

Micro credit loan project

Submitted by:

Pallavi Bajpai

**ACKNOWLEDGMENT**

I have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and references. I would like to extend my sincere thanks to all of them.

I am highly indebted to my institution i.e. Data Trained and the mentors and chat support team for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

I would like to express my gratitude towards my parents & member of Flip and work for their kind co-operation and encouragement which help me in completion of this project.

**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. The Microfinance services (MFS) provided by MFI are Group Loans, Agricultural Loans, Individual Business Loans and so on.

Many microfinance institutions (MFI), experts and donors are supporting the idea of using mobile financial services (MFS) which they feel are more convenient and efficient, and cost saving, than the traditional high-touch model used since long for the purpose of delivering microfinance services. Though, the MFI industry is primarily focusing on low income families and are very useful in such areas, the implementation of MFS has been uneven with both significant challenges and successes.

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.

Exercise: 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 this case, Label ‘1’ indicates that the loan has been payed i.e. Non- defaulter, while, Label ‘0’ indicates that the loan has not been payed i.e. defaulter.

* Conceptual Background of the Domain Problem

A good understanding of programming concepts along with some mathematic basic concepts like statistics , probability are very helpful. Thorough understanding of machine learning and the different models is also very important to solve this problem.

* Review of Literature

Considerable about of online research is done in order to understand the problem and requirement of ML in solving this problem.

* Motivation for the Problem Undertaken

Today, microfinance is widely accepted as a poverty-reduction tool, representing $70 billion in outstanding loans and a global outreach of 200 million clients.

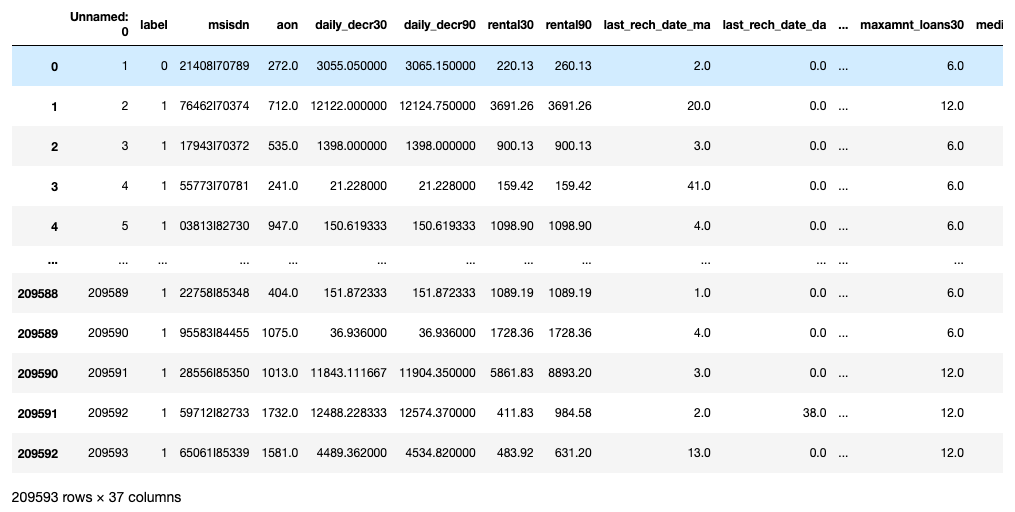
Telecommunications network provider are collaborating with an MFI to provide micro-credit on mobile balances to be paid back in 5 days. The Consumer is believed to be defaulter if he deviates from the path of paying back the loaned amount within the time duration of 5 days. For the loan amount of 5 (in 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).

Motivation behind picking this problem is to analyse the data of customers and make quick decision on if the loan will be repaid on time and should be given or not.

