

# Project Report On

“ Micro-Credit Defaulter Model ”

****

Submitted By:

# Akanksha Raja Parmar

# ACKNOWLEDGEMENT

The internship opportunity I have with Flip Robo Technologies is a great chance

for learning and professional development . I perceive this opportunity as a big

milestone in my career development . I will strive to use gained skills acknowledge

in the best possible way.

I would like to extend my appreciation and thanks for the mentors from

Data Trained Institute and professionals from FlipRoboTechnologies who had extended their suggestions and directions .

# References:

[***https://sklearn.org/supervised\_learning.html#supervised-leaning***](https://sklearn.org/supervised_learning.html#supervised-leaning)

[***https://www.youtube.com/user/krishnaik06***](https://www.youtube.com/user/krishnaik06)

[***https://www.youtube.com/c/CodeWithHarry***](https://www.youtube.com/c/CodeWithHarry)

[***https://www.google.com/***](https://www.google.com/)

[***https://www.github.com/***](https://www.github.com/)

***“ INTRODUCTION ”***

Business Problem:

A client in Telecom Industry is collaborating with an MFI (Microfinance Institution) 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.

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.

Background of domain:

* Microfinance is a banking service provided to unemployed or low-income individuals or groups who otherwise would have no other access to financial services.
* Microfinance allows people to take on reasonable small business loans safely, and in a manner that is consistent with ethical lending practices.
* The majority of micro financing operations occur in developing nations, such as Uganda, Indonesia, Serbia and Honduras.
* Like conventional lenders, micro financiers charge interest on loans and institute specific repayment plans.
* The World Bank estimates that more than 500 million people have benefited from microfinance-related operations.

Indonesia is renowned for its large scale microfinance sector. With a range of commercial banks. More than 56.5 million Micro Small Medium Enterprises (MSME), contributed greater than 50% of Gross Domestic Product (GDP) in 2014 . However, many of them do not have adequate access to the bank financing they need to grow their businesses, particularly in rural areas.

Some rural communities in Indonesia have no choice but to seek out loans from unregulated moneylenders. Micro lenders, particularly those operating under Indonesian banks, as well as social enterprise startups, are also targeting these communities through their high mobile penetration rates and are developing the right digital platforms to reach out to them.

Only around 22% of Indonesians are connected to format financial institutions.

Micro-finance is accessible for people in remote areas and on small island, not just people in the cities.

In 2012, there were 143 million unique mobile subscribers, more than double the number of bank account holders(62 million). Telecommunication operators have more than 300,000 locations at which phone vouchers are sold. Most banks would like to have access to these distribution networks, which would enable them to access the poorest requiring micro-finance.

MOTIVATION FOR PROBLEM UNDER TAKEN:

Based on data provided from our client database, customer’s repayment of loan is assessed based on different factors. By building the model, we can access which customers are highly likely to repay the loan, thereby it will be useful for those needy people who will repay the loan and also prevent the loss to the customer by avoiding loans to the defaulters.

ANALYTICAL PROBLEM FRAMING

MATHEMATICAL MODELLING OF PROBLEM:

Mathematical modeling is simply the method of implementing statistical analysis to a dataset where a Statistical Model is a mathematical representation of observed data.

While analyzing the data, there are an array of statistical models we can choose to utilize.

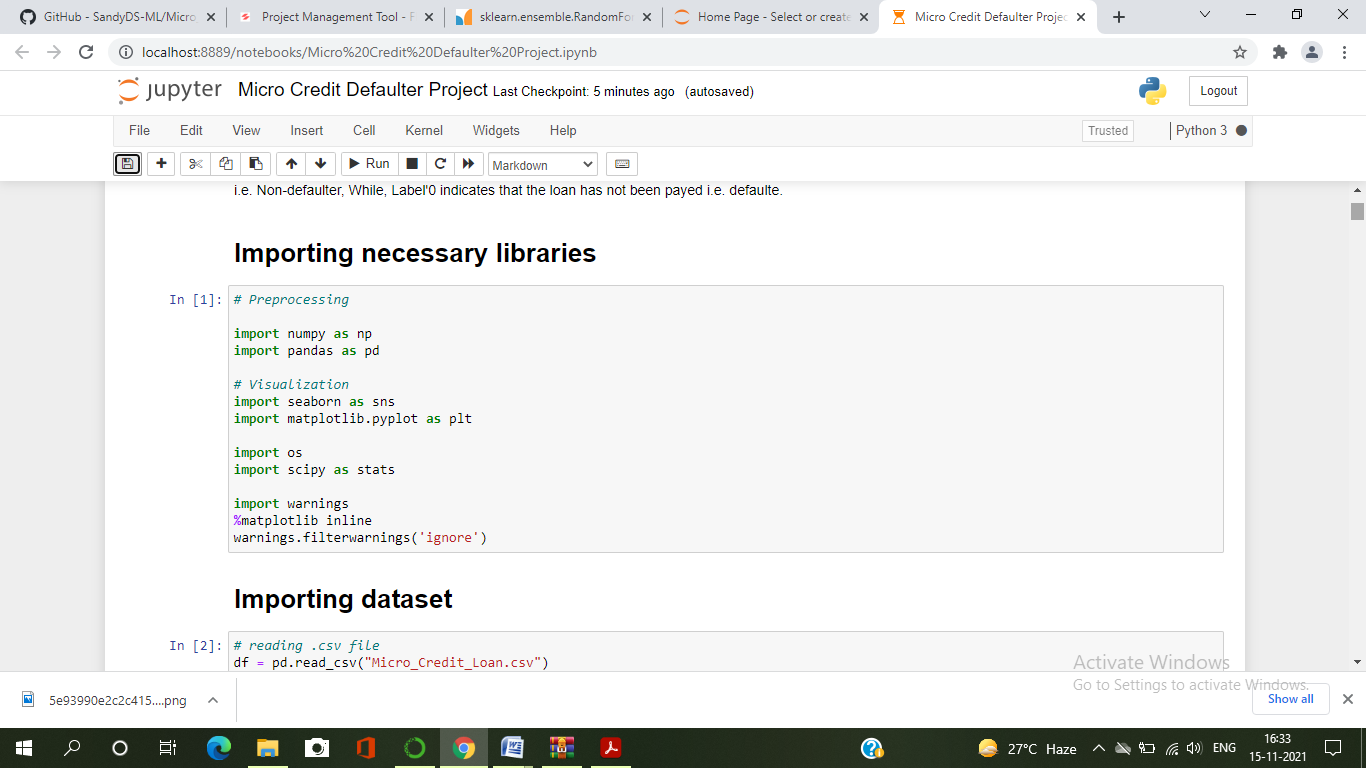
For the given project, we need to predict whether the customer is a defaulter or not.

This is a classification problem. There are wide varieties of classification models like decision tree, extra trees, random forest, gradient boosting, Logistic Regression, bagging classifier.

DATA SOURCE AND FORMAT:

The data has been provided by client in a comma separated values(.csv) format

Libraries required:



* Import numpy as np:

It is defined as a Python package used for performing the various numerical computations and processing of the multidimensional and single dimensional array elements. The calculations using Numpy arrays are faster than the normal Python array.

* Import matplotlib.pyplot as plt:

Matplotlib and Seaborn acts as the backbone of data visualization through Python.

Matplotlib: It is a Python library used for plotting graphs with the help of other libraries like Numpy and Pandas. It is a powerful tool for visualizing data in Python. It is used for creating statical interferences and plotting 2D fraphs of arrays.

* Import seaborn as sns: Seaborn

Seaborn is also a Python library used to replotting graphs with the help of Matplotlib, Pandas and Numpy. It is built on the roof of Matplotlib and is considered as a superset of the Matplotlib library. It helps in visualizing univariate and bivariate data.

* from scipy.stats import zscore
* from sklearn.preprocessing import PowerTransformer
* from sklearn.preprocessing import MinMaxScaler
* from statmodel.starts.outliers\_influence

import variance\_inflation\_factor

* from imbleran.over\_sampling import SMOTE

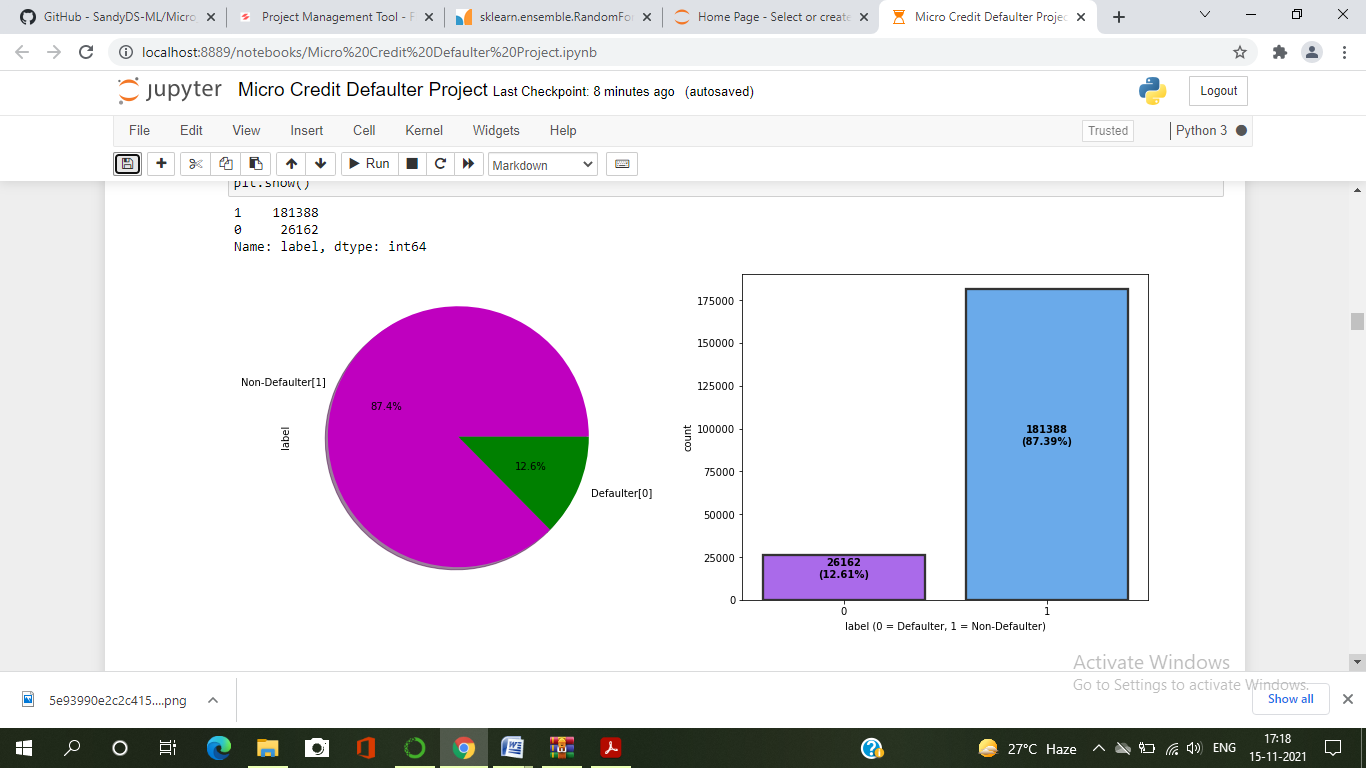
With the above sufficient libraries, we can perform pre-processing and data cleaning. For building my ML models Below libraries are required.

* from sklearn.model\_selection import train\_test\_split
* from sklearn.tree import DecisionTreeClassifier
* from sklearn.ensemble import RandomforestClassifier, ExtraTreesClassifier
* from sklearn.ensemble import GradientBoostingClassifier, BaggingClassifier
* from sklearn.metrics import classification\_report,confusion\_matrix,roc\_curve,accuracy\_score,roc\_auc\_score
* from sklearn.model\_selection import cross\_val\_score

MODELS DEVELOPMENT AND EVALUATION

Identification of possible Problem-solving approaches (Methods):

I have used both statistical and analytical approaches to solve the problem which mainly includes the pre-processing of the data also used EDA techniques and heat map to check the correlation of independent and dependent features. Also, before building the model, I made sure that the input data is cleaned and scaled before it was fed into the machine learning models. The data mainly had class imbalancing issue which looks like below.



From the above we can see that the data set is highly imbalanced, so applied SMOTE method to balance the dataset.

For this particular project we need to predict whether the user paid back the credit loan amount within 5 days of issuing the loan. In this dataset, label is the target variable, which consists of two categories, defaulters and non-defaulters. Which means our target column is categorical in nature so this is a classification problem.

I have used many classification algorithms and got the prediction results. By doing various evaluations I have selected Gradient Boosting Classification as best suitable algorithm to create our final model as if it is giving least difference in accuracy score and cross validation score.

In order to get good performanced to check whether my model getting over-fitting and under-fitting I have made use of the K-Fold cross validation and then hyper parameter tuning on best model. Then I saved my final model and loaded the same for predictions.

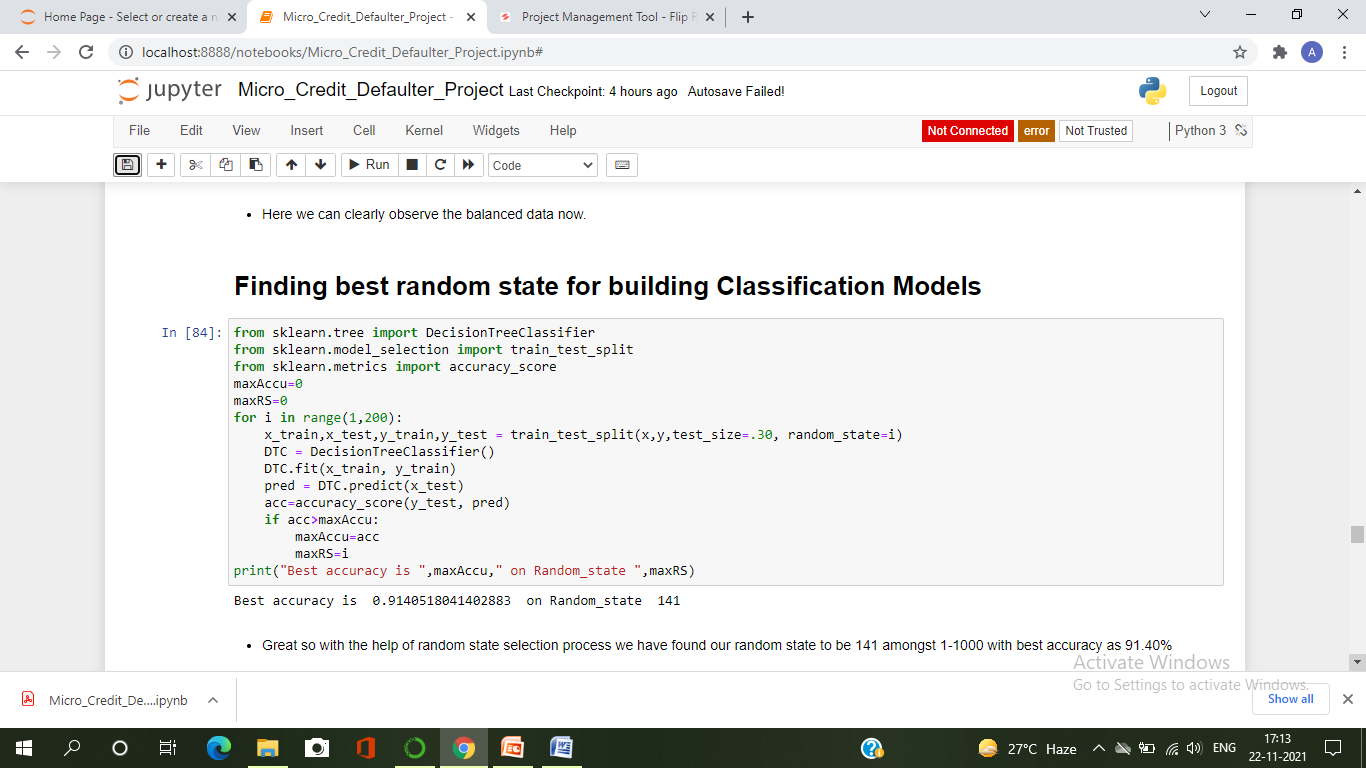
Testing of Identified Approaches(Algorithms)

Since label is my target variable which is categorical in nature, from this I can conclude that it is a classification type problem hence I have used following classification algorithm. After the pre-processing and data cleaning I left with 27 columns including target and with the help of feature importance bar graph I used these idependent features for model building and prediction. The algorithms used on training the data are as follows.

