**“Micro Credit Defaulter Project”**



Submitted by:

RAHUL AHUJA

**INTRODUCTION**

**Business Problem Framing**

This issue related with the Microfinance Institutions and the other lending companies which are facing the defaulter issue (the people which take loans but don’t return on time) so with the help of the data science the companies want to reduce the credit risk by identifying potential defaulters.

**Conceptual Background of the Domain Problem**

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.

**Motivation for the Problem Undertaken**

The Motivation behind taking this research is

* To understand the logic and criterions and motive basis on which the loans are granted by Microfinance Institutions.
* To develop and come up with a model with the help of data science techniques which can correctly predict the potential defaulter which will help companies to recover their loan easily.

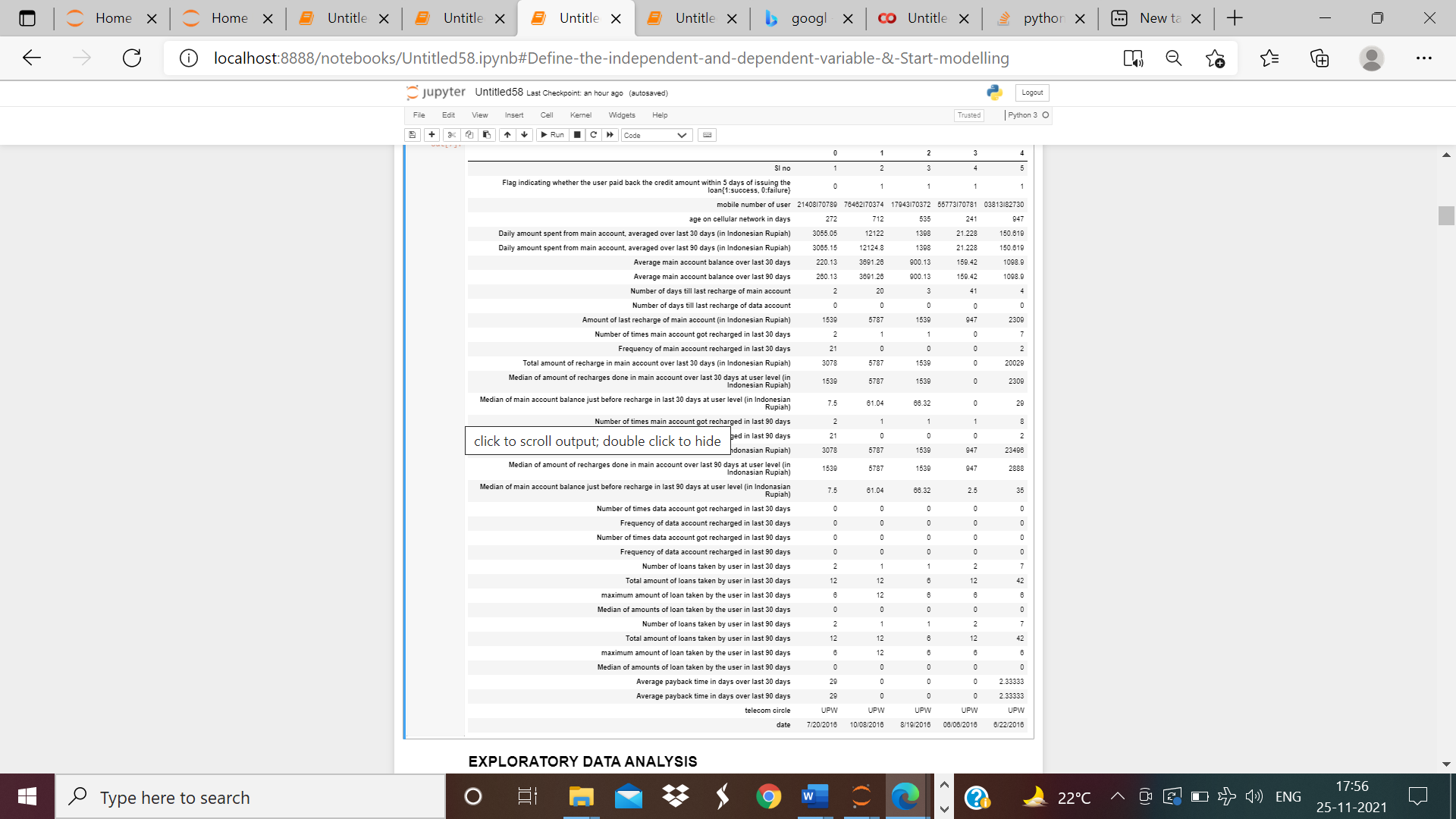
**Analytical Problem Framing**

**Mathematical/ Analytical Modeling of the 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.
* 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 i.e. defaulter.

