

A Project on

## Micro Credit Defaulter



Submitted by

**KUNIGIRI NAGARAJU**

# ACKNOWLEDGMENT

- Firstly I would like to thank FlipRobo for giving this opportunity. Being a fresher it was a wonder opportunity to work on such a project, where it was so much to explore.
- Secondly I would also like to thank my Mentor Gulshana, and Khushboo. She provided us with all the updates and responded quickly whenever the ticket was raised.
- I refer to many articles that help me initially with the project building.
  1. Articles about loan defaulter on [www.analyticsvidya.com](http://www.analyticsvidya.com) were of great help.
  2. I refer to [www.towardsdatascience.com](http://www.towardsdatascience.com) for the approach one should have towards such project.
  3. Whenever I was stuck with coding I refer to [www.stackoverflow.com](http://www.stackoverflow.com)

# INTRODUCTION

- **Business Problem Framing**

- 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.
- 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 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.

- **Conceptual Background of the Domain Problem**

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.

- **Review of Literature**

We have to take the logistic approach to the problem statement whether the customer will pay back the loan in 5 days(Yes or No).

Understand the customer behaviour from the dataset we have.

Build the classification machine learning model to for predicting the outcome.

Draw conclusion from the study.

- **Motivation for the Problem Undertaken**

The objective behind to take this project is to harness the required data science skills.

Improve the analytical thinking.

Get into the real world problem solving mechanics

## **Analytical Problem Framing**

- **Mathematical/ Analytical Modeling of the Problem**

We have 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 paid i.e. Non- defaulter, while, Label '0' indicates that the loan has not been paid .for a loan amount of 5 payback amount should be 6,and for loan amount of 10 payback amount is 12

- Data Sources and their formats

CSV file was used to read the data in python.

Variable	Definition
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

Below data sources was found useful to study the past pattern of customers

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
cnt_ma_rech30	Number of times main account got recharged in last 30 days
cnt_ma_rech90	Number of times main account got recharged in last 90 days
amnt_loans30	Total amount of loans taken by user in last 30 days
amnt_loans90	Total amount of loans taken by user in last 90 days

- **Data Preprocessing Done**

Below were the steps taken in data pre-processing.

- 1. Null values**

We checked for Null values but there were no Null values present in the dataset.

- 2. Negative Values**

There were 9 columns which show negative value. This was unrealistic as average recharge; number of days for last recharge cannot be negative

Below are the columns with negative value :

('aon','daily\_decr30','daily\_decr90','rental30','rental90','last\_rech\_date\_ma',  
'last\_rech\_date\_da','medianmarechprebal30','medianmarechprebal90')

The account balance can be negative hence having negative value in 'rental30' and 'rental90' is a possibility.

The negative values from the data sets were removed. We lost around 3% of the data while removing it.

- 3. Dropping of columns**

Below columns were dropped based on assumption given

'pdate' - This data does not show what date it is referring to and we already have 'aon' to give us the age on cellular network for necessary study hence it was dropped.

'pcircle' - Has only 1 value and will not provide any usefulness for model fitting. Hence the same was dropped.

'msisdn' - mobile number of user is irrelevant for our study to determine the defaulter hence the same was dropped.

'fr\_da\_rech30','maxamnt\_loans30' - This columns were dropped as it has no impact with the target variable as per the correlation study.

