

**Micro Credit Loan**

Submitted by:

Prince aggarwal

**ACKNOWLEDGMENT**

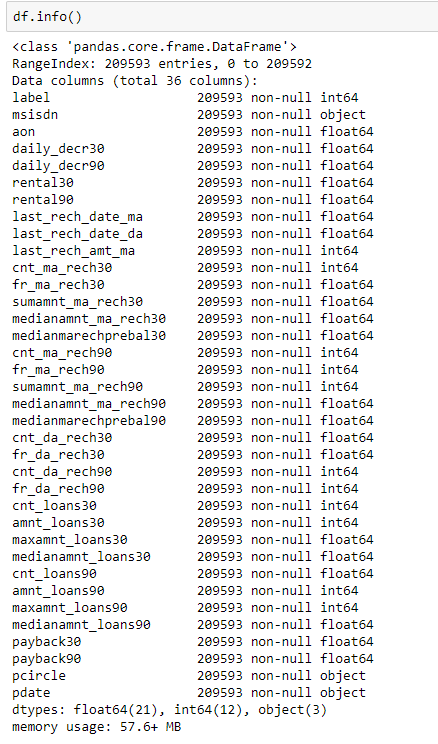
I would like to express my sincere appreciation to all those who provided me the prospect to take a step in this project. I gone through google and git-hub for the references and my old project algorithms helps me understanding the way to solve this use case.

**INTRODUCTION**

* Business Problem Framing
  + There are multiple telecommunication company that are providing micro loan credit to their customers which help the low financial population not to disconnect from the commination world. So, in this use case also there are multiple customers which taken loan and some don’t so I have to predict by taking important features that whether in future customers having this type of inputs and history will able to pay back the loan or not. So, this a real-world use case which will help telecom industry to provide the loan to which type of customers.
* Conceptual Background of the Domain Problem
  + From my prospective, primarily is to take the consideration of the region that from which place the customers are belonging. Because in the remote area most of the customers required the Micro credit and also check their history whether they are able to pay back the loan. Since we know that telecom sector is very much competitive so this data is very helpful in understanding the problem for the lower class people specially by providing them the facility of network and the credit amount provided by the help of MFI and MFS. From this data we get to know that what the criteria to become defaulters and successor are. And the useful information from the data to know how much amount people spend on data recharge or on the main balance recharge.
* Review of Literature
* From the dataset I get to know that it is a classification problem and there are two categories which are successor and the defaulters. And there are so many features which help to find it.
* Micro Credit Loan policy is offered by MFI which is very useful for the low income population, In this current era everyone need to connect virtually so this policy provide the option to low income people to take a next step in this, In India also Jio and BSNL are providing this type of facility to their low income customers which are in need of loan if they don’t have sufficient amount and can payback the amount.
* Motivation for the Problem Undertaken
  + From this project I get to know of different kind of information every recharge done by the user on which kind of recharge user is using mostly and the data service or the main balance the frequency of recharge in 30 day or 90 days. It is really quite interesting to know that each column contributed to make you close to know more about the data and in prediction you can do in many ways

**Analytical Problem Framing**

* Mathematical/ Analytical Modelling of the Problem
* Dataset contain around 209592 rows and 36 columns which make this dataset huge and required lot of feature engineering before applying the proper model on it including an unknown column which is completely irrelevant to dataset .
* There are 0 null values in the dataset
* The correlation heatmap shows that only few columns are correlated i.e. something around 10
* Also, there are lot of outliers in the dataset which need to taken care of and most important is that most of features contain negative values which are not required according to requirement of the project use case.
* The dataset is unbalanced as the target variable label is unbalanced through label=1.
* model is bias towards the label as 1, so this thing is also need to take care, so I used Tress which are not sensitive to imbalance dataset and does not bias.
* There is proper mention that the loan amount is 5 and 10 and the payback amount is 6 and 12, but in the dataset this feature includes 0 value also which is consider as error in the dataset.
* Data Sources and their formats
* Data I get form the Flip Robo the format was in CSV (Comma Separated Values).The number of columns and row are 209593 and columns are 36.
* In the dataset there are no null values that is good sign that gathering this data all fields and column are taken care off.



