

MICRO CREDIT LOAN PROJECT



**Submitted by:**

**DEEPTHI MALLU**

**ACKNOWLEDGMENT**

I take this opportunity to acknowledge everyone who have helped me in every stage of this project.

Firstly, I am indebtedly grateful to my SME MR. Sajid Chowdary sir, who helped me from beginning of my Projects. Am also thankful to my Mentor Shankar Gowda Sir and my whole Data Trained team, where I have learnt Analysing the datasets and building the models using Machine learning and making the projects. Finally, am so thankful to my Flip Robo Technologies team, as they provided me the opportunity to work as intern in their company.

I feel pleasure, to make project report on “Micro Credit Defaulter Model”. It has been my privilege to have a team of project guide who have assisted me from the commencement of this project. The project is a result of my hard work, and determination put on by me with the help of Wikipedia, You Tube videos, skikit-learn.org, reffered some old projects on Kaggle.com.

**INTRODUCTION**

BusinessProblem Framing:

* Microcredit is a method of lending very small sums to individuals to start or expand a small business.
* Microcredit borrowers tend to be low-income individuals living in parts of the developing world; the practice originated in its modern form in Bangladesh.
* Most microcredit schemes rely on a group borrowing model, originally developed by Nobel Prize winner Muhammad Yunus and his Grameen Bank.
* A Microfinance Institution is an organization that offers financial services to low- income populations.
* Micro credit is “Loan of very small amount”. It can be defined as provision of parsimony, credit and other financial services and products of very small amount to the poor in rural, semi-urban and urban areas for enabling them to raise their income levels and improve living standards.
* The institutions that provide Micro Credit are called Micro Credit Institutions. Micro Credit is provided to those individuals that lack collateral, steady employment and a verifiable credit history and therefore cannot meet even the most minimal qualifications to gain access to traditional credit.
* This group of individuals includes artisans, tiny and small industries, grocers, vegetable vendors, rickshaw pullers, roadside retailers and the like. Other activities include farming, poultry, cattle rearing, piggery, fishery etc.

Conceptual Background of the Domain Problem:

Micro credit is “Loan of very small amount”. It can be defined as provision of parsimony, credit and other financial services and products of very small amount to the poor in rural, semi-urban and urban areas for enabling them to raise their income levels and improve living standards.

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).

Review of Literature:

Microcredit is a part of microfinance. The term Microfinance is used for the provision of a wider range of financial services to the very poor.

The innovative idea of Microcredit originated with the Grameen Bank in Bangladesh. In 1976 Professor Muhammad Yunus, launched a research project to examine the possibility of designing a credit delivery system to provide banking services targeted to the rural poor. Yunus began the project in a small town called Jobra, using his own money to deliver small loans at low-interest rates to the rural poor. Grameen Bank was followed by organizations such as [BRAC](https://en.wikipedia.org/wiki/BRAC_(NGO)) in 1972 and [ASA](https://en.wikipedia.org/wiki/Association_for_Social_Advancement) in 1978.  Microcredit reached Latin America with the establishment of PRODEM in Bolivia in 1986; a bank that later transformed into the for-profit BancoSol. So for that he awarded Nobel peace prize in 2006.

Grameen Bank has successfully enabled extremely impoverished people to engage in self-employment projects that allow them to generate an income and, in many cases, begin to build wealth and exit poverty.

Different types of institutions that provides micro credit loans in India are:

* Commercial banks.
* Credit unions.
* Non-governmental organisations (NGOs)
* Sectors of government banks.
* Cooperatives.

Motivation for the Problem Undertaken:

In India we see many families under below poverty Line, by using these techniques we can also implement many such micro credit programs in India.

In [India](https://en.wikipedia.org/wiki/India), the [National Bank for Agriculture and Rural Development](https://en.wikipedia.org/wiki/National_Bank_for_Agriculture_and_Rural_Development) (NABARD) finances more than 500 banks that on-lend funds to [self-help groups](https://en.wikipedia.org/wiki/Self-help_group_(finance)) (SHGs). SHGs comprise twenty or fewer members, of whom the majority are women from the poorest [castes](https://en.wikipedia.org/wiki/Caste) and tribes. Members save small amounts of money, as little as a few rupees a month in a group fund. Members may borrow from the group fund for a variety of purposes ranging from household emergencies to school fees. As SHGs prove capable of managing their funds well, they may borrow from a local bank to invest in small business or farm activities. Banks typically lend up to four rupees for every rupee in the group fund. In Asia borrowers generally pay interest rates that range from 30% to 70% without commission and fees. Nearly 1.4 million SHGs comprising approximately 20 million women now borrow from banks, which makes the Indian SHG-Bank Linkage model the largest microfinance program in the world.

Motivation behind this whole project is that we will analyse the low-income of Indonesia population who have taken a credit facility for telecom. We will train our model by using machine learning models to check the patterns using different Algorithms when a person gets defaulter and who all are the persons who pay on time and are non-defaulters. I hope this model will help in analysing the patterns for micro credit institutions while making their policies for giving credit in telecom sector.

These days many of them are taking the services provided by micro financial services in India for making their business and paying back the loans along with the interests in rural, semi urban and urban areas. These services are considered as poverty reduction tool. So government has to come forward to motivate and make them to know about services and make them benefit and help them by providing services. Even though there are some villages where they are not aware of this type of services. So our government has to take a step to skill the people who are not aware of this type of credits, even some people they worry of paying interests for taken loan amounts.