- 4. PCA (Principal component Analysis)**

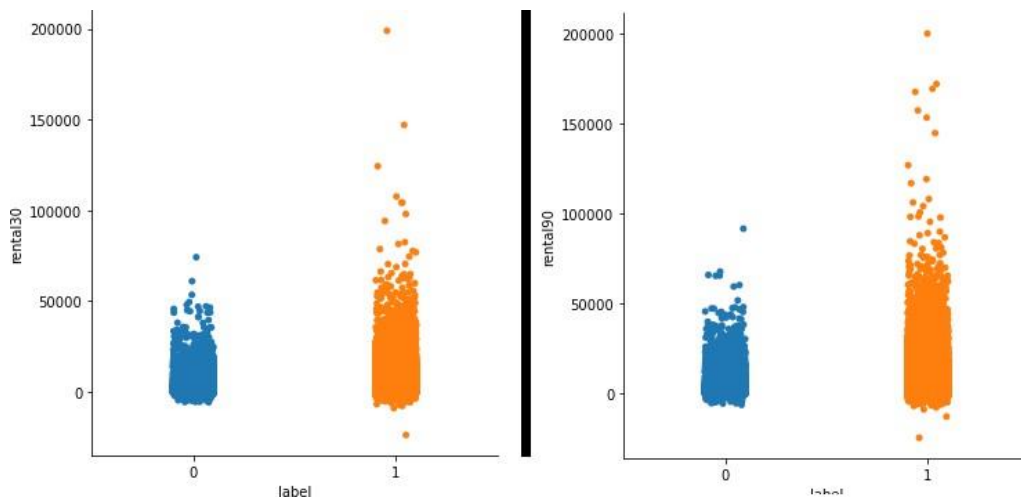
PCA was performed on the input variable as most of the features in the dataset were correlated to each other. This helped us to avoid multi colinearity issue with our model.

- **Data Inputs- Logic- Output Relationships**

To understand pattern of customer behaviour below aspect from the past data was looked into:

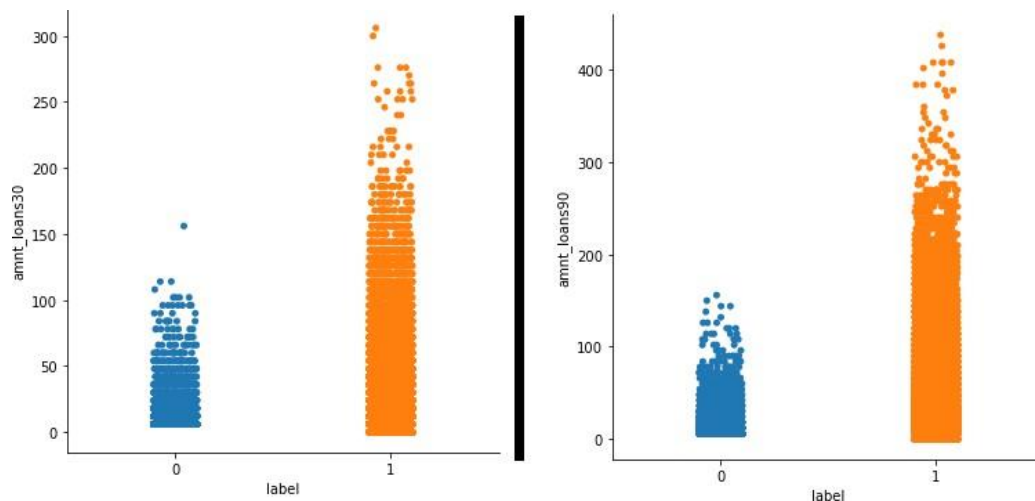
1. **Average main account balance over last 30 days and 90days against the target study**

The below data shows that customer who maintains average balance of over 50000 in last 30 days and 90 days is less likely to default the loan



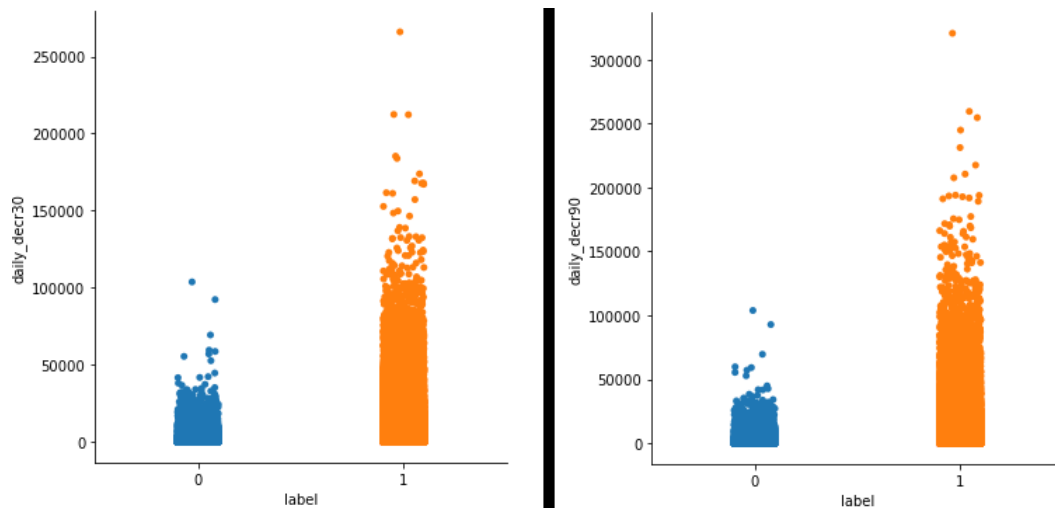
2. **Total amount of loans taken by user in last 30 days and 90 days**

