FLIP ROBO

Micro-Credit Defaulter Mode

Submitted by:

SWATHI R

ACKNOWLEDGMENT

The internship opportunity I had with Flip Robo was a great chance for learning and professional development. Therefore, I consider myself as a very lucky individual as I was provided with an opportunity to be a part of it.

I would like to thank our SME for suggesting this project and for his whole hearted cooperation and constant encouragement throughout the project.

And I also like to thank the data trained mentors and Technical team members for helping me with the technical queries.

And these are the following website which I referred for the reference

1. https://www.kaggle.com/
2. https://scikit-learn.org/
3. www.stackoverflow.com
4. www.google.com
5. www.geeksforgeeks.org

INTRODUCTION

- Business Problem Framing

  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. They are collaborating with an MFI to provide micro-credit on mobile balances to be paid back in 5 days. 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.

- Conceptual Background of the Domain Problem

  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

  From the dataset description I have leant the nature of data.

- Motivation for the Problem Undertaken

  From this project I get to know of different kind of information every recharge done by the user on which kind of recharge user is using mostly and the data service or the main balance the frequency of recharge in 30 day or 90 days. It is really quite interesting to know that each column contributed to make you close to know more about the data and in prediction you can do in many ways

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  After uploading the data I get to know the data by the data.describe() so many information the min value , max value, SD the 25 percentile the 50th percentile the 75 percentile of the data. Then by the help of .skew() I get to know the skewness of the data. Then by the help of correlation function I get to know the correlation of each columns with each other. From the heatmap I can visualized to see them clearly that they are positive correlated or the negative correlated the dark side is show the negative correlation among each other the lighter side represent the positive correlation among the each other.

  .

- Data Sources and their formats

  Data I get form the Flip Robo the format was in CSV (Comma Separated Values).The number of columns and row are 209593 and columns are 36.

  The data descriptions are as follow:-

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 Indonesian Rupee)
medianamnt_ma_rech90	Median of amount of recharges done in main account over last 90 days at user level (in Indonesian Rupee)
medianmarechprebal90	Median of main account balance just before recharge in last 90 days at user level (in Indonesian Rupee)
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

The raw data is taken and performed various steps to reduce skewness, outlier, class imbalance and scaling. There were no null value was present in the dataset but there are some outliers which also get too removed. Many outlier removal and skewness removal methods are tested and best method Is chosen in order to prevent data loss.

- Data Inputs- Logic- Output Relationships

The input data contains 209593 rows and 36 columns.

Label (target variable) depends on all the features of 30 or 90 days mobile loan/repay .

Pcricle,Pdate and msisdn are removed from the table since that is not much effective in predicting the target variable.

- Hardware and Software Requirements and Tools Used

Hardware – Laptop

Software – google colab, jupyter notebook

Libraries- numpy, pandas, seaborn, matplotlib.pyplot,sklearn.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  Classification Model with following algorithms

  - KneighborsClassifier
  - LogisticRegression
  - DecisionTreeClassifier
  - GaussianNB
  - SVC

  Evaluation metrics

  - Accuracy score
  - Precison, recall
  - AUC,ROC
  - F1 score

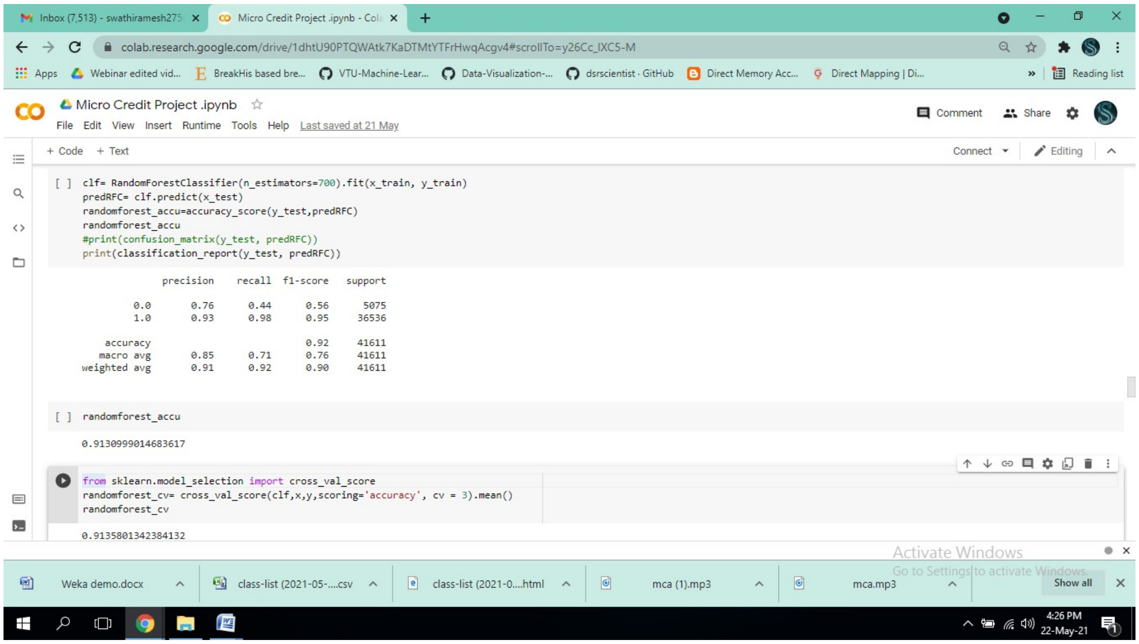
- Testing of Identified Approaches (Algorithms)

