**NAME OF THE PROJECT**

**Micro-Credit Defaulter predicationusing Machine Learning  
Submitted by:**

**SAUNAK MUKHERJEE**



**ACKNOWLEDGMENT**

I would like to express my special gratitude to THE “Flip Robo” team, they have given me this opportunity to deal with a beautiful dataset and it has

helped me to improve analyzation of skills. And I want to express my

huge gratitude to Ms. Sapna Verma (SME Flip Robo), she is the person

who has helped me When I am in difficulties I faced while doing the

project. And also A huge thanks to “Data trained” who are the reason behind my Internship at Fliprobo.

References use in this project:

1. SCIKIT Learn Library Documentation

2. Blogs from towardsdatascience, Analytics Vidya, Medium

3. Andrew Ng Notes on Machine Learning (GitHub)

4. Data Science Projects with Python Second Edition by Packt

5. Hands on Machine learning with scikit learn and tensor flow by

Aurelien Geron

6. Stackoverflow.com to resolve some project related queries.

7. Predicting Credit Default among Micro Borrowers in Ghana

Kwame Simpe Ofori, Eli Fianu

8. Predicting Microfinance Credit Default: A Study of Nsoatreman

Rural Bank, Ghana Ernest Yeboah Boateng

9. A Machine Learning Approach for Micro-Credit Scoring

Apostolos Ampountolas

And also thank you for many other persons who has helped me directly

or indirectly to complete the project

Business Problem Framing

A Microfinance Institution (MFI) which is an organization that offers financial to services to low-income populations. MFS becomes very useful when the targeting especially which is the unbanked poor families living in remote areas

Or not much sources of income. The Microfinance services (MFS)

provided by the MFI are Group Loans, Agricultural Loans, Individual Business of the Loans and so on.

Many microfinance institutions (MFI), experts and donors are supporting to the idea of using mobile financial services (MFS) which they feel they are more convenient and efficient, and cost saving, than the traditional high to the touch model used for since long for the purpose of the delivering microfinance

services. Though, the MFI industry which is the primarily focusing on the low-income families and they 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 the $70 billion in outstanding loans and a global outreach of the 200 million clients.

We are working with the one such client that is in Telecom Industry. They are a fixed wireless which is telecommunications network provider. They have to launched various products and have developed its business and organization based on budget operator model, offering better to the products at Lower Prices to all value conscious customers through the strategy of disruptive innovation that focuses on the subscriber.

Mathematical / Analytical Modelling of the

Problem

Whenever we employ any ML algorithm, statistical models or feature

pre-processing in background lot of mathematical framework work. In

this project we have done lot of data pre-processing & ML model building.

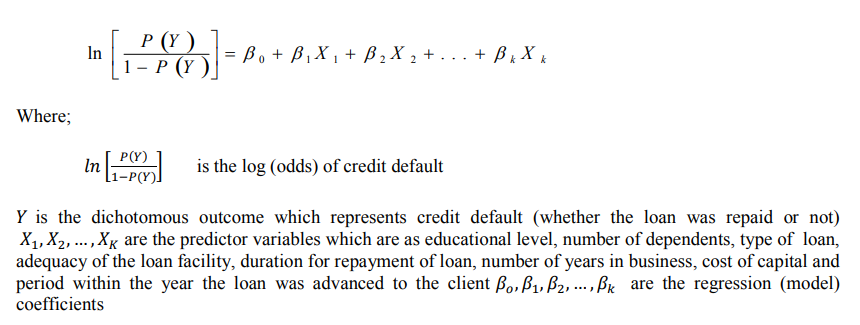
Logistic Regression

The response variable, label, is a binary variable. Therefore, the logistic regression which is the suitable technique to use that it is developed to predict a binary dependent variable as

a function of the predictor variables. The logit, in this model that is the

likelihood ratio that the dependent variable, non-defaulter, is one (1) as

opposed to zero (0), defaulter. The probability, P, of credit default is given

by; 

Decision Tree Classifier

Decision Trees (DTs) are a non-parametric (fixed number of parameters)

supervised learning method used for the classification and regression. The goal is to create the model that predicts the label of a target variable by

learning simple decision rules inferred from the data features.