**Data Sources and their formats**

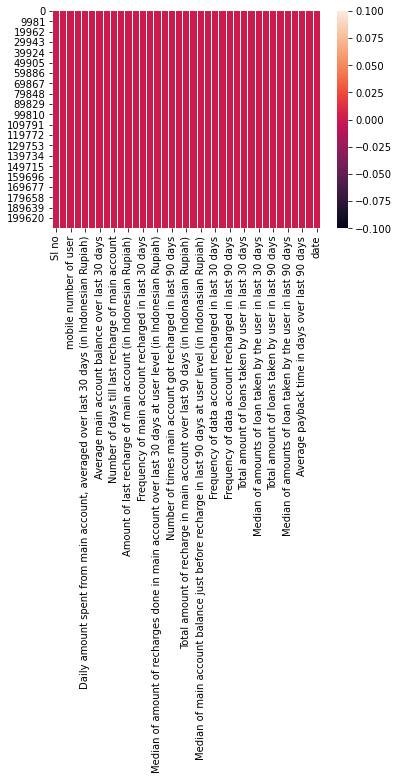
The Data has been shared by client which was available in analysis. The data contains 209593 Rows and 37 Columns which contains the below in formation. The Entire data in in integer and float format except one or two columns which includes (“age on cellular network in days, telecom circle & Date”).



Snapshot of data

**Data Preprocessing Done**

Below are the steps taking for pre-processing and cleaning the data for   
 the analysis.

1. Check the data for the null value. No Null Value found.

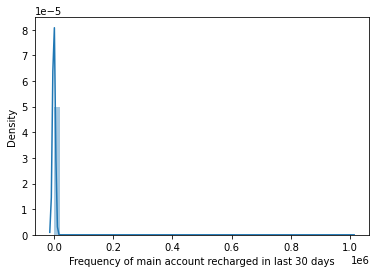
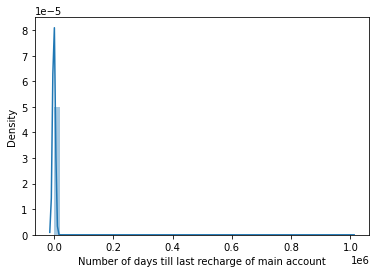
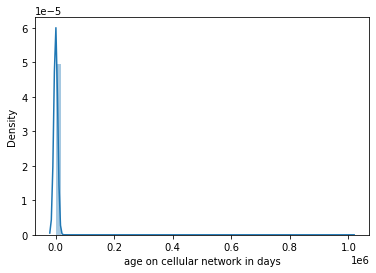
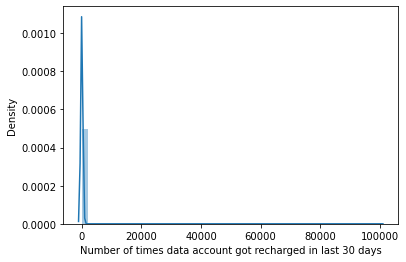
1. We have checked the data thoroughly in which we found that there are lot of outliers present in the data.

For example, if you see column= “Daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah)”. The maximum value 265925 and the 75% value is 7244 Which reflects the presence of outliers. This can be seen in each and every column

