**Project Report:**

**“Micro- Credit Defaulter”**

Creator’s Name:

Trupti Badole



**Acknowledgment:**

It is a Great pleasure to express my Gratitude to Team FlipRobo, for giving me opportunity to work on a fantastic project, which helped me in improving my knowledge, coding skills and my analyzation skills.

Team FlipRobo also gave me opportunity to build PowerPoint Presentation and Project Report, which will help me to share steps taken while building the entire model. It will also help me with the story telling capabilities which are required to be a good Data Scientist. It is just a start I will try to convey my understanding of the project through this report.

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 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.

**Ⅵ**

**Ⅳ**

Ⅱ

**Ⅲ**

Ⅴ

**Background**

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.

**Motivation:**

I will model this data set micro credit defaulters with all given available independent variables. This model will then be used for management of how the customer is considered as defaulter or non-defaulter based on the independent variables. With the help of this prediction model they can decide accordingly and manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be prediction-based insights to the management to understand whether the customer will be paying back the loan amount within 5 days of disbursement of loan or not.

**Analytical Problem Framing:**

In this particular problem we have label as a target column and it has two classes Label ‘1’ indicates that the loan has been paid i.e., non-defaulter, while, Label ‘0’ indicates that the loan has not been paid i.e., defaulter. So clearly it is a binary classification problem and I have to use all classification algorithms for building the model. There were no null values in the dataset. Also, I observed some columns where I found more than 90% zero values so I decided to drop those columns. If I would have kept that columns then it would have affected my model’s performance. To get better insight on the features I have used plotting like distribution plot, pie plot and reg plot, violin plot. With these plotting I was able to understand the relation between the features in better manner. Also, I found huge number of outliers and high skewness present 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 I tunned the best model and saved the best model. At last, I have predicted the label using saved model.

**Enter Title:**

Here you can enter the method or idea to solve the problem.

Here you can enter the method or idea to solve the problem.

**Enter Title:**

Here you can enter the method or idea to solve the problem.

Here you can enter the method or idea to solve the problem.

**Enter Title:**

Here you can enter the method or idea to solve the problem.

Here you can enter the method or idea to solve the problem.

**Insufficient**

**Improvement**

**Data Preprocessing:**

* As a first step I have imported required libraries and I have imported the dataset using csv file provided by the company.
* Then we did all the statistical analysis like checking shape, nunique, value counts, info and description etc….
* Then while looking into the value counts, I found some columns with more than 90% zero values this creates skewness in the model and there are chances of getting model bias so I have dropped those columns with more than 90% zero values.
* While checking for null values I found no null values.
* I dropped Unnamed:0, msisdn and pcircle column as they were not needed.
* Next as a part of feature extraction I converted the pdate column to day, month and year.
* Also, I have dropped some columns when I tried to remove multicolinearity from the dataset using Variance Inflation Factor.
* Since we had all numerical columns so, I used dist plot to see the distribution of each column data.
* I have used violin plot and reg plot for each pair that shows the relation between label and independent features. Also we tried to observe whether the person pays back the loan within the date based on features.
* In maximum features relation with target I observed Non-defaulter count is high compared to defaulters.

