

Micro Credit Defaulter Analysis

Submitted by:

Priyanka Yadav

**ACKNOWLEDGMENT**

I would like to thank FlipRobo Technologies to provide me this valuable data in order to perform this analysis. I would also like to thank my mentor Swati Mahaseth for helping me throughout this project.

**INTRODUCTION**

* Business Problem Framing

The prediction for defaulters and non-defaulters can change the landscape of banking sector as banks can efficiently predict and choose their customers. They can also predict the amount of days they will be needing in order to return the total amount of loan

* Conceptual Background of the Domain Problem

Here by using best machine learning practices we can predict that if a person can return the loan taken by the given institution in a given timeframe (5 days here). And the probability of the repayment of the loan if taken again in future.

* Review of Literature

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.

By the means of this reports, I will be summarizing and explaining how the analysis too place and what are its significance and use case.

* Motivation for the Problem Undertaken

The motivation that driven me towards the timeline of this project was the sheer size and complexity of this dataset. I have tried my best to follow best practices of machine learning throughout the project and follow the steps that doesn’t only predict the best results but also tried to keep the process simple to understand.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

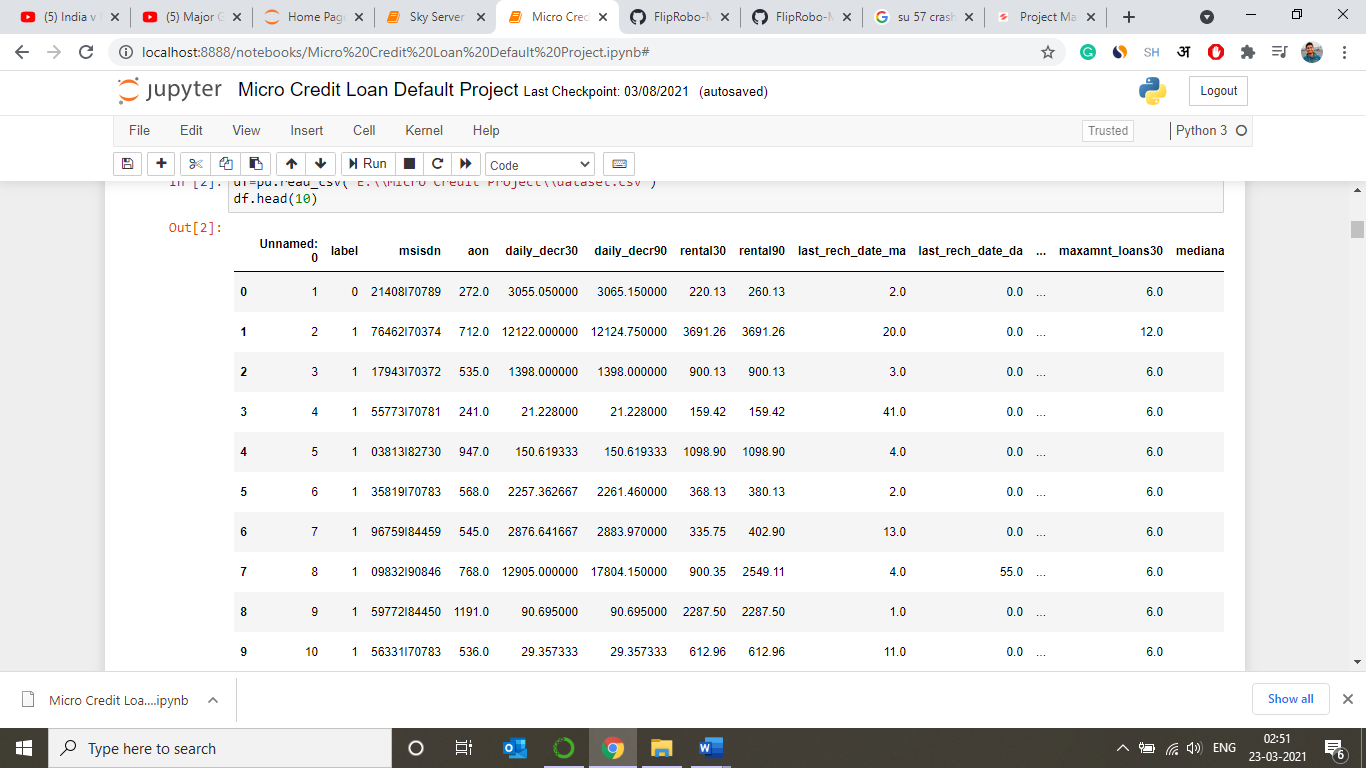
Mathematical Tools Used: Mean, Average, IQR, Standard deviation and Median for gaining insights of the dataset.

Analytical Tools Used: Correlation and Skewness for finding the relationships of dependent and independent variable and checking the distribution of data.

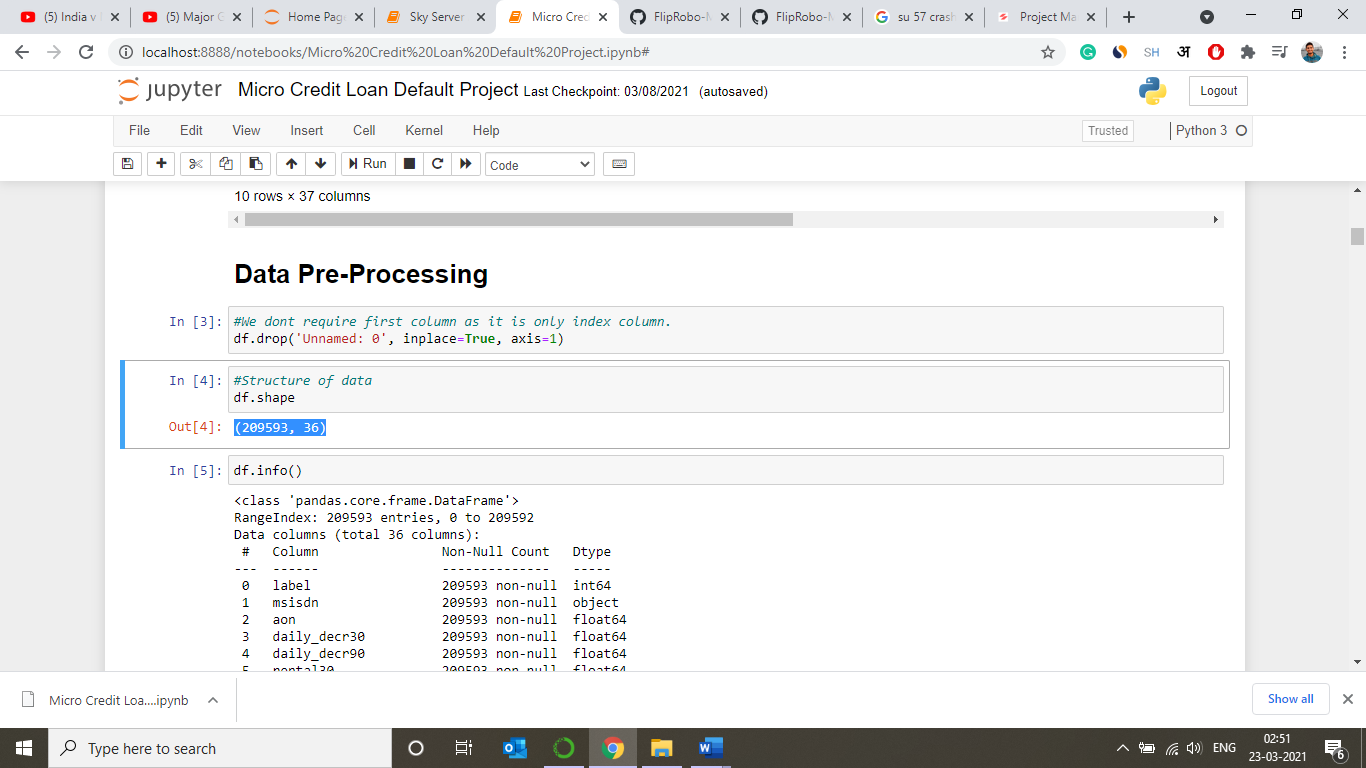
Packages Used: scikit-learn, pandas, seaborn etc.

* Data Sources and their formats

Data was been provided to me by FlipRoboTechnologies. Whereas originally data belongs to an Indonesian Telecom Company.The data contained 34 columns and 209583 entries which are shown below:



Figure



Figure

* Data Preprocessing Done
* For keeping the only the useful data we used feature engineering. We removed the data if uneccesary by qualifying them on certain conditions like their uniqueness, their correlation with target variable and the no of outliers present in that particular variable.
* We then tried the change the data into the form they belong to either by changing their type or by bifurcating the complex variables.
* We also tried to treat the outliers and skewness present in the data.
* We also scaled the data so that the data is standardized.
* Data Inputs- Logic- Output Relationships

After data pre-processing was completed, we were left with 34 columns excluding aon, label, msisdn, pcircle&pdate which were observed as not significant to our final results.

Other than label column all the columns were taken as input variables whereas the label column deciding the whether the person will be a defaulter or not was taken as the target variable.

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

It was pretty clear that machine learning will be used to predict the results for the dataset.

* Hardware and Software Requirements and Tools Used

Hardware Requirements: Laptop or PC with 4GB RAM or plus having i3 or above processor.

Software Requirements: Python and Jupyter Notebook.

Jupyter Tools/ Packages:pandas, numpy, seaborn,matplotlib.pyplot, import warnings, sklearn.preprocessing(Label Encoder & Standard Scaler), sklearn.model\_selection, sklearn.linear\_model etc.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)
* For that, we used various statistical tools like median, correlation, data skewness, standard deviation for data analysis.
* We also used scikit-learnand other numerous analytical tools for achieving the target.
* Testing of Identified Approaches (Algorithms)

In this analysis, we have used three algorithms and three ensemble techniques to boost the predictive scores for better results. With the help of this approach, we will be able to successfully micro credit loan repayment predictions.

