

**Micro – credit Defaulter**

**Model**

Submitted by:

Asheem Siwach

**ACKNOWLEDGMENT**

I would like to thank Flip Robo Technologies for providing me with the opportunity to work on this project from which I have learned a lot. I am also grateful to Mr. Shubham Yadav for his constant guidance and support.

Reference sources are :-

* Google
* Stackoverflow.com
* Notes and repository from DataTrained
* Krish Naik youtube videos

**INTRODUCTION**

* Business Problem Framing

This project includes the real time problem for Microfinance Institution (MFI) offering financial services to low income population. Mobile financial services (MFS) become very useful when targeting the unbanked poor families living in remote areas with negligible sources of income, MFI provides 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

A microfinance institution is an organization that offers financial services to low income populations. Almost all give loans to their members, and many offer insurance, deposit and other services. A great scale of organizations is regarded as microfinance institutes. They are those that offer credits and other financial services to the representatives of poor strata of population (except for extremely poor strata). 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.

Today, microfinance is widely accepted as a poverty-reduction tool, representing $70 billion in outstanding loans and a global outreach of 200 million clients.

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).

We are going to build a machine learning 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.

* Review of Literature

We are going to work on the dataset provided for building a model to predict the nature of future customers –

1. Uploading the dataset
2. Exploratory analysis of features
3. Data -preprocessing, removing unnecessary features
4. Visualisaton of features, features along with target feature
5. Data cleaning by removing outliers or missing values treatment
6. Model pre-processing and model training
7. Tesing multiple algorithms with multiple evaluation metrics
8. Using the evaluation metrics best suited for our problem
9. Hyper-parameter tuning using RandomisedSearchCV for the best model parameter
10. Model prediction and saving the file in csv format
11. Saving the final model

* Motivation for the Problem Undertaken

This project was provided to me by FlipRobo Technology as a part of internship program . This dataset helped me a lot for understanding the the real time problems with data we are going to face. It helps me in developing the skills needed for the future work.

Further diving into the dataset and understanding the current situation of my country due to covid-19 , I was highly motivated to take up this project. Seeing the current covid situation, many people are facing health emergency .In this time they need to call ambulance,police, doctors ,relatives etc but due to lack of sufficient funds they may face calling problem. But this project is based on the solution of this problem ,as some Microfinance Institution are providing small credit loans to the poor familiers in remote areas.

The objective of this project is to prepare a model which can signify the difference in defaulters and non-defaulters and helps our client in further investment and selection of customers.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

The dataset provided for this project contains 209593 rows with 36 columns. Out of these 36 columns , one is target feature which is discrete in nature. This target feature consist of either 1 which represents non-defaulter and 0 which represents defaulter. Other features consists of aon(age on cellular network), msisdn(mobile number of user) , daily\_decr30 and daily\_decr90(daily amount spent from main account over last 30 days and 90 days average),rental30 and rental90(average main account balance over last 30 days and 90 days ), and some more features like these. Many of these features contains 30 days and 90 days average values which are highly correlated . But for the modelling we need to remove those features which are highly correlated .

Understanding the importance of data ,we all agree that the longer period the data, more better results it will give. As we have data of 30 days and 90 days in some feature and the problem of correlation is also there , in such case we should drop the 30 days data to prevent model from the problem of multicollinearity.

Using the describe function we will check the feature information. It will gives us mean,median, standard deviation, percentiles and count values from a numerical feature. If the mean and median of a feature are equal , it means that the data present is normally distributed. But in our dataset almost all the features are having either mean greater or median greater than each other, which indicated that skewness is present in data. Further the standard deviation represents the distribution of data around the mean values. If any feature having standard deviation high ,it shows the variance in the data of feature. While using the percentile ,we can find out the outliers in features.

Afterr checking the describe function we find out that there are a lot of outliers in our dataset and the distribution of data is not normally distributed.

Using the correlation graph ,we find the presence of correlation in independent and dependent features. Some of features are having 0% correlation with independent and dependent feature ,we will remove these features in data-preprocessing while some features with high correlation like features containing 30 days and 90 days data. In these feature we will remove one feature.

* Data Sources and their formats:-

Data Scource- Dataset was provided by FllipRobo Technologies in CSV(comma separated format) which includes a record of 209593 rows and 36 columns.

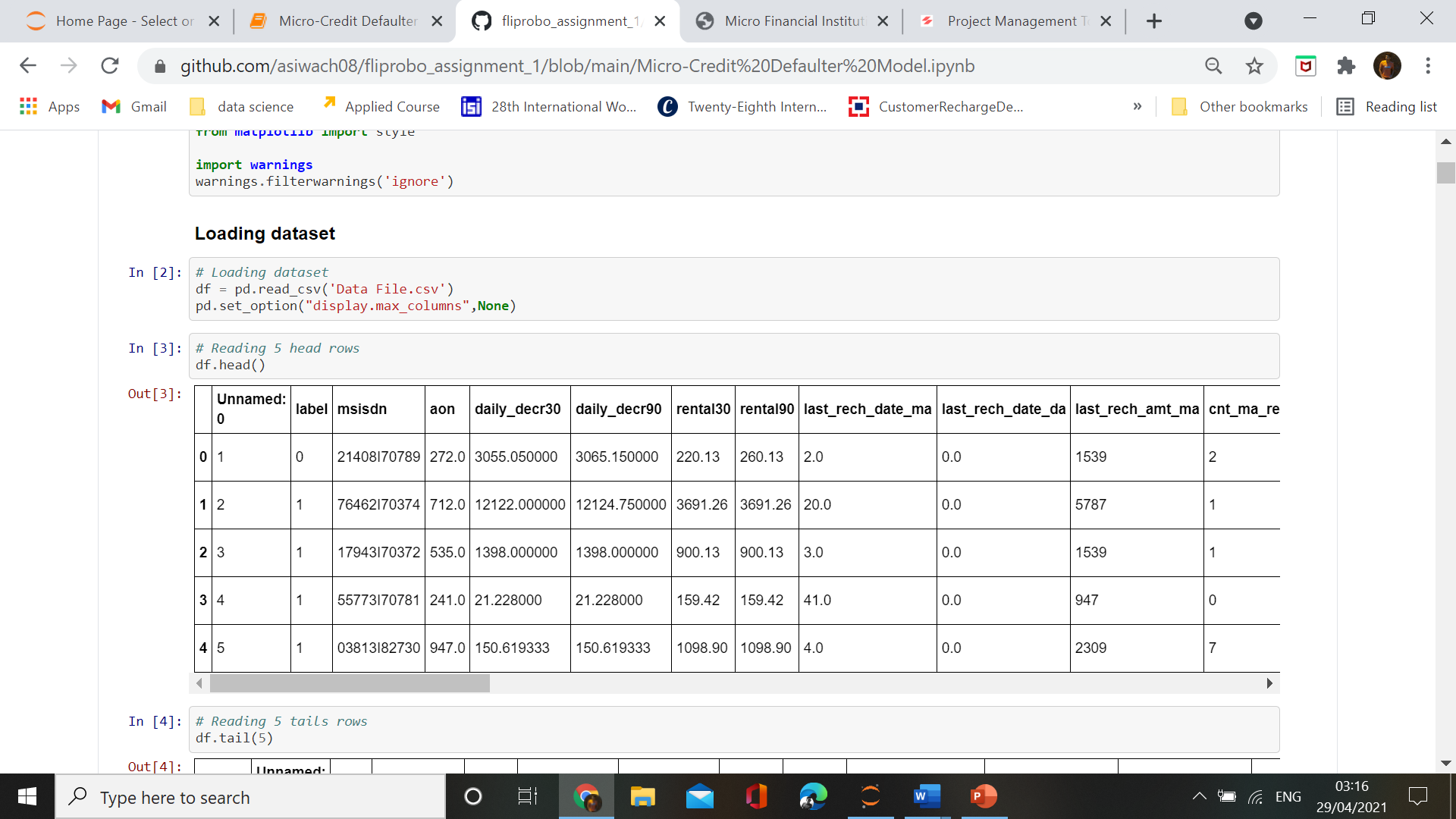
Data Description is as follows:-

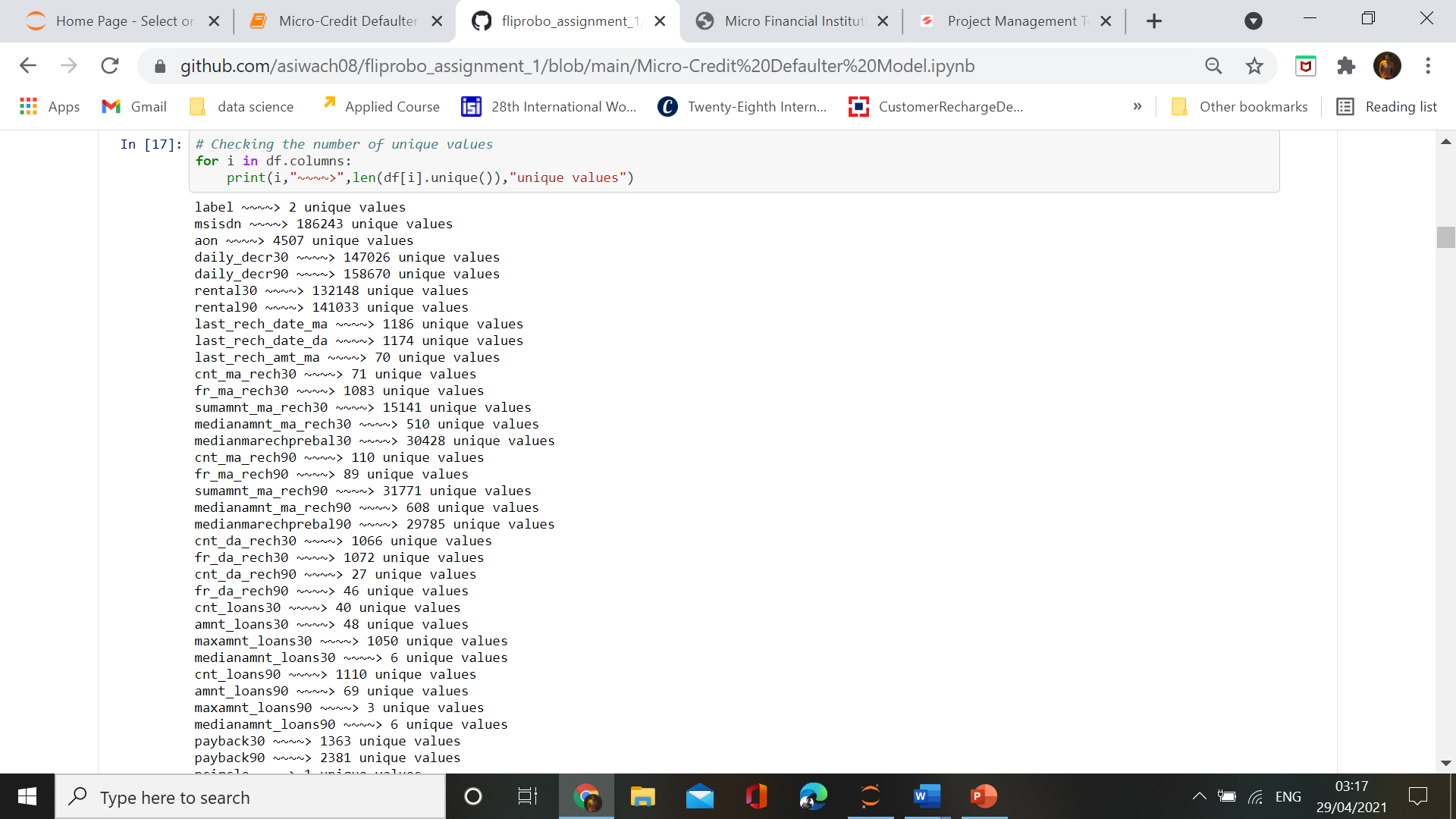
|  |  |
| --- | --- |
| Variable | Definition |
| label | Flag indicating whether the user paid back the credit amount within 5 days of issuing the loan{1:success, 0:failure} |
| msisdn | mobile number of user |
| aon | age on cellular network in days |
| daily\_decr30 | Daily amount spent from main 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 |
| rental90 | 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 |