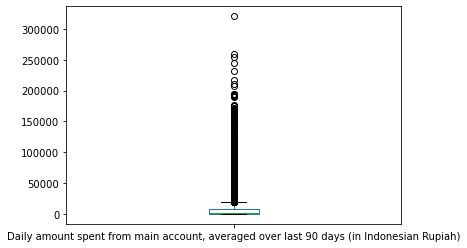
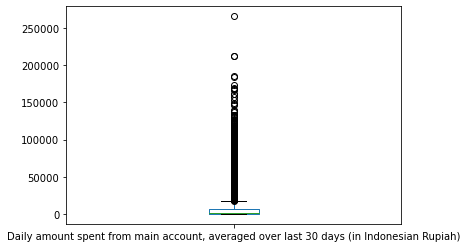
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Flag indicating whether the user paid back the credit amount within 5 days of issuing the loan{1:success, 0:failure}** | **age on cellular network in days** | **Daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah)** | **Daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah)** | **Average main account balance over last 30 days** | **Average main account balance over last 90 days** | **Number of days till last recharge of main account** | **Number of days till last recharge of data account** |
| count | 209593 | 209593 | 209593 | 209593 | 209593 | 209593 | 209593 | 209593 |
| mean | 0.875177 | 8112.343 | 5381.402 | 6082.515 | 2692.582 | 3483.407 | 3755.848 | 3712.203 |
| std | 0.330519 | 75696.08 | 9220.623 | 10918.81 | 4308.587 | 5770.461 | 53905.89 | 53374.83 |
| min | 0 | -48 | -93.0127 | -93.0127 | -23737.1 | -24720.6 | -29 | -29 |
| 25% | 1 | 246 | 42.44 | 42.692 | 280.42 | 300.26 | 1 | 0 |
| 50% | 1 | 527 | 1469.176 | 1500 | 1083.57 | 1334 | 3 | 0 |
| 75% | 1 | 982 | 7244 | 7802.79 | 3356.94 | 4201.79 | 7 | 0 |
| max | 1 | 999860.8 | 265926 | 320630 | 198926.1 | 200148.1 | 998650.4 | 999171.8 |

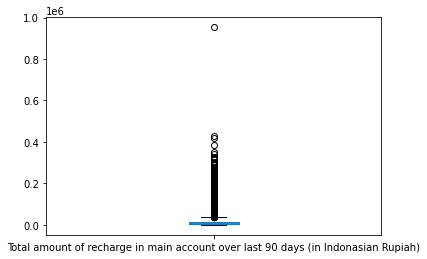
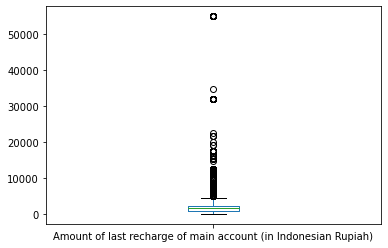
The above table describing the data where we can see the difference between the maximum value and the 75% value which indicates the presence of the outliers. Also we can see the difference between the mean and the median (50%) means that the data is skewed (right skewed).

1. In each column there is a difference between the median and the mean which reflects the presence of skewness of the data. Since the mean is on a higher side, all the columns are rightly skewed.



1. Removal of the Unnecessary columns and converting the date column into day and month.
2. Converting the negative data into zero.
3. Removing of the skewness.
4. Use of label encoder to convert the object data ("mobile number of user") in integer.
5. Removal of the outliers. We will use the rule of 6 sigma instead of 3 sigma because we are not suppose to loose more than 8% of important data.





1. Used Min-Max scalar for standardization of the data
2. Use SMOTE for the balancing of the imbalance data

Data Inputs- Logic- Output Relationships

The data has been categorized into two and that has been given x and y. y data contains the first columns (Flag indicating whether the user paid back the credit amount within 5 days of issuing the loan {1: success (Non-Defaulter, 0:failure(defaulter)}) considering that a person which took the loan has return the loan amount or nots. The purpose is to check and develop a model which can identify that the person will be a defaulter or not if the loan will be provided to him. x data contains the complete data except the first column. We will develop a model and predict the data which we will match the predictions with actual data to verify the accuracy.

**Model/s Development and Evaluation**

Identification of possible problem-solving approaches (methods)

Data Exploration and Cleaning On data exploration, I found that the dataset was imbalanced for the target features (87.5% for non-defaulters and 12.5% for Defaulters). Also, I found that the data had some very unrealistic values such as 999860 days which is not possible. Also, there were negative values for variables which must not have one (example: frequency, amount of recharge etc). We have used the six-sigma method to remove the outliers and the unrealistic values to clean the refine the data. Also, we have converted the non-negative values into zero