From below data study we can conclude that customer who takes average of 50 loans and 90 loans in last 30 and 90 days respectively are less likely to default the loan.



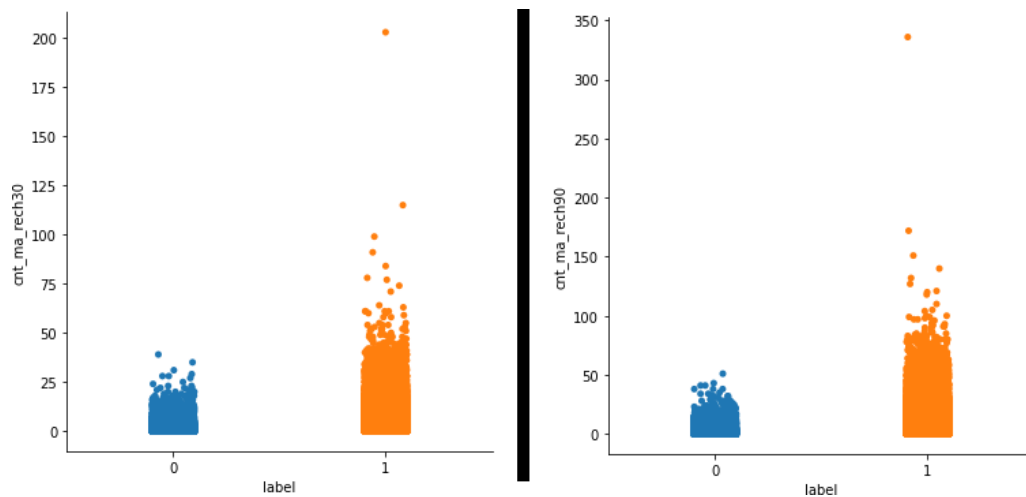
### 3. Daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah).

The below data clearly shows that if customer is spending daily average of over 40000 in last 30 and 90 then they are more likely to pay the loan.



### 4. Number of times main account got recharged in last 30 and 90 days

From the below data we can conclude that if customer is recharging its main account 20 times in the last 30 or 90 days is less likely to default the loan.






## 5. State the set of assumptions (if any) related to the problem under consideration

Some of the features had negative values which, as per the given description, does not hold true. Hence the same was removed.

Below are the columns with negative value :

('aon','daily\_decr30','daily\_decr90','rental30','rental90','last\_rech\_date\_ma',  
'last\_rech\_date\_da','medianmarechprebal30','medianmarechprebal90')

## 6. Hardware and Software Requirements and Tools Used

Rating:	 <a href="#">Windows Experience Index</a>
Processor:	AMD A4-4020 APU with Radeon(tm) HD Graphics 3.20 GHz
Installed memory (RAM):	4.00 GB (3.19 GB usable)
System type:	64-bit Operating System

Python 3

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')

from scipy.stats import zscore

from sklearn.preprocessing import power_transform

from sklearn.preprocessing import StandardScaler

from sklearn.decomposition import PCA

from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.naive_bayes import GaussianNB

