

MICRO CREDT PROJECT



Submitted by:

MANOJ.I.V

**ACKNOWLEDGMENT**

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

**INTRODUCTION**

* Business Problem Framing

Describe the business problem and how this problem can be related to the real world.

Answer: 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. In, real world also, there are many financial corporations and institutions that wants to help the local businesses to grow.

* Conceptual Background of the Domain Problem

Describe the domain related concepts that you think will be useful for better understanding of the project.

Answer: Bank account details, transaction details, total amount of profit for a business, previous loans, fine paid for missing the instalments, total instalments missing, time period and interest of the loan, loan amount, field/domain of business etc.

* Review of Literature

This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.

* Motivation for the Problem Undertaken

Describe your objective behind to make this project, this domain and what is the motivation behind.

Answer: The objective behind the project is to earn money helping the unbanked poor families living in remote areas with not much sources of income. It is type of win-win situation as it provides an opportunity to small vendors to prosper and at the same time the credit company also earns money. So a model have to be build such that it can predict whether the vendor or the party going for the credit can give back the money or not.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Describe the mathematical, statistical and analytics modelling done during this project along with the proper justification.

* Data Sources and their formats

What are the data sources, their origins, their formats and other details that you find necessary? They can be described here. Provide a proper data description. You can also add a snapshot of the data.

Answer: The data can be taken by a survey conducted by micro credit company, open source websites like Kaggel etc. The data is in the form of .csv file it may also be in .json or Excel files. Currently the data was provided in terms of .csv files. There is training file with 1168 columns and 81 rows. There is a test file with 37 columns and 209593 rows. The below shows the column list.

Data columns (total 37 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Unnamed: 0 209593 non-null int64

1 label 209593 non-null int64

2 msisdn 209593 non-null object

3 aon 209593 non-null float64

4 daily\_decr30 209593 non-null float64

5 daily\_decr90 209593 non-null float64

6 rental30 209593 non-null float64

7 rental90 209593 non-null float64

8 last\_rech\_date\_ma 209593 non-null float64

9 last\_rech\_date\_da 209593 non-null float64

10 last\_rech\_amt\_ma 209593 non-null int64

11 cnt\_ma\_rech30 209593 non-null int64

12 fr\_ma\_rech30 209593 non-null float64

13 sumamnt\_ma\_rech30 209593 non-null float64

14 medianamnt\_ma\_rech30 209593 non-null float64

15 medianmarechprebal30 209593 non-null float64

16 cnt\_ma\_rech90 209593 non-null int64

17 fr\_ma\_rech90 209593 non-null int64

18 sumamnt\_ma\_rech90 209593 non-null int64

19 medianamnt\_ma\_rech90 209593 non-null float64

20 medianmarechprebal90 209593 non-null float64

21 cnt\_da\_rech30 209593 non-null float64

22 fr\_da\_rech30 209593 non-null float64

23 cnt\_da\_rech90 209593 non-null int64

24 fr\_da\_rech90 209593 non-null int64

25 cnt\_loans30 209593 non-null int64

26 amnt\_loans30 209593 non-null int64

27 maxamnt\_loans30 209593 non-null float64

28 medianamnt\_loans30 209593 non-null float64

29 cnt\_loans90 209593 non-null float64

30 amnt\_loans90 209593 non-null int64

31 maxamnt\_loans90 209593 non-null int64

32 medianamnt\_loans90 209593 non-null float64

33 payback30 209593 non-null float64

34 payback90 209593 non-null float64

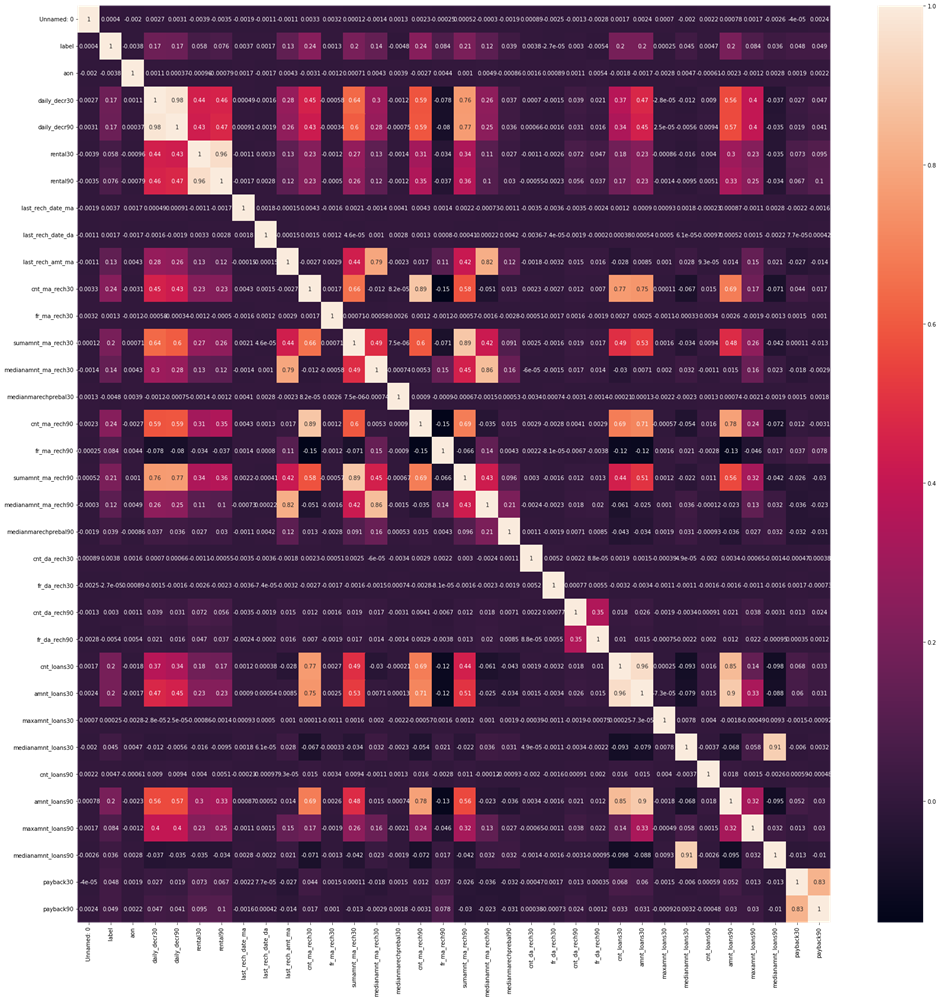
35 pcircle 209593 non-null object

36 pdate 209593 non-null object

dtypes: float64(21), int64(13), object(3)

memory usage: 59.2+ MB

The correlation was performed. It was as shown in the fig.



Form the correlation label was seen:

fr\_da\_rech90 -0.005418

medianmarechprebal30 -0.004829

aon -0.003785

fr\_da\_rech30 -0.000027

maxamnt\_loans30 0.000248

Unnamed: 0 0.000403

fr\_ma\_rech30 0.001330

last\_rech\_date\_da 0.001711

cnt\_da\_rech90 0.002999

last\_rech\_date\_ma 0.003728

cnt\_da\_rech30 0.003827

cnt\_loans90 0.004733

medianamnt\_loans90 0.035747

medianmarechprebal90 0.039300

medianamnt\_loans30 0.044589

payback30 0.048336

payback90 0.049183

rental30 0.058085

rental90 0.075521

maxamnt\_loans90 0.084144

fr\_ma\_rech90 0.084385

medianamnt\_ma\_rech90 0.120855

last\_rech\_amt\_ma 0.131804

medianamnt\_ma\_rech30 0.141490

daily\_decr90 0.166150

daily\_decr30 0.168298

cnt\_loans30 0.196283

amnt\_loans30 0.197272

amnt\_loans90 0.199788

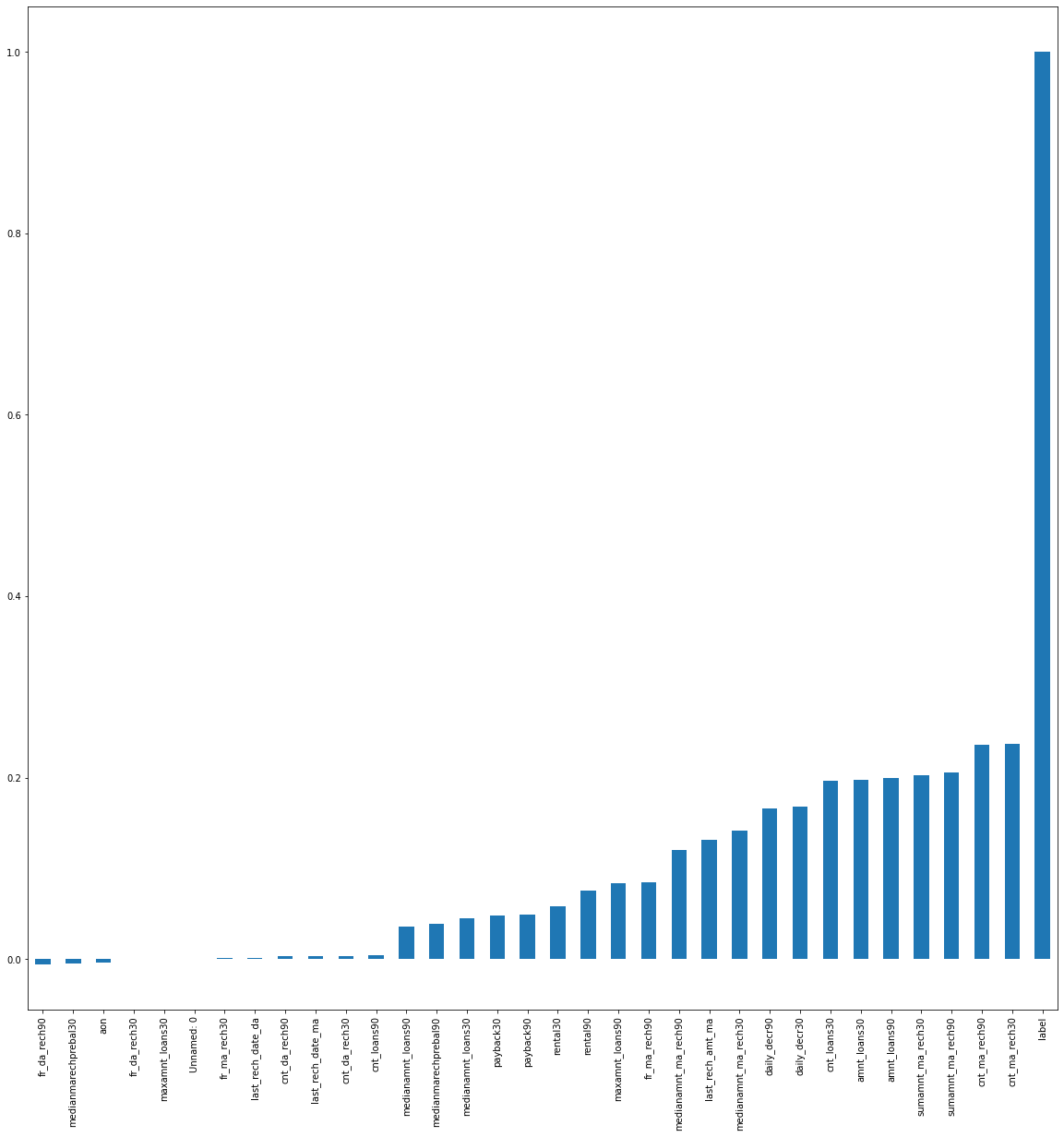
sumamnt\_ma\_rech30 0.202828

sumamnt\_ma\_rech90 0.205793

cnt\_ma\_rech90 0.236392

cnt\_ma\_rech30 0.237331

label 1.000000

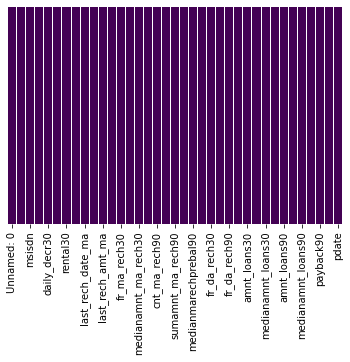


* Data Preprocessing Done

What were the steps followed for the cleaning of the data? What were the assumptions done and what were the next actions steps over that?

Answer: As there are no null values there was no imputation done. The assumptions were unnamed 0, pcircle, msisdn, and year was neglected as the data was correctly to the specific year.

dt.drop(columns=['Unnamed:0','msisdn','pcircle',’year’],inplace=True)



* Data Inputs- Logic- Output Relationships

Describe the relationship behind the data input, its format, the logic in between and the output. Describe how the input affects the output.

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

Here, you can describe any presumptions taken by you.