**Observations:**

* We can see data imbalance in our target column which we will rectify later.
* We can Skewness towards the left, in columns aon, daily\_dect30, daily\_decr90, rental30, rental90, last\_rech\_date\_ma, last\_rech\_amt\_ma, cnt\_ma\_rech30, ft\_ma\_rech30, sunamnt\_ma\_rech30, medianamnt\_ma\_rech30, medianmarechprebal30, cnt\_ma\_rech90, fr\_ma\_rech90, sumamnt\_ma\_rech90, medianamnt\_ma\_rech90, medianmarechprebal90, cnt\_loans30, amnt\_loans30, maxamnt\_loans30, cnt\_loans90, amnt\_loans90, payback30, payback90,.
* maxamnt\_loans90 column is skewed towards the right which denotes median is more than mean
* Here we can see that there are only 2 values present in our target column i.e. 1 and 0 and we can clearly see data imbalance in this column. we will balance the data using oversampling method.
* Here we can see Correlation of our target column with other columns present in the dataset.
* Here looking at plot of label and age on cellular network in days, we can say that defaults can happen even if the user is using services from longtime.
* Here correlation of target column label with columns daily\_decr30, daily\_decr90, rental30, rental90, last\_rech\_amt\_ma, cnt\_ma\_rech30, sumamnt\_ma\_rech30, medianamnt\_ma\_rech30, cnt\_ma\_rech90, sumamnt\_ma\_rech90, medianamnt\_ma\_rech90, medianmarechprebel90, cnt\_loans30, amnt\_loans30, amnt\_loans 90, we can say that There are less no. defaulters and most of the users paid there dues on time.
* Here we can see that the user who took loans in last 30 days paid back most of loans on time.
* Here we can observe that user who took loans in last 90 days was able to half of the loans on time while half users become defaulter as they couldn't pay back loan on time.
* Here we can observe that users paid most of the loans on time from total loans taken in 30 days, but they became defaulter in few loans.
* Here we can see that users paid most of the loans on time from the total loans taken in last 90 days, but users became defaulters in paying back few loans.
* Here we can observe that user was able to pay half of the amount of loan taken by him in last 30 days on time.
* Here we can observe that users paid all small amount loans on time, but were able to pay half of large amount loan on time.
* Here we can observe multicolinearity in some columns which we need to handle later.
* here we can observe that our target column label shares very little relation with all other columns present in the dataset.
* Here we can see that our target column label shares positive relation with all the columns except 3 columns.

**Data Cleaning:**

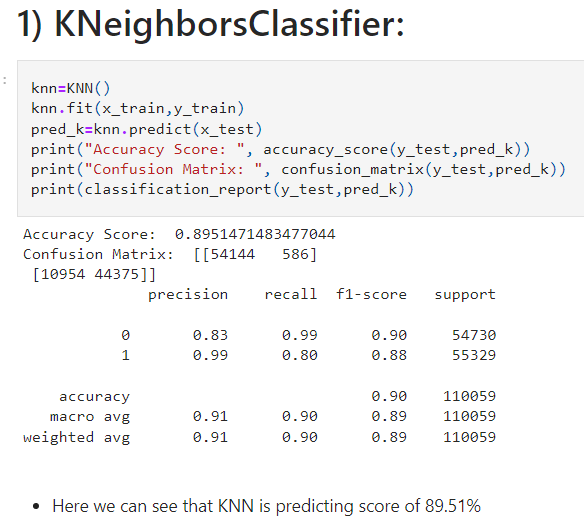
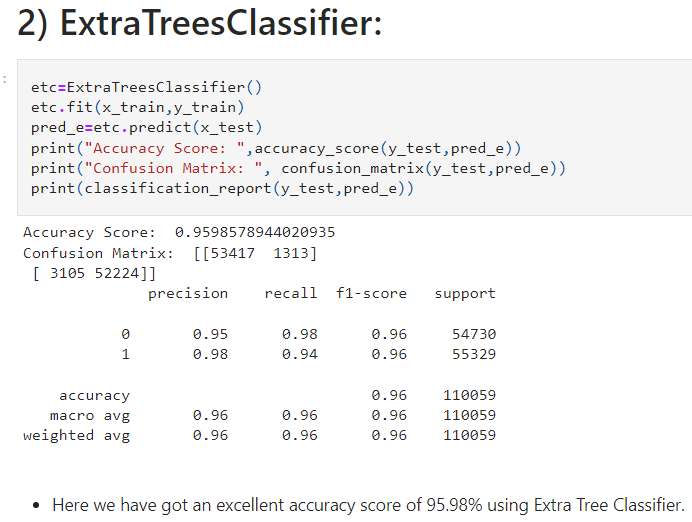
* In the dataset we did not found any null values there were many 0 values, but I found huge amount of outliers and very high skewness.
* To remove outliers we used percentile method. And for removing skewness we used yeo-johnson method.
* Then we used Standard Scaler method to Scale the data.
* As we found issue of Multicolinearity while Multivariate Analysis, we removed Multicolnearity using Variance Inflation Factor.
* While Univariate Analysis we found that our Target column was imbalanced, so we balanced the target column using SMOTE method in Oversampling. Then after Preparing our data we moved to model selection.

