

Micro-Credit Defaulter Prediction

Project

Submitted by:

POOJA MISHRA

ACKNOWLEDGMENT

I would like to acknowledge the contributions of the following people without whose help and guidance this project would not have been completed. I acknowledge the counsel and support of the training institute ‘DataTrained’ Trainers including Dr. Deepika Sharma whose expertise, guidance, support, encouragement, and enthusiasm has made this project possible.

I would also like to thank Flip Robo Technologies who has given me such wonderful opportunity to learn more about Machine Learning and their mentors Shubham Yadav, Sajid Chaudhary for their constant encouragement, valuable suggestions, and cooperation.

This acknowledgement will remain incomplete if I fail to express my deep sense of obligation to my family members and God for their consistent blessings and believe in me.

Thank You.

**INTRODUCTION**

* Business Problem Framing

This Project based on a classic business problem which helps Micro Financing Institutions and other Lending companies reduce Credit risks by recognizing potential Defaulters. This kind of problems are facing many industries which results in financial loss.

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.

* Conceptual Background of the Domain Problem

In this project, we are working into 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.

Machine Learning can help lenders predict potential defaulters before approving their candidature using their past data. The candidates’ income, past debt and repayment behaviour can be important metrics for the same.

The domain that has been illustrated here is on Telecom industry. From where this credit analysis is based on responsible for assessing this risk by thoroughly analysing a borrower’s capability to repay a loan.

* Review of Literature

Machine learning algorithms have a lot to offer to the world of credit risk assessment due to their unparalleled predictive power and speed. In this blog, we will be utilizing machine learning’s power to predict whether a borrower will default on a loan or not and to predict their probability of default.

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).

Here the defaulter is considered as ‘0’ and non-defaulter as ‘1’. With help of this dataset we can predict customers who are most likely to default or not.

* Motivation for the Problem Undertaken

The motivation came for this problem solver is the keen to learn machine learning. This project is real time and very common in telecom industry. Basically, it’s a minor loan but when it comes to large count then it affects the growth of industry that’s why identification of such type of problem is very important. Machine learning is the solution of day to day problem solutions and prediction helps to avoid the foreseen risks.

In this project as well, we are showing solution to very crucial problem which are facing micro financial industries. With help of machine learning we are building 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

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

There are several techniques used in machine learning which contributes in problem solution. Machine Learning problem solving is highly related to statistical, mathematics, programming and analytics leaning.

For analysing this project, the mathematical modelling done includes the understanding of types of values in complete dataset. Based on the problem, understanding the numerical values and non-numerical values, and making them ready for modelling. Analysing each value whether its numeric, non-numeric, symbols, null, object type etc. converting them in numeric model to perform algorithms.

Statistical modelling includes the understanding of statistical parameter analysis which gives more interpretation of dataset. It also predicts what kind of techniques need to perform. More understanding about the range, mean, median, min, max value which helps a lot to understand which techniques need to be perform before implementing algorithm testing.

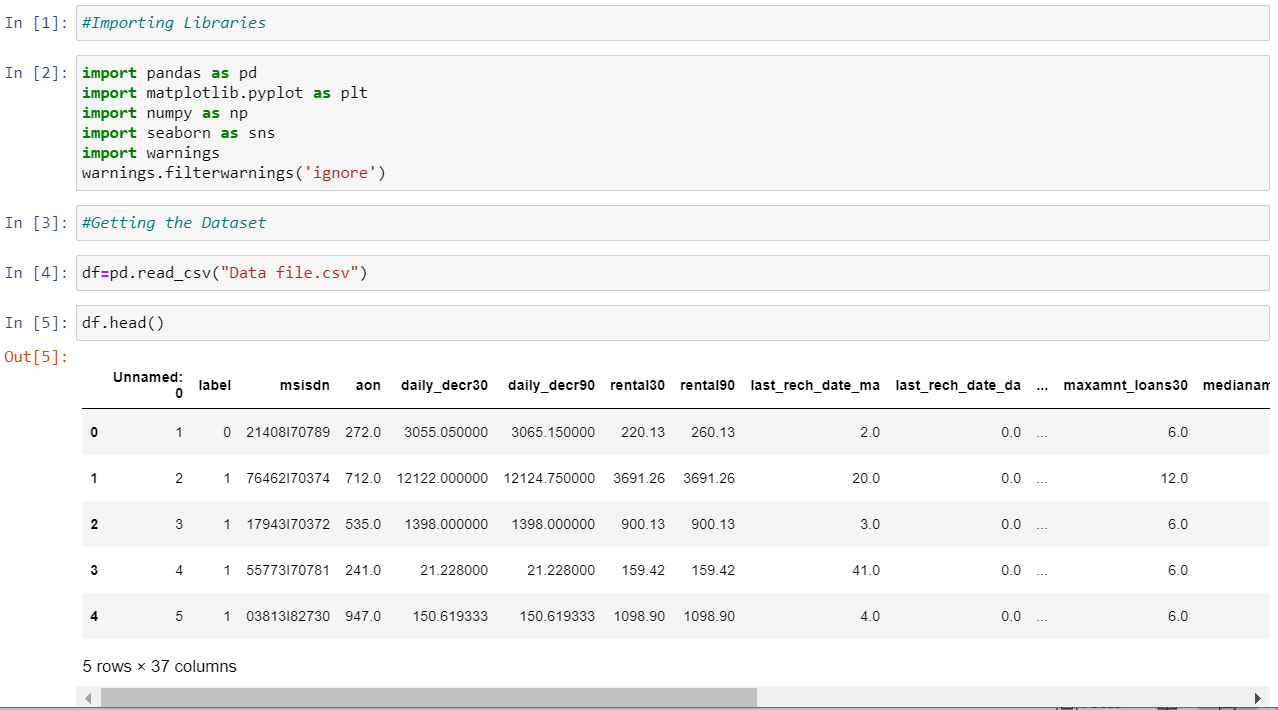
Machine learning is incomplete without analytic understanding as it builds the basic building block to solve every problem.