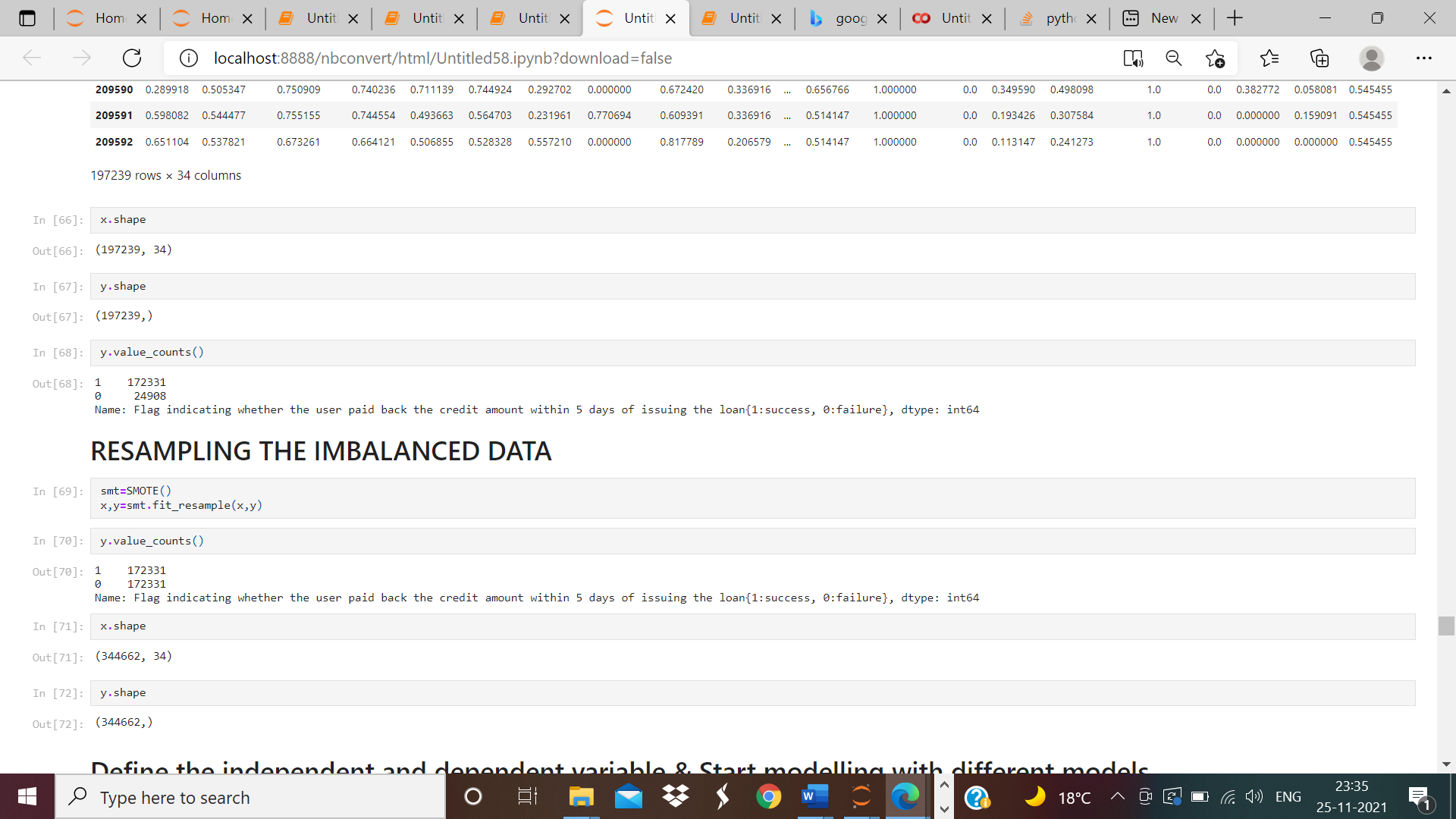
Data Visualization On visualizing data, there were two important insights I gathered.

1. Imbalance of data
2. Distribution was not normal.

The skewed data has been un-skewed with the help of log.1p method.

Data Normalization Since the data was not normal, I normalized all the features except the target variable which was dichotomous (Values '1' and '0').

