

Micro-Credit-Loan

**Submitted by:**

Abhishek Kumar

**ACKNOWLEDGMENT**

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

<https://stackoverflow.com>

<https://arxiv.org/ftp/arxiv/papers/1403/1403.1949.pdf>

<https://pypi.org/project/imbalanced-learn/>

<https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.power_transform.html>

**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.

Here 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).

* Conceptual Background of the Domain Problem

In this case, users have 5 days to pay back the loan, or else they are listed as defaulters, impacting their bank loan eligibility for the future.

* Review of Literature

The project objective is to find out the defaulters. Loan eligibility will be decided based on several parameters (like Daily amount spend & average main account balance in last 30 days, Frequency of recharge for data account & main account in 30/90 days, loan taken in last 90 days & payback time for last 30 days) using different mathematical and statistical tools.

After cleaning the dataset model training and testing process take place using several models like SVM, GaussianNB, KNN, Decision Tree classifier, Random forest classifier etc.

* Motivation for the Problem Undertaken

As we understand the importance of communication and how it affects a person’s life, thus, focusing on providing call and data services and products related to telecom services to low income families and poor customers that can help them in the need of hour.

Using the model, we 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.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

This problem is a type of classification problem where 2 output are defaulter (as “0”) and non-defaulter (as “1”).

The dataset is in CSV format with 37 attributes (36 features as input and 1 target or output).

For a loan amount of 5 payback amount should be 6, and for loan amount of 10 payback amount is 12. We must 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.

* Data Sources and their formats

This Dataset is provided by Flip Robo Technologies (in CSV format)

In this dataset, there are 209593 rows and 37 columns (columns description also provided)

**Column Name 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

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 Indonasian Rupiah)

**medianamnt\_ma\_rech90**: Median of amount of recharges done in main account over last 90 days at user level (in Indonasian Rupiah)

**medianmarechprebal90**: Median of main account balance just before recharge in last 90 days at user level (in Indonasian 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

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Unnamed: 0 209593 non-null int64

1 label 209593 non-null int64

2 msisdn 209593 non-null object

3 aon 209593 non-null float64

4 daily\_decr30 209593 non-null float64

5 daily\_decr90 209593 non-null float64

6 rental30 209593 non-null float64

7 rental90 209593 non-null float64

8 last\_rech\_date\_ma 209593 non-null float64

9 last\_rech\_date\_da 209593 non-null float64

10 last\_rech\_amt\_ma 209593 non-null int64

11 cnt\_ma\_rech30 209593 non-null int64

12 fr\_ma\_rech30 209593 non-null float64

13 sumamnt\_ma\_rech30 209593 non-null float64

14 medianamnt\_ma\_rech30 209593 non-null float64

15 medianmarechprebal30 209593 non-null float64

16 cnt\_ma\_rech90 209593 non-null int64

17 fr\_ma\_rech90 209593 non-null int64

18 sumamnt\_ma\_rech90 209593 non-null int64

19 medianamnt\_ma\_rech90 209593 non-null float64

20 medianmarechprebal90 209593 non-null float64

21 cnt\_da\_rech30 209593 non-null float64

22 fr\_da\_rech30 209593 non-null float64

23 cnt\_da\_rech90 209593 non-null int64

24 fr\_da\_rech90 209593 non-null int64

25 cnt\_loans30 209593 non-null int64

26 amnt\_loans30 209593 non-null int64

27 maxamnt\_loans30 209593 non-null float64

28 medianamnt\_loans30 209593 non-null float64

29 cnt\_loans90 209593 non-null float64

30 amnt\_loans90 209593 non-null int64

31 maxamnt\_loans90 209593 non-null int64

32 medianamnt\_loans90 209593 non-null float64

33 payback30 209593 non-null float64

34 payback90 209593 non-null float64

35 pcircle 209593 non-null object

36 pdate 209593 non-null object

**dtypes: float64(21), int64(13), object(3)**

* Data Preprocessing Done

After loading the dataset:

ds.info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 209593 entries, 0 to 209592

Data columns (total 37 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Unnamed: 0 209593 non-null int64

1 label 209593 non-null int64

2 msisdn 209593 non-null object

3 aon 209593 non-null float64

4 daily\_decr30 209593 non-null float64

5 daily\_decr90 209593 non-null float64

6 rental30 209593 non-null float64

7 rental90 209593 non-null float64

8 last\_rech\_date\_ma 209593 non-null float64

9 last\_rech\_date\_da 209593 non-null float64

10 last\_rech\_amt\_ma 209593 non-null int64

11 cnt\_ma\_rech30 209593 non-null int64

12 fr\_ma\_rech30 209593 non-null float64

13 sumamnt\_ma\_rech30 209593 non-null float64

14 medianamnt\_ma\_rech30 209593 non-null float64

15 medianmarechprebal30 209593 non-null float64

16 cnt\_ma\_rech90 209593 non-null int64

17 fr\_ma\_rech90 209593 non-null int64

18 sumamnt\_ma\_rech90 209593 non-null int64

19 medianamnt\_ma\_rech90 209593 non-null float64

20 medianmarechprebal90 209593 non-null float64

21 cnt\_da\_rech30 209593 non-null float64

22 fr\_da\_rech30 209593 non-null float64

23 cnt\_da\_rech90 209593 non-null int64

24 fr\_da\_rech90 209593 non-null int64

25 cnt\_loans30 209593 non-null int64

26 amnt\_loans30 209593 non-null int64

27 maxamnt\_loans30 209593 non-null float64

28 medianamnt\_loans30 209593 non-null float64

29 cnt\_loans90 209593 non-null float64

30 amnt\_loans90 209593 non-null int64

31 maxamnt\_loans90 209593 non-null int64

32 medianamnt\_loans90 209593 non-null float64

33 payback30 209593 non-null float64

34 payback90 209593 non-null float64

35 pcircle 209593 non-null object

36 pdate 209593 non-null object

dtypes: float64(21), int64(13), object(3)

memory usage: 59.2+ MB

Observations:

#no null or missing values in data set.