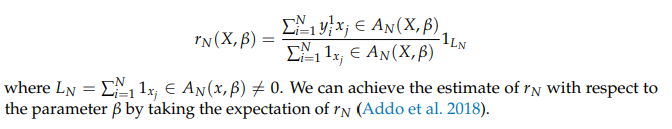
Random Forest Classifier

The random forest classifier is an ensemble method algorithm of the

decision trees wherein each tree depends on randomly selected samples trained for independently, with the similar distribution for all the trees in the

forest. Hence, the random forest is a classifier incorporating a collection of the tree-structured classifiers that decrease overfitting, resulting in an increase in overall accuracy (Geurts et al. 2006). As such, random forest’s accuracy differs based on the strength of each tree classifier and

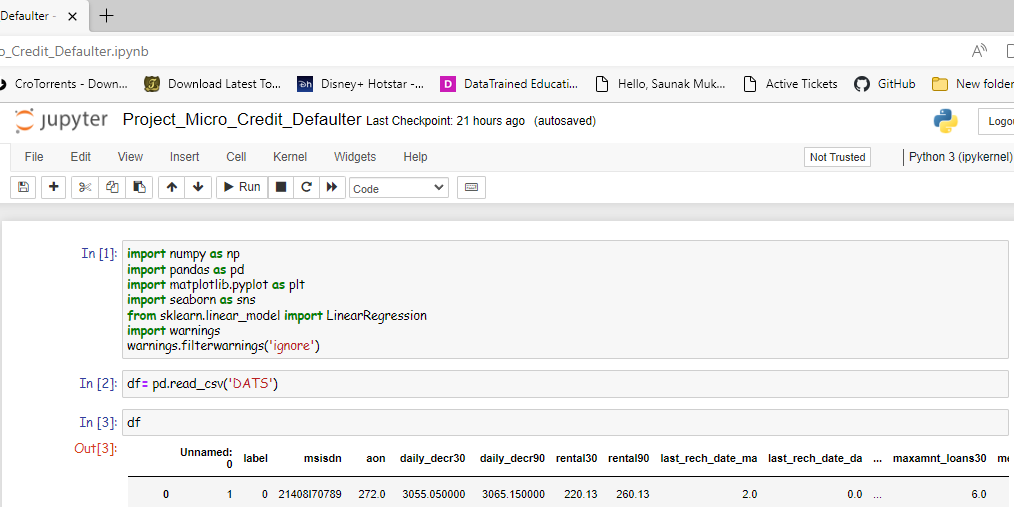
their dependencies



Data Sources and their formats

The data set comes from my internship company – Fliprobo technologies

in excel format



There are 37 columns and 209593 rows in this dataset. The different

features in dataset are as below:

➢label : Flag indicating whether the user paid back the credit amount within 5 days

of issuing the loan {1:success, 0:failure}

➢msisdn : mobile number of user

➢aon : age on cellular network in days

➢daily\_decr30 : Daily amount spent from main account, averaged over last 30 days

(in Indonesian Rupiah)

➢daily\_decr90 : Daily amount spent from main account, averaged over last 90 days

(in Indonesian Rupiah)

➢rental30 : Average main account balance over last 30 days

➢rental90 : Average main account balance over last 90 days

➢last\_rech\_date\_ma : Number of days till last recharge of main account

➢last\_rech\_date\_da: Number of days till last recharge of data account

➢last\_rech\_amt\_ma : Amount of last recharge of main account (in Indonesian Rupiah)

➢cnt\_ma\_rech30 : Number of times main account got recharged in last 30 days

➢fr\_ma\_rech30 : Frequency of main account recharged in last 30 days

➢sumamnt\_ma\_rech30 : Total amount of recharge in main account over last 30 days

(in Indonesian Rupiah)

➢medianamnt\_ma\_rech30 : Median of amount of recharges done in main account

over last 30 days at user level (in Indonesian Rupiah)➢medianmarechprebal30 : Median of main account balance just before recharge in

last 30 days at user level (in Indonesian Rupiah)

➢cnt\_ma\_rech90 : Number of times main account got recharged in last 90 days

➢fr\_ma\_rech90 : Frequency of main account recharged in last 90 days

➢sumamnt\_ma\_rech90: Total amount of recharge in main account over last 90 days

(in Indonesian Rupiah)

➢medianamnt\_ma\_rech90: Median of amount of recharges done in main account

over last 90 days at user level (in Indonesian Rupiah)