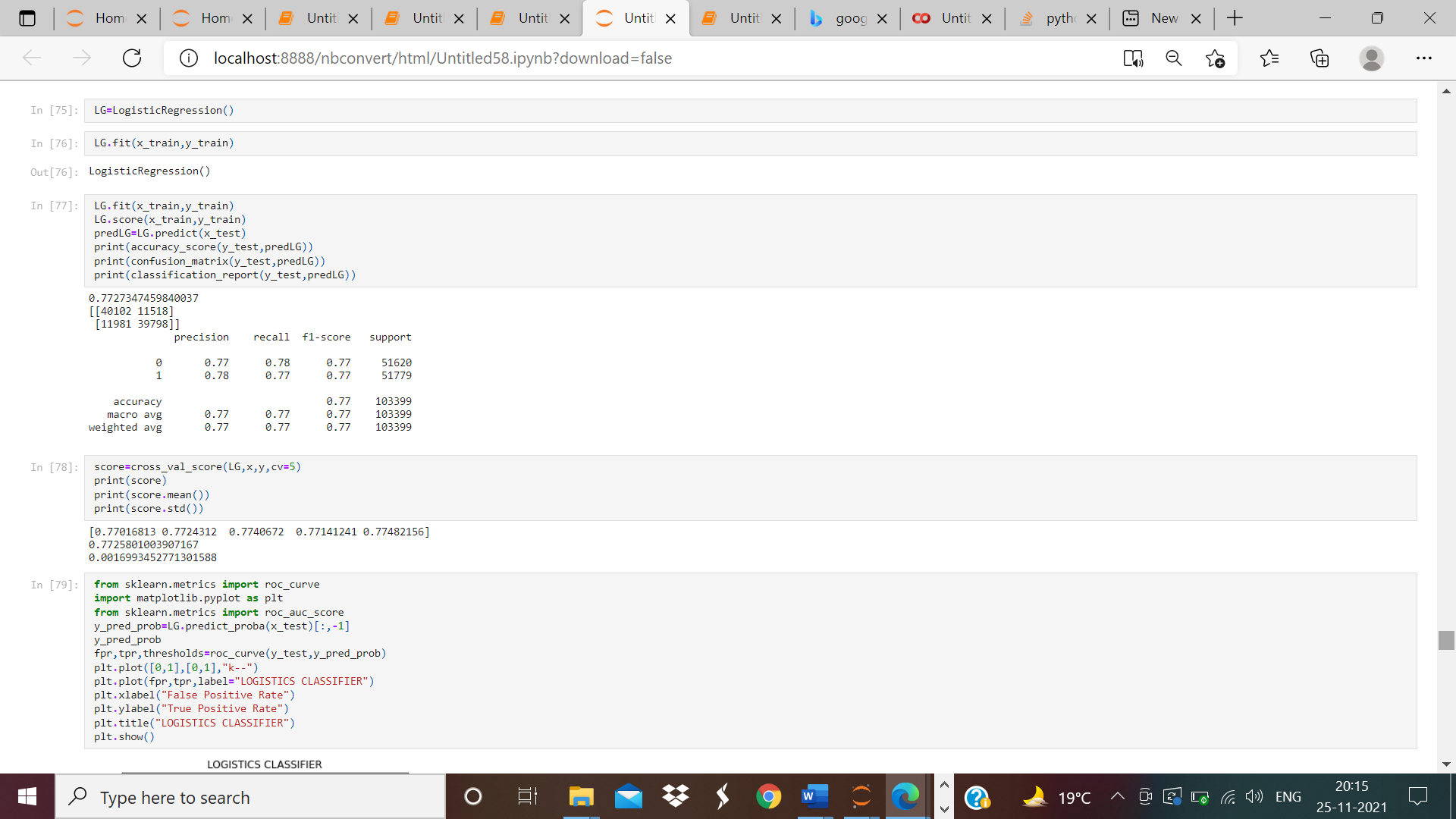
To over-come the imbalance issue in the output column and giving the model a best dataset to teach we have used SMOTE function.

