

PROJECT ON:

**MICRO CREDIT LOAN DEFAULTER**

PREDICTION OF A PROBABILITY FOR EACH LOAN TRANSACTION OF THE CUSTOMER WILL BE PAYING BACK THE LOANED AMOUNT WITHIN 5 DAYS OF INSURANCE OF LOAN

**SUBMITTED BY:**

PRATAP LAVATE

**ACKNOWLEDGMENT**

I take this opportunity to acknowledge everyone who have helped us in every stage of this project.

Firstly, I am indebtedly grateful to our Datatrained teacher Mrs. Deepika Sharma for his support in concept clearing. I am also greatful to our SME Ms. Swati Rustagi for their valuable guidance and support. Without their support this project would not have been completed.

**INTRODUCTION**

* Business Problem Framing
* Build a Classifier 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 issuance of loan.
* 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).
* It can also 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
* This project is based on Microfinance services (MFS). A Microfinance Institution (MFI) is an organization that offers financial services to low income populations.
* 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).
* 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.
* Review of Literature
* In this project the sample data is provided to us from our client database. 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.
* The client wants some predictions that could help them in further investment and improvement in selection of customers. So we build a machine learning model to help them.
* First we do all the data preprocessing steps and do EDA to visualize the data graphically and after that we make a machine learning model in order to improve the selection of customers for the credit.
* Motivation for the Problem Undertaken

Banking is the fastest growing sectors in every country and loan business holds a big share of banks revenue and finding right customers who will repay the loan on time is very important so I believed it would be very interesting to work on this project.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

To predict the outcomes we use the random forest classifier

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

We can understand the working of Random Forest algorithm with the help of following steps −

* **Step 1** − First, start with the selection of random samples from a given dataset.
* **Step 2** − Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
* **Step 3** − In this step, voting will be performed for every predicted result.
* **Step 4** − At last, select the most voted prediction result as the final prediction result.

The following diagram will illustrate its working −



* Data Sources and their formats

1) total 36 variables are present in the dataset in which 3 objective type,12 integer type and 21 float type

2)there is no null values present in the dataset

3)target variable is of classification type

4)target variable is imbalance

* Data Preprocessing Done
* First we check the information of the given dataset and extract information about dataset like number of rows and columns and data types of the features using some code. In this dataset 209593 row with 36 features.
* Drop duplicates rows if present in dataset
* Then we check for the null values present in our dataset. If null values are present then fill it via mean, median or mode. Or also you can remove that rows but kindly check it properly. In this dataset no any null value present.
* After that we check the summary statistics of our dataset. This part tells about the statistics of our dataset i.e. mean, median, max value, min values and also it tell whether outliers are present in our dataset or not
* We also check the correlation of our dataset to check the correlation of the columns with each other and output features. If columns are highly correlated with each other approx 90% or above then remove those columns to avoid multi co-linearity problem.
* Exploring the data variable and cleaning the data one by one features which gives more information about dataset and drop those row which value not realistic and related with outcome of loan.
* We delete the pcircle column because it has only one unique value that tells that collected data is only for one circle.
* We cannot remove outliers because more than 20% of our data are removed.
* Data Inputs- Logic- Output Relationships
* Input were different variables from the past records of the customers and output was where the customers will go defaulter or not, so considering the past data we can create a model to predict that.
* We check the correlation between target variable and attribute variable
* From that we can say that if positive correlation is present then as attribute values increases target values also increases
* From that we can say that if negative correlation is present then as attribute values decreases target values also decreases.
* State the set of assumptions (if any) related to the problem under consideration

1)In logistic regression It assumes that there is minimal or no multicollinearity among the independent variables

2)In decision tree classifier Initially, whole training data is considered as root

3)in random forest we assume that of no formal distributions. Being a non-parametric model, it can handle skewed and multi-modal data.

* Hardware and Software Requirements and Tools Used
* Hardware:
* Processor—Intel (R) Core(TM) i5-2430M CPU @ 2.40GHz
* Installed Memory(RAM)—8.00 GB
* System type—64-bit Operating System
* Software: Windos 10 Pro
* We have used Python Package because it is powerful and general purpose programming language.
* I used Jupyter notebook as my coding environment. I used pandas and Numpy to do handle the data, I used Matplot and Seaborn to do visualization of the data, Scipy is used for some static analysis and finally sk-learn is used to build the model and test it.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)
* Loading the data-we load the data using pandas
* Checking the basic details (Null Value, Dtype, Shape etc.)-checked this attribute in pandas library
* Identifying the target and independent features and perform EDA (Univariate, Bivariate and Multivariate analysis) using Data Visualisation and Statistical approach accordingly.- this is done by various libraries like for visualization we use seaborn, for statistical approach we use stats.
* Perform data cleaning, outliers handling, missing value imputation- for data cleaning we use pandas and for outliers we use zscore
* Do feature Engineering
* Build model- for building model we use scikit learn library
* Evaluate the model
* Again perform feature engineering and Feature selection (if required)
* Perform hyperparameter tunning
* Evaluate the model again
* Make prediction
* Testing of Identified Approaches (Algorithms)
* Logistic Regression,
* Random Forest Classifier,
* Gaussian NB,
* Decision Tree Classifier,
* Extra Trees Classifier,
* Bagging Classifier
* Run and Evaluate selected models

As we know that this dataset is imbalance so first we see the result without doing any sampling and for that I use Logistic Regression with KFold cross validation and hyperparameter tuning.