* Data Sources and their formats

This dataset has been taken from Telecom Industry. It comprises 3 columns and 209593 rows. The industry has 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.

In this Dataset, there are no null values present thus no need to perform null values removal techniques. Here most of the columns are present in float and integer type, only 3 features are object type. Utilizing all features, we’ll able to predict the investment and improvement in selection of customers to whom loan should be given or not given.

Loading dataset into Variables:



* Data Pre-processing Done

Data pre-processing is very important to get the dataset into the best format before performing algorithm. This is very important step in Machine Learning that should not be skipped. It involves three stages: Data Cleaning, Data Transformation, and Feature Engineering which converts complicated dataset into quality data.

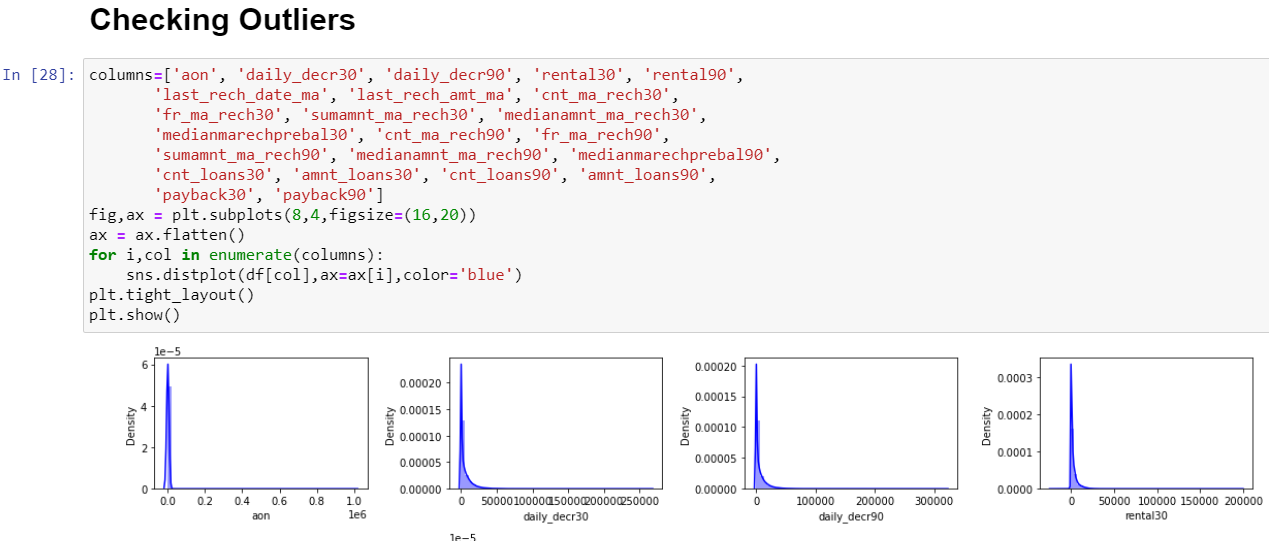
**Data Cleaning:**

In Data cleaning, we understand the data, we perform several activities to clean the data. This stage comprises of following activities:

**-** Dealing with Null Values: In this Dataset, there are no null values present.

**-** Outliers: An outlier is a data point in a dataset that is distant from all other observations i.e. a data point that lies outside the overall distribution of the data set. Depending on the frequency of outliers, we’ll perform activities to remove it.

In this Dataset, we have checked the outliers by plotting boxplot and removed outlier’s column by column by using IQR score technique.



**Data Transformation:**

In Data transformation, we check whether the columns are present in numeric or not?

If not, then we need to perform Label Encoding or One-hot encoding to convert the object type

into numeric.

Here in this dataset, we observed that the columns we need to take ahead for modelling are in numeric. As 3 columns are not in numeric that we’ll remove later as its not contributing any important role in algorithm performance.

Feature Engineering:

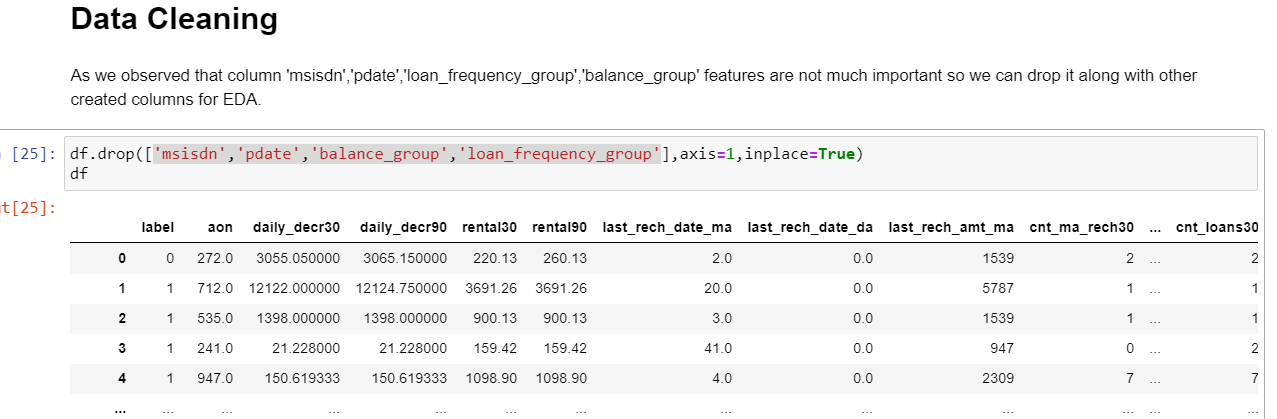
In feature Engineering, is the process of creating new features based upon knowledge about current features and the required task. It involves 2 activities:

-Feature Extraction

-Capturing Feature Relationships

We check which columns are useful in prediction and which are not. The columns which are not contributing in analysis we can drop it.

