**Micro Credit Defaulter Project**

Submitted by:

Tushar Saini

**Acknowledgement**

I would like to express my sincere gratitude to fliprobo technologies for supporting me in this project. I wish to express my sincere gratitude to my SME, Ms. Swati Rastogi, for his enthusiasm, patience, insightful comments, helpful information, practical advise and unceasing ideas that helped me tremendously at all times in my project time. Her immense knowledge, profound experience and professional expertise in Data Science has enabled me to complete this research successfully. Without her support and guidance, this project would not have been possible.

I also wish to express my sincere thanks to the DataTrained for helping me to get trained in the field of Data Science. Without them I don’t know anything about Data Science. They help me to learn the Data Science from scratch. The faculty their helped me a lot in every aspect.

I also like to thank websites like geeksforgeeks.com, towardsdatascience.com, etc for providing the information regarding many topics of data science.

1. **Introduction:-**

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 donorsare 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 MFShas 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.

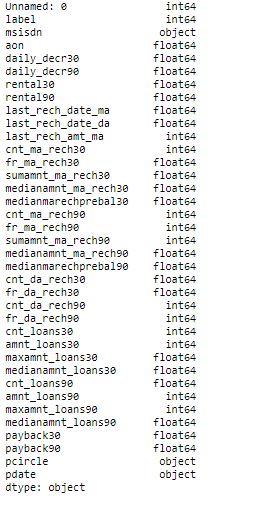
