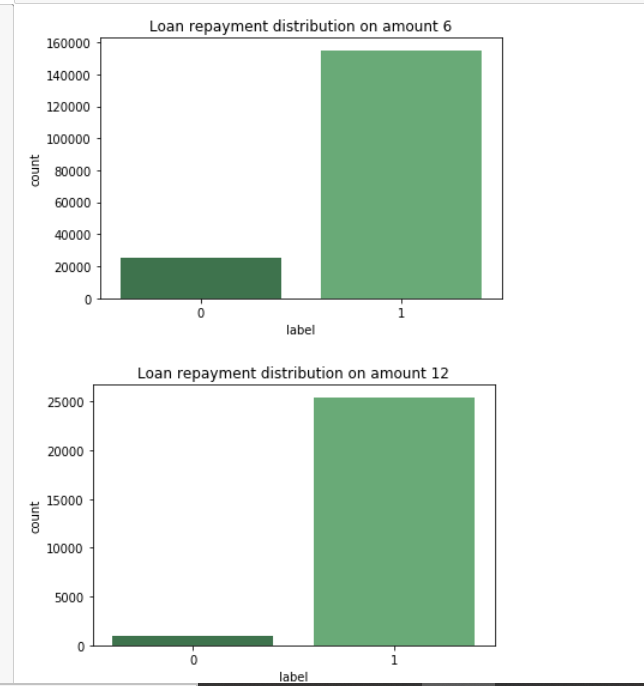
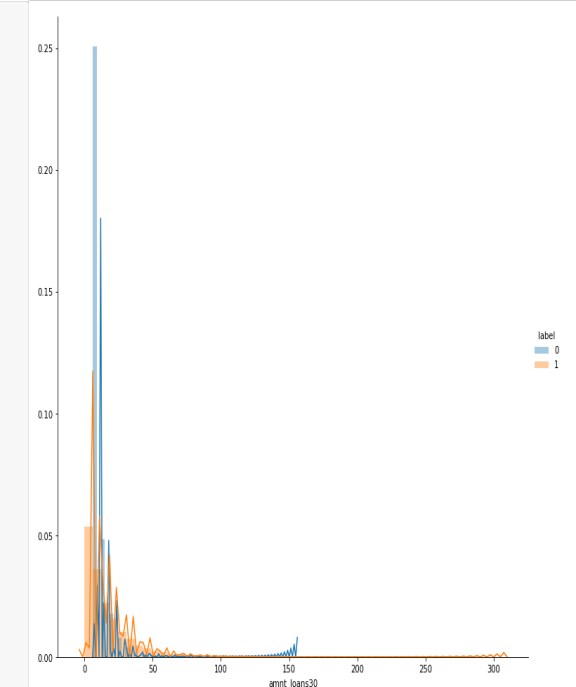
2.a) We have found that there other unwanted values for the feature maxamnt\_loans30 apart from 0 , 6 and 12 which is shown in the below figure , but when we checked maxamnt\_loans90 , there was no other values other 0, 6 and 12 .



To treat this unwanted values , we have made those values of maxamnt\_loans30 as equal as maxamnt\_loans90 values for that particular row . Now our maxamnt\_loans30 looks as desired.



2.b) The histogram and count Plot shows that the mostly non-payers or defaulters are those people who have a loan lower amount , from figure below will tell us that defaulters are more when the amount is less that is 6 and count of defaulters are less when the amount is 12

The histogram shows defaulters(Blue line) whose loan amount is lower(6).

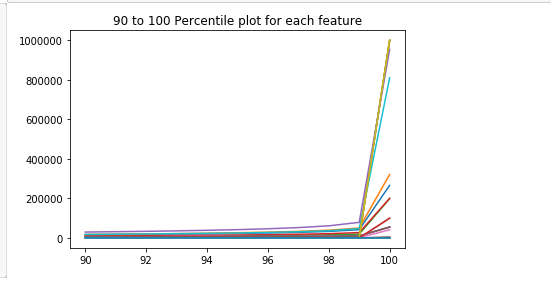
3] Bivariate Analysis:-

We have also performed BoxPlot on the dataset , in order to check the outliers and found there are outliers in several feature which has been treated and will be discussed in **Outlier Treatment** sectionbelow .

**OUTLIER TREATMENT**

As discussed above that plotting BoxPlots we have found that dataset has several features which has outliers, now our challenge in this dataset is how to treat them . As this highly unbalanced dataset, we cannot use use Zscore filtering as this will delete more data which will not be efficient for our modelling and prediction.

Therefore to treat the outliers we have checked the percentile from 90 to 100 for each feature as shown below:-



From the above plot we can see that mainly the values got exploded after 99th percentile which created outliers or separated the data from th usual desired values.

Now to treat them the outliers we have made the values above 99th percentile as equal as 99th percentile which will not only treat outliers but also save the data information by not removing the rows belonging to outliers.

**Multicollinearity Test**

There are various feature in the dataset which has same and equal trend in the contribution to label (class). Therefore in this features having collinearity > 0.95 , are removed as it will reduce the dimension of the dataset and will help in better performance of the Model.