ALGORITHM USED:

* Logistic Regression
* Decision Tree Classifier
* GaussianNB

ENSEMBLE TECHNIQUE USED:

* Gradient Boosting Classifier
* AdaBoost Classifier
* Extra Trees Classifier

We ran these algorithms in order to find best fitting model and best accuracy and other different parameters. The working of these algorithms are defined below:

* **Logistic Regression**

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

* **Decision Tree Classifier**

Decision tree analysis is a predictive modelling tool that can be applied across many areas. Decision trees can be constructed by an algorithmic approach that can split the dataset in different ways based on different conditions. Decisions trees are the most powerful algorithms that falls under the category of supervised algorithms.

Classification decision trees − In this kind of decision trees, the decision variable is categorical. The above decision tree is an example of classification decision tree.

* **Gaussian NB**

Naïve Bayes algorithms is a classification technique based on applying Bayes’ theorem with a strong assumption that all the predictors are independent to each other. In simple words, the assumption is that the presence of a feature in a class is independent to the presence of any other feature in the same class.

Gaussian NB is the simplest Naïve Bayes classifier having the assumption that the data from each label is drawn from a simple Gaussian distribution.

* **Gradient Boosting Classifier**

[Gradient boosting classifiers](https://en.wikipedia.org/wiki/Gradient_boosting) are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting.

* **AdaBoost Classifier**

It combines multiple classifiers to increase the accuracy of classifiers. AdaBoost is an iterative ensemble method. AdaBoost classifier builds a strong classifier by combining multiple poorly performing classifiers so that you will get high accuracy strong classifier. The basic concept behind Adaboost is to set the weights of classifiers and training the data sample in each iteration such that it ensures the accurate predictions of unusual observations.

* **Extra Trees Classifier**

**Extremely Randomized Trees Classifier(Extra Trees Classifier)** is a type of ensemble learning technique which aggregates the results of multiple de-correlated decision trees collected in a “forest” to output it’s classification result. In concept, it is very similar to a Random Forest Classifier and only differs from it in the manner of construction of the decision trees in the forest.

* **Run and Evaluate selected models**

We have tried to fit different algorithms on our dataset and their description and snapshots are pasted below:

* + Applying different algorithms (Code)