**Analytical Problem Framing**

* Data Sources and their formats

The dataset consists of both numerical and categorical variables. There is a total of 37 explanatory variables describing every aspect of the residential home. Overview of the dataset:



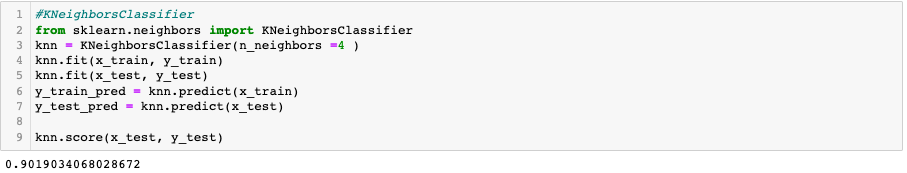
* Data Preprocessing Done
* Fair amount of data cleaning process was also involved. This includes checking for space and null values and then dealing with them. Also our available data has 37 columns.
* Duplicate data was removed.
* Outliers were removed but to an extent to maintain 7-8% of data loss.
* Data was imbalanced, so resampling was done.
* Hardware and Software Requirements and Tools Used
* Python code was written in Jupiter notebook. Below are the libraries needed in the process.
* Pandas, numpy, matplotlib.pyplot, sklearn.preprocessing, sklearn.model\_selection.cross\_val\_score, sklearn.linear\_model, sklearn.linear.svm, sklearn.ensemble.GradientBoostingRegressor, sklearn.metrics.confusion\_matrix,accuracy\_score, sklearn.metrics.f1\_score, scipy.stats, seabor.

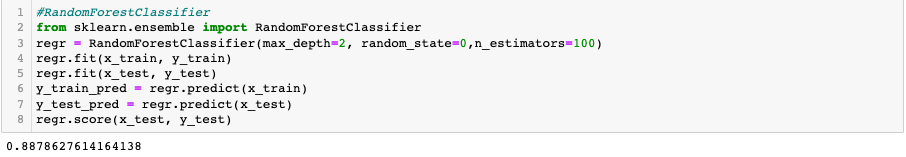
**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)
* Correlation heatmap was drawn to find the correlation between features and accordingly we selected few out of all features.
* Outliers were detected with the help of scatterplot and boxplot and were removed.
* Skewness was checked using histogram.
* Testing of Identified Approaches (Algorithms)

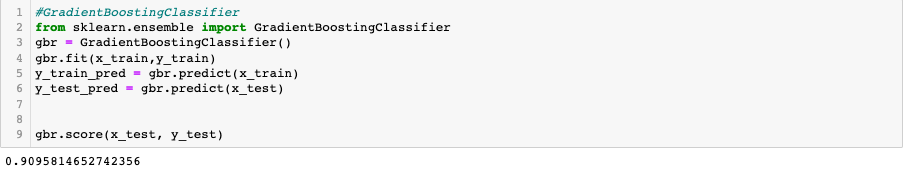
Below algorithms were used for training and testing of data

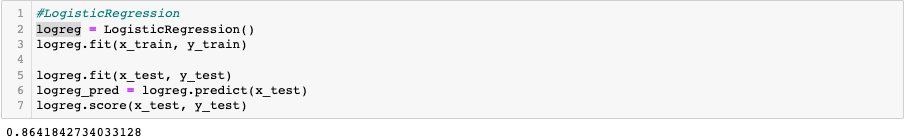
* Random Forest Classifier
* KNeighbors Classifier
* Decision Tree Classifier
* Gradient Boosting Classifier
* Logistic Regression
* Support Vector Classifier
* Gaussian NB
* Run and Evaluate selected models