➢medianmarechprebal90: Median of main account balance just before recharge in

last 90 days at user level (in Indonesian Rupiah)

➢cnt\_da\_rech30 : Number of times data account got recharged in last 30 days

➢fr\_da\_rech30: Frequency of data account recharged in last 30 days

➢cnt\_da\_rech90 : Number of times data account got recharged in last 90 days

➢fr\_da\_rech90 : Frequency of data account recharged in last 90 days

➢cnt\_loans30 : Number of loans taken by user in last 30 days

➢amnt\_loans30 : Total amount of loans taken by user in last 30 days

➢maxamnt\_loans30 : maximum amount of loan taken by the user in last 30 days

➢medianamnt\_loans30 : Median of amounts of loan taken by the user in last 30 days

➢cnt\_loans90 : Number of loans taken by user in last 90 days

➢amnt\_loans90 : Total amount of loans taken by user in last 90 days

➢maxamnt\_loans90 : maximum amount of loan taken by the user in last 90 days

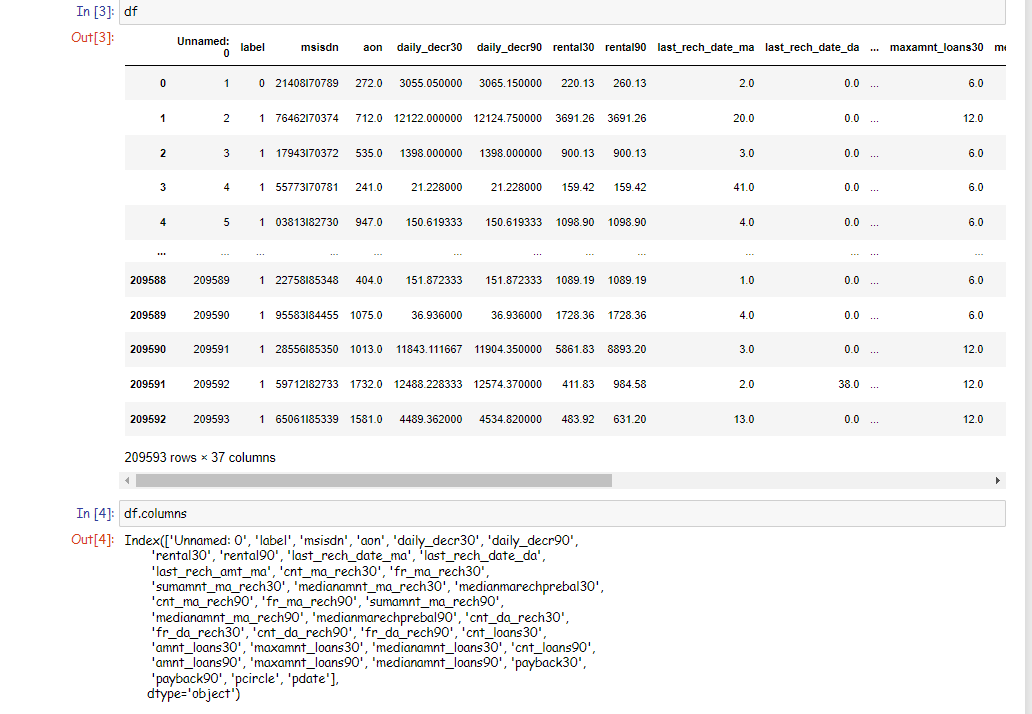
➢medianamnt\_loans90 : Median of amounts of loan taken by the user in last 90 days

