Rainfall prediction with Machine Learning

Parth Sharma, Aabhas Garg

I. Introduction

A. Problem

Rainfall prediction could be considered one of the oldest and important predictions in human history. Humans as well as animals have been using their senses to understand weather data and predict rainfall as water is the most essential life resource. In our computer era as well, this problem has continued but now with even more precise accuracy requirements. The beginning of rainfall predictions must have been around observing and studying the weather through senses each day and creating an intuition to the upcoming weather through memory and intelligence.

With the understanding of scientific methods however new mathematical methods evolved which made the predictions more accurate but only over longer time ranges. As the technologies for measuring weather data improved more data was needed to be calculated which required computers. To process large data artificial intelligence and machine learning can be used to make more accurate predictions.

The weather data such as humidity, temperature, location, wind speed and direction is recorded over a place after intervals and the rainfall is also measured to create a dataset.

Different machine algorithms can be used to target this problem over varied datasets and variables. But different models can yield different results on datasets based on where the data is recorded, which parameters affecting rainfall has been considered and the size of data. Therefore, predicting the rainfall pattern for a place requires study of the data and accurate comparison of different models and their scores.

This paper is based on our project which involved studying a dataset, processing the data and creating different models.

Then the scores of the different models are calculated and compared with each other.

B. Motivation and background

Rain affects several aspects of our life. Whether it is personal decisions affecting an individual for making daily decisions or big industrial and corporate decisions, a lot depends on relying on rain and weather. Studying rain and its cause and effects became the main motivation for this paper. Several problems which are caused by rainfall can be avoided and solved with accurate predictions some of which are discussed in this section.

Rain affects our daily lives throughout the year. From planning trips for business and other trips to making other daily decisions rain and weather forecasting can help a lot with all these decisions.

Climate change has become a major worry across in the globe and is drastically affecting weather sensitive areas. The problem can only be tackled using new technology and science. As our climate is becoming warmer the weather is becoming more extreme in places and this brings an increased number of natural calamities each year. Many calamities such as flooding, cyclones, water logging and extreme high tides come besides heavy and unexpected rainfall. Predicting rainfall and bad weather can help save a number of lives in such scenarios and also the more accurate the prediction is, the more lives can be saved.

Agriculture is greatly affected by rainfall. Artificial irrigation is still not that developed in many countries and rainfall is still required for irrigation and crop cultivation. Farmers nowadays rely heavily on planning their crop planting and irrigation on weather forecasts. Some crops such as rice and wheat requires a lot of water and hence are planted in the rainy season, their irrigation however has to be planned according to rain patterns. Fertilizers are also used in dry conditions because water can wash away the fertilizer which can result in huge losses. Farmers can benefit from forecasts for deciding which day they should fertilizers and also pesticides. Field workability is also to be planned on a cloudy but with rainless day.

Construction projects both on county wide or individual have to be planned keeping rain in mind. Huge investments can be washed away because of unpredictable rains. Activities such as digging, cementing or road construction are needed to be planned on dry days. And unpredicted rainfall if predicted early can save a lot of money as a lot of construction material and structures can be preserved early. Also, it can save a lot on labor costs.

Various sports such as cricket and other outdoor sport events rely on rainfall predictions. The accuracy of such short term rainfall predictions can cause a lot of loss or profit for the venue as the rainfall can cancel a game resulting in the cancellation of a match. Accurate rainfall prediction can help to shift the match at right times to avoid rainfall from affecting the match and its viewership.

All these applications of rainfall prediction motivated us to work on this project and the research that followed.

II. Previous Work related to the project

Various research has been done on rainfall prediction and we have studied and cited some in our paper. In this section a few are discussed.

In a recent paper published in Ain Shams Engineering journal a case study of Terrenganu, Malayasia was done. Various machine learning methods are used, and a comparative study is done based on the results and an efficient model was developed. Several methods for comparing the methods were used such as comparing errors and scores. Four different ML algorithms namely Linier regression, boosted decision tree regression, decision forest regression and neural network regression were used. The performance indictors to compare the algorithms were mean absolute error, mean square error, relative absolute error, relative square error and coefficient of determination. Boosted decision tree was found to be the best with the least error. The study helped us with implementing our algorithms in our paper.[1]

In a paper published in 2013 in international journal of engineering research and technology, a model of rainfall prediction was made based on data mining principles which was dynamic in nature.[2]

In a paper published in 2016 in International Journal of Computer Application, a vivid descriptive study of different rainfall prediction studies was done. The motivation behind the project was agriculture in India. It describes thirteen different methods used for prediction with clear steps and flow charts. It helped us through our project to study the methods. But it does not provide any empirical data on the accuracy of these methods and doesn't provide a comparative study. It is research providing great reference and hence cited here.[3]

III. Proposed Methodology

The flow chart in Figure 1 describes the tentative solution we propose for the problem.

In the first step the data is explored such as the shape of data and column labels. The data is converted into binary to provide easy modeling values. Then the data is plotted to see the imbalances.

Then the imbalanced data is balanced by oversampling or subsampling. Then, some data processing is done by plotting correlation matrices and other plots.

Then the best features are selected by methods such as random foresting.