Here in this dataset, the columns which are not contributing in analysis are 'msisdn','pdate','balance\_group','loan\_frequency\_group',’pcircle’ thus we’ll drop it.



Feature Rescaling:

Min-max scaler is the standard approach for scaling. For normally distributed features standard scaler could be used, which scales values around a mean of 0 and a standard deviation of 1.

Here in this dataset we found values of numerical columns are not standard so for simplicity we use min-max scaler for all features. Resulting we get our dataset ready with standardised data to perform all algorithms.

* Data Inputs- Logic- Output Relationships

Understanding the relationship behind data inputs and its relationship with target feature will help to perform modelling.

We can understand the relationship of target feature with other features by checking correlation and plotting different types of by analysing the data visualization.

**Correlation Matrix:**

Correlation analysis shows the correlation of target feature i.e. ‘label’ with all remaining

features. The columns which are taken in consideration to perform modelling are as follows:

['aon', 'daily\_decr30', 'daily\_decr90', 'rental30', 'rental90',

'last\_rech\_date\_ma', 'last\_rech\_amt\_ma', 'cnt\_ma\_rech30',

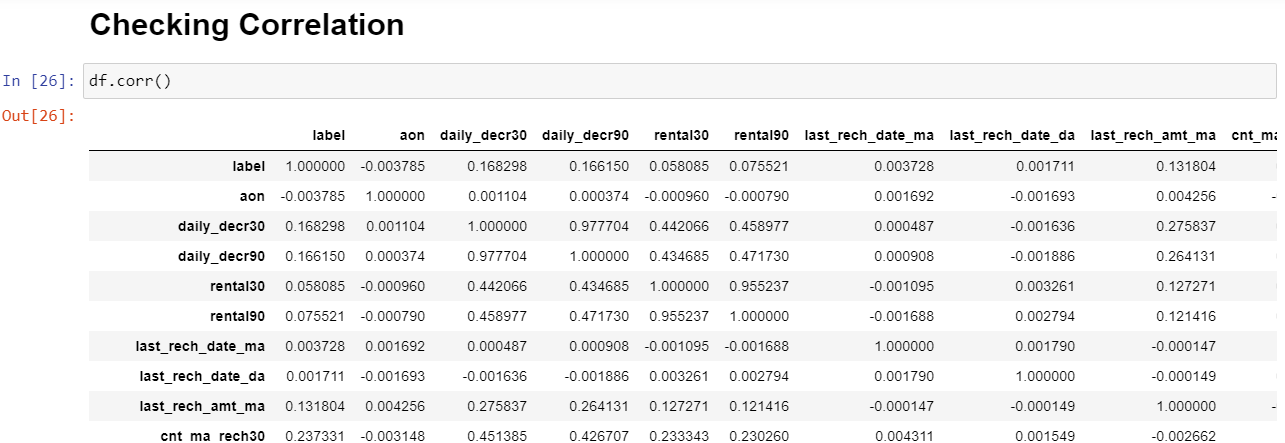
'fr\_ma\_rech30', 'sumamnt\_ma\_rech30', 'medianamnt\_ma\_rech30',

'medianmarechprebal30', 'cnt\_ma\_rech90', 'fr\_ma\_rech90',

'sumamnt\_ma\_rech90', 'medianamnt\_ma\_rech90', 'medianmarechprebal90',

'cnt\_loans30', 'amnt\_loans30', 'cnt\_loans90', 'amnt\_loans90',

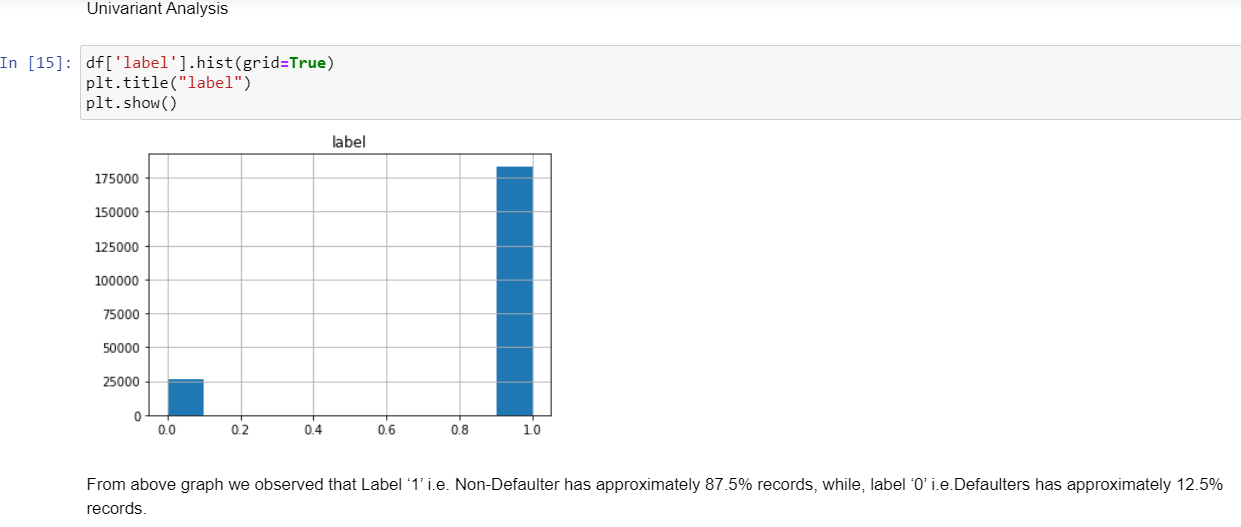
'payback30', 'payback90']