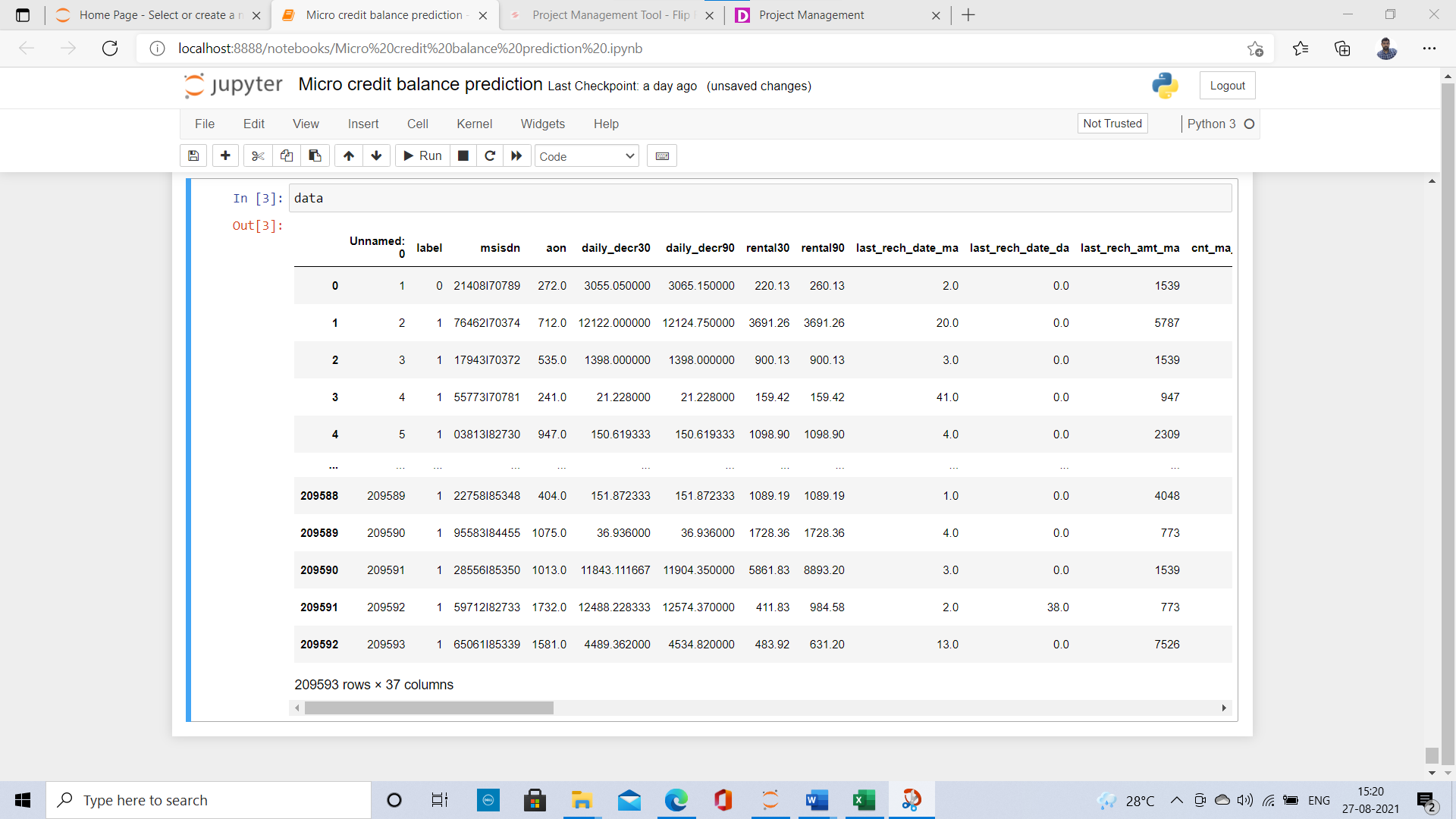
**Analytical Problem Framing**

Mathematical/ Analytical Modelling of the Problem:

We need 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 the dataset, the 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. There are total 209593 rows and 37 columns in the dataset. There are No Null values in the dataset. Maximum number of columns are of Float type and Integer type and very few are of object type. There is lots of skewness and outliers present in our dataset which need to be cleaned using many skew removing techniques.

Data Sources and their formats:

Our dataset consists of Features and label. The label indicates the loan payment weather the consumer paid back the amount balance. 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). On the basis of that the label is given as 1 and 0. 1 is success and 0 is failure. Which means the consumer paid back the loan amount or not.



Our dataset consists of different columns which are:

|  |  |
| --- | --- |
| 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 users |
| 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 Pre-processing Done:

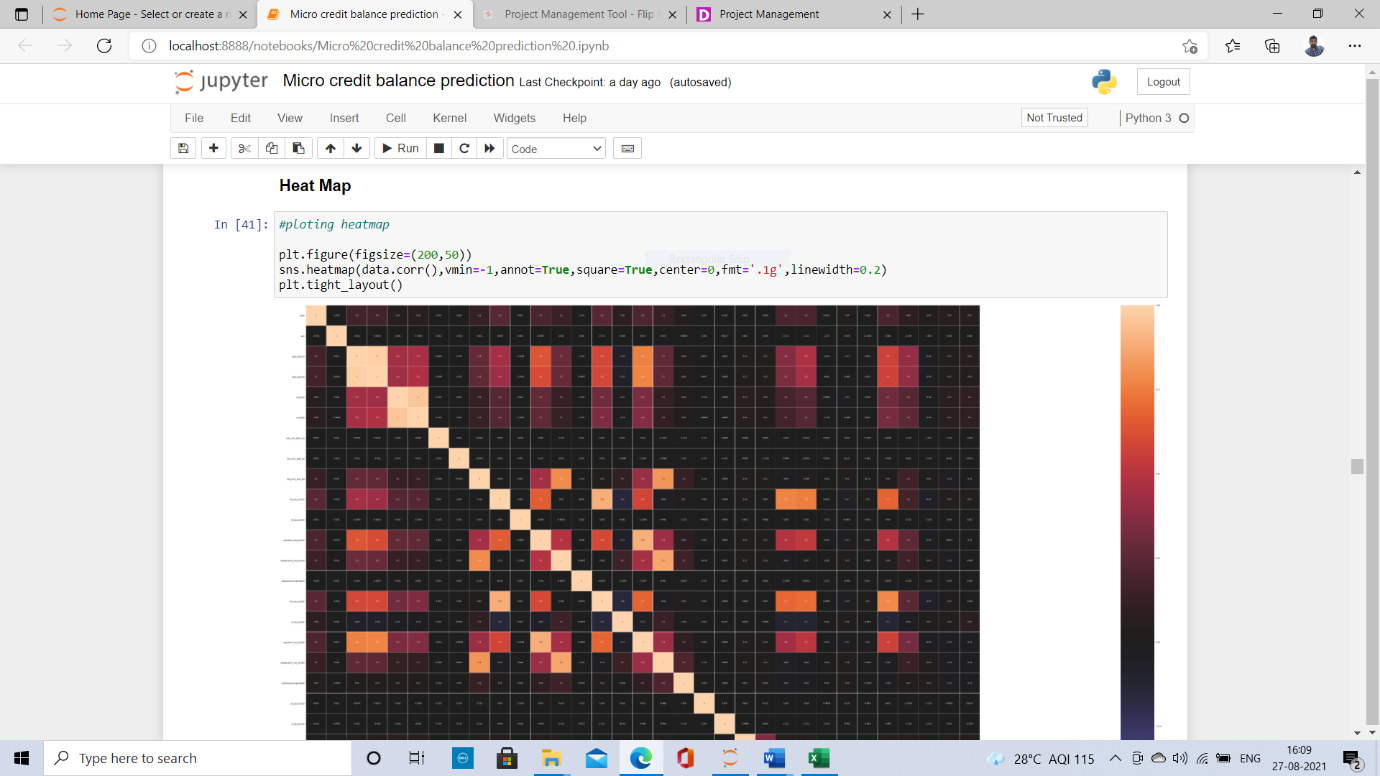
First I have imported all the libraries like pandas, Numpy, matplotlib, seaborn, etc… by using pandas technique I imported the CSV file. After that I checked shape of the data, next I have checked for the Null Values using is-null () method and then I have checked for type of the data using info method which gives the information about the data. After that I have checked for the stats of the columns.

In the dataset there is a column with msisdn which gives the information of the consumers mobile number which is not required for our prediction so I have dropped that column and there are more two columns of object type they are pcircle and date pcircle which is a telecom circle which s of same throughout the column and the date column which are not required for our prediction so I have dropped those two columns. In our dataset all the columns has skewness and outliers present which need to be handled.