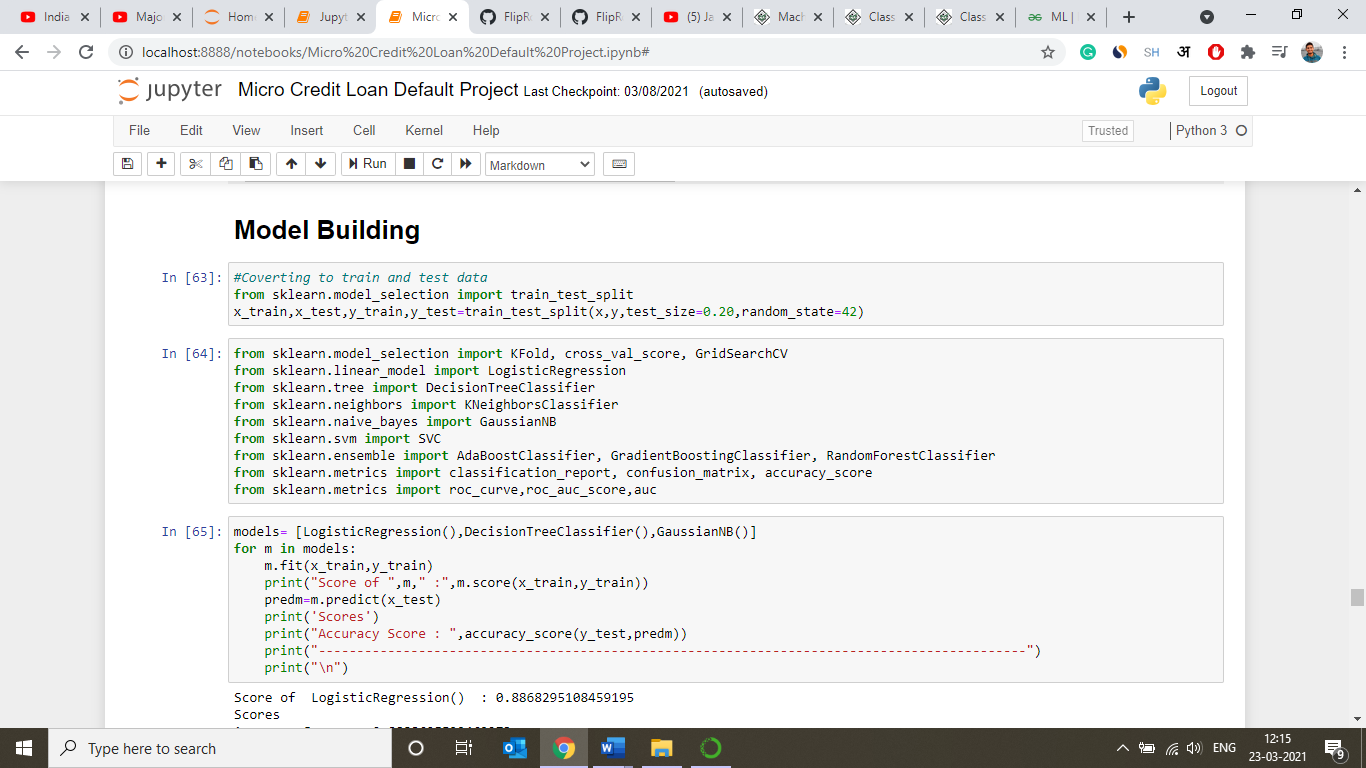


Figure 3: Using Different Models

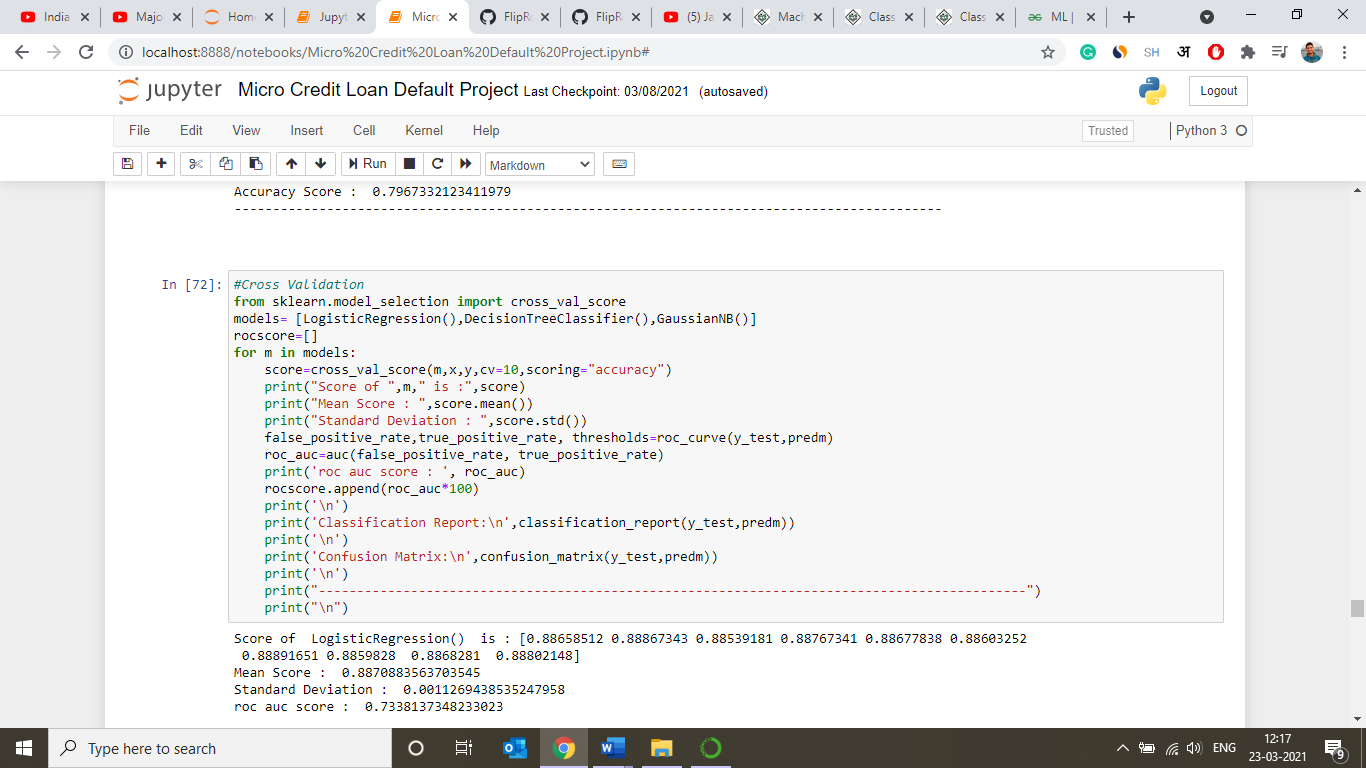


Figure 4: Cross Validating all the models

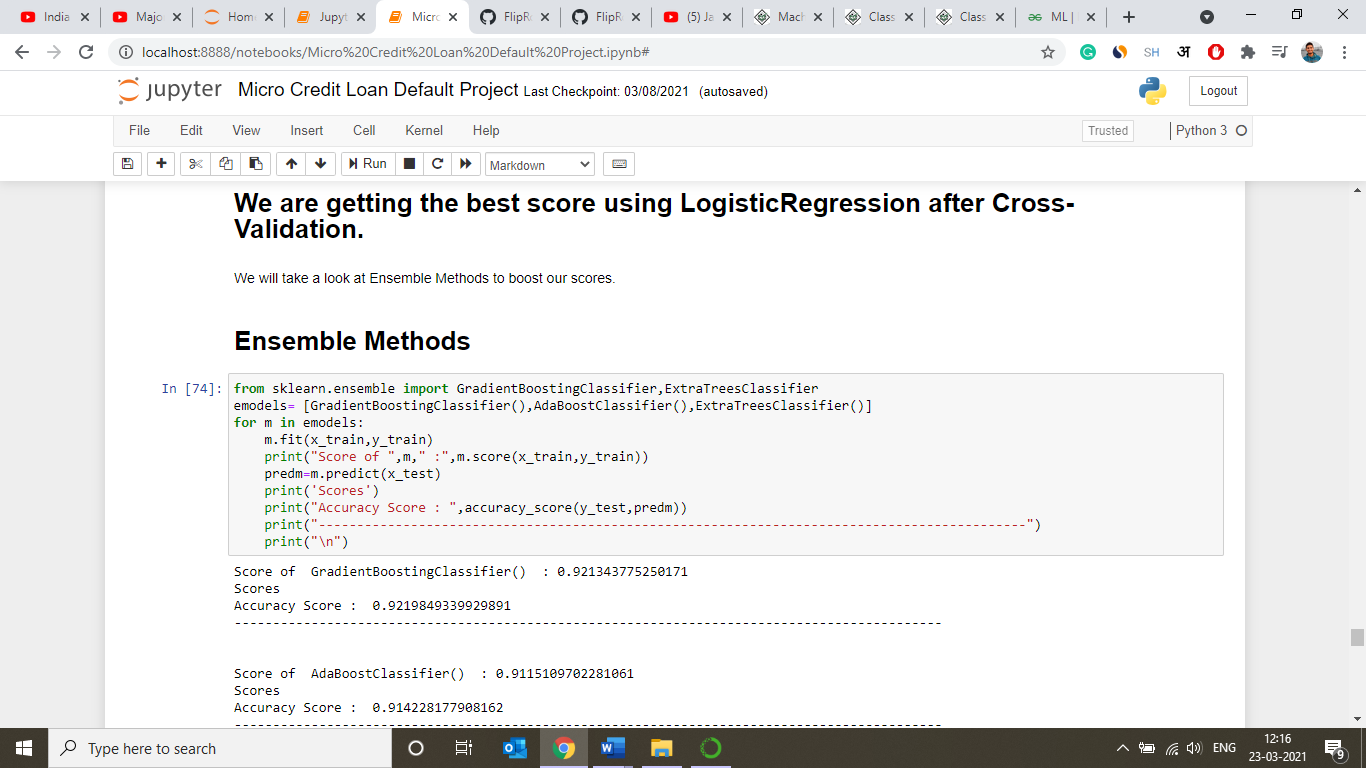


Figure 5: Using Ensemble Methods to boost the scores

* + Output (Scores)