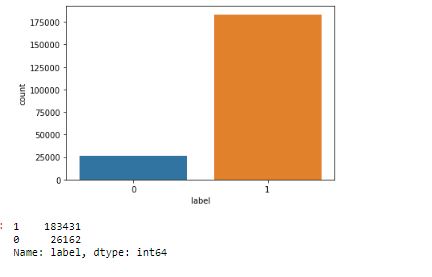
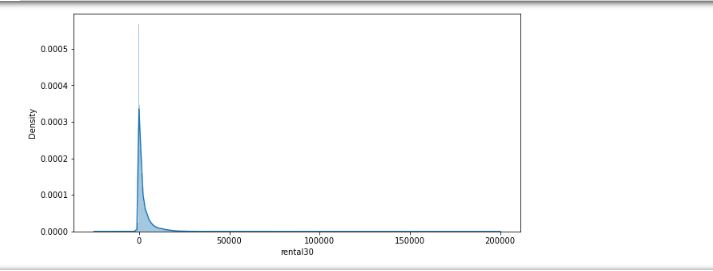
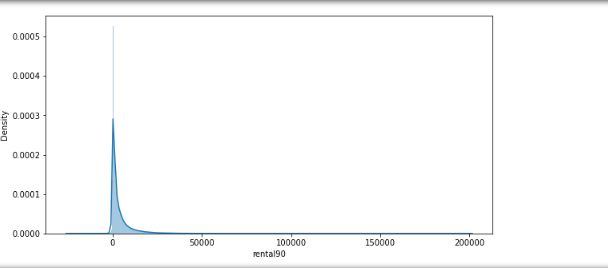
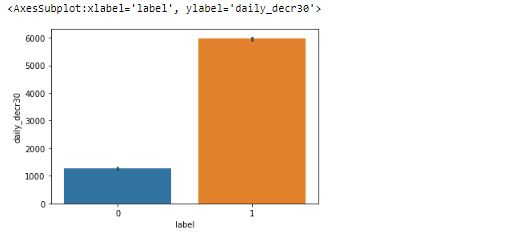
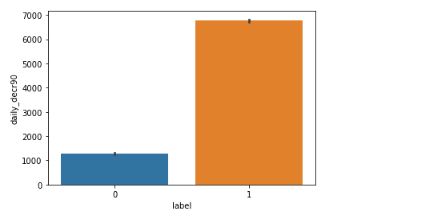
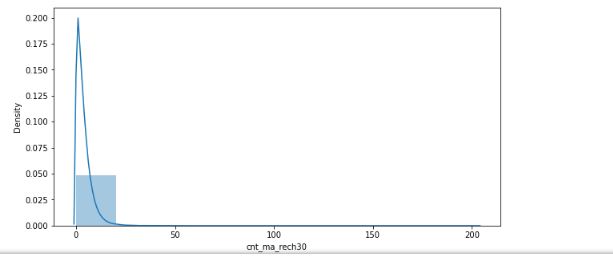
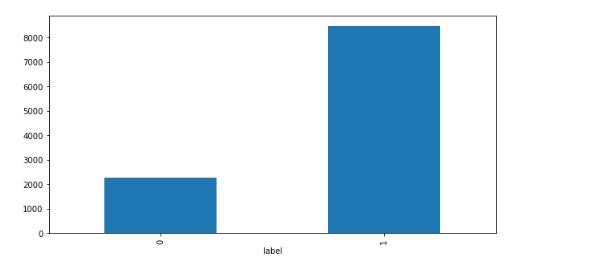
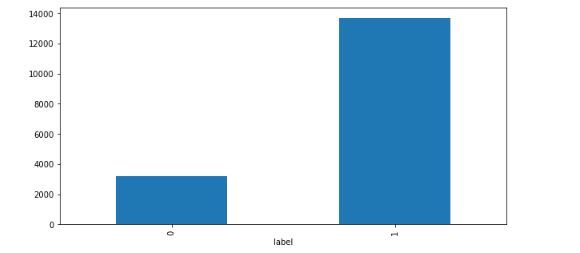
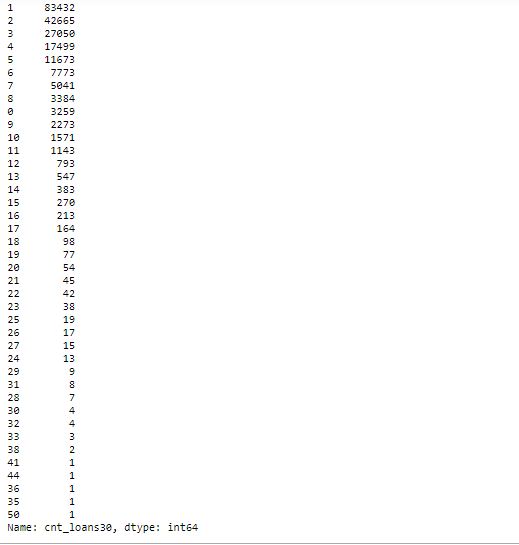
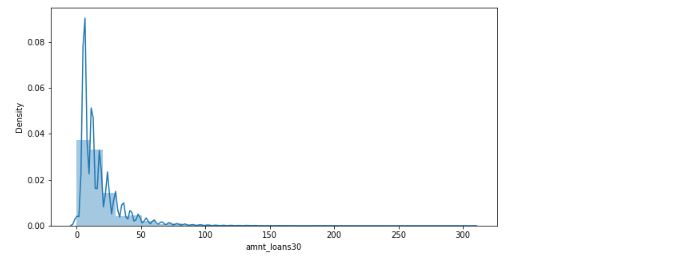
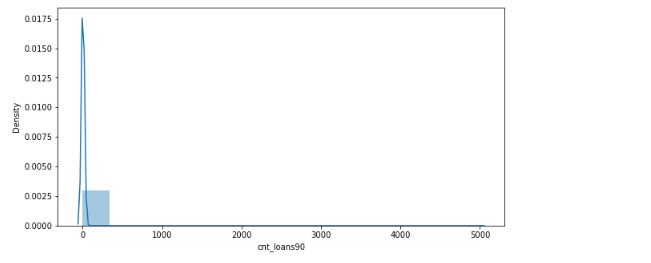
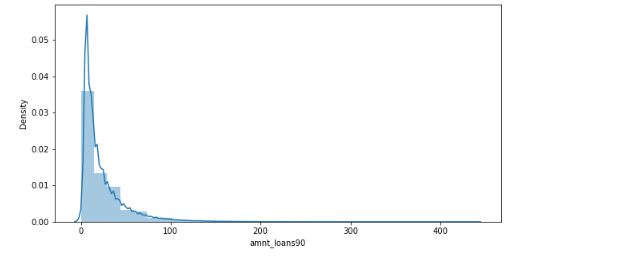
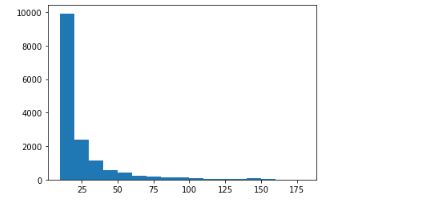
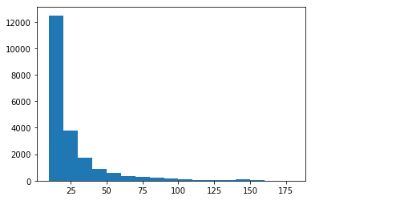
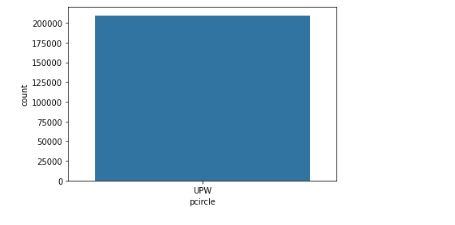
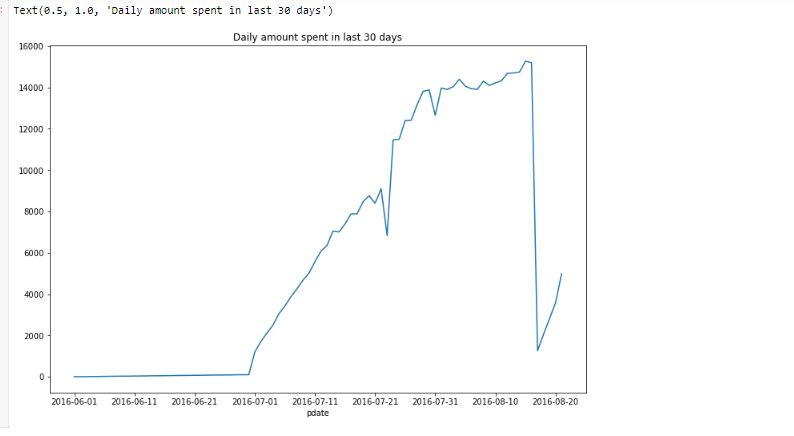
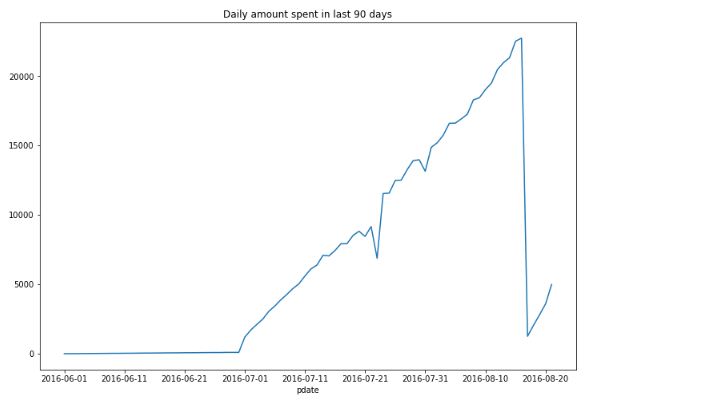