1. Random Forest Classifier
2. Decision Forest Classifier
3. Extra Trees Classifier
4. Gradient Boosting Classifier
5. Bagging Classifier

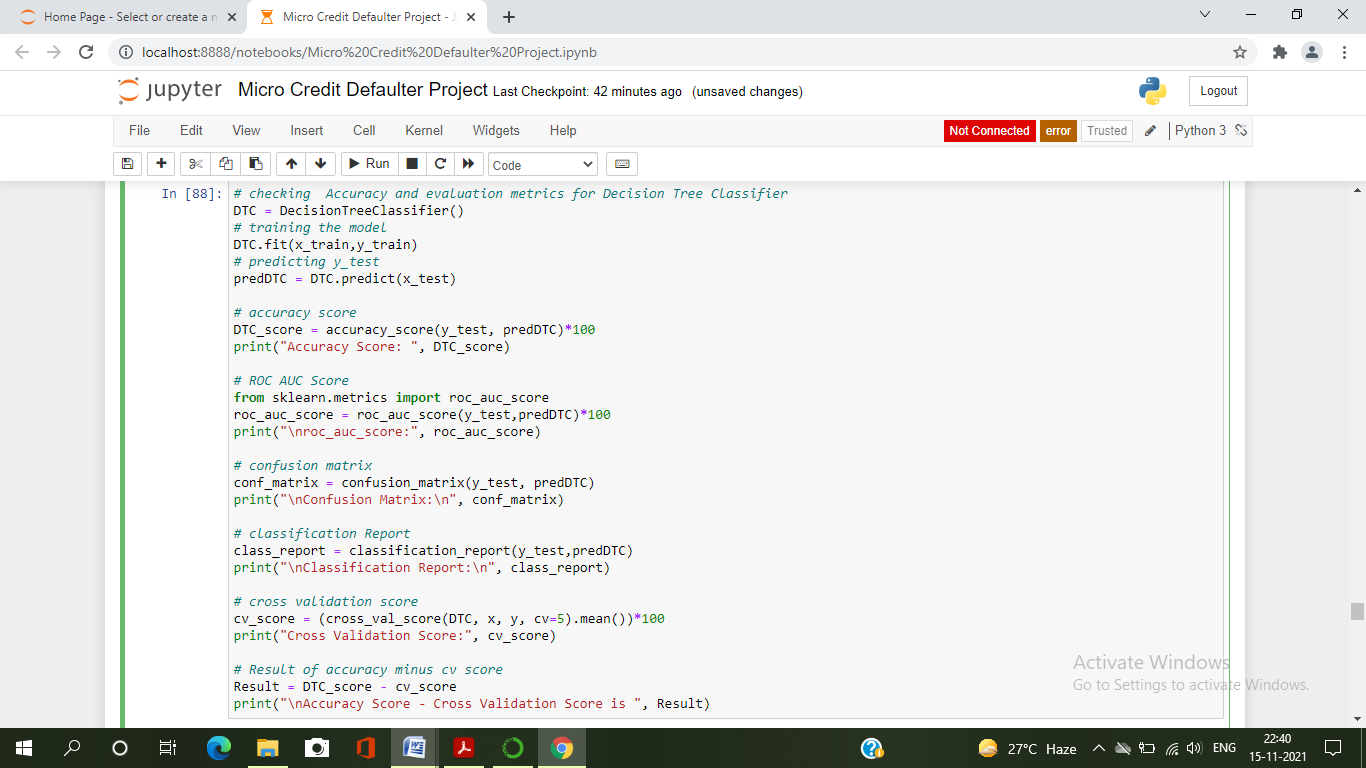
Run and evaluate selected models

I have used 5 classification algorithms after choosing random state amongst 1-1000 numbers. I have used Decision Tree Classifier to find best random state and the code is as below.

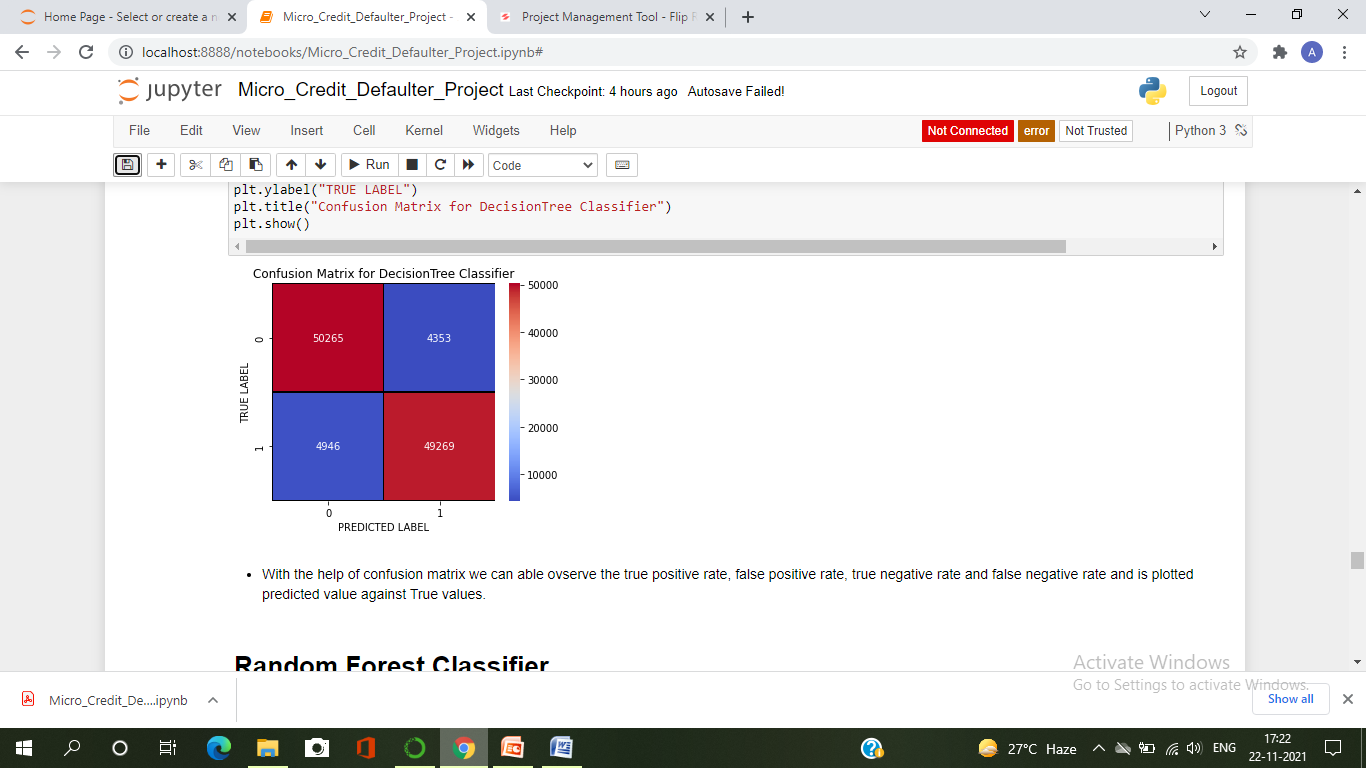


Model Building:

1. DecisionTree Classifier

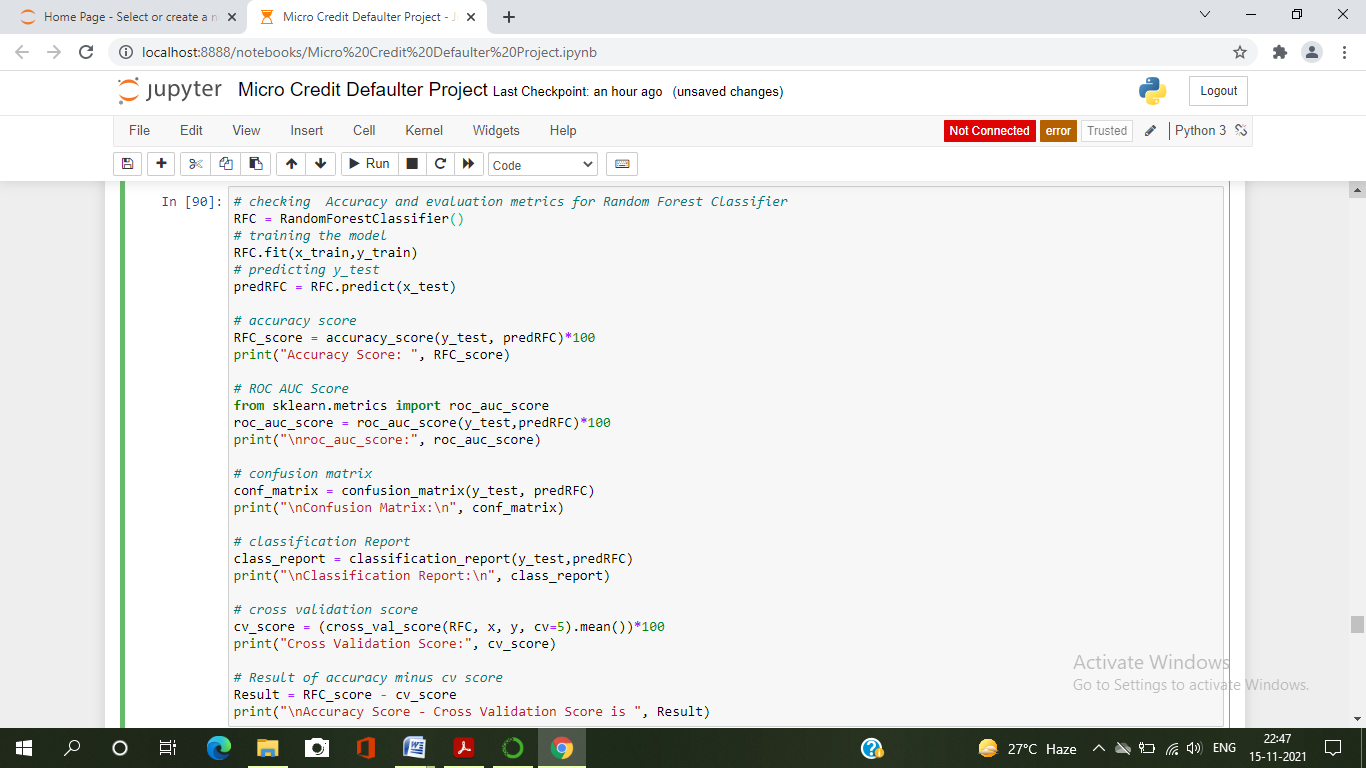


* Created the Decision Tree Classifier model and checked for its evaluation metrics and it is giving accuracy 91 %

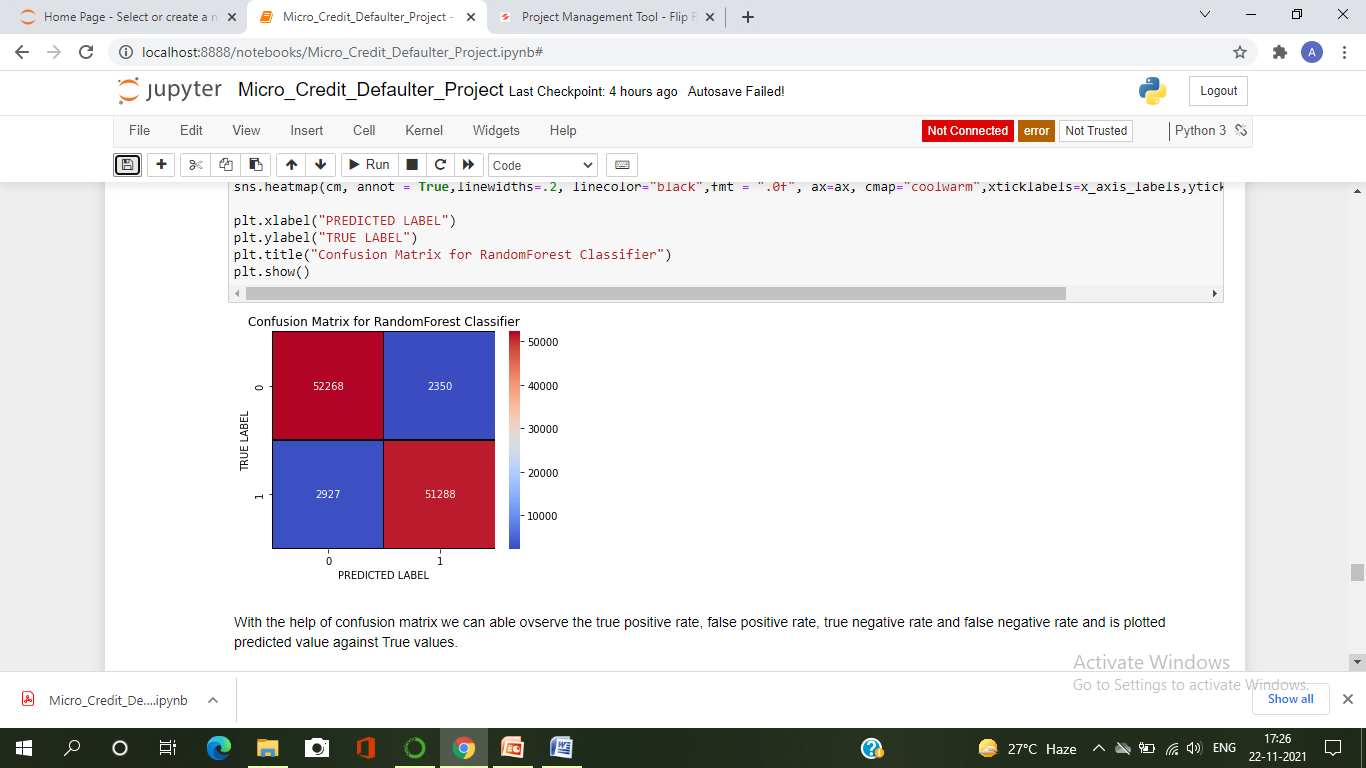


* With the help of confusion matrix we can able observe the true positive rate, false positive rate, true negative rate and false negative rate and is plotted predicted value against True values.

1. Random Forest Classifier

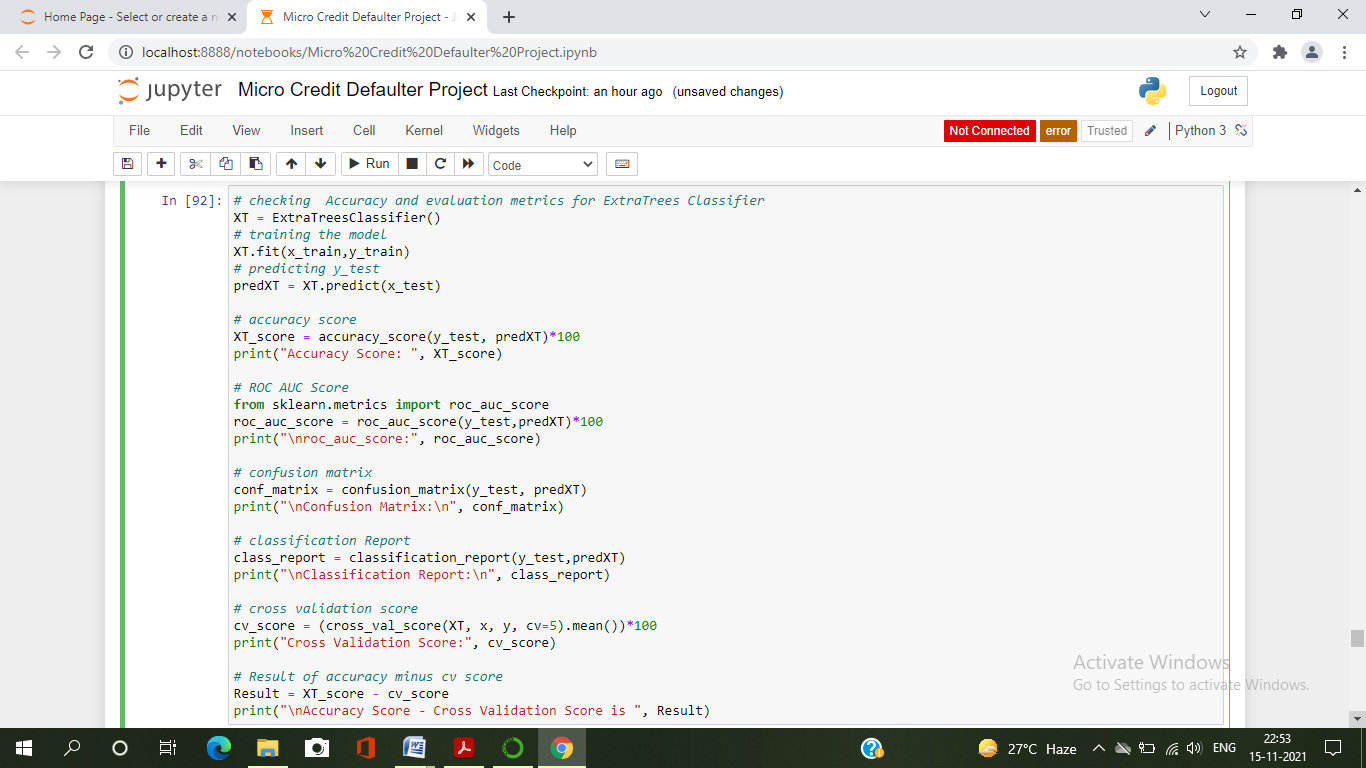


* Created Random Forest Classifier Model and checked for its evaluation metrics. The model giving 95% accuracy

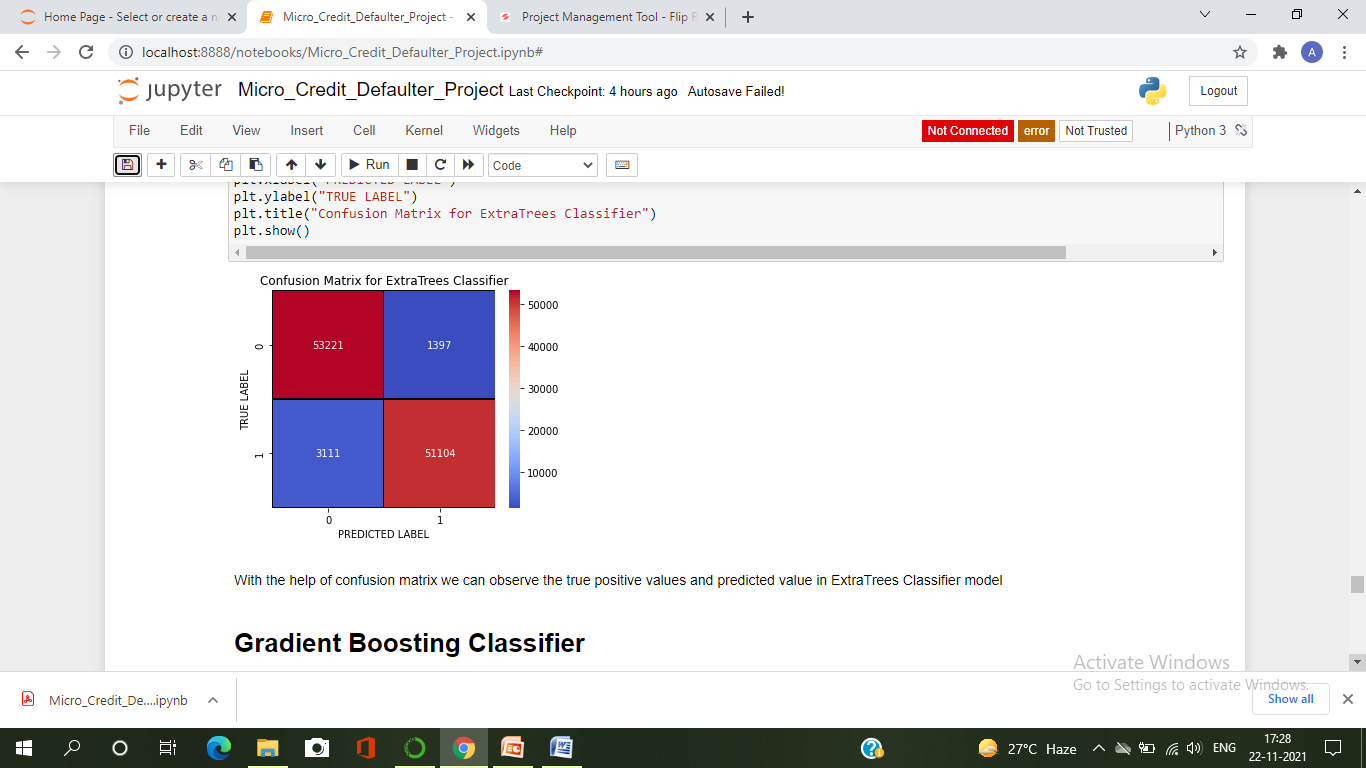


* With the help of confusion matrix we can able ovserve the true positive rate, false positive rate, true negative rate and false negative rate and is plotted predicted value against True values.