* Hardware and Software Requirements and Tools Used

Listing down the hardware and software requirements along with the tools, libraries and packages used. Describe all the software tools used along with a detailed description of tasks done with those tools.

The different libraries and packages used are:

1. Pandas, 2. Numpy, 3. Matplotlib, 4. Sklearn and 5.Dtale etc.

Pandas: for importing the dataset

Matplotlib and Dtale: For graphing

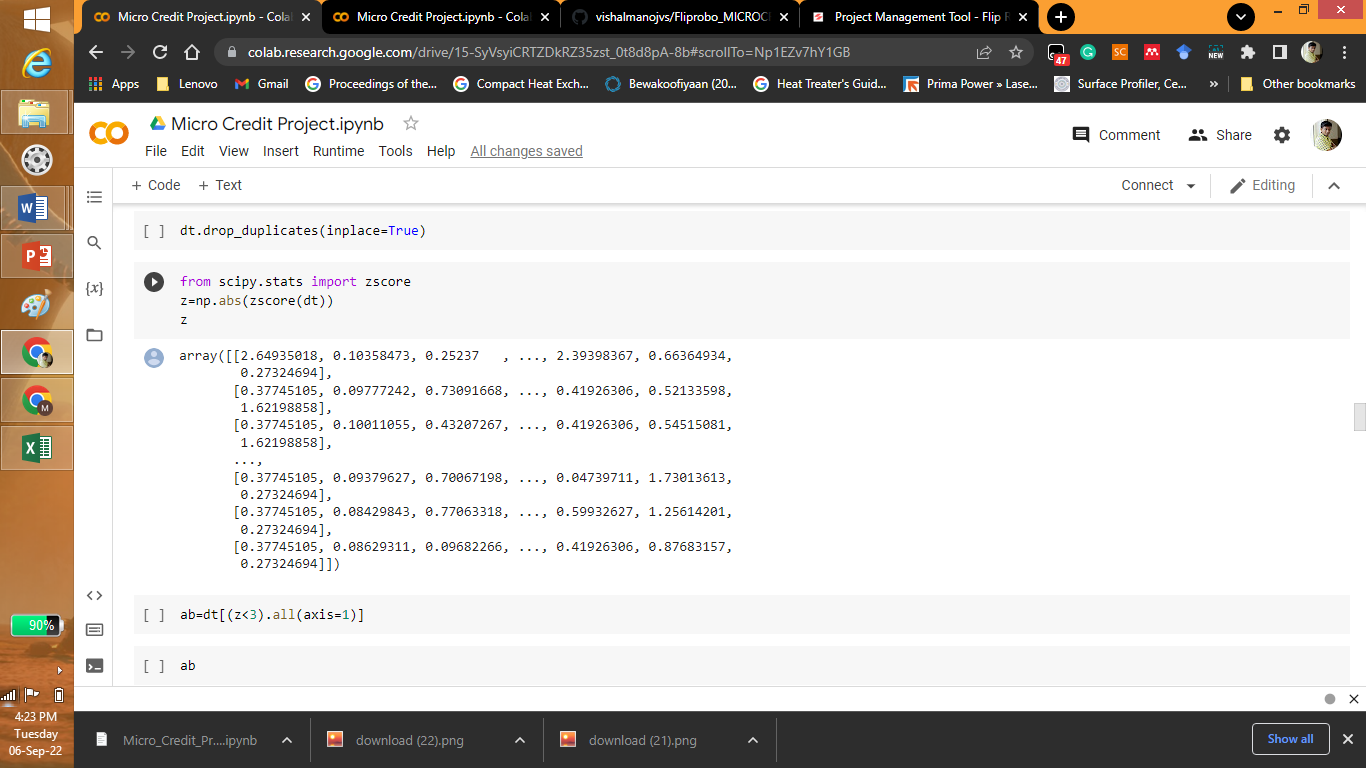
Sklearn: Modelling

**Model/s Development and Evaluation**

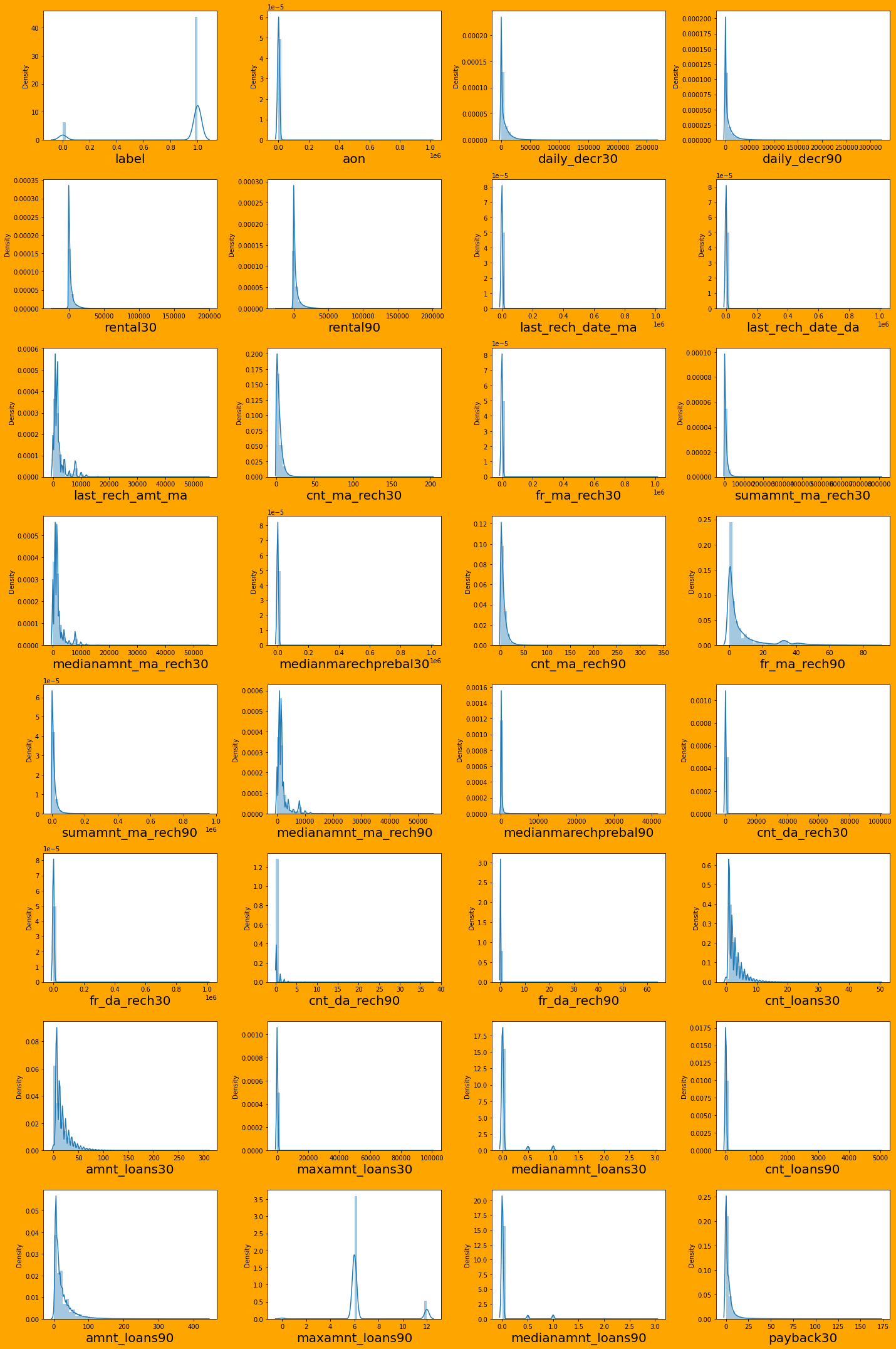
* Identification of possible problem-solving approaches (methods)

Describe the approaches you followed, both statistical and analytical, for solving of this problem.

Answer: Both box plot and kde plots were ploted and zscore was applied.



****

****

* Testing of Identified Approaches (Algorithms)

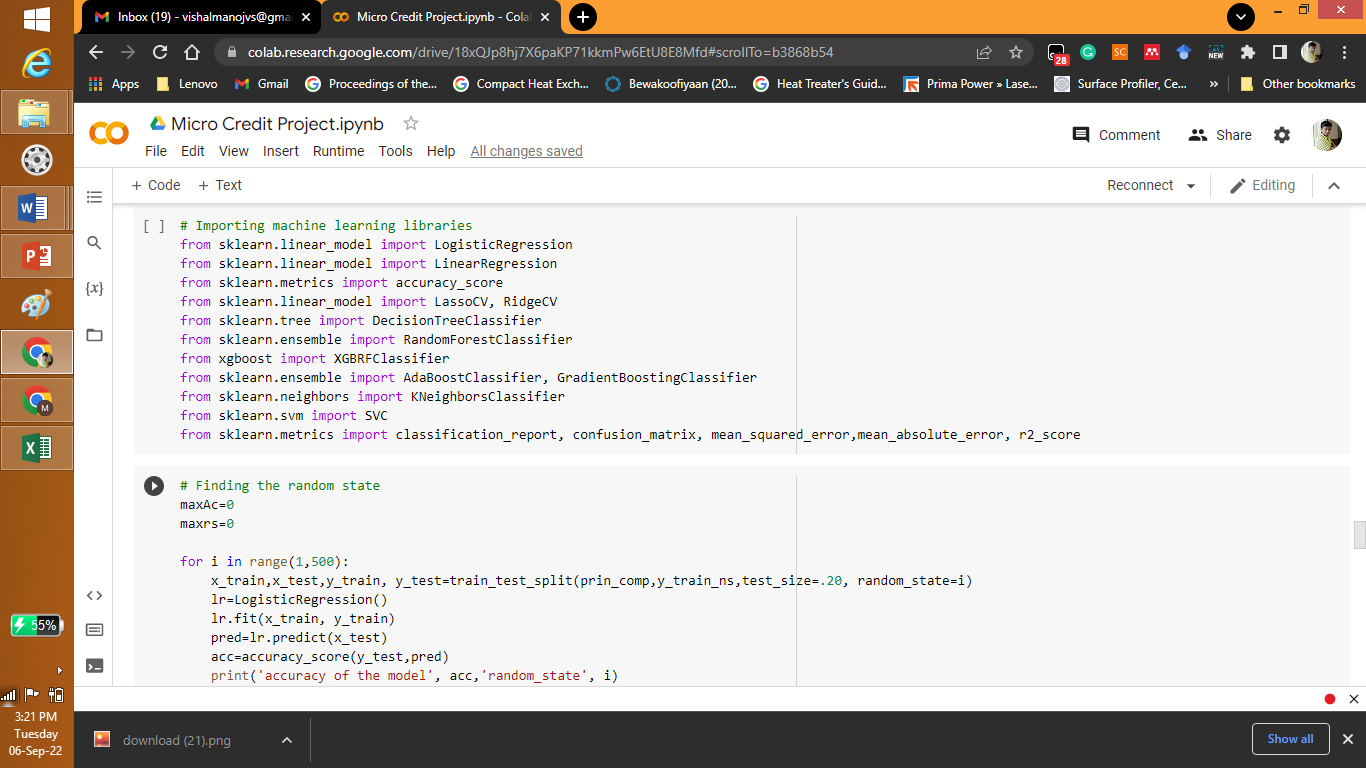
Listing down all the algorithms used for the training and testing.