So above image described the dataset i.e. total features, total count, their types and memory usage of the dataset.

* Data Pre-processing Done

There were no null value was present in the dataset but there are some outliers needs to be treated as we can not remove them using z score method because they will remove a large amount of data and we can not afford to do so because we can not loose more than 7 % data therefore all the outliers values are replaced by their upper bridge value . Then I used updated data for the correlation for splitting it into x and y with the help of standard scalar it will transform the data in such way that its distribution will have a mean value 0 and standard deviation of 1. In case of multivariate data, this is done feature-wise (in other words independently for each column of the data). There are multiple features which are highly positive corelated with each other so by plotting the correlation heatmap I tried to remove the high positive correlated feature and using one of them.

* Data Inputs- Logic- Output Relationships
* By using seaborn heatmap I tried to plot correlation heatmap as well as the null value heatmap which show the covariance of each feature to each other and also clarifies that there are no null values.
* There are 10 features which are positively correlated with each other more than 80%, means if one feature is increasing the second one is also increasing,
* So, these features need to remove which reduce the problem of overfitting.
* State the set of assumptions (if any) related to the problem under consideration

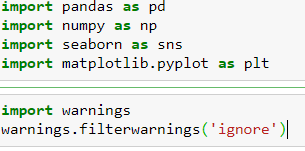
I assumed that all the features which contain the currency should be in same scale and their currency should be same which will help in feature scaling.

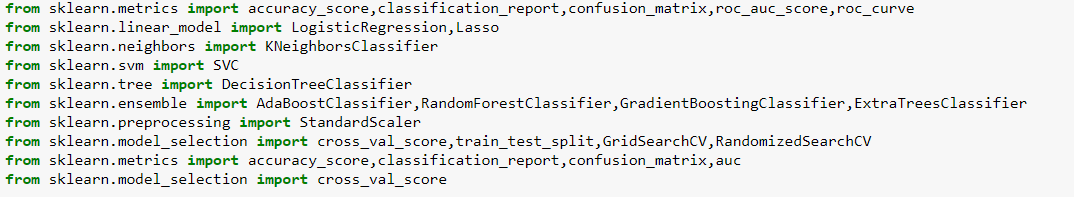
* Hardware and Software Requirements and Tools Used

**Hardware** – Laptop

**Software** - anaconda jupyter notebook

**Libraries and tools** –





**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

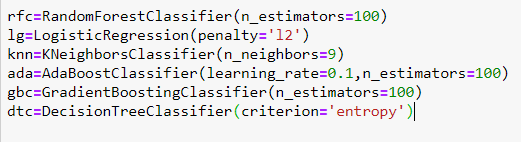
Diagram

Description automatically generated

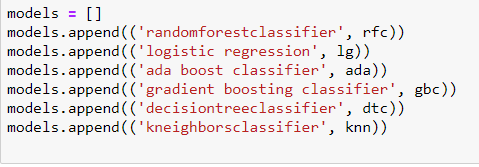
**Descriptive statistics** are used to describe the basic features of the data in a study which are mean count max standard deviations 25% , 75% , 50 % it all help me to understand the data in terms of statistically for the problem solving

