

Micro credit defaulter project

Submitted by:

Rebanta Dutta

**ACKNOWLEDGMENT**

I would like to acknowledge some of the websites from where I have taken help:

* <https://www.analyticsvidhya.com/blog/2022/02/a-quick-guide-to-bivariate-analysis-in-python/>
* https://medium.com/mlearning-ai/univariate-bivariate-and-multivariate-data-analysis-in-python-341493c3d173

**INTRODUCTION**

* Business Problem Framing

The microfinance institutions provides micro credit of mobile balances to the customers for 5days. The machine-learning model is to be built to identify where the customer would become a defaulter or not.

* Conceptual Background of the Domain Problem

Microfinance institutions provides small loans and other financial services to low income households. The concerned company in this project provide micro credit of telecom customers and demands to identify whether the customer would be a defaulter.

* Review of Literature

In the Indonesian market, there are more than 40000 micro finance institutions, who provides small loans to the customers. However, there are even large number of customers who end up defaulting the loan amount. This is a serious issue for the institutions and therefore a machine learning is to be built on the concept of prediction of the defaulter.

* Motivation for the Problem Undertaken

The primary motivation behind development of this machine-learning model is to help the micro finance institution recover the loans for development of both business and the national economy.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

After loading the dataset, I checked for the data types and found that there are 3 columns that are of object type and the rest were a mixture of float and int type. I performed label encoding to convert the object type into numeric type.

I performed EDA using countplot to visualise the change in the independent variables with respect to the dependent variable. From the plot I also figured that there is imbalance in the target or the output variable.

I also used distribution plot to visualise the distribution of the variables and found that some of the variables have skewness present in them.

Thirdly, I used Boxplot to identify presence of any outliers in the data. From the plot, I found that large number of outliers are present. I used Zscore method to eliminate some of the outliers while keeping in kind that the data loss is not more than 8%.

Lastly, I used the heatmap to visualise the correlation between the variables present in the dataset.

* Data Sources and their formats

The dataset contains a mixture of int, float and object type data. Each data in the dataset has some value and description.

|  |  |
| --- | --- |
| 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 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 |

Formats of the data are as follows:



* Data Preprocessing Done

Data cleansing is important as it helps to improve the accuracy of the model.

Step1: null values were checked

Step2: label encoding was performed to convert object type data into numeric type

Step 3: boxplot was used to visualise any presence of outliers in the dataset

Step 4: using zscore outliers were removed

Step 5: variance inflation factor was used for checking presence of multicollinearity

Step 6: Smote technique was used for balancing the data

Step 7: skewness was removed from the data using the log1p transformation.

* Data Inputs- Logic- Output Relationships

From the heatmap, I have identified there were large number of columns in the dataset on which the target or the output variable was dependent. Further, some of the columns, which has low correlation with the target variable was dropped.

* Hardware and Software Requirements and Tools Used

The Jupyter notebook was used for analysis and building of the machine learning model. In addition to this different libraries were used like the matplotlib, seaborn, pandas and numpy.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

After identification of object type data, I have used Label encoding technique to convert it into numeric type

Skewness was controlled using log1()

VIF was checked to reduce presence of multicollinearity in the dataset

* Testing of Identified Approaches (Algorithms)