Total number of columns having float value= 21

Total number of columns having integer value= 13

Total number of columns having object value (pdate , pcircle, msisdn ) = 3

so we can drop obect columns also Unnamed is also not providing any information (working as index value) so we can drop it as well.

ds.isnull().sum()

Unnamed: 0 0

label 0

msisdn 0

aon 0

daily\_decr30 0

daily\_decr90 0

rental30 0

rental90 0

last\_rech\_date\_ma 0

last\_rech\_date\_da 0

last\_rech\_amt\_ma 0

cnt\_ma\_rech30 0

fr\_ma\_rech30 0

sumamnt\_ma\_rech30 0

medianamnt\_ma\_rech30 0

medianmarechprebal30 0

cnt\_ma\_rech90 0

fr\_ma\_rech90 0

sumamnt\_ma\_rech90 0

medianamnt\_ma\_rech90 0

medianmarechprebal90 0

cnt\_da\_rech30 0

fr\_da\_rech30 0

cnt\_da\_rech90 0

fr\_da\_rech90 0

cnt\_loans30 0

amnt\_loans30 0

maxamnt\_loans30 0

medianamnt\_loans30 0

cnt\_loans90 0

amnt\_loans90 0

maxamnt\_loans90 0

medianamnt\_loans90 0

payback30 0

payback90 0

pcircle 0

pdate 0

dtype: int64

# total null value in dataset is 0

Total number of columns having object value (pdate , pcircle, msisdn ) = 3

so we can drop obect columns also Unnamed is also not providing any information (working as index value) so we can drop it as well.

drop\_cloumns= ['Unnamed: 0','pcircle','msisdn','pdate']

for i in ds[drop\_cloumns]:

ds.drop (i, axis=1, inplace=True)

object value columns and Unnamed: 0 (which is working just like index and providing any further informations) droped from the dataset

now,

ds.shape

(209593, 33)

new datset (after droping object value columns and unnamed columns) having 33 columns now

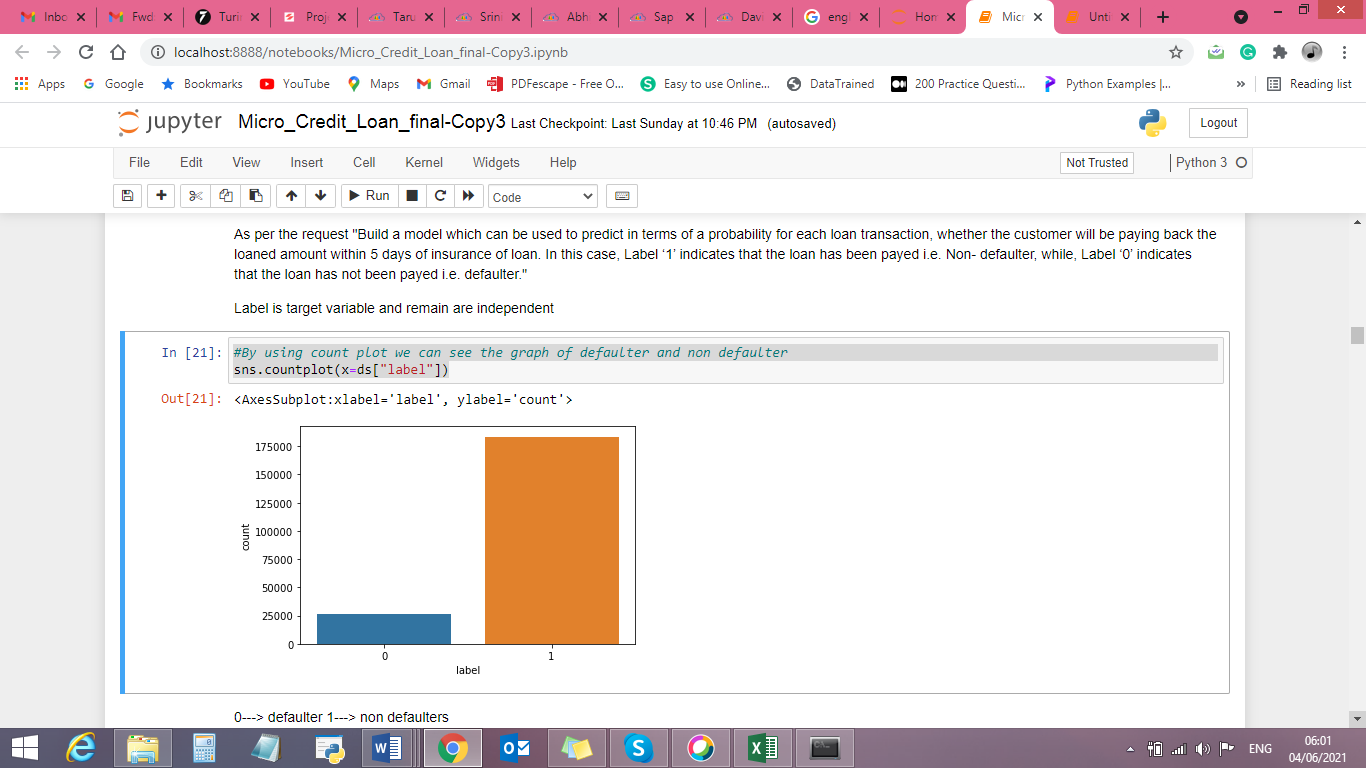
ds.keys()