➢payback30: Average payback time in days over last 30 days

➢payback90: Average payback time in days over last 90 days

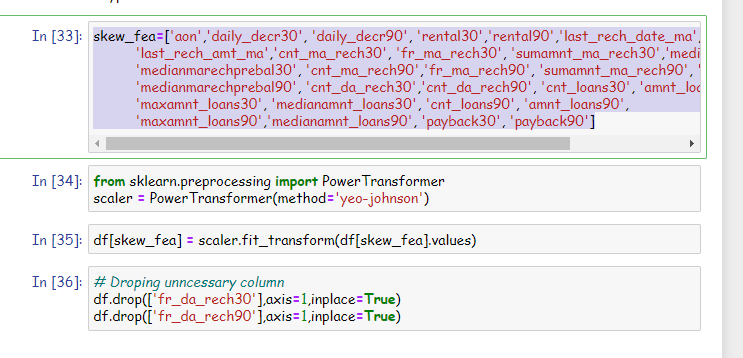
➢pcircle: telecom circle

➢pdate: date



Skewness in features & it’s transformation

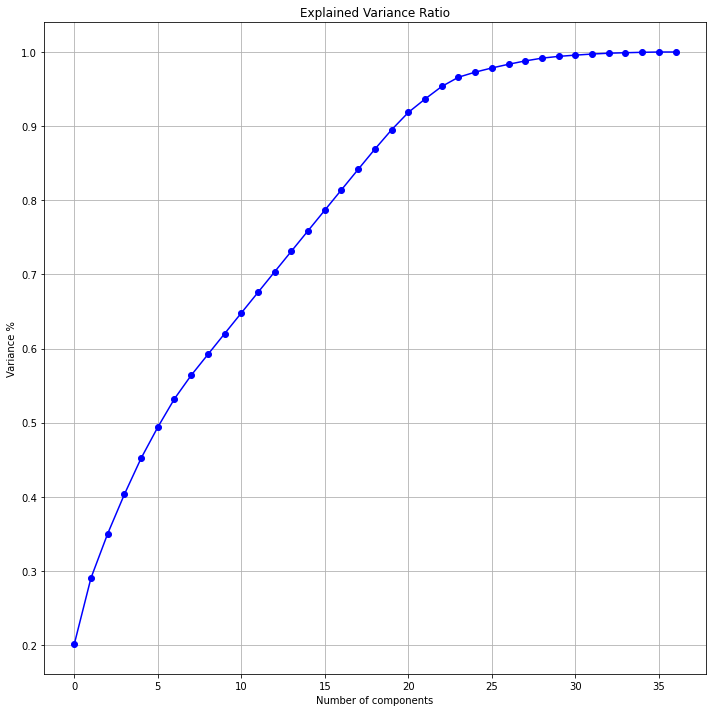
Considerable amount of the skewness found in most features by skew () function. Power transformer from sklearn.preprocessing library used to the transform skewness in features.



For most Independent feature VIF is the exceed permissible limit of 10. PCA

is applied to remove multicollinearity among features

we can see that 11 principal of the components attribute for 90% of variation in a data. We shall pick the first 11 components for our prediction.



Testing of Identified Approaches (Algorithms)

The different classification algorithm used in this project to build ML

model are as below:

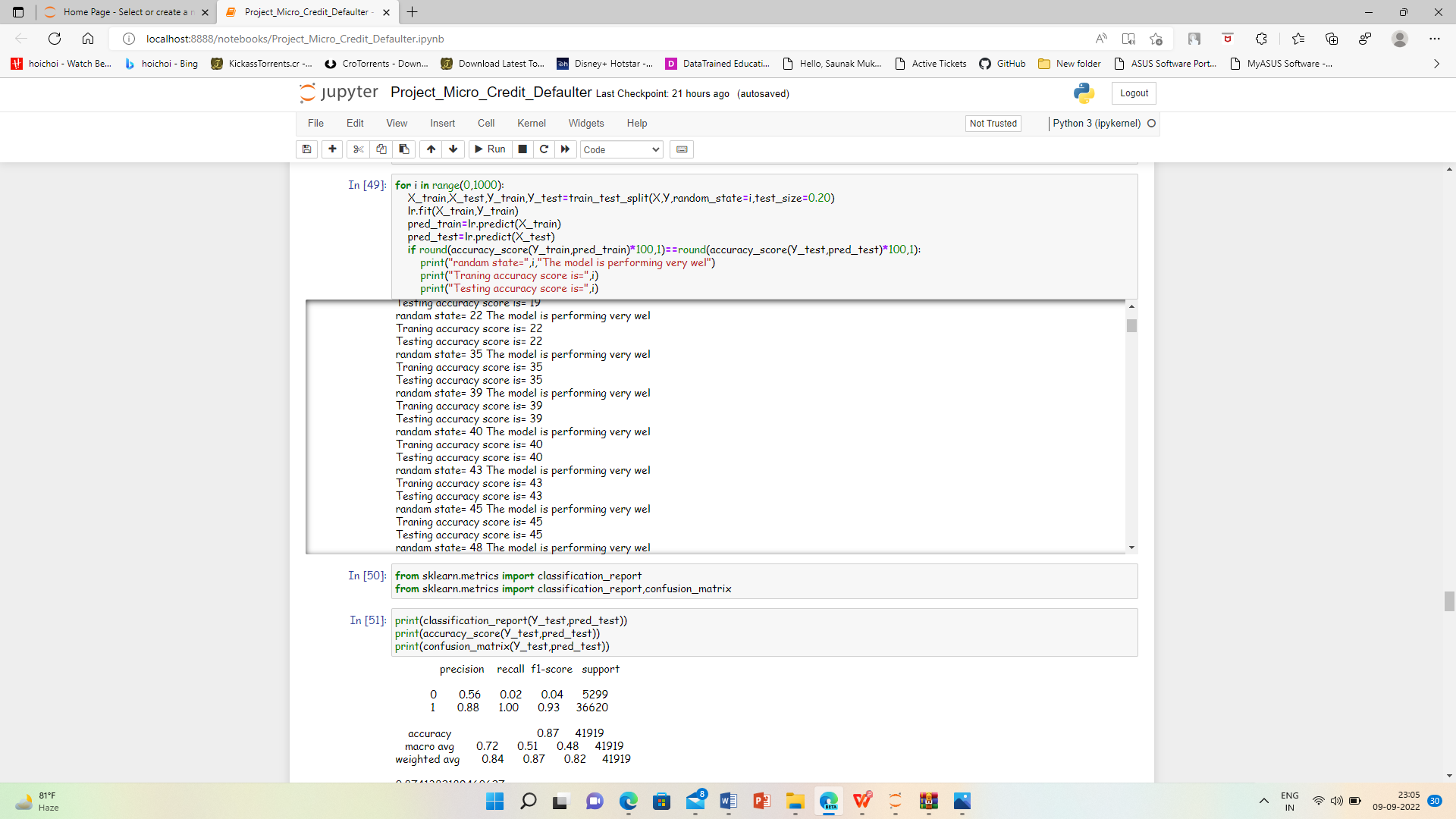
❖Logistics Regression

❖Random Forest Classifier

❖Extra Tree Classifier

RUN AND EVALUATE SELECTED MODELS

1. LOGISTICS REGRESSION



2. RANDOM FOREST CLASSIFIER

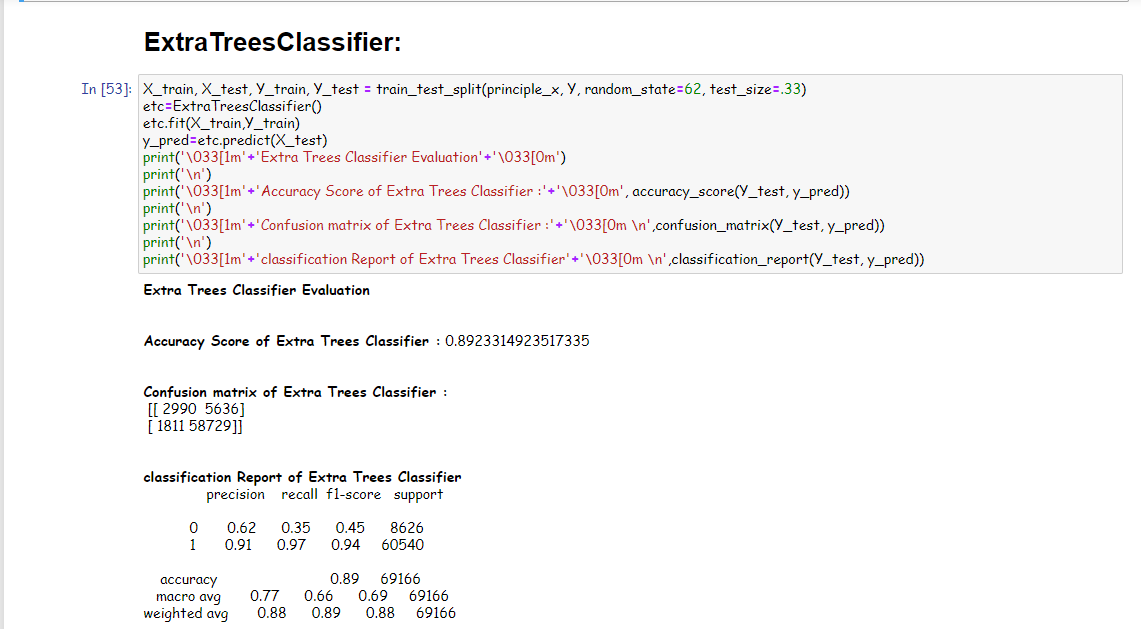
Model is train on Random Forest Classifier and evaluation matrix is as

follow:



3. EXTRA TREE CLASSIFIER

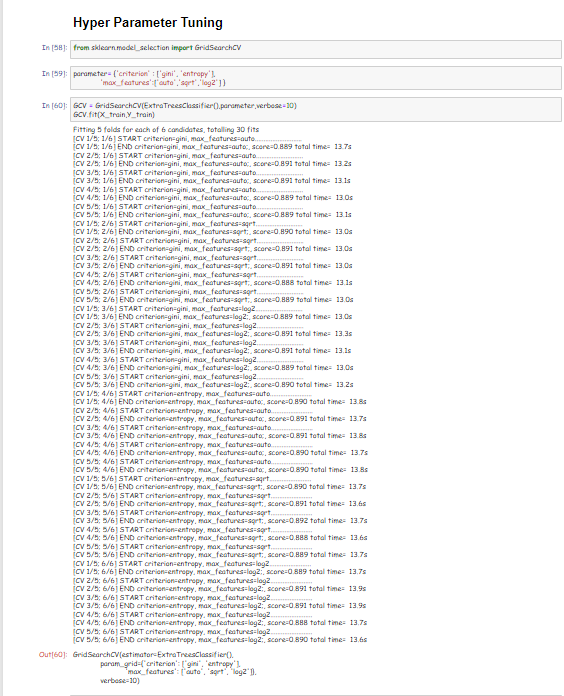
Model is train on Extra Tree Classifier and evaluation matrix is as follow:



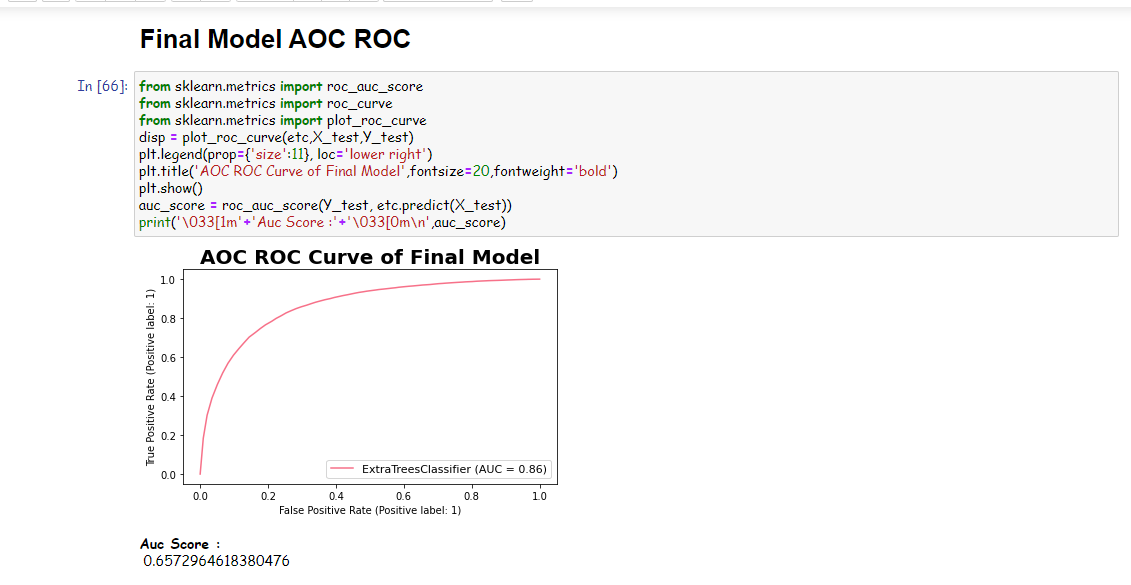
We can see Extra Tree Classifier gives maximum AUC. It also gives us

highest accuracy score and cross validation score. Hyper parameter

tuning perform on this model to enhance accuracy of model.