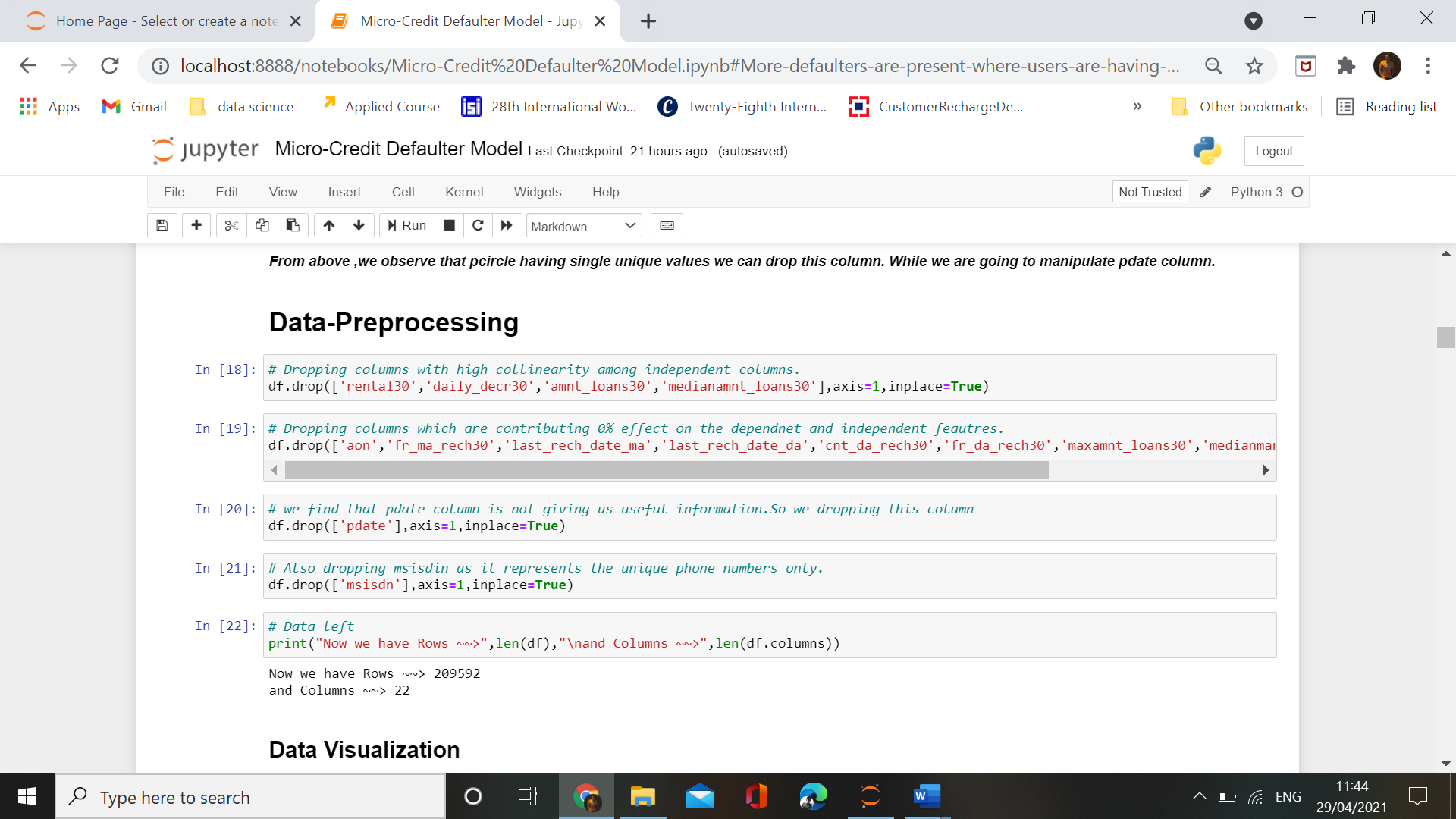
**Data Format –**



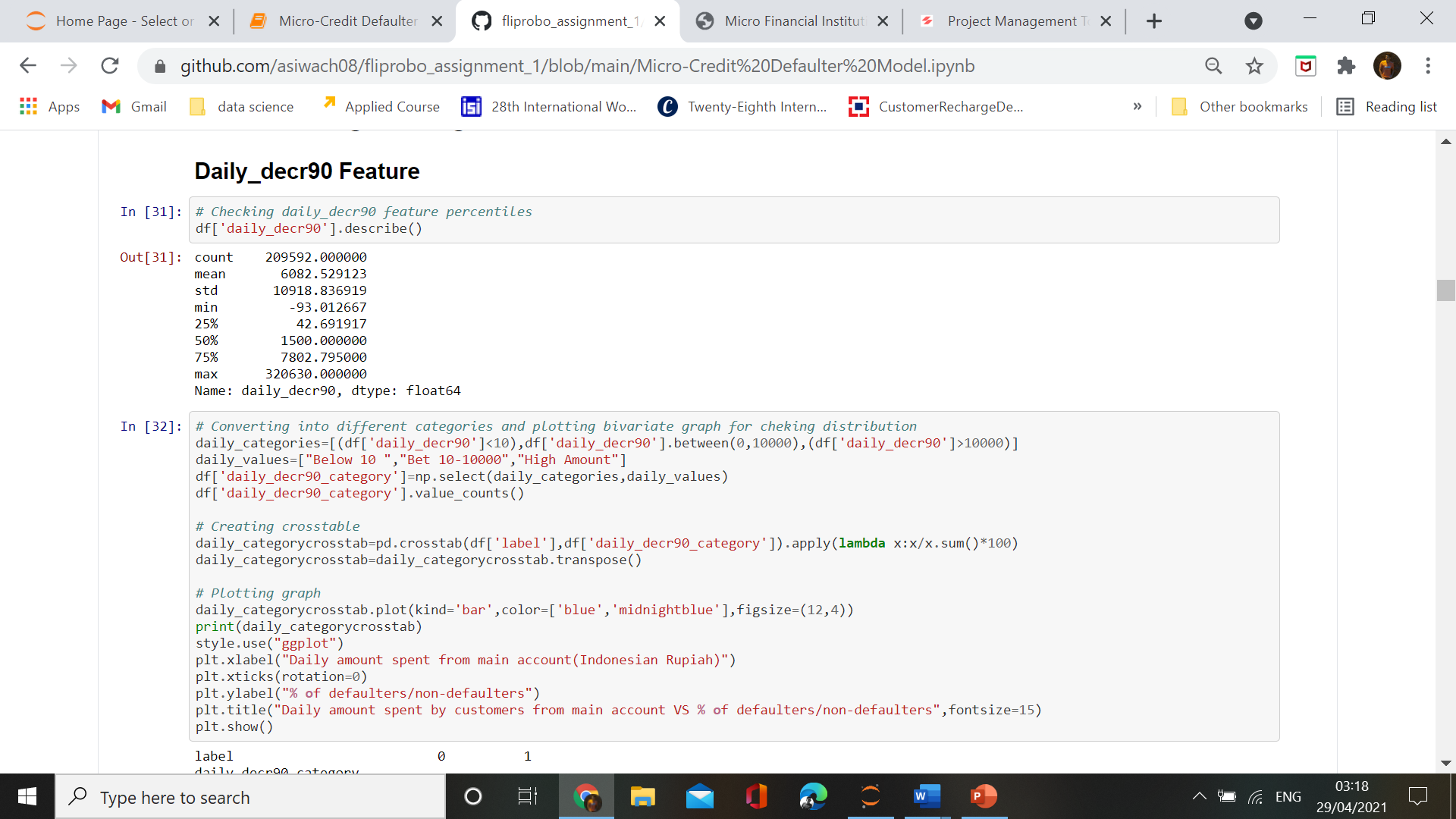
* Data Preprocessing Done



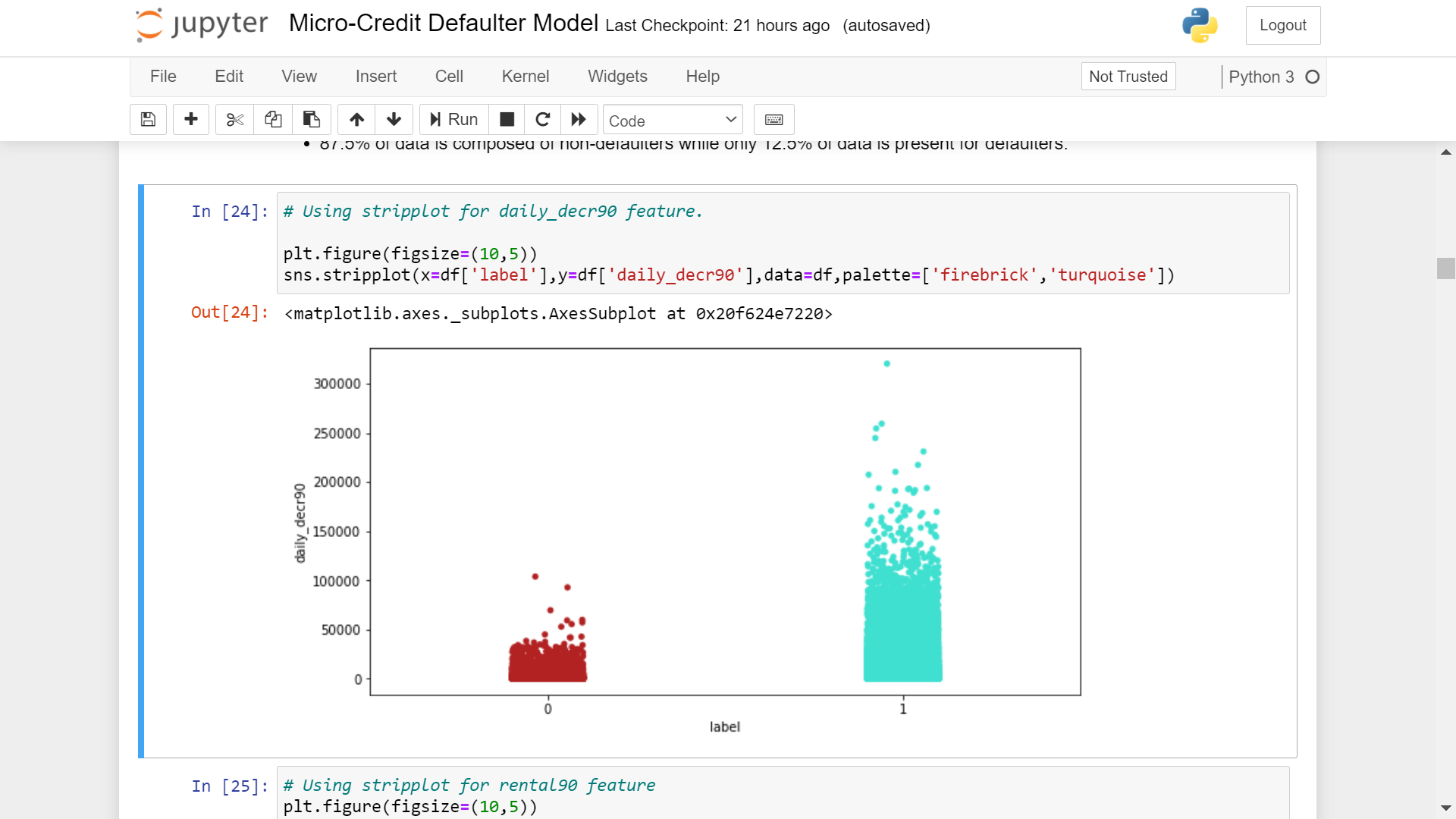


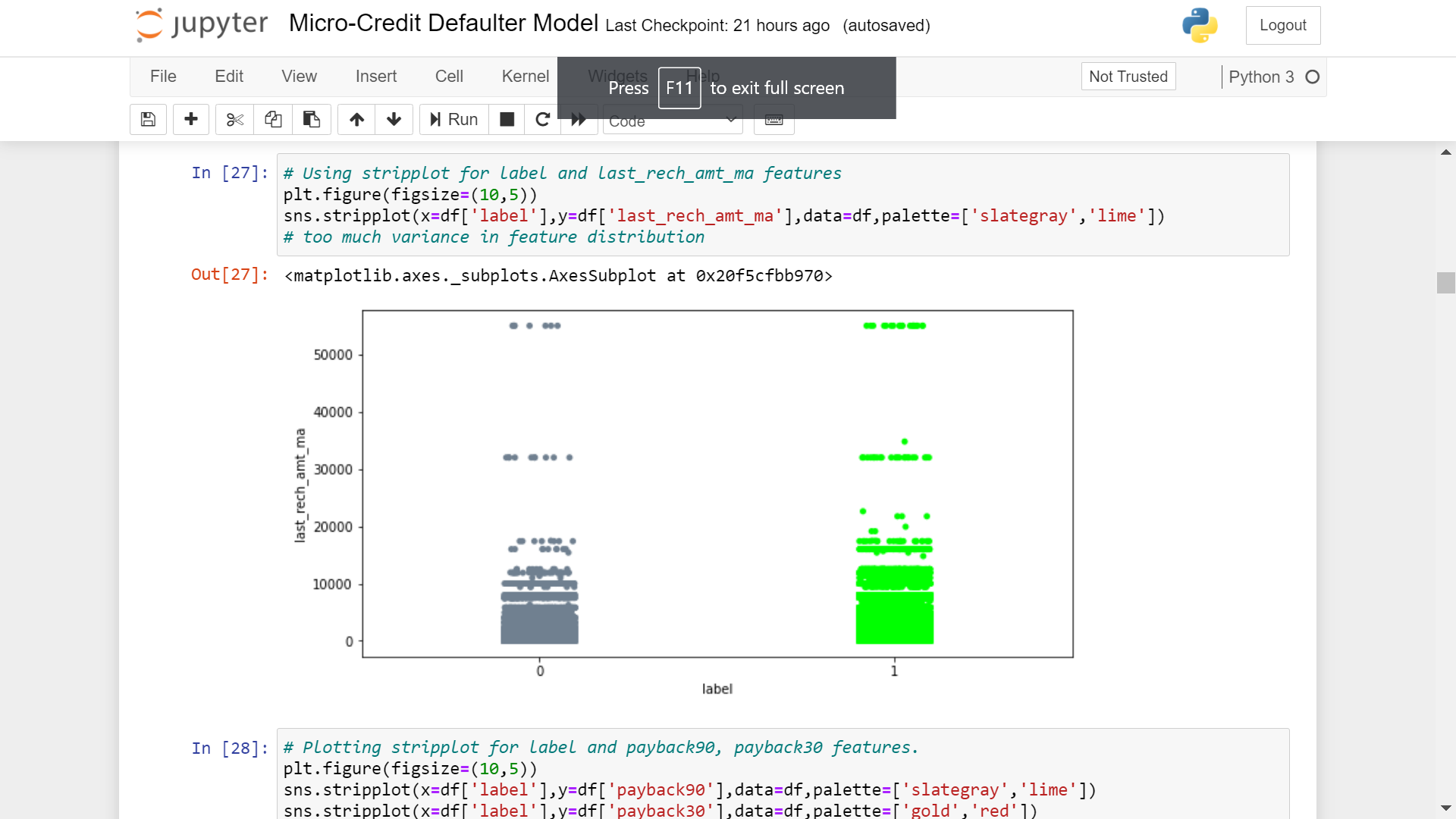


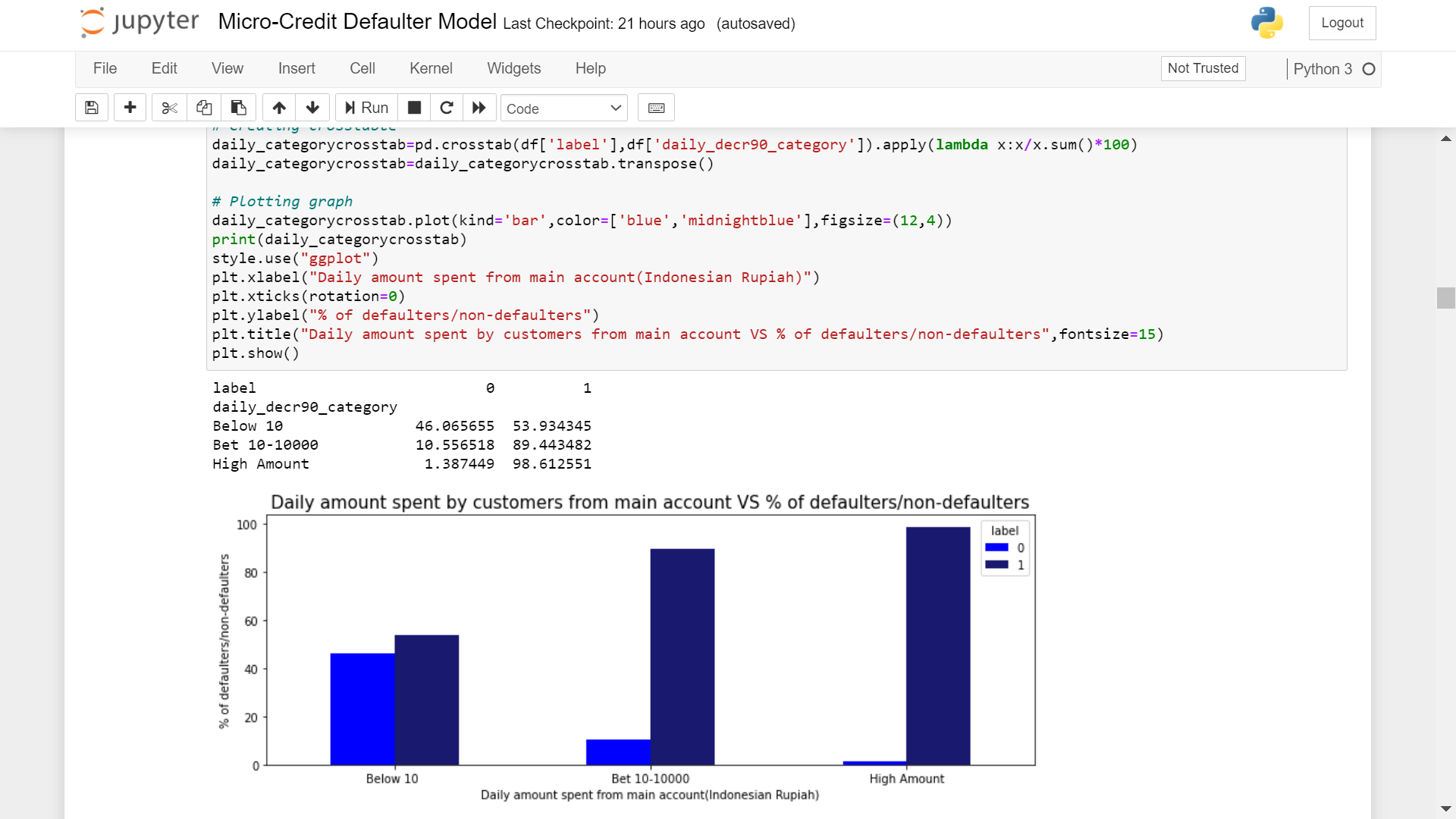
**Feature- engineering**



* Data Inputs- Logic- Output Relationships







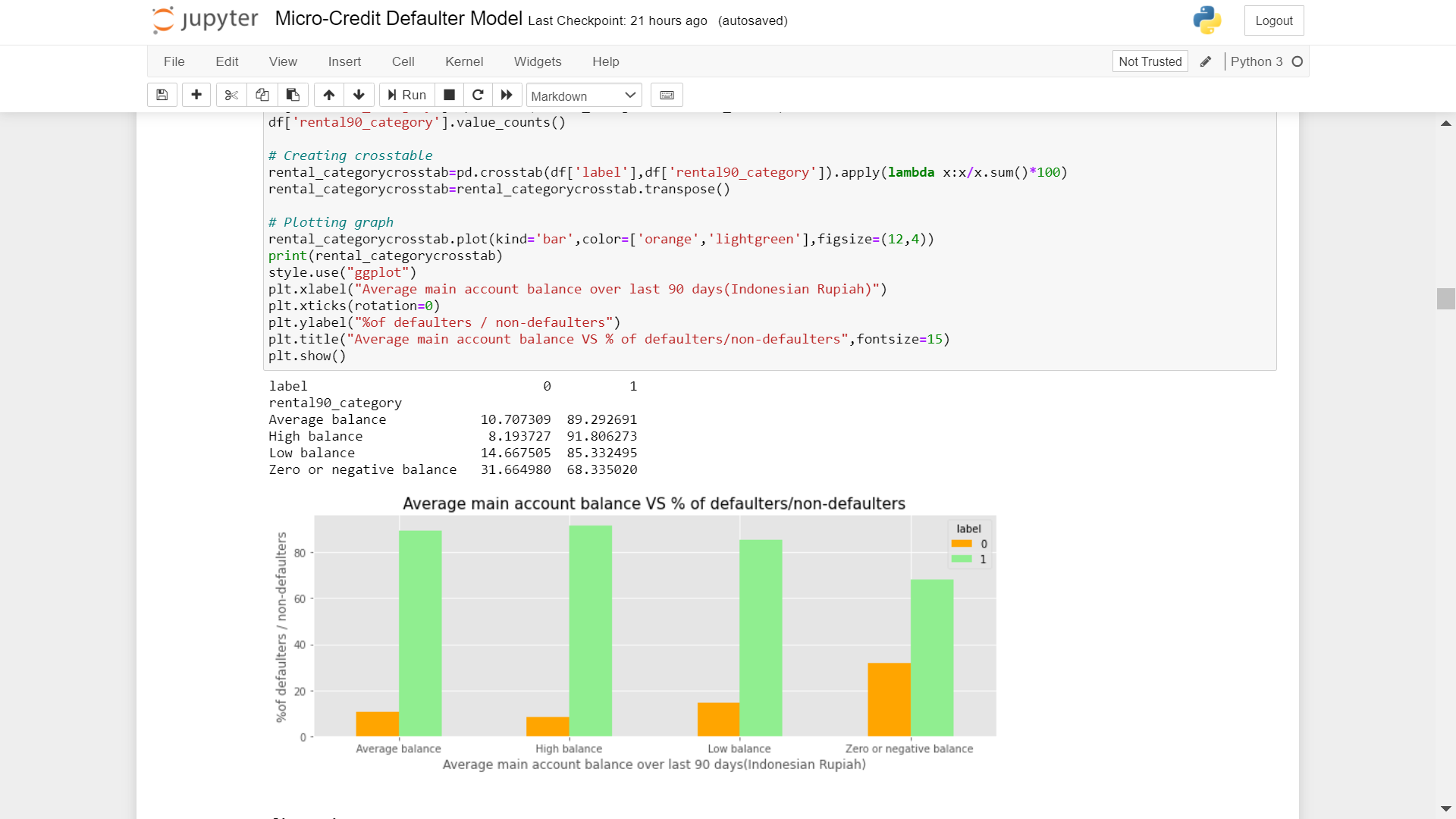
Observations:-

\* 46% defaulters are present among those who are spending averge daily amount below 10(Indonesian Rupiah).

\* 10% defaulters are present among those who are spending averge daily amount between 10-10000(Indonesian Rupiah).