from sklearn.model_selection import cross_val_score

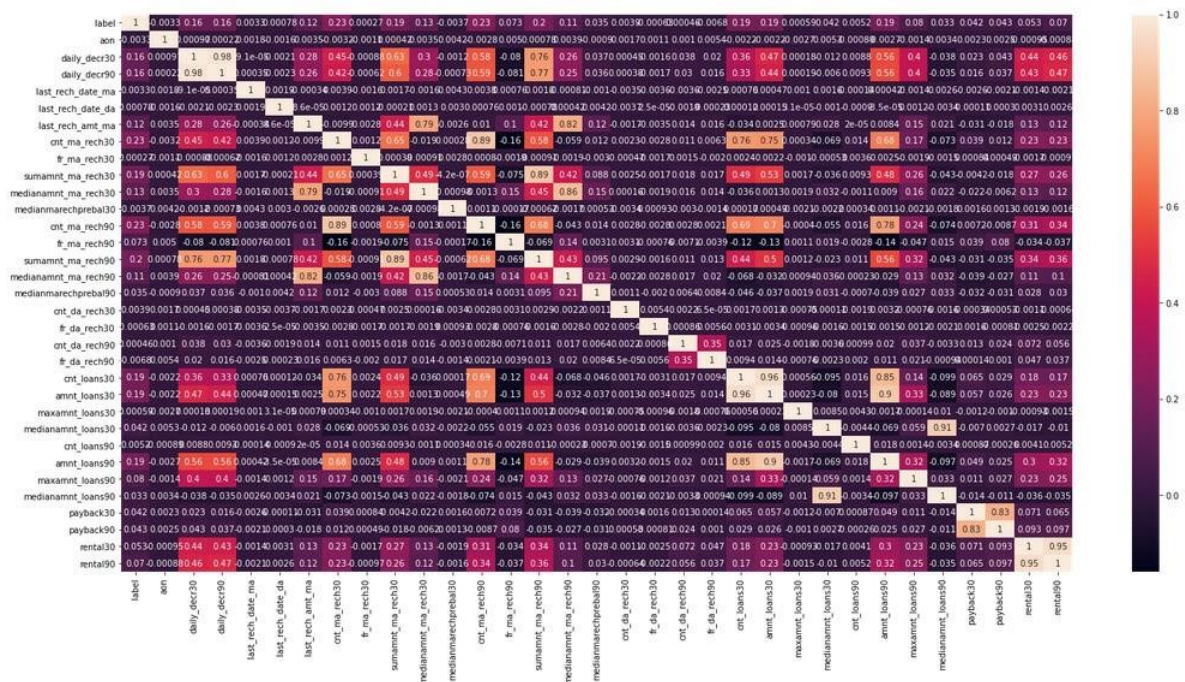
from sklearn.metrics import roc_curve, roc_auc_score

from sklearn.model_selection import GridSearchCV

import joblib
```

# Model/s Development and Evaluation

## 7. Identification of possible problem-solving approaches (methods)



From the above graph it was evident that many features are correlated. Hence PCA was used.

## 8. Testing of Identified Approaches (Algorithms)

Below is the snapshot of all the algorithm used.

```
lg=LogisticRegression()  
rfc=RandomForestClassifier(n_estimators=100)  
dt=DecisionTreeClassifier()  
gnb=GaussianNB()
```

## 9. Run and Evaluate selected models

Below is the snapshot of the algorithm used for solving the given problem

```
lg=LogisticRegression()  
rfc=RandomForestClassifier(n_estimators=100)  
dt=DecisionTreeClassifier()  
gnb=GaussianNB()
```

Snapshot below is the code used for each algorithm

```
model=(lg,rfc,dt,gnb)  
for m in model:  
    m.fit(x_train,y_train)  
    pred=m.predict(x_test)  
    print('Accuracy score of',m)  
    print(accuracy_score(y_test,pred))  
    print(confusion_matrix(y_test,pred))  
    print(classification_report(y_test,pred))
```

```
model=(lg,rfc,dt,gnb)  
for m in model:  
    score=cross_val_score(m,x,y,cv=5)  
    print('Mean Accuracy of', m)  
    print(score.mean())  
    print('\n')
```