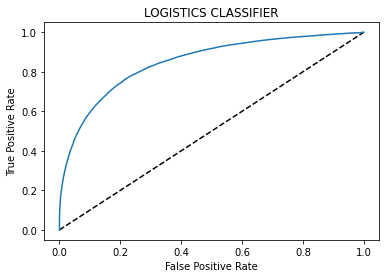


**Run and evaluate selected models**

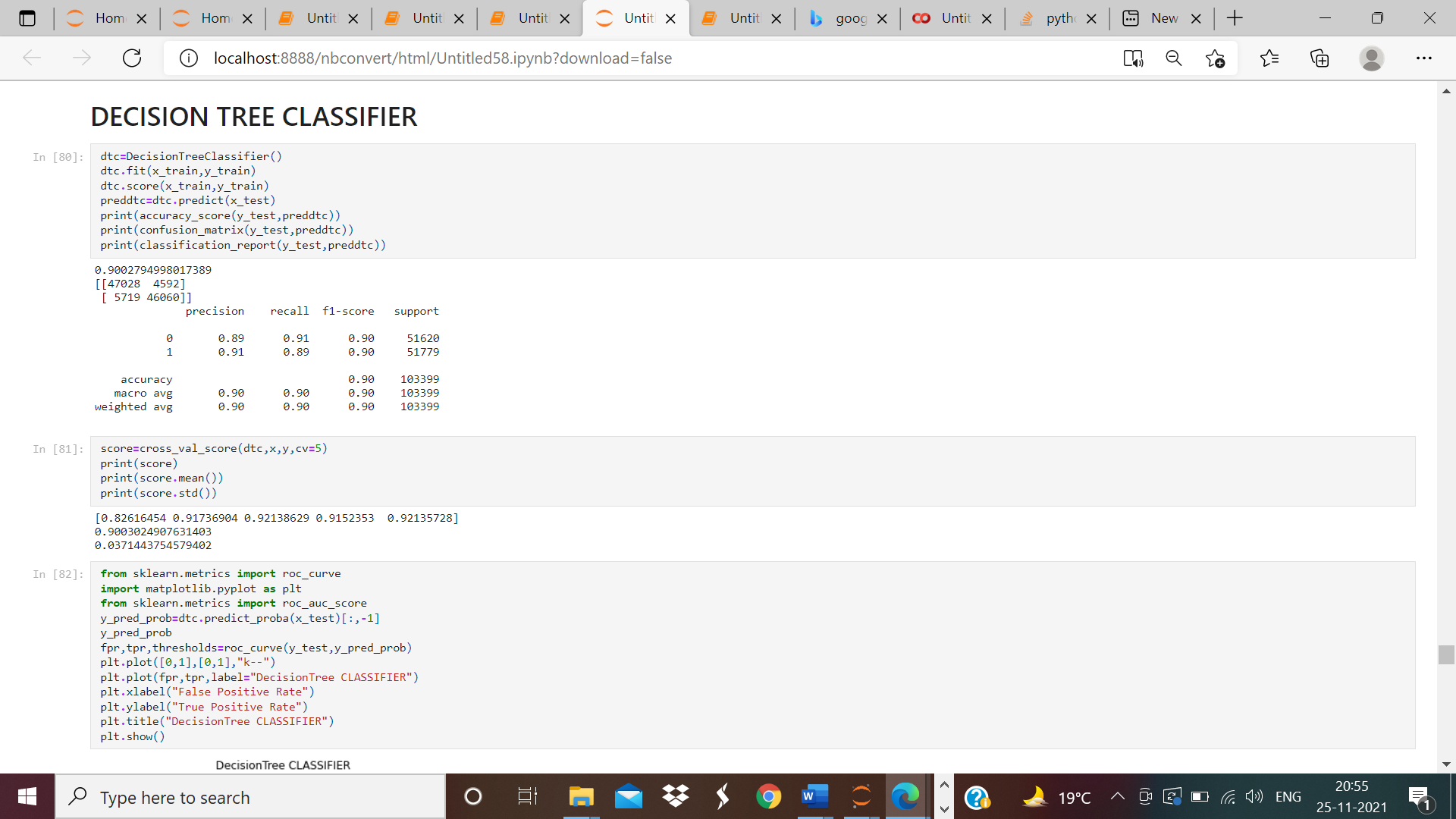
Build Models Since it was a supervised classification problem, I built 5 models to evaluate performance of each of them: a. Logistic Regression b. Decision Tree c. Random Forest d. Ada Boost Classifier e. Gradient Boost Classifier Since the data was imbalanced, accuracy was not the correct performance metric. Instead, I focused on other metrics like precision, recall and ROC-AUC curve.

