

Micro Credit Project

Submitted by:

Mandar M Joshi

**ACKNOWLEDGMENT**

Micro Credit Project is a project with a finance domain so domain knowledge helps to understand the column values more effectively. Also study materials and training provided for this projects help to understand the basic requirement and actual predictions for the project.

**INTRODUCTION**

* Business Problem Framing

Micro Credit defaulter model is basically used get prediction to provide loan for its customer or not on the basis of different features.

* Conceptual Background of the Domain Problem

Domain knowledge helps to understand to add or remove features as per the business requirement. In order to only feed features that seems to be important to train and check out with the predictions for the new data point.

* Review of Literatur

Mainly the research done for this particular project is to understand the dataset as it contains near about 2lakh rows and 37 column. Trying to find out what exactly data is given in order to reach the goal by doing some domain expertise knowledge thinking and using open source search on the same.

* Motivation for the Problem Undertaken

The best motivation to carry out this particular project is basically to create actual model on a given dataset which predict either this particular candidate can eligible for the loan or not based on the past experience.so I will the person who create a model to achieve the goal is the best motivation for me for this particular project.

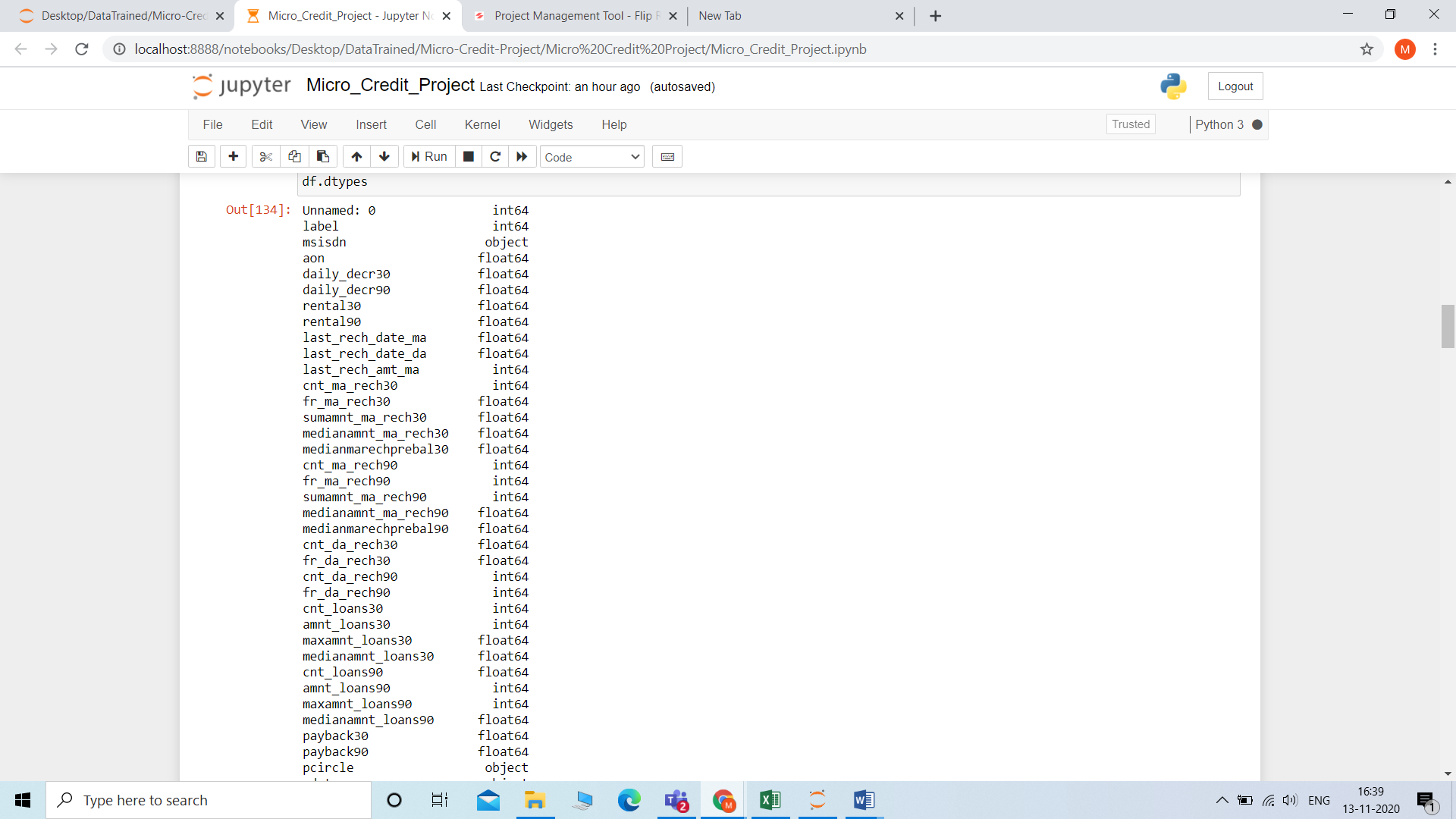
**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