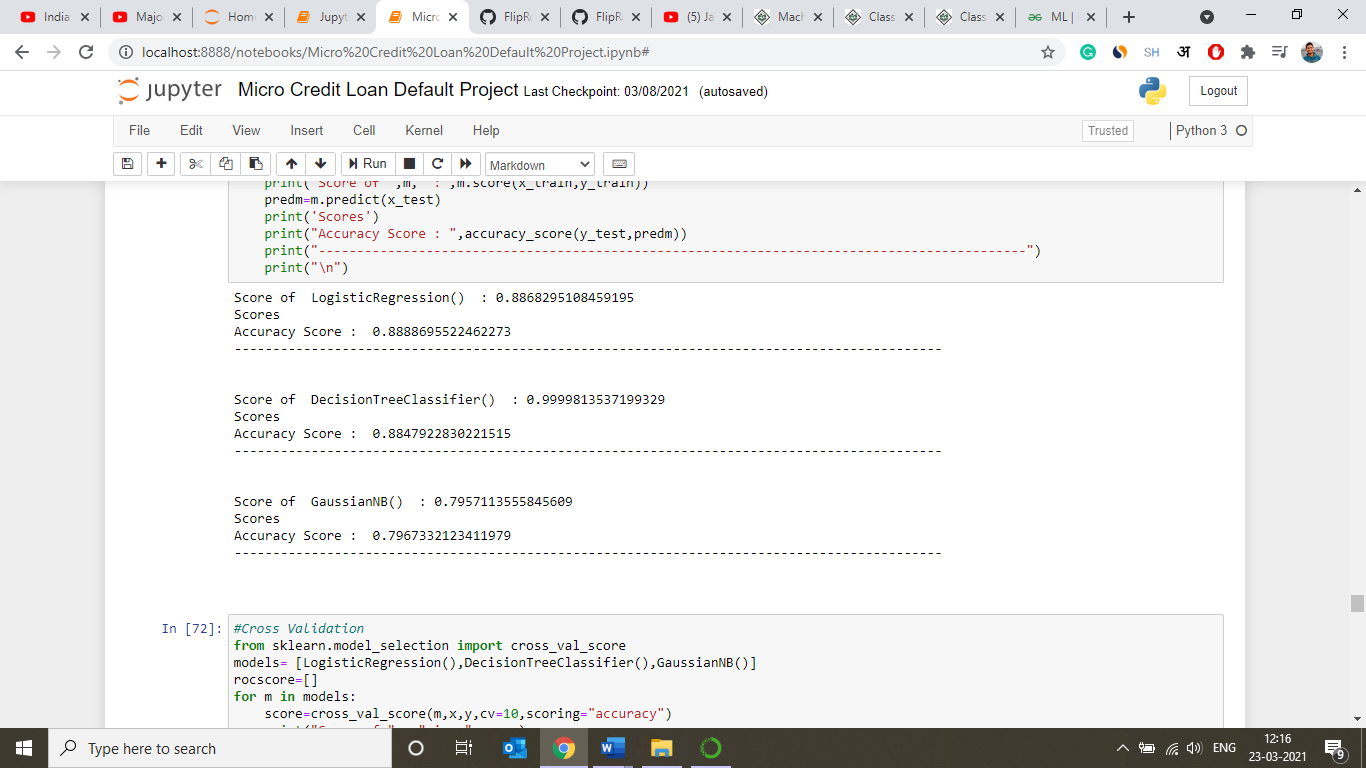


Figure 6: Accuracy Score of diffrent models

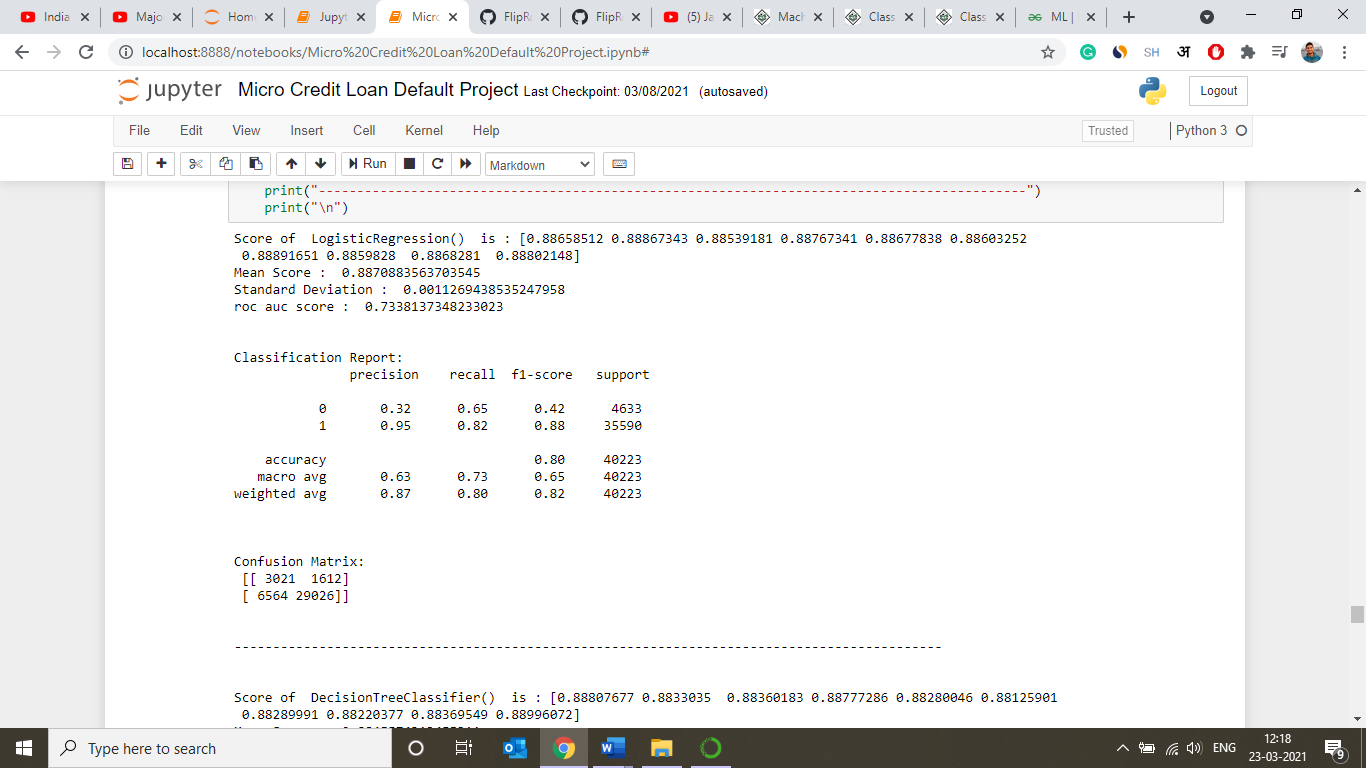


Figure 7: Cross Val Score for Logistic Regression

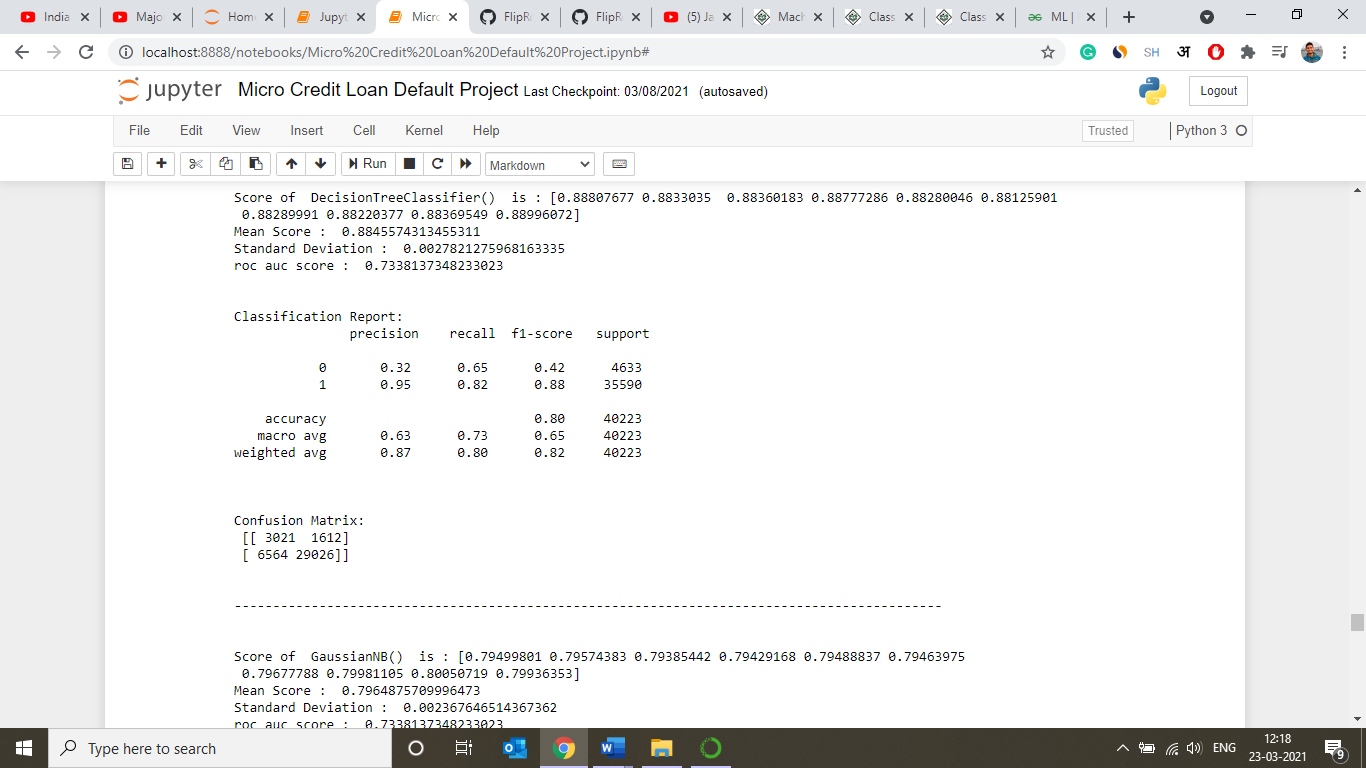


Figure 8:Cross Val Score for Decision Tree Classifier

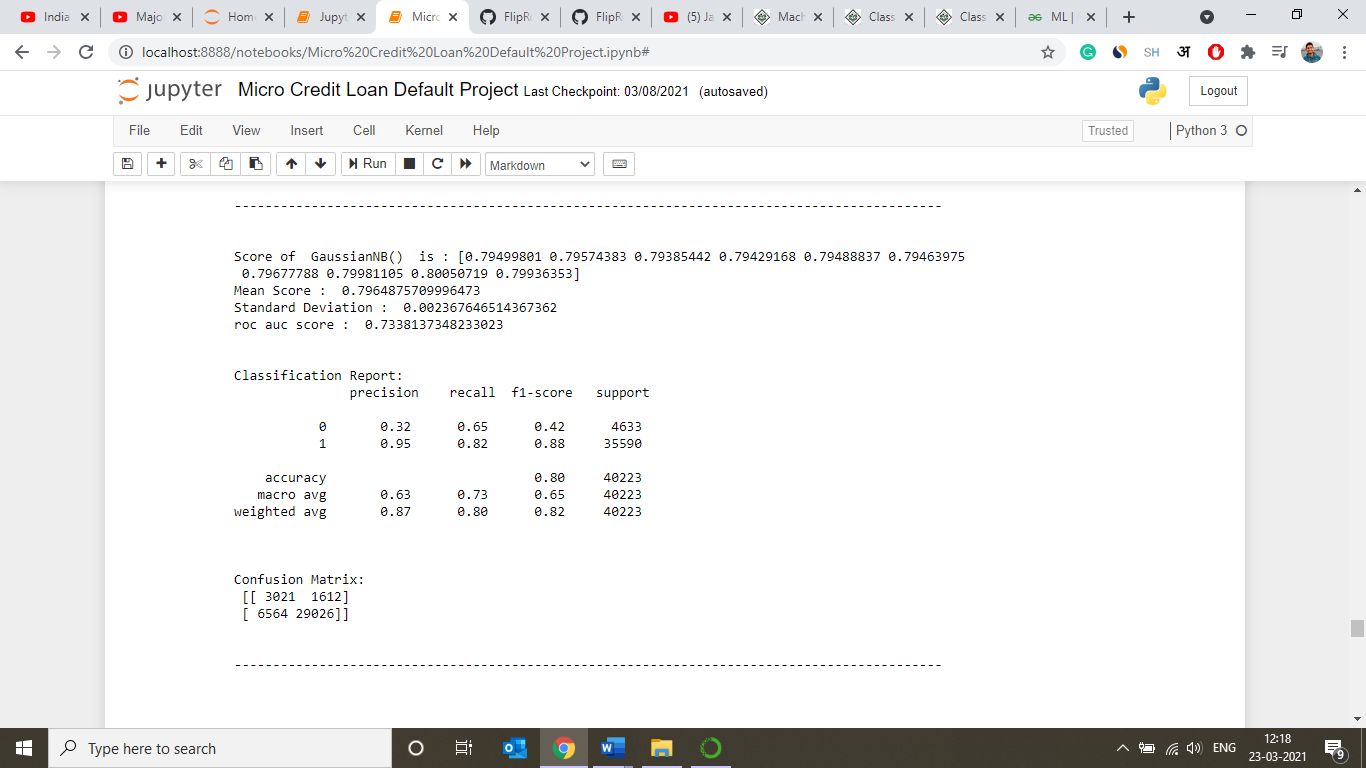


Figure 9: Cross Val Score for Gaussian NB Classifier

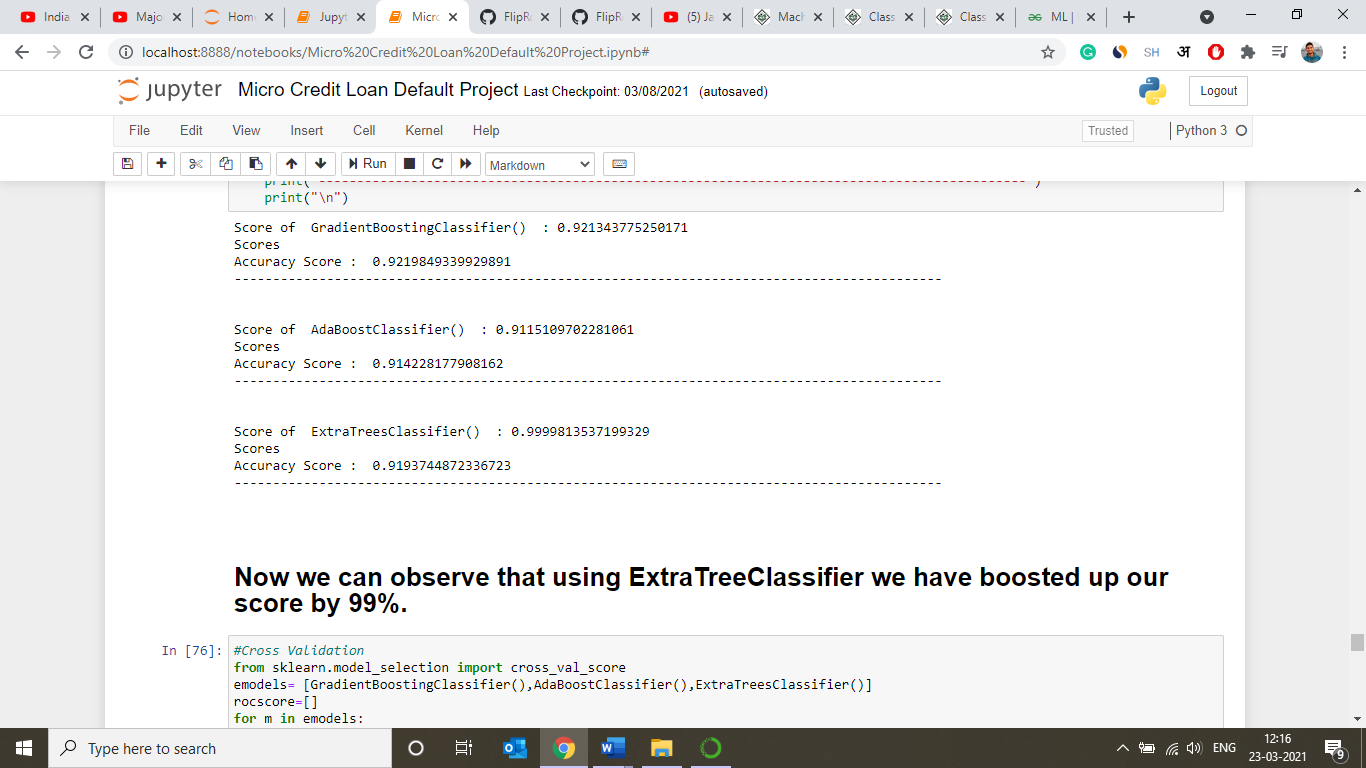


Figure 10: Scores using Ensemble Techniques

* **Key Metrics for success in solving problem under consideration**

When building and optimizing your classification model, measuring how accurately it predicts your expected outcome is crucial. However, this metric alone is never the entire story, as it can still offer misleading results. That's where these additional performance evaluations come into play to help tease out more meaning from your model.

The various metrics used to evaluate our model used in this analysis are:

* **Confusion Matrix**

Evaluation of the performance of a classification model is based on the counts of test records correctly and incorrectly predicted by the model. The confusion matrix provides a more insightful picture which is not only the performance of a predictive model, but also which classes are being predicted correctly and incorrectly, and what type of errors are being made.