From above correlation analysis we observed that few of the features are positively correlated, few are negatively correlated, and few features are not showing any correlation with target variable.

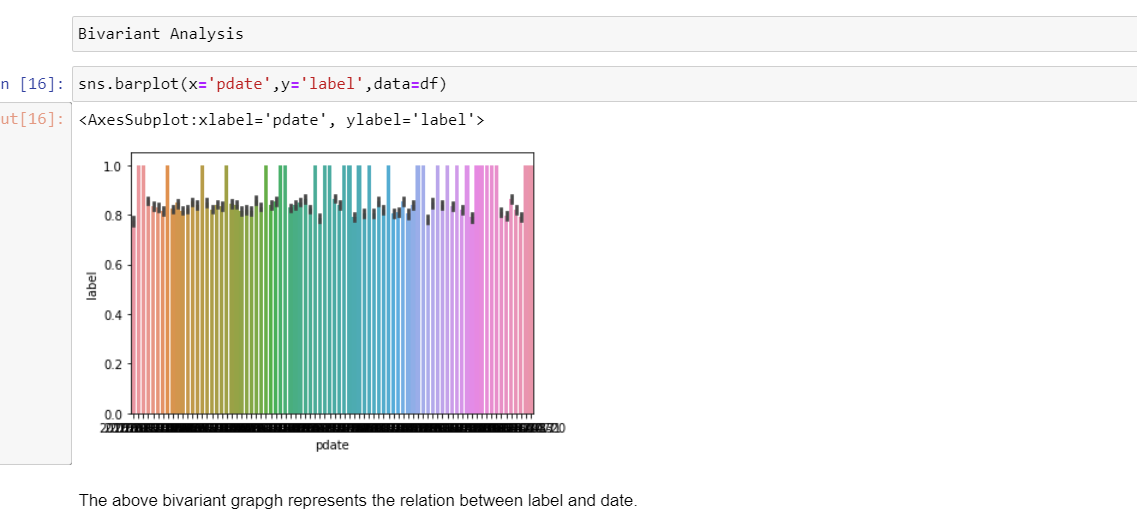
**Data Visualization:**

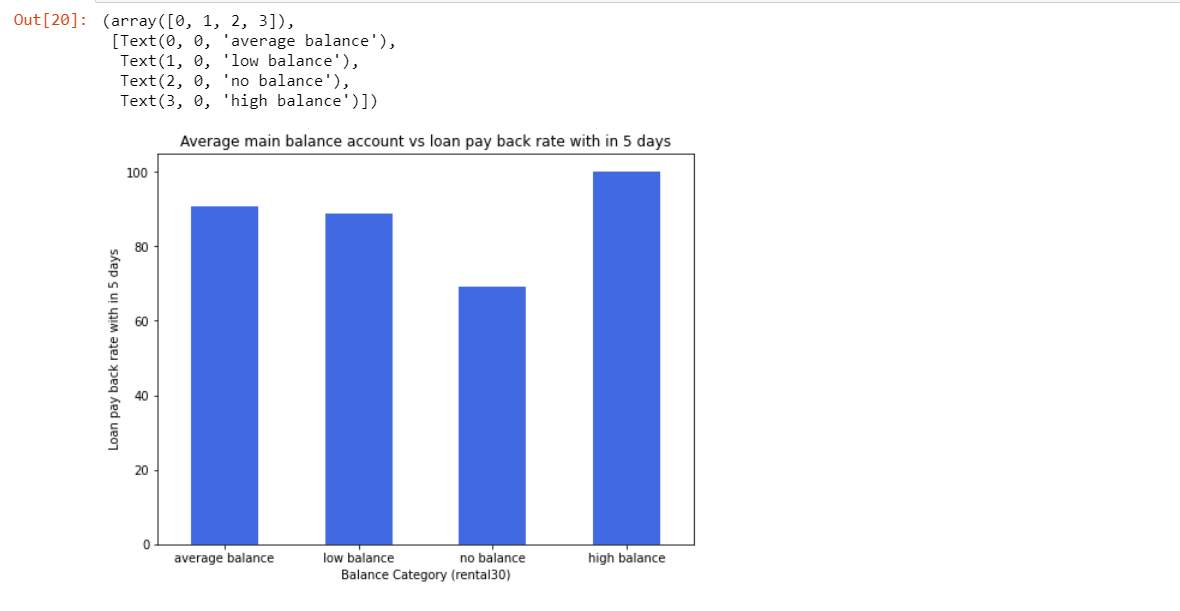
In this portion we can plot different graph using different columns and try to visualize the data using matplotlib and seaborn library.

We have plotted hist plot, bar graph, count plot to visualise the relationship between all features. 