\* While those who are spending high amount are having very less percentage of defaulters.

Above observations suggests that defaulters are present in those who are spending daily average of less amount.

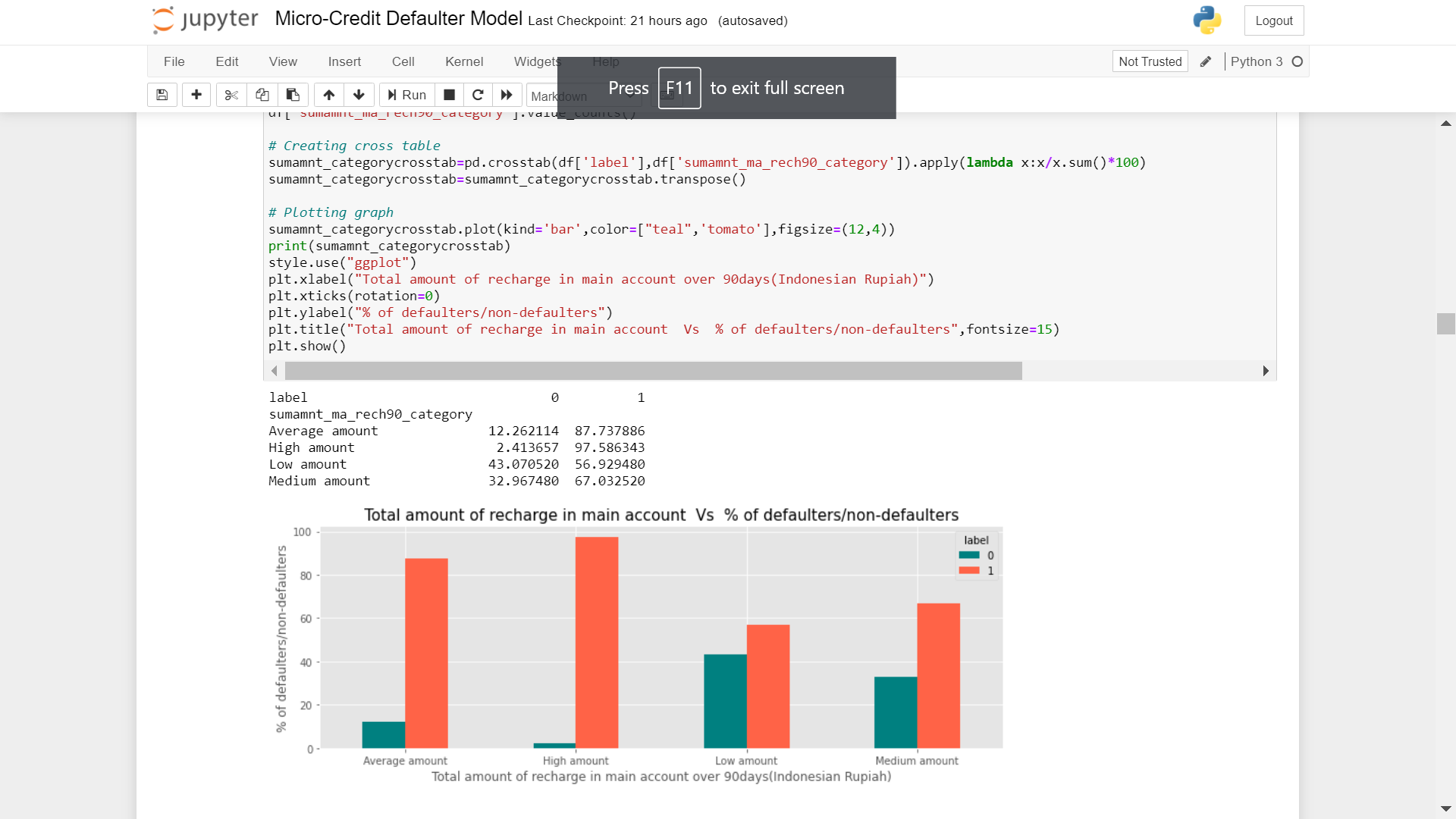


Observations:-

\* 31% of zero or negative main account balance users are defaulters.

\* While only 8% of high main account balance users are defaulters.

More defaulters are present where users are having zero/negative balance or low balance.

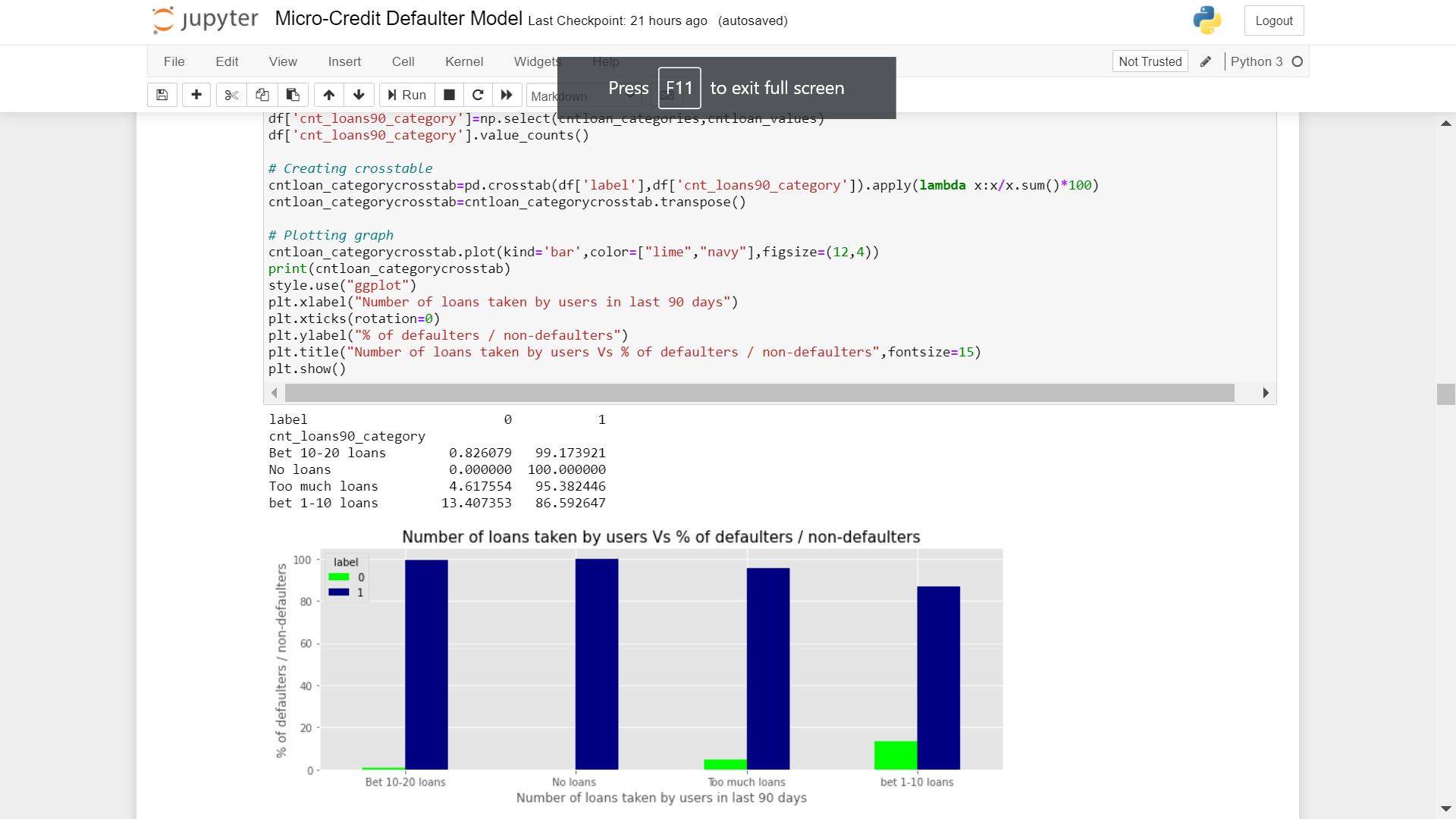


Observations:-

\* 43% of low amount of recharge in main account are defaulters.

\* 32% of median amount of recharge upto 1000(Indonesian Rupiah) are defaulters.

Recharge of main account below 1000(Indonesian Rupiah) are having maximum cases of defaulters.



Observations:-

\* Those who are not taking any loans ,they cann't be defaulters.

\* Those who take loans below 10 times in 90 days are among with highest defaulter's percentage.

\* While as the number of times loans taken by users increases, defaulter's percentage also decreases.

* Hardware and Software Requirements and Tools Used:-

There is no such requirement for hardware ,but I have used intel i5 8th generation processor.