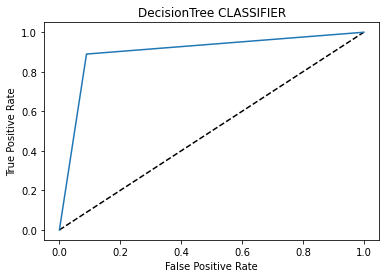
**Logistics Regression:-**





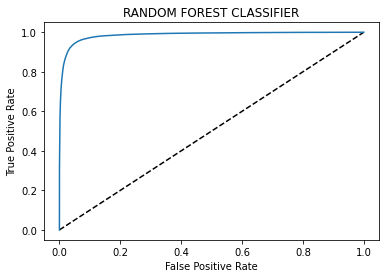
**DECISION TREE CLASSIFIER**



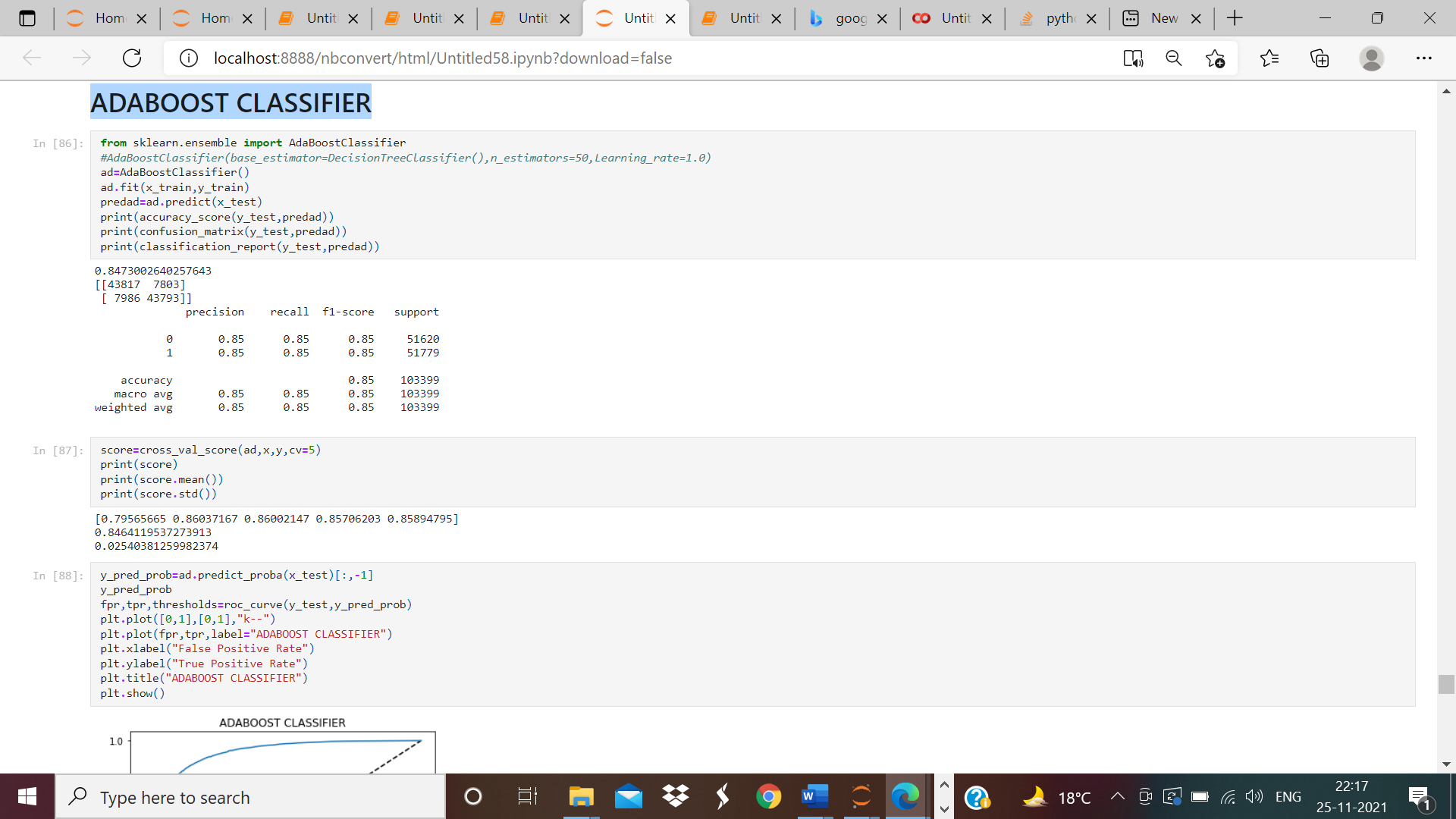


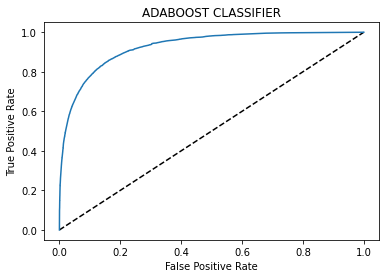
# RANDOM FOREST CLASSIFIER

# 



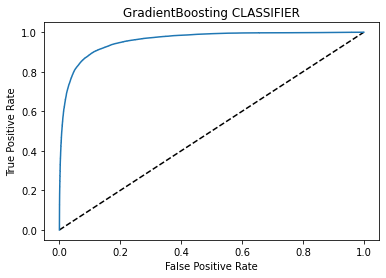
# ADABOOST CLASSIFIER





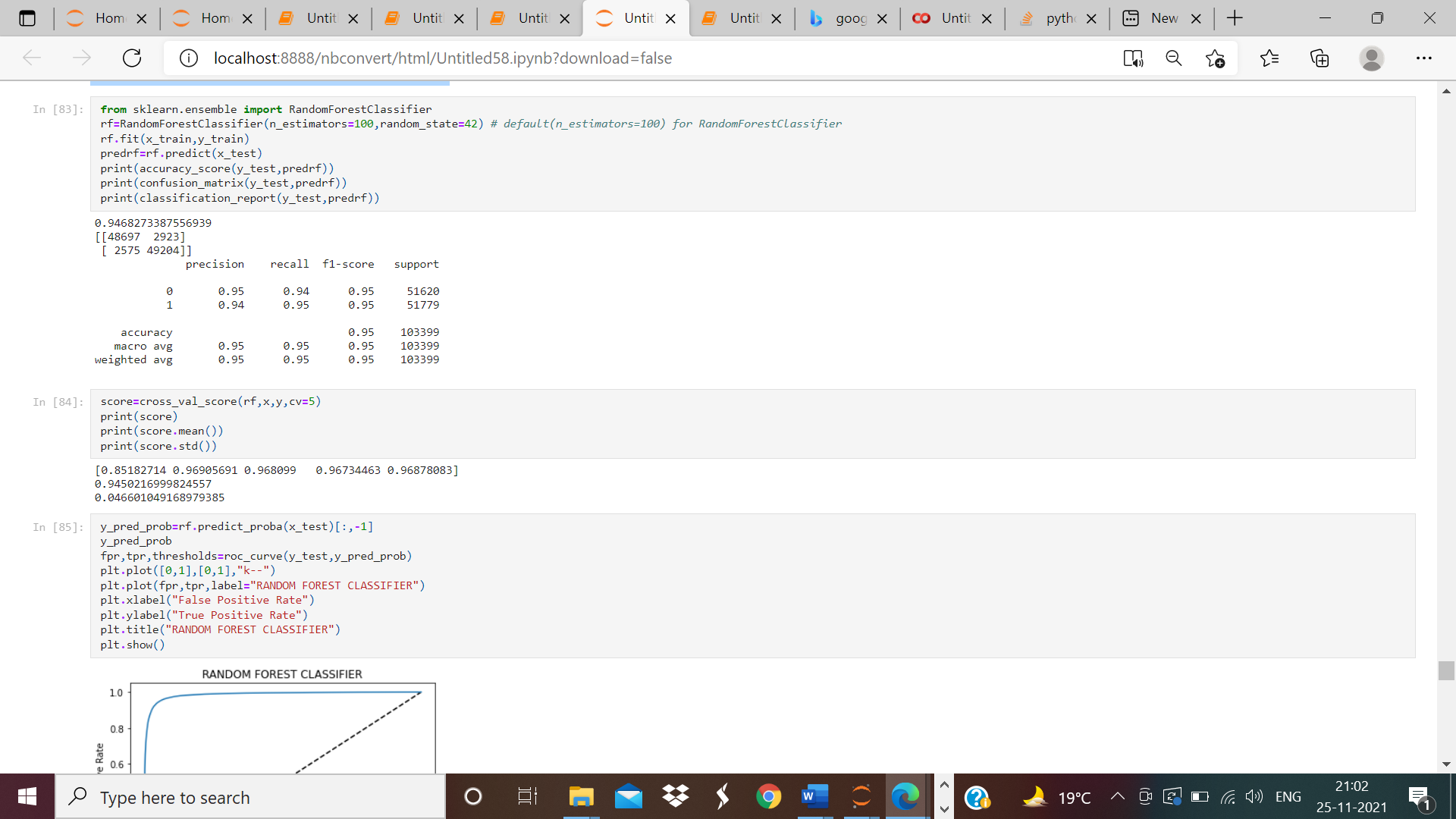
# GRADIENT BOOSTING CLASSIFIER

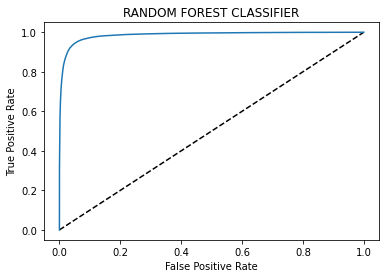
# 



**CONCLUSION**

According to the performance metrics, Random Forest scores highest in accuracy. Also, the curve is tending towards the ideal shape. Below is the accuracy score and the Aur-Roc curve image. Below image clearly shows that random forest classifier model provides 95% accuracy score and 95% validation score. So we can see that this is the best fit for this model





MODEL PREDICTIONS