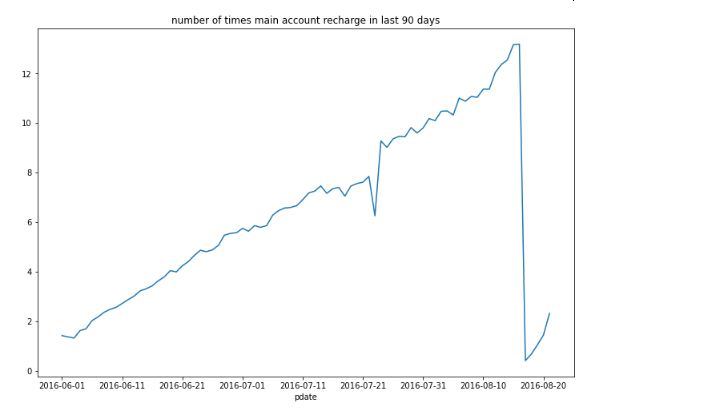
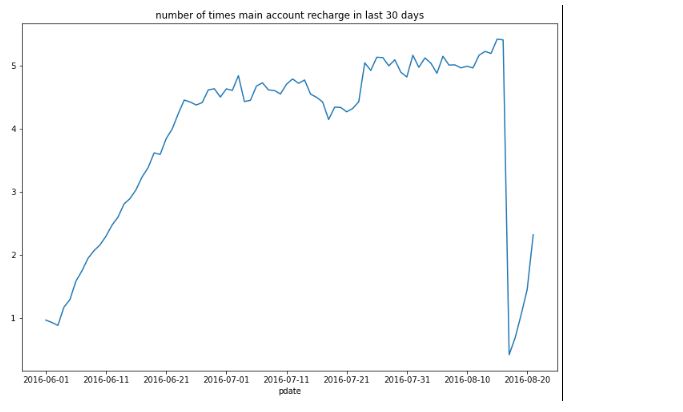
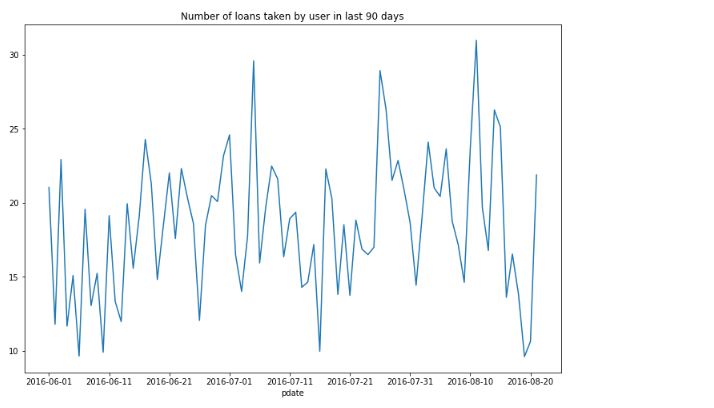
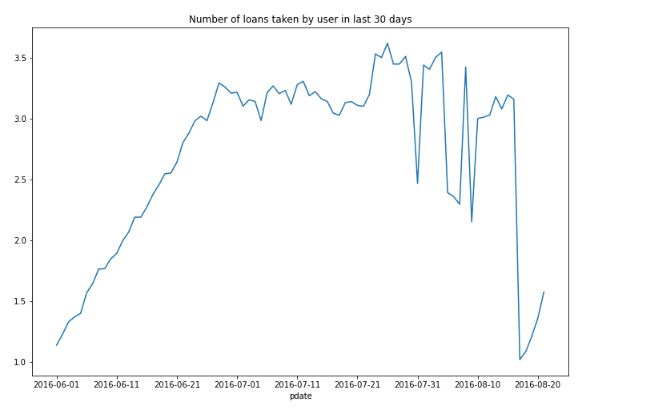
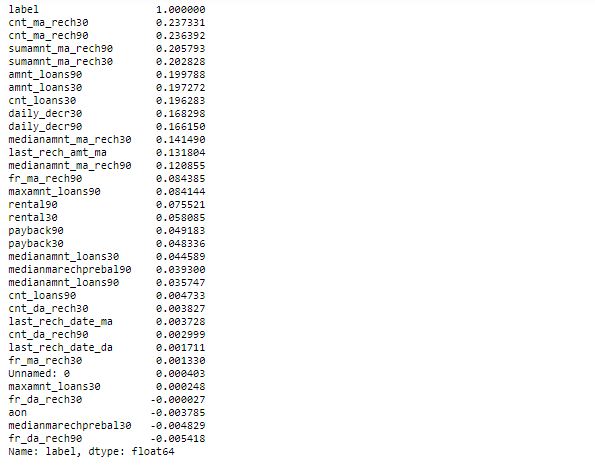
They arecollaborating with an MFI to provide micro-credit on mobile balances to be paid back in 5 days. The Consumer is believed to be defaulter if he deviates from the path of paying back the loaned amount within the time duration of 5 days. For the loan amount of 5 (in Indonesian Rupiah), payback amount should be6(in Indonesian Rupiah), while, for the loan amount of 10(in Indonesian Rupiah), the payback amount should be 12(in Indonesian Rupiah).