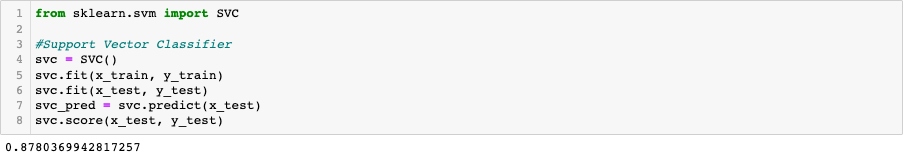


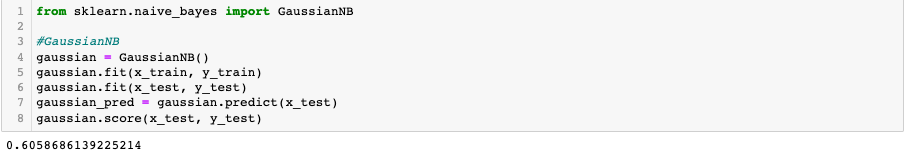


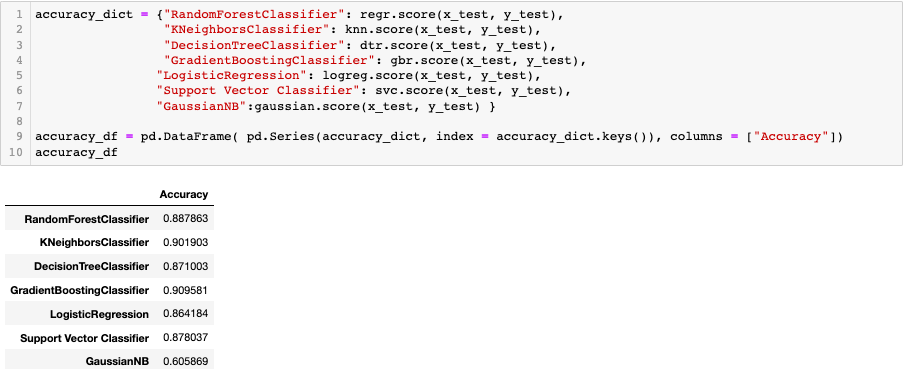










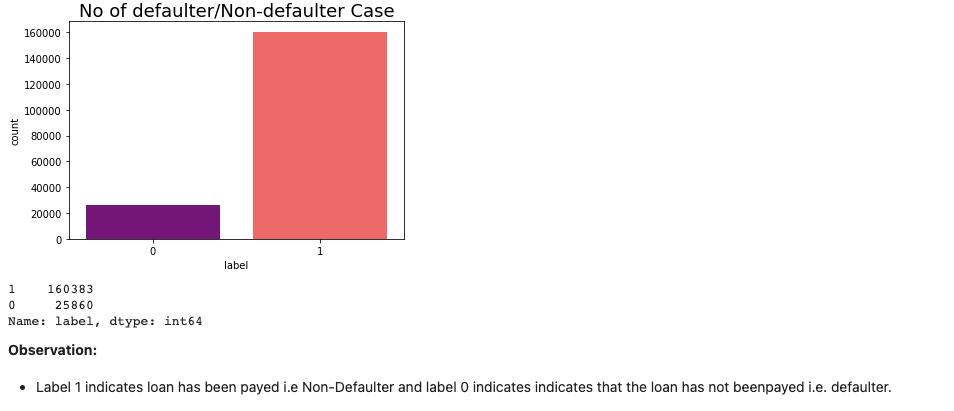


* Key Metrics for success in solving problem under consideration

Accuracy score was the key metric used to finalize the model.