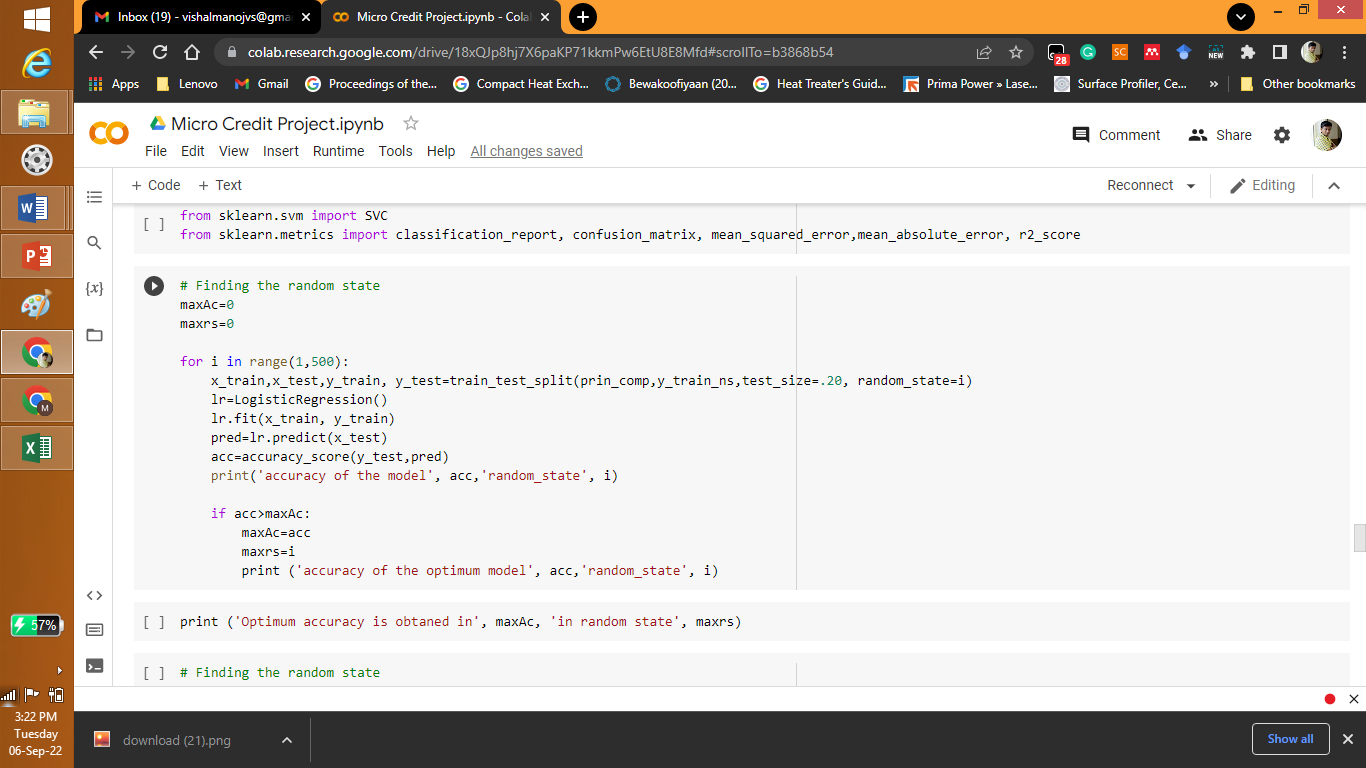
Answer: The algorithms used were Logistic regression, KNN, Random forest, decision tree, EGboost, AdaBoost, Gradient Boost, servo vector classification (SVC). The SVC model takes more time for compution so it was neglected in ipython note book.

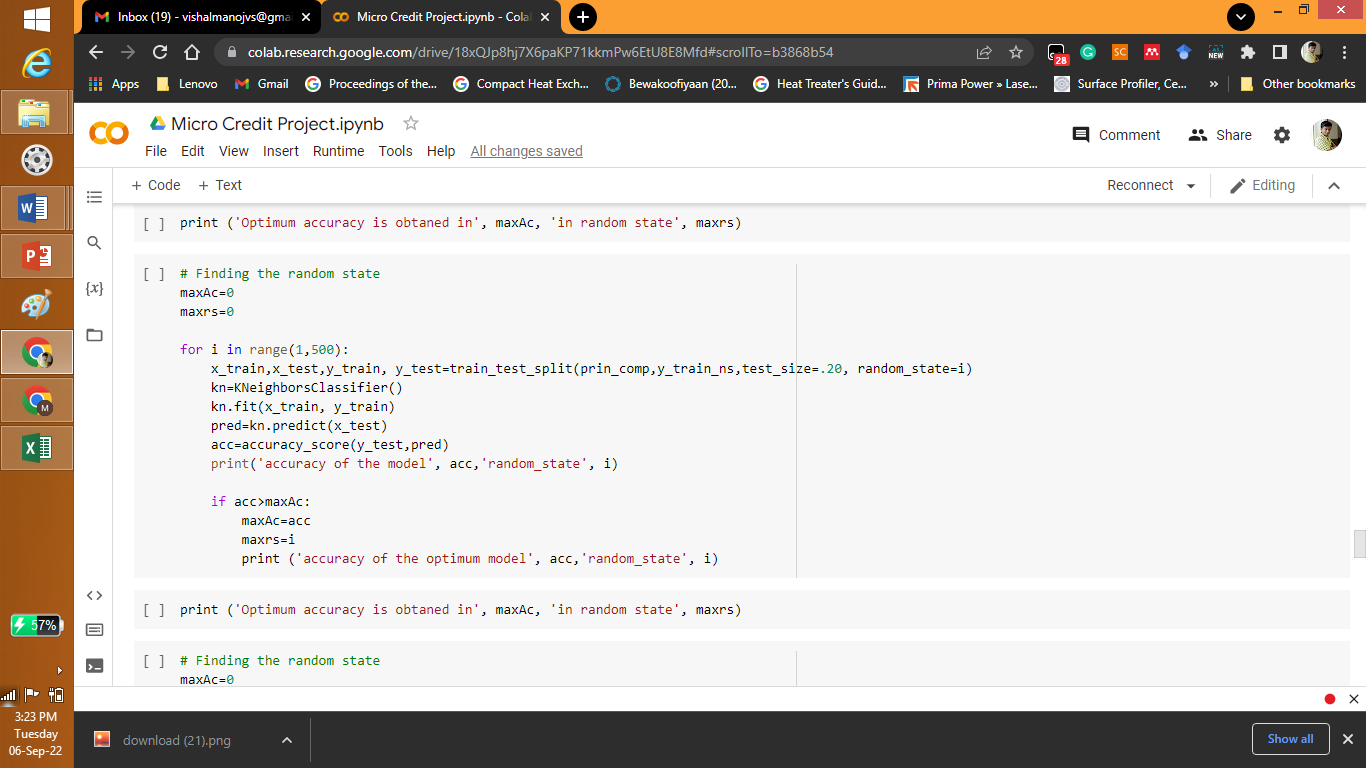
* Run and Evaluate selected models

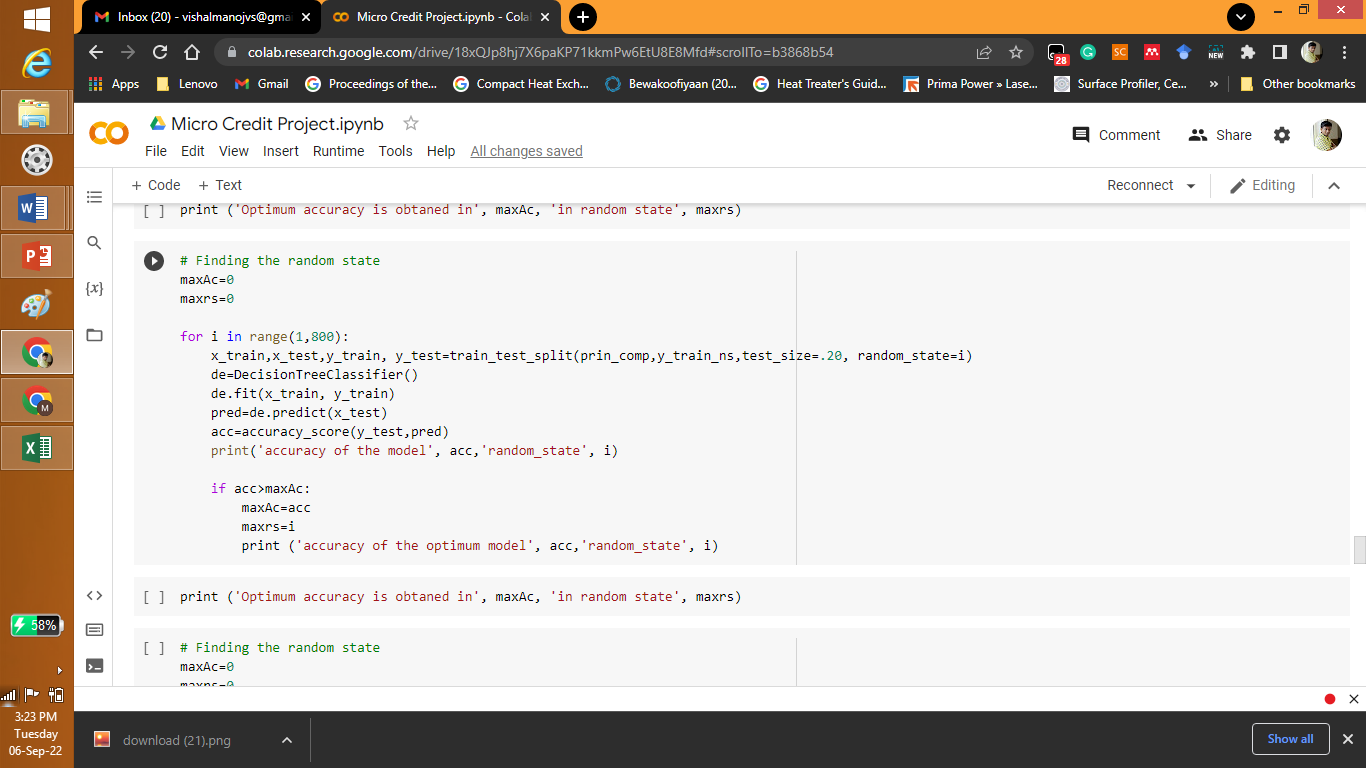
Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.

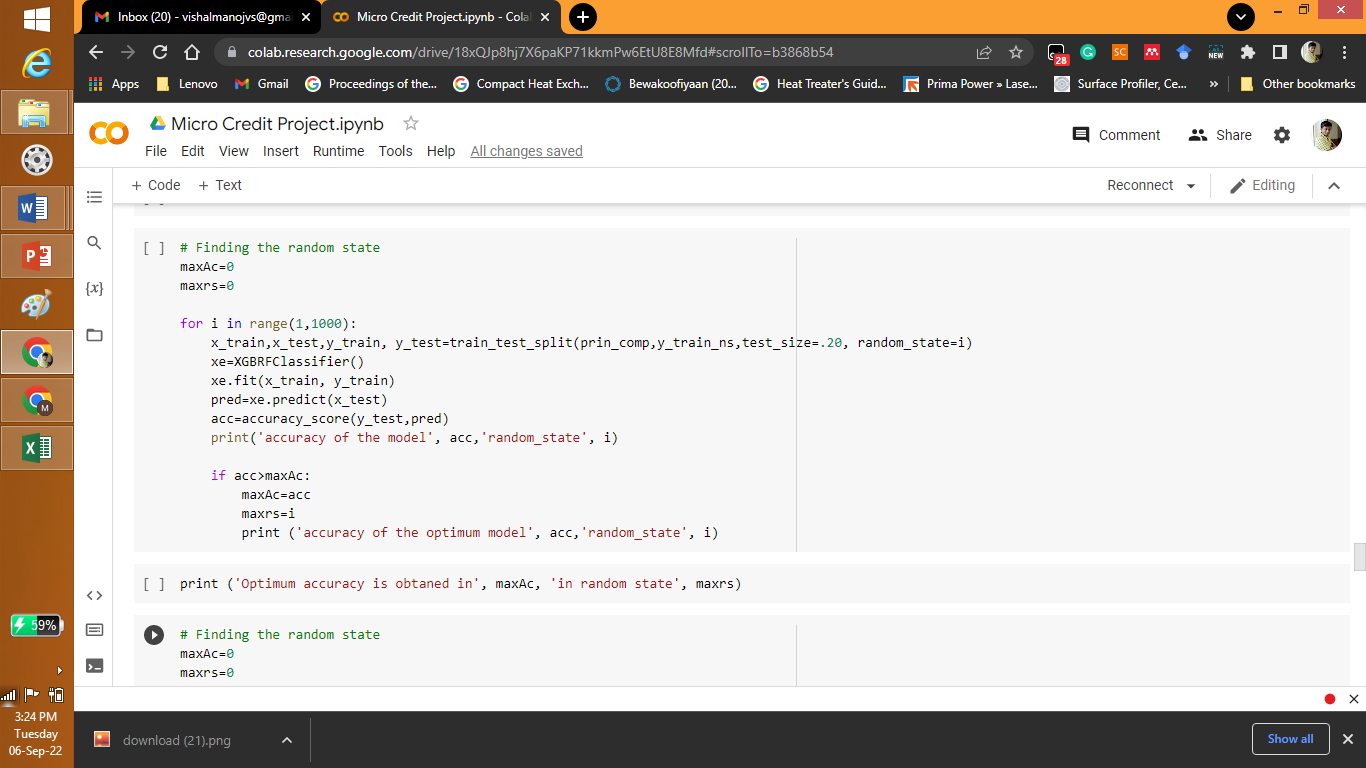
Answer: First the models were imported and random values were found.