As this project has a huge number of rows which includes the outliers it is necessary to check the zscore in order to remove the outlier.also need to settle down the data at a single node in order to help the machine for the training.as this completely involved

* Data Sources and their formats

In the given data there are 3 kind of a columns which contains some values in a categorical and numerical data which actually don’t require so its better to delete the entire column in order to make machine more effectively.



* Data Preprocessing Done

Data preprocessing done by deleting non important column,removing outliers,dealing with the null values(but there are no null values are present in this dataset),to scale the data at a single node, to remove outliers if any

* Data Inputs- Logic- Output Relationships

The basic of the data input and output is we have to give input data only in the form of numbers which can be further read by the model.There are some categorical values are present,strings or any other datatypes are present to in order to make model to predict accurate prediction need to train a model on right features to avoid wrong predictions.

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

No assumptions

* Hardware and Software Requirements and Tools Used

Hardware Requirements:-We have to use minimum 8gb ram laptop of mine but suggesting to work on Google collab as the size of the dataset is pretty larger.

Software Requirements:- I have used numpy,pandas,scipy liabraries extensively for data preprocessing. Used matplotlib and seaborn library for the graphical representation.Scikit-Learn library to choose a model.also used some data preprocessing techniques such as preprocessing,decomposition library.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

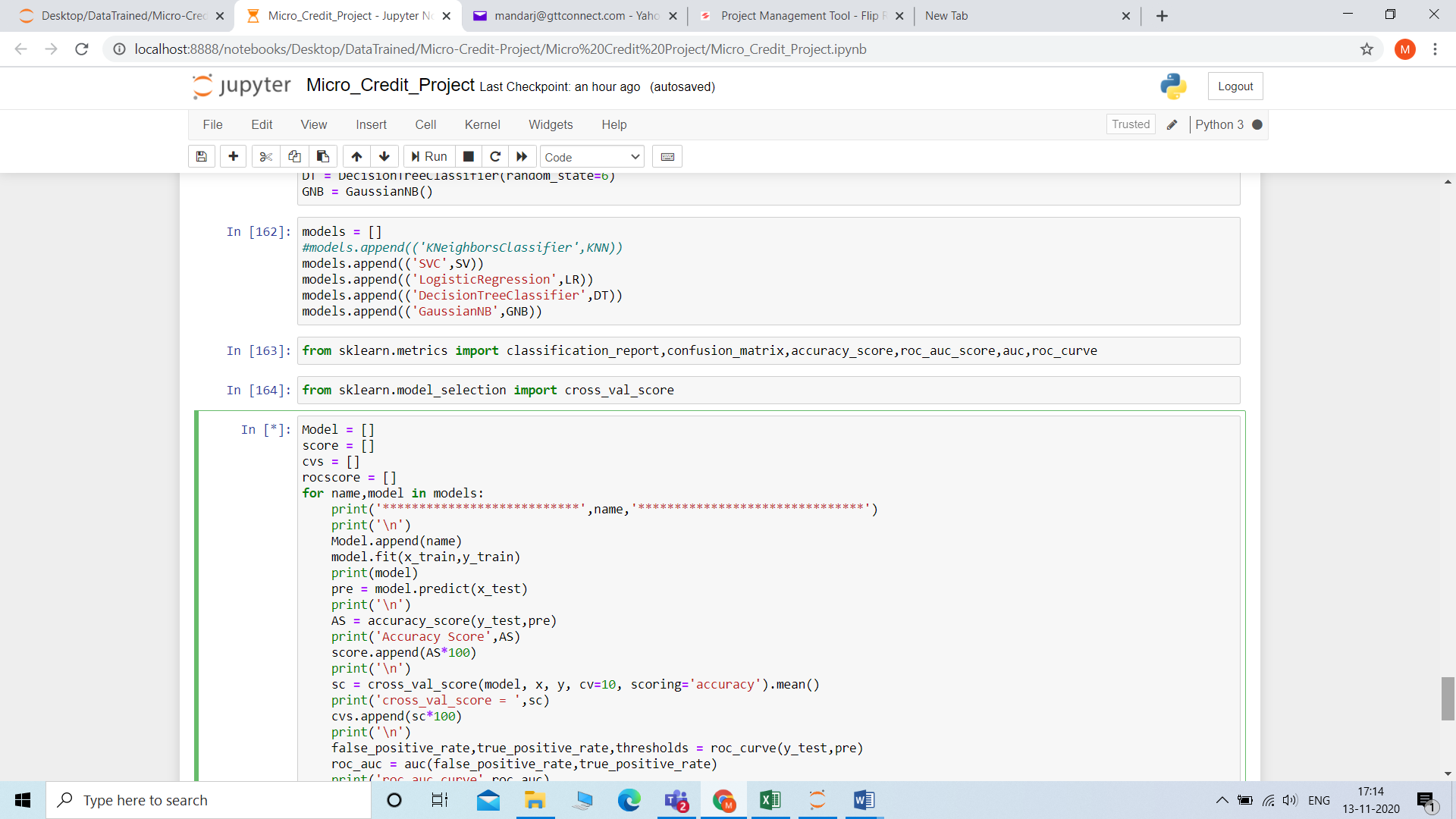
Problem solving approaches I keep in mind to feed a cleaned data to a model use multiple algorithms.

* Testing of Identified Approaches (Algorithms)

Model is still performing a training so I have already given algorithms such as svm,naïve bayes,logistic regression,decision tree classifiers.