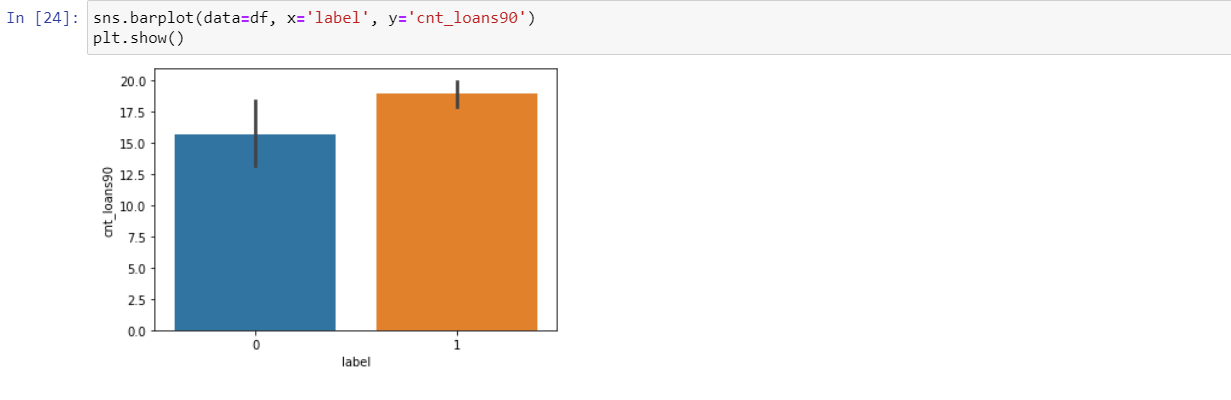
Above graph shows that relationship with target feature showing the approximate % for defaulters and non-defaulters.

Bivariant Analysis





With above plot graph we observed that the high balance level people are 100% who are loan within 5 days. The people with average balance and low balance are around 90% who are paying loan within 5 days and 10-12% who are paying the loan within 5 days. The people with no balance are around 30% who are not able to pay loan.



From above graph we observed that number of loans taken by users in last 90 days and who are defaulters are around 15times while number of loans taken by users in last 90 days and who are not defaulters are around 20 times.

* Hardware and Software Requirements and Tools Used

Machine learning comes with an extensive collection of ML tools, platforms, and software.

Well, in this problem solution we have used following tools that helps to make this project successful as per my possibility.

Scikit-Learn

Scikit-Learn is built on top of the three main Python libraries viz. NumPy, Matplotlib, and SciPy. Along with this, it will also help you with testing as well as training your models.

The Libraries are as listed:

1. For Data loading and Visualisation:

**Numpy** - NumPy is very useful for handling linear algebra, Fourier transforms, and random numbers.

**Pandas** - Pandas are turning up to be the most popular Python library that is used for data analysis with support for fast, flexible, and expressive data structures designed to work on both “relational” or “labeled” data.

**Matplotlib** - The library helps to generate histograms, plots, error charts, scatter plots, bar charts with just a few lines of code.

**Seaborn** – Used for visualization.

1. For Normalization:

**Min-max scaler** is the standard approach for scaling. We use this library to balance the dataset and make it normal for further algorithm performance.

1. **train\_test\_split**: We use it to perform test train spliting.
2. Algorithm Libraries:

In this problem solution we have used algorithm libraries:

* LogisticRegression – for logistic regression algorithm

***from sklearn.linear\_model import LogisticRegression***

* RandomForestClassifier – For Random Forest Classifier algorithm

***from sklearn.ensemble import RandomForestClassifier***

* DecisionTreeClassifier – For Decision Tree Classifier algorithm

***from sklearn.tree import DecisionTreeClassifier***

* SVC – to perform Linear Support Vector Machine algorithm***.***

***from sklearn.svm import SVC***

* GradientBoostClassifier – To perform gradient Boost classifier algorithm

***from sklearn.ensemble import GradientBoostingClassifier***

* classification\_report confusion\_matrix– for getting all classification reports: Precision, Recall, F1-score, Support values and confusion matrix values.

***from sklearn.metrics import classification\_report, confusion\_matrix***

* cross\_validation\_score – To get cross validation score for each used algorithm.

***from sklearn.model\_selection import cross\_val\_score***

* roc\_auc\_score, roc\_curve - To get ROC – AUC curve for all algorithms.

***from sklearn.metrics import roc\_curve,roc\_auc\_score***

* RandomizedSearchCV – For hyper parameter tuning performance.

***from sklearn.model\_selection import RandomizedSearchCV***

Jupyter Notebook

Jupyter notebook is one of the most widely used machine learning tools among all. It is a very fast processing as well as an efficient platform. Moreover, for this problem solution I have used python programming.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

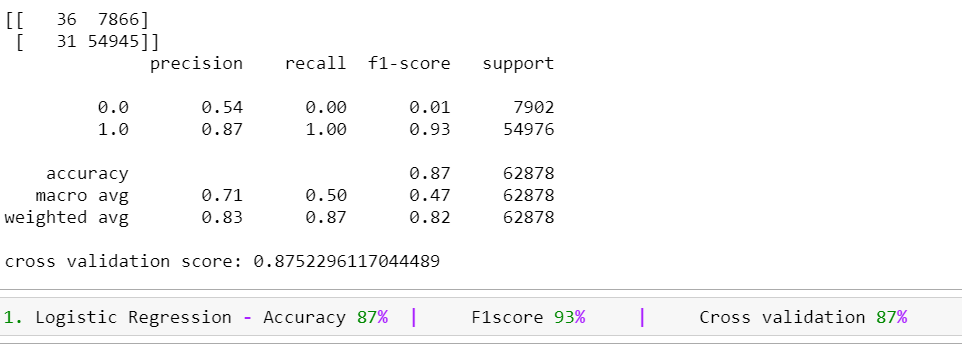
Building classifier Models Since it was a supervised classification problem (target output is classification type i.e. 0 and 1), I built 5 models to evaluate performance of each of them: a. Logistic Regression b. Random forest c. Decision Tree d. Linear SVM e. Gradient Boost Classifier. To get best accuracy I have performed all correct performance/ elevation metric like accuracy, precision, recall and ROC-AUC curve.