Data Inputs- Logic- Output Relationships:

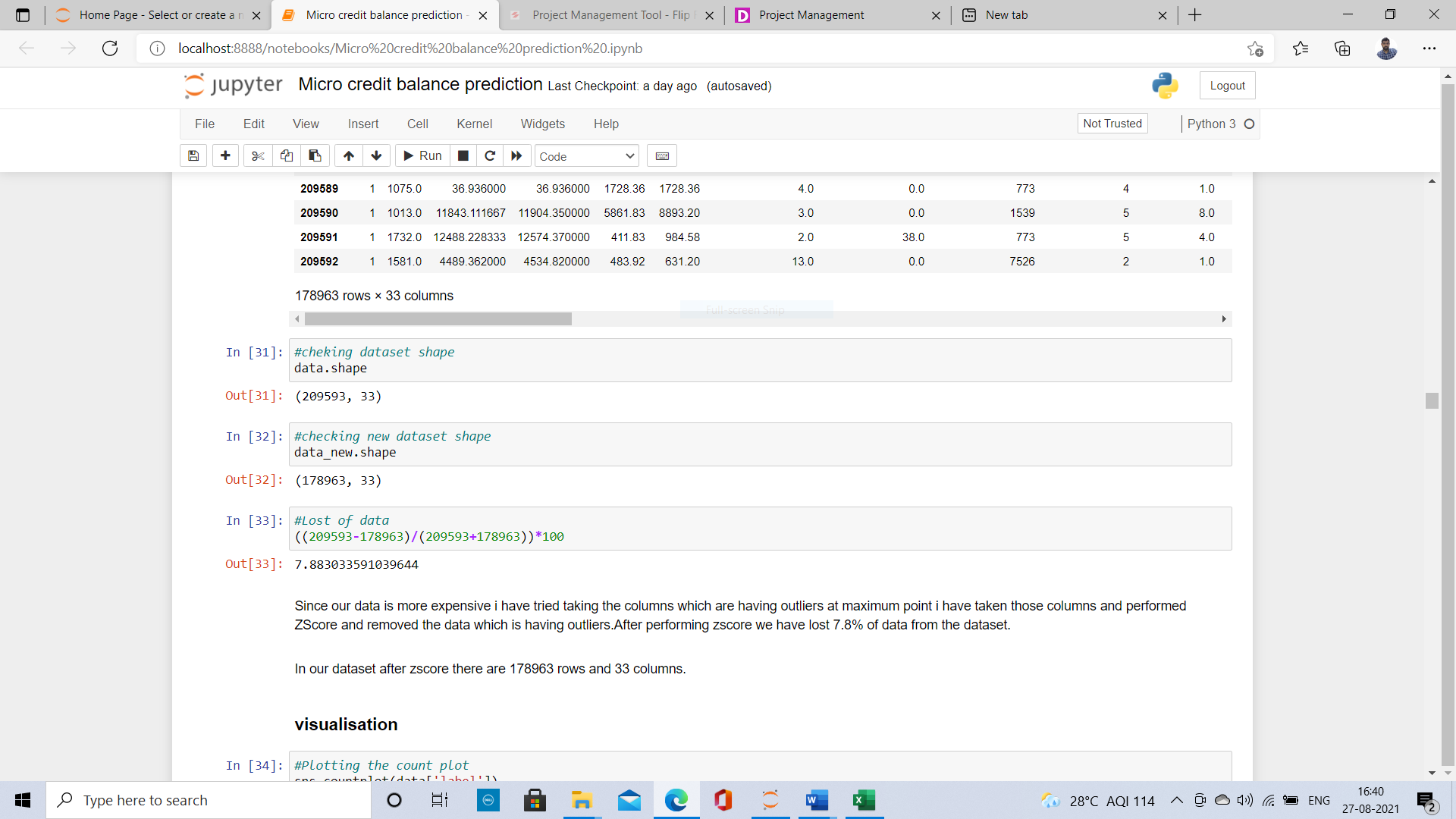
The dataset consists of features and label. The features are independent and label is dependent as the values of our independent variables changes as according our label varies.

To find the relation between our features and label, I have used correlation matrix and heap map plot for visualising the correlation. From that I found that our label is not that correlated with the features. Very less of 10% and 20% correlated with some of the features.



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

In our columns we have many outliers present when we apply Z-Score for removing the outliers on all the columns we are losing more than 12% of data. As the data is more expensive we cannot loose the data more than 8%. So for that reason what I did is I have checked all the outliers by plotting the box plot then the features which are having more outliers which are at maximum point I took those columns and performed Z- score on those features after that we have lost 7.8% data which are having outliers.



Hardware and Software Requirements and Tools Used:

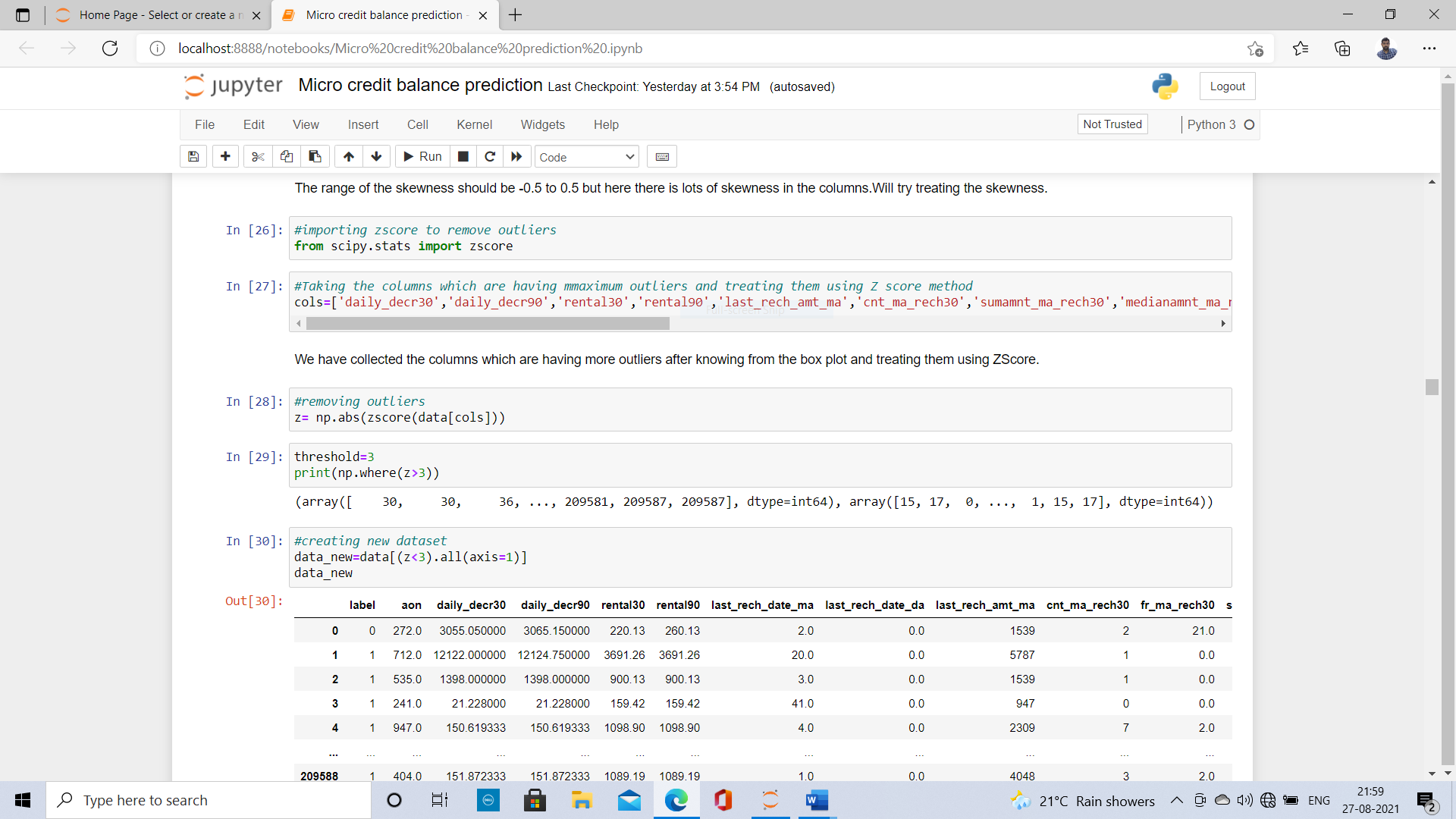
I have used my laptop, Jupiter Notebook which is having GUI interface. Imported necessary libraries from python such as pandas, NumPy, seaborn, matplotlib, then imported the required model libraries from Scikit learn to import our algorithms, after that imported the package called imblearn for up-sampling the label using SMOTE.