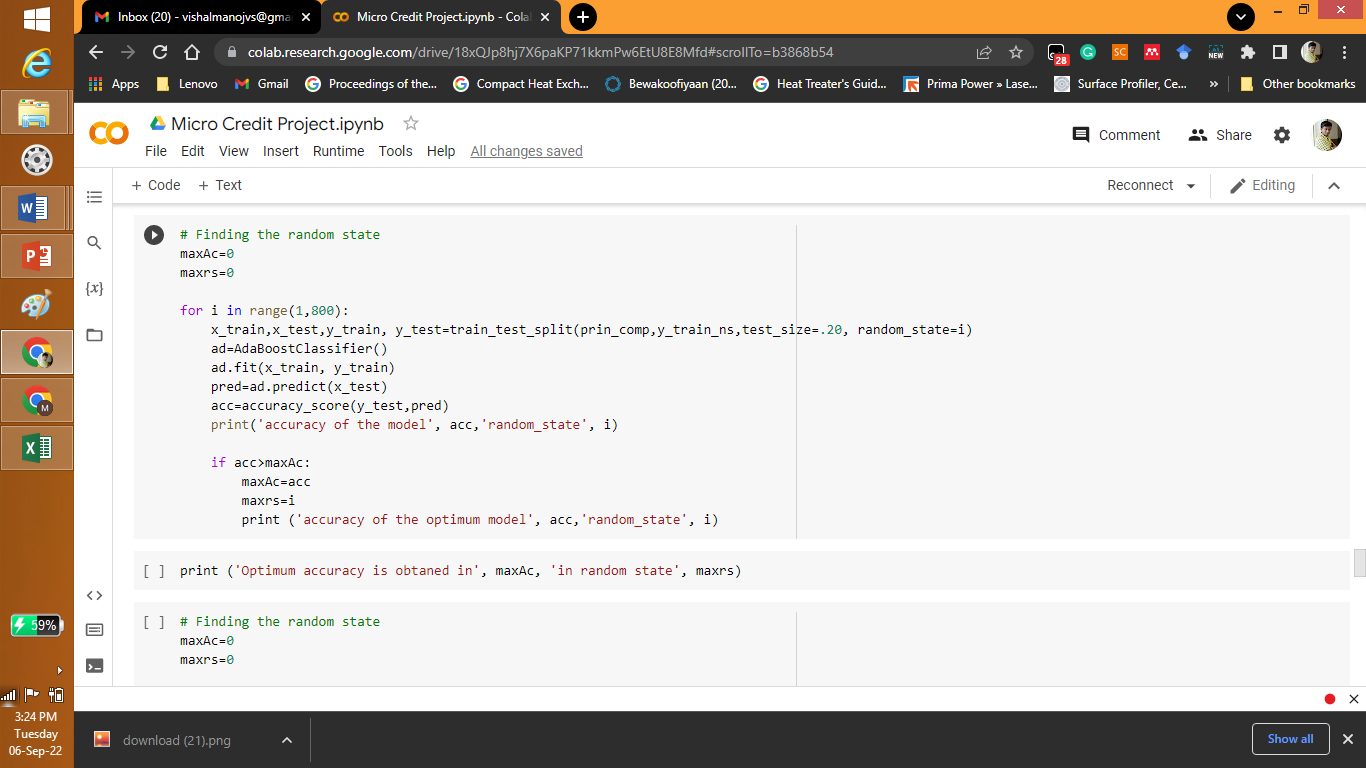


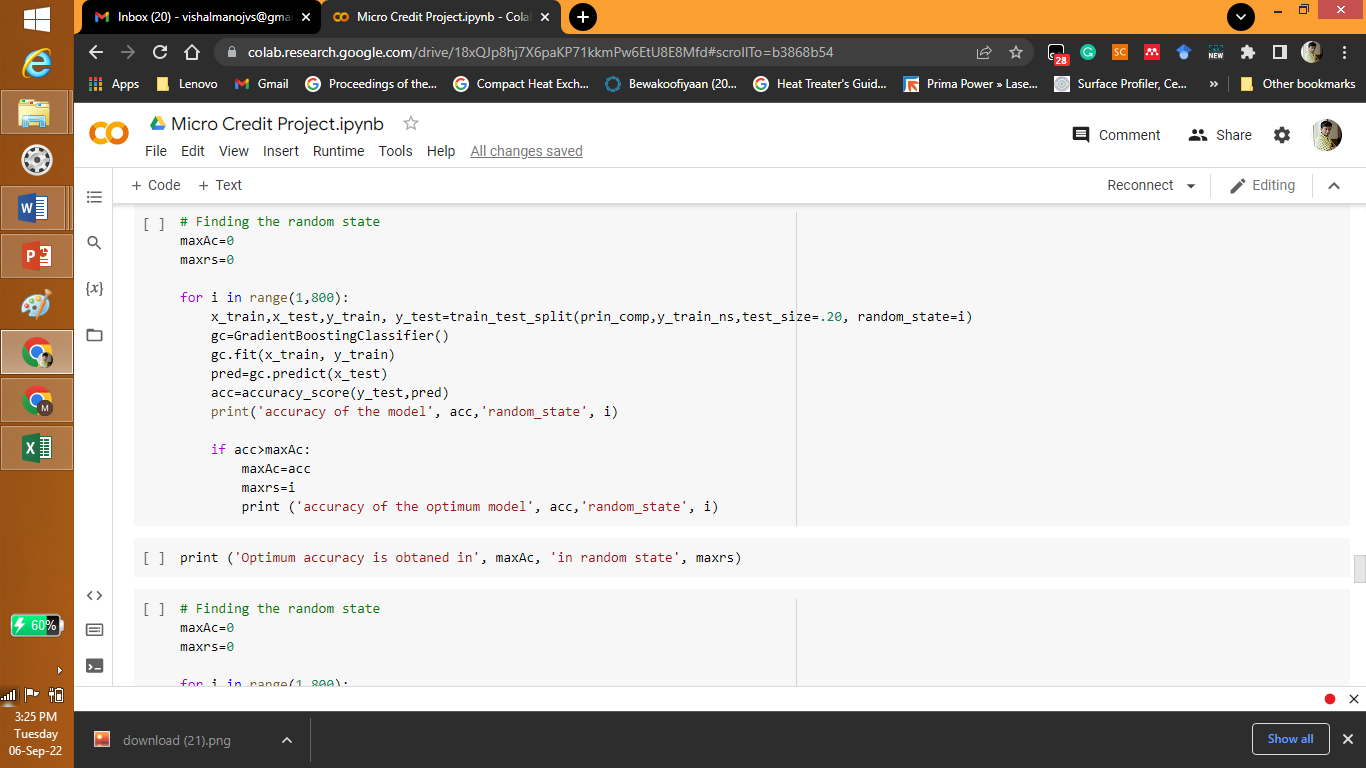


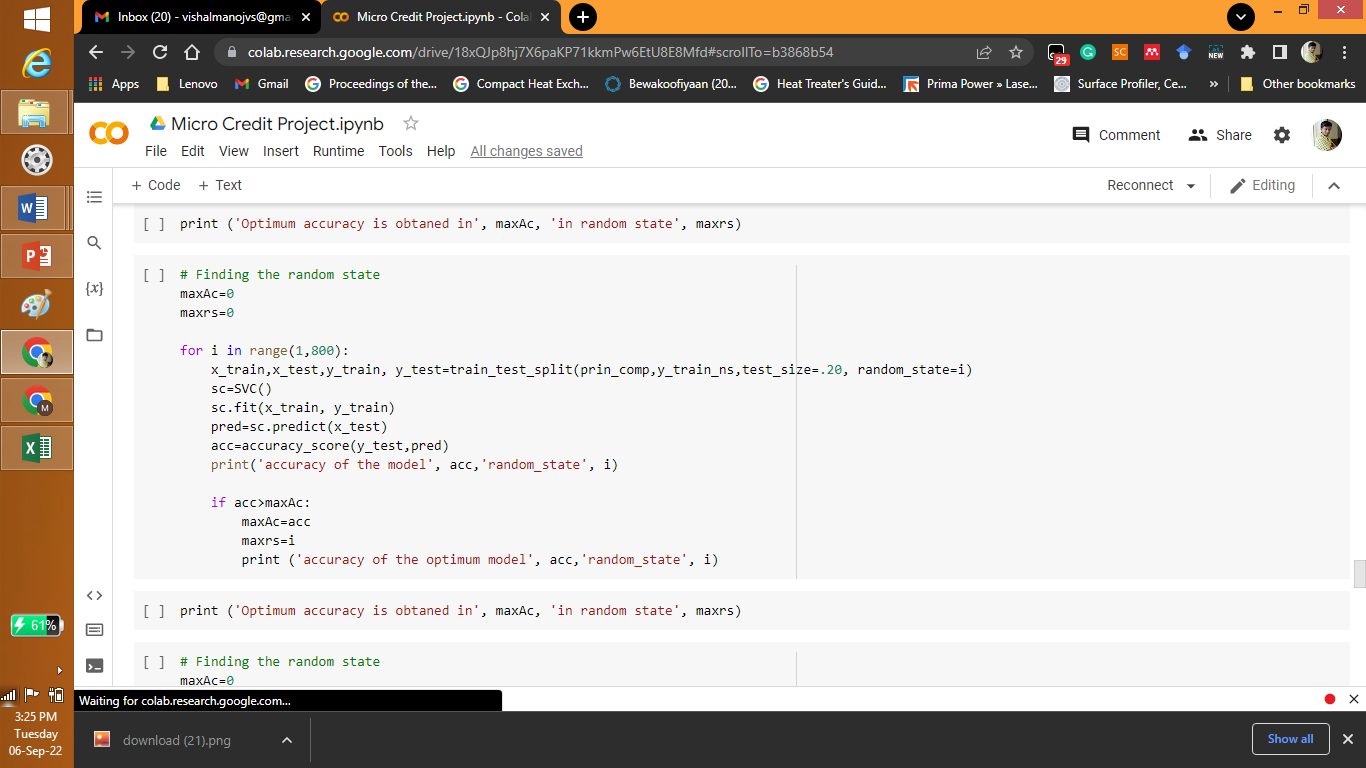


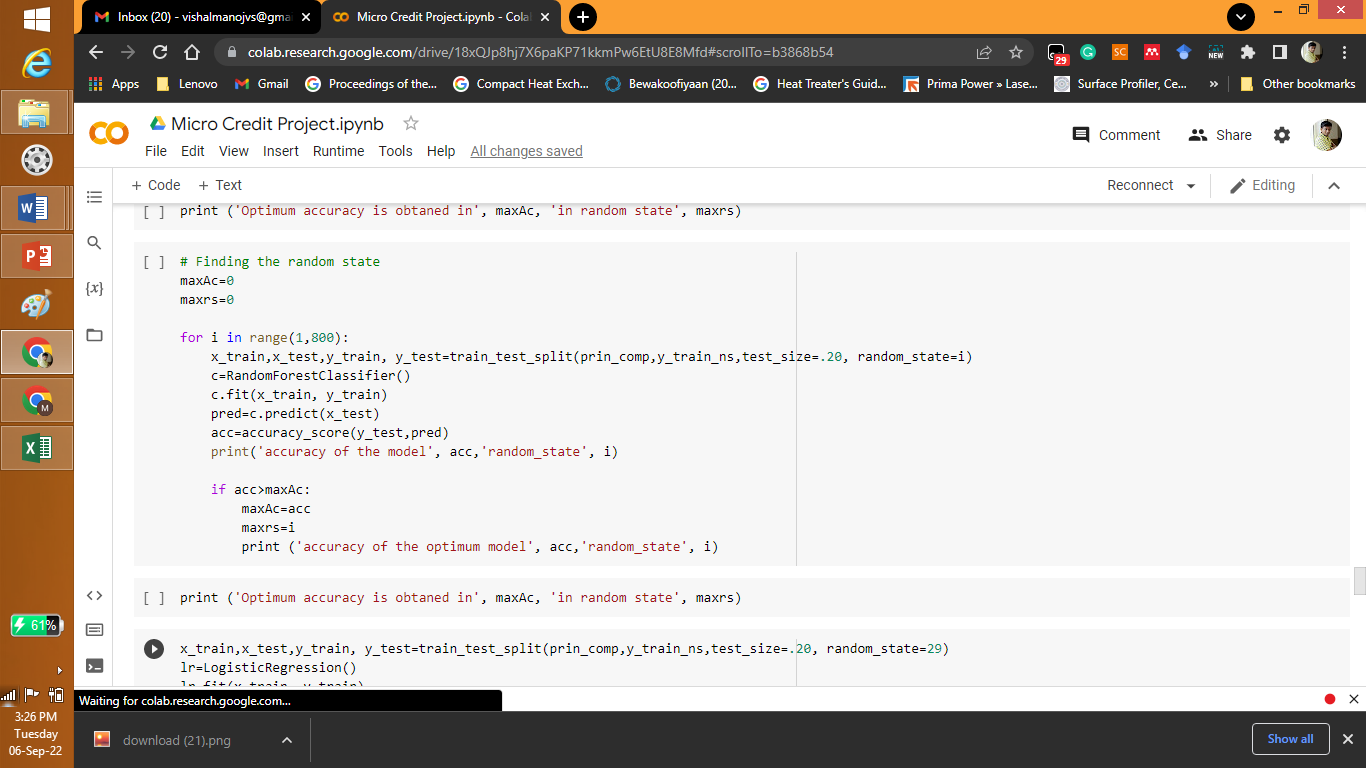


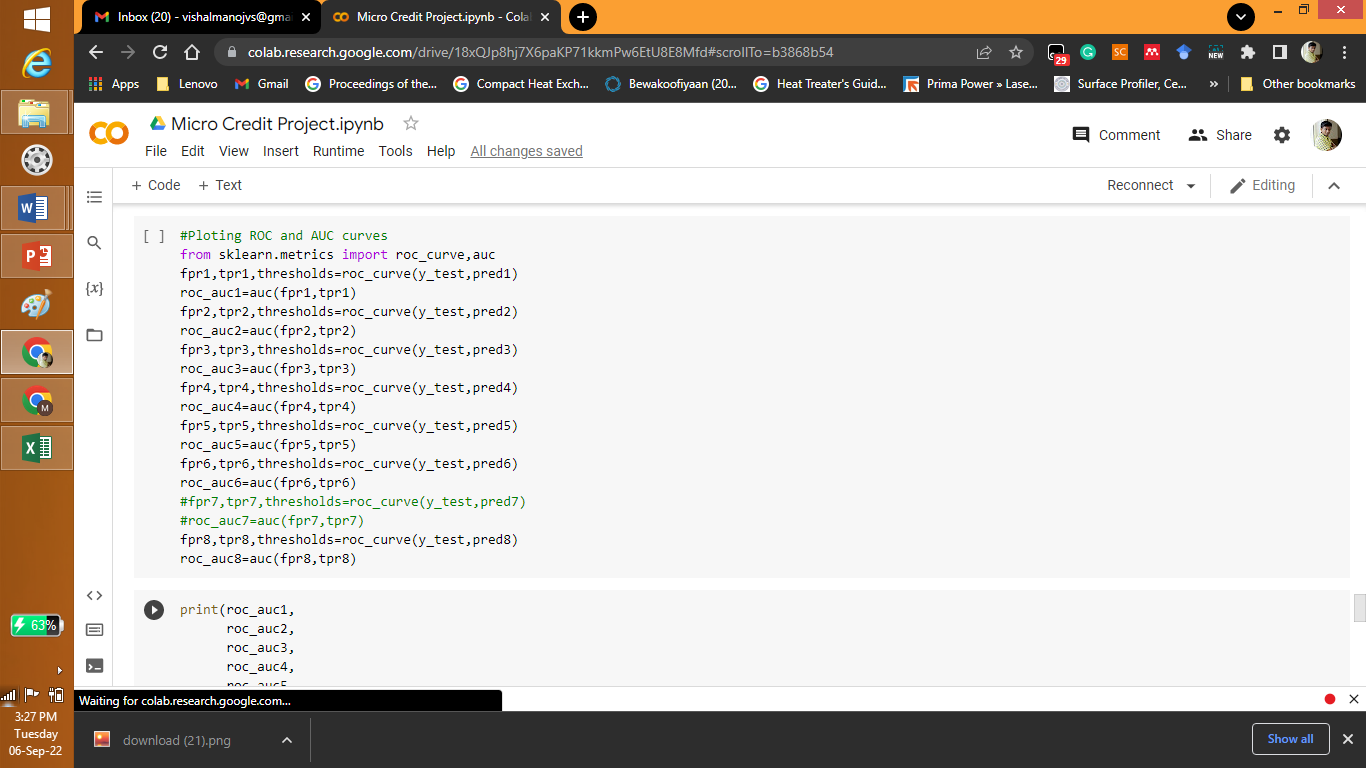


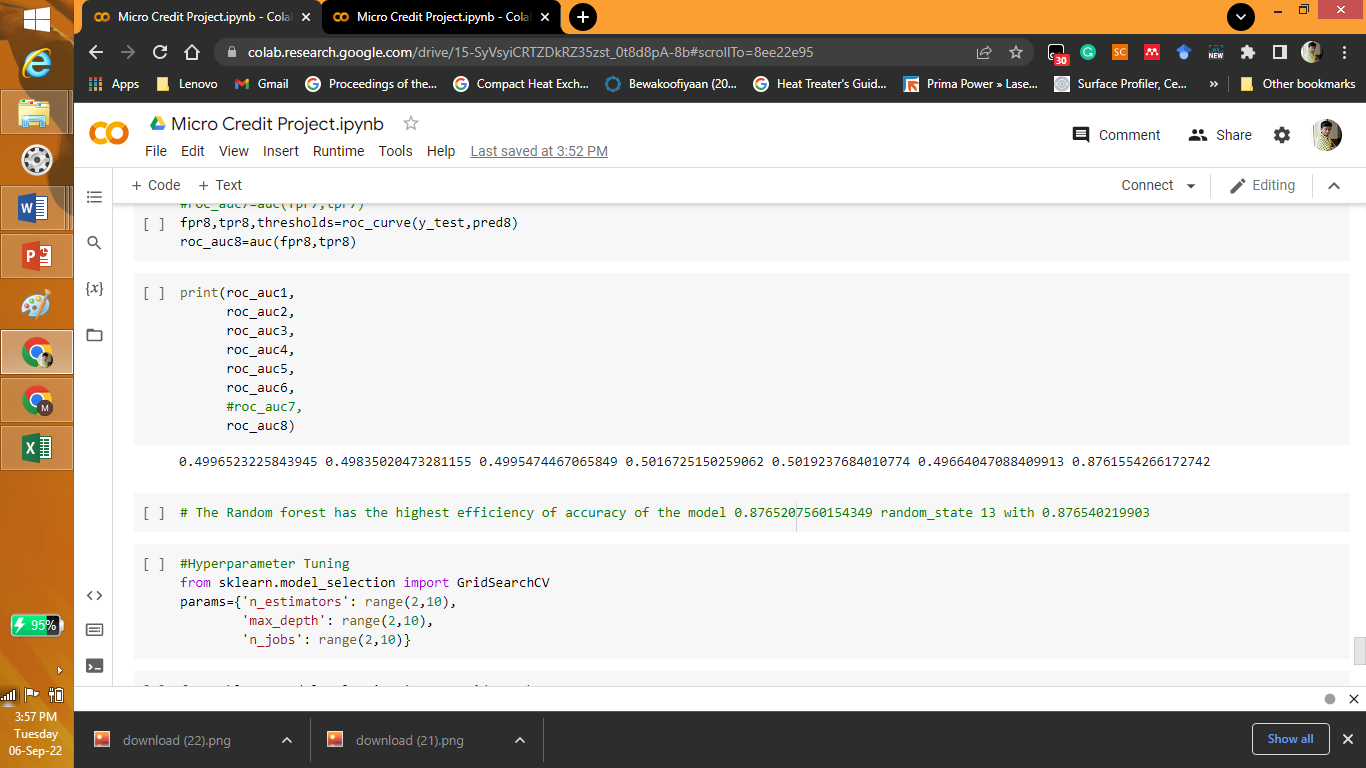