The data is split into train and test and trained for different models such logistic regression etc. The different ROC curves and accuracy values are observed.

Then the decision regions are plotted to compare the models.

Finally, direct comparison of models is done by comparing them on accuracy scores and area under the ROC.

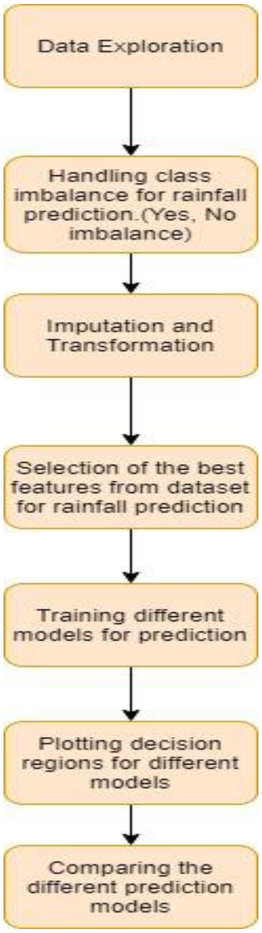

Data Exploration
Handling class imbalance for rainfall prediction.(Yes, No imbalance)
Imputation and Transformation
Selection of the best features from dataset for rainfall prediction
Training different models for prediction
Plotting decision regions for different models
Comparing the different prediction models

Figure 1

The dataset has 23 features that can affect the target feature which is the feature holding data if it will rain tomorrow. The models are to be trained with these features. The shape of the data originally is (142193,24) which means it has 23 columns (features) and 142193 rows (data points).

The data is preprocessed and as a first step the data is checked for imbalances. Then, to balance the dataset the resample utility from sklearn library is used (Figure 2,3).

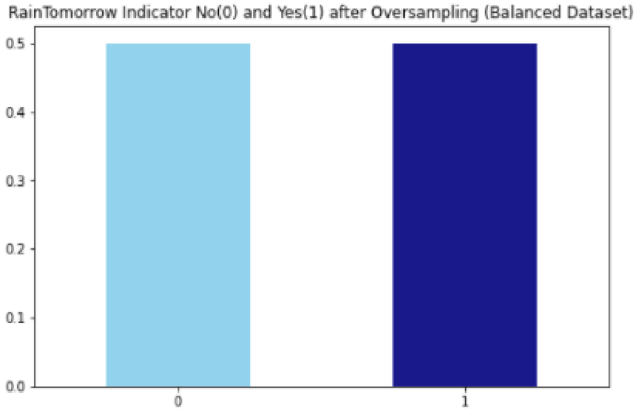


Figure 2

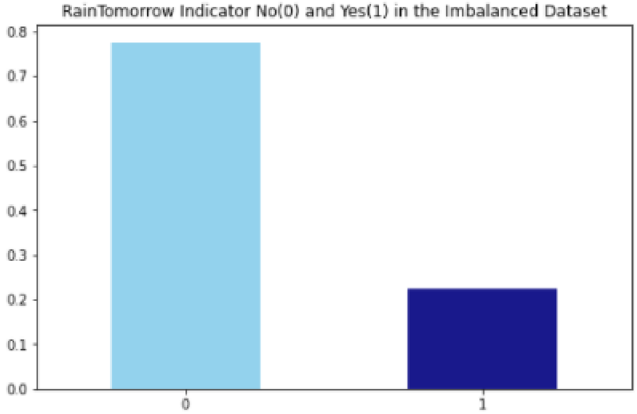


Figure 3

The missing values are now checked using seaborn library. We observed that no feature has less than 50 percent missing values. Hence, all the data is imputated and used for training.

The binary values are converted into numerical values using label encoding. Labels contained ranges of numerical values which are now checked for outliers. The outliers are removed using the MICE package.

The correlation between the features is found out by creating a correlation matrix (Figure 4). The correlation matrix shows that there is correlation between some features, but none is very near to 1 so we don't omit any feature here.

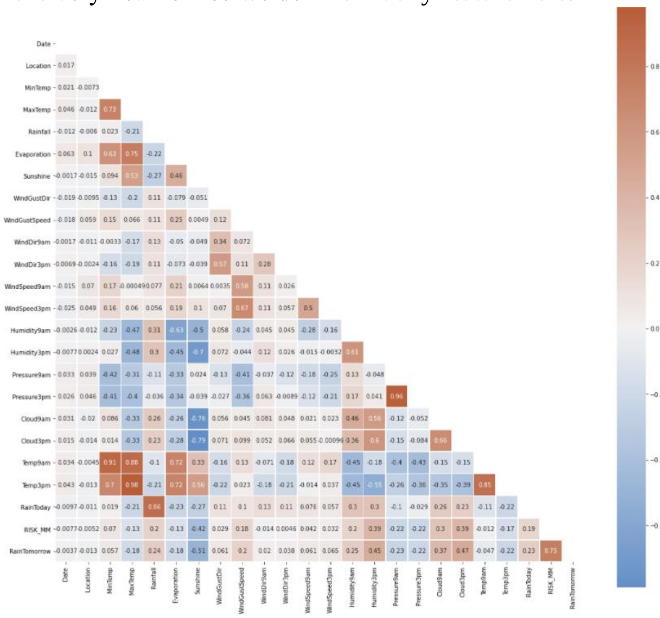


Figure 4

To be sure we create a pair diagram to find out the overlapping clusters in the data points. It shows clear distinction between the 'yes' and 'no' clusters and minimal overlap. Finally, the best features are selected using the filter method (chi square value) and wrapper method (random forest) after normalizing the values using MinMaxscaler in sklearn library. Hence, we proceed with model training and testing.