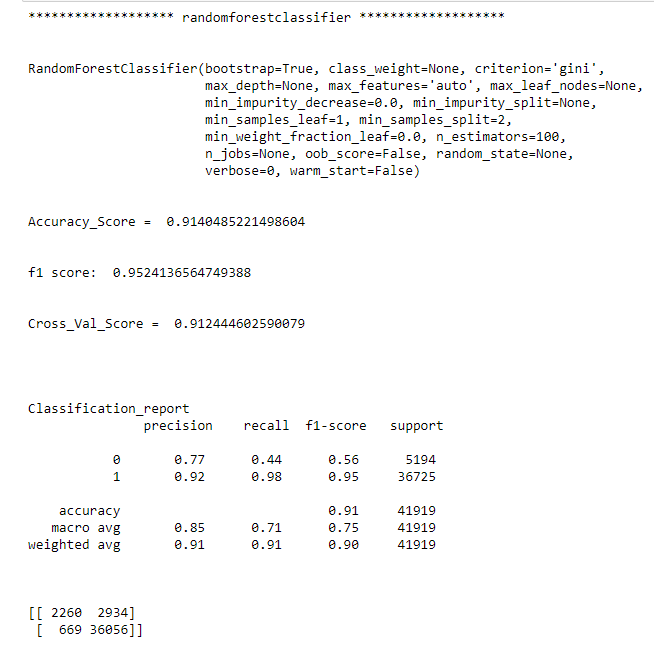
* Data Preparation includes the data cleaning and describing the data which I followed and had done and remove the error data and replace the negative values in the dataset.
* EDA involves the visualization which is helpful to get more insight from the data and the outliers and which features are more important and highly covariant.
* Modelling involves creating the model with suitable algorithm which provide the best result, I tried multiple algo and also apply hyperparameter tuning and cross validation as the data set is imbalance so that’s why cross validation I used, Also to maintain the balance of the dataset I used SMOTE Tomek which handle the imbalance dataset.
* Model Evaluation, for this I used confusion metrices and mainly focus on False positive and tried to reduce the False positive which is type 2 error and plot ROCAUC curve which is also covering most of the area under curve.
* Testing of Identified Approaches (Algorithms)