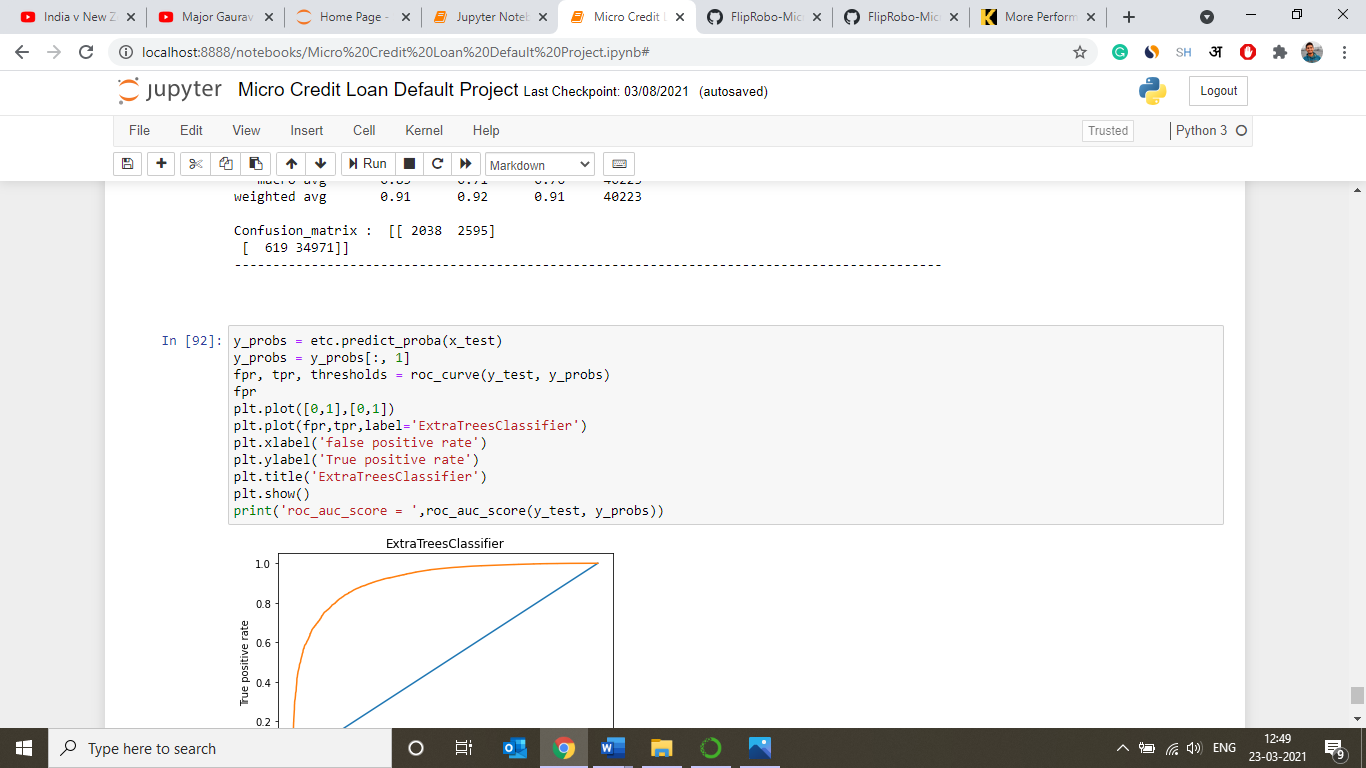
****

Figure 11: Final Confusion Matrix

* **Accuracy Score**

Accuracy is one metric for evaluating classification models. Informally, **accuracy** is the fraction of predictions our model got right.

* **Precision**

It is the ratio of True Positives to all the positives predicted by the model.

* **Recall**

It is the ratio of True Positives to all the positives in your Dataset.

* **F1 Score**

F-Measure provides a single score that balances both the concerns of precision and recall in one number. A good F1 score means that you have low false positives and low false negatives, so you’re correctly identifying real threats, and you are not disturbed by false alarms.

* **Roc\_auc Score**

ROC is a major visualization technique for presenting the performance of a classification model. It summarizes the trade-off between the true positive rate (tpr) and false positive rate (fpr) for a predictive model using different probability thresholds.

A ROC curve plots the true positive rate (tpr) versus the false positive rate (fpr) as a function of the model’s threshold for classifying a positive.