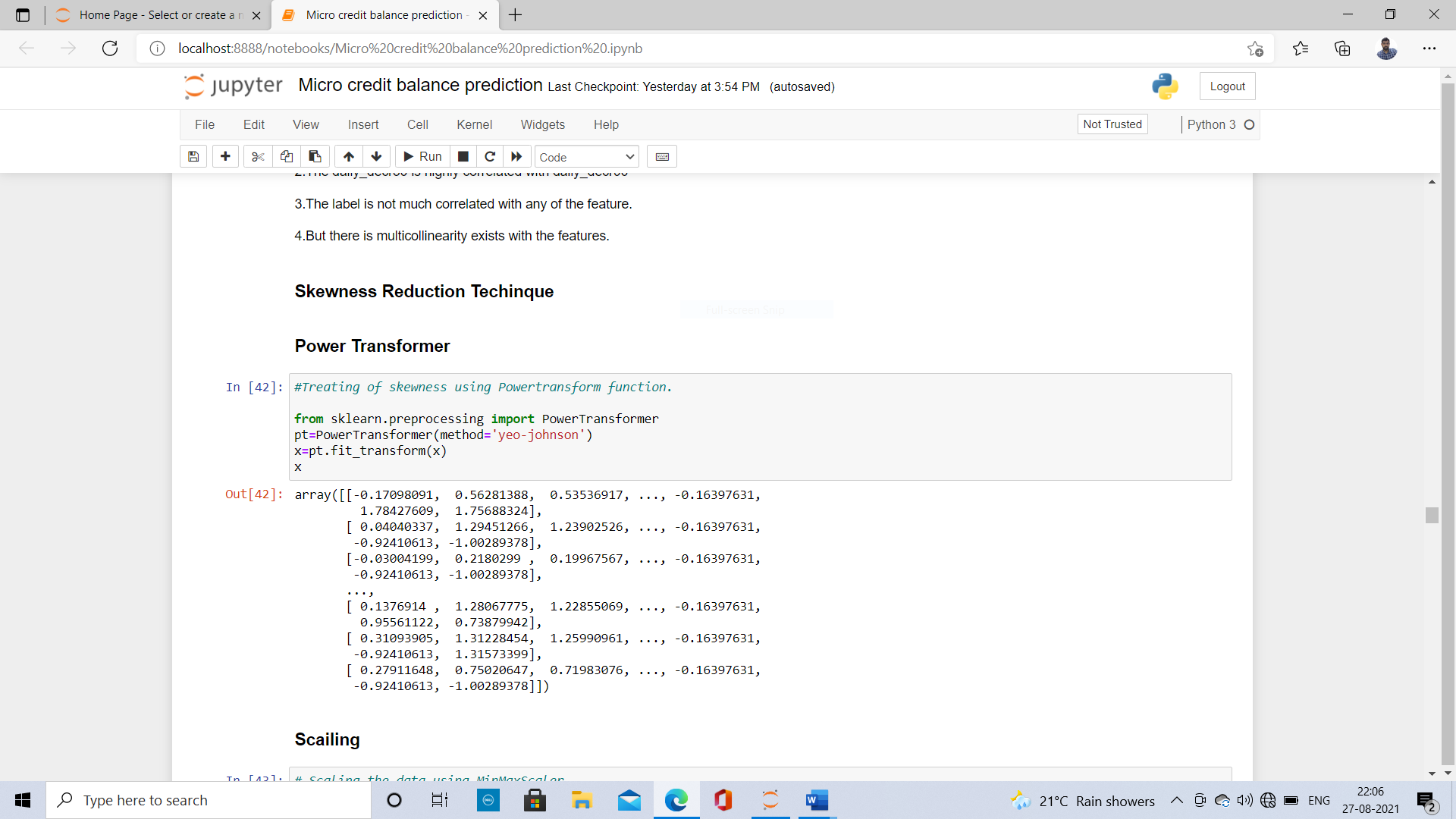
**Model/s Development and Evaluation**

Identification of possible problem-solving approaches (methods):

I have plotted Box plot and distribution plot on all the columns and checked the skewness and outliers present in the data after plotting I got to know that there is lots of skewness in the columns and presence of outliers in all the columns. So in order to remove the outliers I have applied Z-Score and removed the outliers which has max effect on the label prediction.