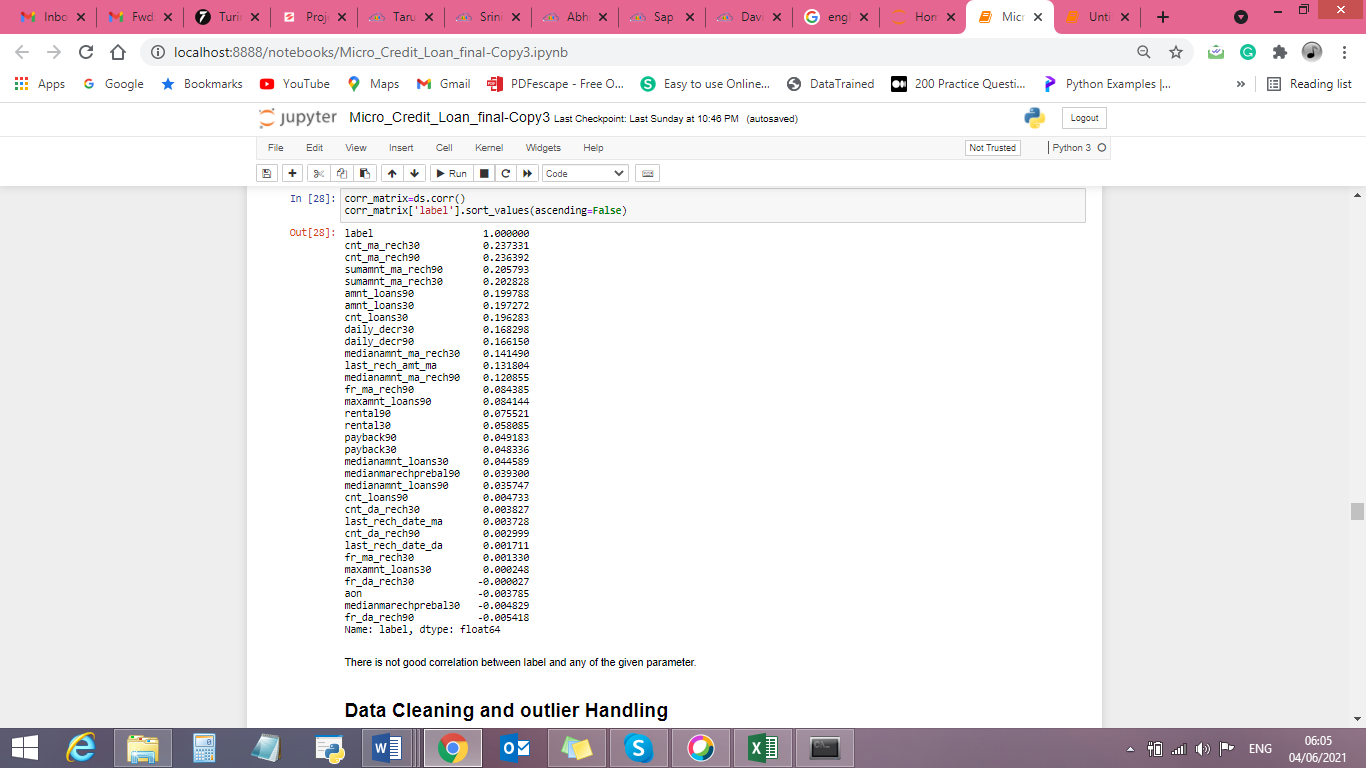
Index(['label', 'aon', 'daily\_decr30', 'daily\_decr90', 'rental30', 'rental90','last\_rech\_date\_ma', 'last\_rech\_date\_da', 'last\_rech\_amt\_ma','cnt\_ma\_rech30', 'fr\_ma\_rech30', 'sumamnt\_ma\_rech30','medianamnt\_ma\_rech30', 'medianmarechprebal30', 'cnt\_ma\_rech90','fr\_ma\_rech90', 'sumamnt\_ma\_rech90', 'medianamnt\_ma\_rech90','medianmarechprebal90', 'cnt\_da\_rech30', 'fr\_da\_rech30', 'cnt\_da\_rech90', 'fr\_da\_rech90', 'cnt\_loans30', 'amnt\_loans30','maxamnt\_loans30', 'medianamnt\_loans30', 'cnt\_loans90', 'amnt\_loans90','maxamnt\_loans90', 'medianamnt\_loans90', 'payback30', 'payback90']

* Data Inputs- Logic- Output Relationships

As per the request "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."