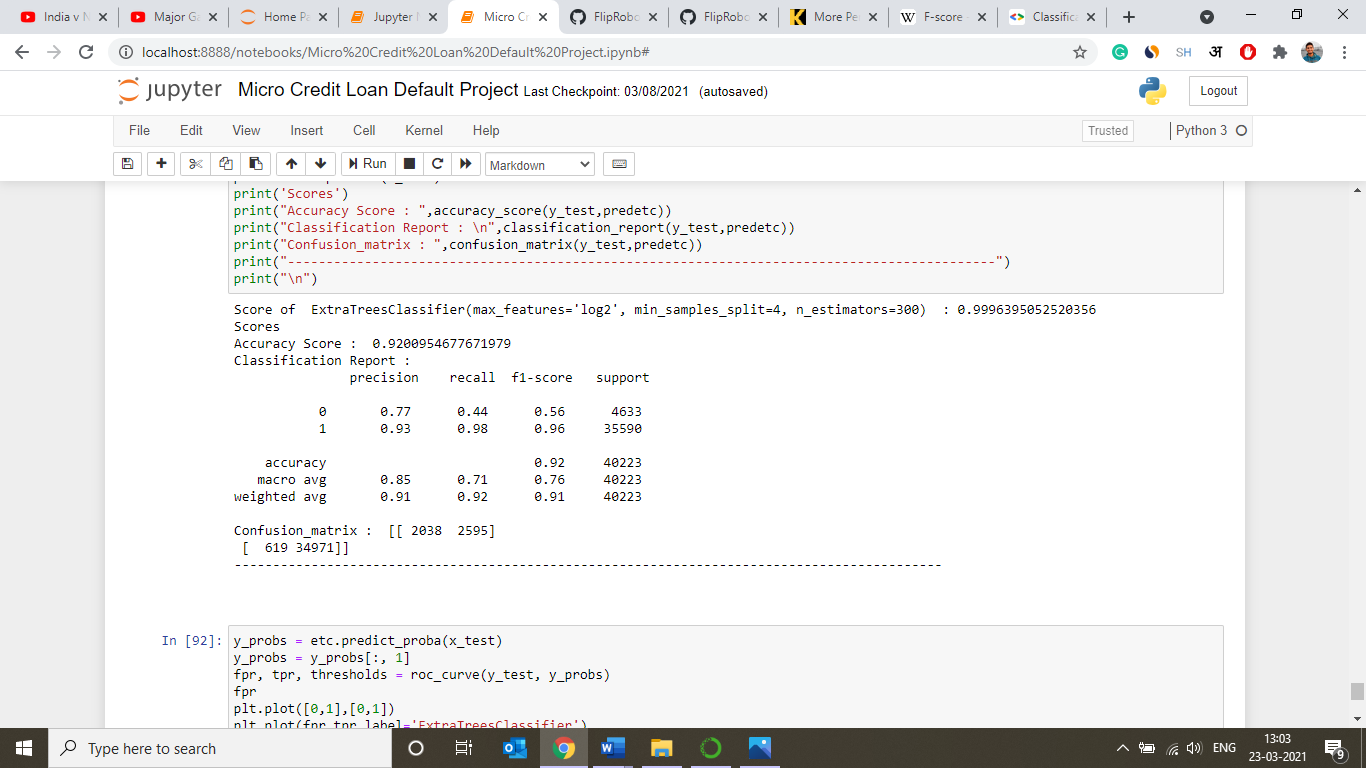


Figure 12: Accuracy, Precision, Recall, f1 score

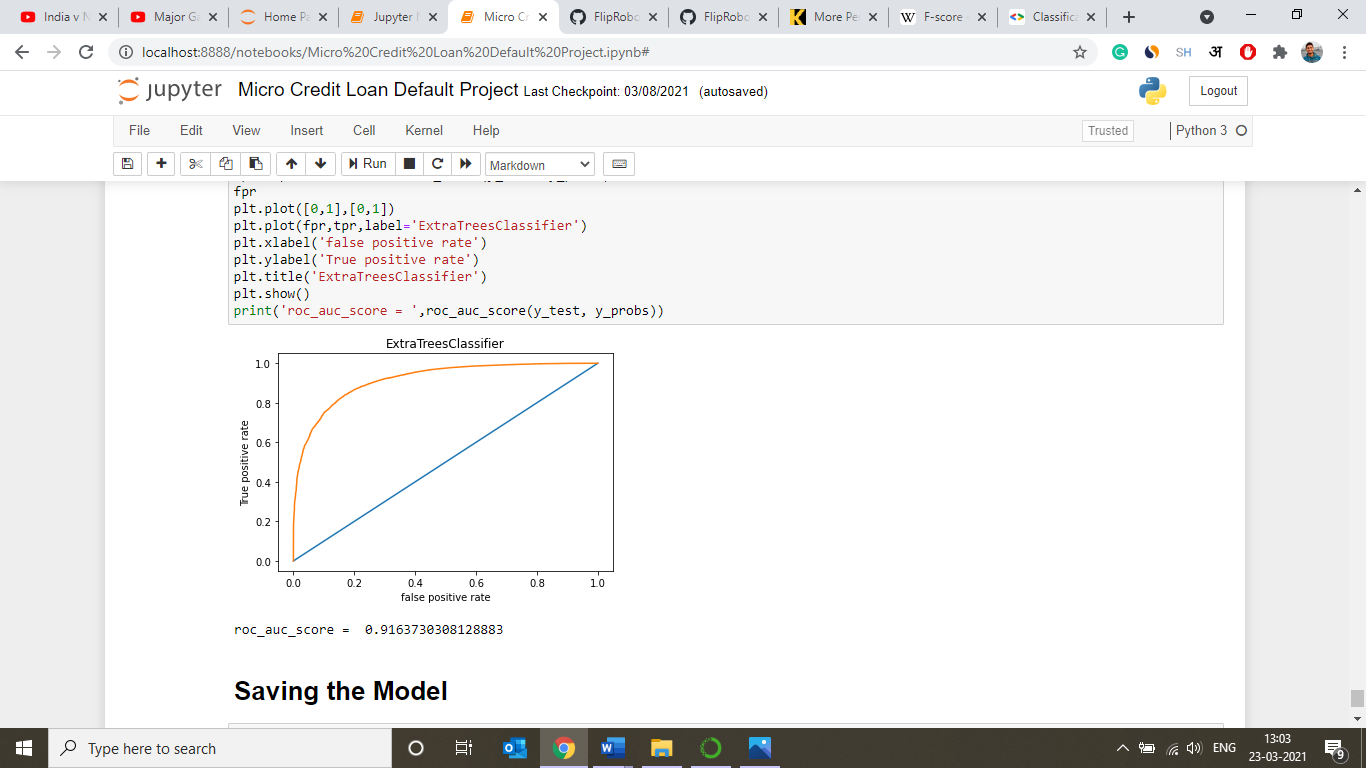


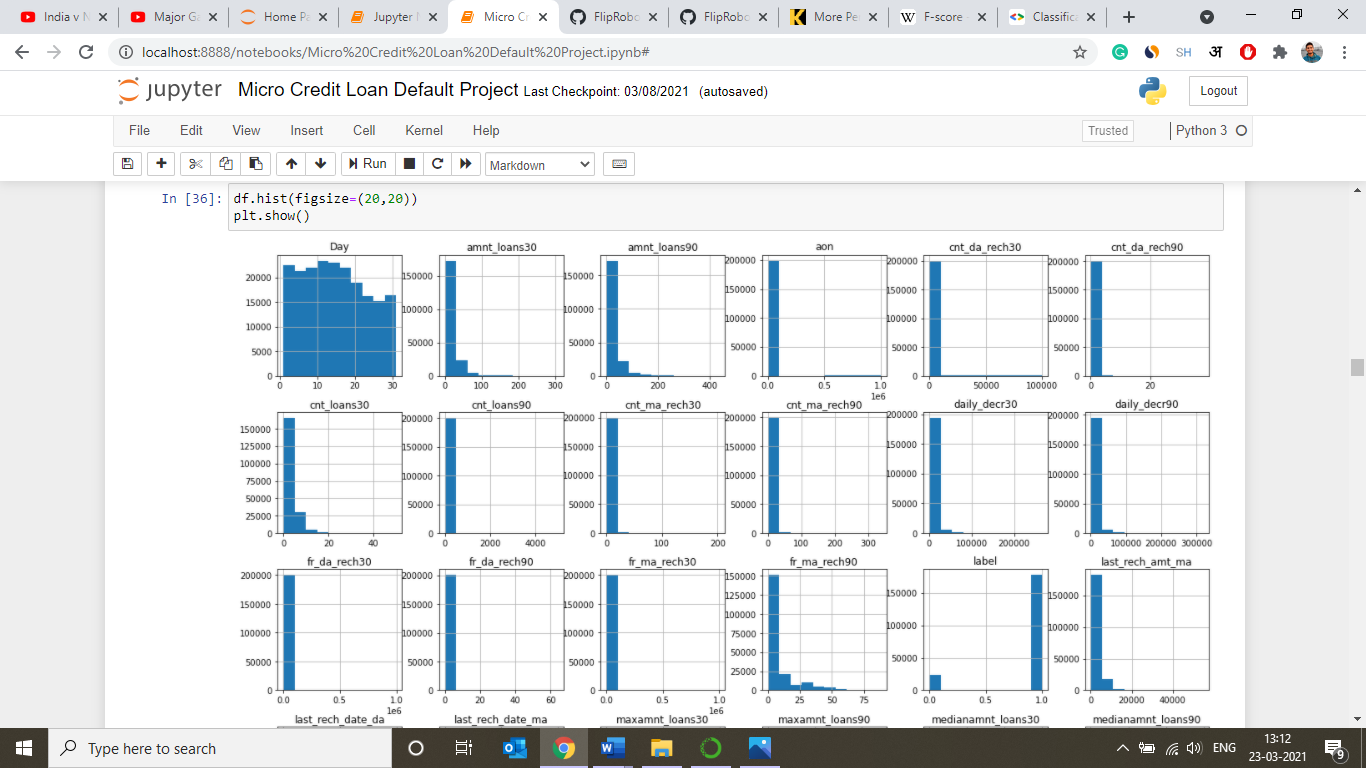
Figure 13: auc\_roc score

* **Visualizations**

In this analysis, we have used numerous number of graphs for getting insights of the data that was given to us so that we can interpret the data to get desired results.

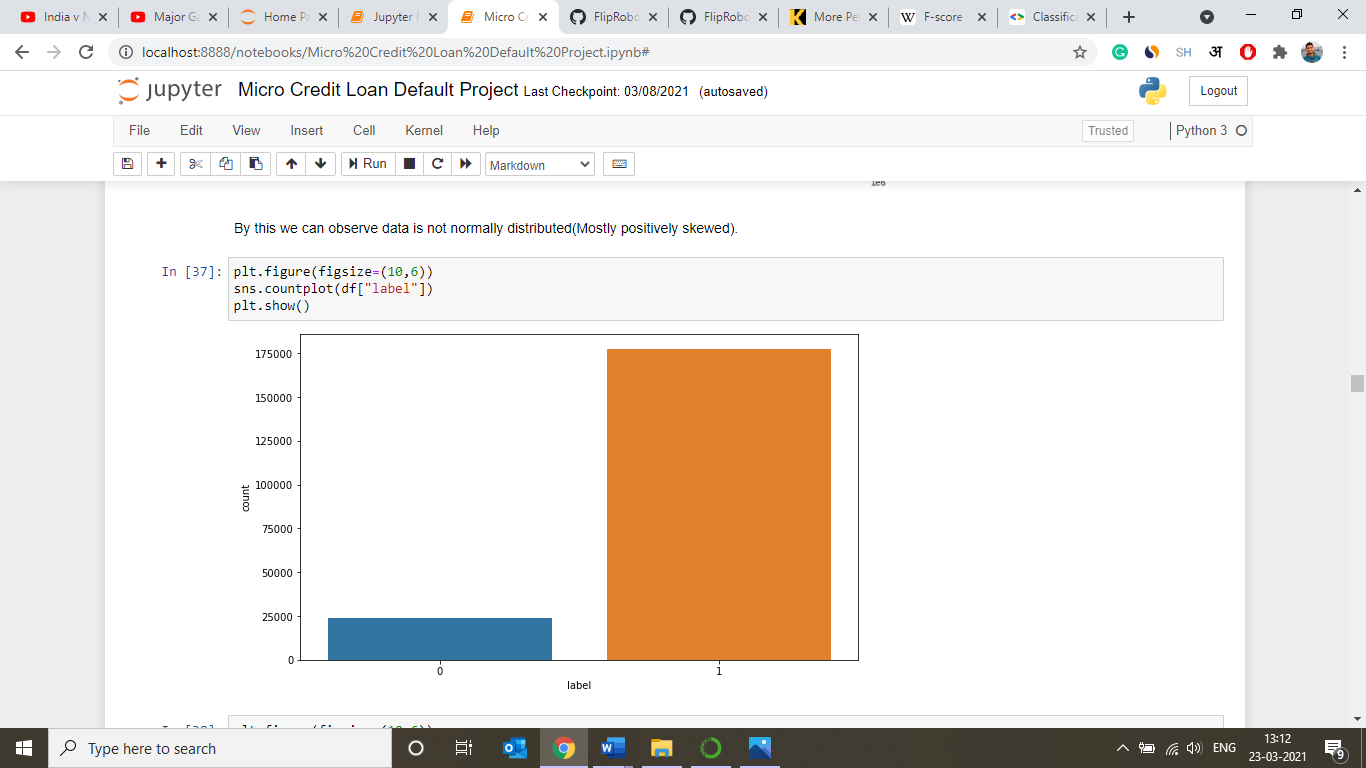
* **Distribution Plots**

For finding distribution of different variables.



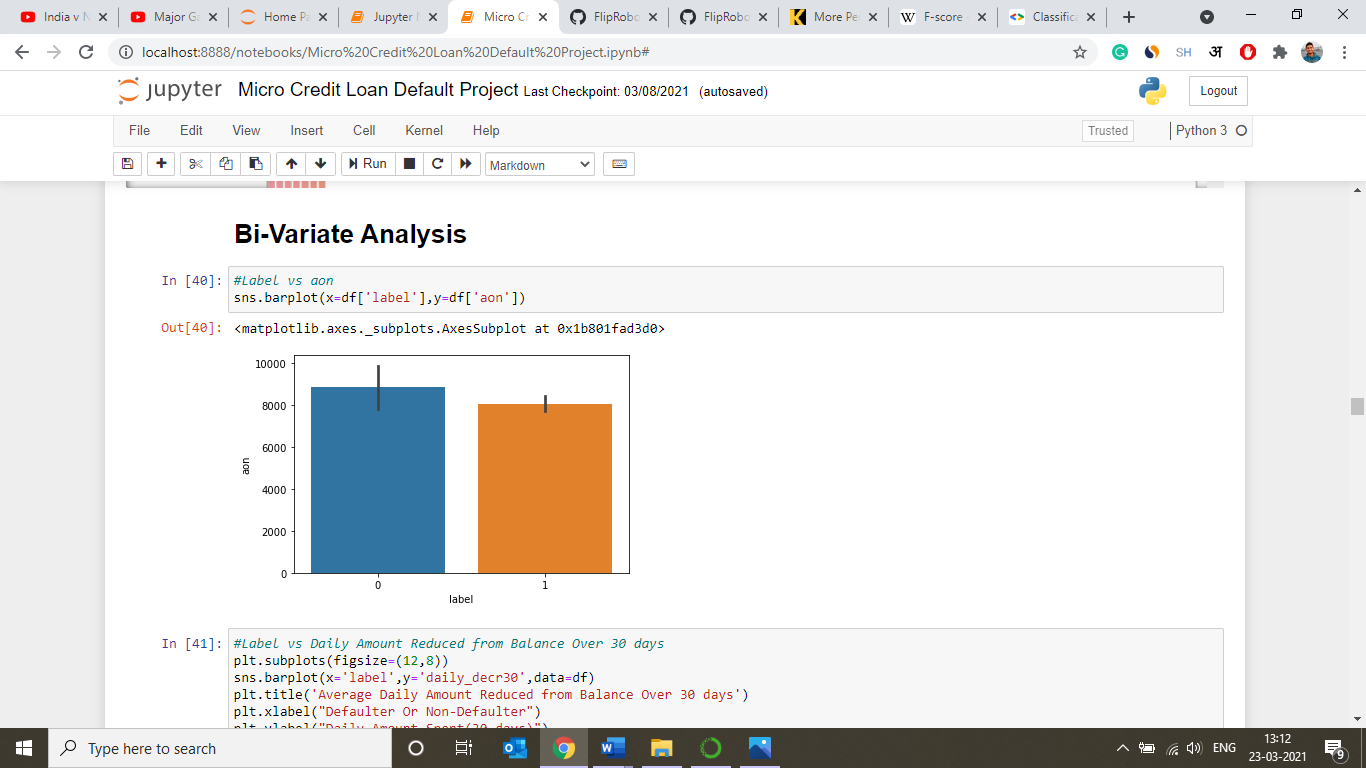
* **Count Plots**

For finding count of defaulters and non-defaulters in dataset.



