**MICRO CREDIT CARD MODEL**

**(loan prediction)**

**Submitted by:**

**BHUVANA SHREE N**

**ACKNOWLEDGMENT**

I would like to express my special thanks of gratitude to Intrnforte for giving this internship project on the topic (Micro-Credit Defaulter), which helped me in doing a lot of Research and which help me learn way too many things. The references I used for completion of this project are mentioned here Kaggle & Github [www.stackoverflow.com](http://www.stackoverflow.com)[www.geeksforgeeks.org](http://www.geeksforgeeks.org) are the websites which helped me in completing the project.

**Problem statement:** 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 the Telecom Industry. They are a fixed wireless telecommunications network provider. They have launched various products and have developed their 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.

**Problem Context**: Telecom Industries 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 microcredit 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.

**The Objective:** We have to 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.

* Develop a machine learning model to predict loan repayment probabilities.
* Compose a comprehensive project report covering the entire process.
* Adhere to the outlined steps, encompassing data cleaning, exploration, visualization, feature engineering, model training, hyperparameter tuning, and appropriate metric selection.
* Try to infer unique patterns from the data and try to generate new features.
* Use 45 models for training, do proper hyperparameter tuning and choose the right evaluation metrics to finalize your model.
* Test your predictions on multiple metrics like log loss, Recall and Precision.

Taking up this project, will help me figure out the need and lending the loan for right and needed ones.

|  |  |
| --- | --- |
| 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\_decr90 | Daily amount spent from main account, averaged over last 90 days |
| 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 |
| 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 |
| medianamnt\_ma\_rech30 | Median of amount of recharges done in main account over last 30 days at user level |
| 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 |
| medianamnt\_ma\_rech90 | Median of amount of recharges done in main account over last 90 days at user level |
| medianmarechprebal90 | Median of main account balance just before recharge in last 90 days at user level |
| 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 |