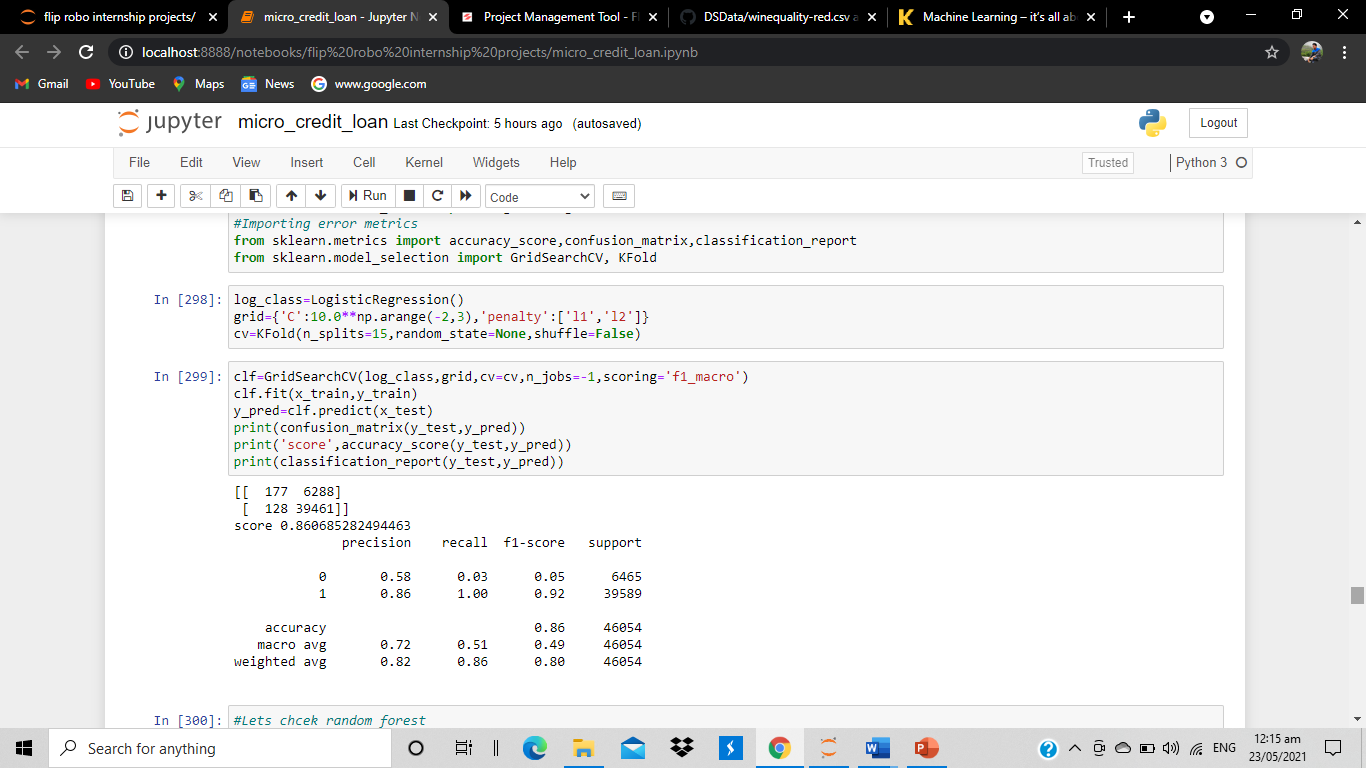
We also use Random Forest Classifier as our evaluation model without using hyperparameter tuning because our dataset is too large and it takes more than hour to give the result.

After that I use some sampling technique and use severals models with sampled training data and see the results.