```
model=(lg,rfc,dt,gnb)  
for m in model:  
    y_pred_prob=m.predict_proba(x_test)[:,-1]  
    fpr,tpr,thresholds=roc_curve(y_test,y_pred_prob)  
    auc_score=roc_auc_score(y_test,m.predict(x_test))  
    print('AUC Score of', m)  
    print(auc_score)  
    print('\n')  
    plt.plot([0,1],[0,1], 'k--')  
    plt.plot(fpr,tpr,label=m)  
    plt.xlabel('False Positive rate')  
    plt.ylabel('True Positive rate')  
    plt.title(m)  
    plt.show()  
    print('\n\n')
```

## 10. Key Metrics for success in solving problem under consideration

Metrics used were:

1. Accuracy check
2. Cross validation
3. ROC AUC Curve

	Model	Acc Score	Cross Val Score	ROC_AUC_curve
0	LogisticRegression	88	88	55
1	RandomForestClassifier	90	90	66
2	DecisionTreeClassifier	85	86	66
3	GaussianNB	86	54	65

## 11. Visualizations

### 1. Model building

#### A. Logistic Regression

**Accuracy score:** 0.8892772745175266

**Confusion Matrix:**

```
[[ 806 6171]
 [ 576 53383]]
```

**Classification Report:**

	precision	recall	f1-score	support
0	0.58	0.12	0.19	6977
1	0.90	0.99	0.94	53959
accuracy			0.89	60936
macro avg	0.74	0.55	0.57	60936
weighted avg	0.86	0.89	0.85	60936

## B. RandomForestClassifier

**Accuracy score:** 0.909150584219509

**Confusion Matrix:**

```
[[2402  4575]
 [ 961 52998]]
```

**Classification Report:**

	precision	recall	f1-score	support
0	0.71	0.34	0.46	6977
1	0.92	0.98	0.95	53959
accuracy			0.91	60936
macro avg	0.82	0.66	0.71	60936
weighted avg	0.90	0.91	0.89	60936

## C. Decision Tree Classifier

**Accuracy score:** 0.8556682420900618

**Confusion Matrix:**

```
[[ 2871  4106]
 [ 4689 49270]]
```

**Classification Report:**

	precision	recall	f1-score	support
0	0.38	0.41	0.39	6977
1	0.92	0.91	0.92	53959
accuracy			0.86	60936
macro avg	0.65	0.66	0.66	60936
weighted avg	0.86	0.86	0.86	60936

## D. Gaussian NB

**Accuracy score:** 0.8668110804778785

**Confusion Matrix:**