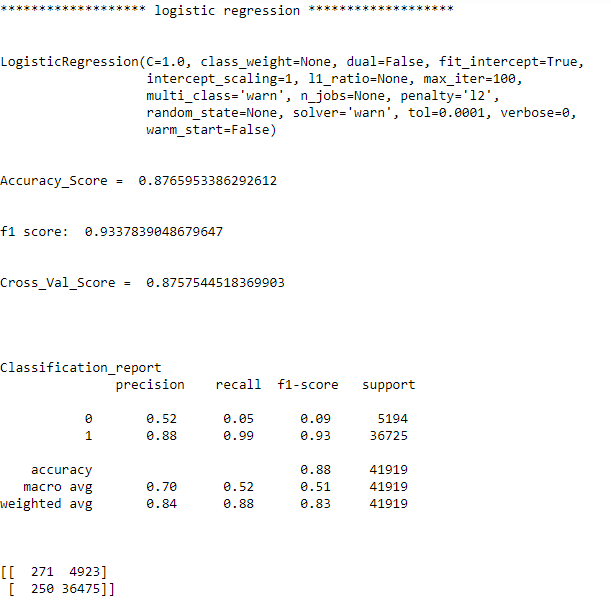
**Algorithm Used:**

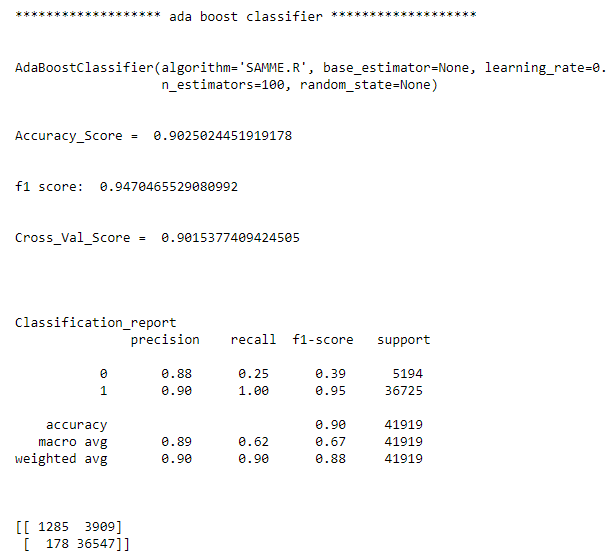


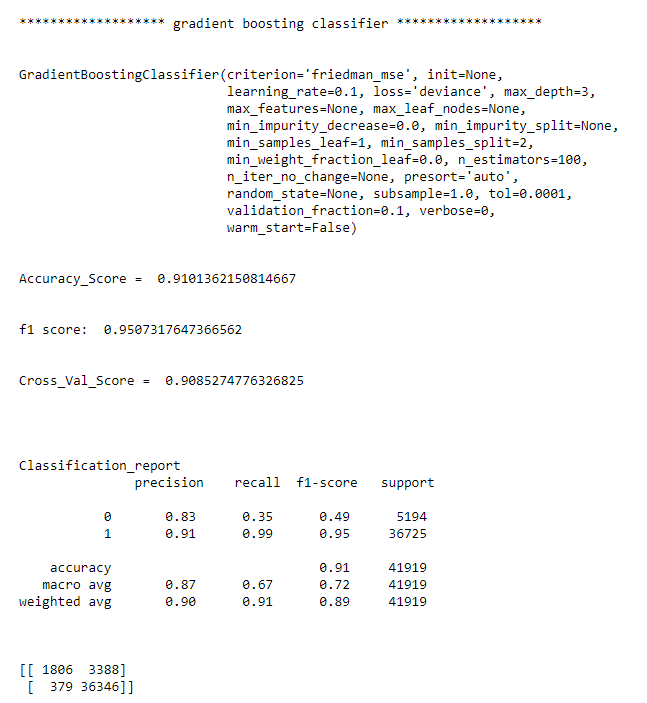
* Run and Evaluate selected models

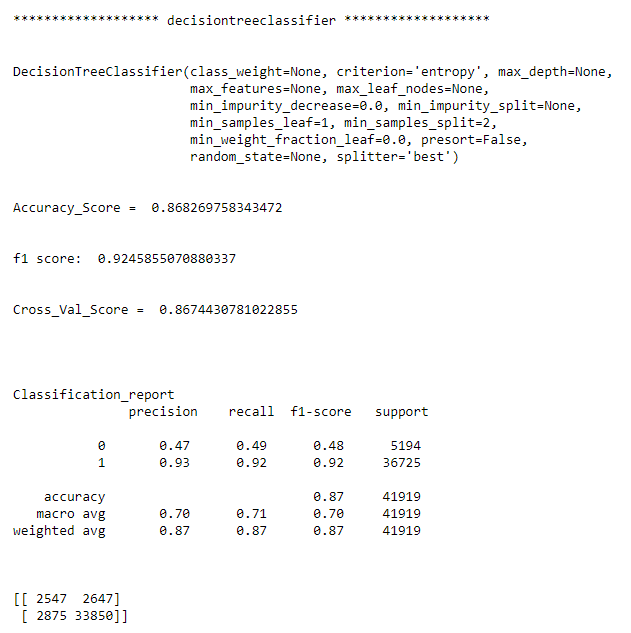


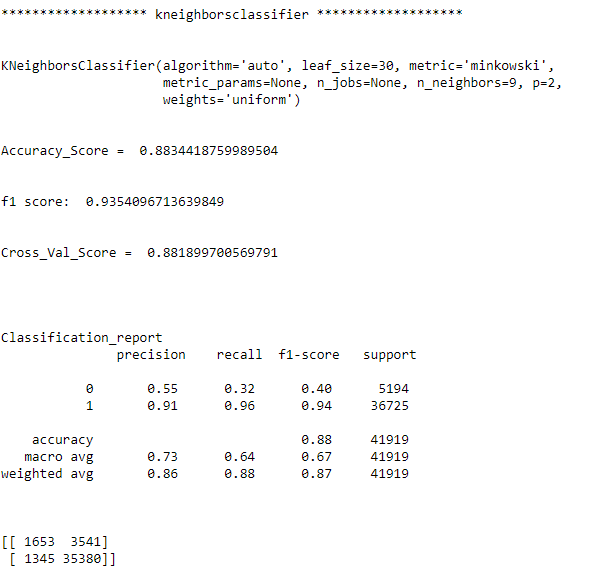


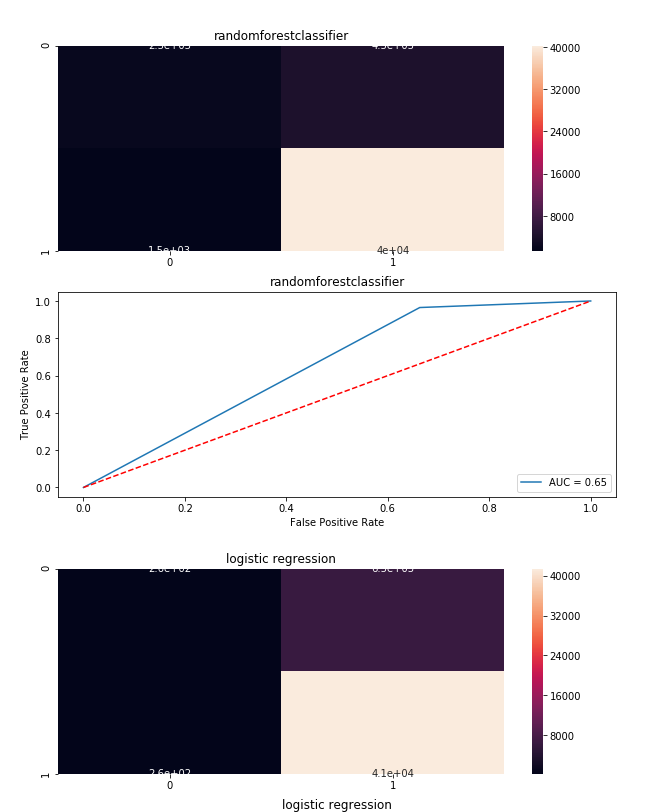


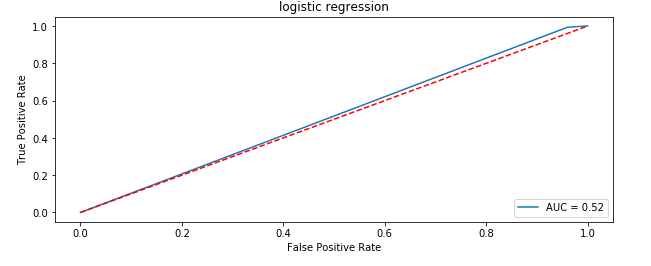