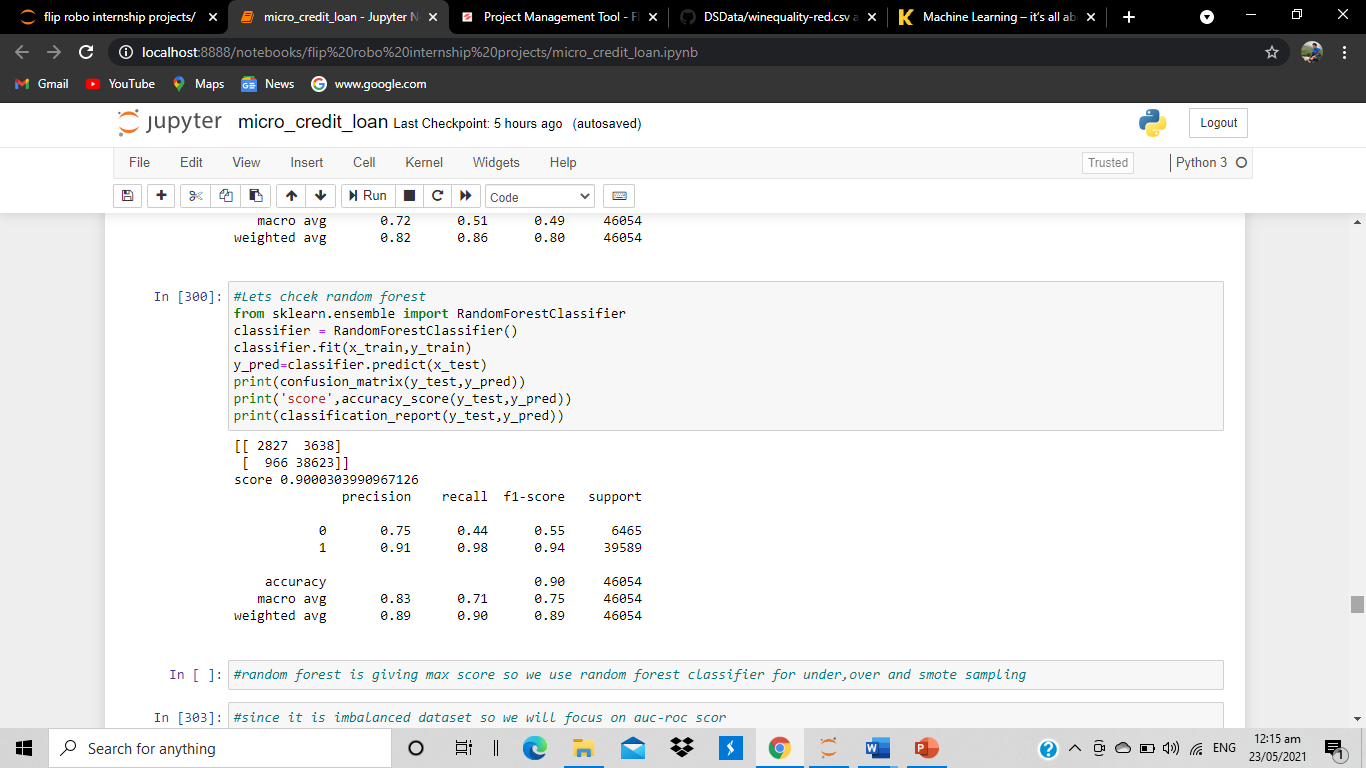
There are three sampling technique which I have used to handle the imbalance problem of our dataset.After using Sampling technique to balance our dataset and then we apply several models and see the results.

* **Under Sampling:** By using the under sampling we are trying to reduce the points of maximum labels.
* **Over Sampling:** By using the over sampling we are trying to increase the points of minimum labels.
* **SMOTE:** SMOTET is a hybrid method which uses on under sampling method (Tomek) in with over sampling method (SMOTE).

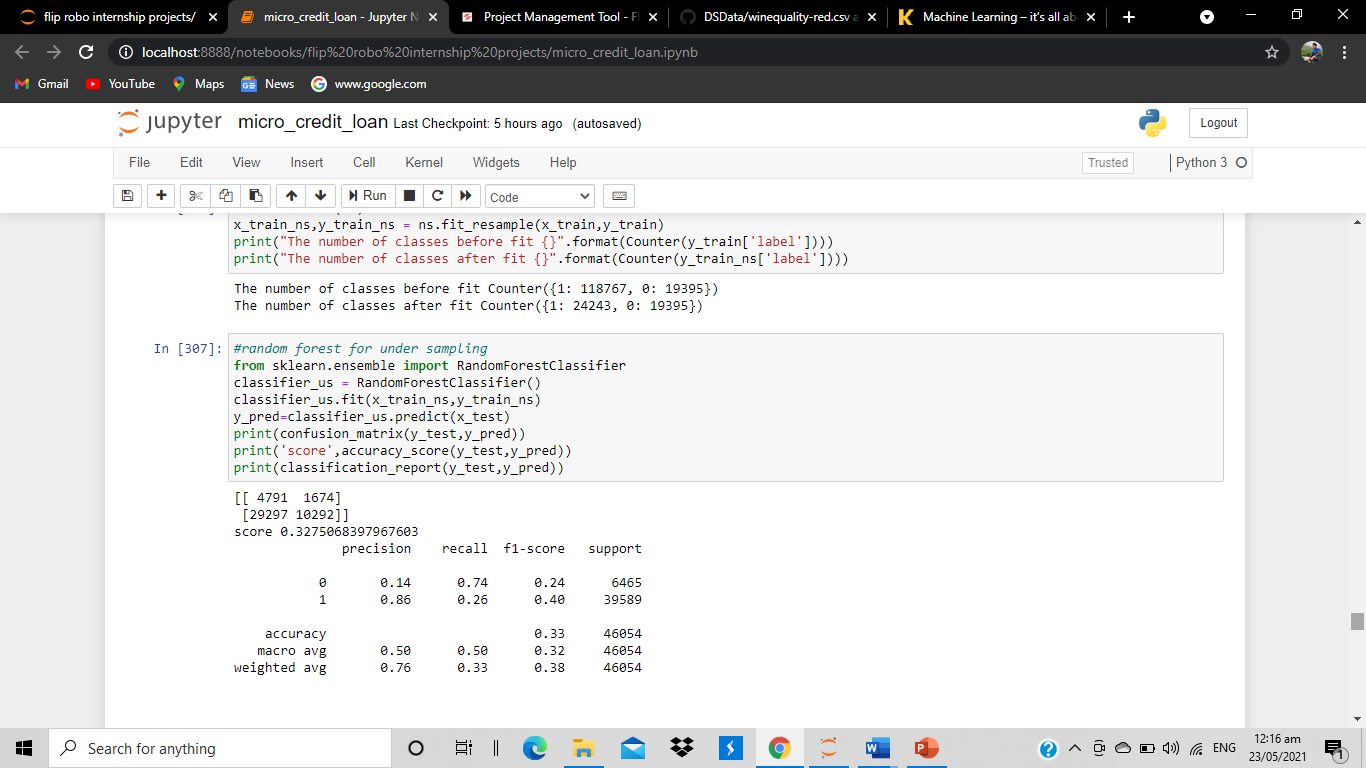
1)we use grid search cv for logistic regression and we get accuracy score of 86% but here we didn’t use resampling techniques