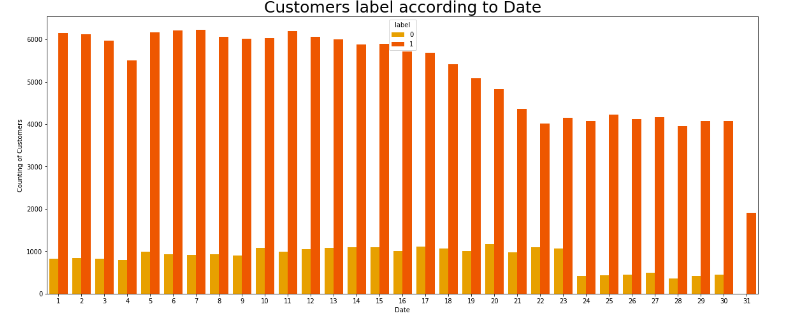
* Visualizations



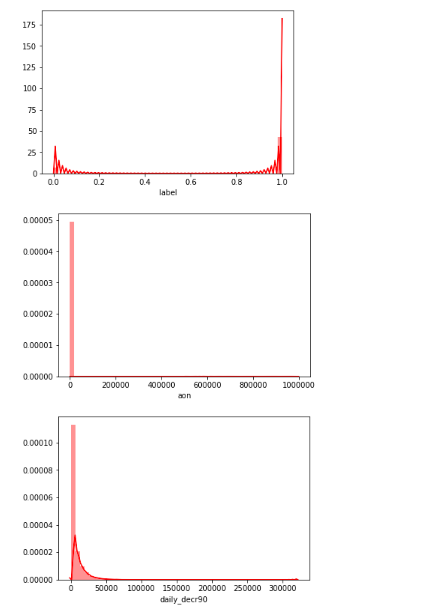


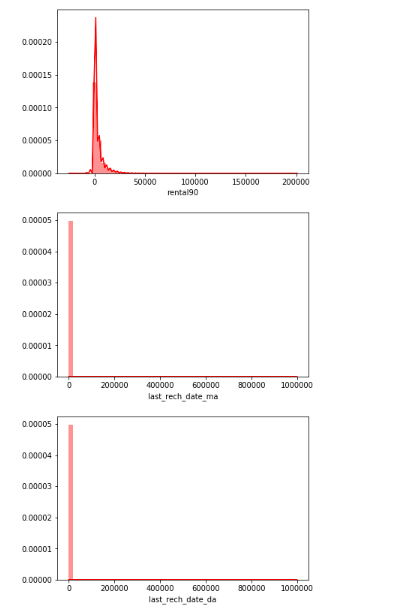


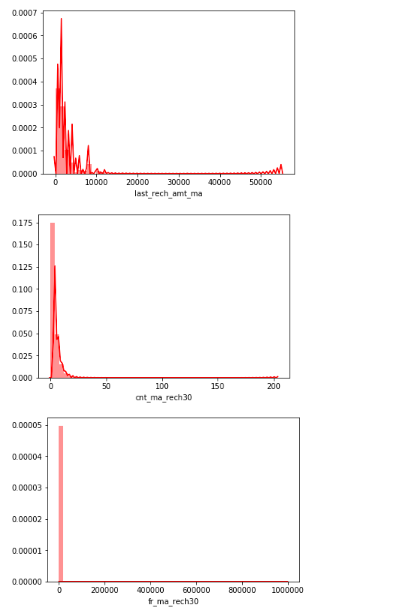


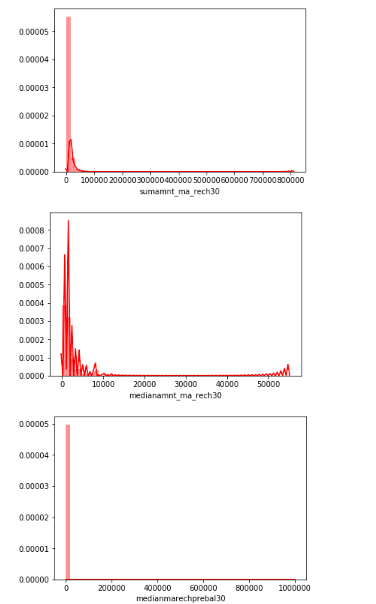


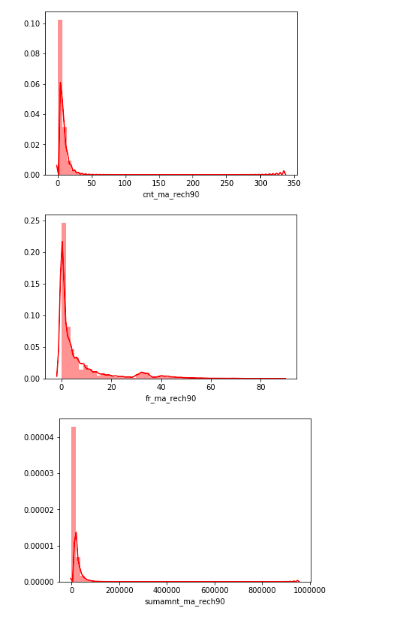


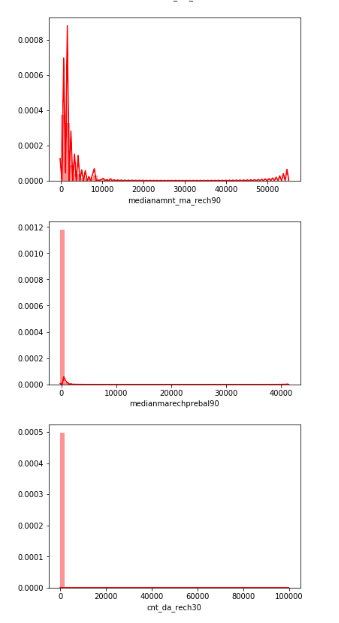


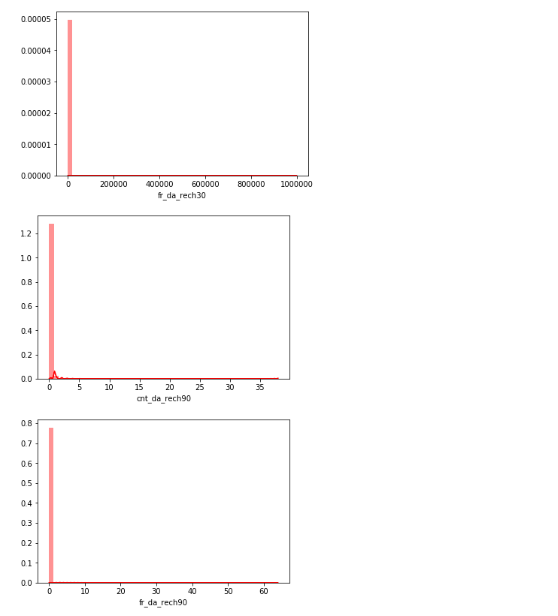


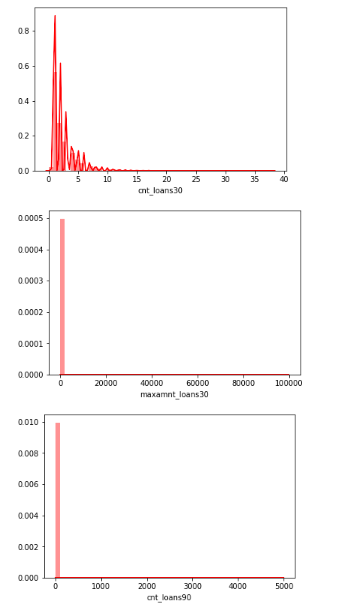


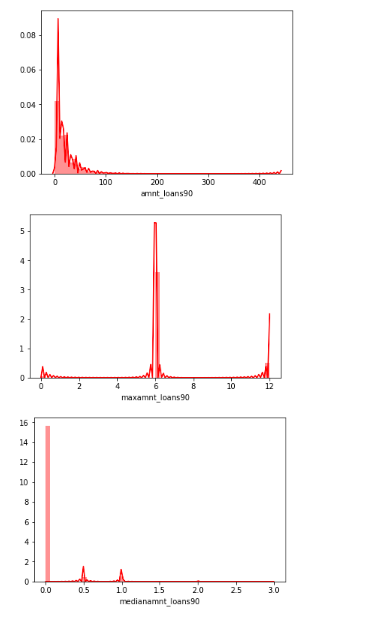


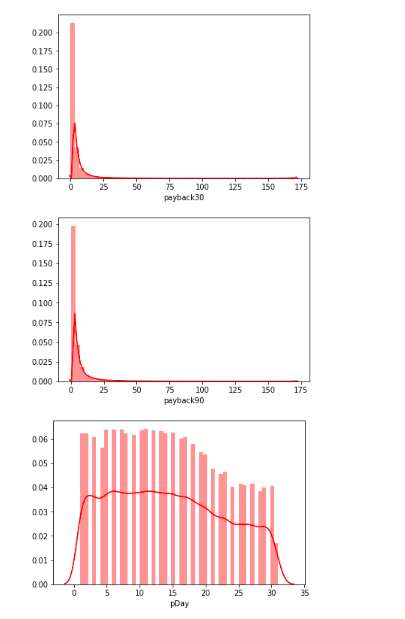


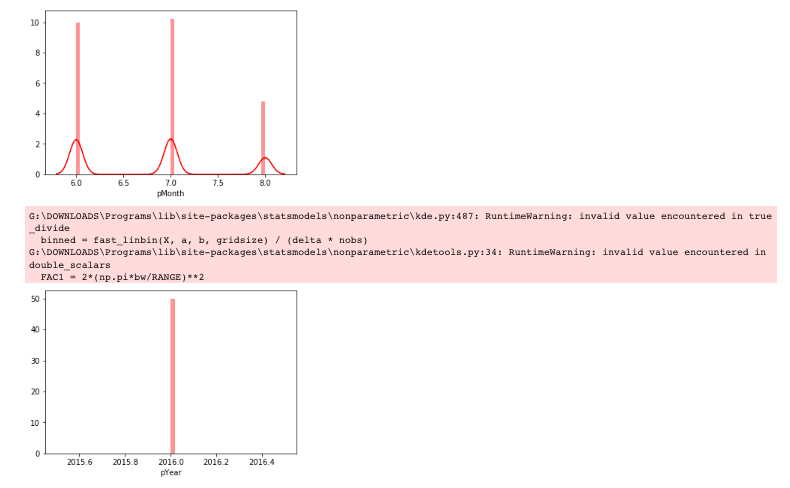


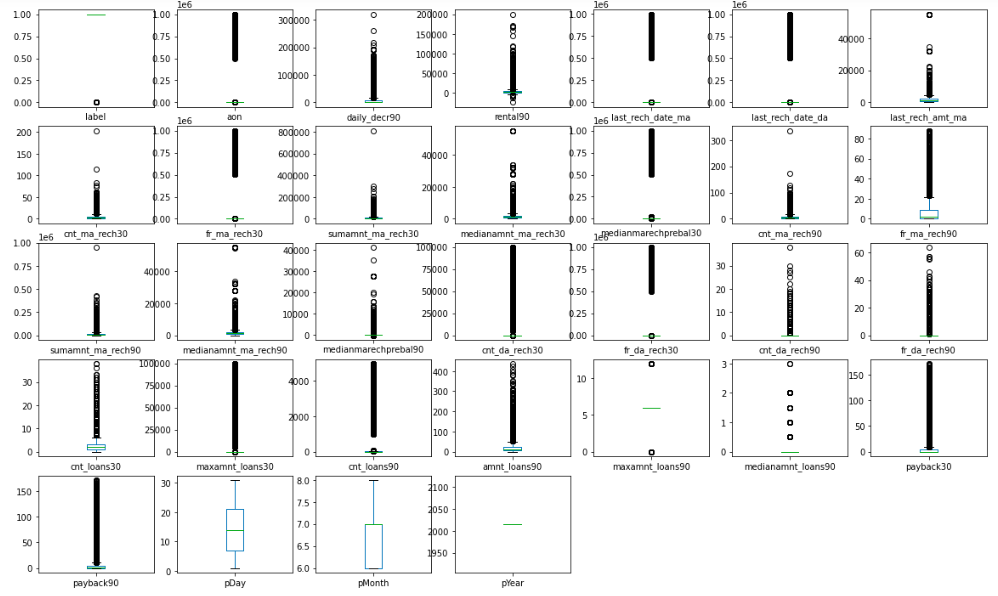












If different platforms were used, mention that as well.

* Interpretation of the Results
* No fraud 1 160383
* Fraud. 0 25860
* daily\_decr30 and daily\_decr90 features are highly correlated with each other.