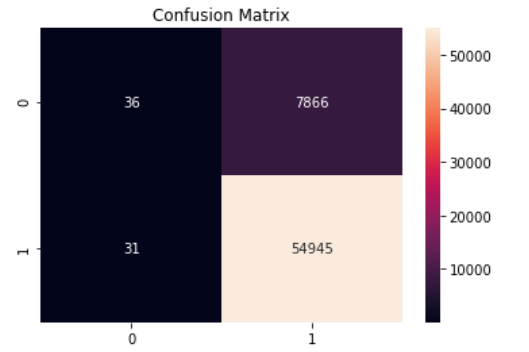
* Testing of Identified Approaches (Algorithms) and
* Run and Evaluate selected models

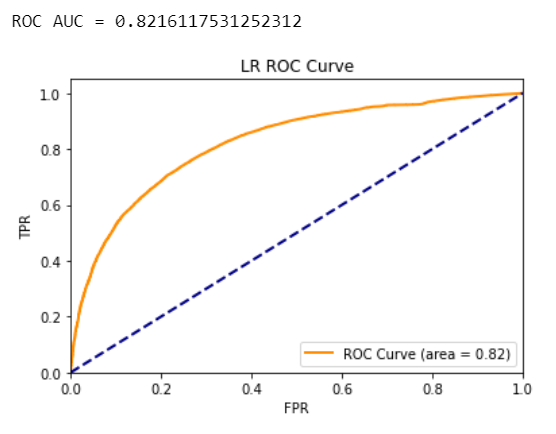
# **Analysis of output of each model**

**1.Logistic Regression:**

With Logistic Regression, I got accuracy of 87% with F1score 93% and cross validation 97%.