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

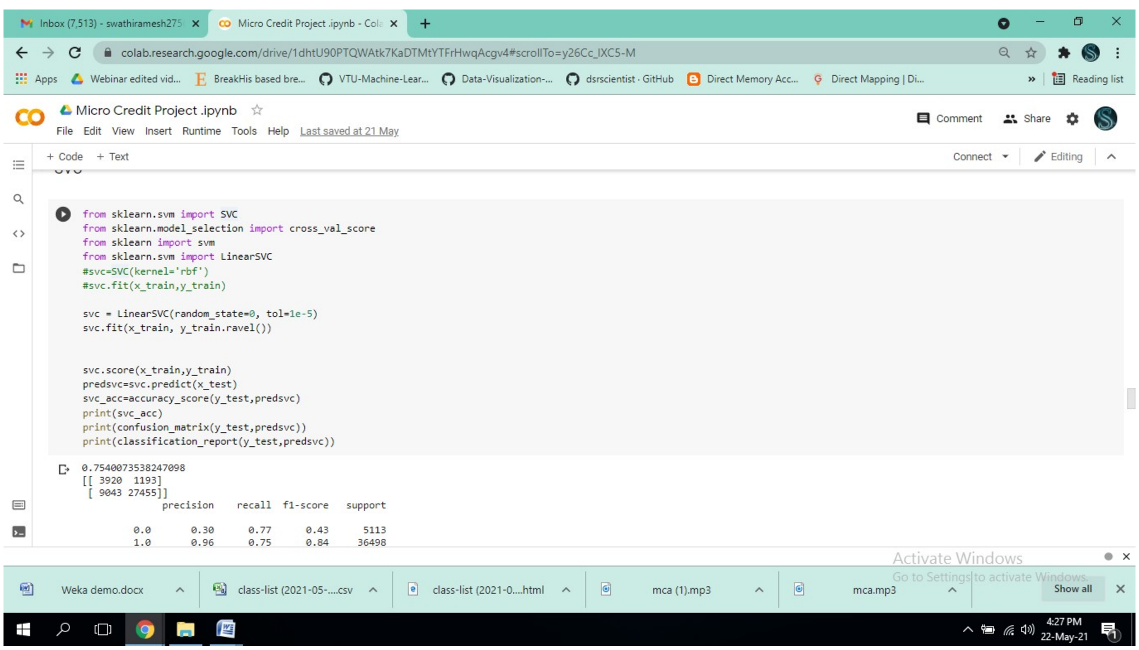
  - KNN=KNeighborsClassifier()
  - LR=LogisticRegression()
  - DT=DecisionTreeClassifier()
  - GNB=GaussianNB()
  - Svc= LinearSVC()