3 Extra Trees Classifier

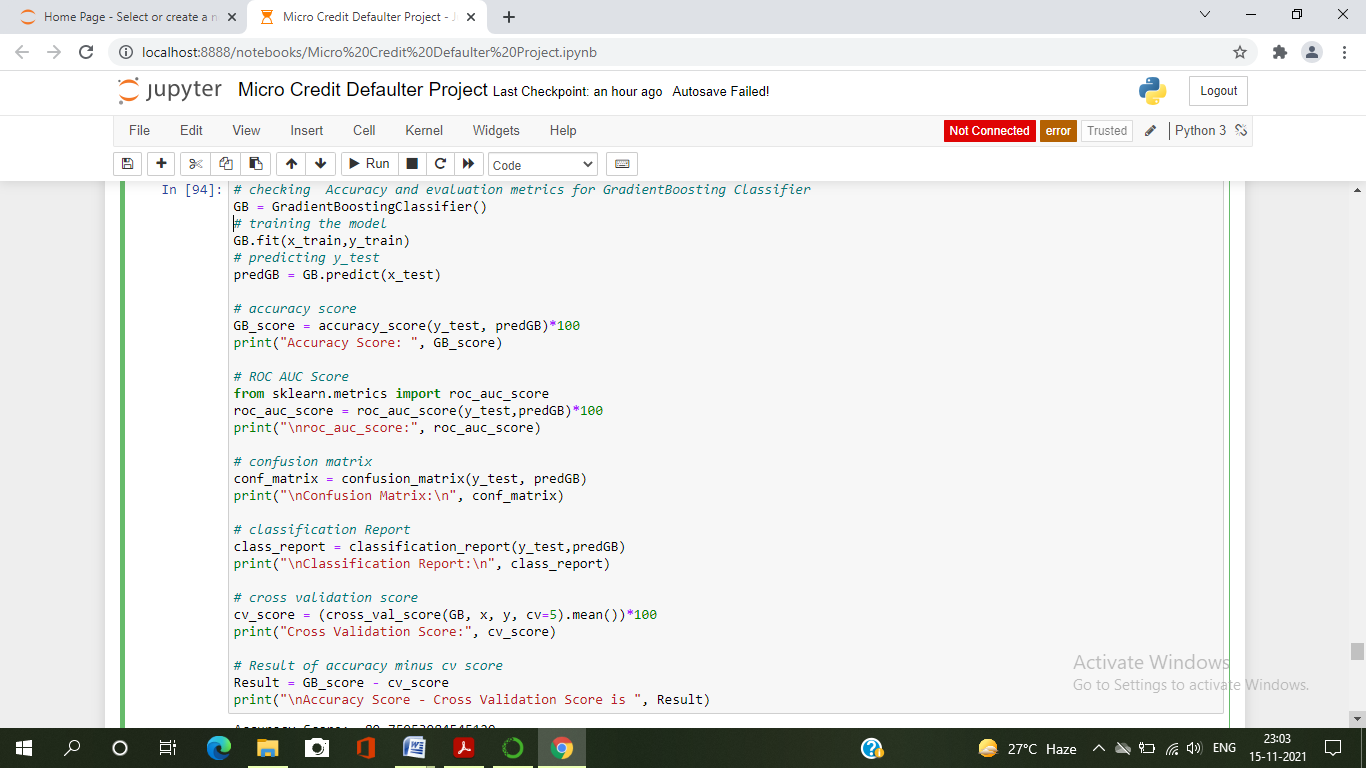


* Created the ExtraTrees Classifier model and checked for it's evaluation metrics . The ExtraTrees Classifier model giving 95% accuracy



* With the help of confusion matrix we can observe the true positive values and predicted value in ExtraTrees Classifier model

1. Gradient Boosting Classifier

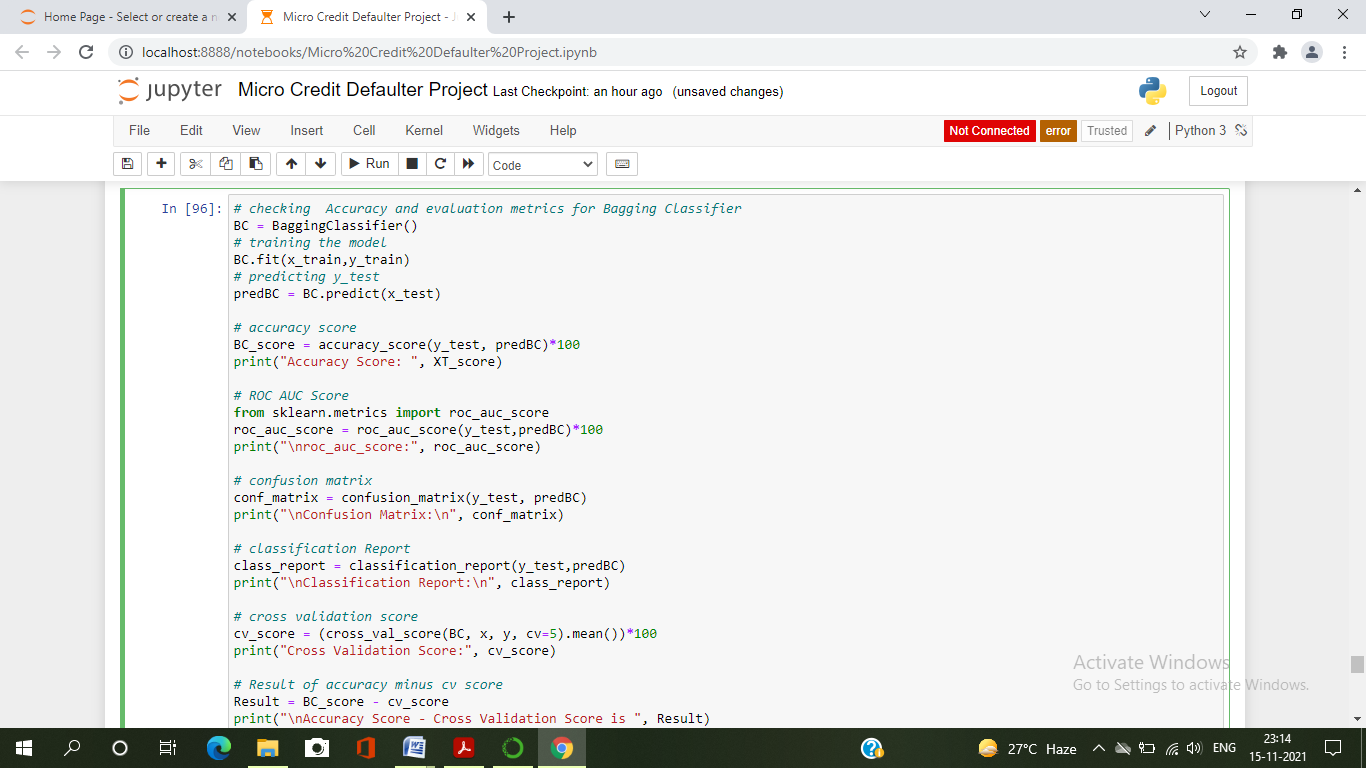


* Created the Gradient Boosting Classifier model and checked for its evaluation metrics. The Gradient Boosting Classifier model giving 89% accuracy.

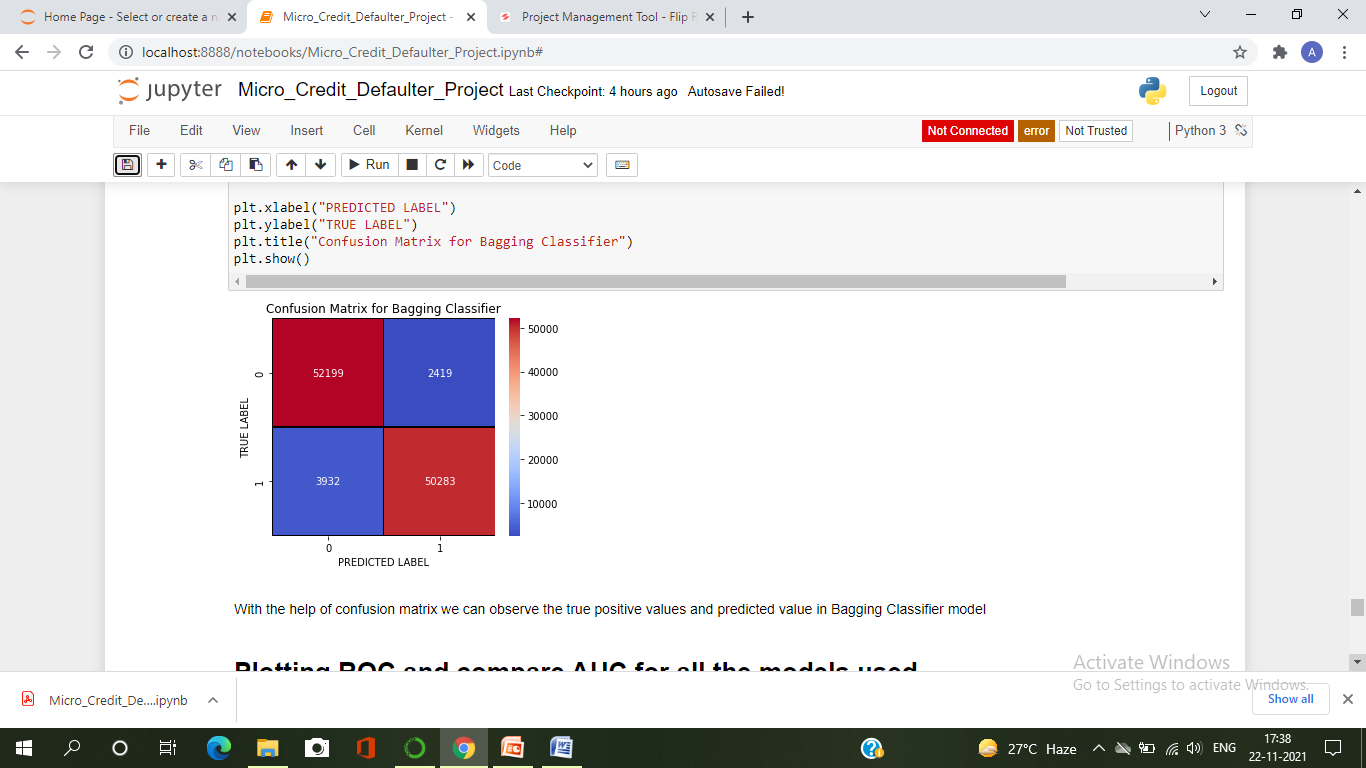


* With the help of confusion matrix we can observe the true positive values and predicted value in GradientBoosting Classifier model

1. Bagging Classifier



* Created the Bagging Classifier model and checked for its evaluation metrics. The Bagging Classifier model giving 95% accuray



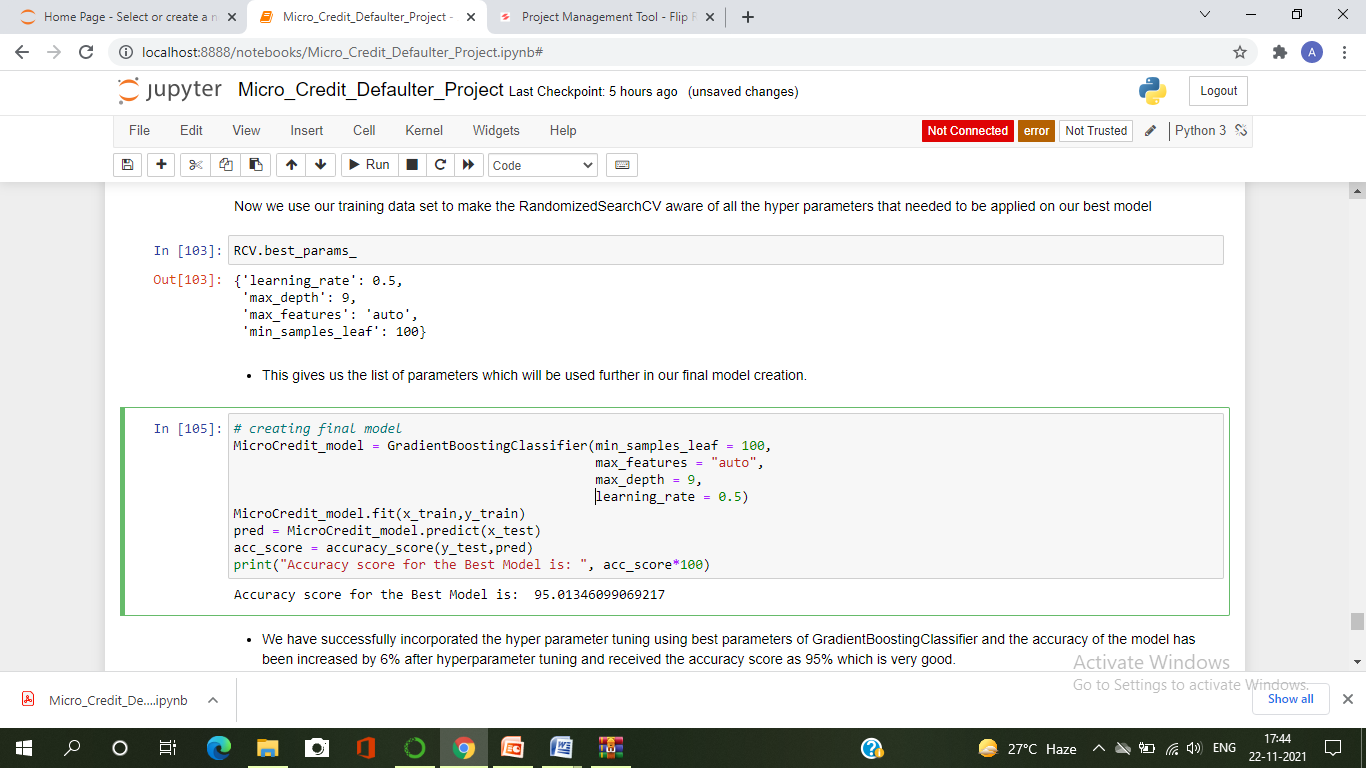
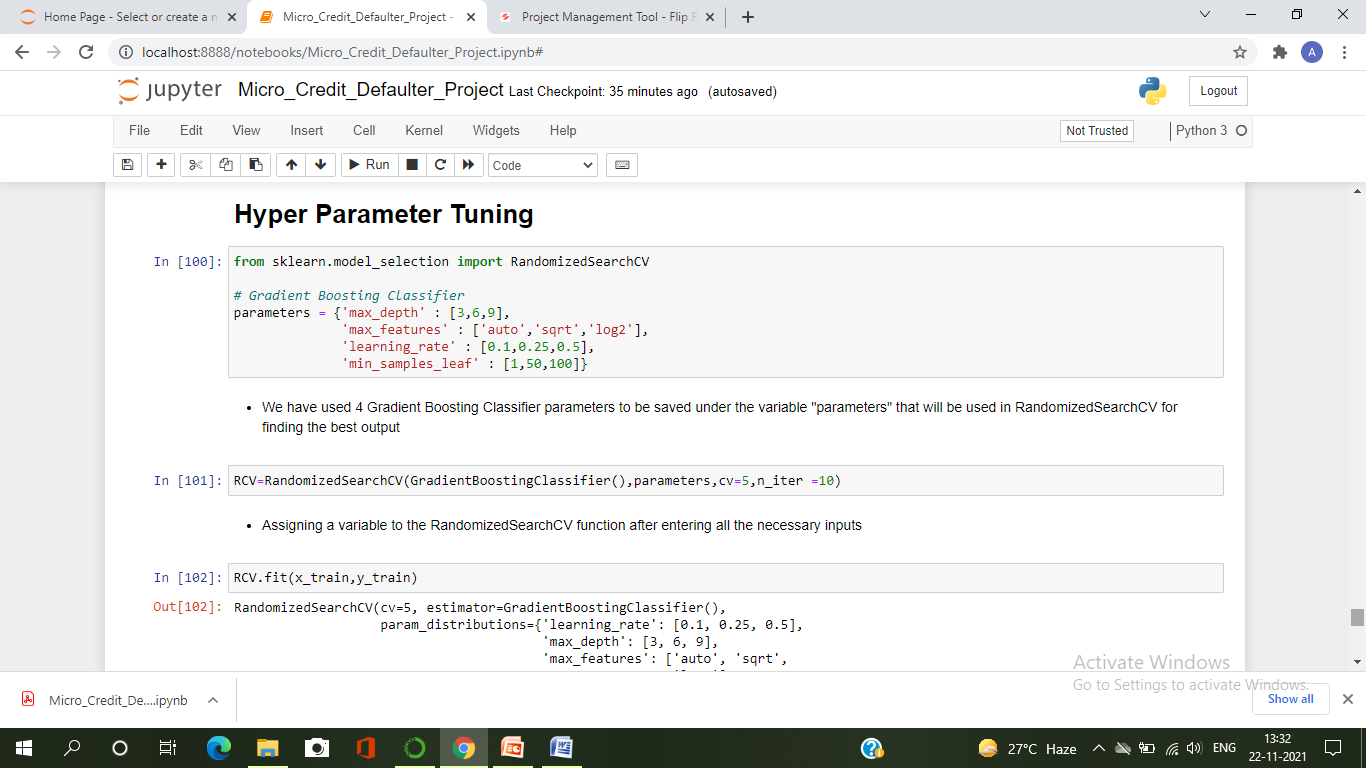
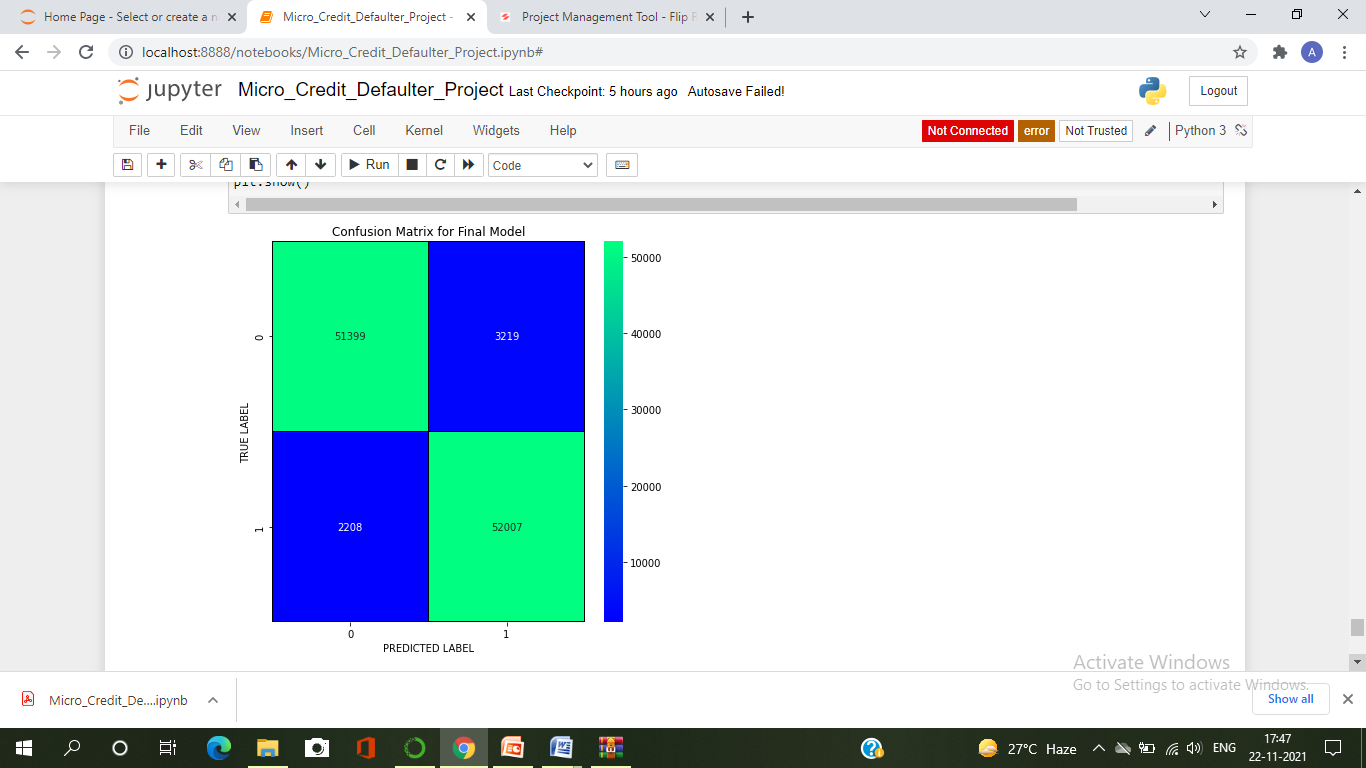
* With the help of confusion matrix we can observe the true positive values and predicted value in Bagging Classifier model