Label is target variable and remain are independent



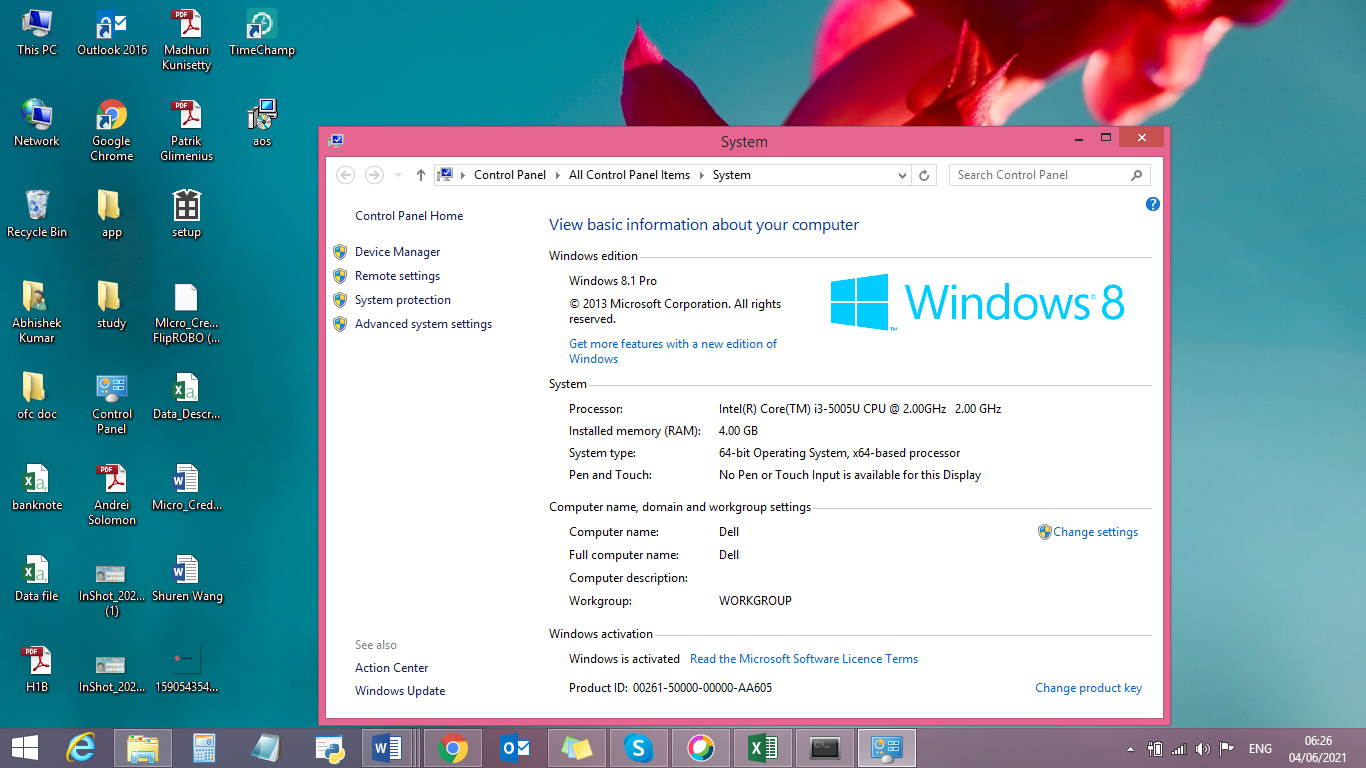
Checked correlation between output parameter (i.e. label) and input parameters (i.e. other than label)

* State the set of assumptions (if any) related to the problem under consideration

No assumptions

* Hardware and Software Requirements and Tools Used

Listing down the hardware and software requirements along with the tools, libraries and packages used. Describe all the software tools used along with a detailed description of tasks done with those tools.



Imported basic libraries (for EDA, data visualization, mathematical and statistical analysis, training the data, creating the model, checking the accuracy and saving the model)

import numpy as np

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.linear\_model import LogisticRegression

from sklearn.naive\_bayes import GaussianNB

from sklearn.tree import DecisionTreeClassifier

from sklearn.svm import LinearSVC

from sklearn.svm import SVC

from sklearn.neighbors import KNeighborsClassifier

from sklearn.ensemble import RandomForestClassifier

from sklearn.ensemble import AdaBoostClassifier

from sklearn.ensemble import GradientBoostingClassifier

from sklearn.metrics import accuracy\_score,confusion\_matrix,classification\_report

from sklearn.preprocessing import StandardScaler

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import PowerTransformer

from sklearn.preprocessing import power\_transform

from sklearn.model\_selection import GridSearchCV

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

from sklearn.metrics import roc\_curve

from sklearn.metrics import roc\_auc\_score

import joblib

import warnings

warnings.filterwarnings("ignore")

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Analytical Approach 🡪 Based on type of data by performing EDA I have decided which model to be used for this data.

Statistical Approach 🡪 Data should be in scaled manner, it should not be distorted, skewness should resolved, outliner should removed.

* Testing of Identified Approaches (Algorithms)