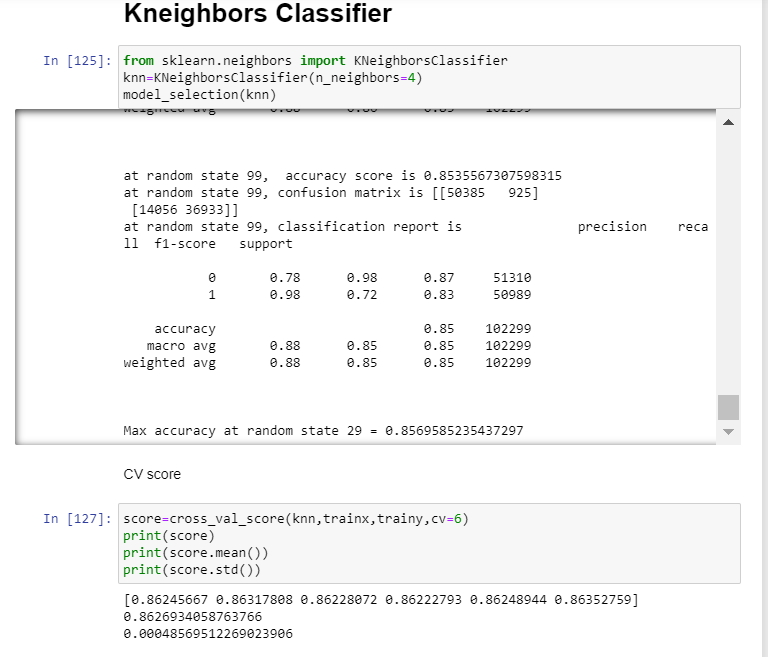
KNeighbor classifier

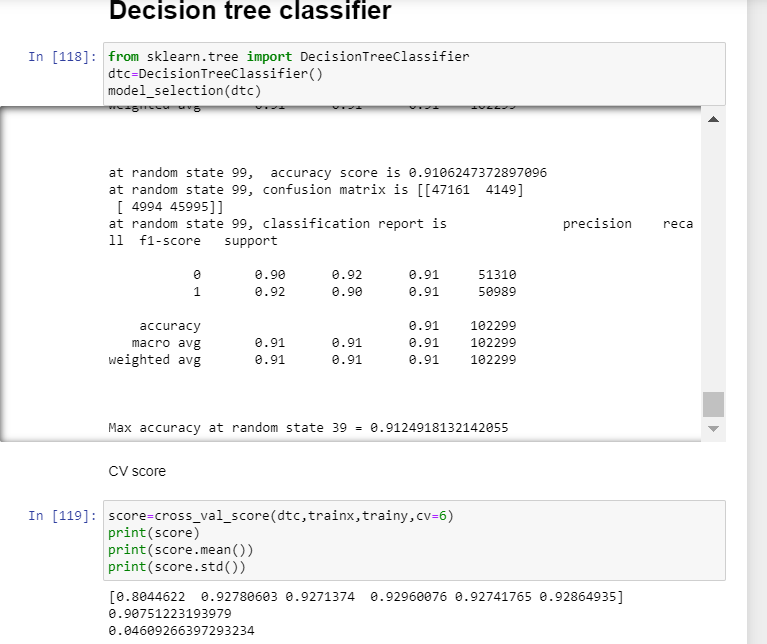
Decision tree classifier

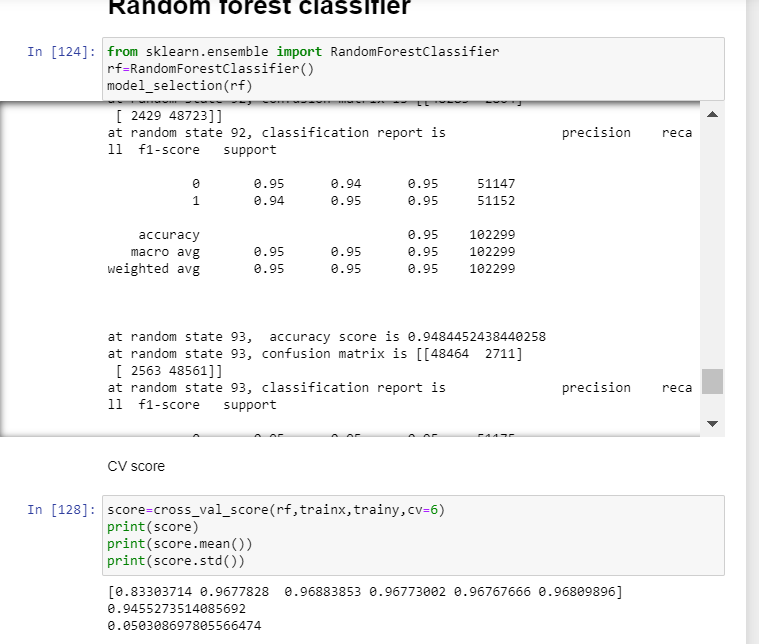
Logistic Regeression

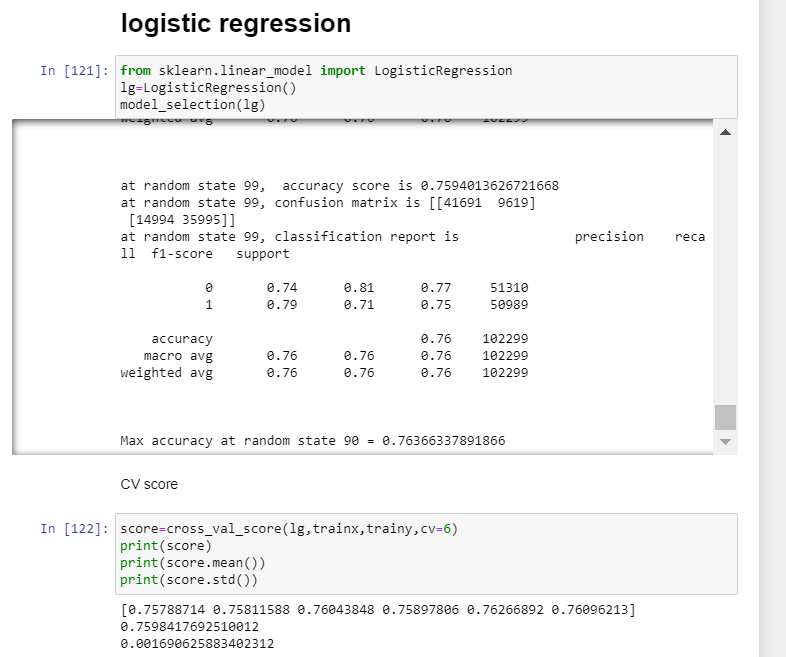
Random Forest Classifier

* Run and Evaluate selected models







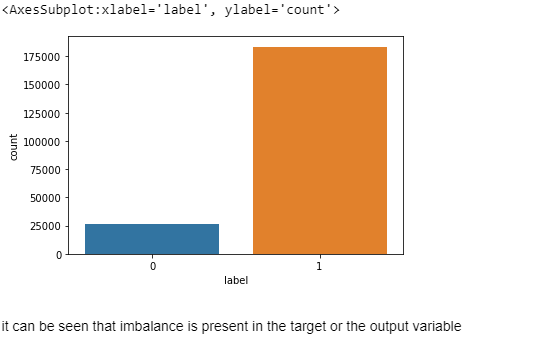