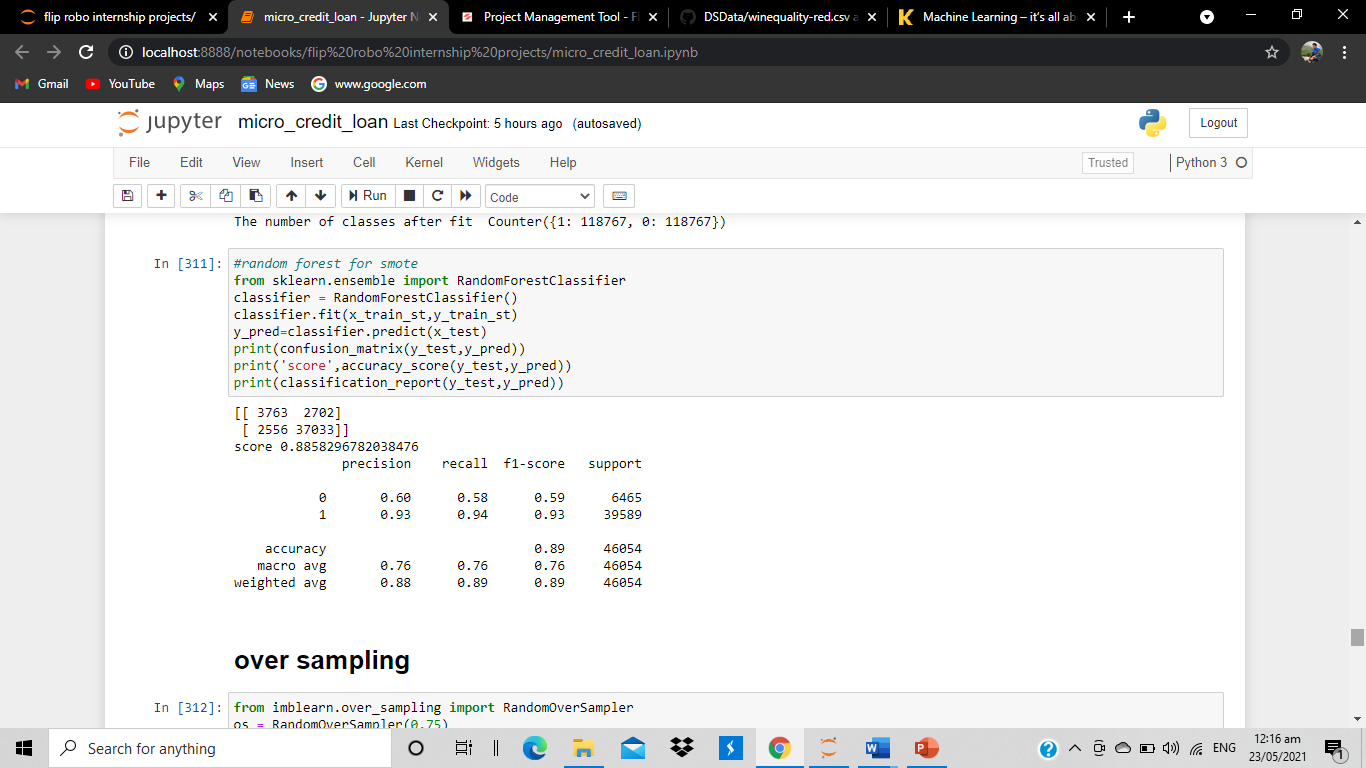
2)we use random forest we got 90% accuracy