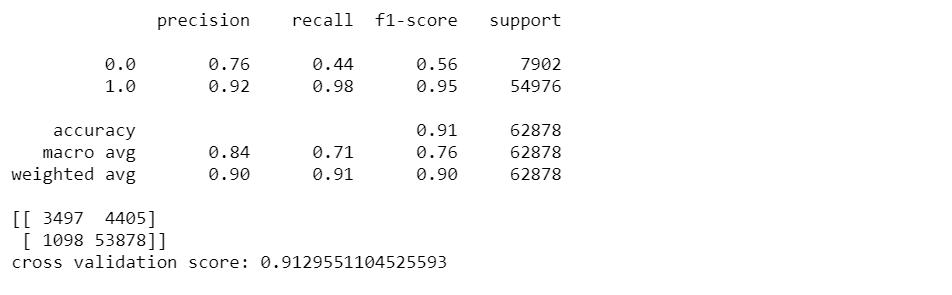


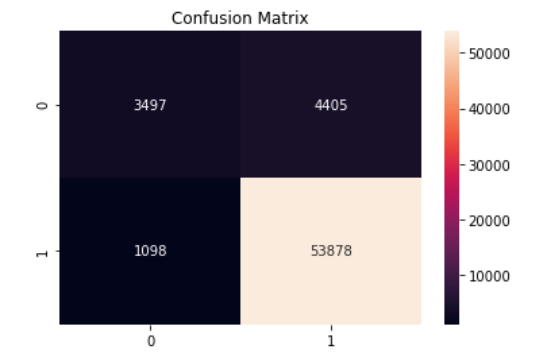


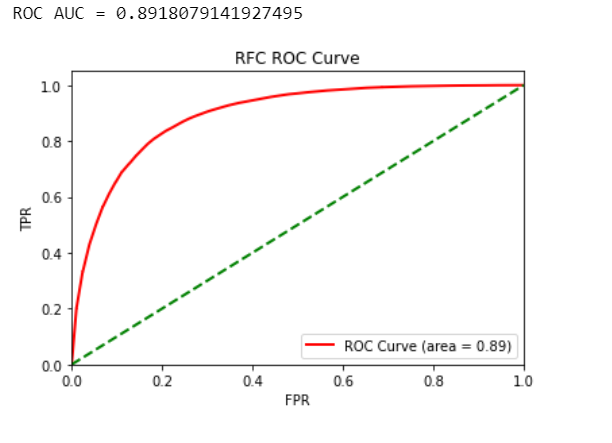
**2) Random Forest Classifier**

With Random Forest Classifier I am getting accuracy of 91% and F1 score is 95%

Cross Validation Report is 91%





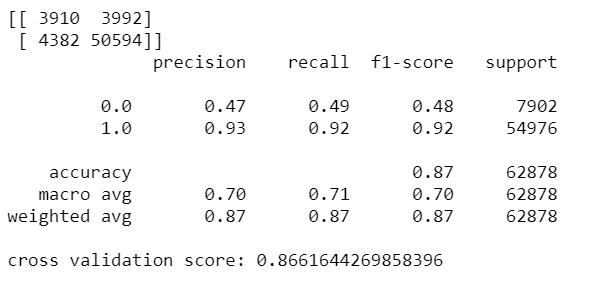


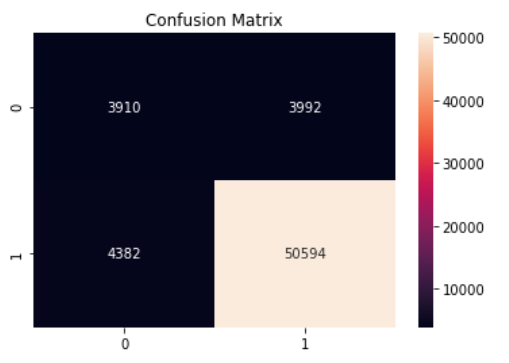


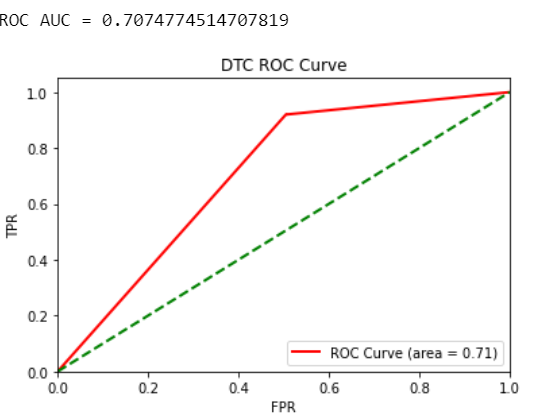
**3) Decision Tree Classifier**

With Decision Tree Classifier I am getting accuracy of 87% and F1 score is 92%

Cross Validation Report is 86%





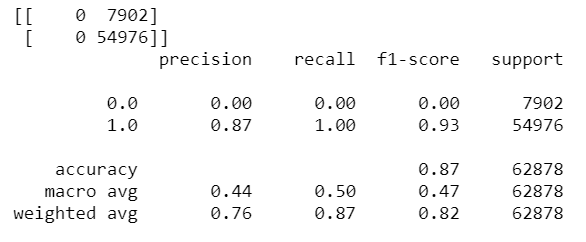


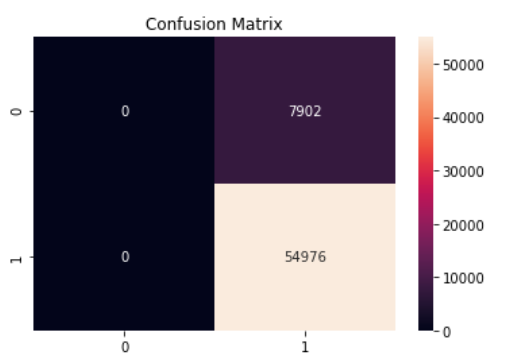


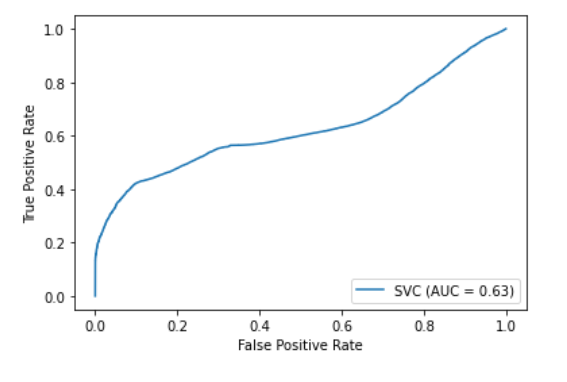
**4) Linear Support Vector Machine**

With SVC Classifier I am getting accuracy of 87% and F1 score is 93%

Cross Validation Report is 87%





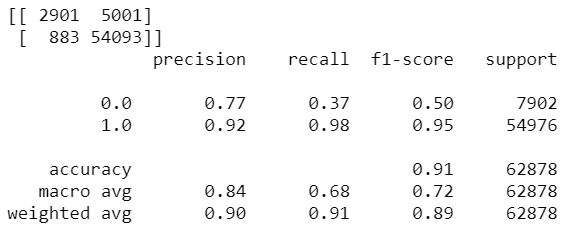


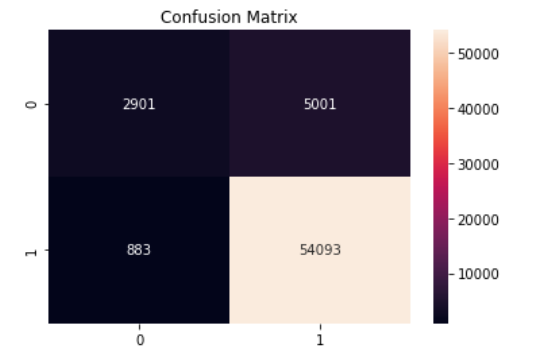


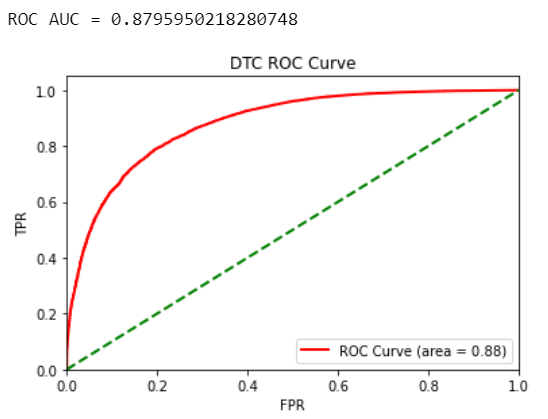
**5) Gradient Boost Classifier**

With SVC Classifier I am getting accuracy of 91% and F1 score is 95%

Cross Validation Report is 90%









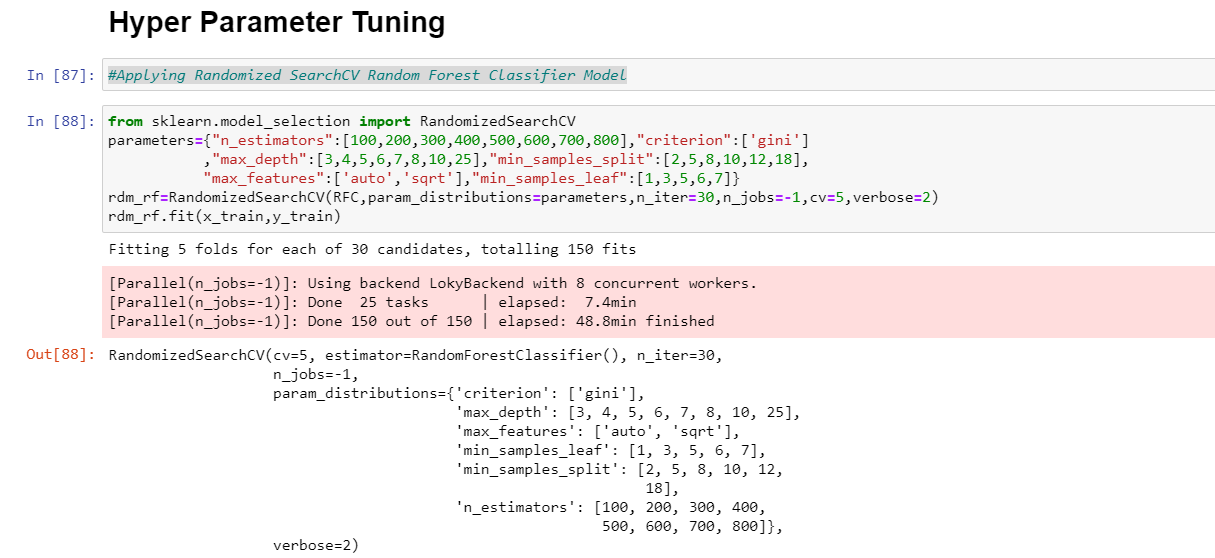
* Interpretation of the Results

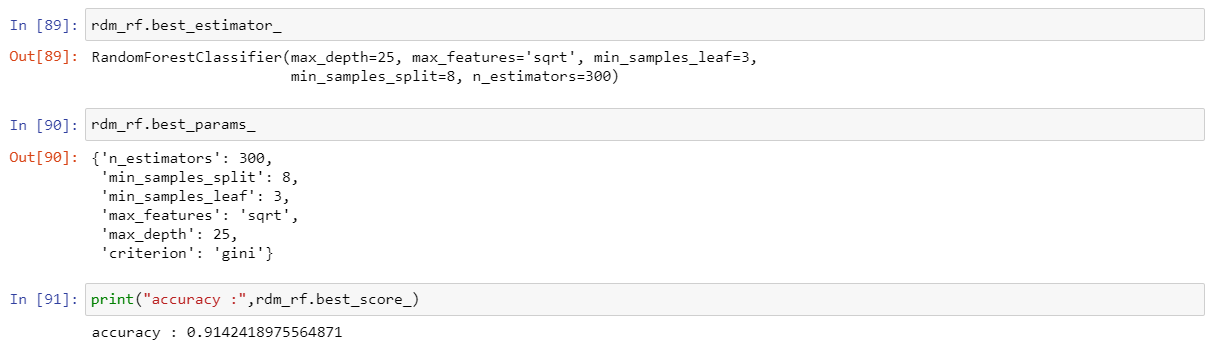
After performing 5 modelling algorithms as explained and elaborated above with all key metrics comes with following observation:

Considering all metrics and accuracy performance we observed that the best accuracy we are getting with Random Forest and Gradient Boost Classifier. Least accuracy we observed with Logistic Regression and Linear SVM.

**Hyper Parameter Tuning**

We’ll be performing Randomized Search to narrow down the possibilites to re-check the accuracy of this model so that we can conclude that Random Forest is the best fit model for this problem. This approach is more suited to get the best model slection confirmation.