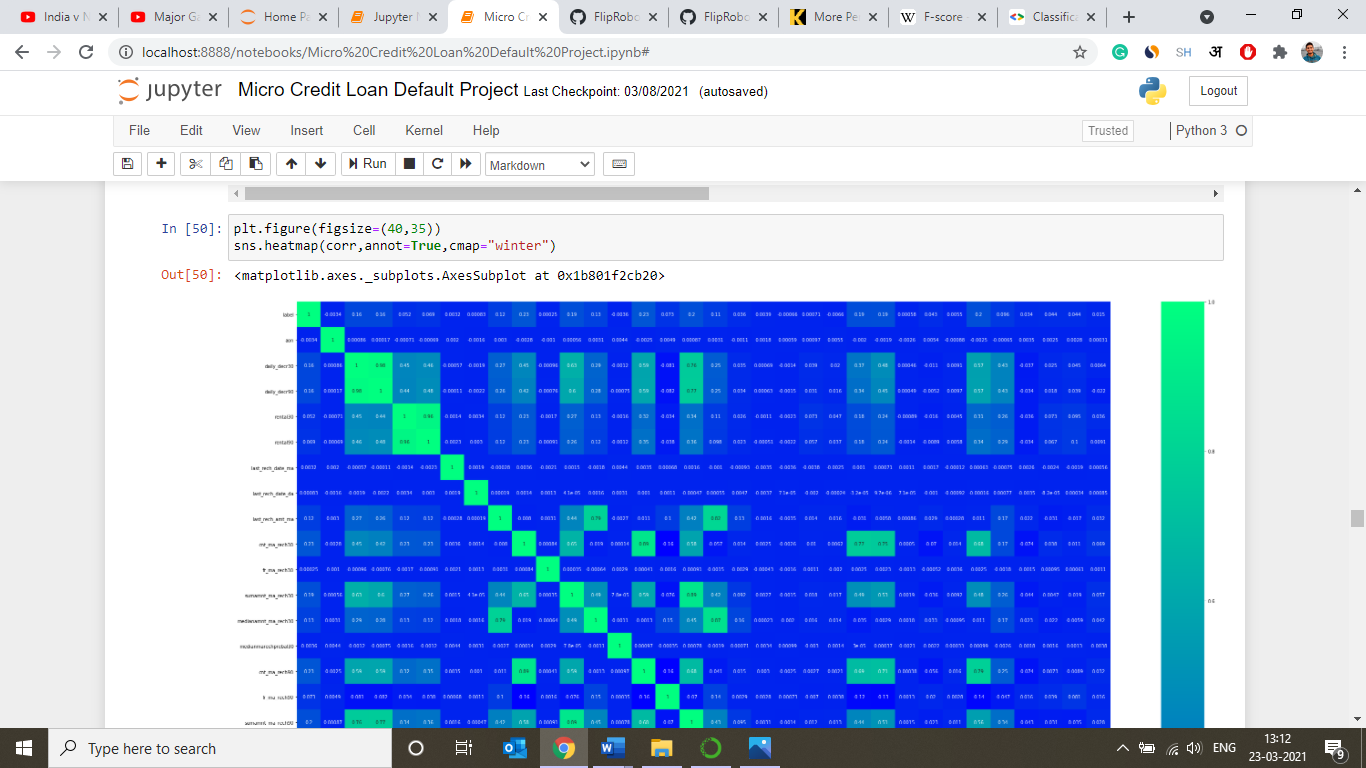
* **Bar Plots**

For finding relation between two variables of dataset.



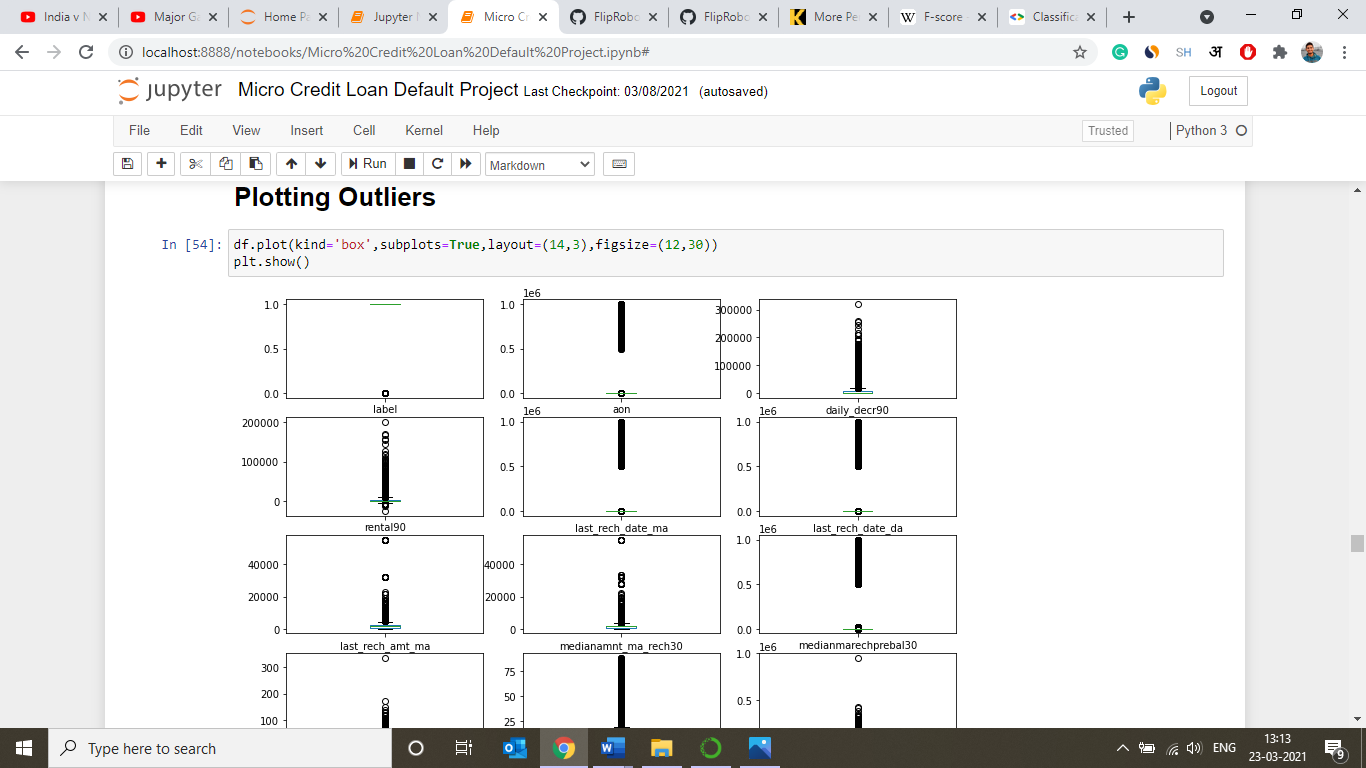
* **Heatmaps**

Used in this analysis, for finding correlation between the independent variables and target variable.