Logistic Regression-

GaussianNB-

Linear SVC-

decision tree classifier-

Kneighors classifier-

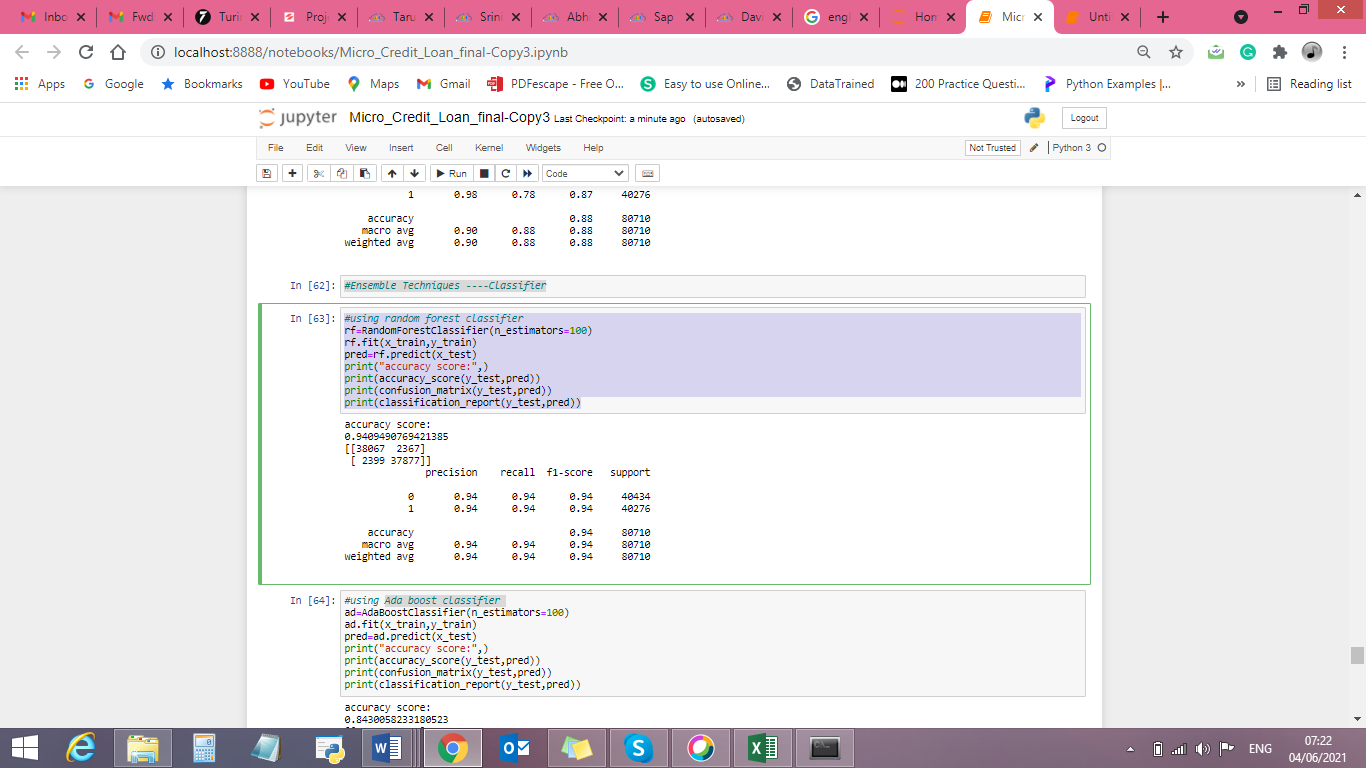
#Ensemble Techniques ----Classifier

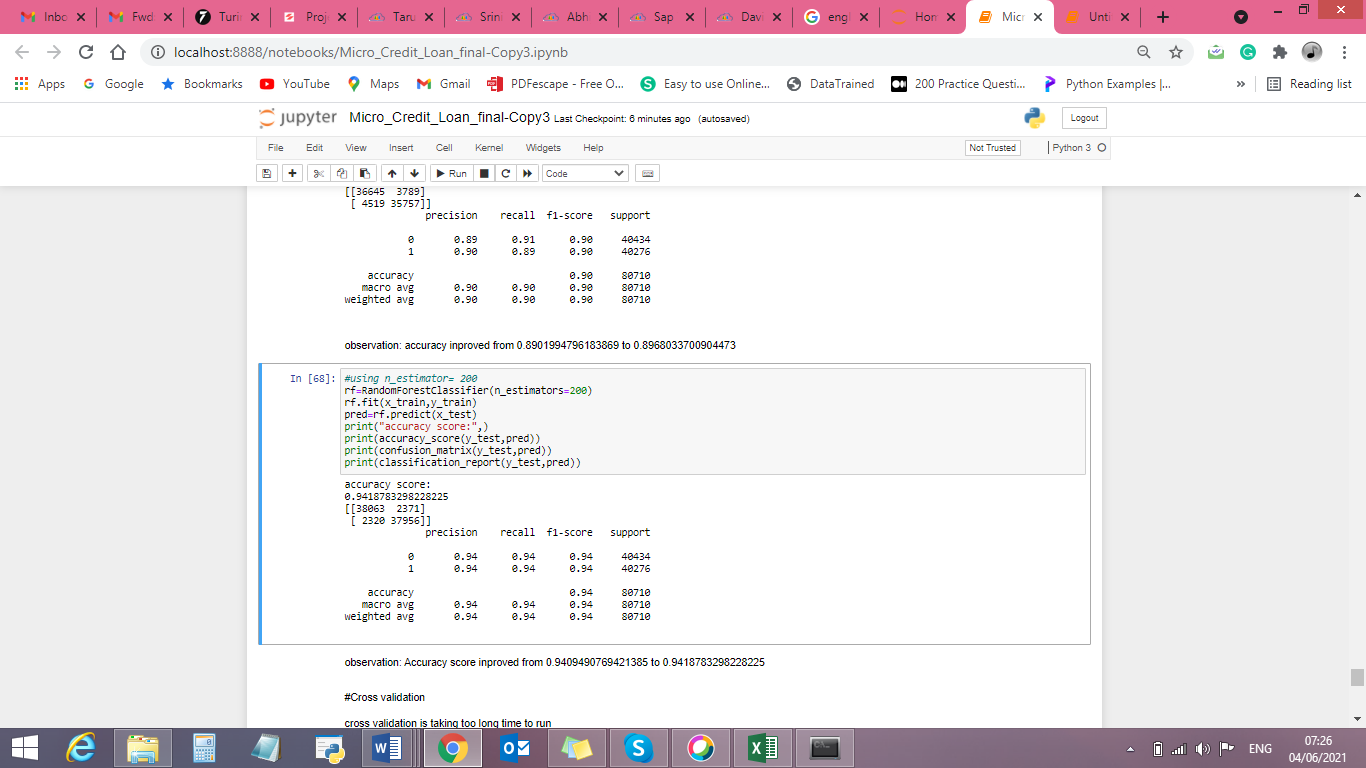
random forest classifier

Ada boost classifier

Gradient Boosting Classifier