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.

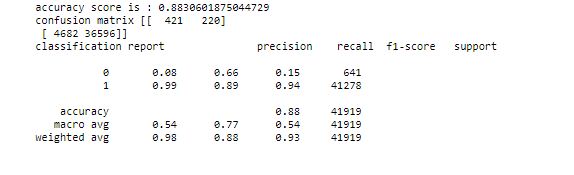
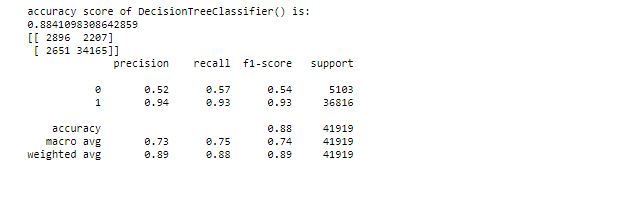
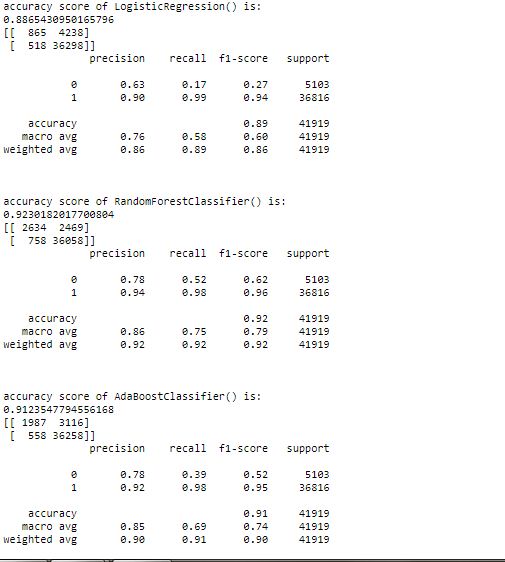
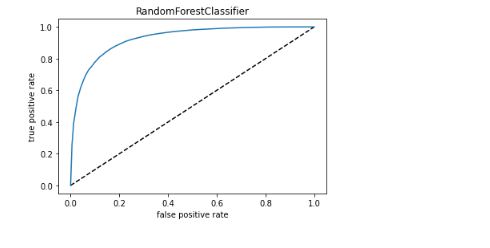
1. **Analytical Problem Framing:-**