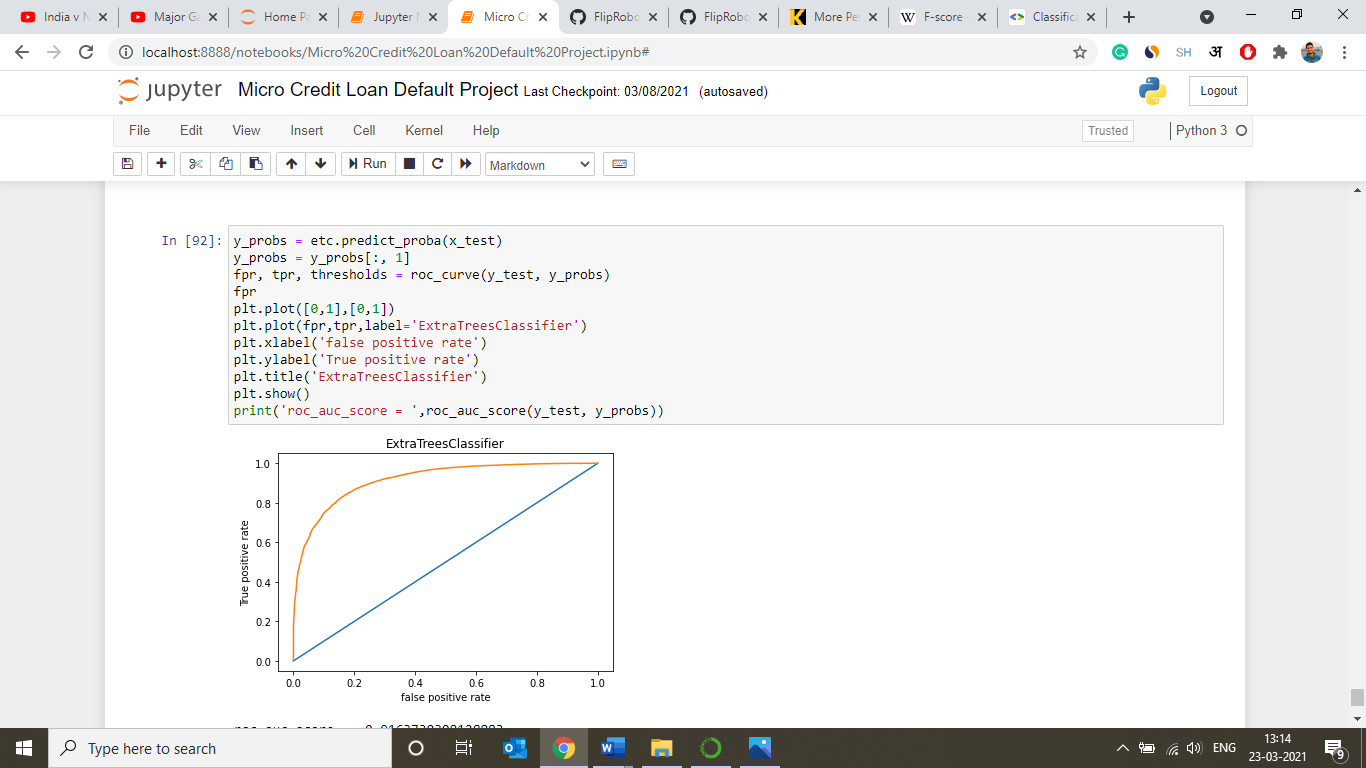
* **Box Plots**

Used for plotting outliers present in different variables of the dataset.

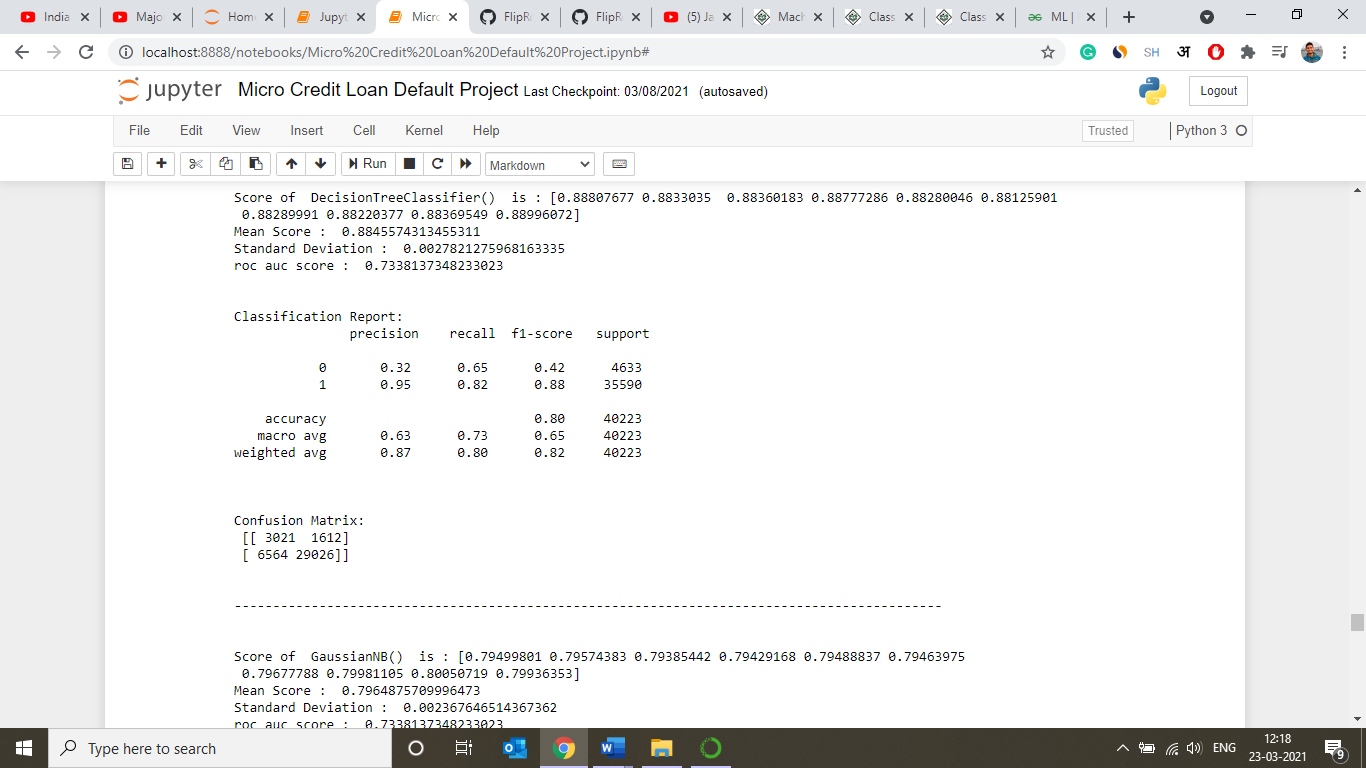


* **Line Plot**

Used here for plotting auc-roc curve.



* **Interpretation of the Results**



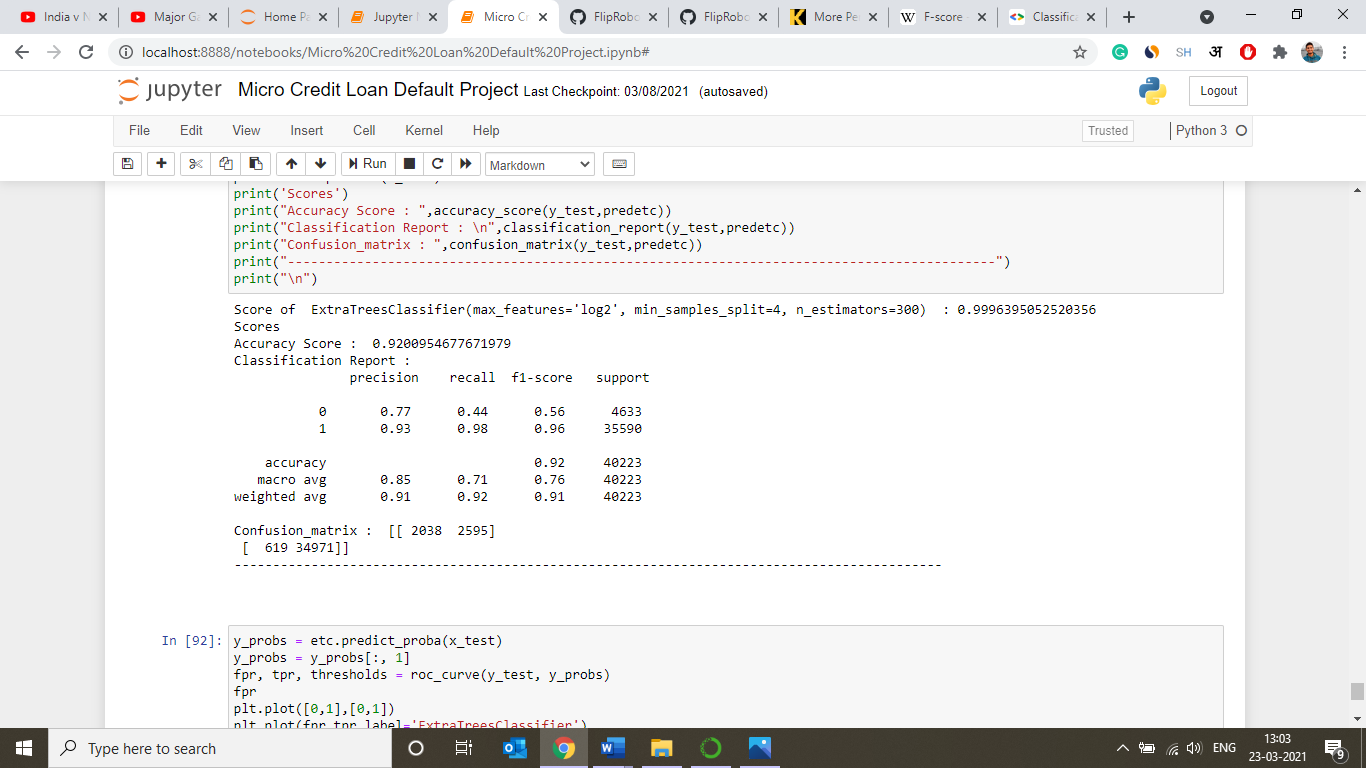


Figure 14: Scores for best fitted model using Extra Trees Classifer

We got our best scores using Decison Trees Classifier. By comparing it with results and metrics of other algorithm we came to a conclusion that it will be the best fit to our dataset.

**CONCLUSION**

* **Key Findings and Conclusions of the Study**
* After using different models for classification, we concluded that DecisionTrees Classifier was best suited to train the dataset for Micro Credit Loan Case study.
* The datasets was imbalanced in nature.
* Some data had to be omitted from the datasets in order to remove multi-collinearity, and others reasons like unrelevancy to the dataset.
* The count of outliers in the data was quite high.
* Variables were not normally distributed.
* **Learning Outcomes of the Study in respect of Data Science**
* The visualization helped us in various ways like giving insights to the datasets that how it is arranged, correlated, how balanced it is and how much outliers it is containing.
* As the datasets was huge it created problems when we tried to manipulate it or cleanse it. As when we tried to cleanse it we had to take care that data doesn’t loose its relevancy and structure. So, we tried to remove much of the unnecessary data from the datasets that was in our reach during the given time.
* And finally, we came to know that the best algorithm used to train the machine for this the dataset is DecisionTree Classifier as all the values along the metrics were highest.
* **Limitations of this work and Scope for Future Work**
* Limitations that I encountered during this project was limited time and lack of proper processor because it would have sped up the whole project.
* As the time will pass by there will be new more efficient ML algorithms that will come to existence that can further improve the prediction but for the time being and in the near future this prediction based approach can help a lot of banks in order to improve their efficiency and increase their profits.