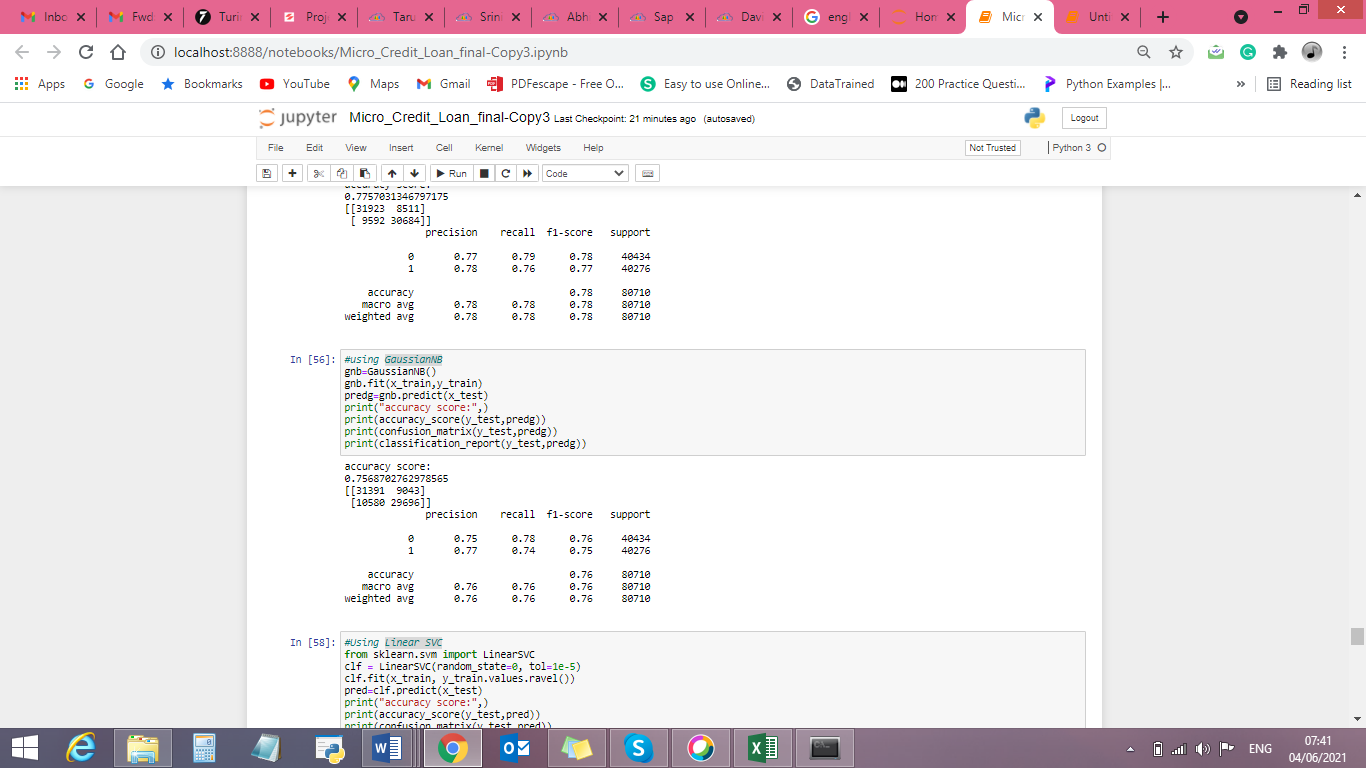
* Run and Evaluate selected models

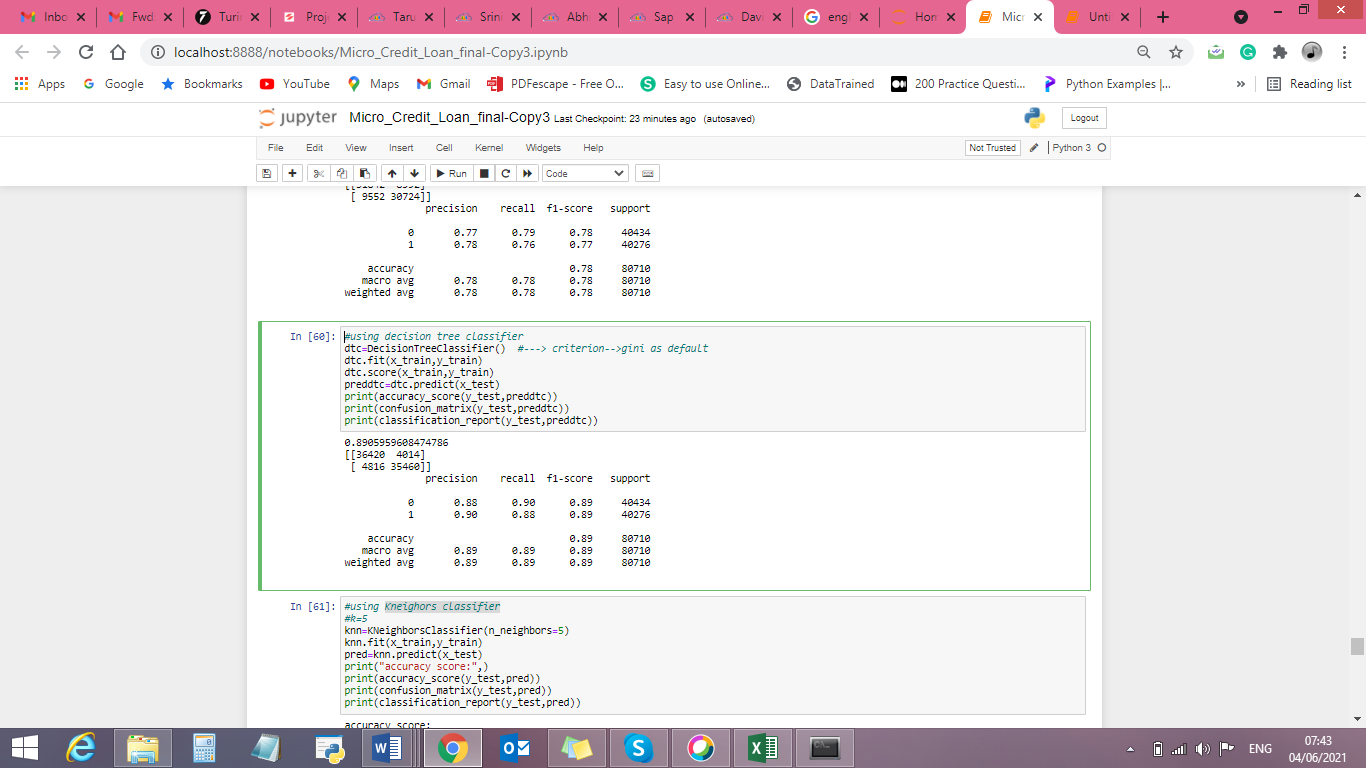


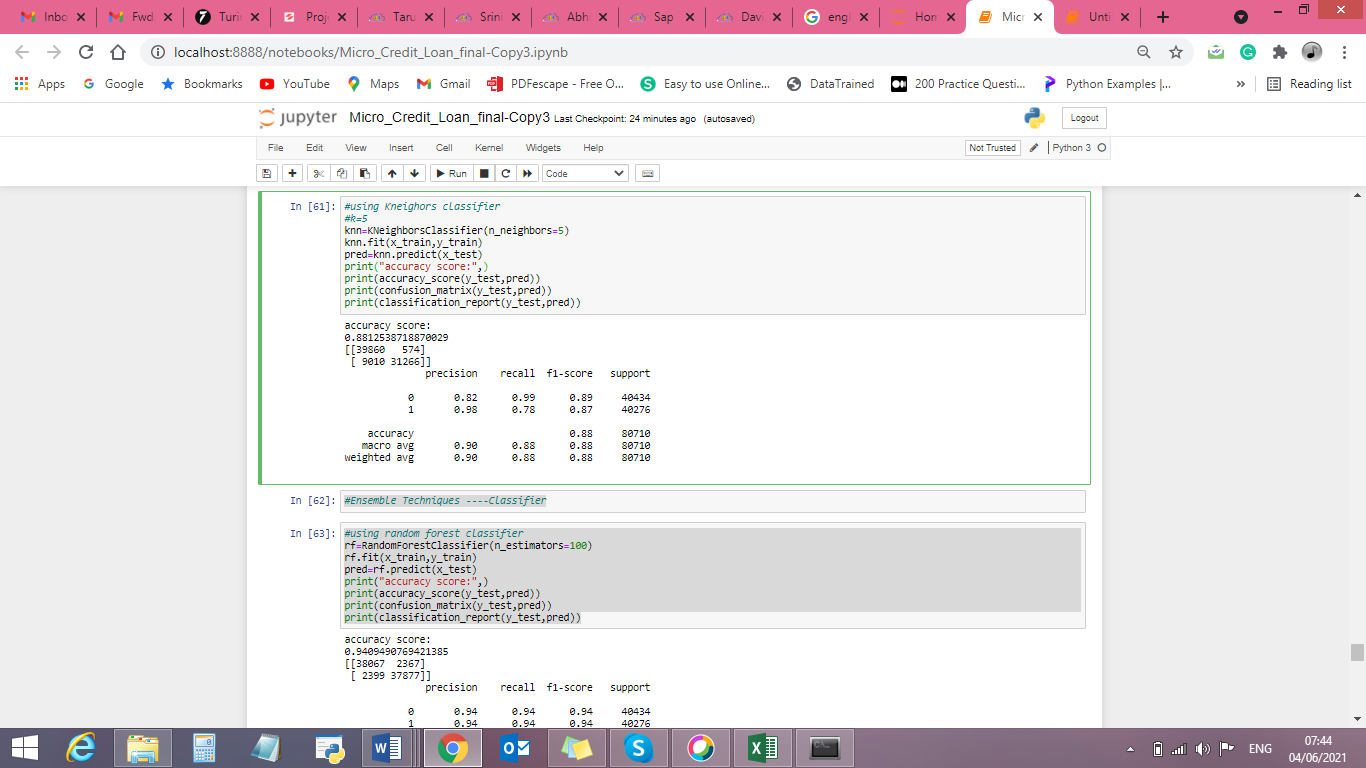


* Key Metrics for success in solving problem under consideration

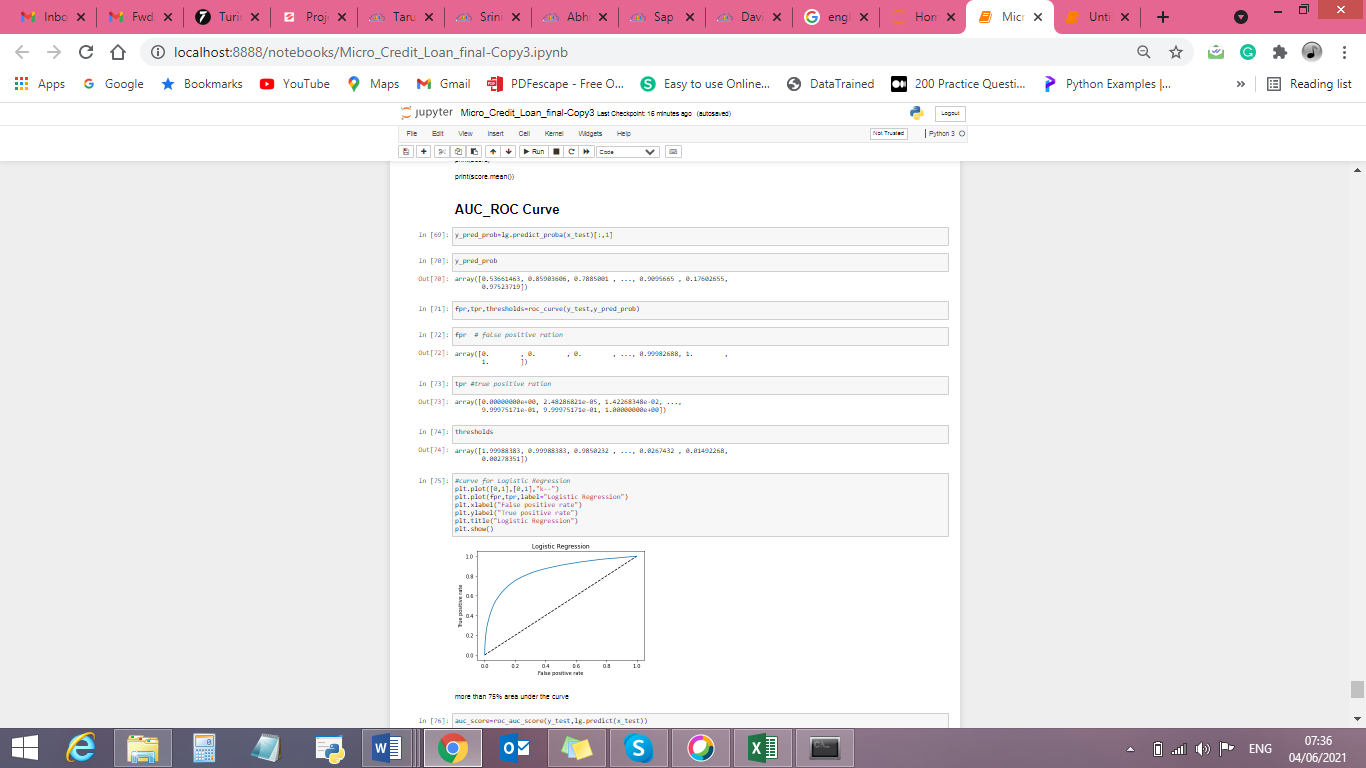
Key Metrices used were the Accuracy Score and AUC & ROC Curve as this was binary classification problem and we focus more on AUC & ROC curve metrices to observe True Positive Rate and False Positive Rare, for users who paid the loan and falsely marked as default and will their affect the credit score