As we have already discussed , that this dataset is highly unbalanced , therefore we cannot go for modelling with this unbalanced dataset . To avoid this issue we need to balance the dataset and for that we have use **SMOTE (OverSampling)**

***Before Over Sampling, counts of label '1': 183431***

***Before Over Sampling, counts of label '0': 26162***

***After Over Sampling, the shape of train\_X: (366862, 31)***

***After Over Sampling, the shape of train\_y: (366862,)***

***After Over Sampling, counts of label '1': 183431***

***After Over Sampling, counts of label '0': 183431***

**We have successfully balanced the dataset using SMOTE as final count for both the labels is *183431***

**Modelling**

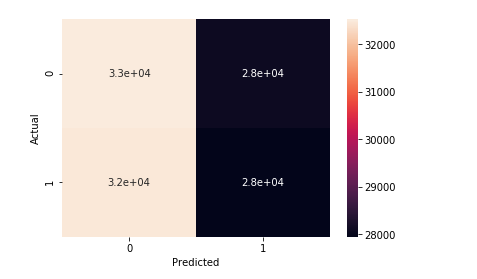
In this section we will discussed the model that we have performed on this dataset , its performance matrix , confusion matrix and at last this section will also show which model is considered best for this model and based on what criteria .

Note:- We have used random search for all the models.

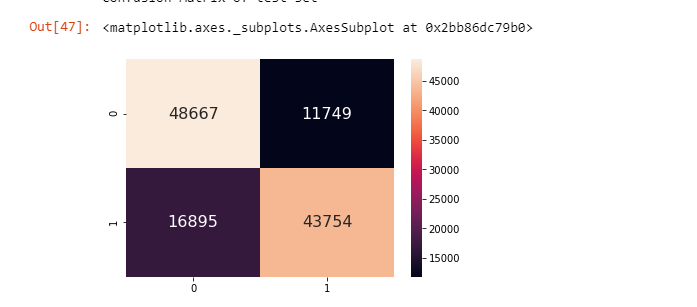
***Lets Start!***

***Lets see the confusion matrix and auc score for all the models***

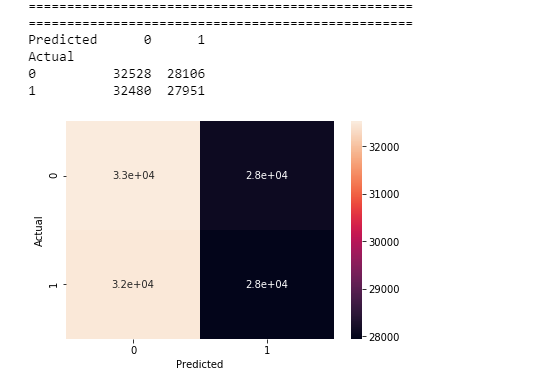
***1.Naive Bayes –***



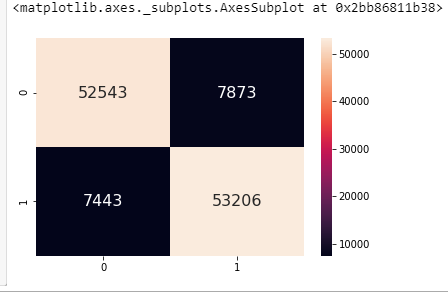
***2.SVM :-***



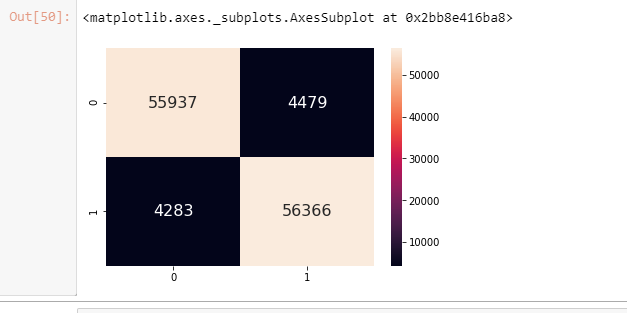
***3. Logistic Regression:-***



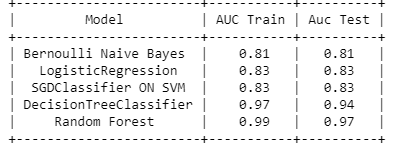
***4.Decision Tree***



***5.Random Forest***



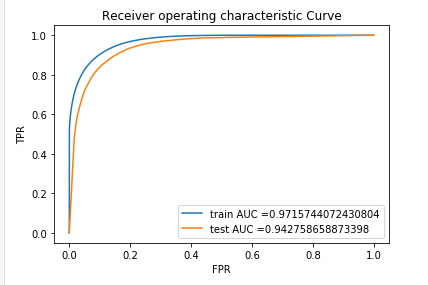
**Pretty Table for AUC SCORE:-**



After observing the above statistics, Decision Tree comes out to be our best model as the AUC score is more significantly high.

Therefore, we will be taking Decision Tree (with all its parameter tested above) as our best model and will use for prediction for the unseen data

The AUC ROC curve for the decision tree is shown below



***FINAL PREDICTION***

***By using Decision Tree as final model we have predicted the probability of defaulter and non defaulter on the unseen dataset (Test data) and final data frame for the actual label and predicted probabilities are shown below .***