**Software:** Jupyter Notebook (Anaconda 3)

Libraries used in project:

* **For data loading and analysis**
* **a). import pandas as pd –** Pandas is a [software library](https://en.wikipedia.org/wiki/Software_library) written for the [Python programming language](https://en.wikipedia.org/wiki/Python_(programming_language)) for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and [time series](https://en.wikipedia.org/wiki/Time_series).
* **b). import numpy as np –** Numpy is a library for the [Python programming language](https://en.wikipedia.org/wiki/Python_(programming_language)), adding support for large, multi-dimensional [arrays](https://en.wikipedia.org/wiki/Array_data_structure) and [matrices](https://en.wikipedia.org/wiki/Matrix_(math)), along with a large collection of [high-level](https://en.wikipedia.org/wiki/High-level_programming_language) [mathematical](https://en.wikipedia.org/wiki/Mathematics) [functions](https://en.wikipedia.org/wiki/Function_(mathematics)) to operate on these arrays
* **For data visualization**
* **a). import matplotlib.pyplot as plt –** Matplotlib is a plotting library used for data visualization.
* **b). import seaborn as sns -** Seaborn is also a plotting library. It is more advanced than matplotlib but works with matplotlib.
* **Model-preprocessing libraries**
* **from sklearn.preprocessing import LabelEncoder –** ML algorithm understands numerical data only, so for converting categorical data into numerical data LabelEncoder library is used.
* from imblearn.combine import SMOTETomek- This library is used to handle the imbalanced dataset. By using over-sampling and under-sampling it balances the data.
* from collections import Counter – This library is used to find out the balance ratio.
* from sklearn.preprocessing import StandardScaler – StandardScaler library is used to scale the data. As different columns have different range of data, dure to this difference model will become biased. It will give more importance to those features who have high range. So to remove this possibility we use StandardScaler.
* **from sklearn.linear\_model import LogisticRegression -** The library sklearn can be used to perform logistic regression in a few lines as shown using the LogisticRegression class. It also supports multiple features. It requires the input values to be in a specific format hence they have been reshaped before training using the fit method.
* **from sklearn.tree import DecisionTreeClassifier -** Decision Tree is a white box type of ML algorithm. It shares internal decision-making logic, which is not available in the black box type of algorithms such as Neural Network. Its training time is faster compared to the neural network algorithm. The time complexity of decision trees is a function of the number of records and number of attributes in the given data. The decision tree is a distribution-free or non-parametric method, which does not depend upon probability distribution assumptions. Decision trees can handle high dimensional data with good accuracy.
* **from sklearn.ensemble import RandomForestClassifier -** A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max\_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree.
* **from sklearn.model\_selection import train\_test\_split,cross\_val\_score *–***

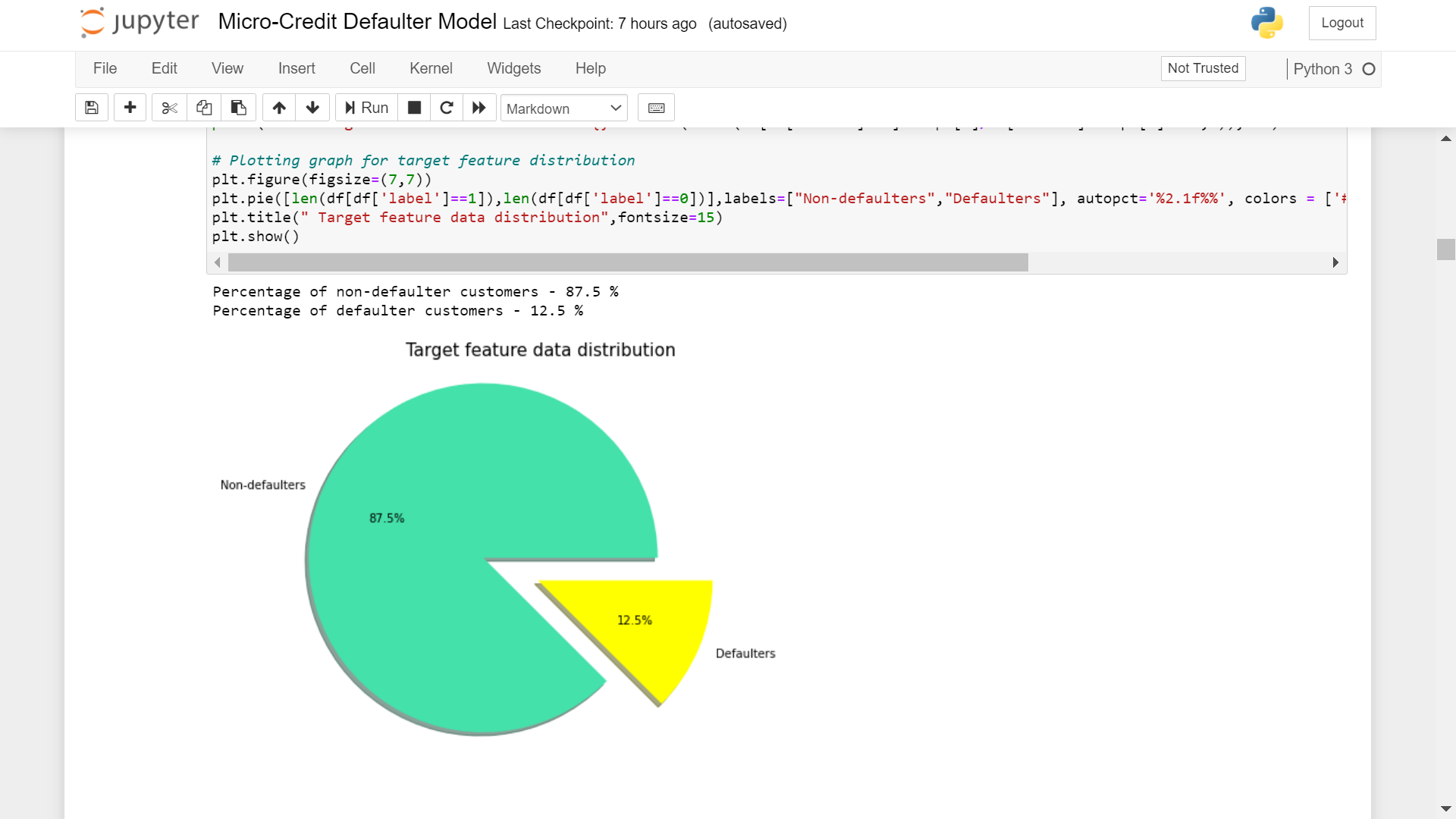
Train\_test\_split is a function in Sklearn model selection for splitting data arrays into two subsets: for training data and for testing data. With this function, you don't need to divide the dataset manually. By default, Sklearn train\_test\_split will make random partitions for the two subsets.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods):-

As we all know ,known target data is treated under Supervised Machine Learning Model and Supervised ML models deals with continuous and discrete data type. If our target data is continuous in nature then Regression Algorithms will be used but if data is discrete in nature/categorical data ,then Classification algorithms will be used.

Let’s check our target data for identification of problem-



So observing the target data ,we find that there are 2 categories ,non-defaulters and defaulters in our data ,which clearly means that this is Classification problem.