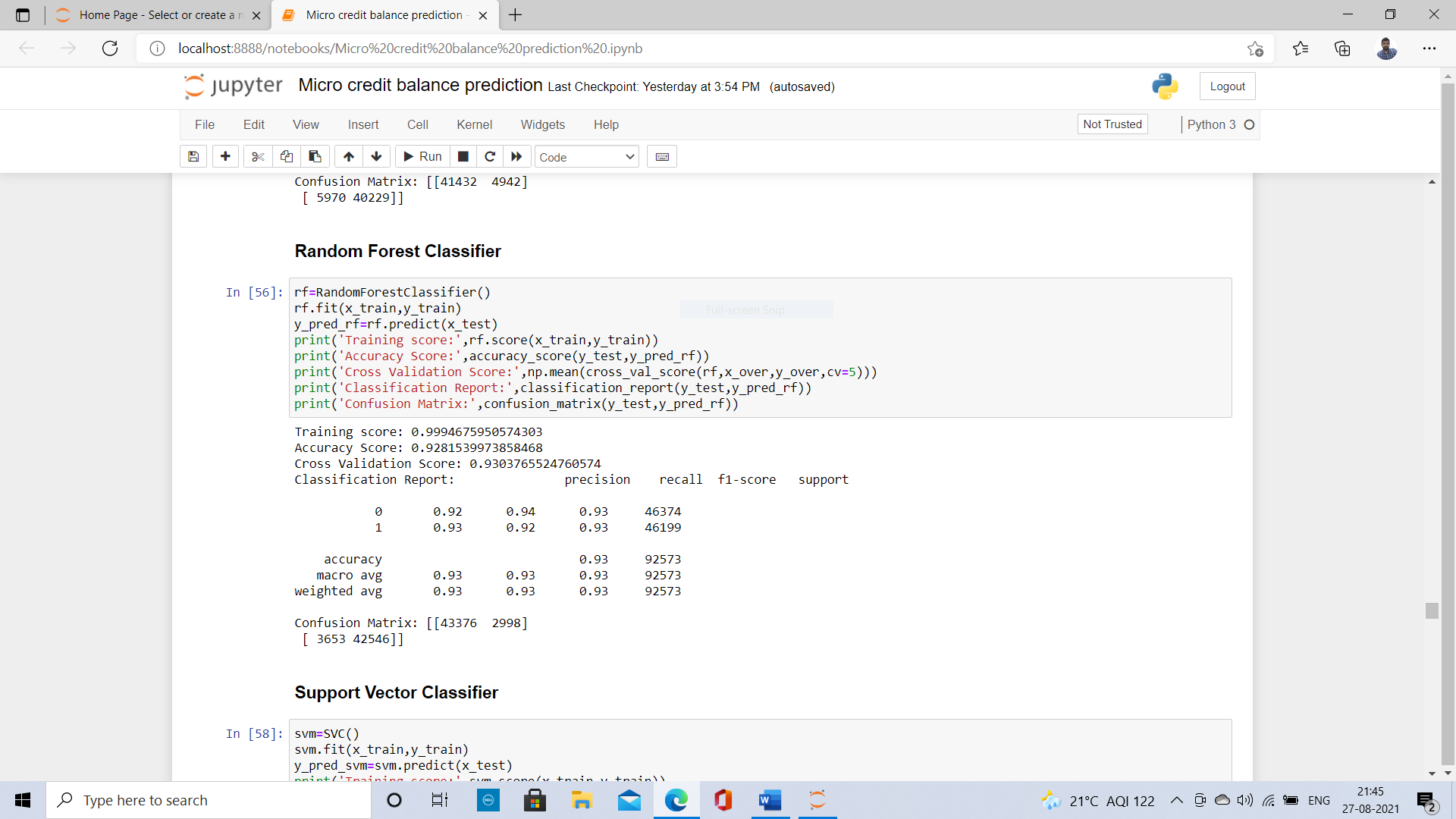


And I have applied power transformer for removing the skewness.



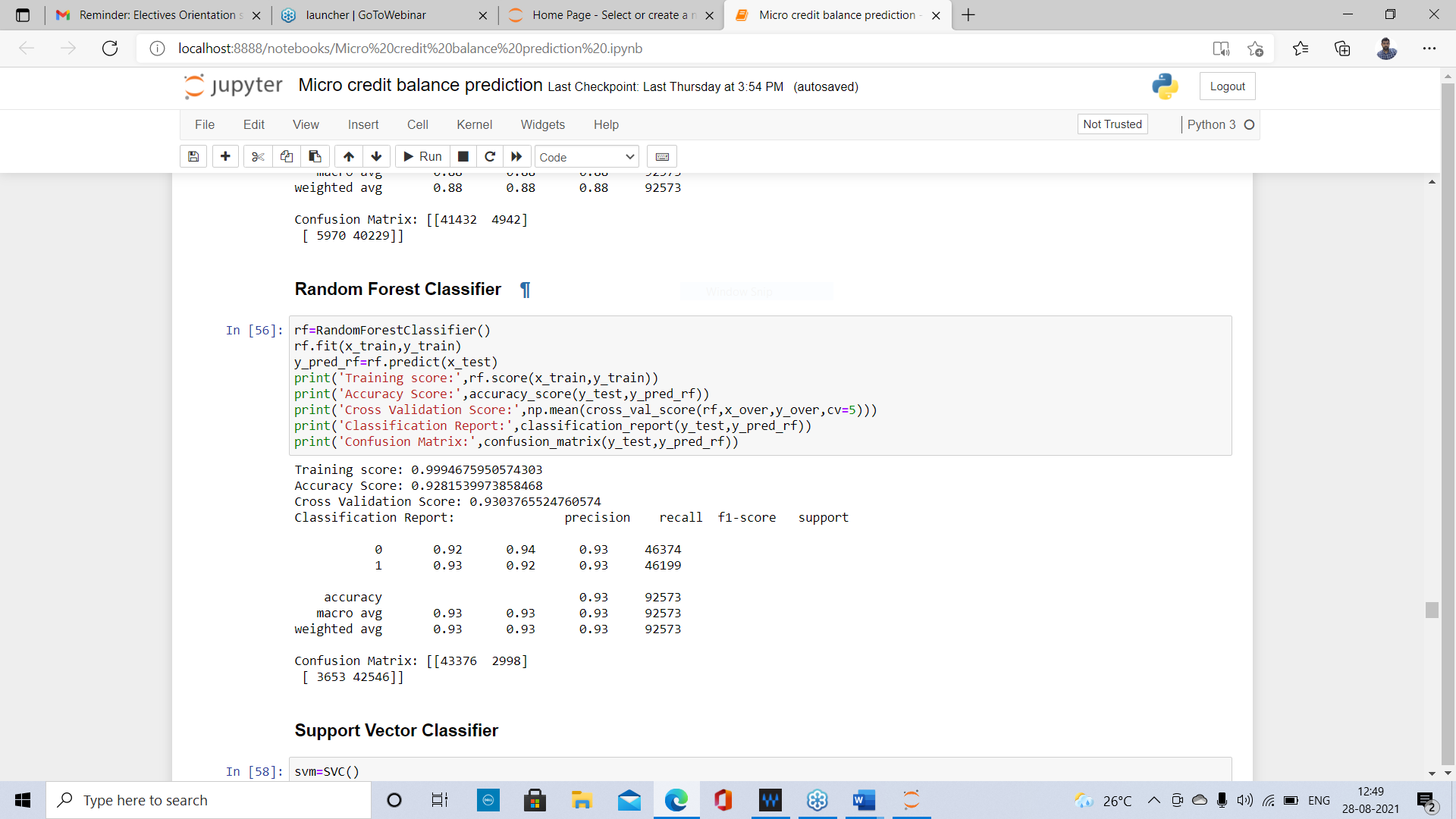
Testing of Identified Approaches (Algorithms):

Since the label has two classes this type of problem comes under binary classification. So I have used all the classification Algorithms to predict our label which gives the different patterns on different model. The Algorithms I have used are Logistic Regression, K-Neighbors Classifier, Decision Tree Classifier, Random Forest Classifier, Support Vector classifier and Gaussian Navie Bayes. The metrics I have used for evaluating the models are accuracy score, cross validation score, precision, recall, Confusion matrix and ROC AUC score.