Model Selection :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | | Accuracy\_Score | | Cross\_validation\_Score | Difference |
| RandomForestClassifier | | 95.15 | | 94.95 | 0.19 |
| DecisionTreeClassifier | | 91.45 | | 90.91 | 0.55 |
| ExtraTreesClassifier | | 95.85 | | 96.29 | -0.43 |
| GradientBoostingClassifier | | 89.89 | | 89.86 | 0.02 |
| BaggingClassifier | 95.85 | | 93.63 | | 0.52 |

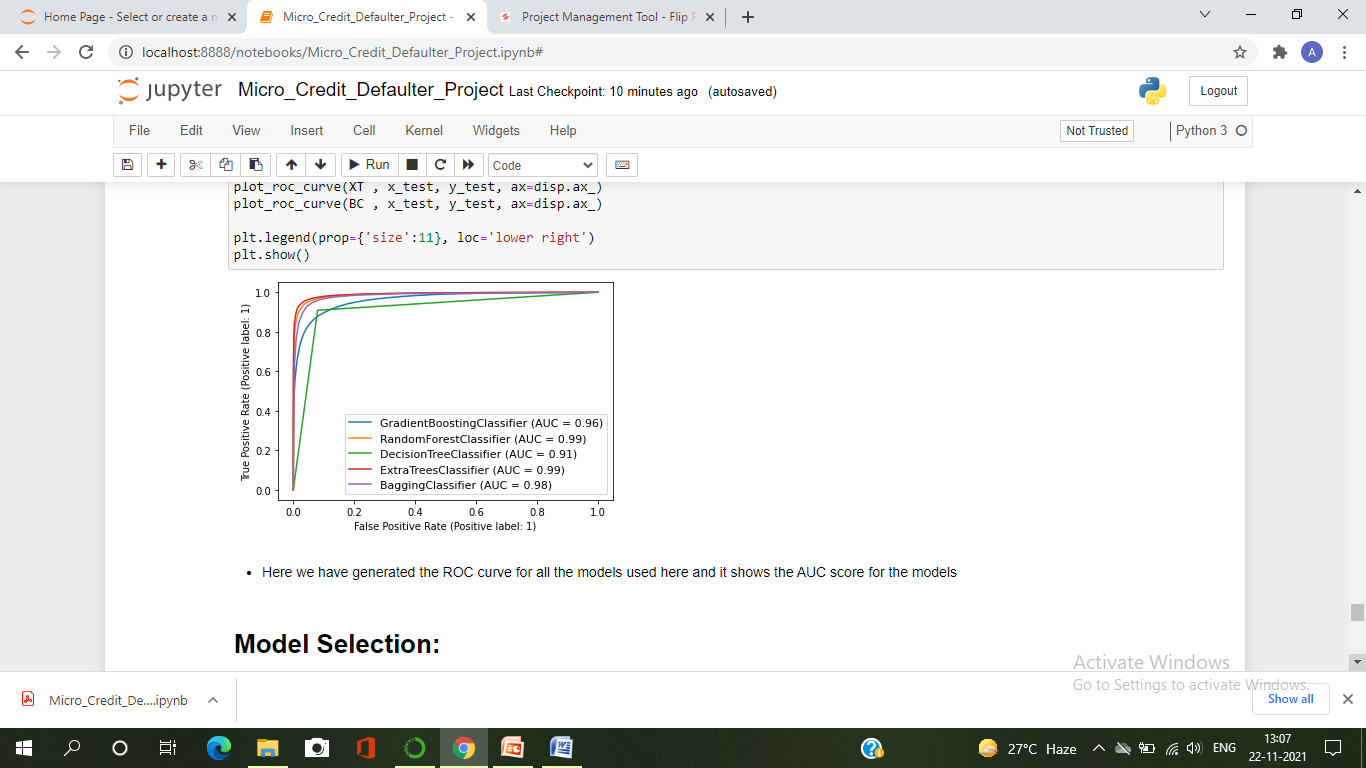
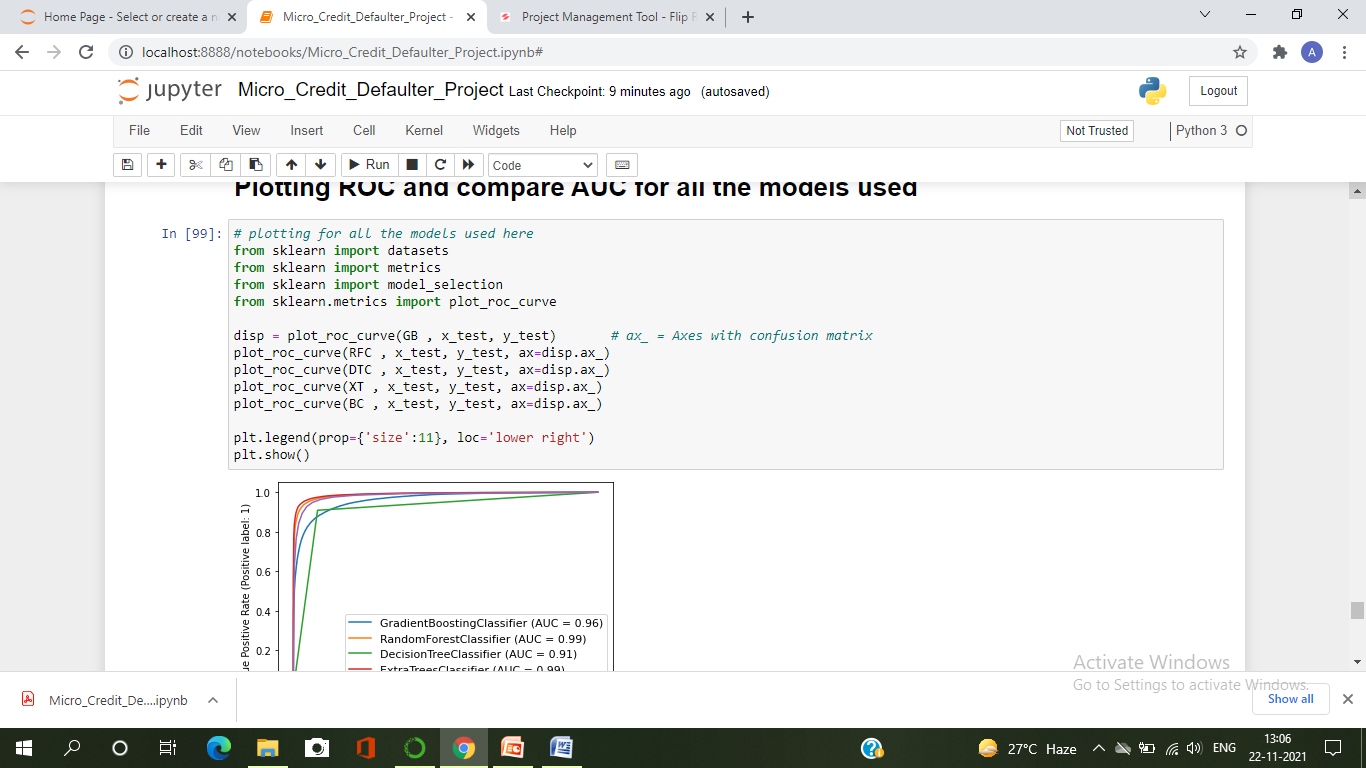
* From the difference between accuracy and cross validation score, Gradient Boosting Classifier has least difference compared to other models. So, we can conclude that Gradient Boosting Classifier as our best fitting model. Performed Hyper Parameter Tuning to increase the best model accuracy.

Hyper Parameter Tuning

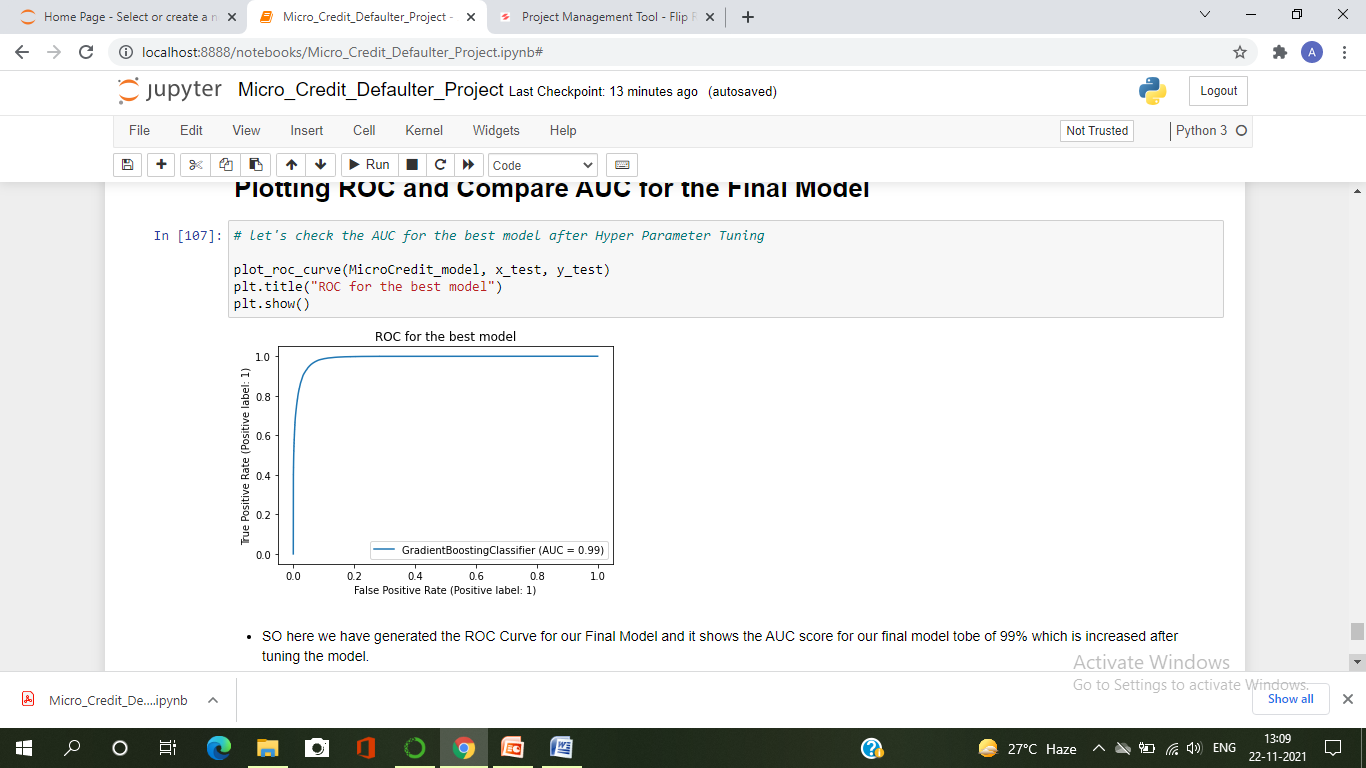
 

I have successfully incorporated the hyper parameter tuning using best parameters of Gradient Boosting Classifier and the accuracy of the model has been increased after hyperparameter tuning and received the accuracy score which is very good. With the help of confusion matrix, we can able to see actual and predicted values for the final model. And also, we can understand the number of times we got the correct outputs and the number of times my model missed to provide the correct prediction.

ROC-AUC Curve for all the models used :



ROC-AUC Curve for the final model



I have generated the ROC Curve for all the models and for the best model and it shows the AUC score for the models. The AUC score for my final model to be of 99% which is increased after tuning the model.

* Key Metrics for success in solving problem under consideration

The key metrics used here were Accuracy Score, Precision, Recall, F1score, Cross Validation Score, ROC AUC Score and Confusion Matrix. We tried to find out the best parameters and also to increase our scores by using Hyper Parameter Tuning and used Randomized SearchCV method.

* Accuracy score : means how accurate our model is that is the degree of closeness of the measured value to a standard or true value. It is one metric for evaluating classification models. Accuracy is the ratio of number of correct predictions into number of predictions.
* Precision: is the degree to which repeated measurements under the same conditions are unchanged. It is amount of information that is conveyed by a value. It refers to the data that is correctly classified by the classification algorithm.
* Recall: is how many of the true positives were recalled (found) .Recall refers to the percentage of data that is relevant to the class. In binary classification problem recall is calculated as below:

Recall = Number of True Positives/ (Total number of True Positives + Total Number of False Negative)

* F1 Score: is used to express the performance of the machine learning model (or classifier). It gives the combined information about the precision and recall of a model. This means high F1-score indicates a high value for both recall and precision.
* Cross Validation Score: is a technique in which we train our model using the subset of the data-set. It is used to protect against overfitting in a predictive model, particularly in a case where the amount of data may be limited. In cross-validation, you make a fixed number of folds (or partitions) of the data, run the analysis on each fold, and then average the overall error estimate. It is used to estimate the performance of ML models.
* ROC-AUC Curve: The Receiver Operator Characteristic(ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR against FPR at various threshold values.

The Area UnderCurve(AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.