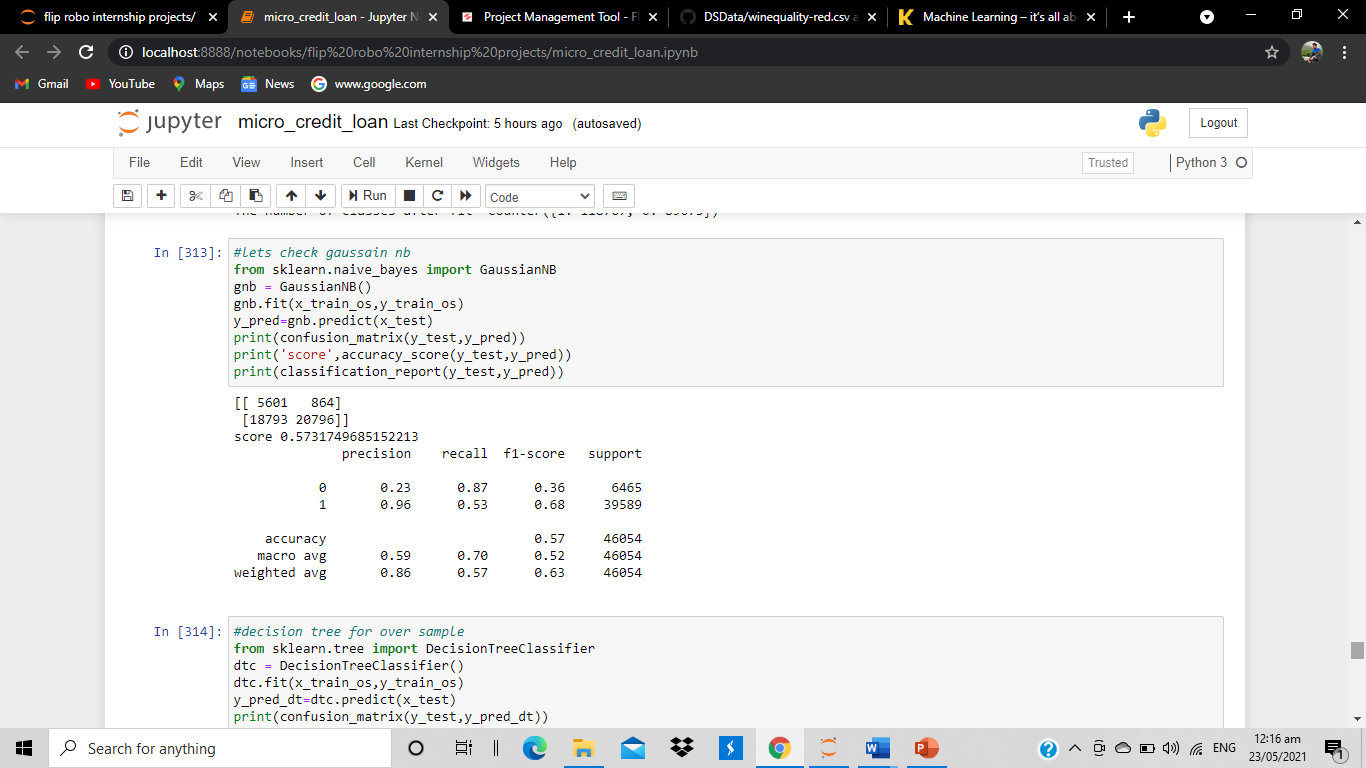
3)class is imbalance so we use under sampling technique so in random forest for under sampling we got 32% accuracy



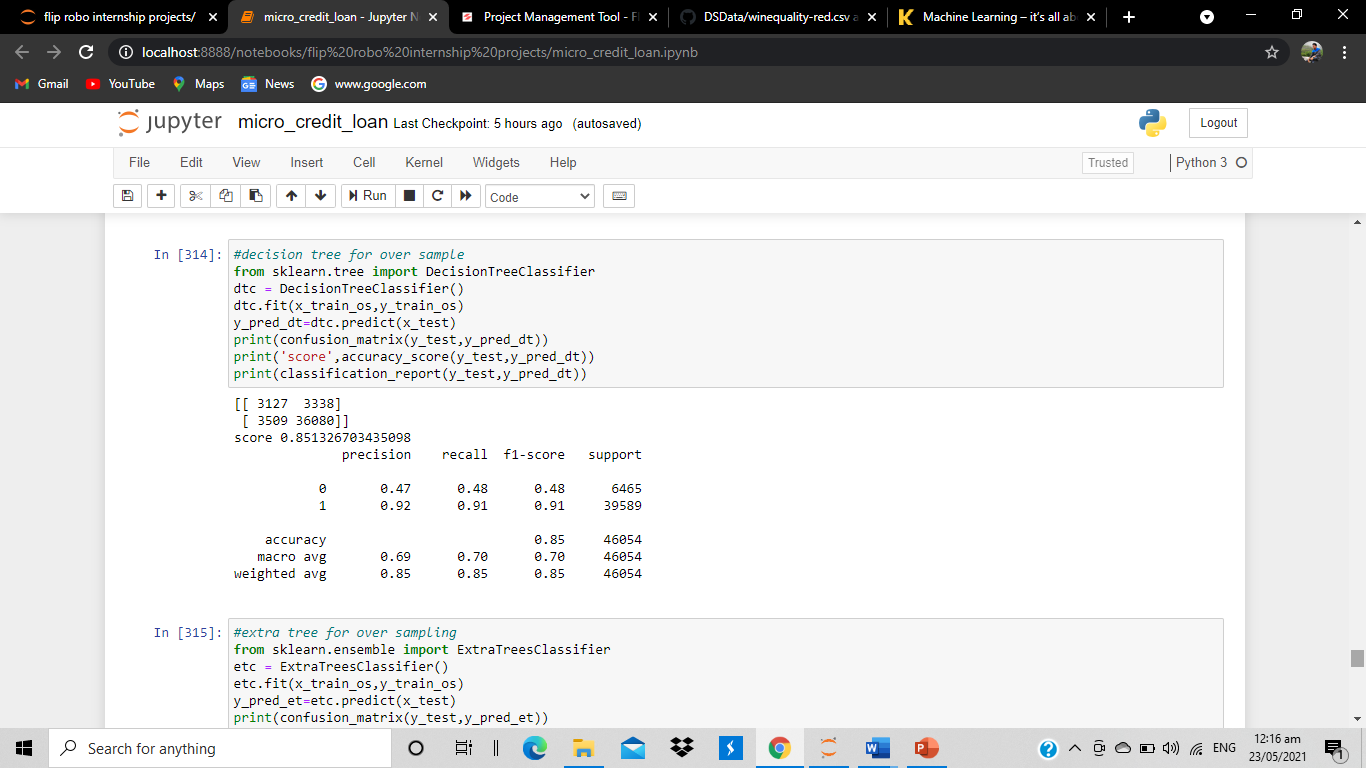
4) we use then smote technique and check for random forest and it gives 88% accuracy