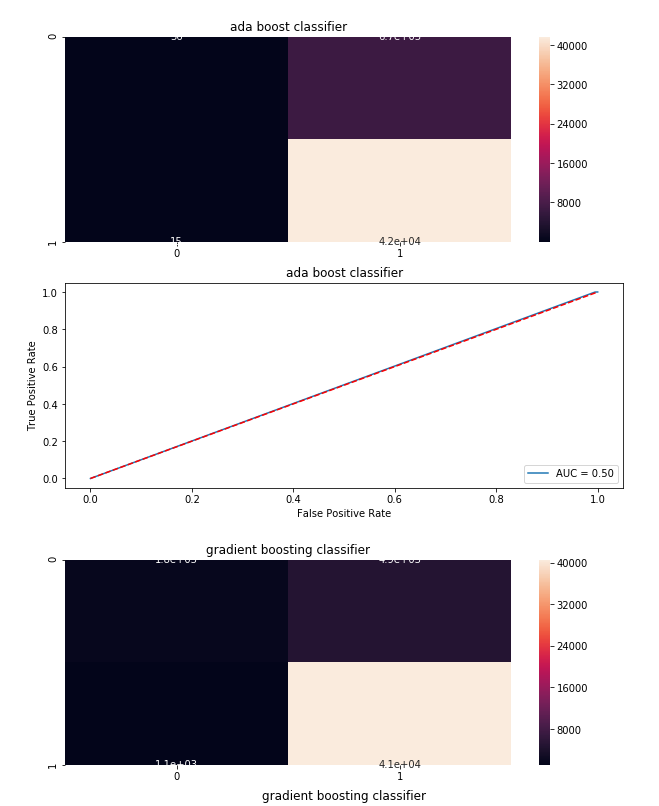


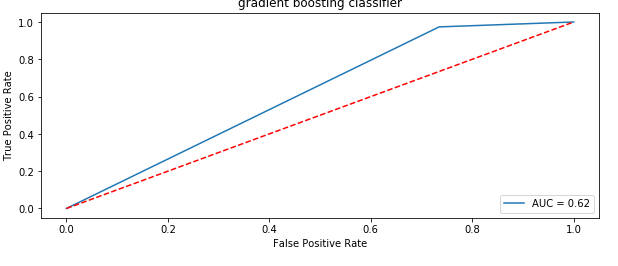


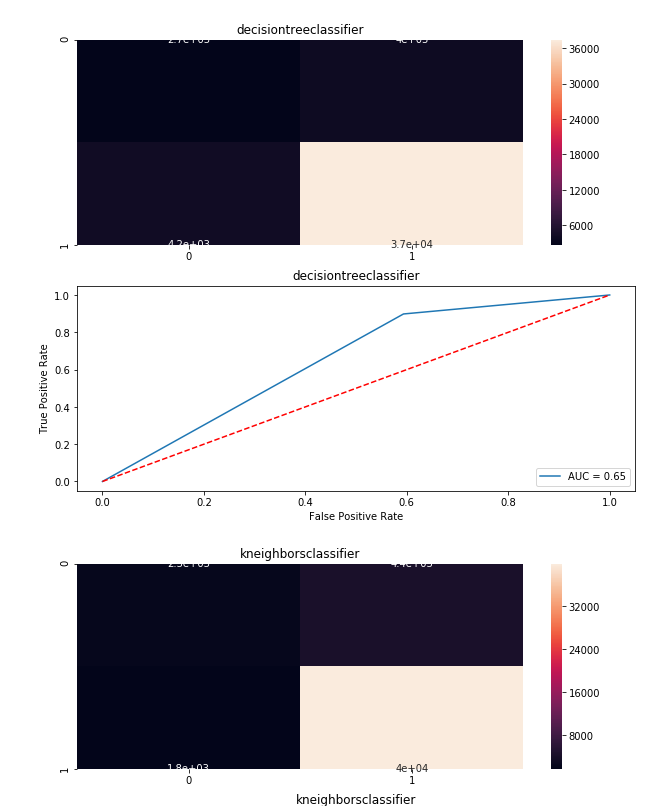


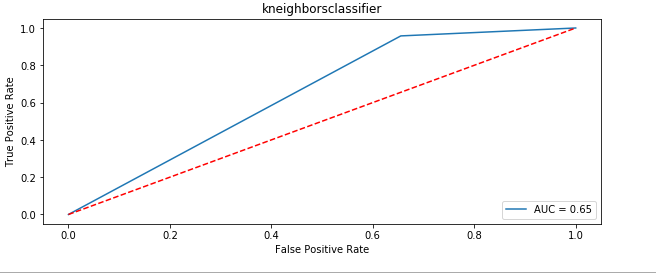


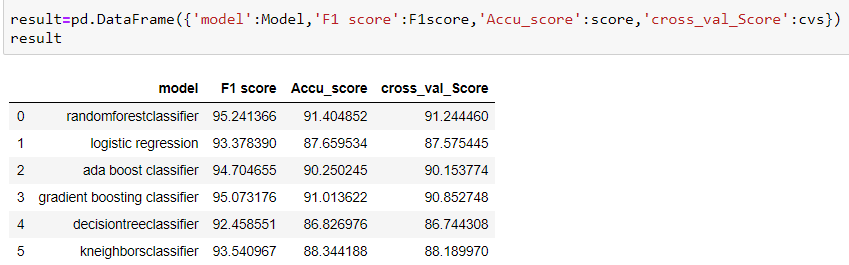




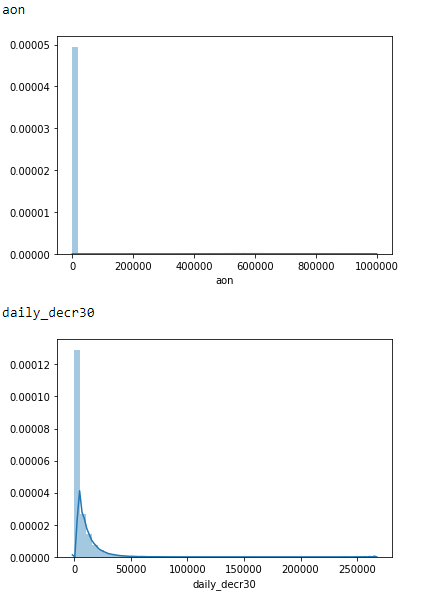


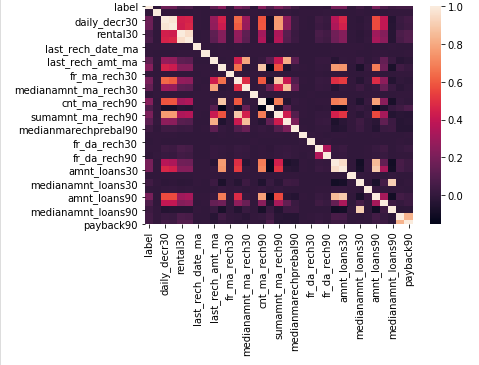




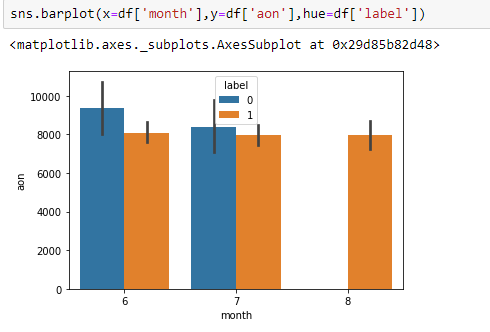


* Key Metrics for success in solving problem under consideration
* In classification problem there are various metrics that are accuracy score, confusion matrix, classification report, F1 score ,cross validation which help to check the efficiency of the model
* Precision: can be seen as a measure of quality, **higher** **precision** means that an algorithm returns more relevant results than irrelevant ones
* So, in this case accuracy score is good but most important is confusion matrix in which we must decrease the False Positive that is type 2 error.
* Visualizations
* Bar Plot
* Count Plot
* Line Plot
* Box Plot
* Heat Map
* Distplots