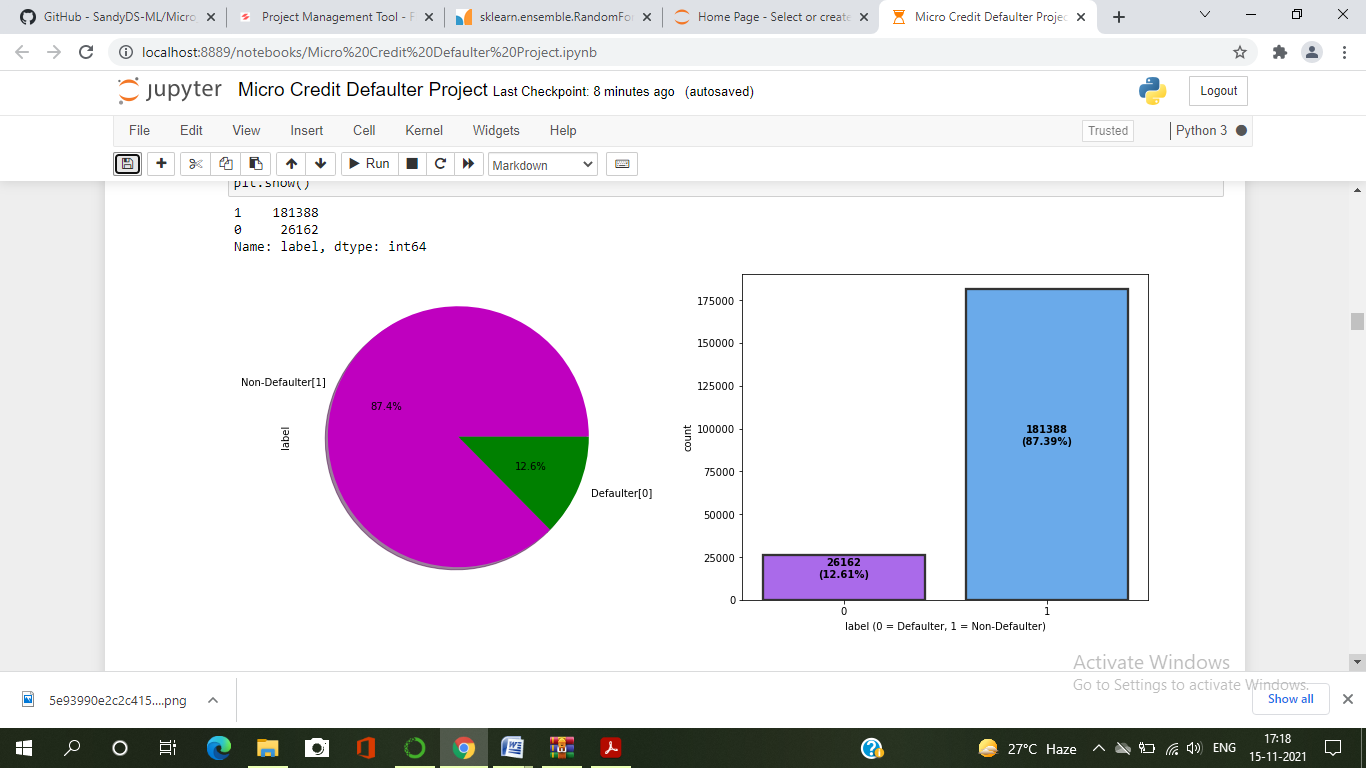
* Confusion Matrix: is one of the evaluation metrics for machine learning classification problems, where a trained model is being evaluated for accuracy and other performance measures. And this matrix is called the confusion matrix since it results in an output that shows how the system is confused between the two classes.

Visualizations

I used pandas profiling to get the over viewed visualization on the preprocessed data. Pandas is an open-source Python module with which we can do an exploratory data analysis to get detailed description of the features and it helps in visualizing and understanding the distribution of each variable. I have analysed the data using both univariate and bivariate analysis. In univariate analysis I have used distribution plot, pie plot and count plot and in bivariate analysis I have used bar plots. These plots have given good pattern.

Univariate Analysis :

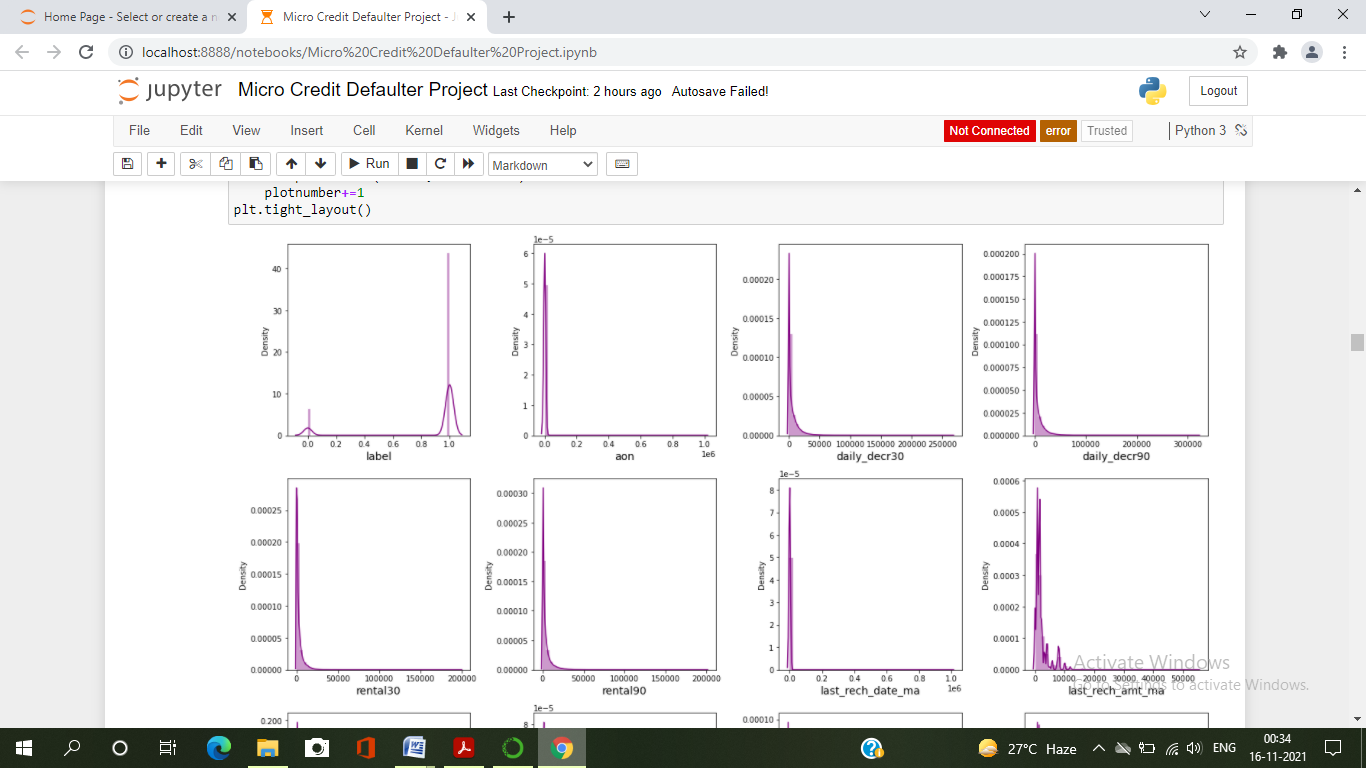
Visualizing label whether the user paid back the credit amount within 5 days of issuing the loan or not.

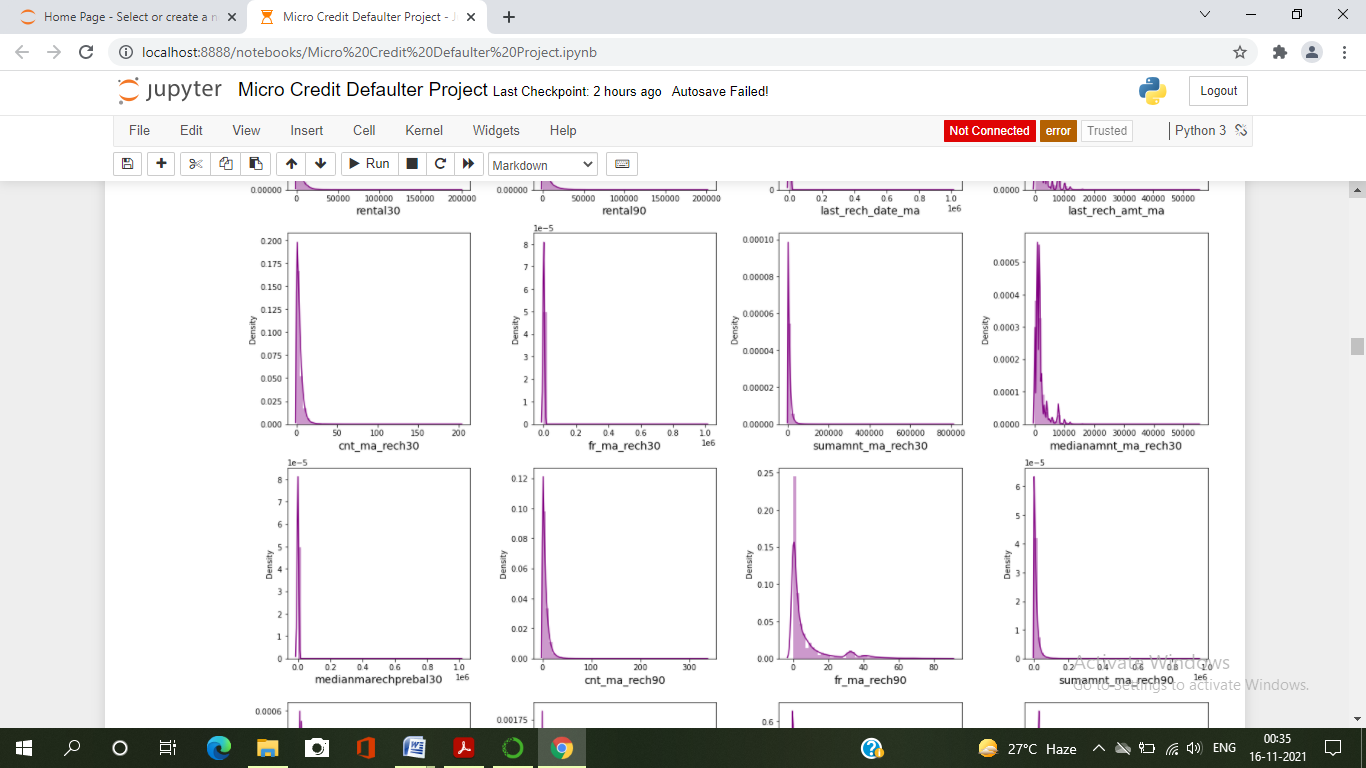


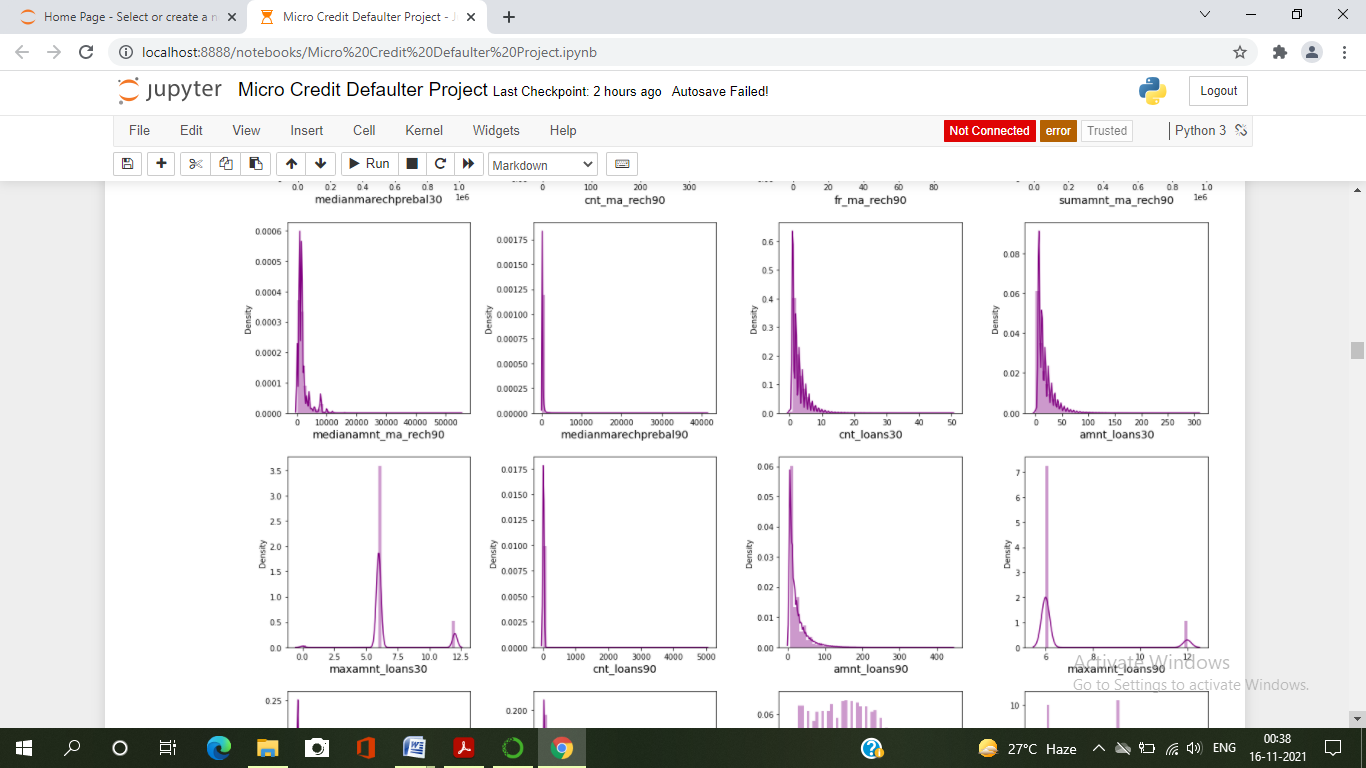
***Observation***:

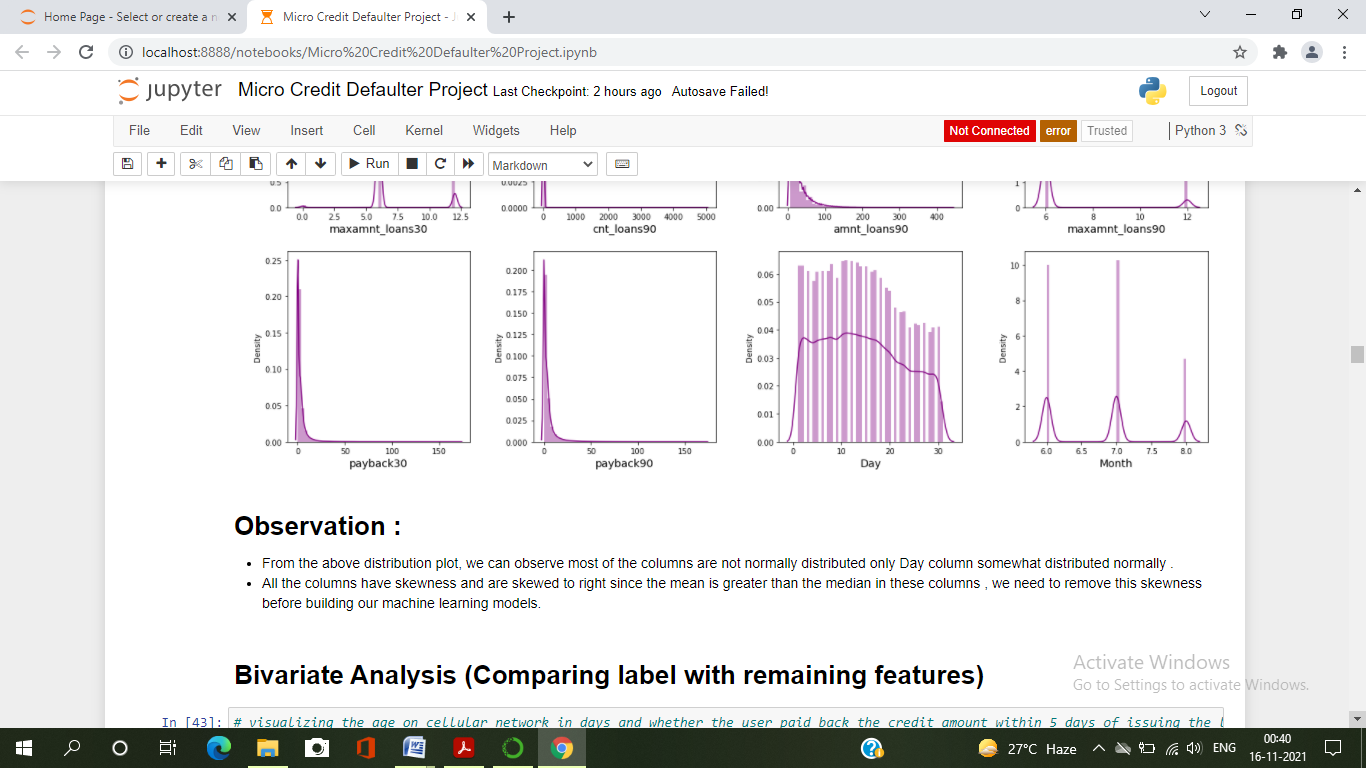
* From the above plots we can observe around 87% of the loan has been paid by the user and only 12% of the loan failed to pay . Also the dataset is highly imbalanced, so we need to work on that or else our model will be more biased towards success and make false interpretation.