```
[[ 2717  4260]
 [ 3856 50103]]
```

### Classification Report:

	precision	recall	f1-score	support
0	0.41	0.39	0.40	6977
1	0.92	0.93	0.93	53959
accuracy			0.87	60936
macro avg	0.67	0.66	0.66	60936
weighted avg	0.86	0.87	0.87	60936

## 2. Cross Validation

Mean Accuracy of LogisticRegression()  
0.8831719339250993

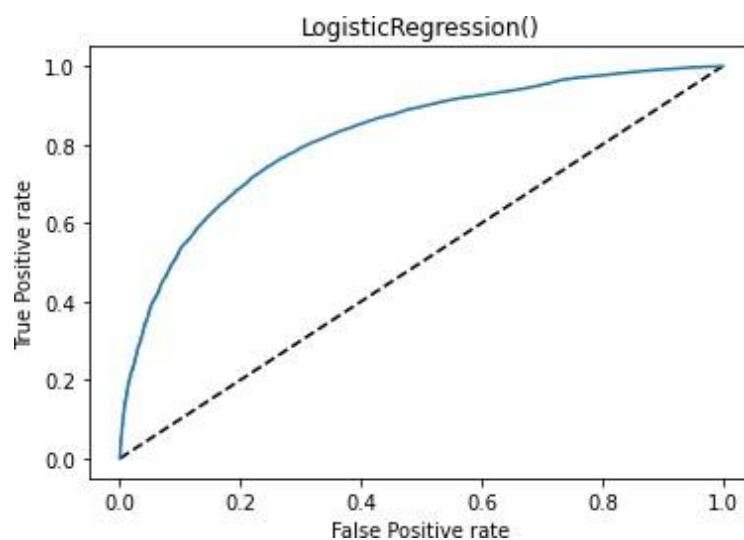
Mean Accuracy of RandomForestClassifier()  
0.9093043975906294

Mean Accuracy of DecisionTreeClassifier()  
0.8640550508345323

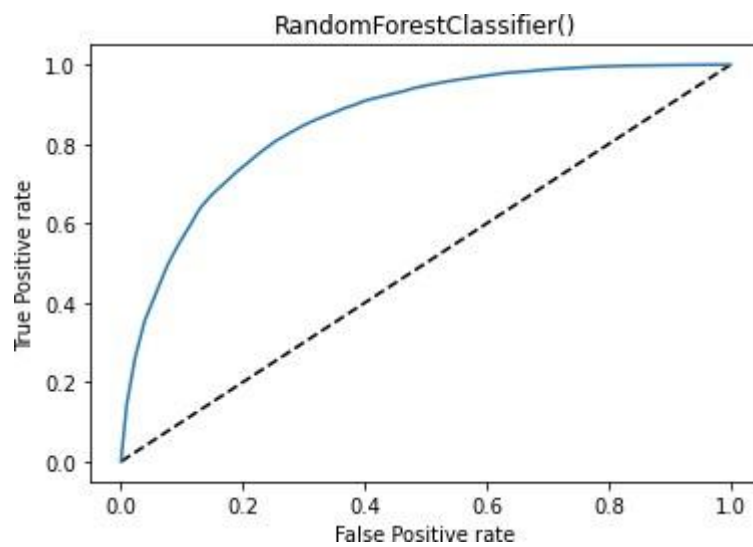
Mean Accuracy of GaussianNB()  
0.5444247150115098

## 3. ROC\_AUC

AUC Score of LogisticRegression()  
0.5524238296291843

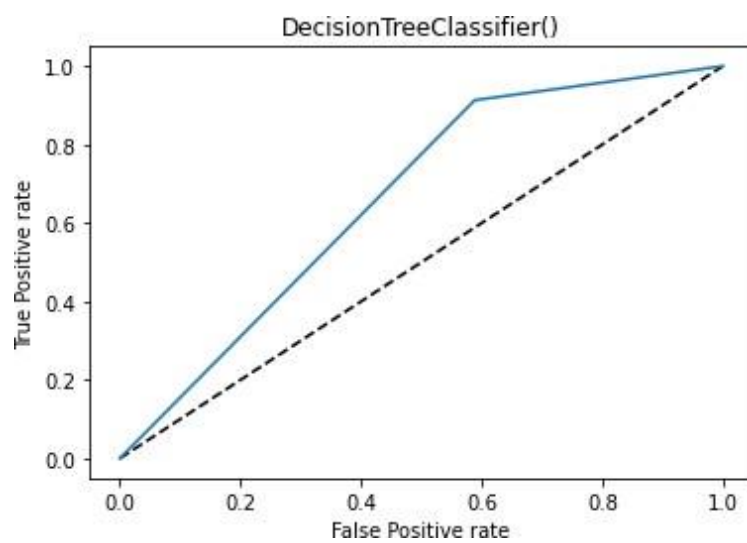


AUC Score of RandomForestClassifier()  
0.6632321123595657

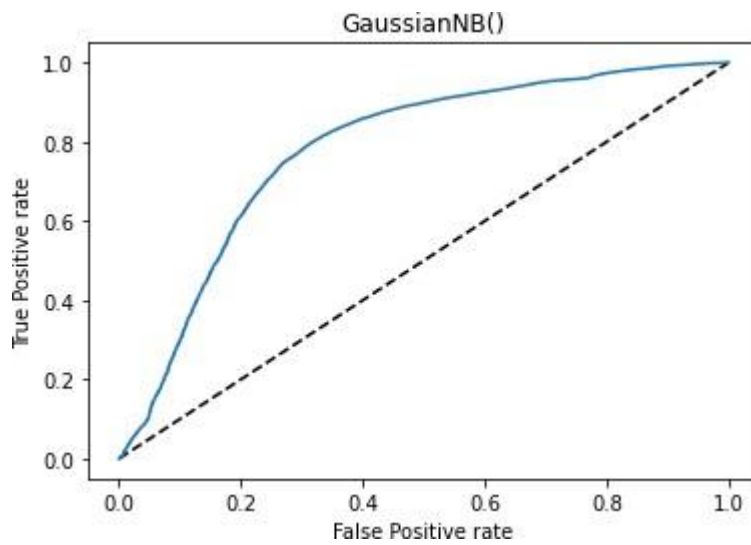


---

AUC Score of DecisionTreeClassifier()  
0.6622977997061523



AUC Score of GaussianNB()  
0.6589803612536406



---

#### 4. HyperTuning of Random Forrest Classifier Model

**Accuracy score:** 90.58028095050544

**Confusion Matrix:**

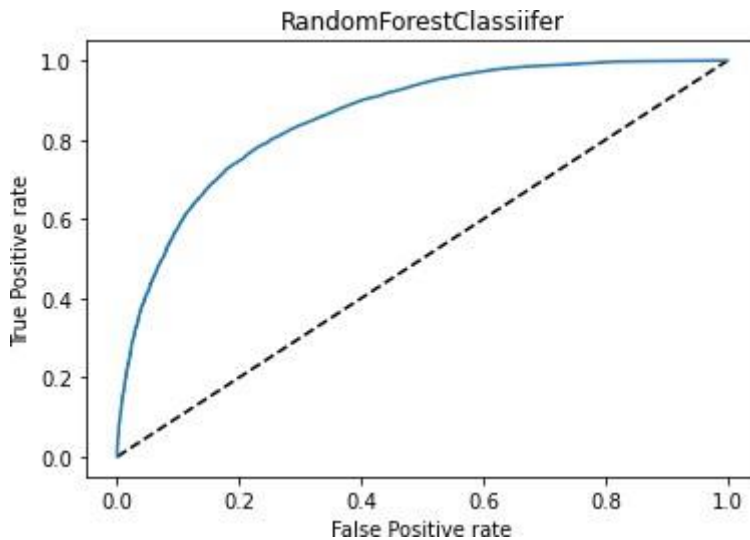
```
[[ 1620  5357]
 [  383 53576]]
```

**Classification Report:**

	precision	recall	f1-score	support
0	0.81	0.23	0.36	6977
1	0.91	0.99	0.95	53959
accuracy			0.91	60936
macro avg	0.86	0.61	0.65	60936
weighted avg	0.90	0.91	0.88	60936

**AUC Score of RandomForestClassifier(max\_depth=9,n\_estimators=250)**  
0.613023067963394





## 1 Interpretation of the Results

From all the model Random Forest has given us the best output of accuracy of 90.50% with ROC\_AUC score of 61.30%.

**The lower ROC\_AUC score is because our data is highly imbalanced.**

## CONCLUSION

### 2 Key Findings and Conclusions of the Study

With the dataset we had we are able to study below behaviour pattern of the customer

1. Customer who maintains average balance of over 50000 in last 90 days is less likely to default the loan.
2. Customer who takes average of 90 loans in last 90 days are less likely to default the loan.