* Testing of Identified Approaches (Algorithms):-

In our project ,we have used the following algorithms:-

1. Logistic Regression

2. DecisionTree Classification

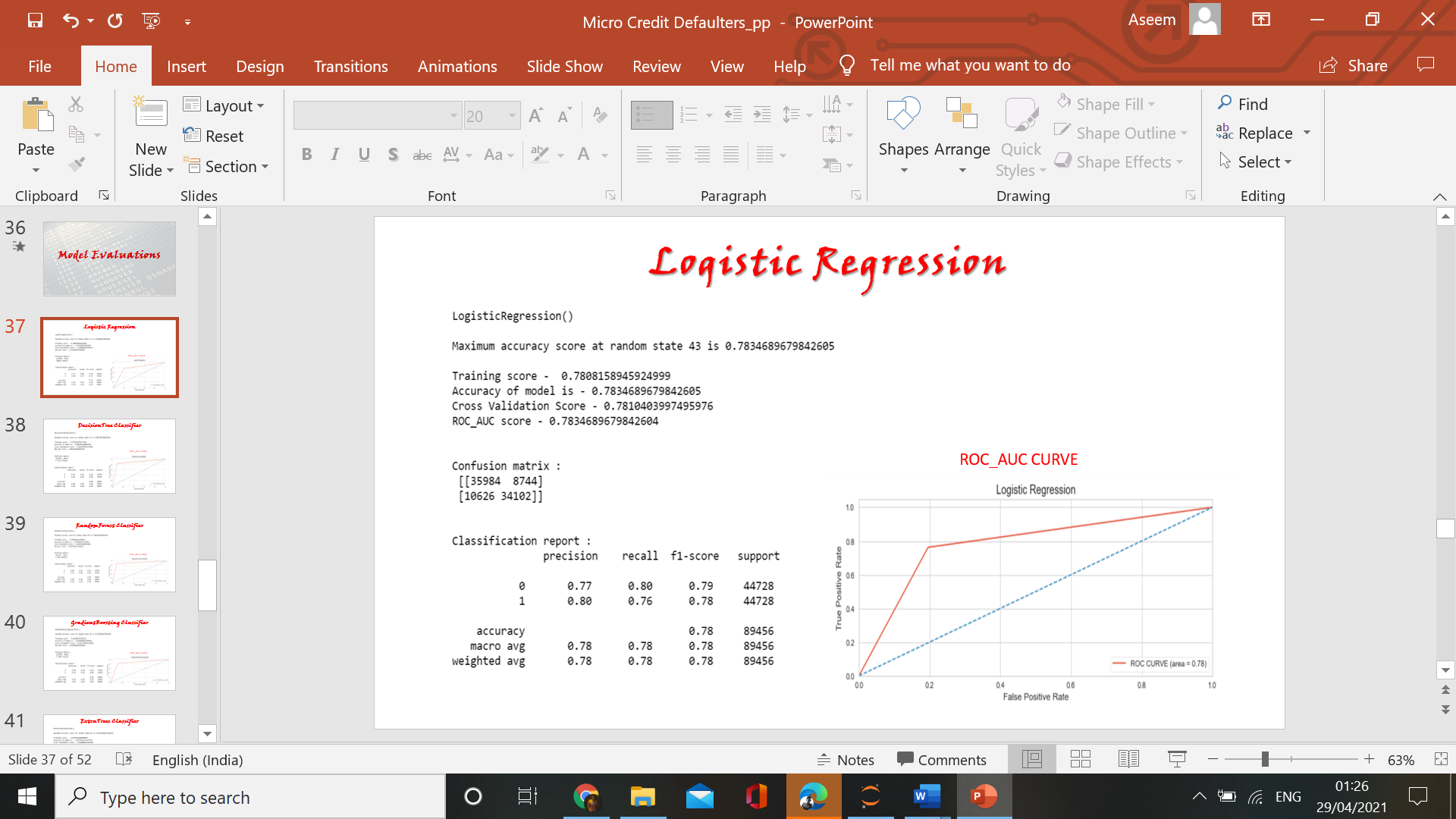
3. RandomForest Classification

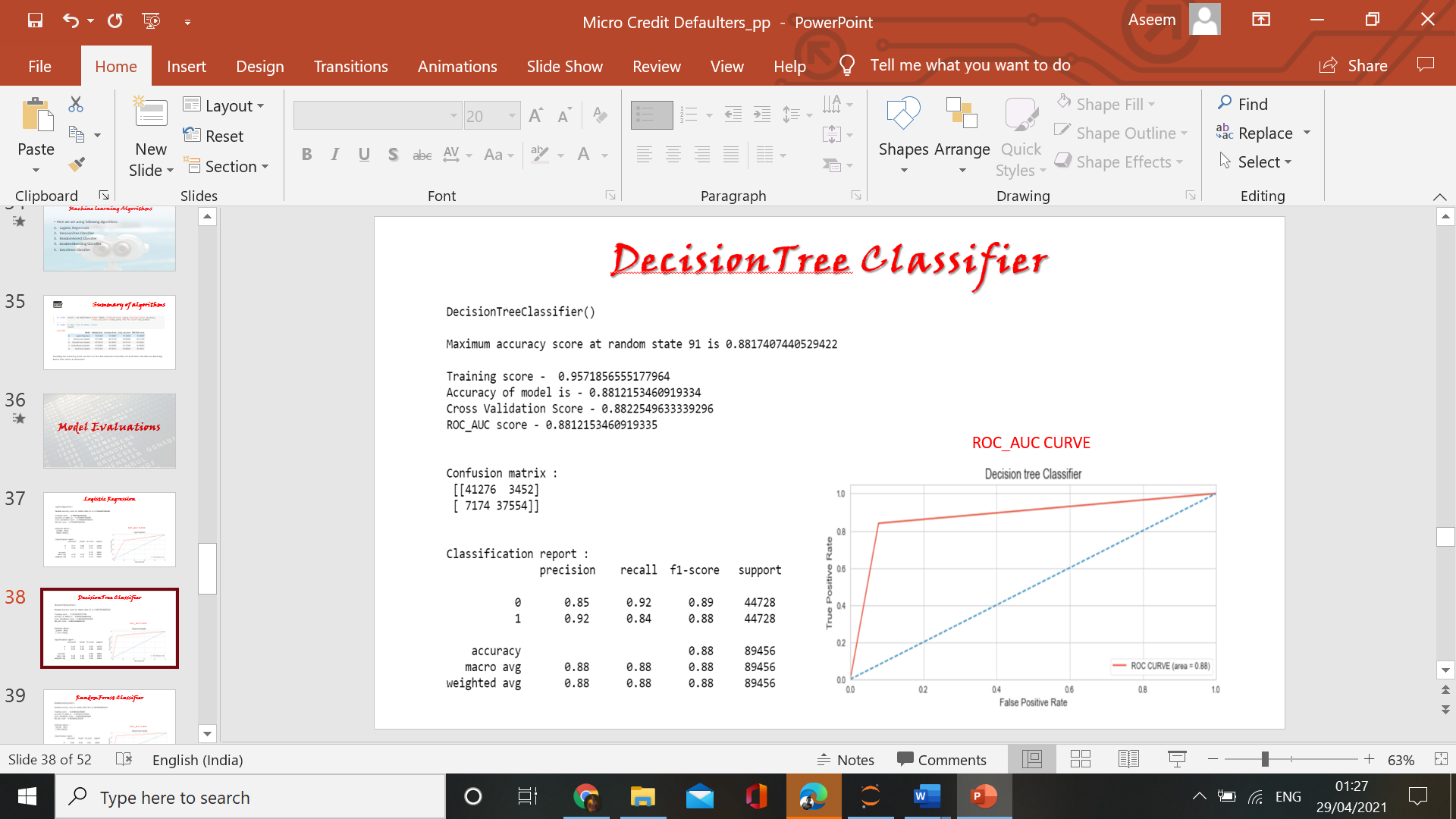
4. GradientBoosting Classification

5. ExtraTrees Classification

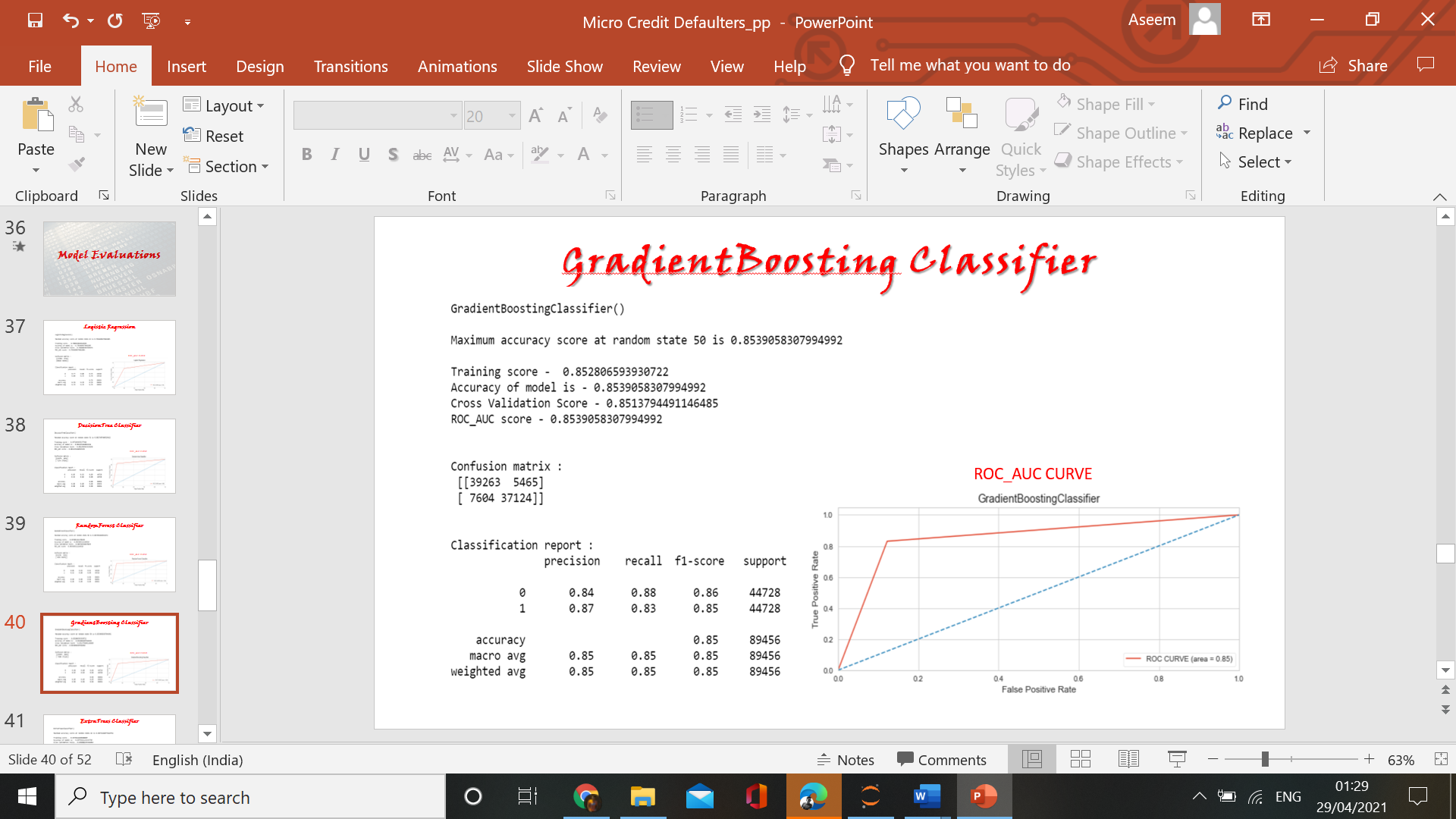
There are some more algorithms which can be used to retrieve better result ,but due to inefficiency of our system we have not used them.