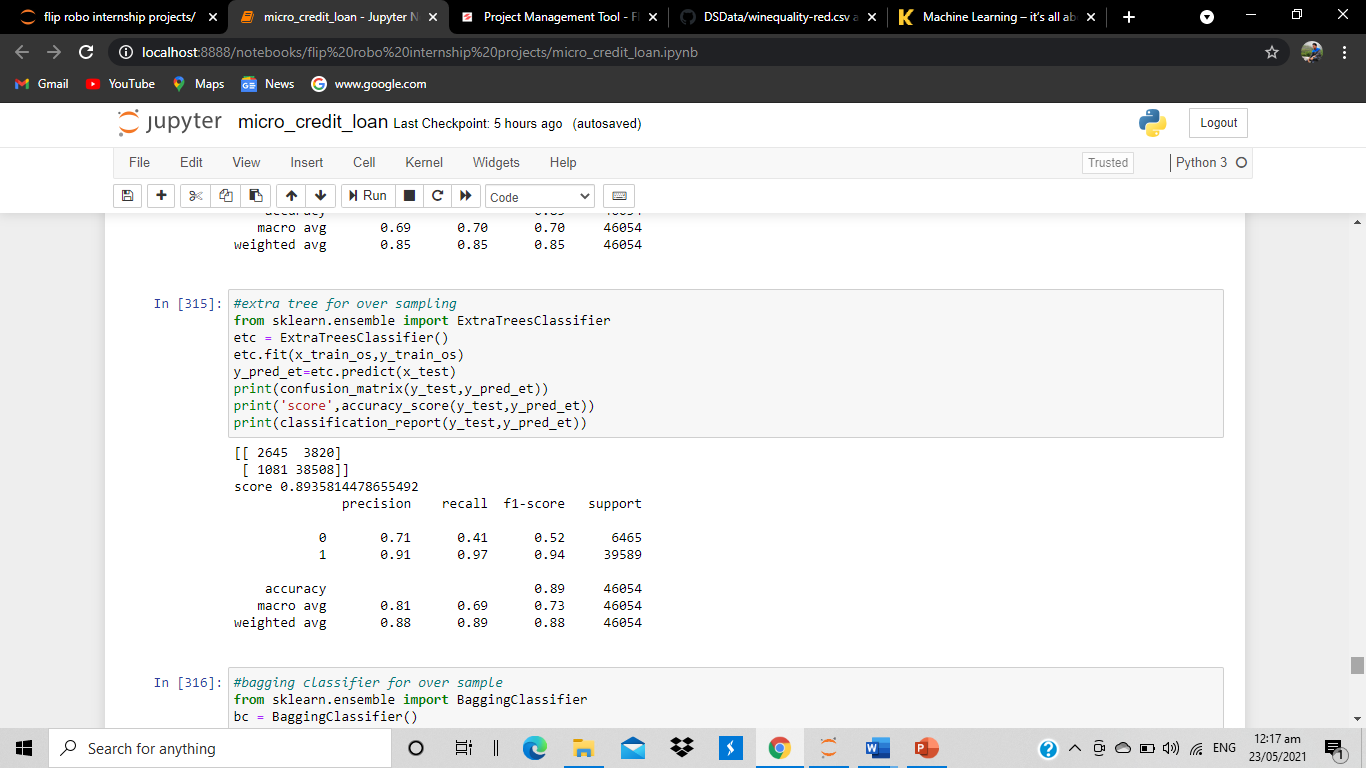
5)for over sampling we use gaussian nb classifier it gives 57% accuracy