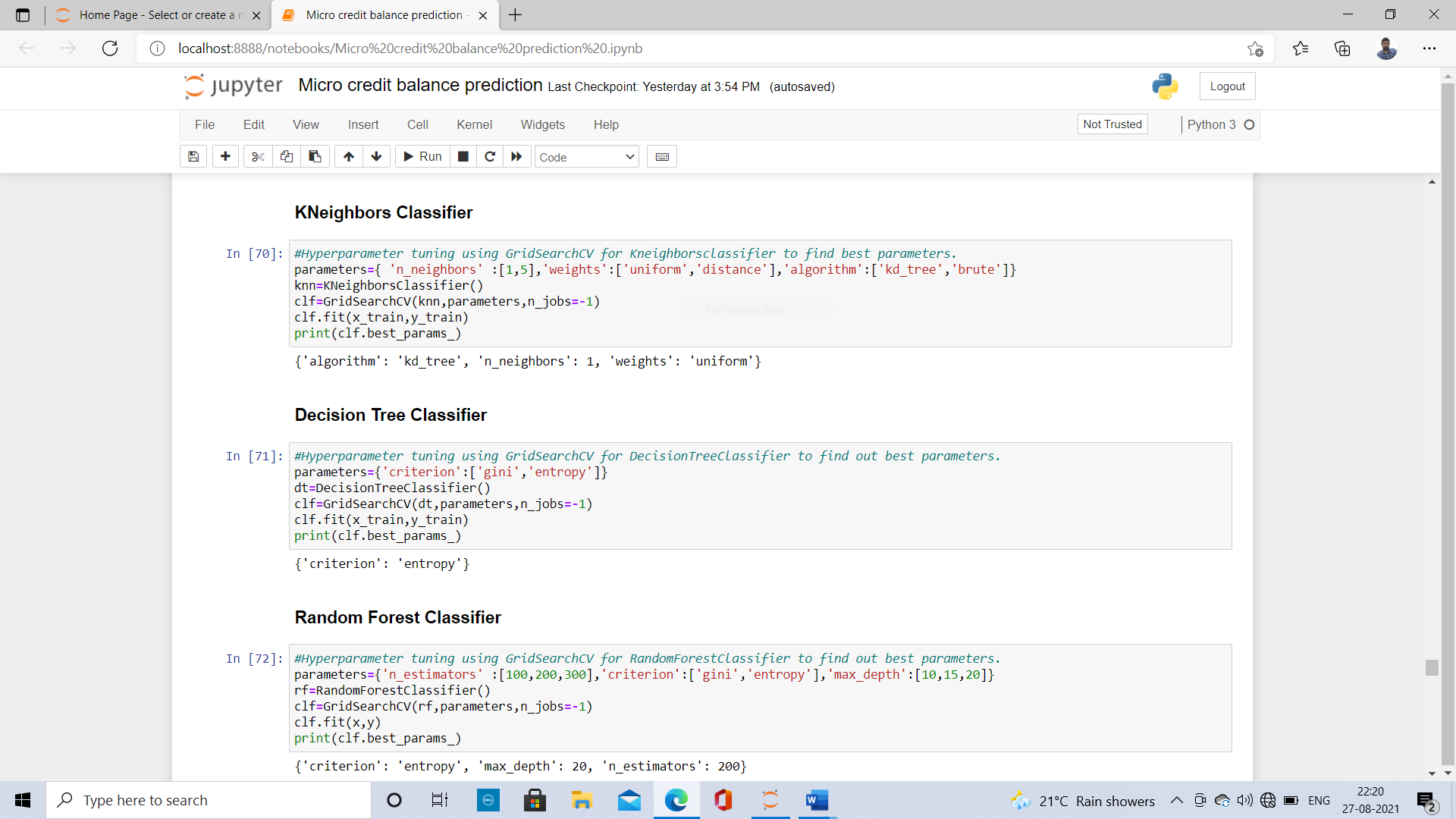
Run and Evaluate selected models:

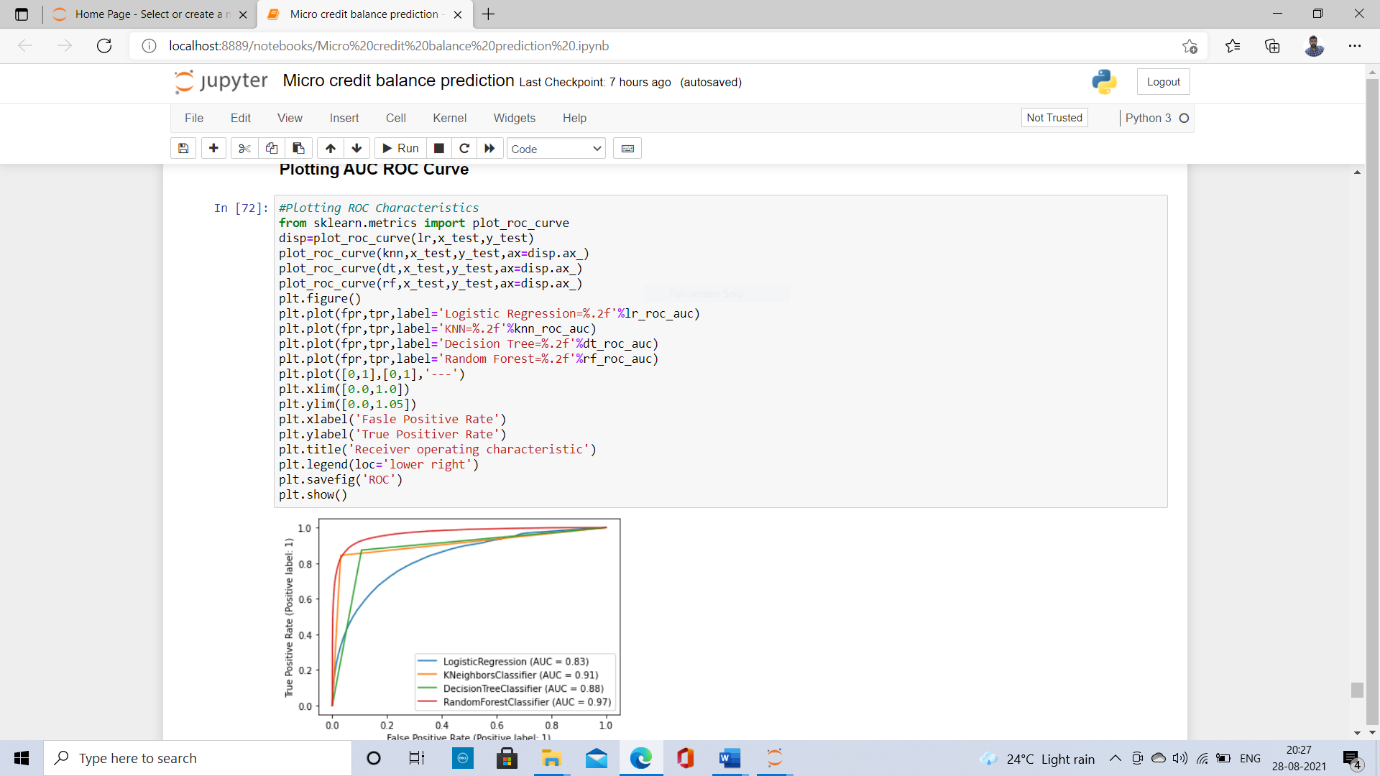
After splitting the features and label in x and y variables the columns in x are scaled then splitted x and y as train and test data using train test split. After that fitting, all the algorithms then used various evaluation metrics like accuracy score, cross validation score, classification report which includes precision, recall, F1 score which is the combination of precision and recall for two classes in label.