Distribution plots to check skewness in numerical columns





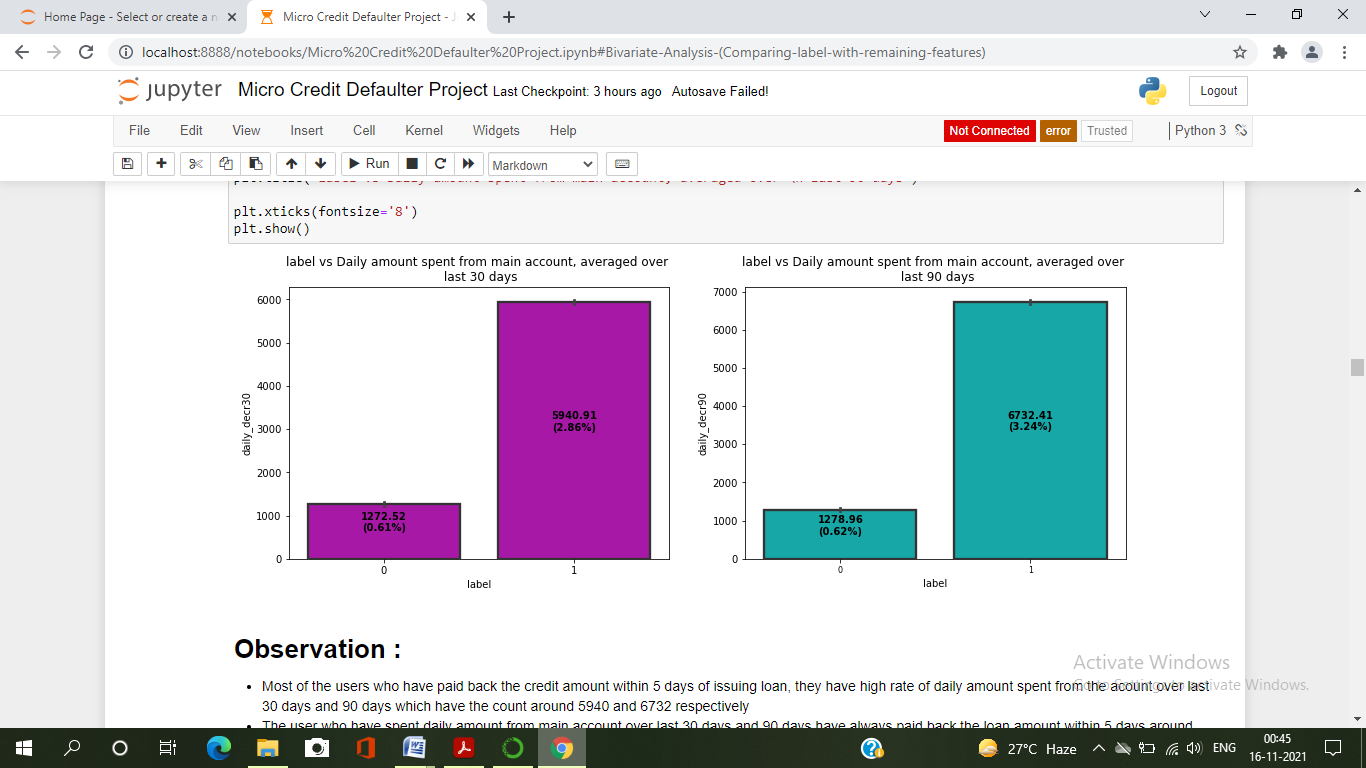




Observation :

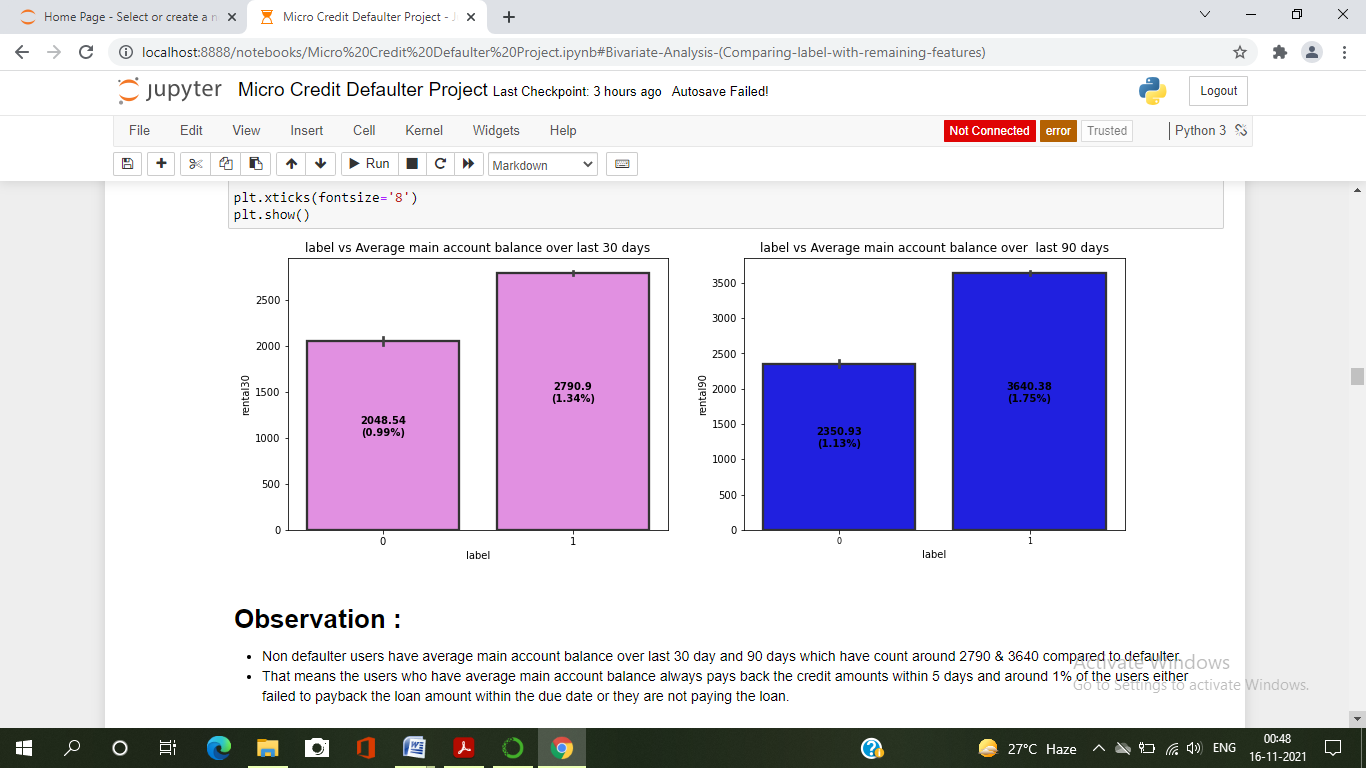
* From the above distribution plot, we can observe most of the columns are not normally distributed only Day column somewhat distributed normally .
* All the columns have skewness and are skewed to right since the mean is greater than the median in these columns , we need to remove this skewness before building our machine learning models.

# Bivariate Analysis (Comparing label with remaining features)



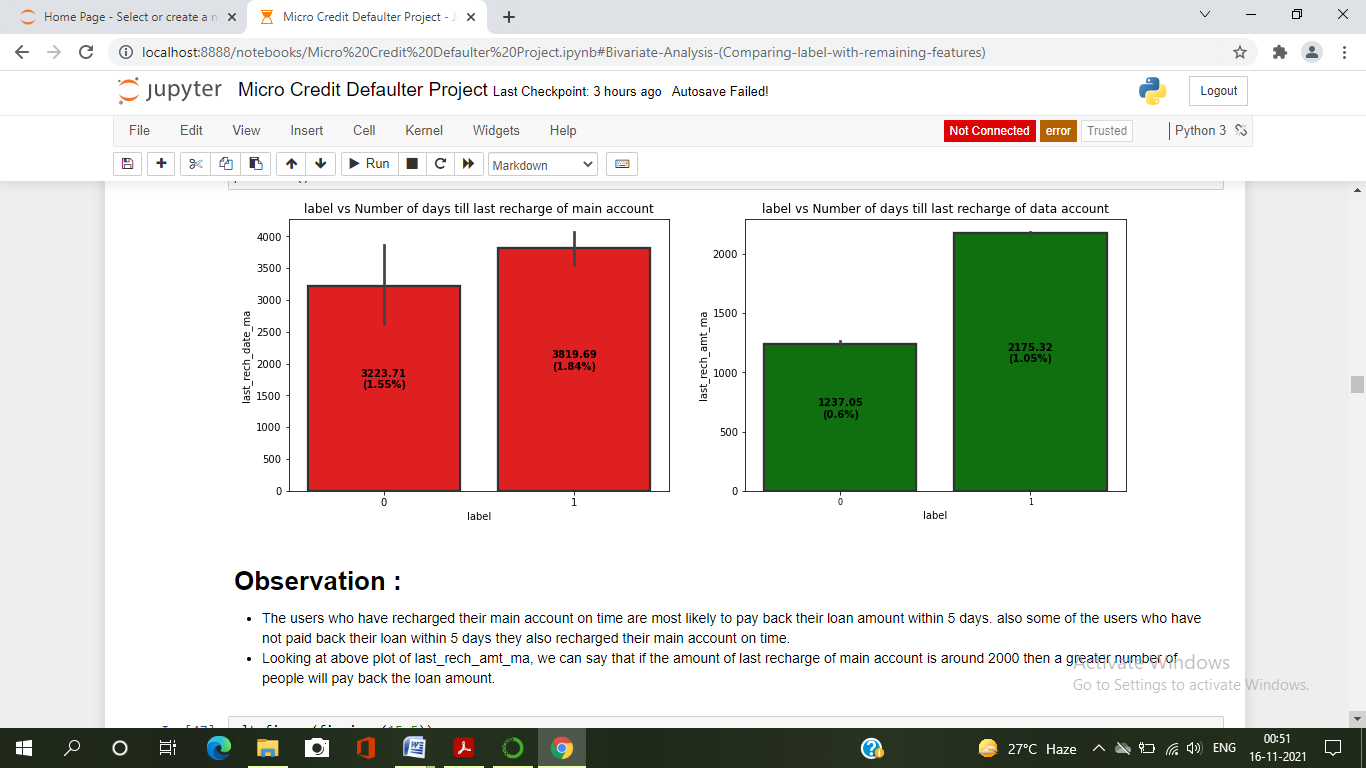
# Observation :

* Most of the users who have paid back the credit amount within 5 days of issuing loan, they have high rate of daily amount spent from the acount over last 30 days and 90 days which have the count around 5940 and 6732 respectively
* The user who have spent daily amount from main account over last 30 days and 90 days have always paid back the loan amount within 5 days around 0.6% of the users failed to pay back the loan within due date.



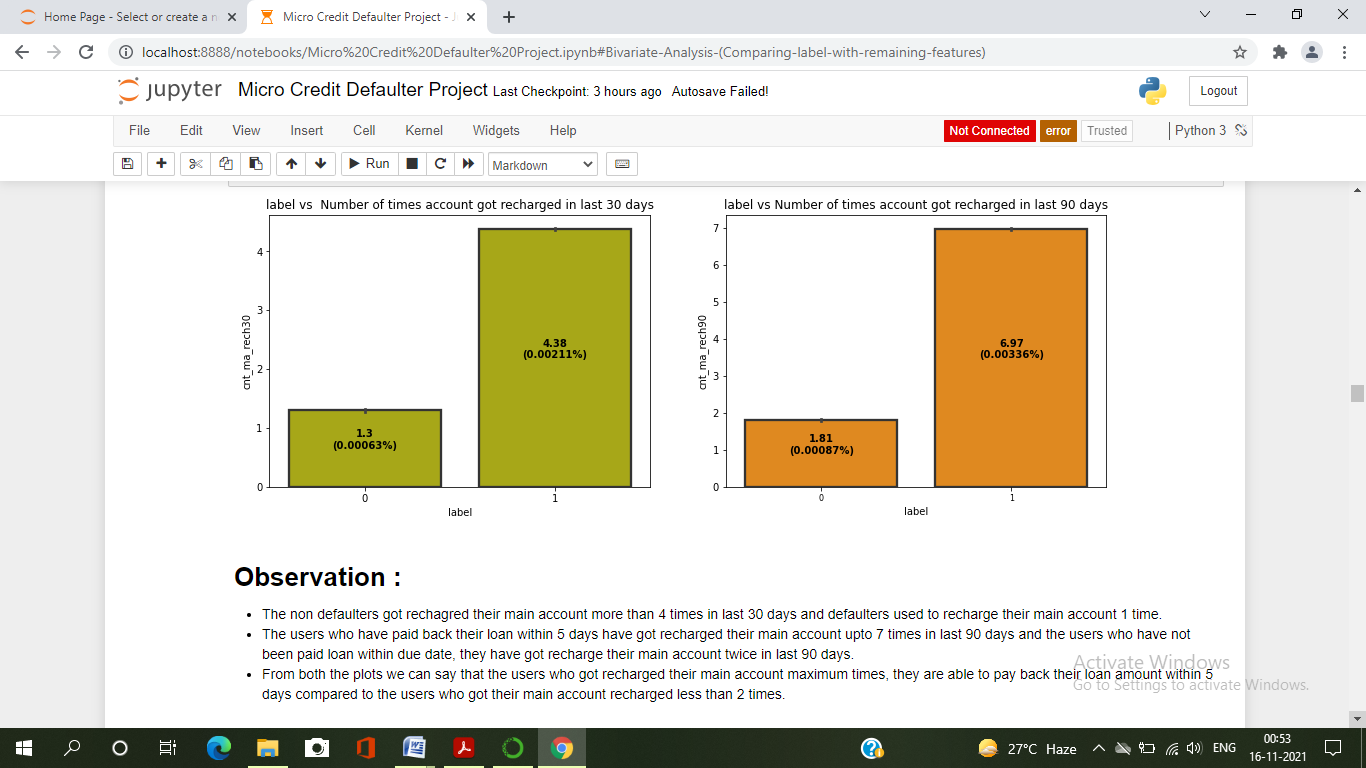
# Observation :

* Non defaulter users have average main account balance over last 30 day and 90 days which have count around 2790 & 3640 compared to defaulter.
* That means the users who have average main account balance always pays back the credit amounts within 5 days and around 1% of the users either failed to payback the loan amount within the due date or they are not paying the loan.



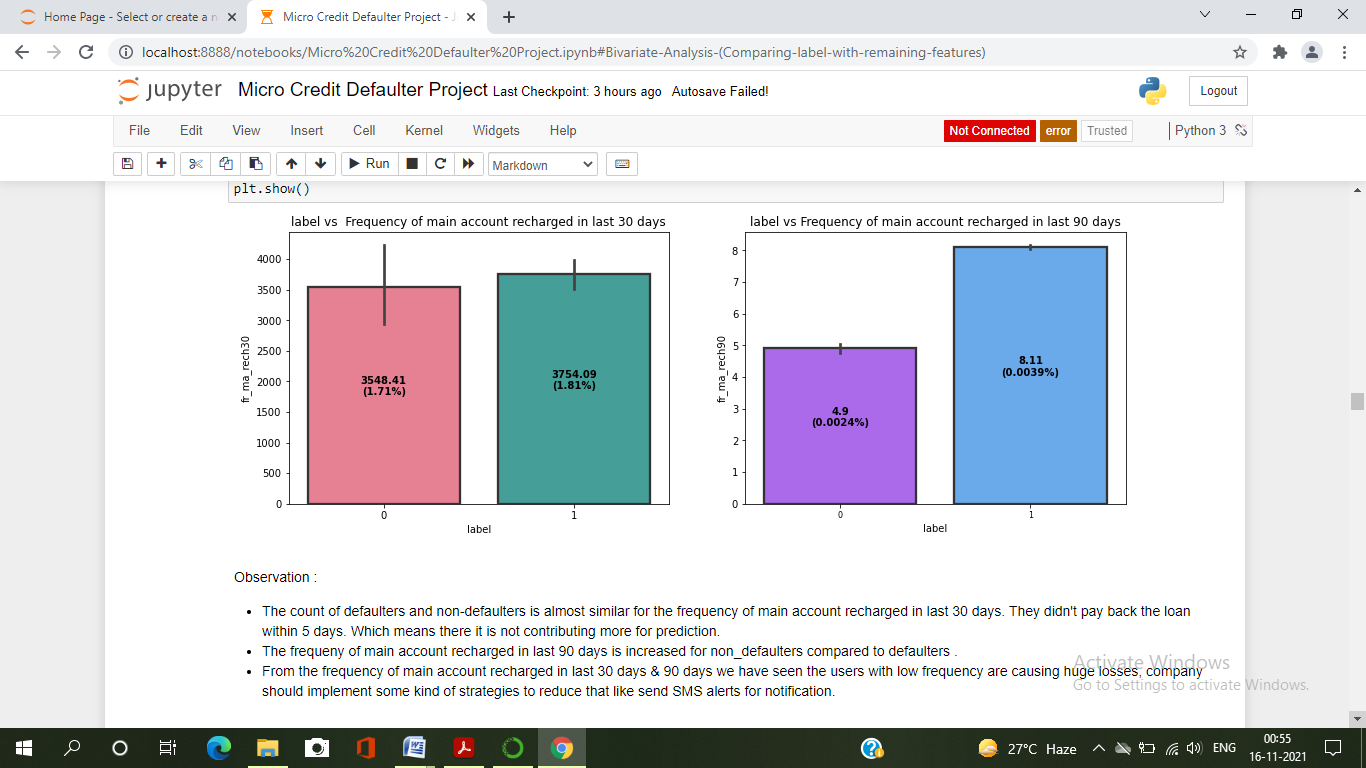
# Observation :

* The users who have recharged their main account on time are most likely to pay back their loan amount within 5 days. also some of the users who have not paid back their loan within 5 days they also recharged their main account on time.
* Looking at above plot of last\_rech\_amt\_ma, we can say that if the amount of last recharge of main account is around 2000 then a greater number of people will pay back the loan amount.