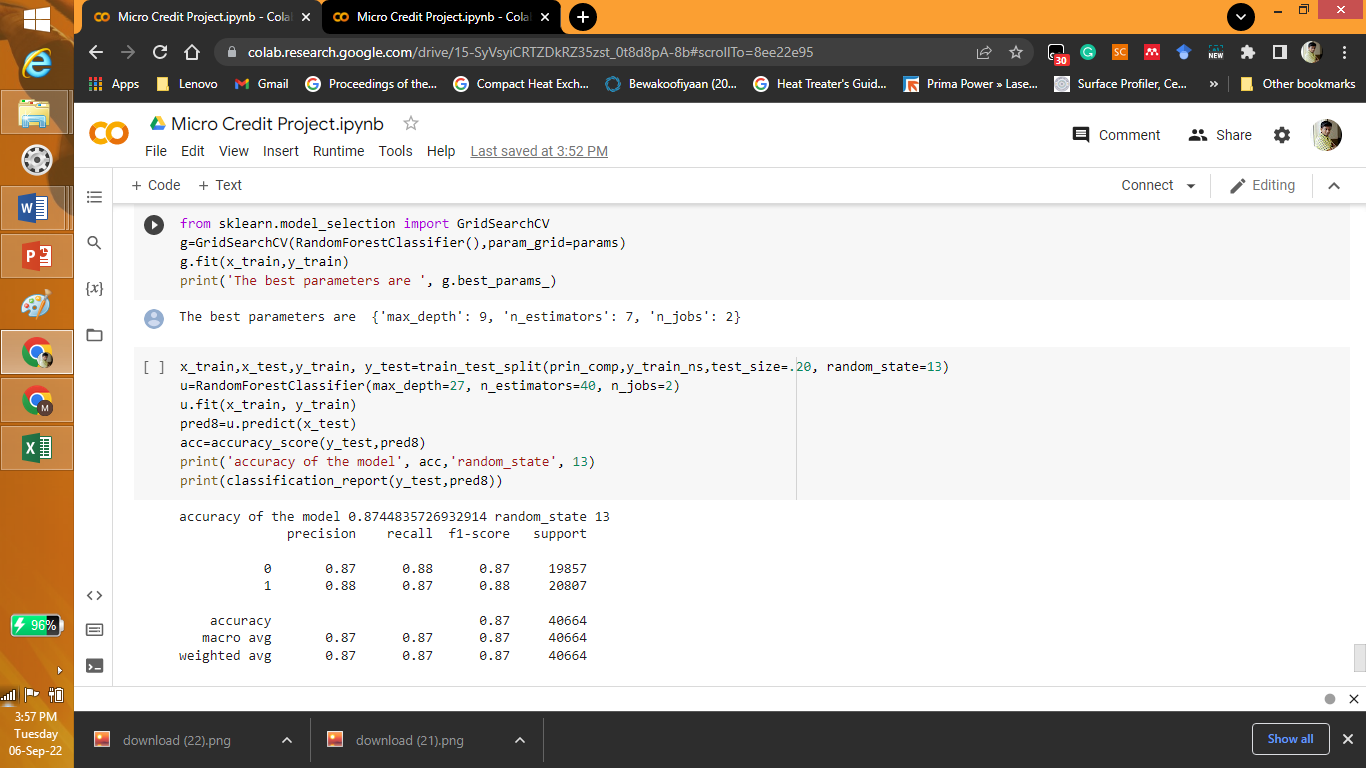








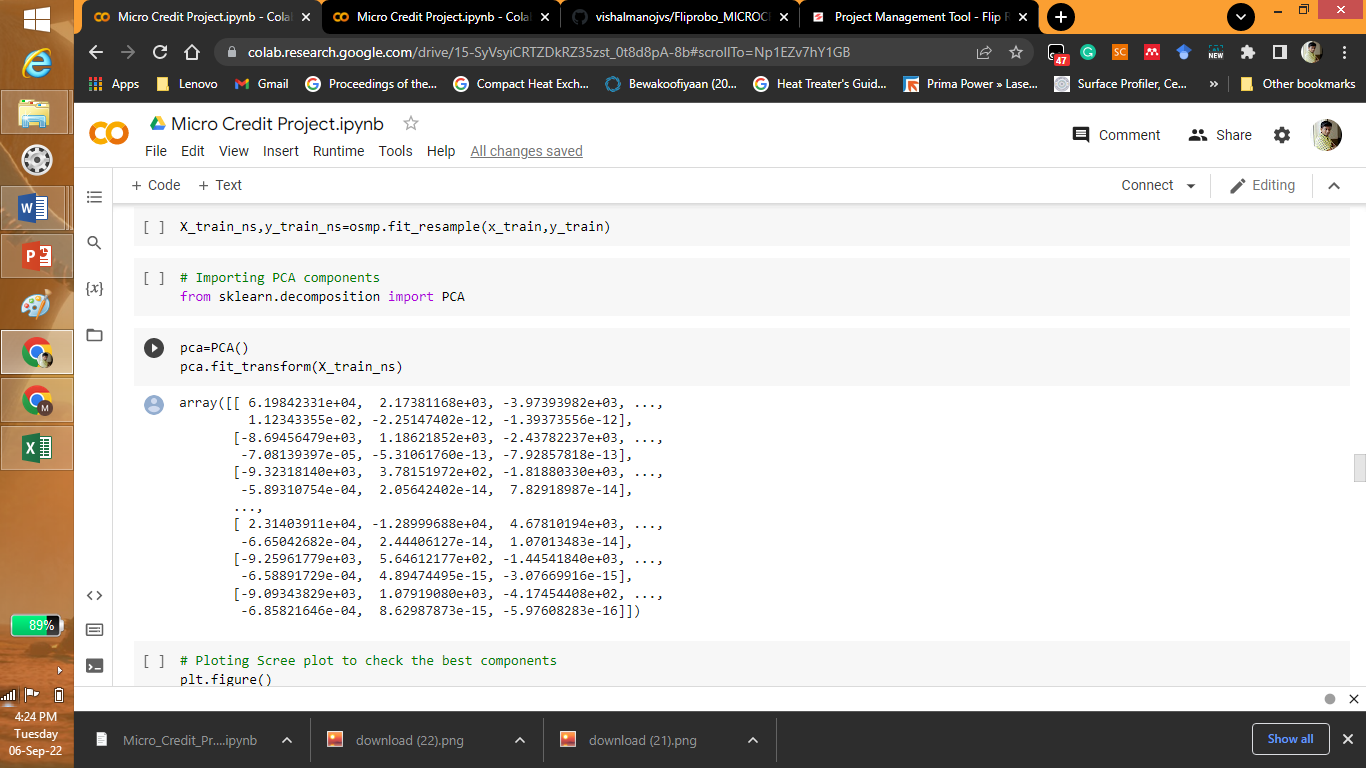
The Random forest has the highest efficiency of ROC\_AUC of the model 0.8765207560154349 random\_state 13 with 0.876540219903.