label 0

msisdn 0

aon 0

daily\_decr30 0

daily\_decr90 0

rental30 0

rental90 0

last\_rech\_date\_ma 0

last\_rech\_date\_da 0

last\_rech\_amt\_ma 0

cnt\_ma\_rech30 0

fr\_ma\_rech30 0

sumamnt\_ma\_rech30 0

medianamnt\_ma\_rech30 0

medianmarechprebal30 0

cnt\_ma\_rech90 0

fr\_ma\_rech90 0

sumamnt\_ma\_rech90 0

medianamnt\_ma\_rech90 0

medianmarechprebal90 0

cnt\_da\_rech30 0

fr\_da\_rech30 0

cnt\_da\_rech90 0

fr\_da\_rech90 0

cnt\_loans30 0

amnt\_loans30 0

maxamnt\_loans30 0

medianamnt\_loans30 0

cnt\_loans90 0

amnt\_loans90 0

maxamnt\_loans90 0

medianamnt\_loans90 0

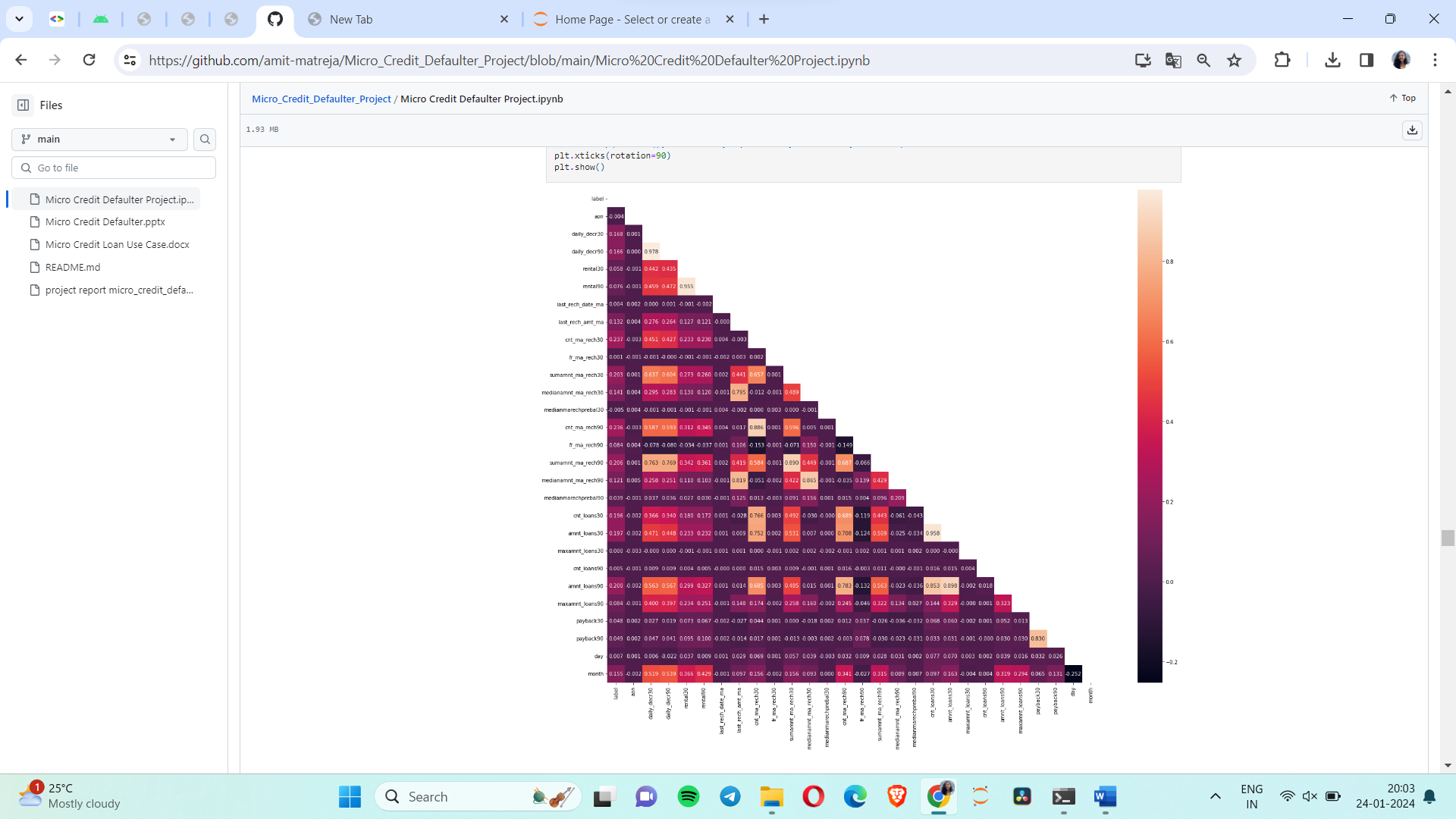
payback30 0

payback90 0

pcircle 0

There were no null value was present in the dataset but there are some outliers which also get too removed, approximately 48128 outliers get removed from the data. After that categorical are change to integer or float with the help of **LabelEncoder**. Then I used updated data for the correlation for splitting it into x and y with the help of standard scalar it will transform the data in such way that its distribution will have a mean value 0 and standard deviation of 1.

* Data Inputs- Logic- Output Relationships



* Hardware and Software Requirements and Tools Used
* **Hardware** – Laptop (Windows 11, 16 GB)
* **Software** - anaconda jupyter notebook
* **Libraries**- numpy, pandas, seaborn, matplotlib.pyplot, warning
* Testing of Identified Approaches (Algorithms)

Listing down all the algorithms used for the training and testing.

* KNN=KNeighborsClassifier(n\_neighbors=6)
* LR=LogisticRegression()
* DT=DecisionTreeClassifier(random\_state=6)
* GNB=GaussianNB()
* Run and Evaluate selected models

# 1) KNeighborsClassifier:

Accuracy Score: 0.8951471483477044

Confusion Matrix: [[54144 586]

[10954 44375]]

precision recall f1-score support

0 0.83 0.99 0.90 54730

1 0.99 0.80 0.88 55329

accuracy 0.90 110059

macro avg 0.91 0.90 0.89 110059

weighted avg 0.91 0.90 0.89 110059

* Here we can see that KNN is predicting score of 89.51%
* **2) ExtraTreesClassifier:**

Accuracy Score: 0.9598578944020935

Confusion Matrix: [[53417 1313]

[ 3105 52224]]

precision recall f1-score support

0 0.95 0.98 0.96 54730

1 0.98 0.94 0.96 55329

accuracy 0.96 110059

macro avg 0.96 0.96 0.96 110059

weighted avg 0.96 0.96 0.96 110059

* Here we have got an excellent accuracy score of 95.98% using Extra Tree Classifier.

# 3) Gradient Boosting Classifier:

Accuracy Score: 0.8982273144404365

Confusion Matrix: [[50129 4601]

[ 6600 48729]]

precision recall f1-score support

0 0.88 0.92 0.90 54730

1 0.91 0.88 0.90 55329

accuracy 0.90 110059

macro avg 0.90 0.90 0.90 110059

weighted avg 0.90 0.90 0.90 110059

* Here we have got accuracy score of 89.82% using GradientBoostingClassifier.