Key Metrics for success in solving problem under consideration:

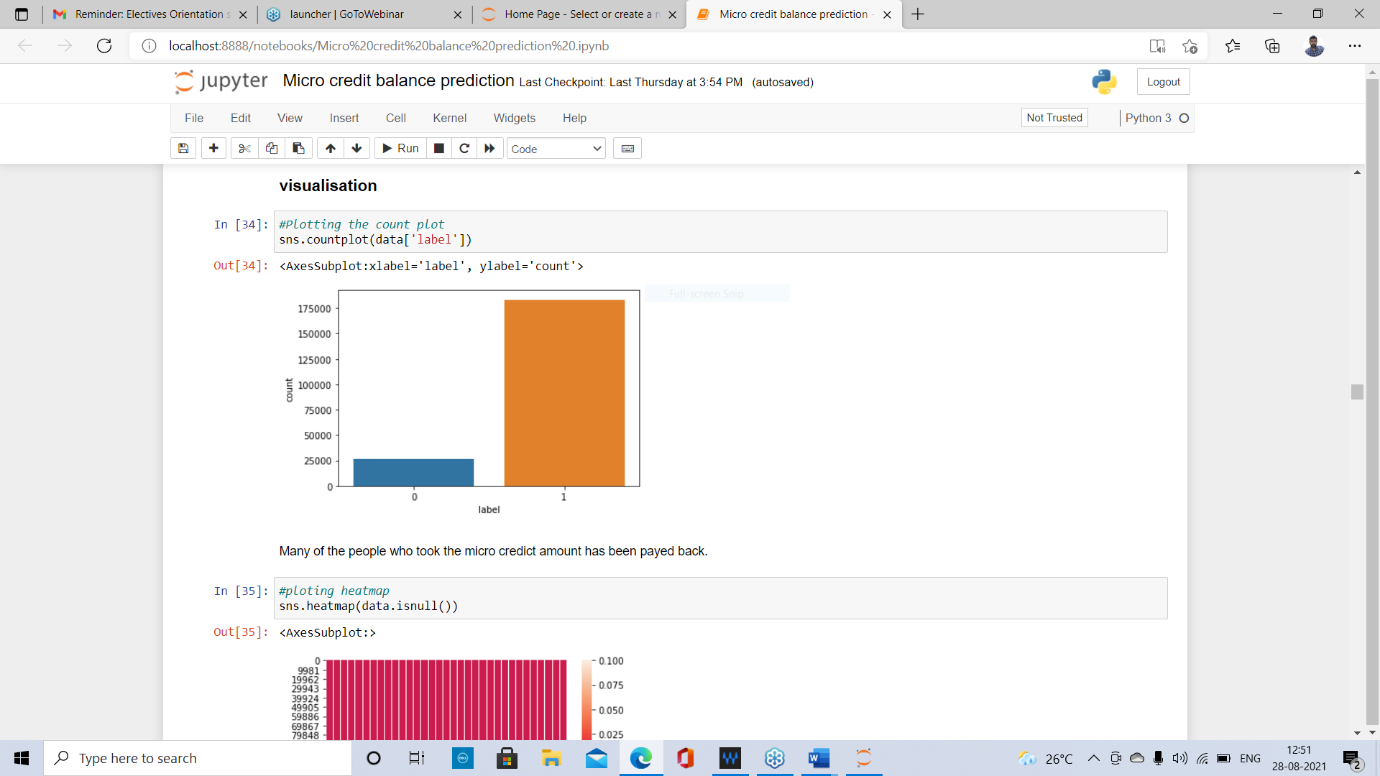
After Checking the models evaluation I have considered Logistic Regression, K Neighbors classifier, Decision Tree Classifier and Random Forest as my final models and I have applied hyperparameter tuning for improving the scores. By using Grid search CV we are going to pass different parameters for all the fitted algorithms which improves the evaluation of models.

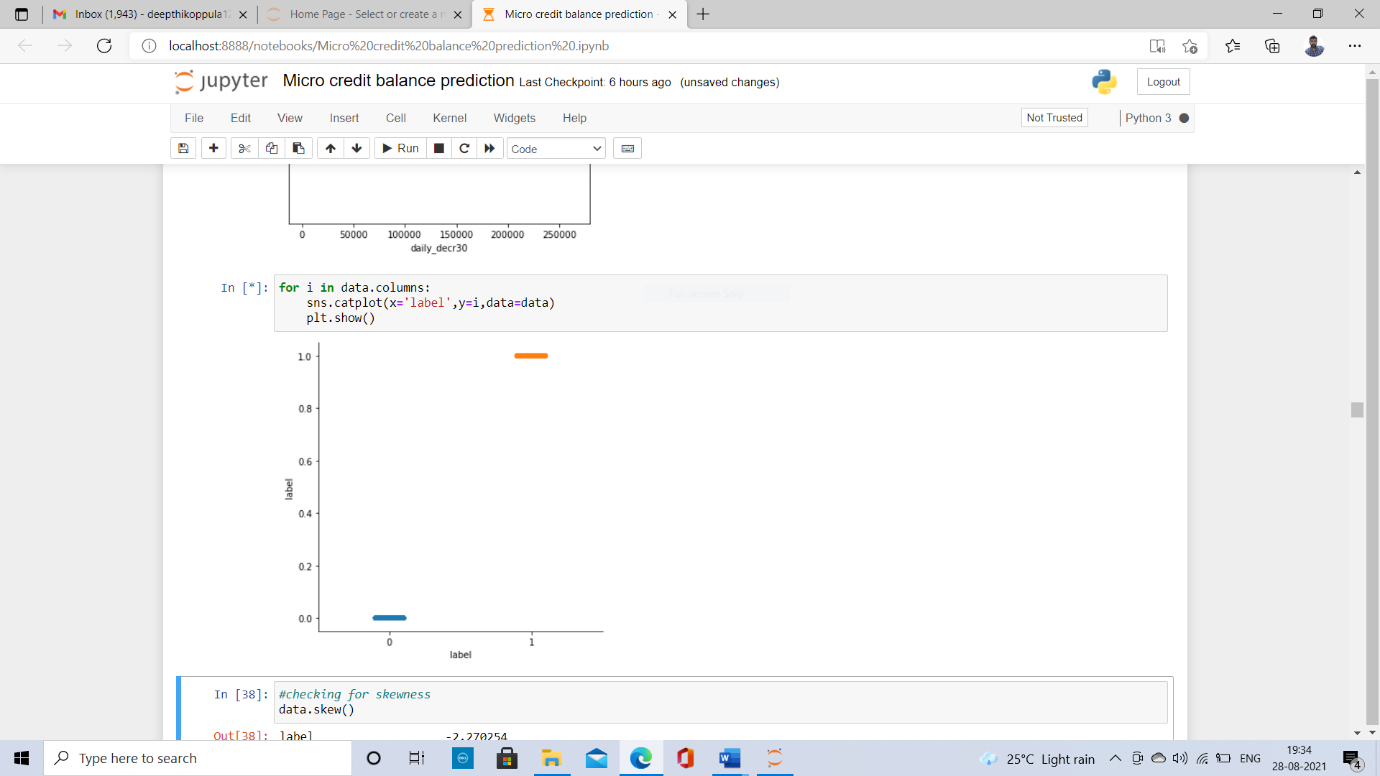




Visualizations:

First I have visualised for finding outliers and skewness of the data in the columns using box plots and distribution plots respectively. After plotting I found out that there are large number of outliers and skewness in the data of all the columns. I have tried to fix them using some techniques. Then I plotted count plot on label since it has two classes then I got to know that the label is unbalanced which indicates success and failure of loan payment. Then I made it balanced by using SMOTE. Then I have plotted heat map on is null method to check any null values in the dataset by plotting I got to know that there are no null values. Then I used to cat plot for checking the relation between all the features and label by using for loop on all the columns. But I found there is no much relation exists between the features and label. Then I have checked for correlation then obtained the percentage how much our label is correlated with all the features based on those values I plotted heat map and visualised.