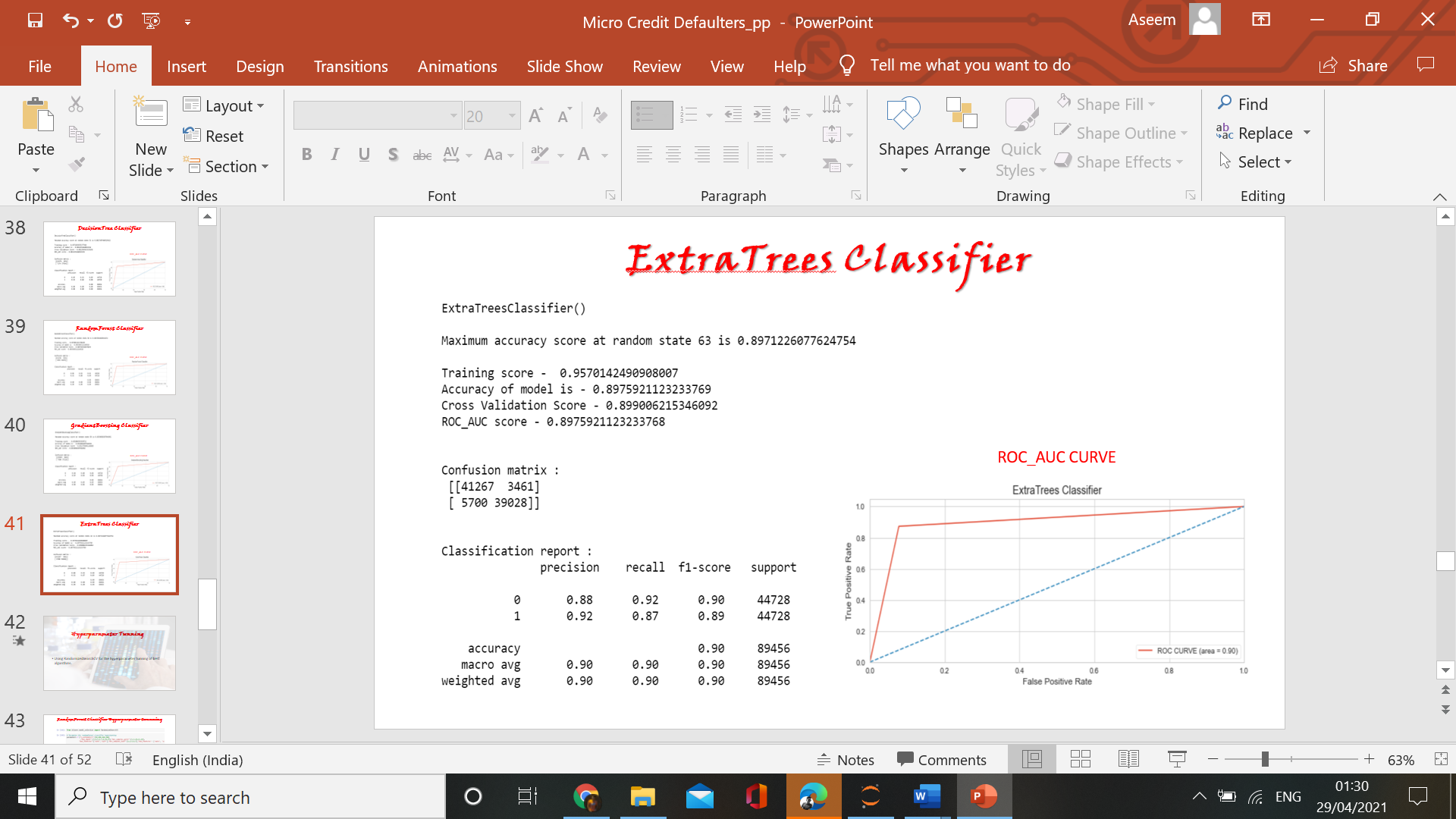
* Run and Evaluate selected models:-



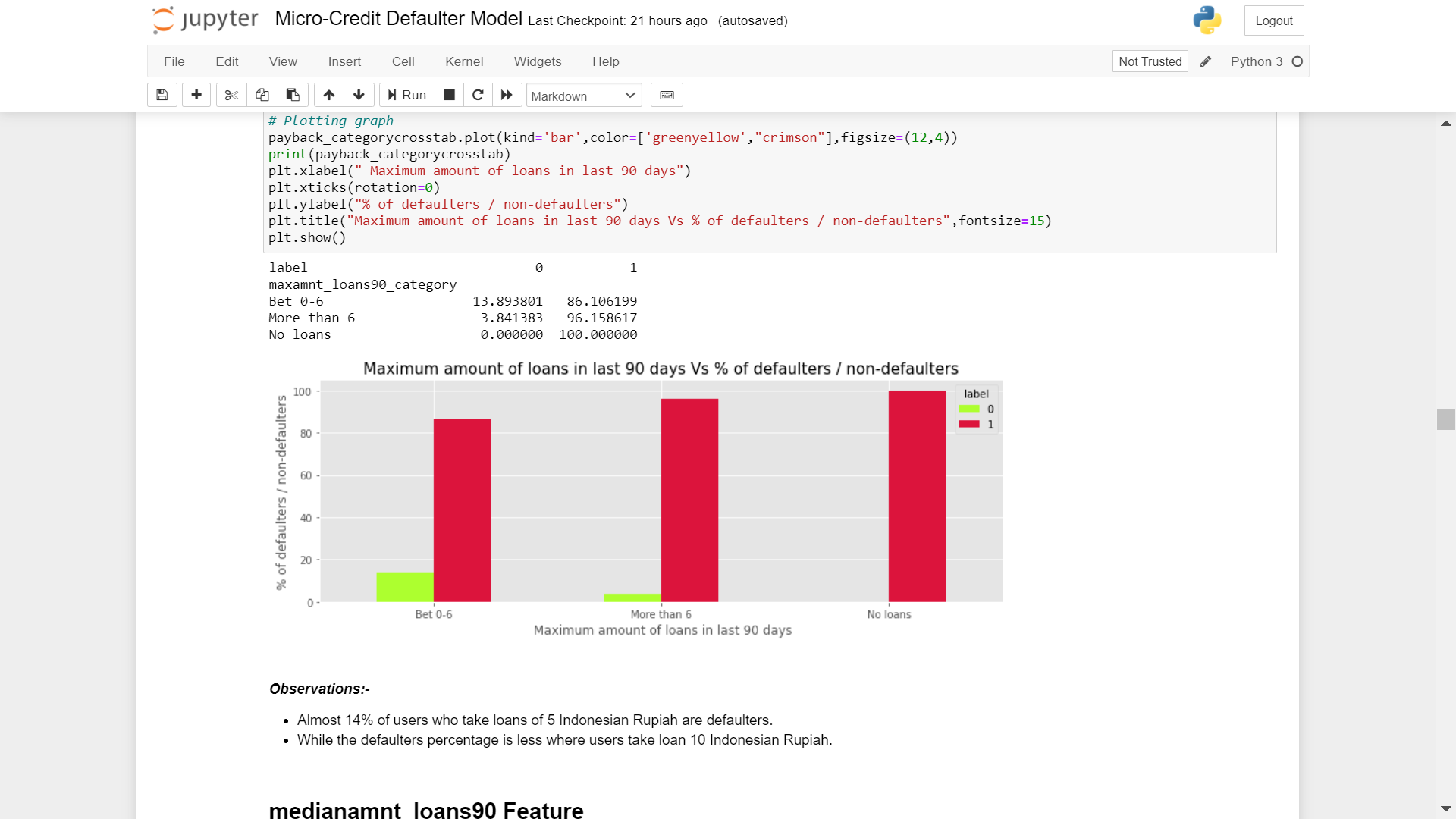








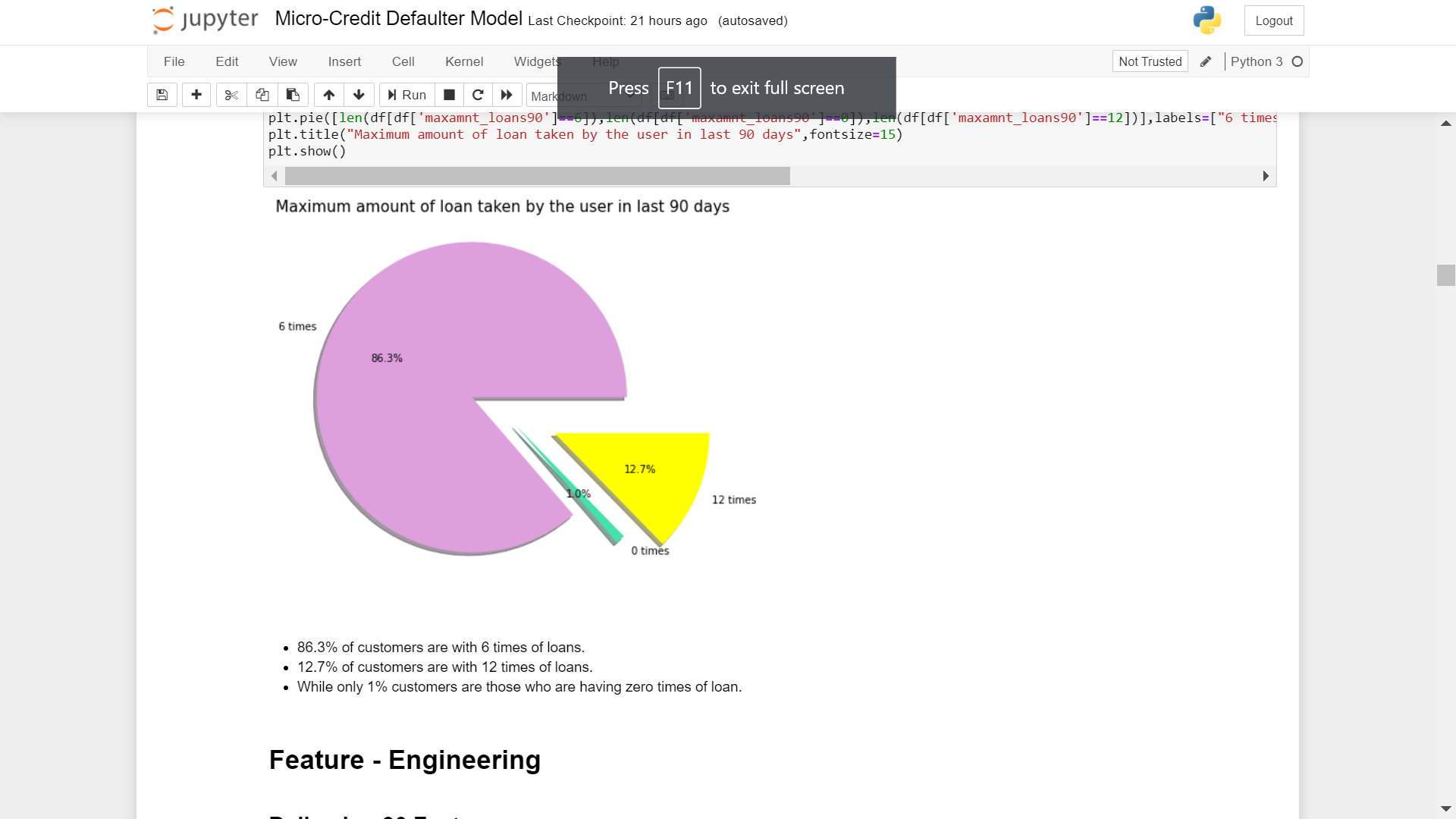
* Key Metrics for success in solving problem under consideration :-
* [**Accuracy score**](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html#sklearn.metrics.accuracy_score)**-** It is used when the True Positives and True negatives are more important. Accuracy can be used when the class distribution is similar.
* **Precision score -** Quantifies the number of positive class predictions that actually belong to the positive class.
* **Recall score -** Quantifies the number of positive class predictions made out of all positive examples in the dataset.
* **F1 score –** Measure provides a single score that balances both the concerns of precision and recall in one number. It is used when we have imbalance in dataset.
* **ROC Curve** - It essentially shows the true positive rate (TPR) against the false positive rate (FPR) for various threshold values.
* **AUC -** AUC calculates the area under the ROC curve, and therefore it is between 0 and 1.
* Visualizations



Observations:-

\* Almost 14% of users who take loans of 5 Indonesian Rupiah are defaulters.

\* While the defaulters percentage is less where users take loan 10 Indonesian Rupiah.

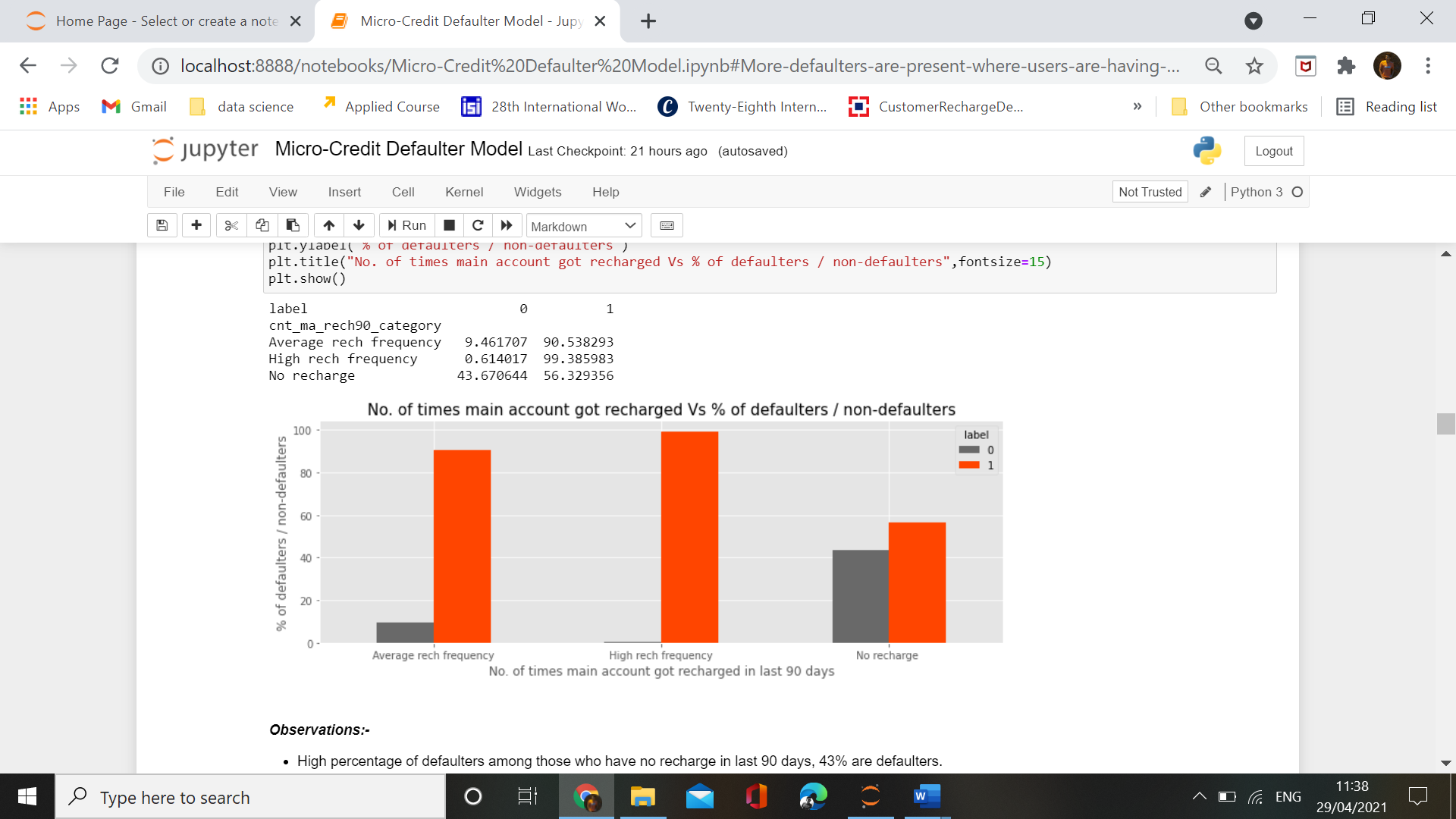


Observations:-

\* 86.3% of customers are with 6 times of loans.