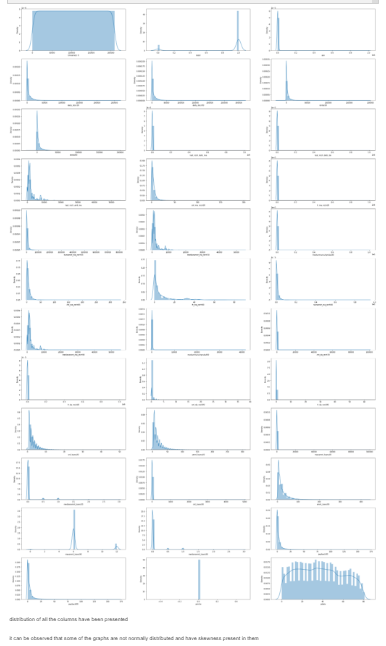
* Key Metrics for success in solving problem under consideration

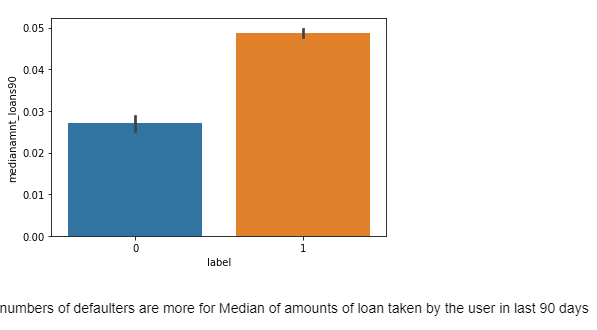
The key metrics that were used for problem solving are:

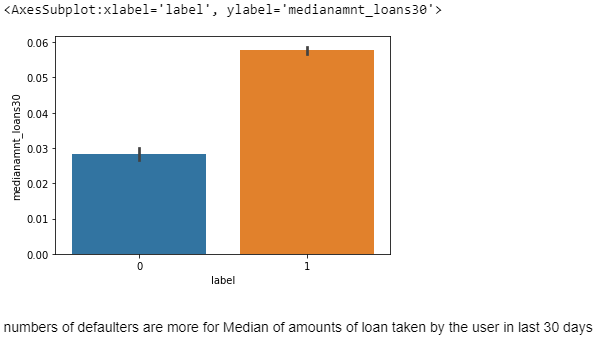
Accuracy score, classification report, confusion matrix, rocauc score.