Interpretation of the Results:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Index | Models | Accuracy score | F1- score | ROC AUC Score |
| 0 | Random Forest Classifier | 0.9138733756062782 | 0.92 | 0.9281403650674581 |
| 1 | KNeighbors Classifier | 0.9067654715737851 | 0.91 | 0.8665676027720839 |
| 2 | Decision Tree classifier | 0.8837133937541184 | 0.89 | 0.8821040440133816 |
| 3 | Support Vector classifier | 0.7727415121039612 | 0.78 | 0.7726922211921375 |
| 4 | Logistic Regression | 0.7579315783219729 | 0.76 | 0.8825779985183999 |

After all findings I got know that Random Forest classifier is the best fit model for our label prediction. So I have considered Random forest as my final model and saved it.

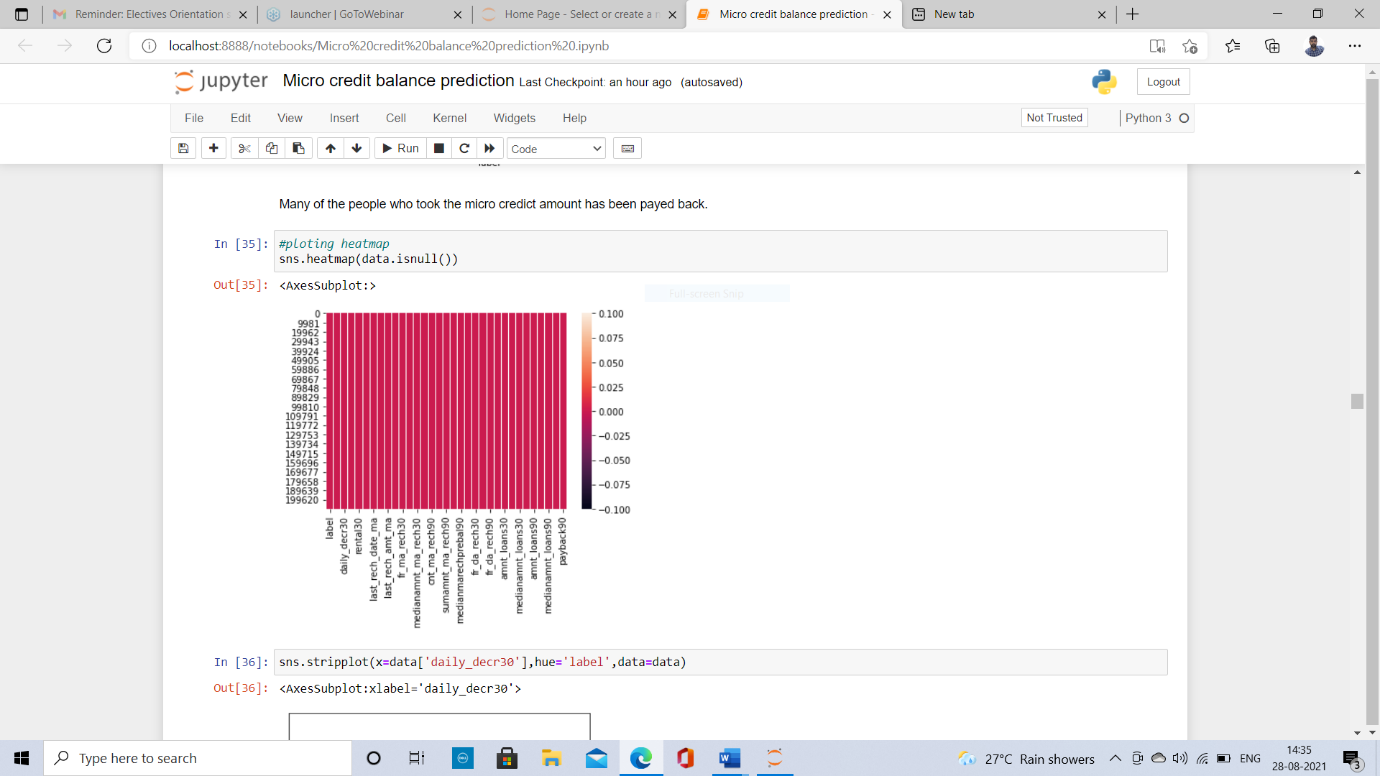
**CONCLUSION**

Key Findings and Conclusions of the Study:

Today, microfinance is widely accepted as a poverty-reduction tool. Providing, Micro credit loans to low-income population was a good idea, which was first implemented in Bangladesh in 1976, by Muhammad Yunus, who is a professor(economist), who was started with a group of women borrowing $27 to finance the group's own small businesses. The women repaid the loan and were able to sustain the business. Now various financial institutions, banks, NGO’s are coming up with great credit facilities in many sectors such as agriculture, small-scale business, telecom service etc. By providing such type of schemes for low-income people like daily wagers, small scale industries, etc… so that the people can sustain themselves in their business or fields or they can fulfil their basic needs. So the micro financial institutions must increase their Donors for funding them and check out themselves weather they can provide all such type loans for people who will apply for that. They should be in a position to provide all type loan sanctions to low-income people. If and when needed, governments should adjust regulatory frameworks to permit all types of financial institutions to offer services to poor people. The Government has to conduct many programs on micro credit loans for the people to get awareness in such type of getting credits.

Learning Outcomes of the Study in respect of Data Science:

After importing the dataset, I have checked for the Null values but there are no null values in the dataset. Even by using heat map I have visualised the data for null values but there are no null values.



For checking the outliers and skewness I have plotted boxplots and distribution plots for checking of data distribution in each column. After that I found that each and every column has lots of skewness and outliers present. To remove the outliers I have applied Z score on all the features.

There are presence of outliers and skewness in our label but we cannot deal with it since we have to predict that remain same so I treat the label column, but when passed all the features for applying Z score there is almost 12% data got lost, so since our data is more expensive, I have treated the columns which are having the outliers at the maximum point, after performing I have lost the data of 7.8% which is minimal. Then for removing skewness I have used power transformer and tried removing the skewness, after that I have scaled the data since all the columns as different data our machine learning models may not perform better so I have scaled the data. Since our label is imbalance, I have used one of the up-sampling technique called SMOTE for balancing the data, In that our classes of the label gets balanced.

Then I have splitted the variables into train and test data and fixed the random state and splitted the train data as 70% and test data into 30%. Where our models predict on test data. Since our label has 2 classes it comes under binary classification, I have used classification algorithms for fitting the data. After that I have checked the patterns of each model and their performances. Other than logistic regression all the other models tooks lots of time to fit the training data since we have huge data. Where I have waited almost hours of time. Compared to all other models Support vector classifier took more than 7 to 8 hours to execute. So after that I have used Grid search CV for improving the performance of the models on various parameters which is also a time consuming process. After that I have evaluated the model performance on best parameters after that I have checked ROC AUC scores and Roc curves for all the executed models.

Finally after all evaluations the Random Forest Classifier is giving the best patterns and good model performance compared to all the other models. So I have considered Random Forest as my final model.

Limitations of this work and Scope for Future Work:

Machine learning offers a high level of modelling freedom, it tends to overfit the data. A model overfits when it performs well on the training data but does not perform well on the evaluation data. In the dataset our data is not properly distributed in some of the columns many of the values in the columns are 0’s which are not realistic. Because I have seen in some of the columns even the person didn’t take loan but the label says that he paid back the loan amount, which is not correct. So because of that data our models may not make the right patterns and the performance of the model also reduces. So that issues need to be taken care then our prediction may be much more-better than this. In Future if we fix this type of problems then our model performance will be much better than now what we predicted.