\* 12.7% of customers are with 12 times of loans.

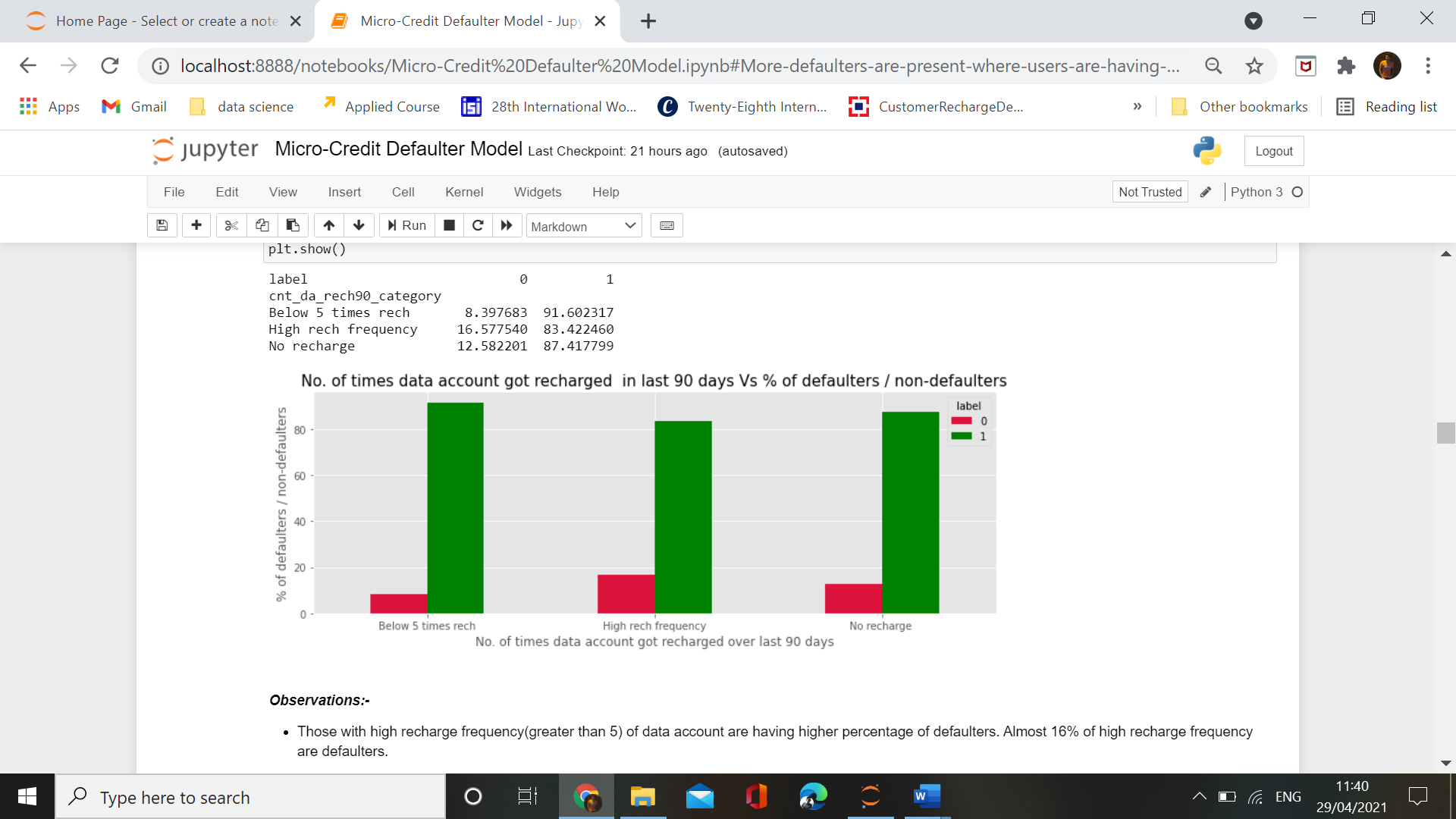
\* While only 1% customers are those who are having zero times of loan.



Observations:-

\* High percentage of defaulters among those who have no recharge in last 90 days, 43% are defaulters.

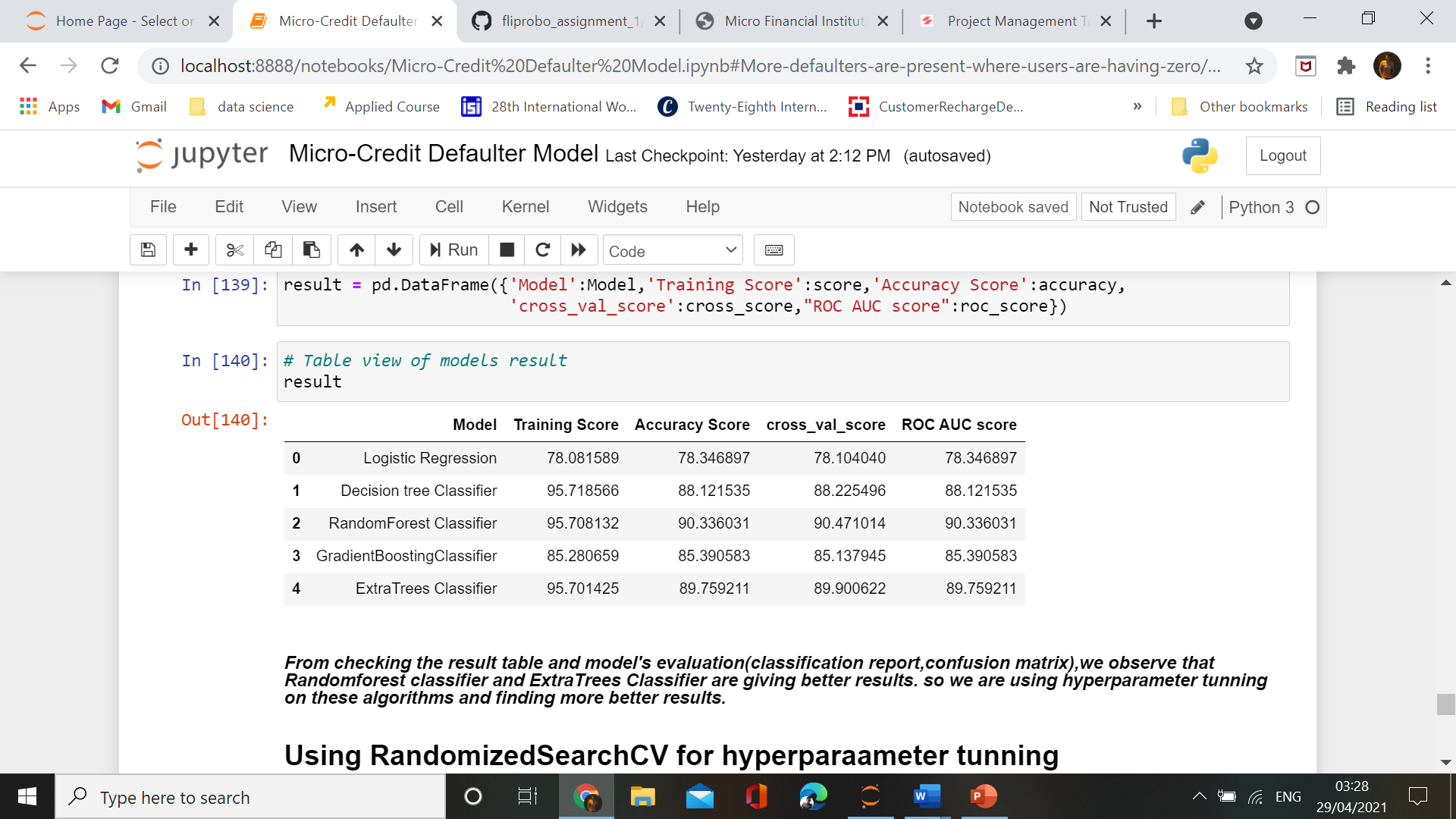
\* While there are some cases in which users are having an average of more than or near about 3 recharges per day but the percentage of defaulters among them are least.



Observations:-

\* Those with high recharge frequency(greater than 5) of data account are having higher percentage of defaulters. Almost 16% of high recharge frequency are defaulters.

* Interpretation of the Results



**CONCLUSION**

* Key Findings and Conclusions of the Study
* 86.3% of Users consits of those who takes loan upto 1-6 times in 90 days.
* While only 1% Users are those who doesn't take any loans.
* 32% of New Users(upto 30 days) are defaulters.
* 46% defaulters are present among those who are spending averge daily amount below 10(Indonesian Rupiah).
* While those who are spending high amount are having very less percentage of defaulters.
* 31% of zero or negative main account balance(over last 90 days) users are defaulters.
* 25.6% of users who had not recharged main account more than a month ago are defaulters.
* only 7.5% of users are defaulters from those who have not recharged the data account from a long time.
* 42% of Users who are having low amount(<=500Indonesian Rupiah) of recharge are defaulters.
* 43% of low amount(<=500Indonesian Rupiah) of recharge in main account are defaulters.
* 43% are defaulters among those who have no recharge in last 90 days, .
* Those with high recharge frequency(greater than 5) of data account are having higher percentage of defaulters. Almost 16% are defaulters.
* Those who take loans less than 10 times in 90 days are among with highest defaulter's percentage.While as the number of times loans taken by users increases, defaulter's percentage also decreases.
* Almost 14% of users who take loans of 5 Indonesian Rupiah are defaulters.
* While the defaulters percentage is less where users take loan of 10 Indonesian Rupiah.
* Learning Outcomes of the Study in respect of Data Science

Human eyes are not capable to understand the numerical data so easily .i.e data visualization helps a lot in understanding the data behavior and helping us to find whats going on with data.

Using the data cleaing we find out that there are a lot of outliers in our data .If not treated ,these outliers will hinder the results and we will not be able to predict precisely ,but data science has solved all these hurdles and making the data more clean which gives more accurate results and helps in predicting the future outcomes.

* Limitations of this work and Scope for Future Work
* The major limitation in dataset I faced was presence of lots of outliers and as data was expensive I couldnot drop it and with those outliers my analysis is not very accurate.
* We could use other algorithms also ,which may give better results,but due to inefficiency of my system,i can’t do that.
* There are a lot of other techniques like feature\_importances,higher hyperparameter tunning using GridSearchCV ,to retrieve better results.
* This model will help my client in signify the difference between defaulters and non-defaulters and also helps in improvement of business in selection valuable customers.