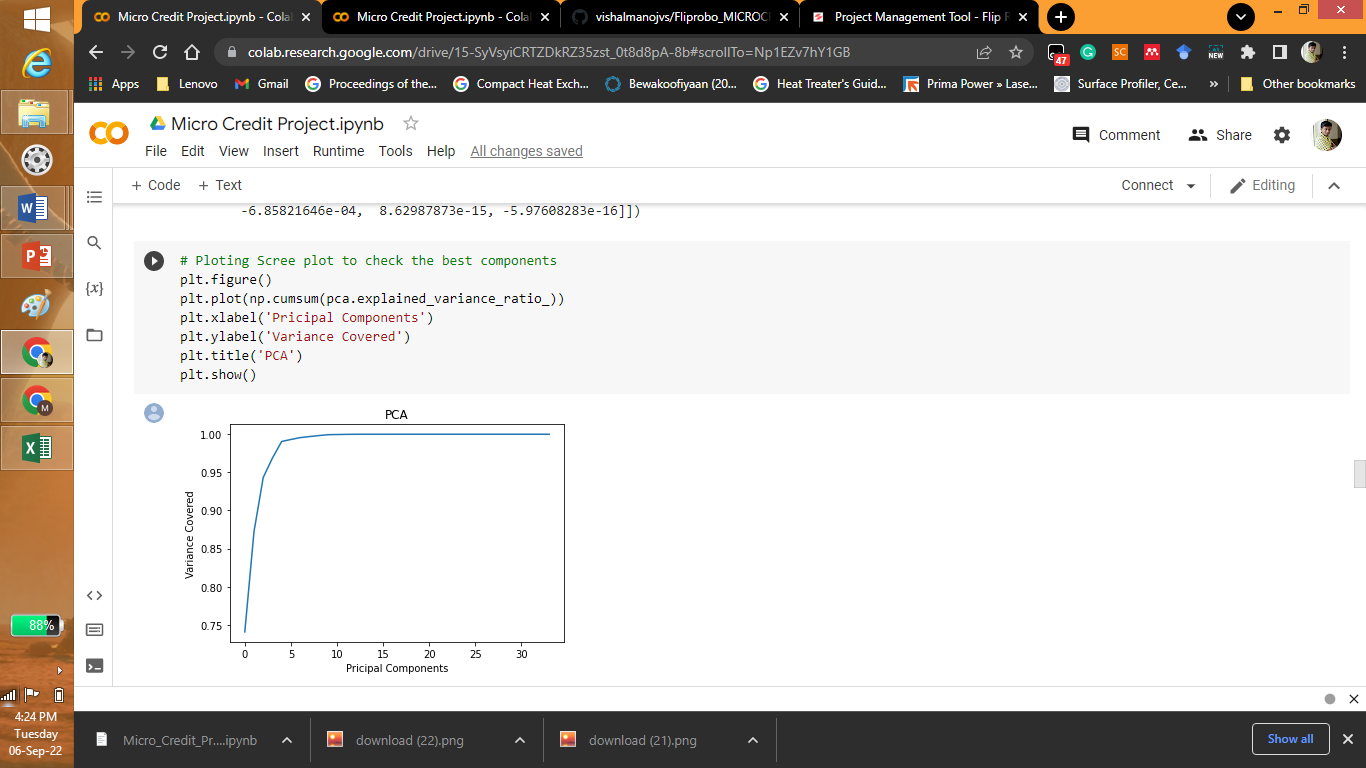


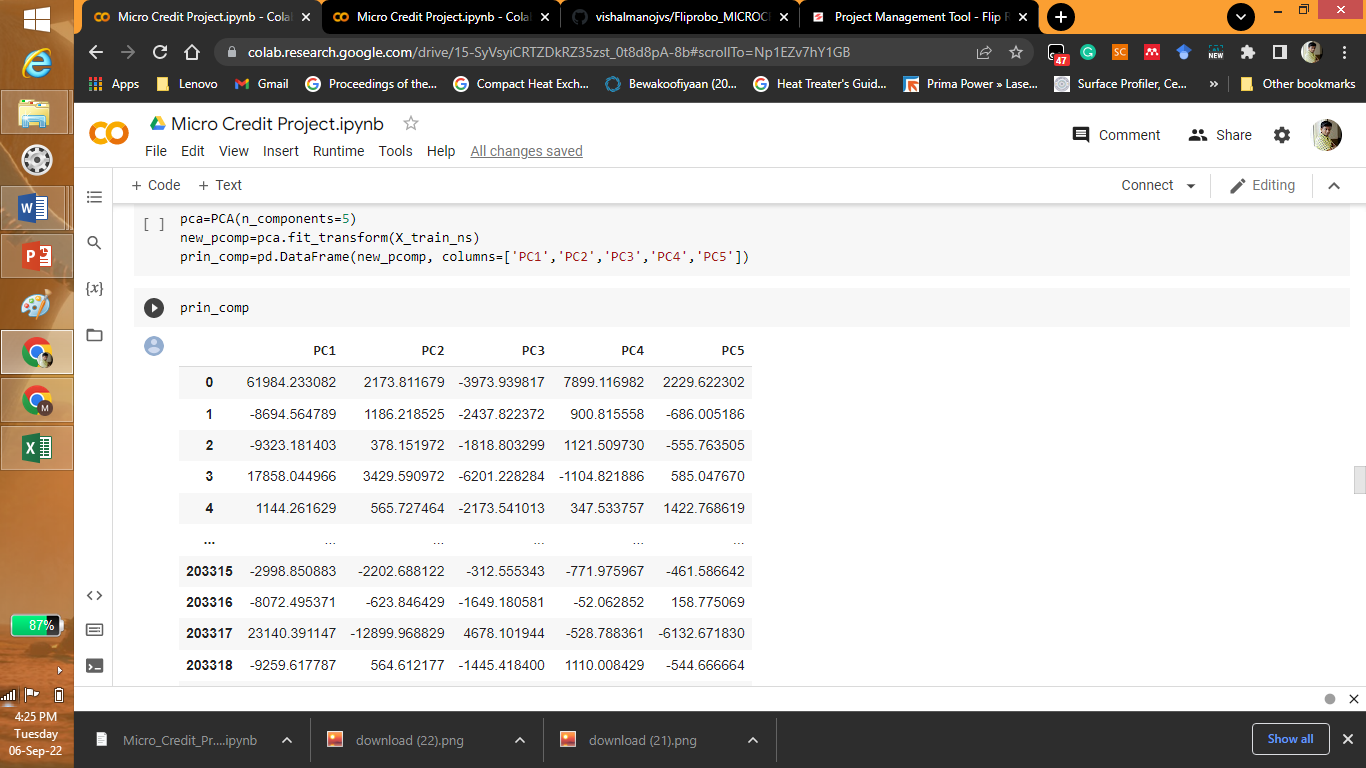
* Key Metrics for success in solving problem under consideration

What were the key metrics used along with justification for using it? You may also include statistical metrics used if any.

Answer: Principle component analysis was used for dimensionality reduction.