- ## Run and Evaluate selected models

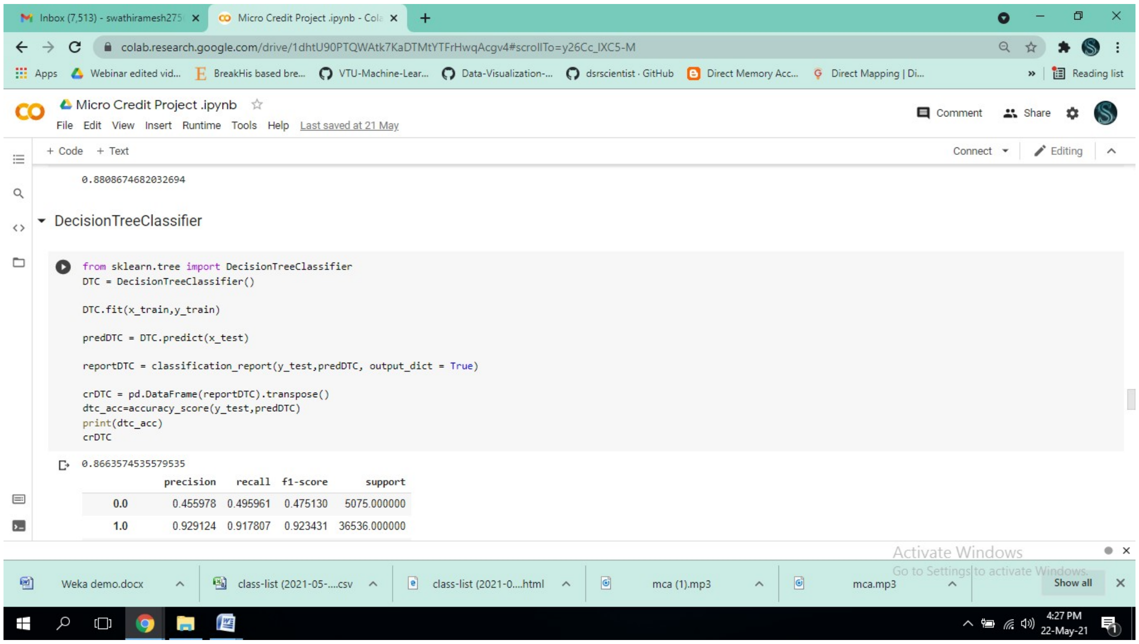
  Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.