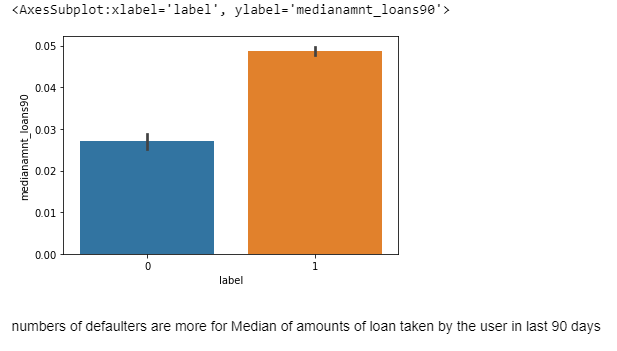
* Visualizations

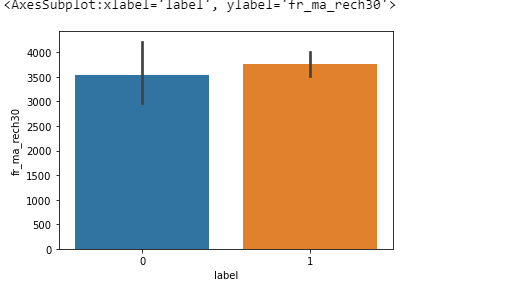


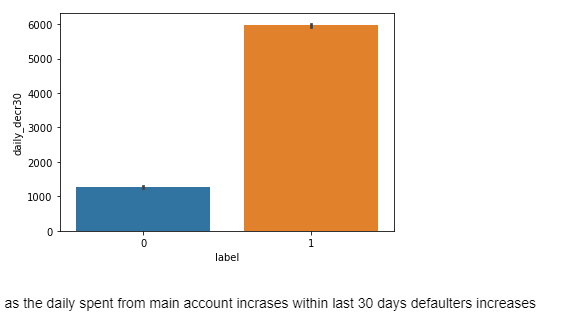


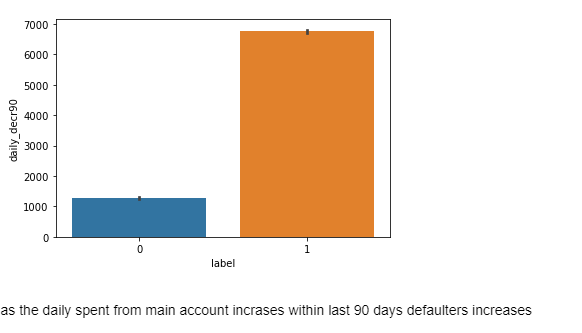


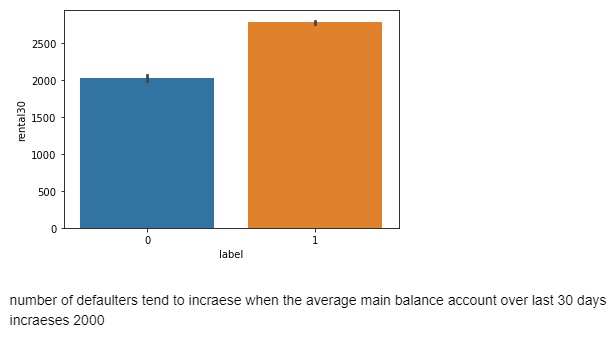


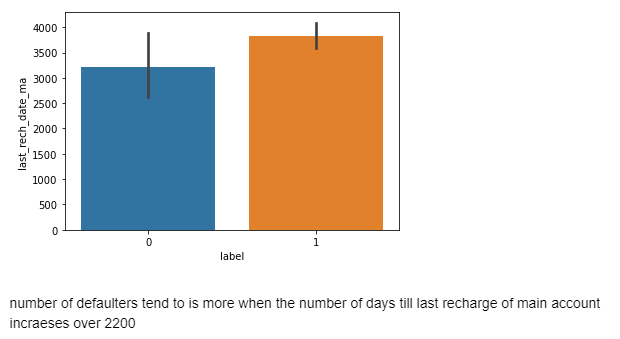


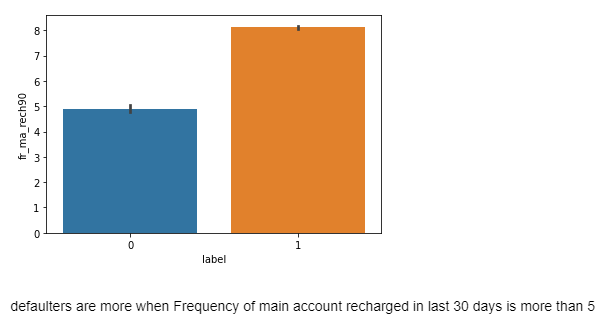


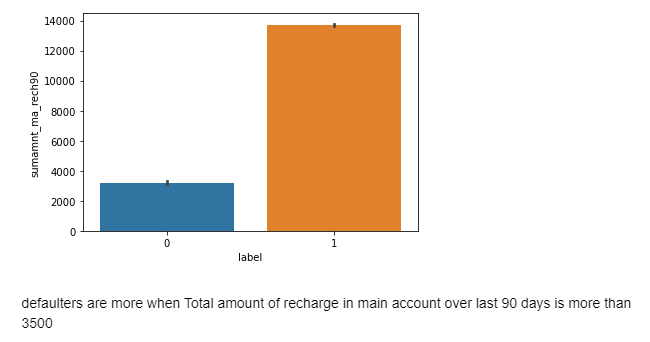


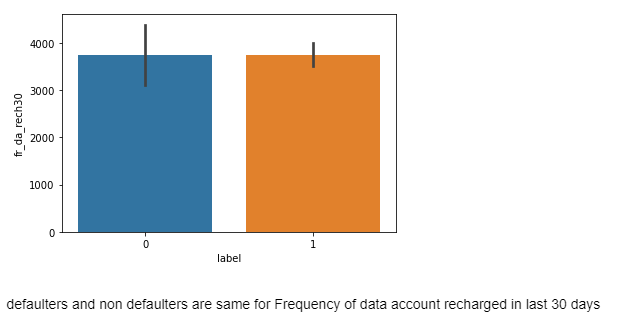


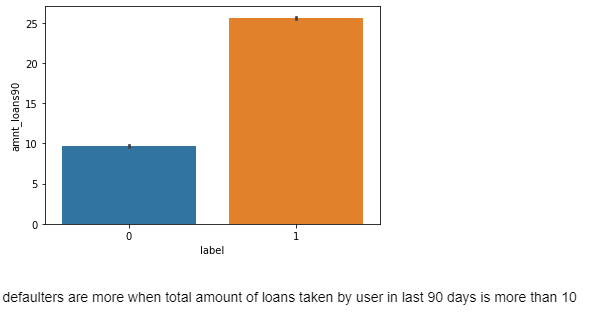


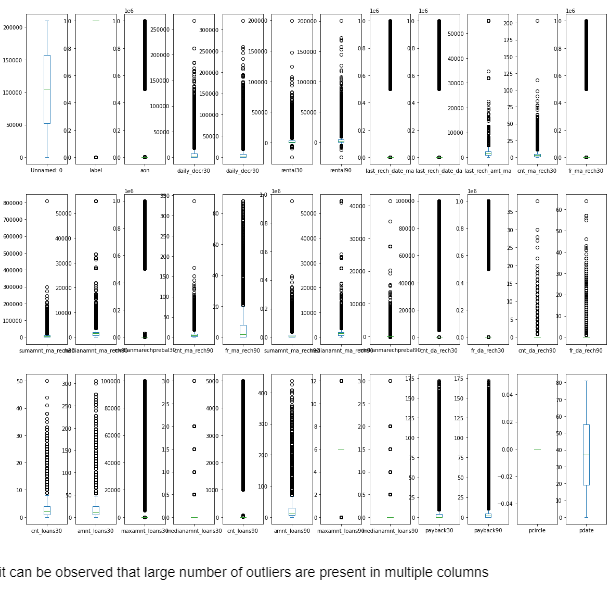


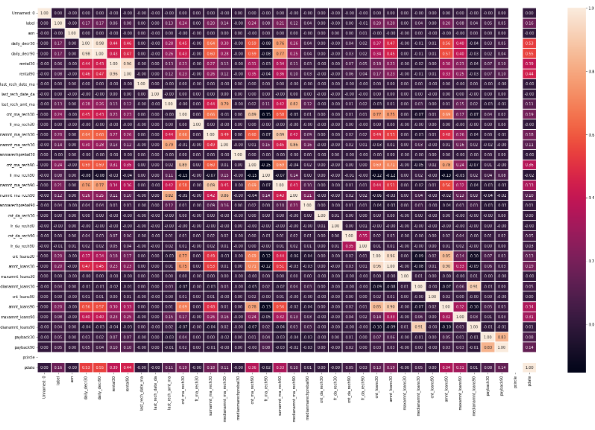












* Interpretation of the Results

It can be interpreted that the output variable has both positive and negative correlation with some of the input or the independent variables.

**CONCLUSION**

* Key Findings and Conclusions of the Study

It can be concluded that building and selecting the most appropriate ML algorithm is important for accurate prediction of the defaulter.

* Learning Outcomes of the Study in respect of Data Science

Data cleaning, visualisation is an important part of the process because it helps to improve accuracy of the model.

* Limitations of this work and Scope for Future Work

The limitations was in aspect of presence of outliers because removing all outliers is leading to large scale data loss. So we removed outliers to some extent.