* **Data Visualization :-**

By analyzing the data, I have found some important features of that dataset. The analyzed information that I found are given below:-

* There are 209593 rows and 37 columns in the dataset and it doesn’t have any null values.
* I also observe that the datatypes of columns in the dataset are a mixer of ‘float’, ‘int’, ‘object’.
* Out of 209593 clients, 183431 are non-defaulter that means they have payed the loan back within specified period of time. 
* Majority of the people have very less(nearly zero) account balance from last 30 days and 90 days. 
* Average daily spent in last 30 of a 'non-defaulter' is more than a defaulter.
* For a non-defaulter the average spent in 30 days is approx 6000 while it is only approx 1500 for a defaulter. 
* Average daily spent in last 90 of a 'non-defaulter' is more than a 'defaulter'.
* For a non-defaulter the average spent in 90 days is approx 7000 while it is only approx 1500 for a defaulter.
* Majority of the people have recharged their account almost 20 times in the last 30 days and 90 days.
* The non-defaulter made a total amount of recharge in main account over last 30 days is over 8000
* The defaulter made total amount of recharge in main account over last 30 days is approx 2500.
* The non-defaulter made a total amount of recharge in main account over last 90 days is almost 14000.
* The defaulter made total amount of recharge in main account over last 90 days is approx 2500.
* I observe that 83432 times people take the loan 1 time in last 30 days.
* I also observe that the number of times loan taken are increasing, the number of people are decreasing.
* Majority of the people have taken the loan of less than Rs50 in the last 30 days.
* Maximum number no of loans taken by the user are approximately 250.
* Majority of the people had taken approximately 0 to Rs50 amount of loan in past 90 days.
* Maximum number of people had payed back their loan before 25 days in last 30 days.
* Maximum number of people had payed their loan back before 30 days in last 90 days.
* There is only 1 telecom circle that is UPW.
* From 21-07-2016 to 20-08-2016 the daily amount spent is very high in last 30 days.
* From 21-07-2016 to 20-08-2016 the daily amount spent is very high in last 90 days.
* After 21-06-2016 the number of times the main account got recharged is increasing in last 30 and 90 days.
* In last 30 days from 01-06-2016, the number of loans taken by user are increasing.
* We can also observe that in last 90 days the frequency of number of loans taken by user from 01-06-2016 is fluctuating. So it keeps increasing and decreasing. 
* I observe that 'Number of times main account got recharged in last 30 days' is the most positively correlated column with our target column that is 'label'.
* We can also observe that 'Frequency of data account recharged in last 90 days' is the most negatively correlated column with our target column that is 'label'.
* **Data Pre-Processing:-**
* Now we can drop column 'pdate' as we make separate date, month and year.
* We can also drop column 'unnamed: 0' as it is for no use.
* We can also drop 'msisdn' as the mobile number will be different for every customer.
* We can also drop 'pcircle' as there is only 1 value in that column.
* The data is highly distributed. So we use ‘Standard Scalar to normalize the data.
* After that we Saw there is skewness present in the dataset, so we used power transformation and log transformation method to remove the skewness in the dataset.
* We didn’t remove the outliers from the dataset as after removing it, we are losing around 20% of our data.
* After that we saw there are some null values present in the dataset, So we replace the null values with the mode as mode is the most occurring value in that column. We are assuming that the the null value is only the most occurring value in the column.
* **Hardware and Software Requirements and Tools Used:-**
* We used anaconda jupyter notebook to do research on this project.
* I use python to do my coding for the dataset.
* We use different kind of libraries present in the jupyter notebook like pandas, numpy, sklearn, seaborn, matplotlib, joblib, etc.
* I have used dell inspiron 1440 which has 3GB Ram and windows 7.
* I use pandas to import that data make it in a dataframe.
* Sklearn is used to import all the model and methods that are used for pre-processing and model building of the dataset.
* Matplotlib and Seaborn are used to plot various graphs and figures to visualize the data.

1. **Model/s Development and Evaluation:-**

* For building a machine learning model first I split the dataset into independent and target variable.
* Independent variable that is ‘x’ contains all the other columns except ‘label’ because ‘label’ is our target value which we have to predict.
* Target variable which is ‘y’ contains the ‘label’ column.
* Now I find the best random state for one model and I use that random state for other models also.
* First I use ‘Logistic Regression’ as it is a classification problem. I make a loop and take 0-100 random state. That loop will give me the accuracy score for every random state from 0-100.
* At random state 81 the testing accuracy and training accuracy is highest.
* Now I test this random state with other models also.
* I used a total of 5 models which are ‘Logistic Regression’, ‘Random Forest Classifier’, ‘SVC’, ‘AdaBoostClassifier’, ‘Decision Tree Classifier’.
* ‘Random Forest Classifier’ gives me the best accuracy score that is 92.30%.
* Now I check whether my models are underfitted or overfitted. For that I use cross validation score and taking the k-fold value of 5.
* The Minimum difference in accuracy score and cross validation score is for DecisionTreeClassifier, AdaBoostClassifier, RandomForestClassifier and SVC (0) so they are our best model.
* I am using ‘Random Forest Classifier’ as it is giving me the best accuracy score which is 92.30%.
* We have checked the AUC-ROC Curve to check the performance of the model. It is showing that our model is doing good in distinguishing between the various classes.
* After that I have done hyperparameter tuning of the model to increase the model accuracy score.
* After doing the hyperparameter tuning our model score has increased from 92.30% to 92.35%.

1. **Conclusion:-**
   * From the above project I find that the majority of the people that are taking this micro credits are paying back within the specified time.
   * These micro credits are very helpful for the people especially for the people who are in emergency and need these micro credits on urgent basis.
   * I made a machine learning model to check whether the customer will be paying the micro credit back within time or not. My model is 92.34% correct.
   * I have learned that through visualization we can understand the core of the data. Visualization is very important for a data science project as it will give us information about the dataset.
   * I have used 5 algorithims. What I have learned that it might be possible some algorithms can be used for a specific problem. So we have to check what is the problem through which we can choose our algorithm.