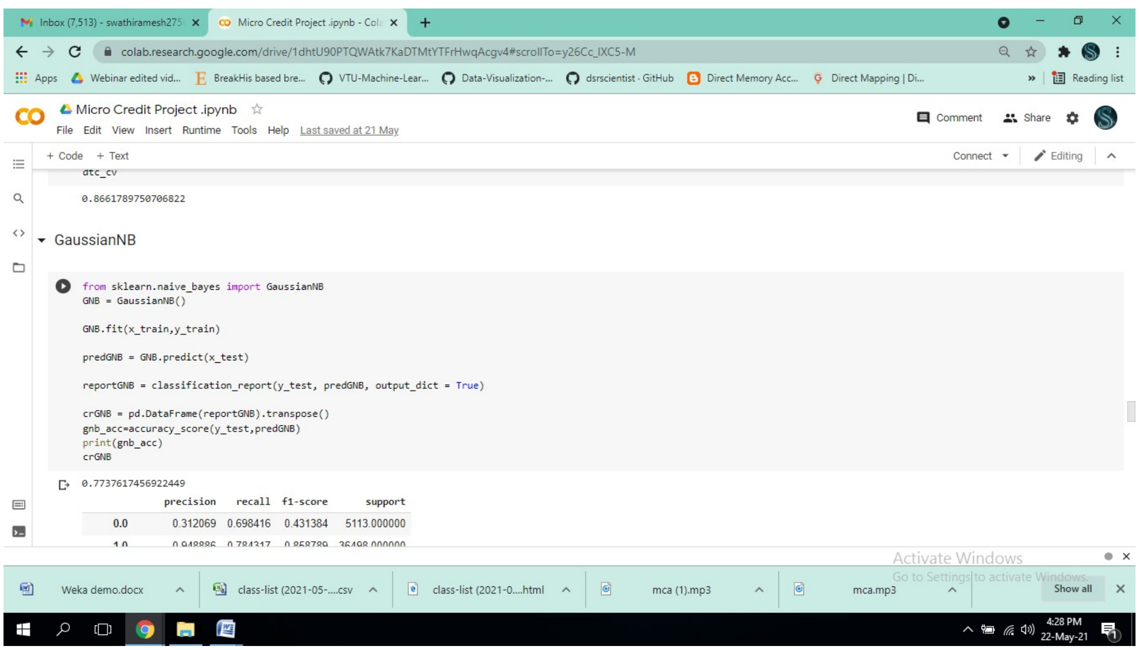
1.Random Forest



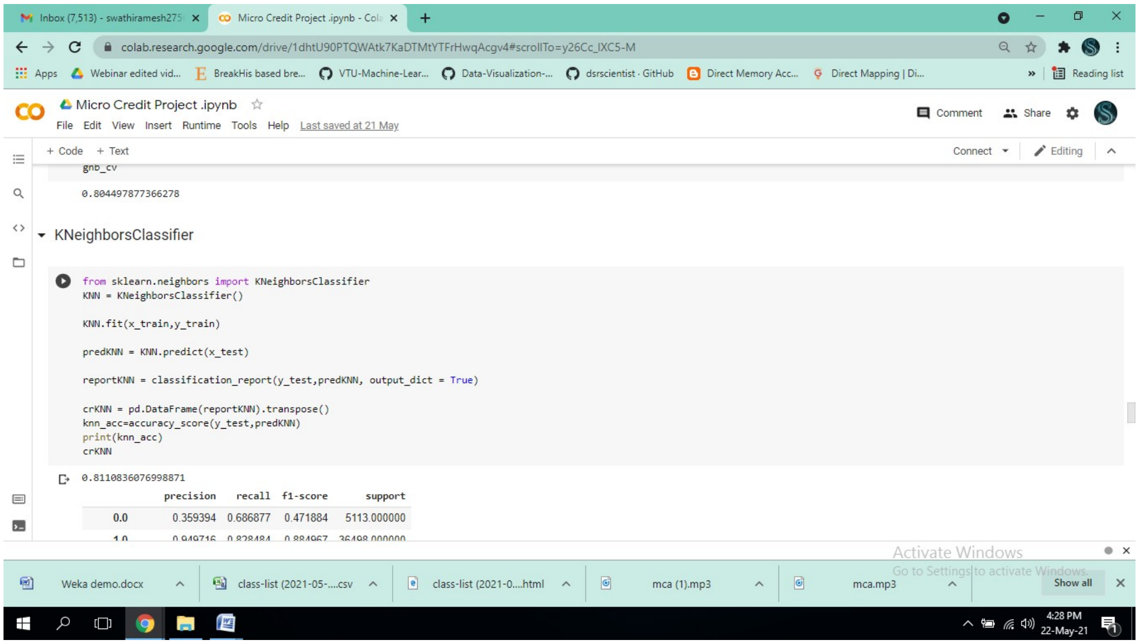
2.SVC



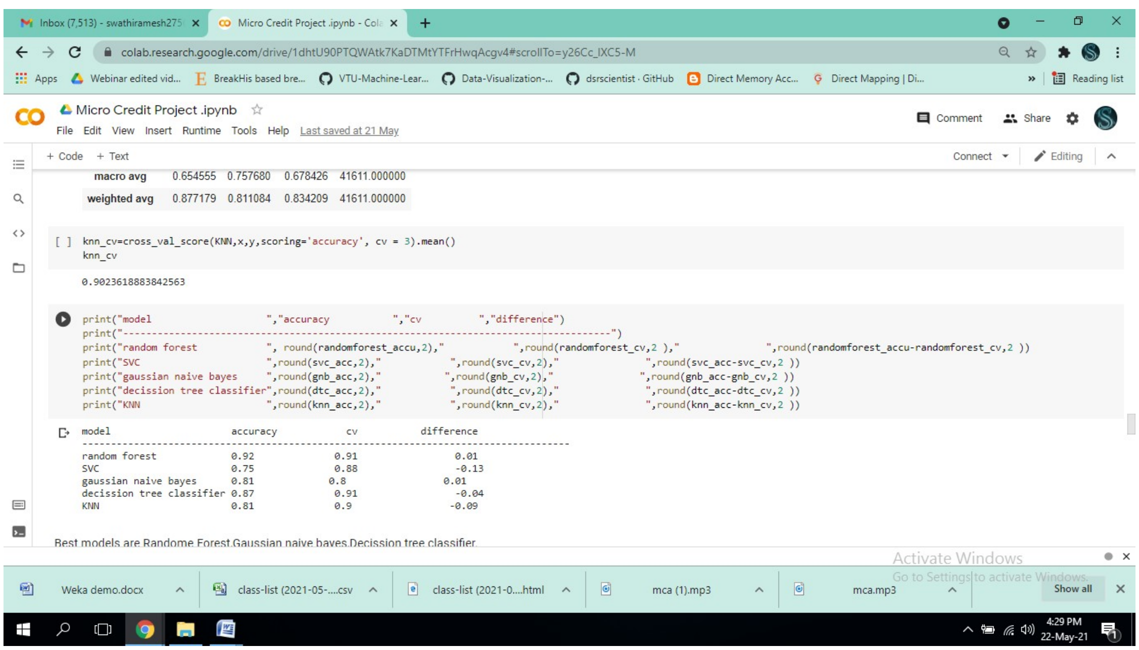
3. Decision Tree



4.Gaussian NB



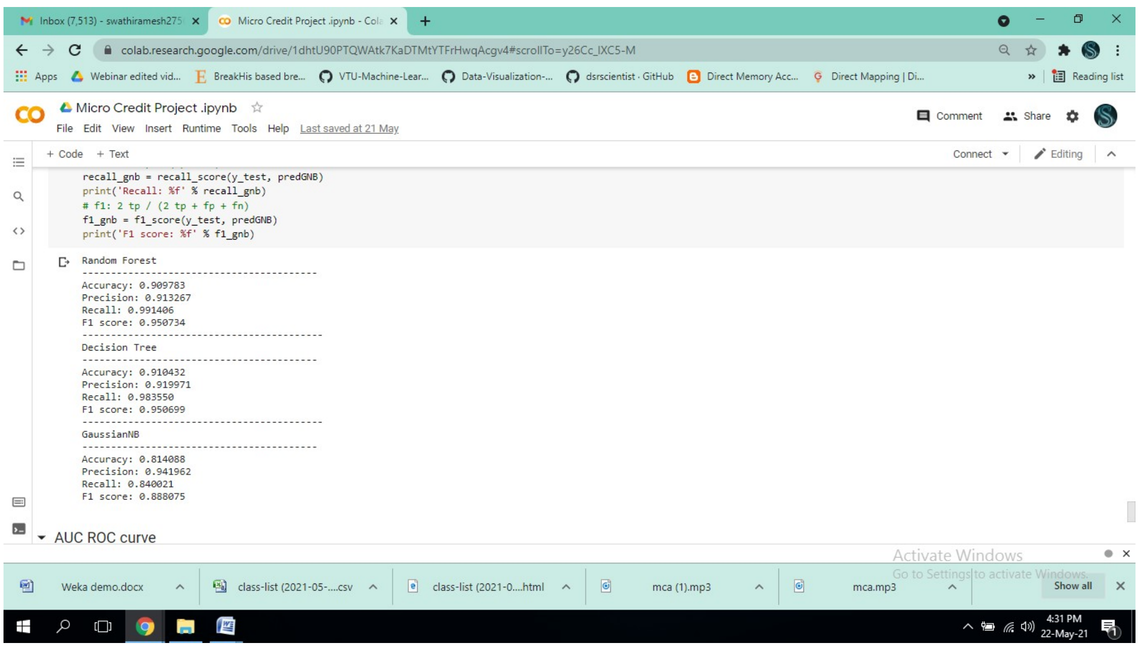
5.KNN