# 4) Decision Tree Classifier:

Accuracy Score: 0.9099937306353865

Confusion Matrix: [[50191 4539]

[ 5367 49962]]

precision recall f1-score support

0 0.90 0.92 0.91 54730

1 0.92 0.90 0.91 55329

accuracy 0.91 110059

macro avg 0.91 0.91 0.91 110059

weighted avg 0.91 0.91 0.91 110059

* Here we have got the accuracy score of 90.99% using DecisionTreeMatrix

# 5) Random Forest Classifier:

Accuracy Score: 0.9523800870442217

Confusion Matrix: [[52430 2300]

[ 2941 52388]]

precision recall f1-score support

0 0.95 0.96 0.95 54730

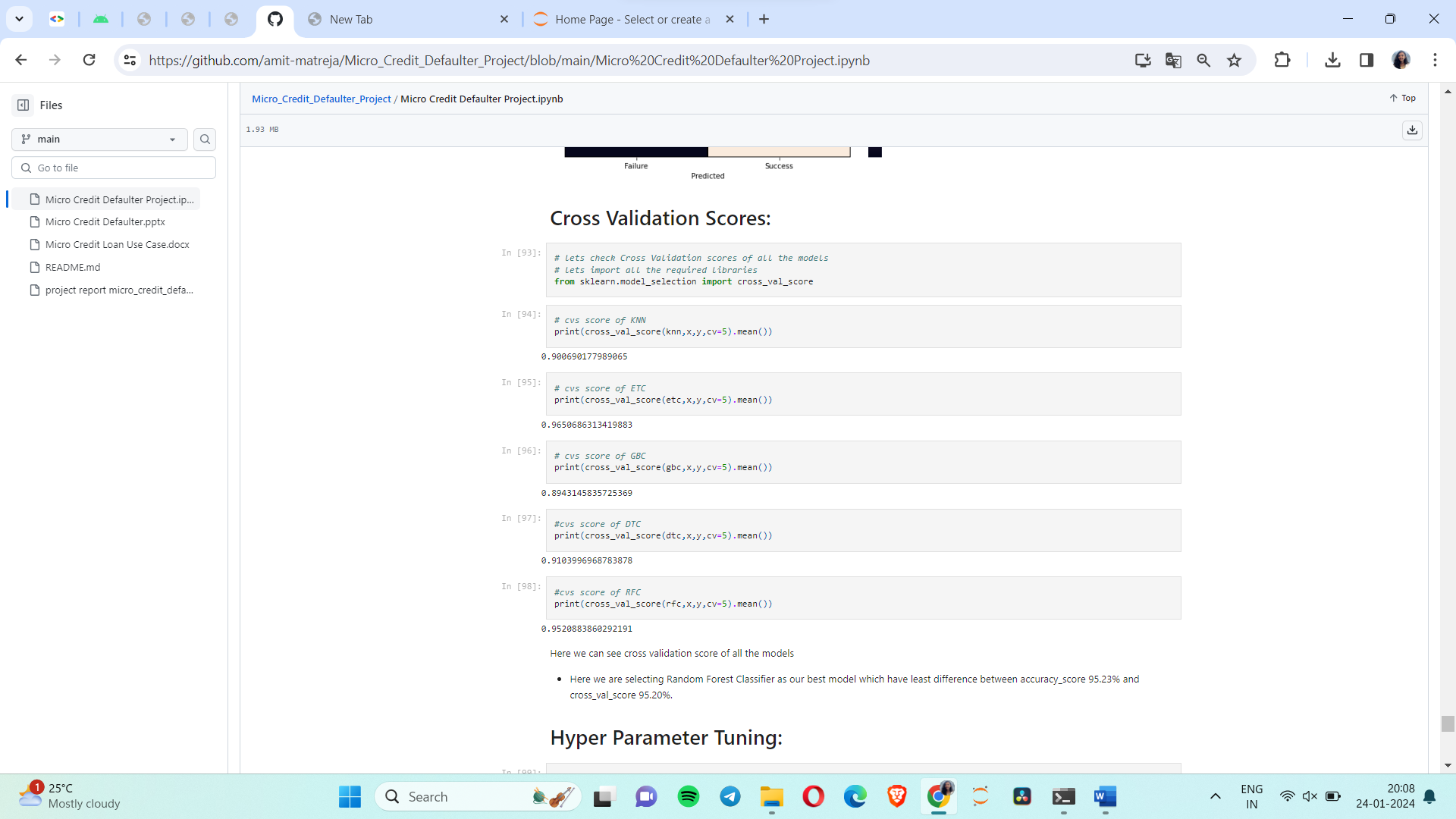
1 0.96 0.95 0.95 55329

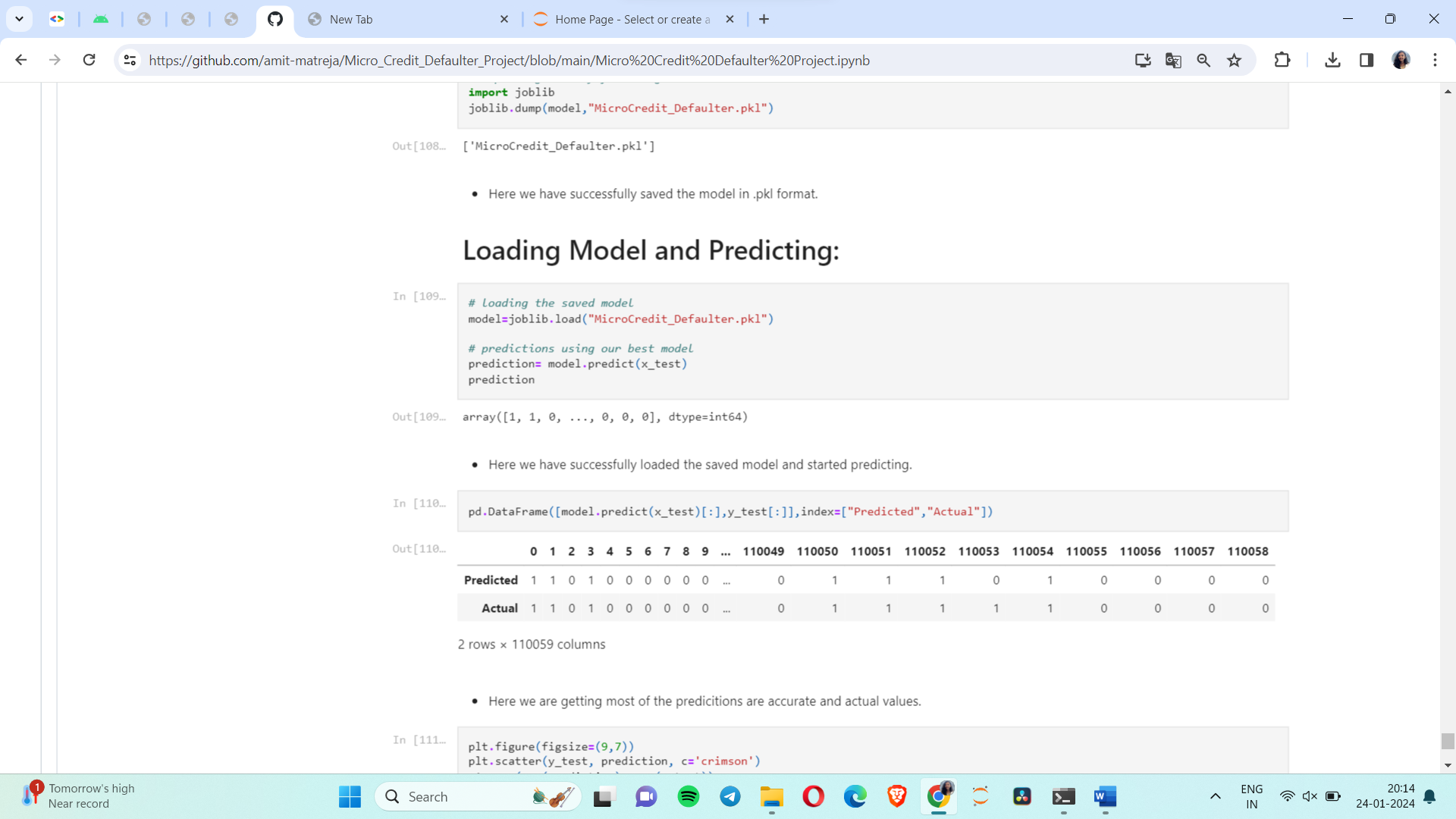
accuracy 0.95 110059

macro avg 0.95 0.95 0.95 110059

weighted avg 0.95 0.95 0.95 110059

* Here we have got accuracy score of 95.23% using Random Forest Classifier

****



Conclusion

From this dataset I get to know that each feature play a very import role to understand the data. Data format plays a very important role in the visualization and Appling the models and algorithms.

By selected the features which are correlated to each other and are independent in nature. Visualization helped us in understanding the data by graphical representation it made things easy for us to understand what data is trying to say.

From this project I learned about the micro credit industry, visualization, data cleaning, handling outliers and using various algorithms on huge dataset. This was the first time I worked on such huge dataset