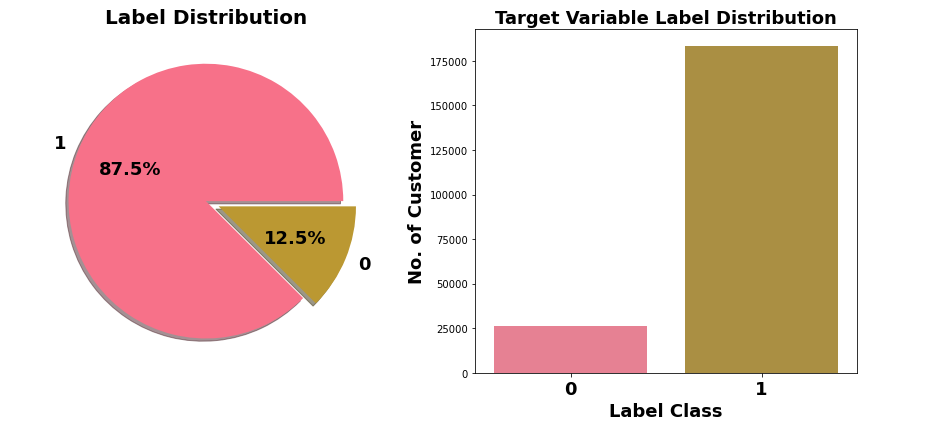


Final AUC ROC curve:



VISUALIZATIONS

Lets see target variable distribution before balancing data



Here target variable Label class 1 represent non-defaulter while Label

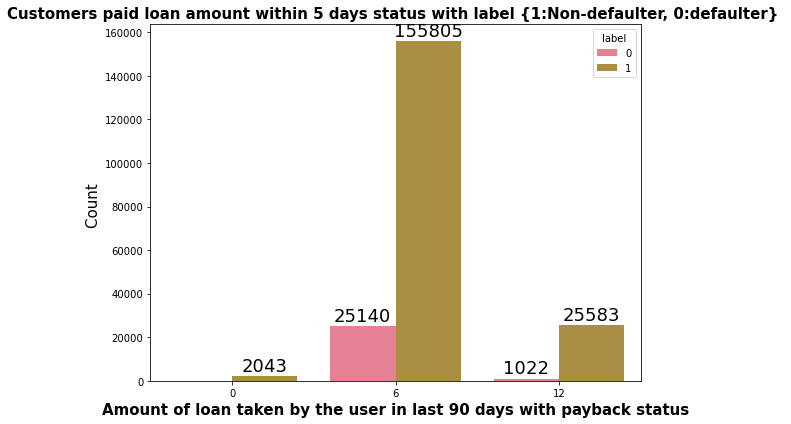
class 0 represent defaulter i.e., Loan not paid. We can see Most of

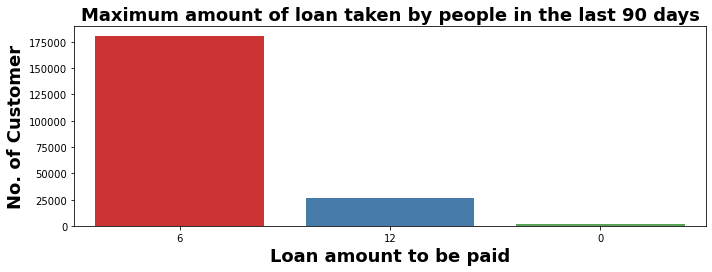
customers are non-defaulter while very few are defaulter. From ML

model building point of view target variable is imbalanced which need

to balance using balancing techniques.

Very few defaulters in case of customers who have taken loan in amount of 12.



n 30 days, maximum number of people had taken 6Rs as the loan amount and the number of people is 179192 whereas the number of the people had not taken loan and their number is 4291.In 90 days, there are maximum number of people had taken 6Rs as the loan amount and the number of people is 180944 whereas the number of people had not taken loan and their number is 2043.Maximum number of people had taken 12Rs as the loan amount within 90 days and their number is 26605 whereas for 30 days the number of people who had taken 12Rs is 26109 respectively.

Final model is saved using joblib library