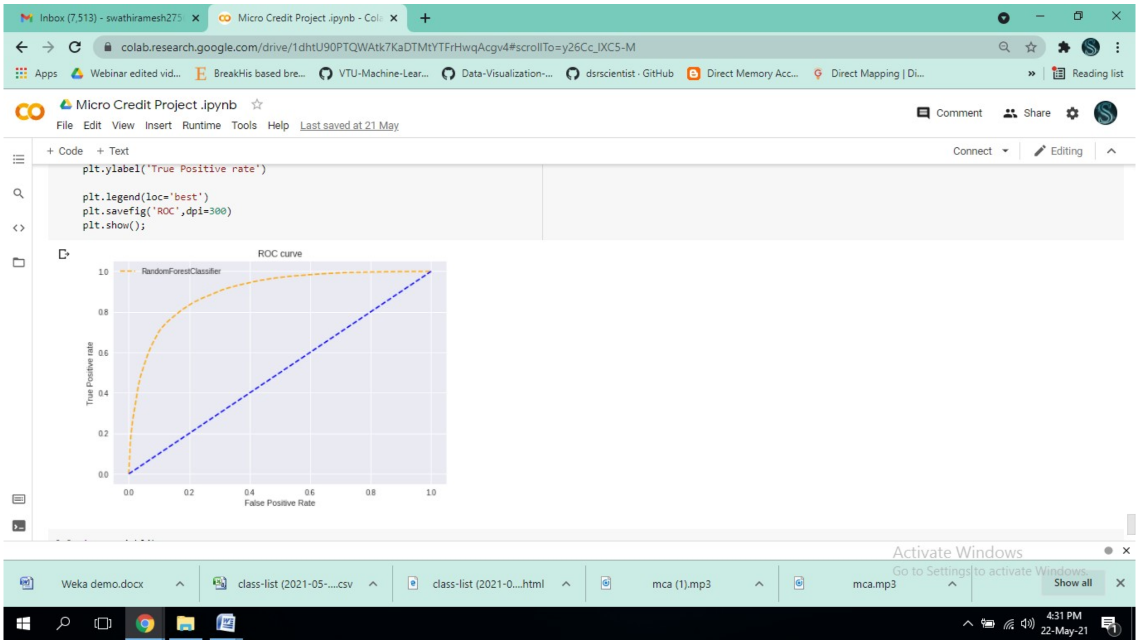


Evaluation using Accuracy score and cross validation.

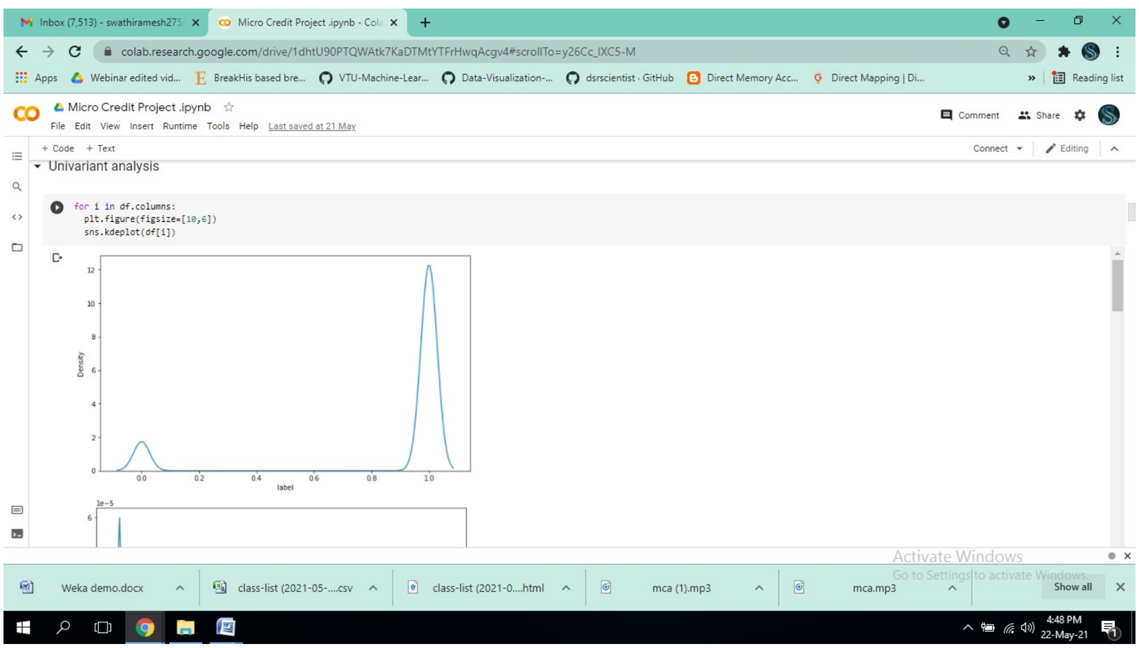
- Key Metrics for success in solving problem under consideration
- **Precision**: can be seen as a measure of quality, **higher precision** means that an algorithm returns more relevant results than irrelevant ones
- **Recall** is used as a measure of quantity and high recall means that an algorithm returns most of the relevant results.