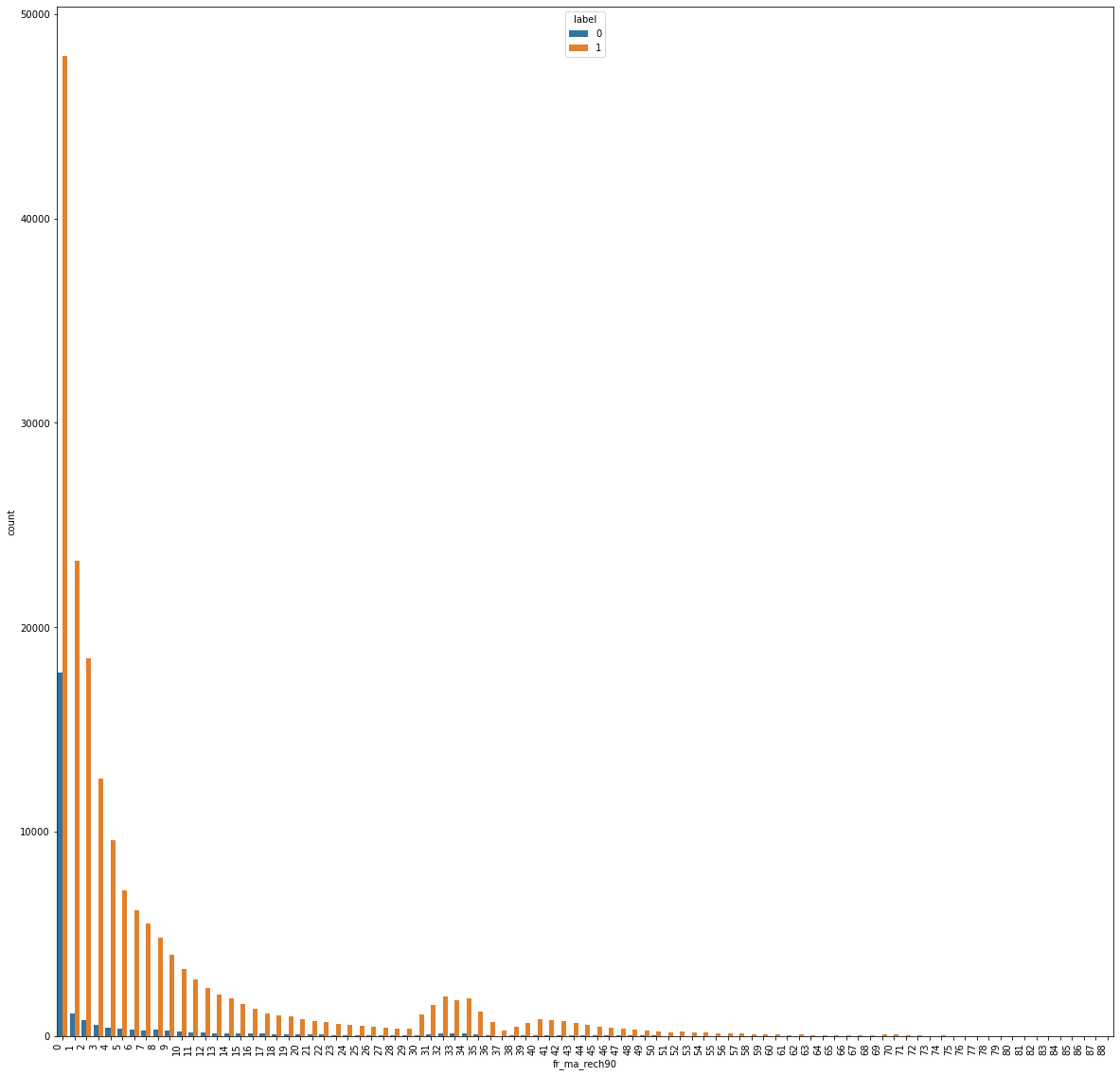


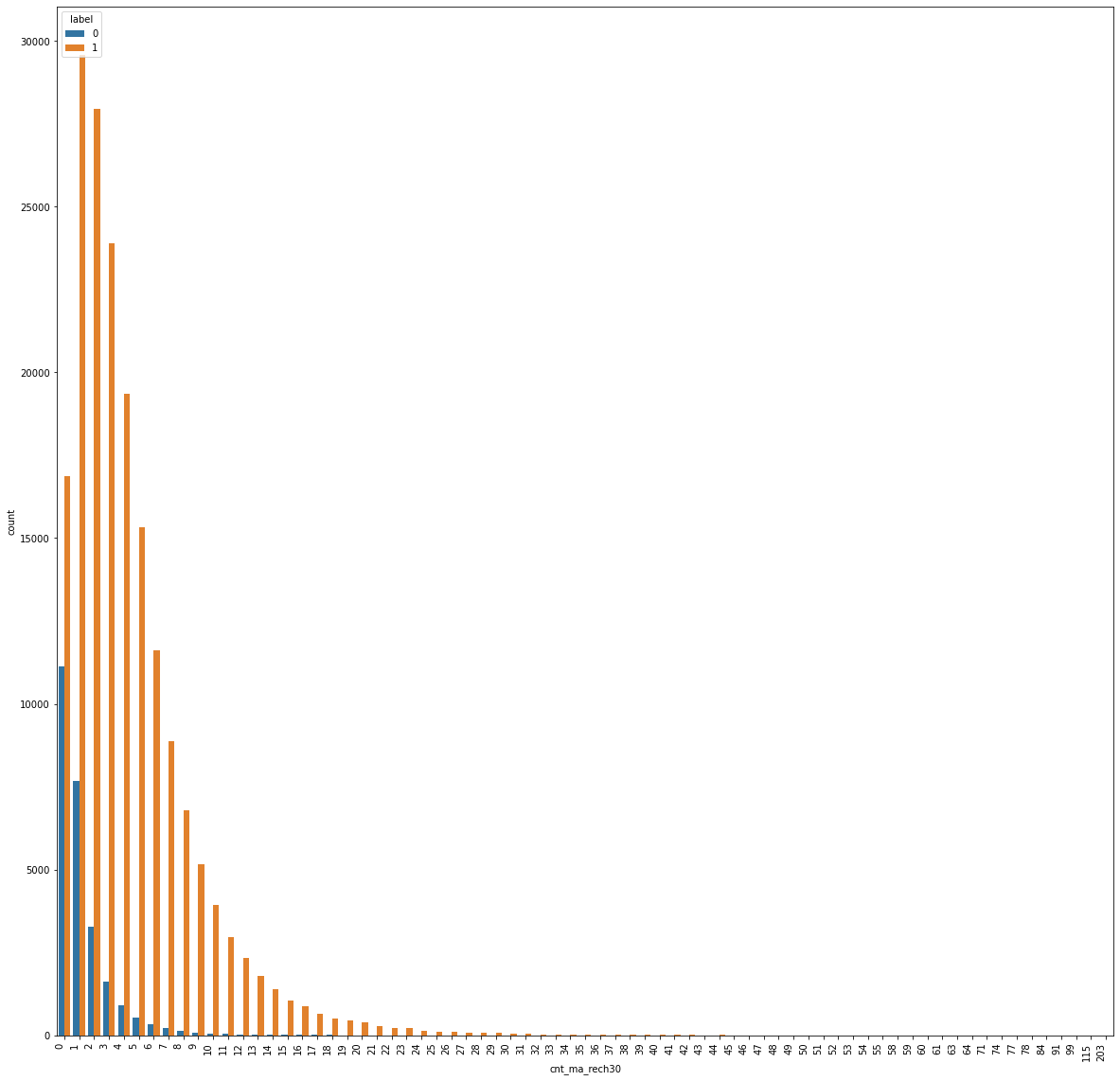
* Visualizations

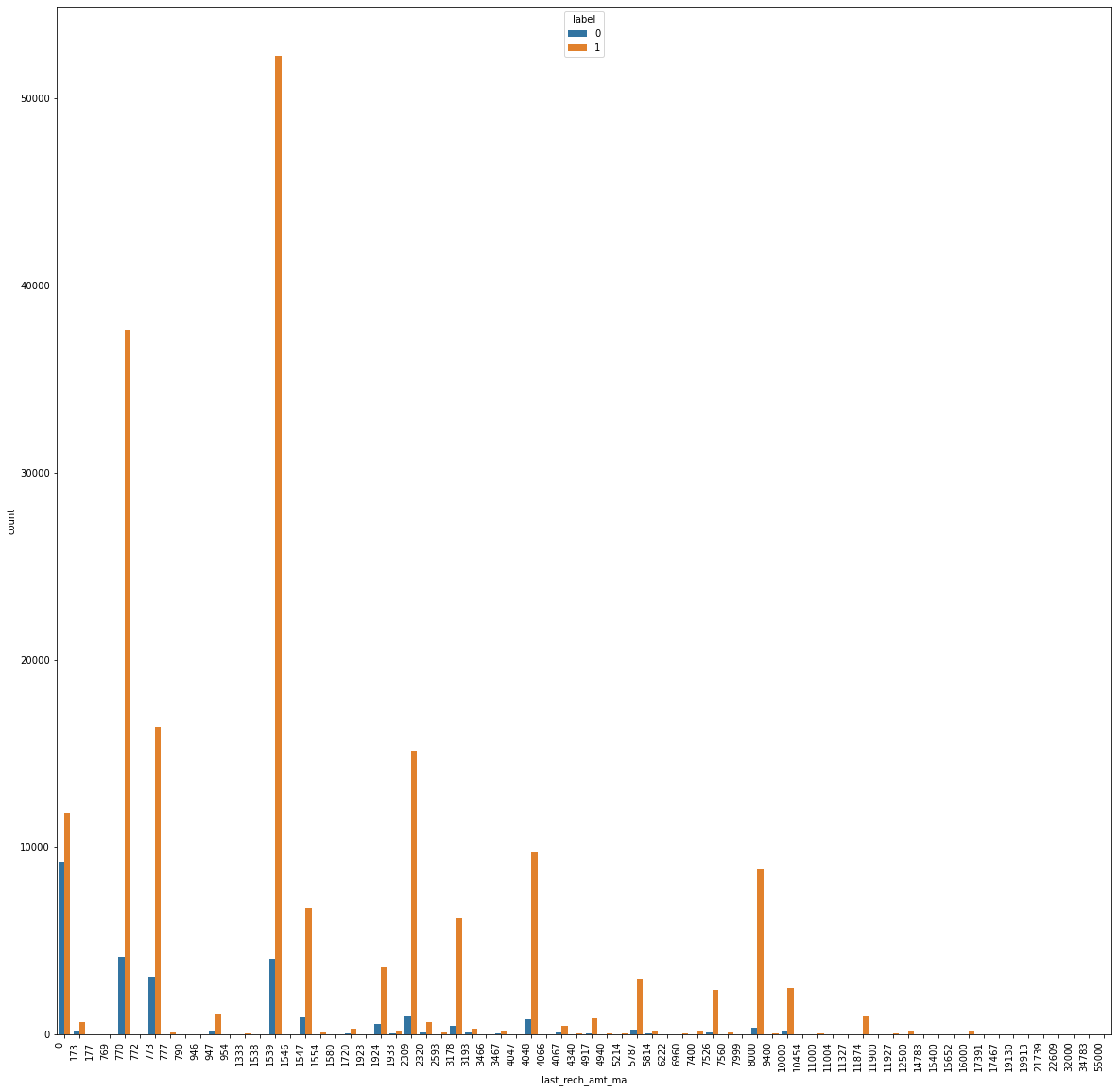
Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.

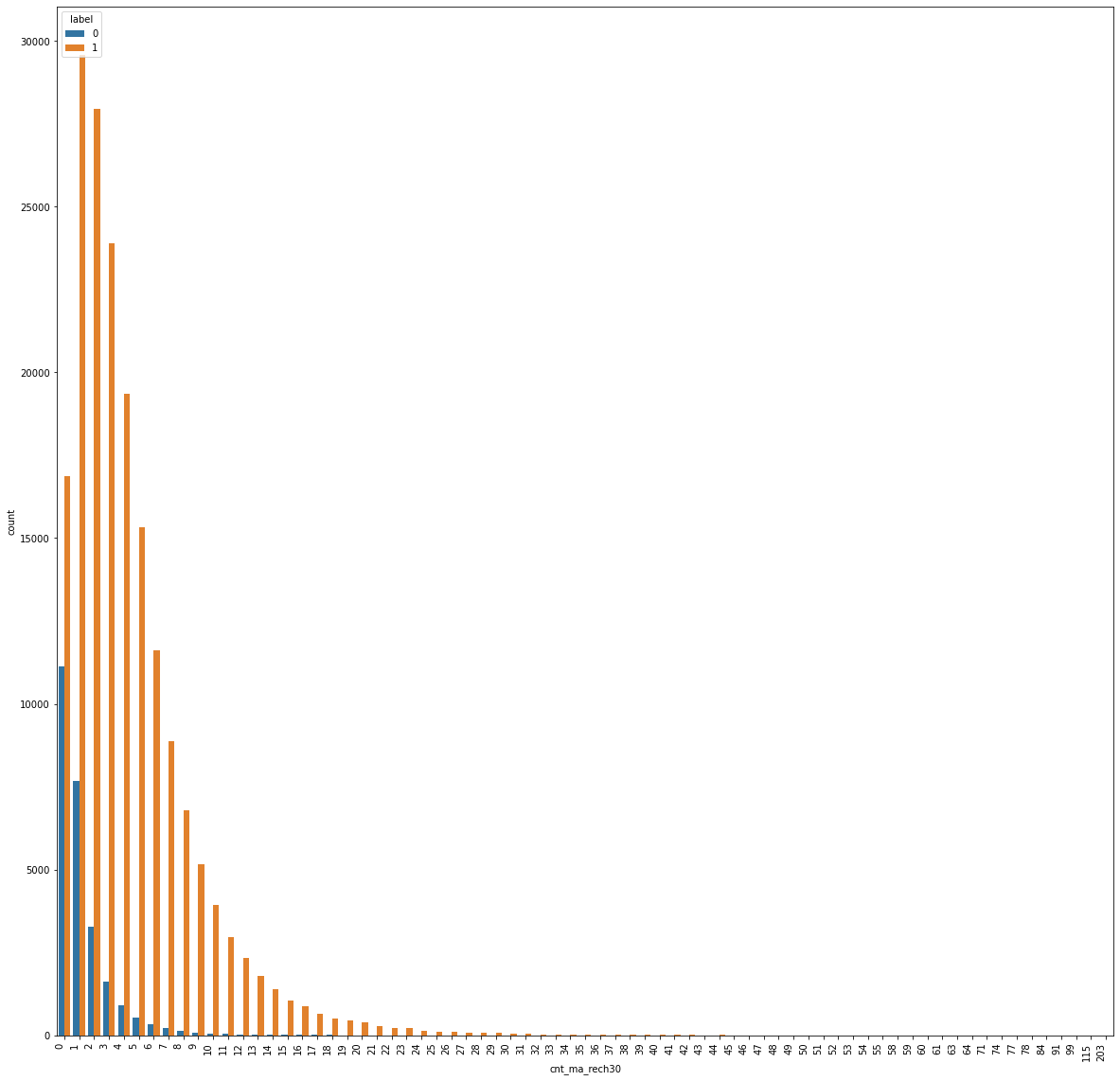
If different platforms were used, mention that as well.