**CONCLUSION**

* Key Findings and Conclusions of the Study

After performing 5 algorithms I can conclude that Random Forest is giving the best results in comparison to other algorithms. Considering all evaluation and performance metrics, Random Forest is getting the highest accuracy more than Gradient Boost Classifier as well.

We came to confirmation that Random Forest is best fit model by performing hyper parameter tuning on it which gives same accuracy of 91-92%.

* Learning Outcomes of the Study in respect of Data Science

If I talk of learning then definitely I can say that with problem I came up with more understaning of Machine Learning concepts. Regarding Visualization I learnt new techniue by splitting the results and plotting it on graph which results in more justification.

## The result of this problem is random forest classifier as it meets the best accuracy and all

## performce metrics. As well the curve score is also maximum for it and tending towards the

## ideal shape than other models so I can say that it is the best fit model to perform this

## telecom defaulter problem.

* Limitations of this work and Scope for Future Work

## I tried to find out best accuracy score for this problem i.e. 91% for Random Forest Classifier.

## Definitely, I’ll try to upskill myself to learn more and practice well to give better

## performance going ahead.

## For this problem, In order to decrease loss to the telecom company in case of defaulters, the

## company should take some marketing strategies like sms alerting and notifications and

## others on the people with all loan levels and especially on low & high level people notifying

## them to pay the loan back within five days of time.

## 

Thank You.