- **Accuracy score** is used when the True Positives and True negatives are more important. **Accuracy** can be used when the class distribution is similar
- **F1-score** is used when the False Negatives and False Positives are crucial. While F1-score is a better metric when there are imbalanced classes.
- **Cross_val_score** :- To run **cross-validation** on multiple metrics and also to return train **scores**, fit times and **score** times. Get predictions from each split of **cross-validation** for diagnostic purposes. Make a scorer from a performance metric or loss function.
- **roc _auc _score :-** ROC curve. It is a plot of the false positive rate (x-axis) versus the true positive rate (y-axis) for a number of different candidate threshold values between 0.0 and 1.0

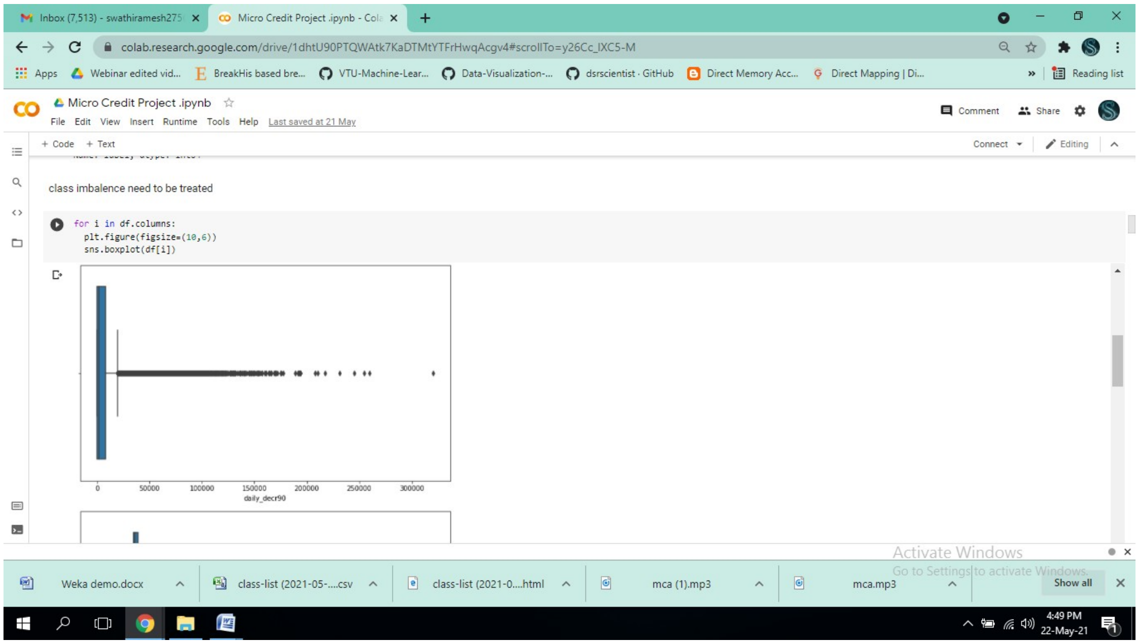




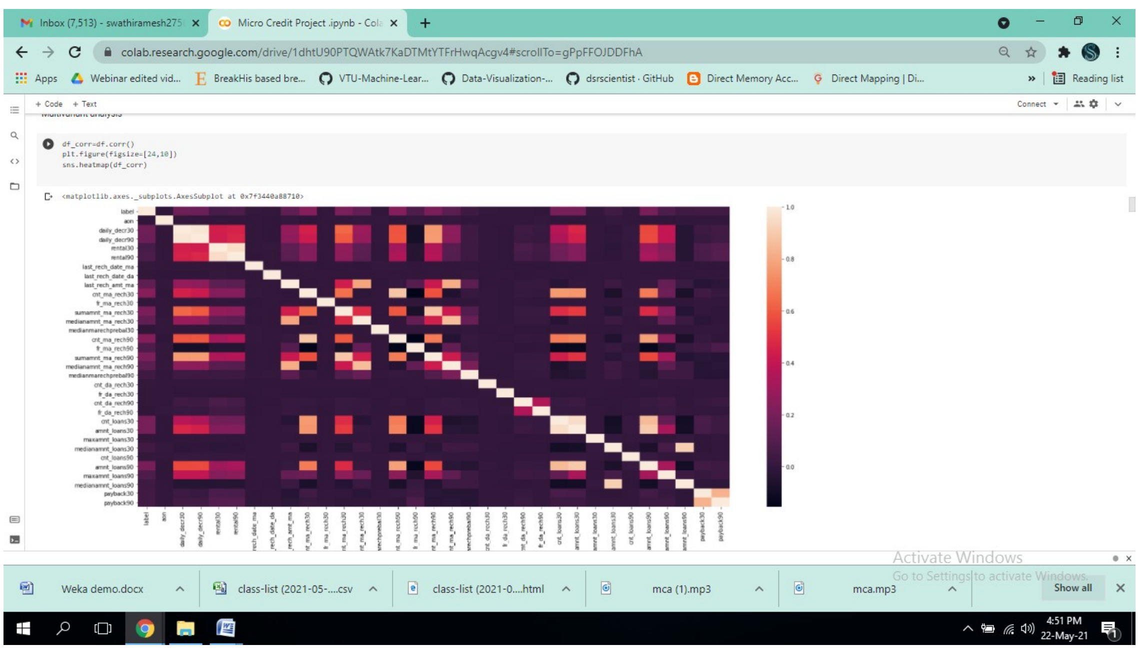
- Visualizations



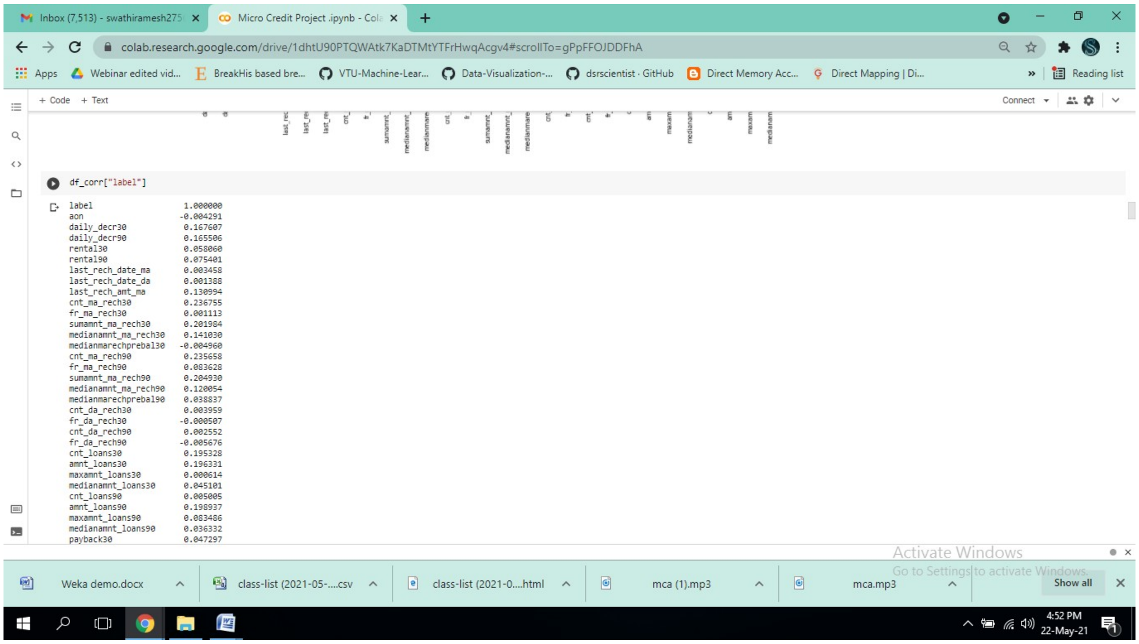
Kdeplot is made for all the columns to check the distribution



Box plot is made to check the outliers and distribution of data



Heat map is drawn to visualize the relationship among the variables.



To check the data which are highly correlated with label.

- ## Interpretation of the Results

Data is highly skewed from the mean so skewness removal methods need to be followed.

Box plot tells that there are many outliers are present in the data so need to remove the outliers too.

Heat map tells that which data is highly correlated with class are more important in constructing the model.

CONCLUSION

- ## Key Findings and Conclusions of the Study

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. Importance of removing the skewness and outlier is important. Finding the best parameters for the algorithm also plays a important role in performance and accuracy of the model.

- ## Learning Outcomes of the Study in respect of Data Science

  Learnt how to process the large number of data. Tried and learnt more about distribution of the data. The power of visualization is helpful for the understanding of data into the graphical representation its help me to understand that what data is trying to say, Data cleaning is one of the most important step to remove missing value or null value fill it by mean median or by mode or by 0.Setting a good parameters is more important for the model accuracy. Finding a best random state played a vital roll in finding a better model.

- ## Limitations of this work and Scope for Future Work

  The techniques to increase the speed of the model need to be constructed. The future model can be constructed with the most co related data with the target variable in order to increase the speed of the model.