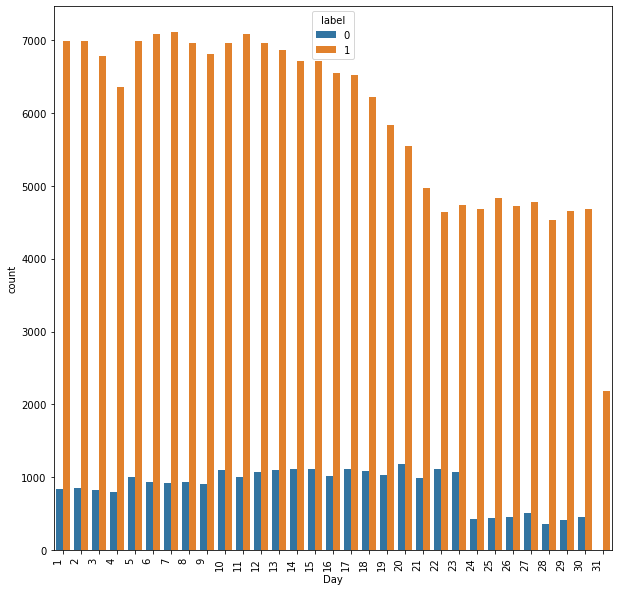
Answer: Only python was used for visualization:

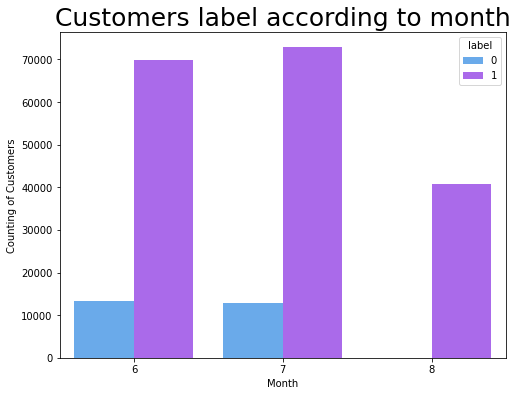


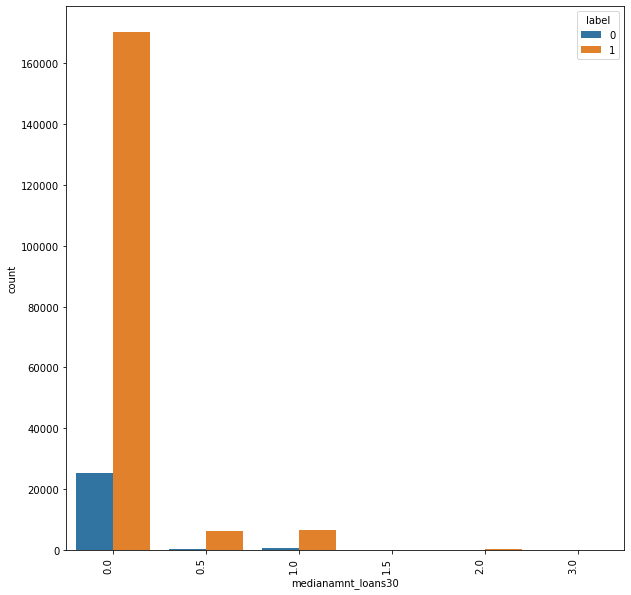
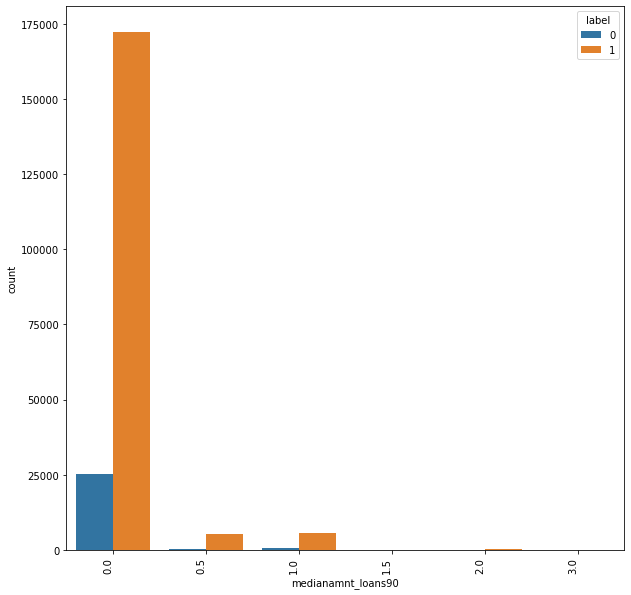
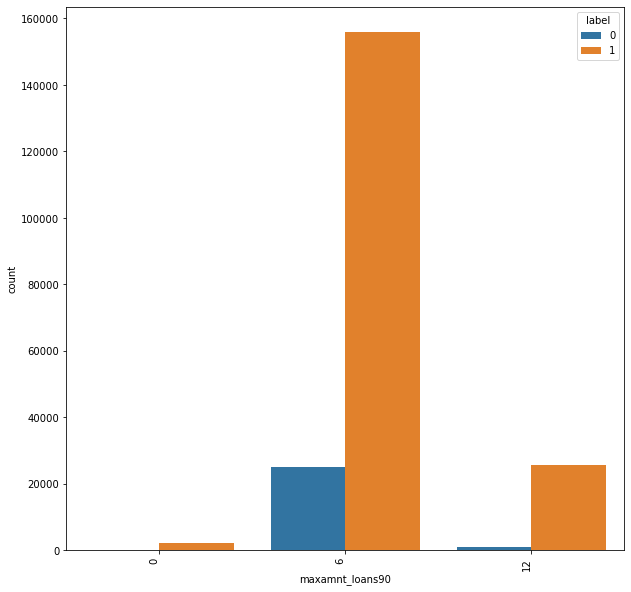
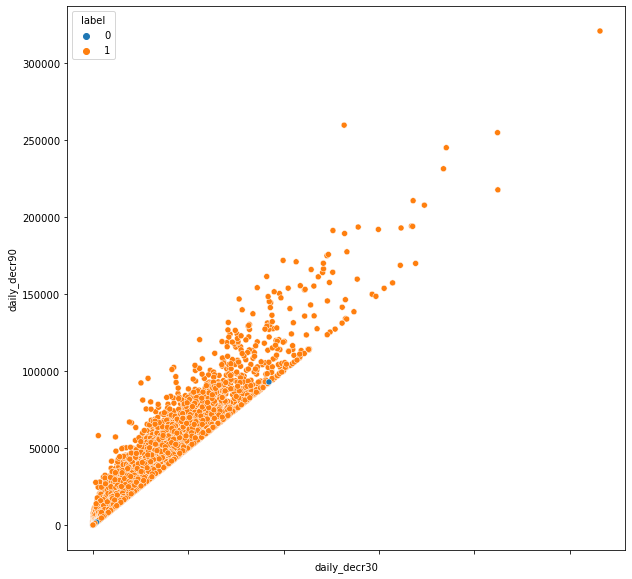
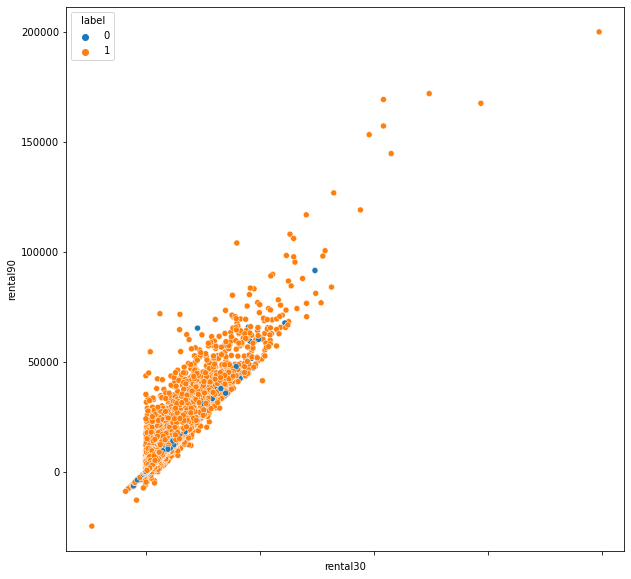
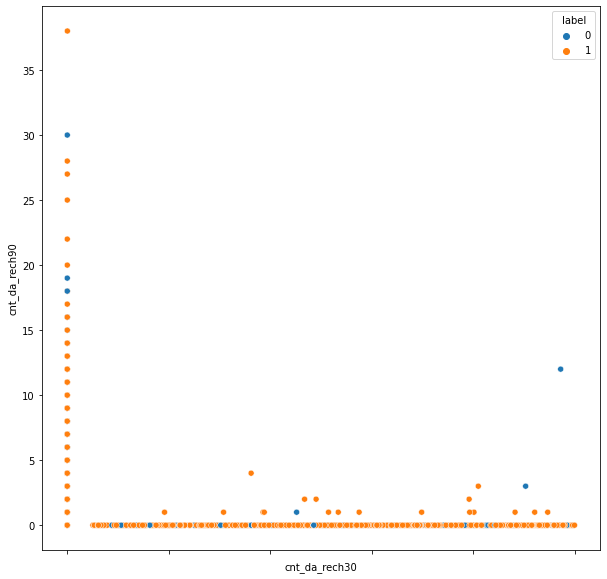
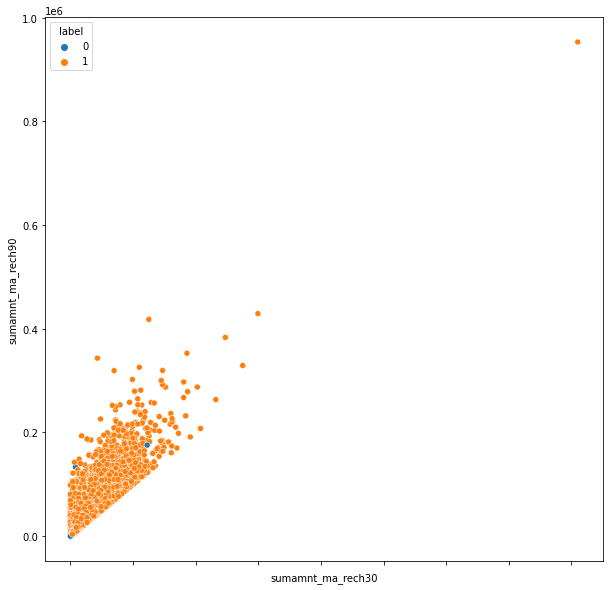
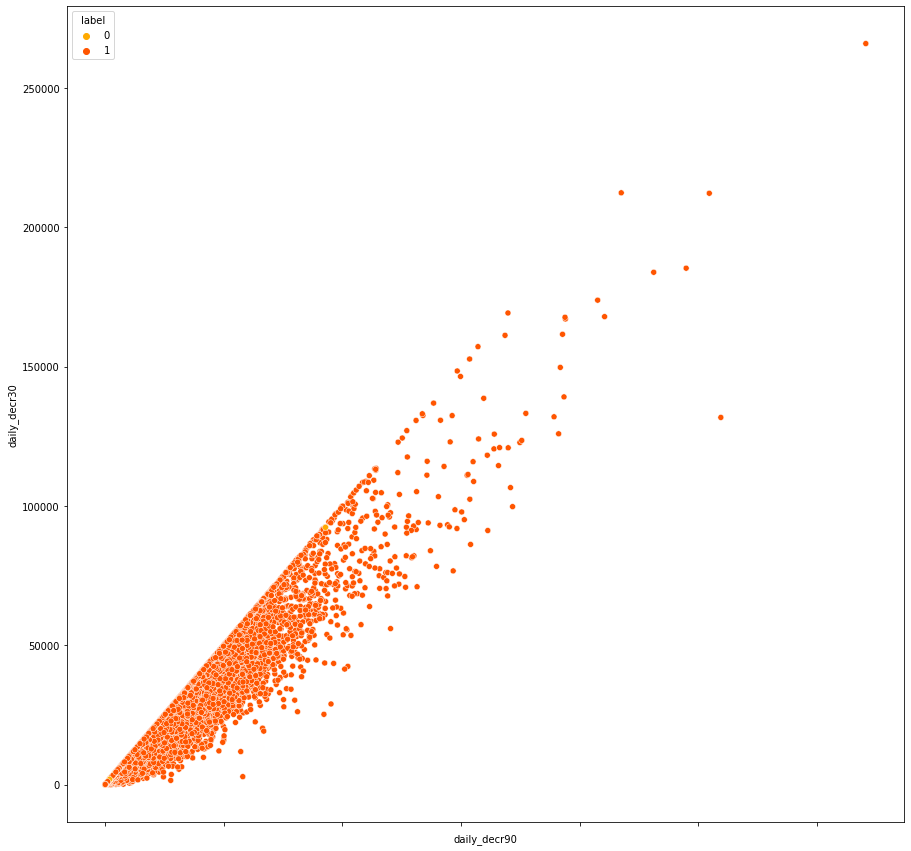












* Interpretation of the Results

Give a summary of what results were interpreted from the visualizations, pre-processing and modelling.

Answer: In a month the credits taken from 0th -20th days is more and from 20-30 day the credit decreases. The month of july has the highest loans and the month of august has the least loans.

**CONCLUSION**

* Key Findings and Conclusions of the Study

Describe the key findings, inferences, observations from the whole problem.

Answer: In a month the credits taken from 0th -20th days is more and from 20-30 day the credit decreases. The month of july has the highest loans and the month of august has the least loans.

Learning Outcomes of the Study in respect of Data Science

List down your learnings obtained about the power of visualization, data cleaning and various algorithms used. You can describe which algorithm works best in which situation and what challenges you faced while working on this project and how did you overcome that.

Answer: The column analysis has to be done in details as it involves financial decisions. Every column has to be studied and then dropped if not related.

* Limitations of this work and Scope for Future Work

What are the limitations of this solution provided, the future scope? What all steps/techniques can be followed to further extend this study and improve the results.

Answer: The dataset is imported in the python. Explorative data analysis is performed. The statistical analysis is performed by z-score method. The data loss is tried reduced to be minimal. The SMOTE method is used to reduce the bias in the data. The principle component analysis was also done. All the machine learning models were applied. The SVC model needs more computing time apart from that the random forest is found to have the highest efficiency. The accuracy of the model 0.8748278575644305 random\_state 13.The solution provided only has the accuracy of 87 approximately by using neural networks this accuracy can be increased