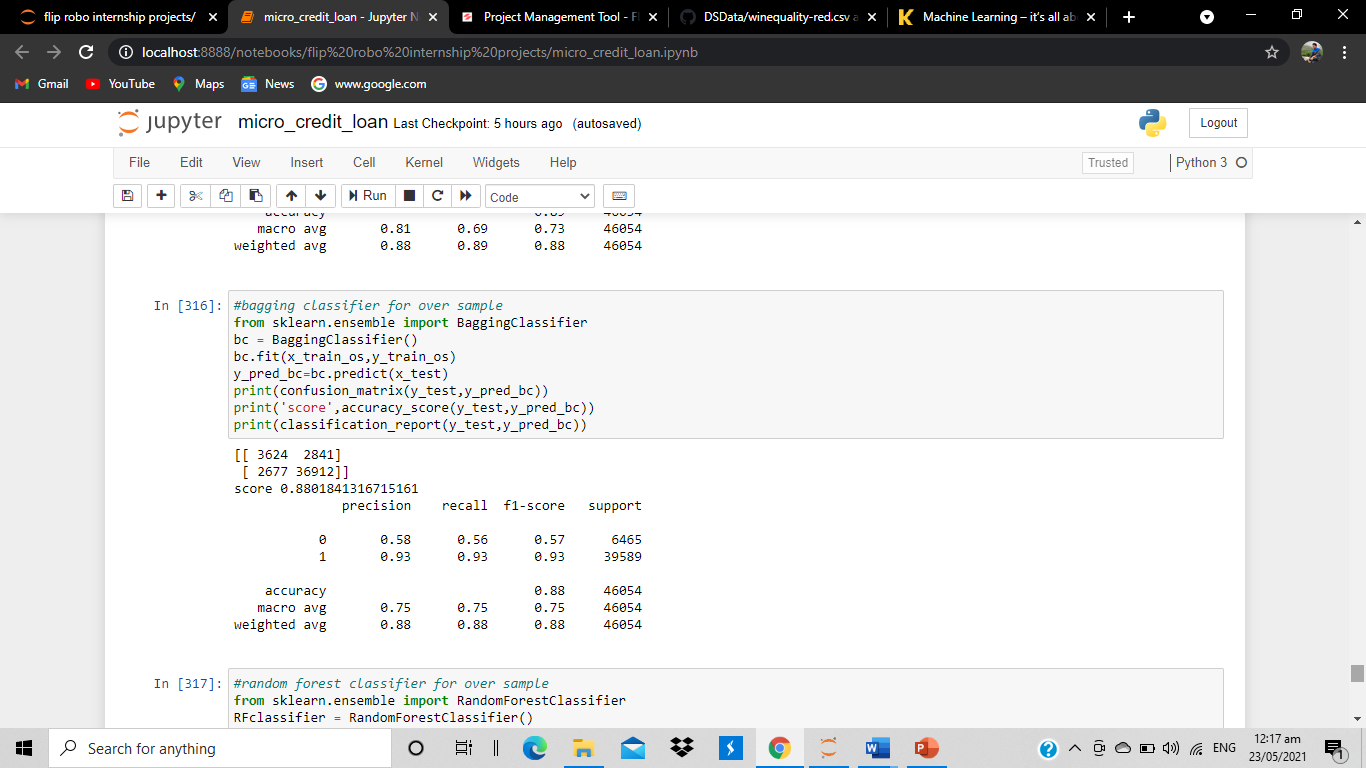
6)decision tree for over sample accuracy got 85%



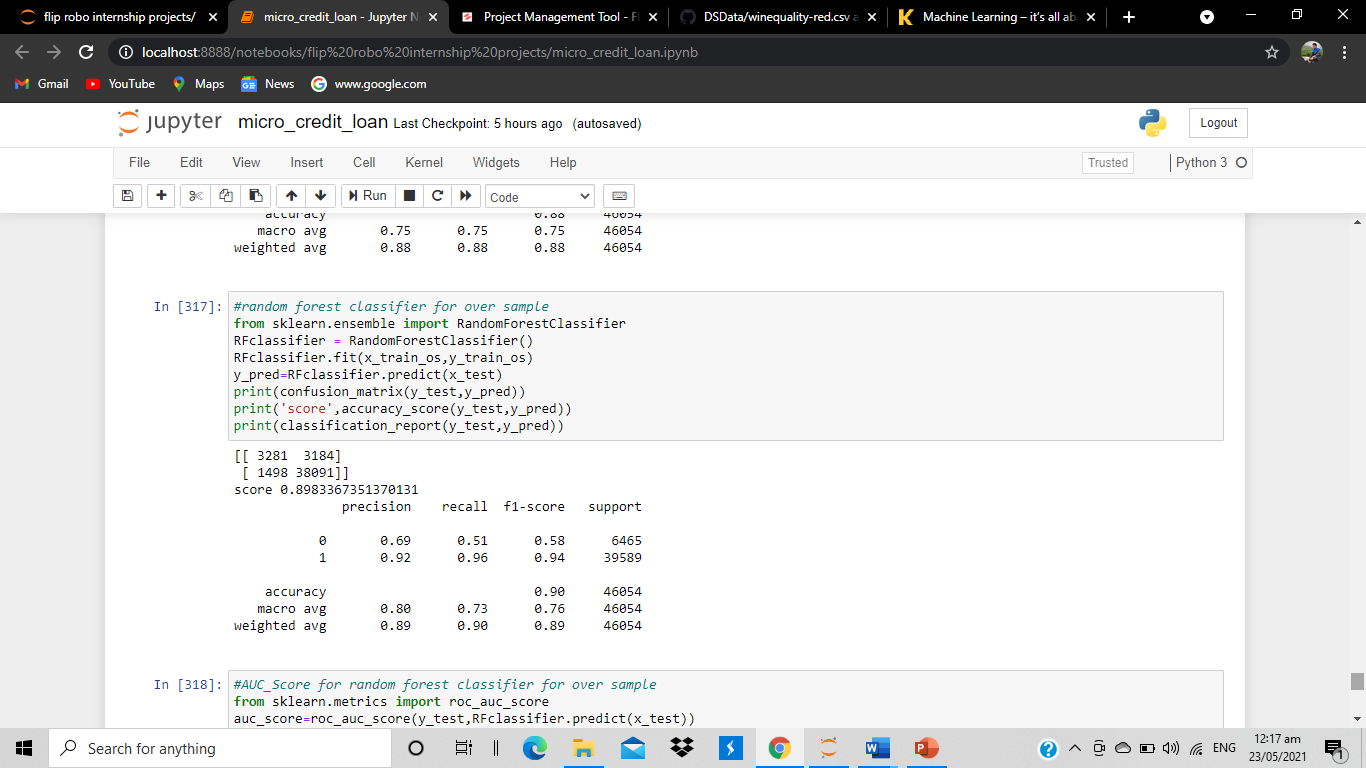
7)extra tree for over sample accuracy got 89%