* Run and Evaluate selected models

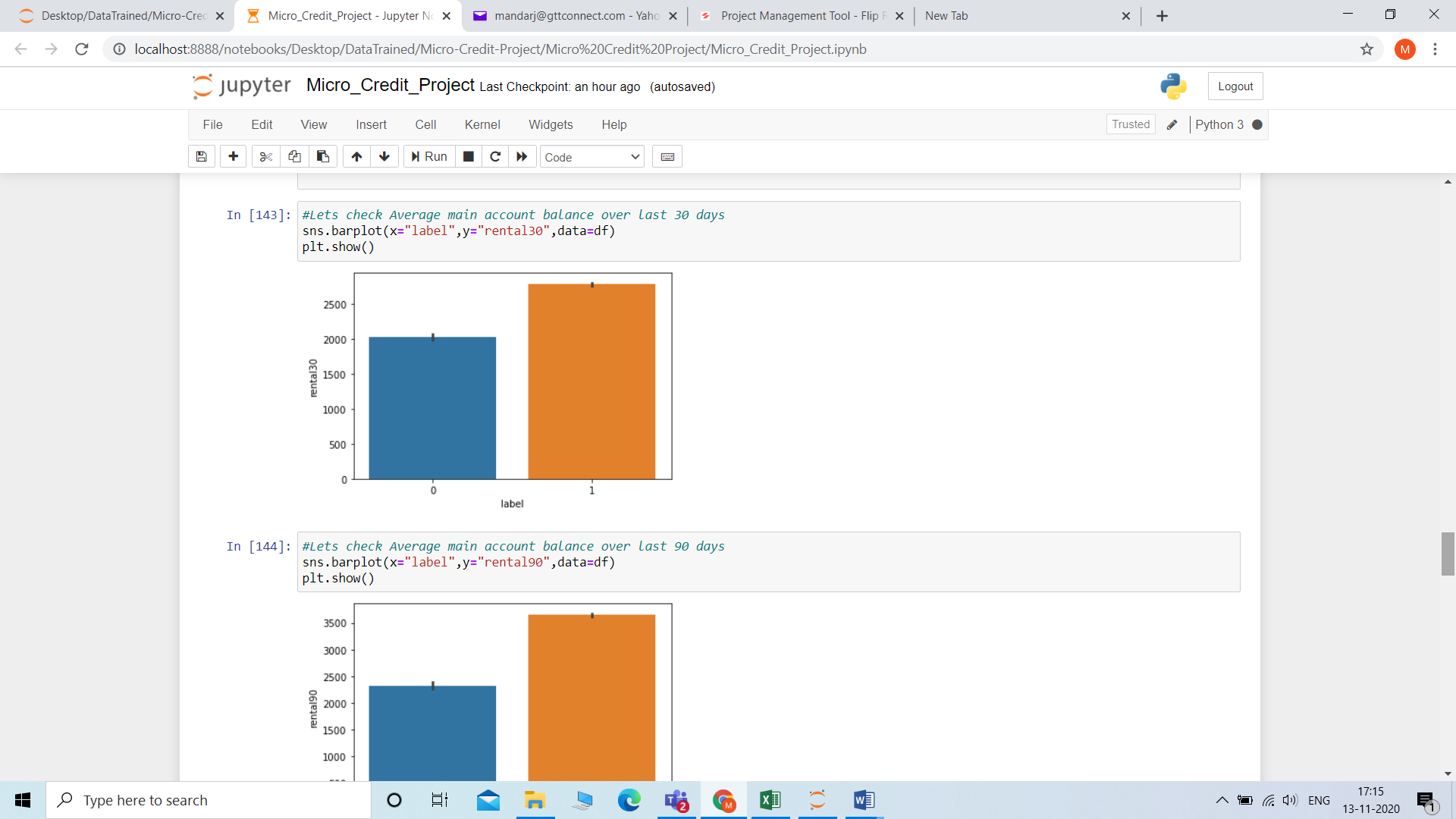
Using a for loop given a model with the various metrices such as auc roc curve ,accuracy score,cross val score,classification report etc