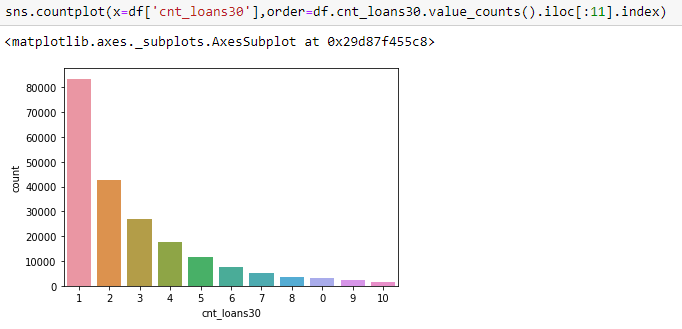


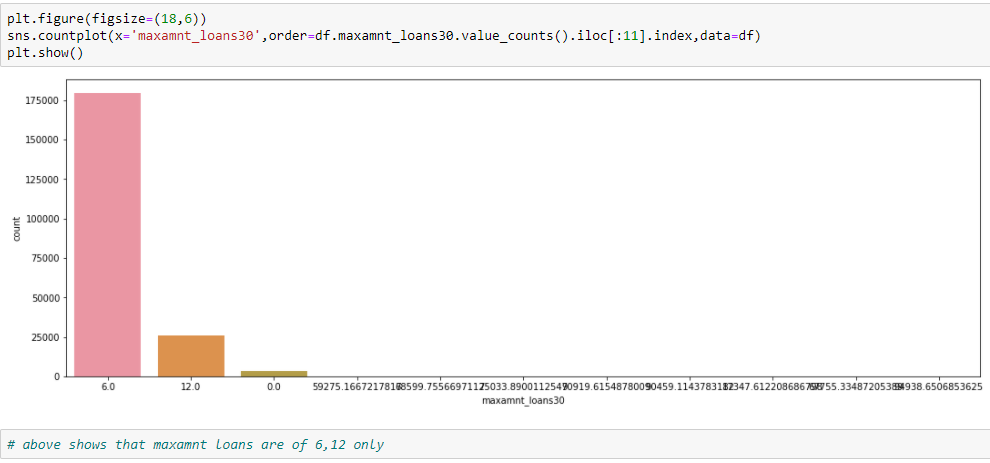


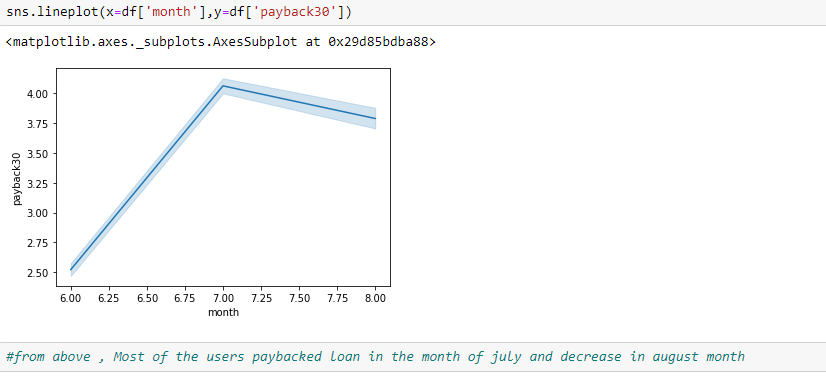


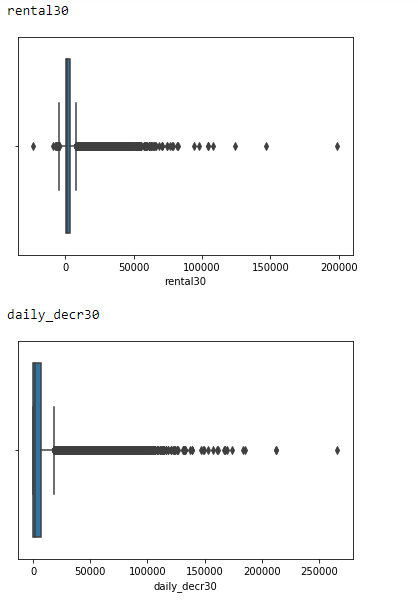


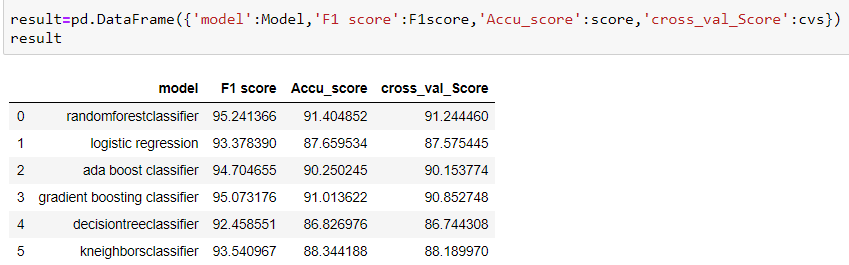










* Interpretation of the Results
* the correlation among different columns since darker side represents the negative correlation and the higher side represent the positive correlation.
* From the above graph we can clearly see that there is no defaulter in 8th month i.e. August month of 2016 year.
* In June month there are more defaulter as compared to non-defaulter and their age on cellular network is also more.
* In the above count plot, most of the user take loan 1,2 and 3 time, these cover 90% of the total data. So, the data is for one year that is 2016 and especially for 3 months that are June, July an August so the count of taking loan is less.
* Clearly visible outliers in the boxplots of rental30 and daily\_decr\_30
* Result : 

**CONCLUSION**

* Key Findings and Conclusions of the Study
* From this dataset I get to know that each feature play a very import role to understand the data. Data format plays a very important role in the visualization and Appling the models and algorithms.
* From the statistical analysis it is clearly see that there is difference between the 75th percentile and max so this is an indication of the outliers in the dataset.
* There is no defaulter in 8th month i.e. August month of 2016 year
* In June month there are more defaulter as compared to non-defaulter and their age on cellular network is also more.
* Removing unwanted column like msisdn number which does not play any imp role because that is unique.

* Learning Outcomes of the Study in respect of Data Science

the power of visualization is helpful for the understanding of data into the graphical representation its help me to understand that what data is trying to say, Data cleaning is one of the most important step to remove missing value or null value fill it by mean median or by mode or by 0. Also, because of outliers and some error in the data set lot of other statically analysis and function are used to make the data suitable to feed for the various model. Because of large data set time of execution is very large which may up to 3 hours also.

In my finding, for imbalance dataset Random Forest perform not as good as compared to other decision tree, as decision tree is because of using the tree internally it may also handle missing values and outliers and scaling is also not required.

* Limitations of this work and Scope for Future Work

The data is only for year 2016, if the data contain more information about 3-4 years then it will more precise and the output is more valuable and accurate. To improve, there need to gather more data for defaulter also, so that by using their inputs we can able to predict the defaulter in future also.

These large datasets takes a very large time for the execution as they need top end specifications for the machines