**Model Selection:**

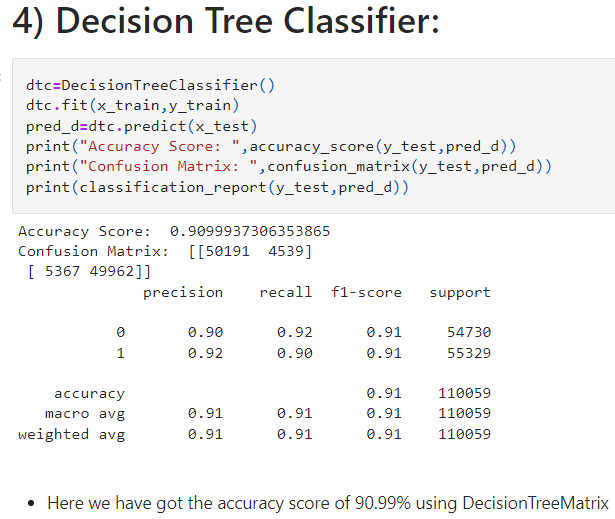
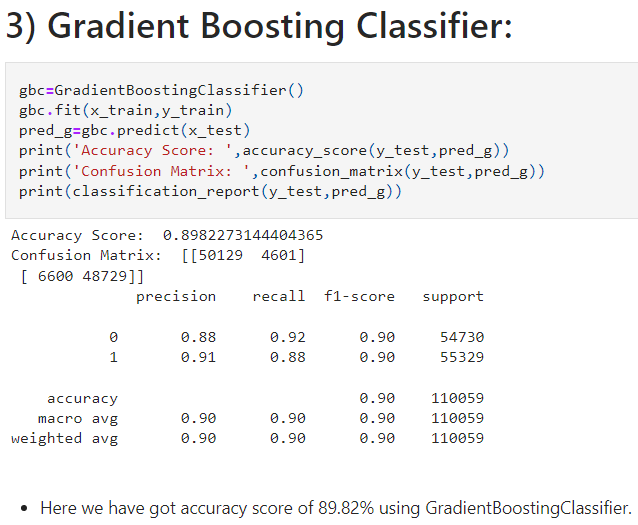
* Since Label was our target column and it was a Categorical column, so this particular problem was a Classification problem. And we used all Classification algorithms to build our model. We Tried multiple models and to avoid the confusion of overfitting we went through cross validation. Below are the list of Classification algorithms I have used in my project.By looking into the least difference of accuracy score and cross validation score I found RandomForestClassifier as a best model. Other models which i tried were:
* KNeighborsClassifier
* ExtraTreesClassifier
* GradientBoostingClassifier
* DecisionTreeClassifier