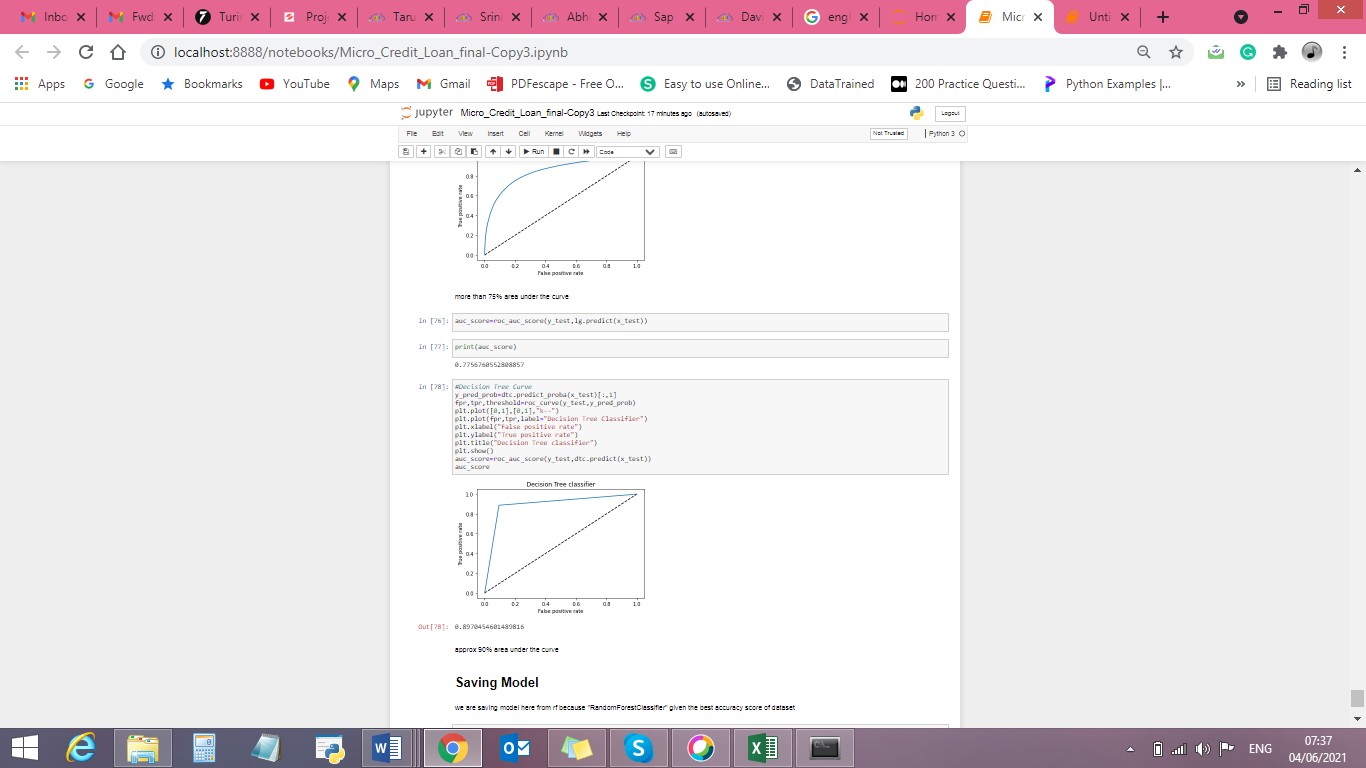






* Visualizations





* Interpretation of the Results

we are selecting "RandomForestClassifier" because it is giving the best accuracy score of dataset (approx. 95%)

**CONCLUSION**

* Key Findings and Conclusions of the Study

By using several data parameters, we can find or predict loan defaulter and non-defaulters by which financial form can help more needy people.

* Learning Outcomes of the Study in respect of Data Science

List down your learnings obtained about the power of visualization, data cleaning and various algorithms used. You can describe which algorithm works best in which situation and what challenges you faced while working on this project and how did you overcome that.

Load the data- Data is categorical.

Check the basic details (Null Value, Dtype, Shape etc.)- no null value

Identify the target and independent features and perform EDA (Univariate, Bivariate and Multivariate analysis) using Data Visualisation and Statistical approach accordingly- lable is target parameter and rest are independent features

Perform data cleaning, outliers handling, missing value imputation- outlier is present so by using IQR resolving outliers.

Build model- creating test, train and split

Evaluate the models- used several models in which Random forest giving best accuracy.

Performed hyper parameter tuning

Check the prediction score using accuracy score & get AUC\_ROC Curve score.

Save file using joblib library.

* Limitations of this work and Scope for Future Work

We are getting accuracy 94% but for remain 6% we are not still sure about a person will pay the loan or not.

We can get more parameters related to financial status from the customer to get better accuracy.