* Key Metrics for success in solving problem under consideration

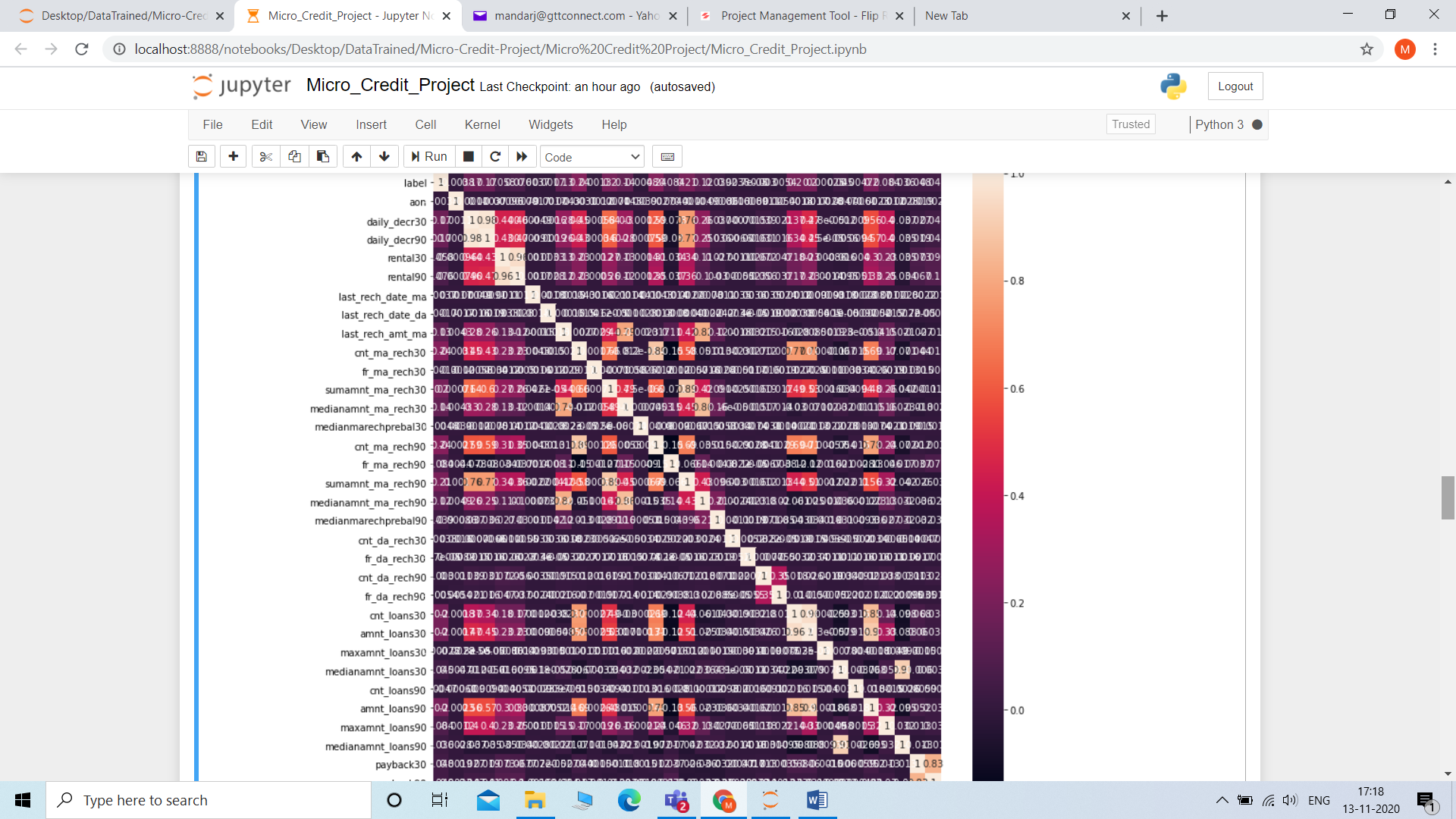
Made a assumption to avoid overfitting and underfitting if the accuracy score will be more than 85 percent then will go ahead with the regularization technic.

* Visualizations



Used bar plot to check out various parameters such as #Lets Average main account balance over last 90 days, Lets check Number of loans taken by user in last 30 days.

Used Heatmap to check the correlation of the columns to understand either the columns are co-related with each other or not



* Interpretation of the Results

Due to the limitation of some hardware requirement the model is still training (it is taking a very long time in training )but achieved my 1st model accuracy about 86 percent

**CONCLUSION**

* Key Findings and Conclusions of the Study

Model is predicting about 86 percent the right predictions.

* Learning Outcomes of the Study in respect of Data Science

Learn a lot from this project such as EDA,Data preprocessing,dealing with outliers,model selection,model evaluation parameters.

* Limitations of this work and Scope for Future Work

Only the factor is need to work on how to reduce the training time for this particular dataset as after doing PCA method also it takes a very long time for training .