8)bagging classifier for over sampling accuracy=88%



9)random forest classifier for over sample accuracy 90%



### **Recall and Auc Score is high when I use Random forest Classifier with over sampled data. So I select random forest Classifier model with over sampled data as my final model. And I save my model as a pickle file with the help of joblib.**

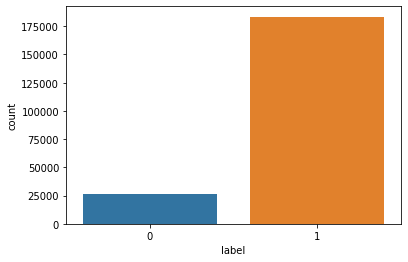
* Key Metrics for success in solving problem under consideration

### Although it is a classification problem so we use accuracy score, classification report and confusion matrix as our evaluation metrics along with AUC ROC score.

### As we know that our dataset is imbalance so we do not much focus upon the accuracy score. We mainly see the precision and recall value of our model.

### Precision talks about all the correct predictions out of total positive predictions. Recall means how many individuals were classified correctly out of all the actual positive individuals.

* Visualizations
* It is the graphical representation of data that is used to check about the presence of outliers, patterns, distribution of the data, etc. There are different data visualisation libraries in python that include matplotlib, seaborn, etc. We will make use of the seaborn and matplotlibtolibrary to visualise the dataset.
* **Checking the count of Defaulter and Non-Defaulter Customers*:***



Label 1 indicates loan has been payed i.e Non-Defaulter and label 0 indicates indicates that the loan has not been payed i.e. defaulter.

**1 160383**

**0 25860**

86% of the sample are in Non defaulter category, and 14% of the sample is in defaulter category. We will take this into account when splitting the data into a training and test set. We also set a seed to make this blog reproducible

**Correlation Matrix:**



1-daily\_decr30 and daily\_decr90 features are highly correlated with each otheer.

2-rental30 and rental90 features are highly correlated with each other.

3-cnt\_loans30 and amount\_loans30 columns are highly correlated with each other.

4-amount\_loans30 is also highly correlated with amount\_loans90 column.

5-medianamnt\_loans30 and medianamnt\_loans90 is highly correlated with each other.

6-We have to drop one of the features which are highly correlated with other feayures. And if we dont do this then our model will face multicolinearity problem.

* Interpretation of the Results
* From the above interpretation we come to know that this is classification based problem so we have learned to build a complete machine learning model for classification based problem. msisdn, aon and maxamnt\_loans90 features are good for prediction level. 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.
* We will also check the most important features according to decision tree. This information can bring insights by applying feature engineering to improve the classification of the target.
* We also visualize the data and see the outcomes of our result that what percentage of customers are in defaulter case, which column is most correlated with target column and much more.

**CONCLUSION**

* Key Findings and Conclusions of the Study
* Key Findings and Conclusions of the Study
* There are no null values in the dataset.
* The dataset is imbalanced. Label ‘1’ has approximately 86% records, while, label ‘0’ has approximately 14% records.
* maxamnt\_loans90 columns gives information about customers with no loan history.
* msisdn features some values which might not be realistic. So drop the row which contain not realistic value
* There are some rows which is repeated means duplicate entries are present in our dataset.
* Sampling data gives the better precision and recall value along with auc score
* The collected data is only for one area circle.
* Learning Outcomes of the Study in respect of Data Science

In this project we learn how to build a machine learning model for classification based problem.

We also learn how to handle the imbalanced dataset for machine learning model. Because when we over sampling and use this over sampled data to build a ML model then it gives the better result.

The goal of any machine learning problem is to find a single model that will best predict our wanted outcome. Rather than making one model and hoping this model is the best/most accurate predictor we can make, ensemble methods take a myriad of models into account, and average those models to produce one final model.

On doing this project the biggest problem I have faced is that I am not able to use GridSearchCV. Because when I use GridSearchCV then my system takes too much time to give the result as our dataset is too large . So If I uses GridSearchCV then our result improves.

So based on all the learning and outcomes our Random Forest Classifier Model with over sampled data gives the best result so we save this model as our final model by using Joblib as a pickle file.

* Limitations of this work and Scope for Future Work
* The dataset is belongs to only one area circle so our model trained for the data of only one area circle and also this data is too old it is from 2016 and contains only three months of data.
* As Today I uses i5 processor computer so I am not able to use GridSearchCV and cross validation because my system takes too much time to give the result. If some how I uses GridSearchCV and Cross validate the model then our model result will be surely improved.
* I would conclude the project report by hoping that now you have understood every step that is required to be done to build a machine learning model. We have built the classification model for classifying the labels that is which customers is in label 1 that is Non defaulter and which one is in label 0 that is defaulter customer and then evaluated it using different error metrics.