* rental30 and rental90 features are highly correlated with each other.
* cnt\_loans30 and amount\_loans30 columns are highly correlated with each other.
* amount\_loans30 is also highly correlated with amount\_loans90 column.
* medianamnt\_loans30 and medianamnt\_loans90 is highly correlated with each other.
* We have to drop one of the features which are highly correlated with other feayures. And if we dont do this then our model will face multicolinearity problem.
* msidn is a mobile number of user and mobile number is unique for every customers. There are only 186243 unique number out of 209593 so rest of the data is duplicates entry so we have to remove those entry.
* The pdate column tells the date when the data is collect. It contains only three month data.
* Summary statistics shows all the statistics of our dataset i.e. mean, median and other calculation.
* Mean is greater than median in all the columns so aur data is right skewed.
* The difference between 75% and maximum is higher that's why outliers are removed which needs to be removed.

**CONCLUSION**

* Key Findings and Conclusions of the Study

Accuracy score was highest in Gradient Boosting Classifier hence we trained the model with the same.

* Learning Outcomes of the Study in respect of Data Science
* Correlation heatmap was drawn to find the correlation between features and accordingly we selected few out of all features.
* Outliers were detected with the help of scatterplot and boxplot and were removed.
* Skewness was checked using histogram.
* Limitations of this work and Scope for Future Work

Outliers of all features were not removed as it was giving huge data loss.