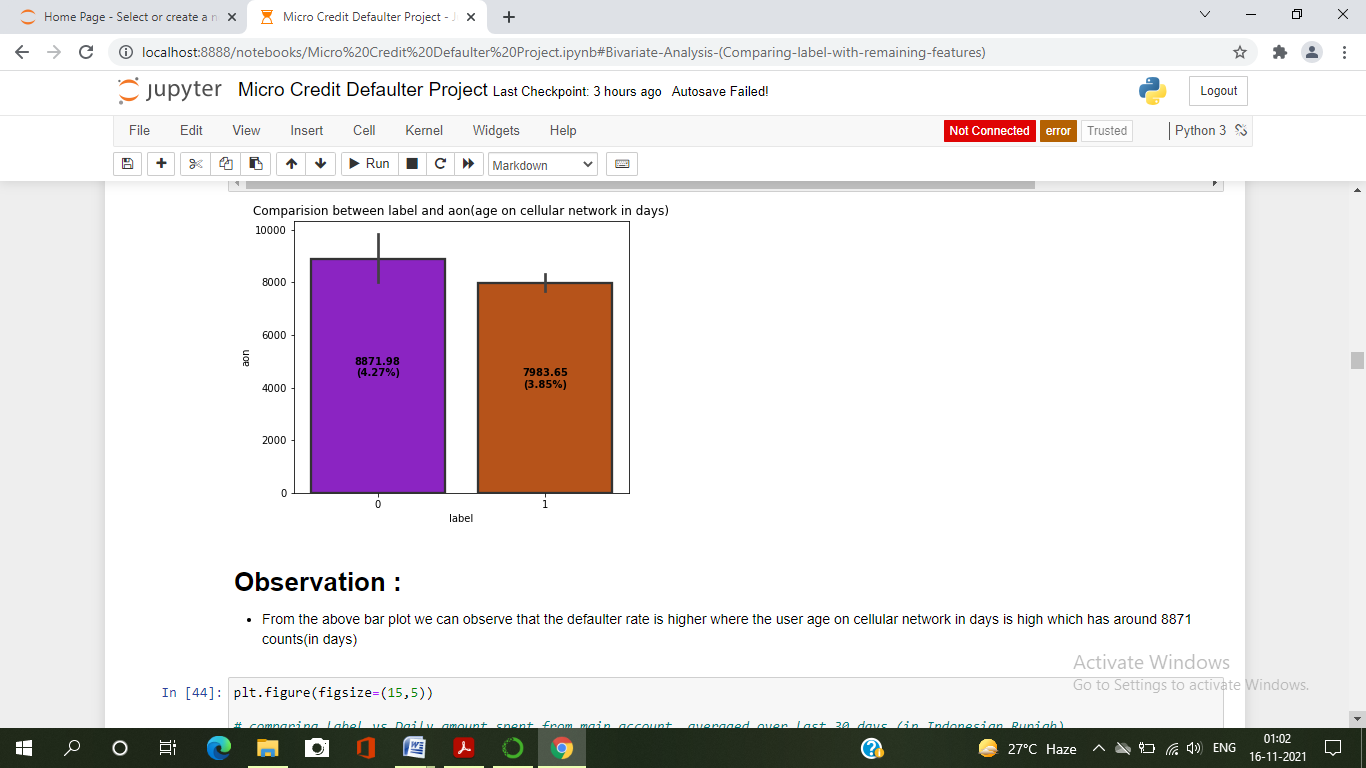
# Observation :

* The non defaulters got rechagred their main account more than 4 times in last 30 days and defaulters used to recharge their main account 1 time.
* The users who have paid back their loan within 5 days have got recharged their main account upto 7 times in last 90 days and the users who have not been paid loan within due date, they have got recharge their main account twice in last 90 days.
* From both the plots we can say that the users who got recharged their main account maximum times, they are able to pay back their loan amount within 5 days compared to the users who got their main account recharged less than 2 times.



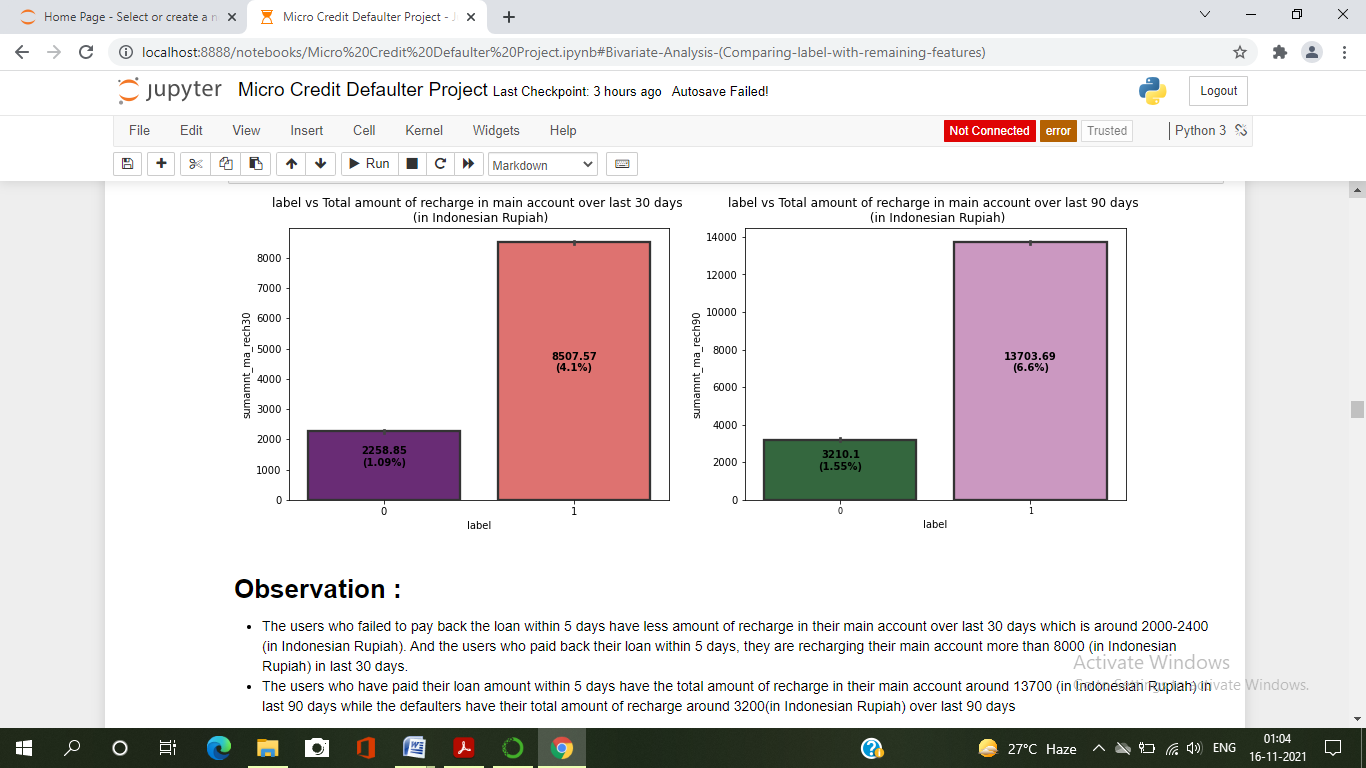
Observation :

* The count of defaulters and non-defaulters is almost similar for the frequency of main account recharged in last 30 days. They didn't pay back the loan within 5 days. Which means there it is not contributing more for prediction.
* The frequeny of main account recharged in last 90 days is increased for non\_defaulters compared to defaulters .
* From the frequency of main account recharged in last 30 days & 90 days we have seen the users with low frequency are causing huge losses, company should implement some kind of strategies to reduce that like send SMS alerts for notification.



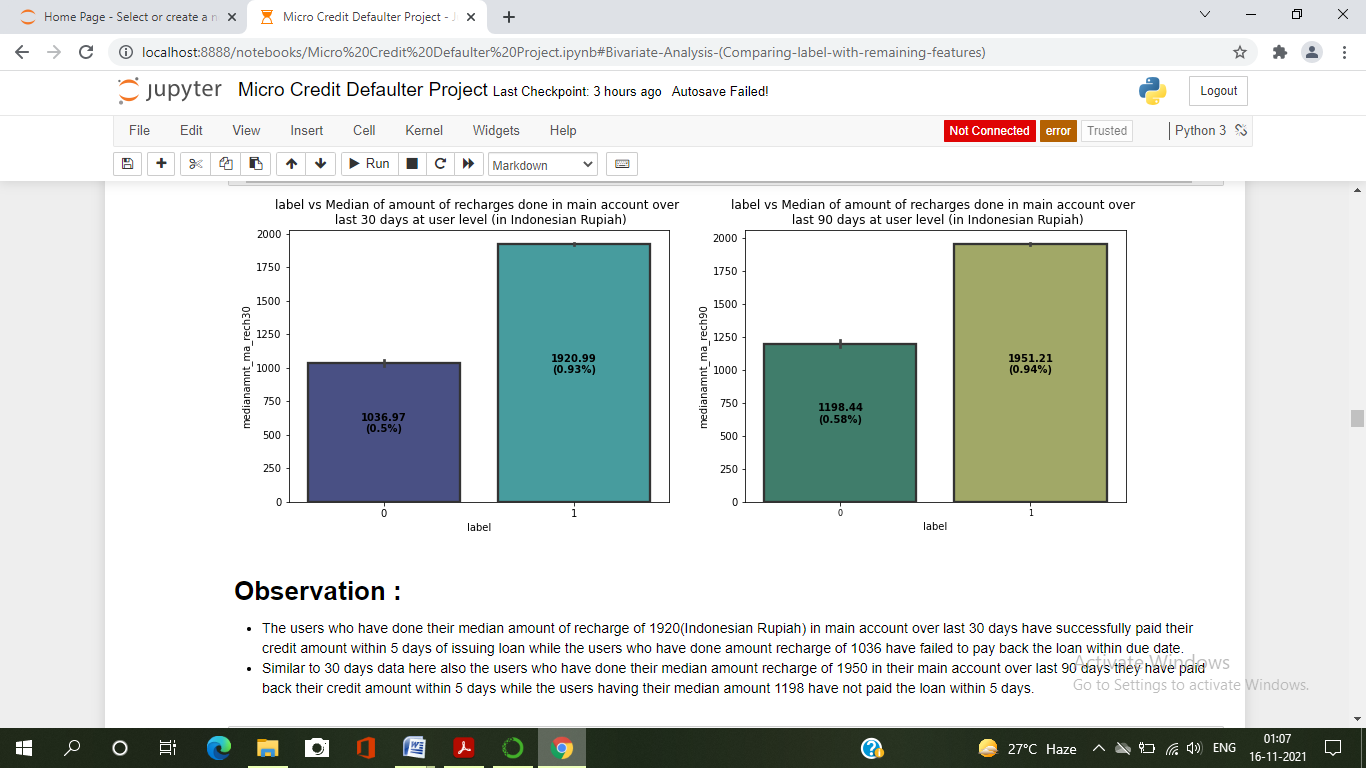
# Observation :

* From the above bar plot we can observe that the defaulter rate is higher where the user age on cellular network in days is high which has around 8871 counts(in days)



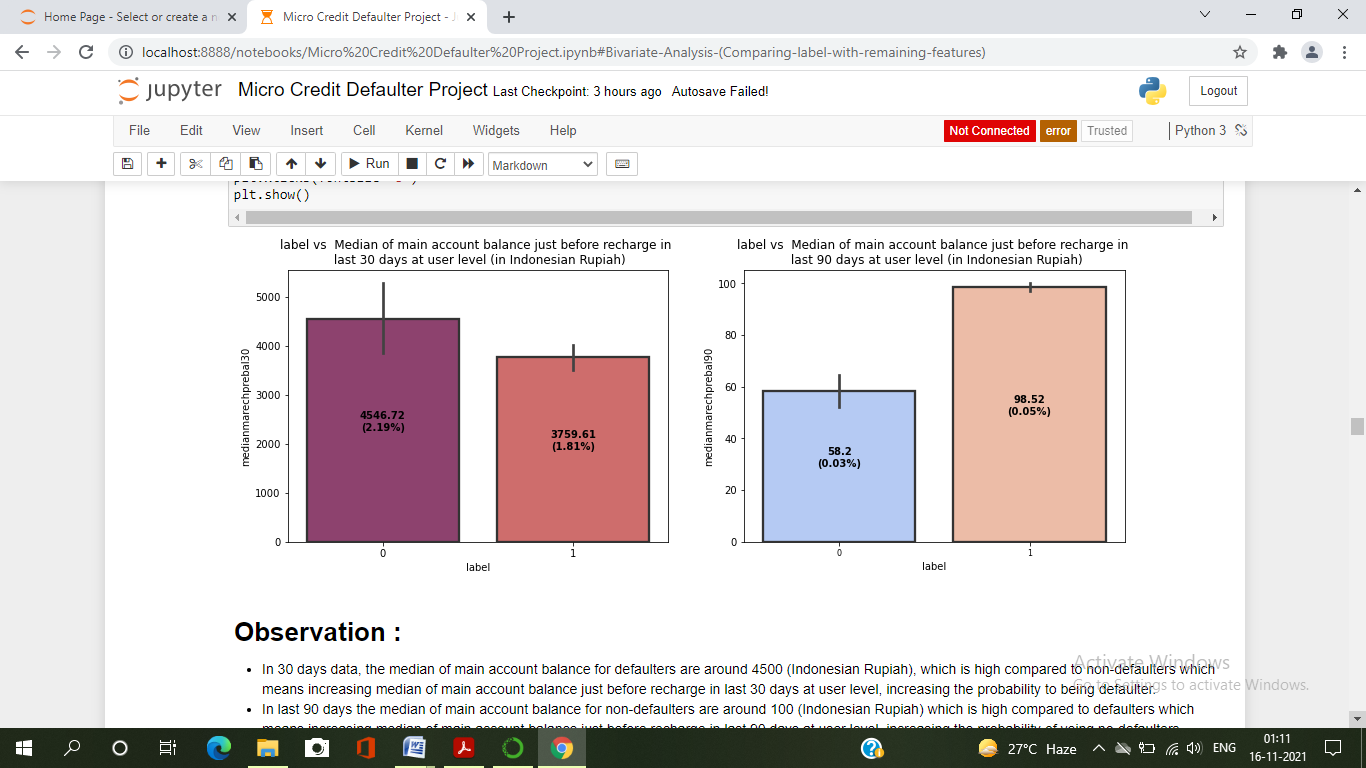
# Observation :

* The users who failed to pay back the loan within 5 days have less amount of recharge in their main account over last 30 days which is around 2000-2400 (in Indonesian Rupiah). And the users who paid back their loan within 5 days, they are recharging their main account more than 8000 (in Indonesian Rupiah) in last 30 days.
* The users who have paid their loan amount within 5 days have the total amount of recharge in their main account around 13700 (in Indonesian Rupiah) in last 90 days while the defaulters have their total amount of recharge around 3200(in Indonesian Rupiah) over last 90 days.



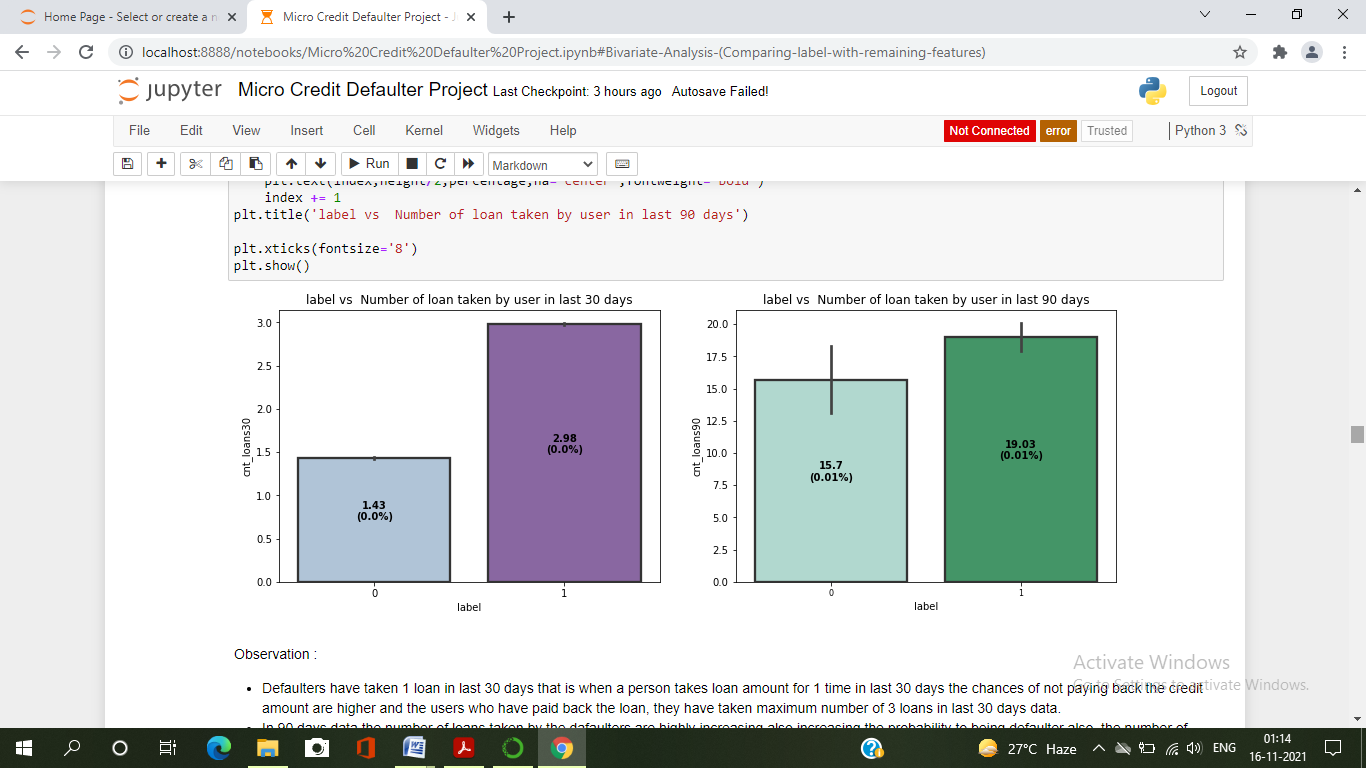
# Observation :

* The users who have done their median amount of recharge of 1920(Indonesian Rupiah) in main account over last 30 days have successfully paid their credit amount within 5 days of issuing loan while the users who have done amount recharge of 1036 have failed to pay back the loan within due date.
* Similar to 30 days data here also the users who have done their median amount recharge of 1950 in their main account over last 90 days they have paid back their credit amount within 5 days while the users having their median amount 1198 have not paid the loan within 5 days.