3. Customer spending daily average of over 40000 in last 90 then they are more likely to pay the loan.
4. Customer who are recharging its main account 20 times in the last 90 days is less likely to default the loan.
5. If company is issuing loan 1<sup>st</sup> time to the customer they should look at their average balance, average daily spending and recharge ability over last 90 days. If all this parameters are low then they are chances customer mite default the loan.

6. Customers who are using the telecom services for a long time are mostly non- defaulters and company should try and retain those customer by introducing value added services.
7. Company do not have high number of defaulters. A reminder service should be introduce to remind customer about the timely repayment of their loan.

### 3 Learning Outcomes of the Study in respect of Data Science

Being a complete fresher to the data science domain, this project was very helpful and challenging as well

1. There was a lot this dataset has to offer in terms of understanding the pattern customer follows.
2. Visualization helped to understand the customer behaviour.
3. Feature Engineering was a good learning and it help to study how some feature are related and not related to each other.
4. It was also helpful to know speciality of some of the algorithm.
5. There is lot of assistance on the google if you are stuck with any problem or code, which is a good thing. 'Google' will help you with 'code' but 'logic' has to come from us to study, understand, implement and deliver the project.

### 4. Limitations of this work and Scope for Future Work

The result could have been better. Since it was a class imbalance case, I could have worked on oversampling and under sampling using SMOTE and NearMiss respectively to check if that can offer me a better result.

However due to system limitation, it took lot of time to compute the output especially for Hypertuning the model. Hence dropped the idea to use the same.

# Thank You