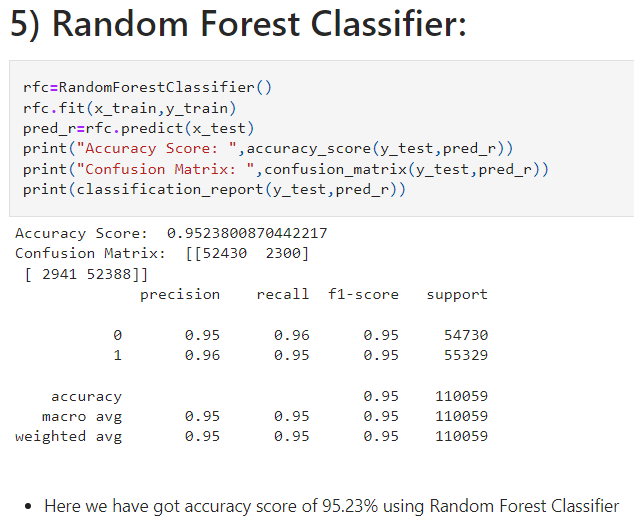
**Models Accuracy Scores:**

**Plan：**

**02**



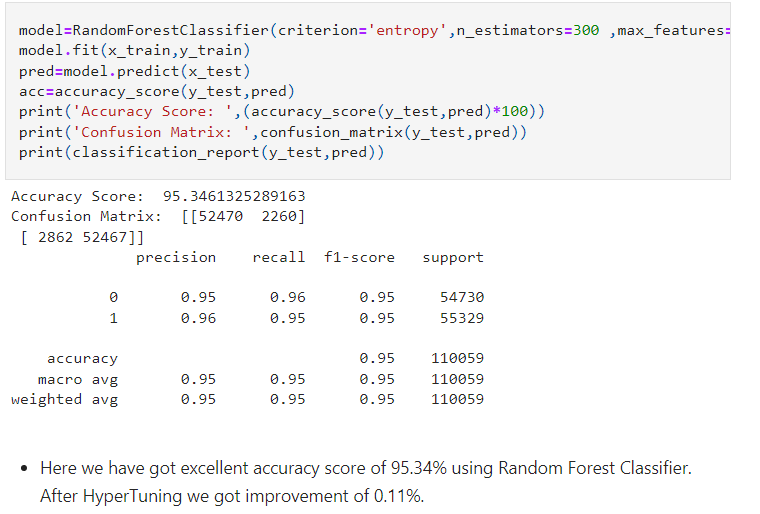


**Plan：**

**Cross Validation Scores:**

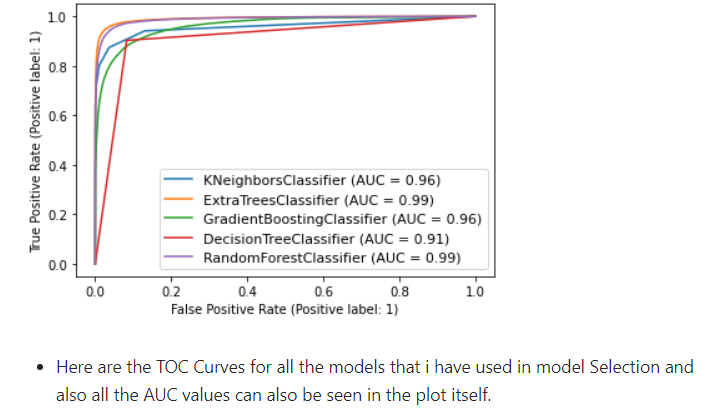
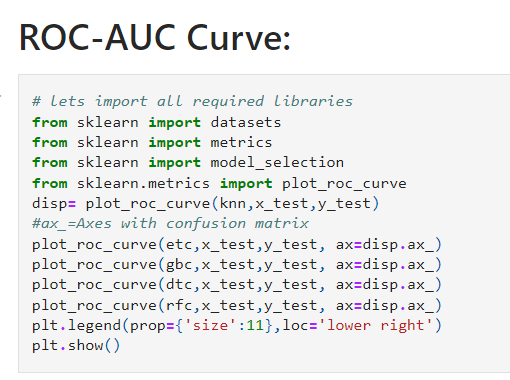