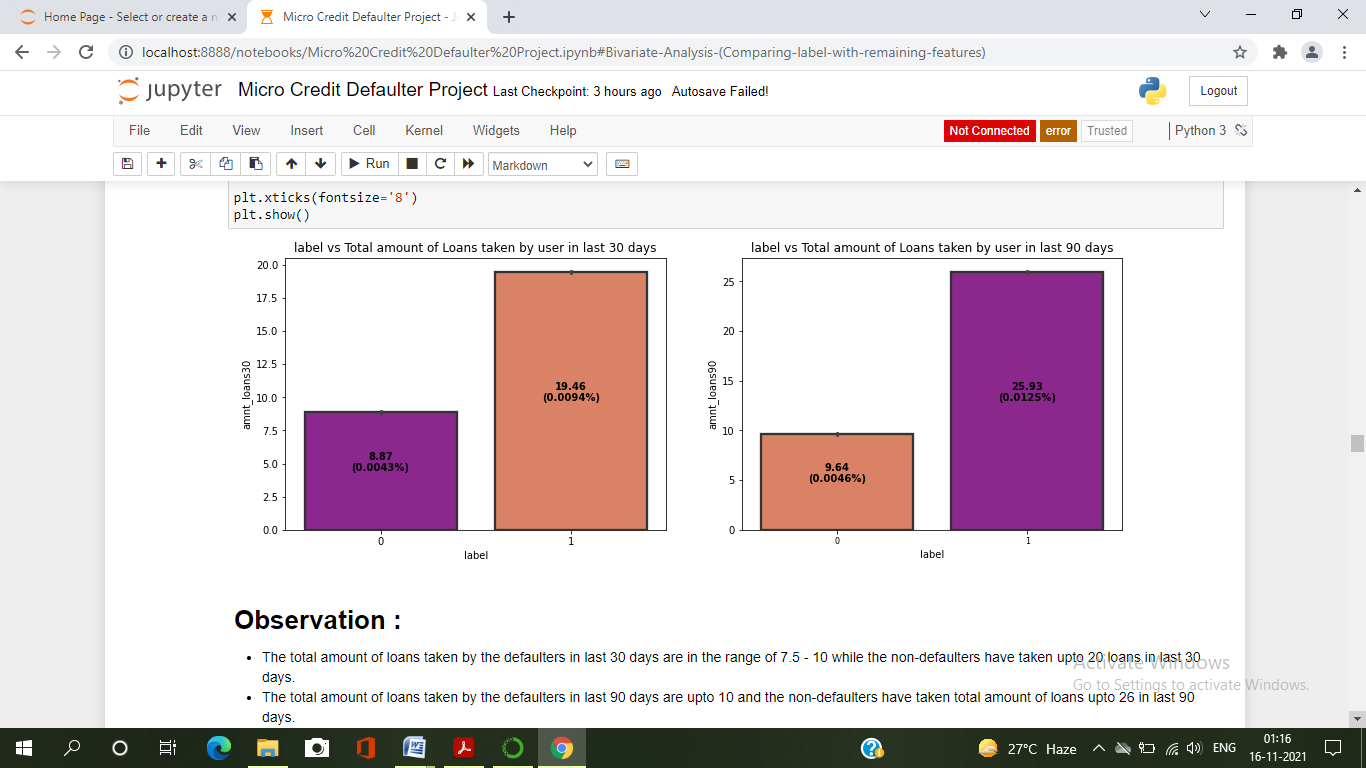
# Observation :

* In 30 days data, the median of main account balance for defaulters are around 4500 (Indonesian Rupiah), which is high compared to non-defaulters which means increasing median of main account balance just before recharge in last 30 days at user level, increasing the probability to being defaulter.
* In last 90 days the median of main account balance for non-defaulters are around 100 (Indonesian Rupiah) which is high compared to defaulters which means increasing median of main account balance just before recharge in last 90 days at user level, increasing the probability of being no-defaulters.



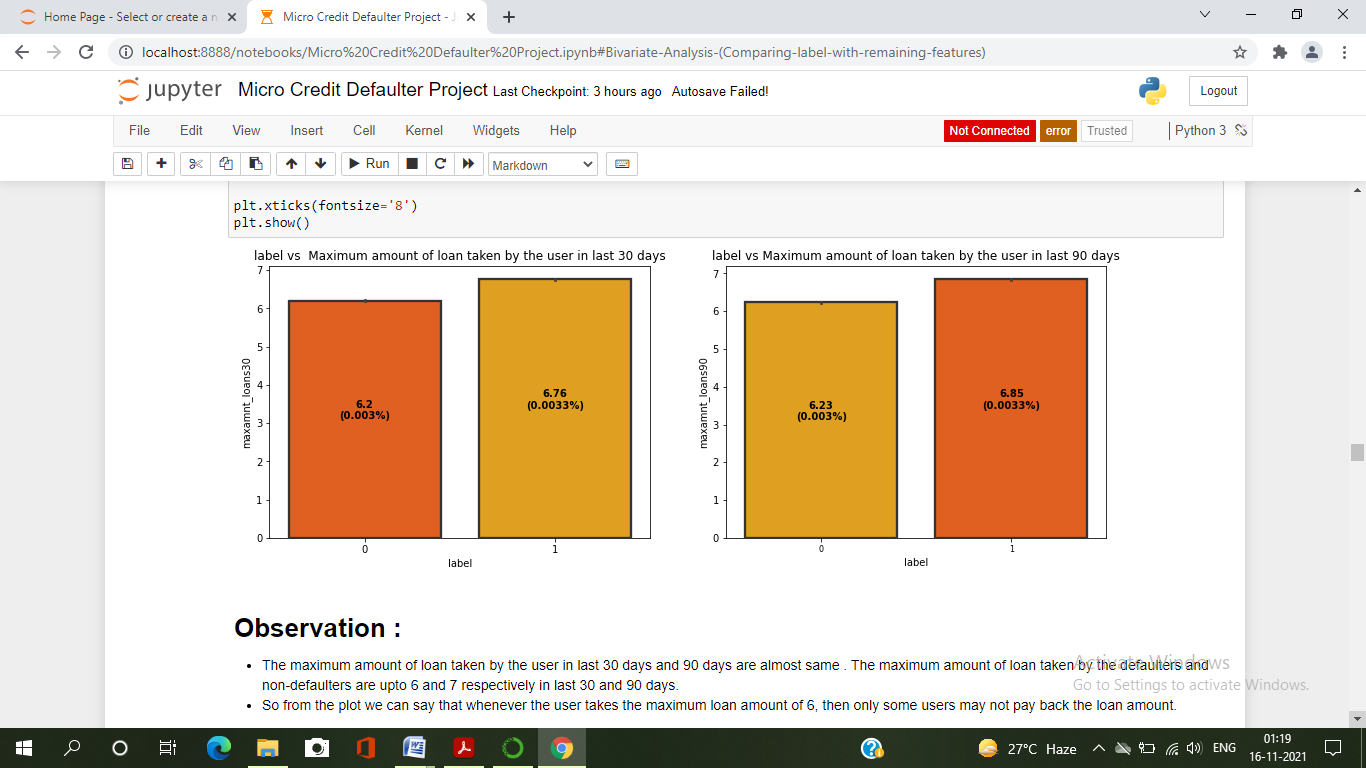
Observation :

* Defaulters have taken 1 loan in last 30 days that is when a person takes loan amount for 1 time in last 30 days the chances of not paying back the credit amount are higher and the users who have paid back the loan, they have taken maximum number of 3 loans in last 30 days data.
* In 90 days data the number of loans taken by the dafaulters are highly increasing also increasing the probability to being defaulter also, the number of loans taken by non-defaulters being decreased in last 90 days when compared to 30 days data.



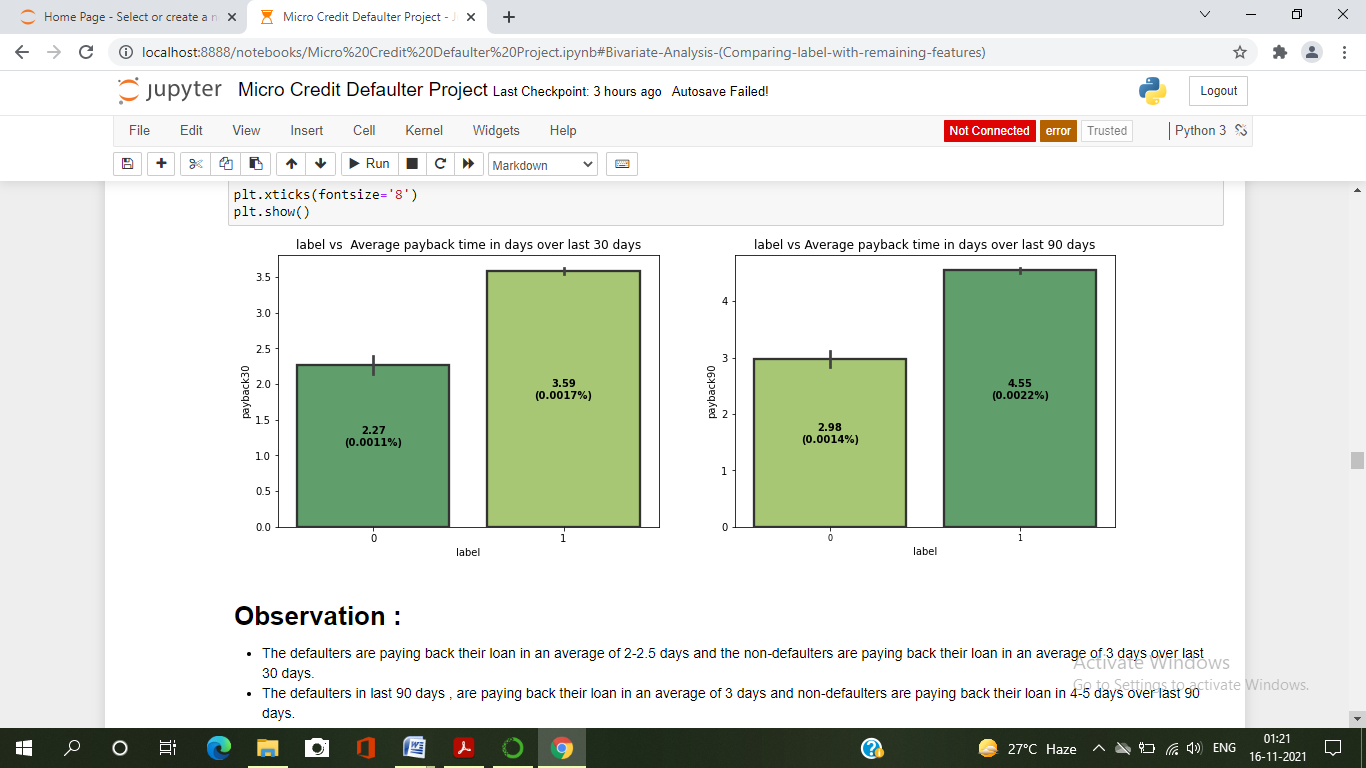
# Observation :

* The total amount of loans taken by the defaulters in last 30 days are in the range of 7.5 - 10 while the non-defaulters have taken upto 20 loans in last 30 days.
* The total amount of loans taken by the defaulters in last 90 days are upto 10 and the non-defaulters have taken total amount of loans upto 26 in last 90 days.
* So, from the above plot we can conclude that when the total number of loans taken by the users in last 90 days is below 10 then the chances of not paying back the loan amount are high.



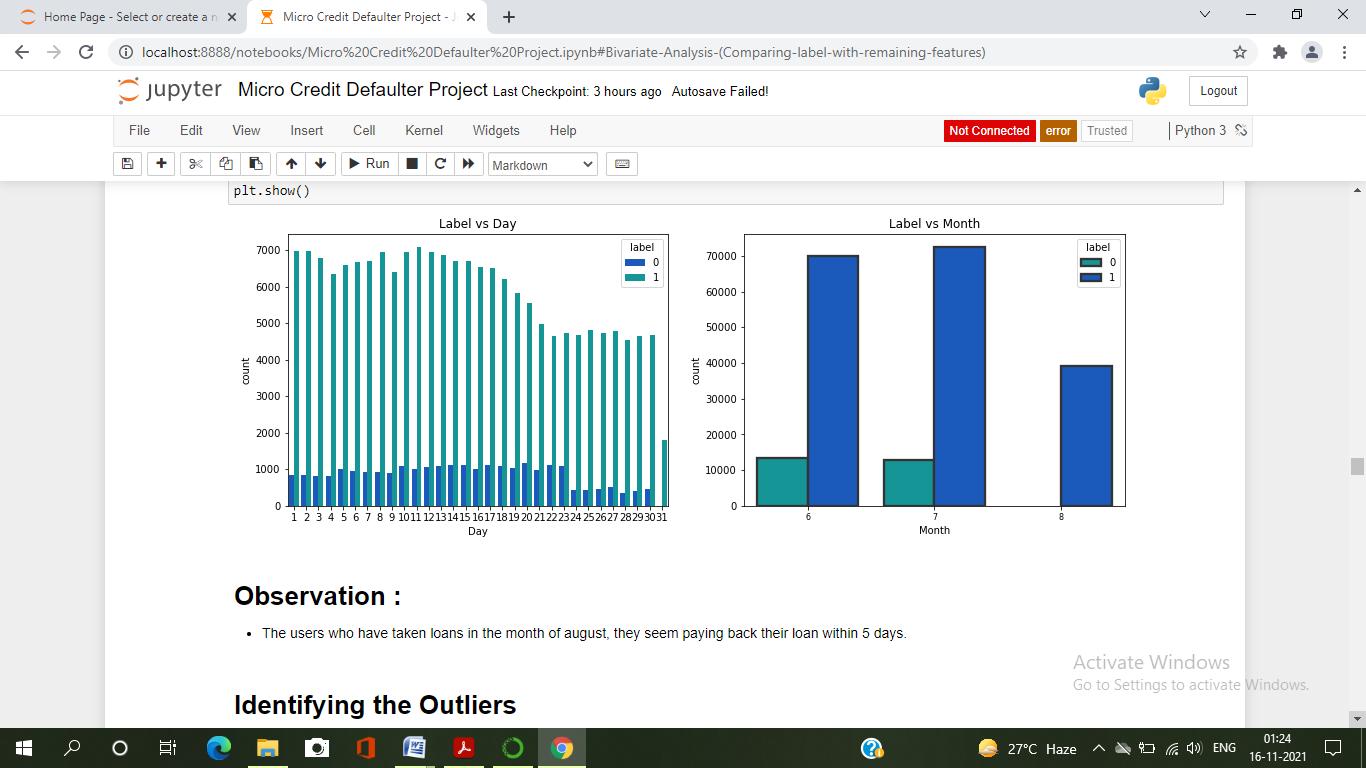
# Observation :

* The maximum amount of loan taken by the user in last 30 days and 90 days are almost same . The maximum amount of loan taken by the defaulters and non-defaulters are upto 6 and 7 respectively in last 30 and 90 days.
* So from the plot we can say that whenever the user takes the maximum loan amount of 6, then only some users may not pay back the loan amount.



# Observation :

* The defaulters are paying back their loan in an average of 2-2.5 days and the non-defaulters are paying back their loan in an average of 3 days over last 30 days.
* The defaulters in last 90 days , are paying back their loan in an average of 3 days and non-defaulters are paying back their loan in 4-5 days over last 90 days.
* It is seem from the plot that when an average payback time is below 3 days over last 30 & 90 days, then defaulters rate is high.



# Observation :

* The users who have taken loans in the month of august, they seem paying back their loan within 5 days.

Interpretation of the Result

Visualizations:

I have used distribution plot to visualize the numerical variables. Used bar plots to check the relation between label and the features. The heat map and bar plot helped me to understand theCorrelation between dependent and independent features. Also , heat map helped to detect the multicollinearity problem and feature importance . Detected outliers and skewness with the help of box plots and distribution plots respectively. And I found some of the features skewed to right . I got to know the count of each column using bar plots.

Pre-processing: The dataset should be cleaned and scaled to build the ML models to get good predictions. I have performed few processing steps which I have already mentioned in the pre-processing steps where all the important features are present in the dataset and ready for model building.

Model Building : After cleaning and processing data, I performed train test split to build the model. I have built multiple classification models to get the accurate accuracy score, and evaluation metrics like precision

Recall, confusion matrix, f1 score. I got Gradient Boosting Classifier as best Model which gives 89% accuracy score. I checked the cross- validation score ensuring there will be no overfitting. After tuning the best model Gradient Boosting Classifier, I got 95% accuracy score and also got increment in AUC-ROC curve . Finally, I saved my final model and got the good predictions results for defaulters.

* We can note that there is less data about defaulters and more about those who did repay their loan. Hence can say that the data is imbalanced
* With increase in Age on Network, defaulting rate is higher.
* The data is collected based on different parameters for two time periods. One observation is for 30 days and other is for 90 days. Analyzing the parameters separately
* For 30 Days :
  1. With the increase in Average Main balance, there is a probability of defaulting.
  2. Defaulters recharged Main account max number between 1 and 2 times. Whereas re-payers recharged for 4 plus times.
  3. On an average the defaulters have recharged for a max of 1000 Indonesian Rupiah for Main Balance.
  4. Defaulters has recharged the data account for a maximum of 200 to 250 times. With increase in No. of times data accounts recharge, probability of defaulting is high.
  5. A defaulter may default after 2 days, re payers took average of 3.5 days.
  6. Defaulters took 1 loan, re payers took 3 loans.
* For 90 Days:

1. The defaulters have spent a max of 1000 from main account, Repayers has spent 7000 rupiah.
2. Defaulters average main account balance=2000 to 2500, Repayers average main account balance=3500
3. Defaulters recharged main account for 2 times,

Re-payers recharged main account for 7 times.

1. Defaulters frequency of main account recharge is 5,

Re-payers frequency of main account recharge is 8.

Conclusion

* Key Finding and Conclusions of the Study :

The defaulting rate is higher in old customers. Defaulters recharge for the main account less no. of times but does recharge for data account more no. of times.

Re payers recharge the main account more no. of times when compared to defaulters.

* Learning Outcomes of the Study in respect of Data Science

One of the challenges I faced while data cleaning is outlier removal, in most of the scenarios Z-score will be used as outlier removal technique since it performs quite well with less data loss. In our data set, Z-score has caused 19% data loss. Then I tried another famous technique called Inter Quartile Range(IQR) it caused around 62% data loss.

Another technique is replacing the outlier data with mean or median. But when we observe this data set there is a huge difference between minimum and maximum values. If we calculate mean or median it won’t give appropriate values as it includes the outlier value(maximum ones). So not using this approach.

As we are not dropping the outliers, another approach is capping or winsorization of outliers. Using percentile capping. Values that are less than the value at 10th percentile are replaced by 10th percentile value, and the value greater than 90th percentile are replaced by 90th percentile value.

The other challenge is when I used the imblanced data, the accuracy was very high but there was bias in predictions. So I used imblearn to reduce the imbalances in the target variable.

* Limitations of this work and Scope for Future Work:

This data set contains data of the year 2016 belonging to PSW telecom circle.

If we get data of other years along with other telecom companies we can predict on varied scenario.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*