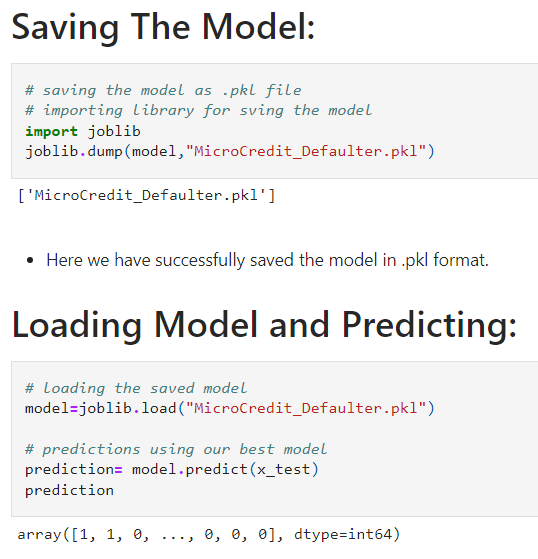
**Hyper Parameter Tuning:**



**03**

**Plan：**

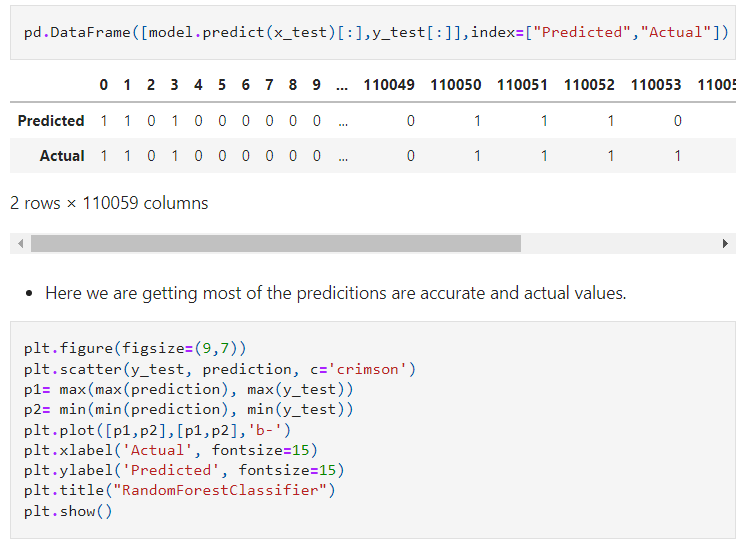


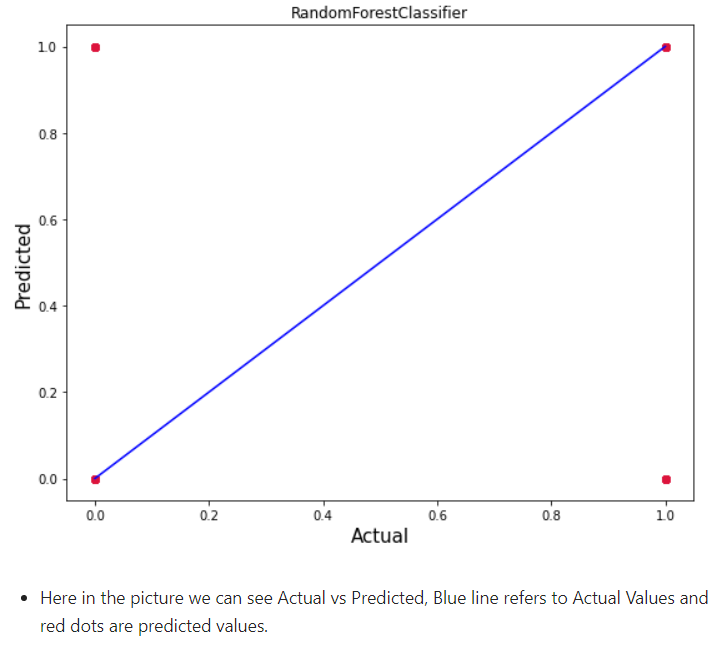


**Plan：**

**05**

**Predictions:**





**Conclusion:**

* In this project report, we used machine learning algorithms to predict the Micro-Credit Defaulters. We used proper procedure to analyze the dataset and finding the correlation between the features.
* Here we selected the features which are correlated to each other and are independent in nature. Visualization helped us in understanding the data by graphical representation it made things easy for us to understand what data is trying to say.
* Data cleaning is one of the most important steps to remove unrealistic 0 values and columns which had morethan 90% 0 values.
* Using these feature we deployed 5 algorithms to find the best model and a hyper parameter tunning was done to the best model and we succeded in improvement of accuracy score.
* Then we saved the best model and predicted the label. Our model’s performance felt good when we saw the predicted and actual values were almost same it felt really good observing good performance by our model.
* To conclude, the Project Micro Credit Defaulter , We hope this study will move a small step ahead in providing some methodological and empirical contributions to crediting institutes, and presenting an alternative approach to the valuation of defaulters.



**06**

**Plan：**



**Enter target title**