For our comparative study we use eight different machine learning algorithms which now we will discuss one by one.

Linear regression is a supervised machine learning algorithm which uses linear approach for modelling data. It works by fitting the data onto a linear equation. The numerical data we preprocessed in the above steps is fitted onto a linear equation and the prediction model is trained and tested. Sklearn library is used for the model.

Decision tree is also a type of supervised learning algorithm used for predicting by learning specific decision rules while going through the data. This model is also used from the sklearn library.

We use the neural network from sklearn library to create a prediction model too.

Random forest is also a supervised machine learning algorithm which uses ensemble learning to combine many classifiers to make predictions. We also used RandomforestClassifier from sklearn.

Light GBM is an open source distributed gradient boost algorithm which uses decision tree algorithms for ranking and predicting the data. We made a LightGBM model using the library LightGBM.

CatBoost is also a gradient boosting algorithm which uses a permutated algorithm is compared to other boosting methods. We also use it from the CatBoost library.

XGBoost is an efficient and portable gradient boosting algorithm designed for speed and performance. We also use this from the XGBoost library.

We found that XGBoost, CatBoost and Random foresting perform better as compared to others. However, linear regression performs the best when it comes to processing time.

Naïve Bayes is a group of supervised machine learning algorithms which uses bayes theorem for prediction. We used Gaussian Naïve Bayes algorithm which is a quicker alternative.

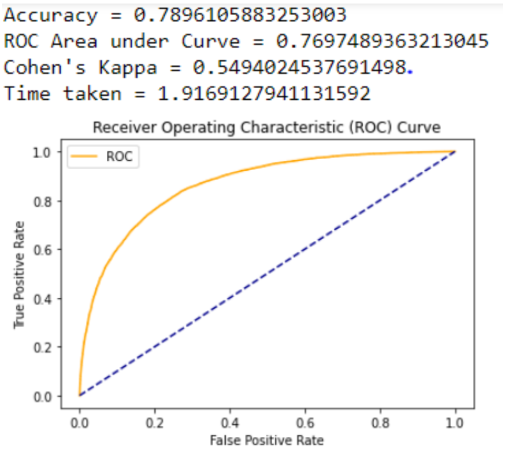
IV. Result & Analysis

The following section is based on the result and analysis of our comparative study. First, we will discuss the results of the different algorithm and then we will follow it with the comparative graphs and figures.

A. Hyperparameter Analysis

On applying the different algorithms, the following results were produced. The different scores calculated for each algorithm is accuracy, Roc area under curve, Cohen's Kappa and Time taken. Accuracy is the score measured by how good an algorithm is in defining patterns between test variable and training variables. The area under the curve gives the measure of the usefulness of the model or the ability to classify between the classes. Cohen's Kappa score measures the reliability between two raters and time taken is the processing time for the training the testing.

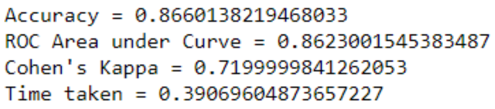
(Figure 5) Linear regression gives the least accuracy among all which can be seen in the accuracy score. For linear regression the Roc area under the curve is also the least. Cohen's Kappa score is also the least which shows lowest reliability. The only benefit shown is the time taken which is the least.



```
Accuracy = 0.7896105883253003
ROC Area under Curve = 0.7697489363213045
Cohen's Kappa = 0.5494024537691498.
Time taken = 1.9169127941131592
```

Figure 5

The figure below (Figure 6) shows the results for the decision tree algorithm. The accuracy improved over linear regression as it is a more accurate algorithm. The ROC Area under Curve is also increased as the model's usefulness is also more. Cohen's Kappa score is also more. The time taken is increasing with the complexity of the algorithm which maybe a downside.



```
Accuracy = 0.8660138219468033
ROC Area under Curve = 0.8623001545383487
Cohen's Kappa = 0.7199999841262053
Time taken = 0.39069604873657227
```

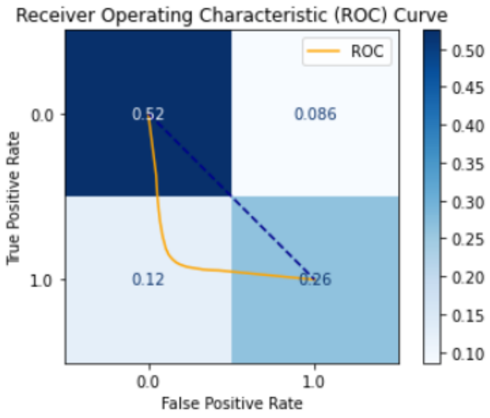
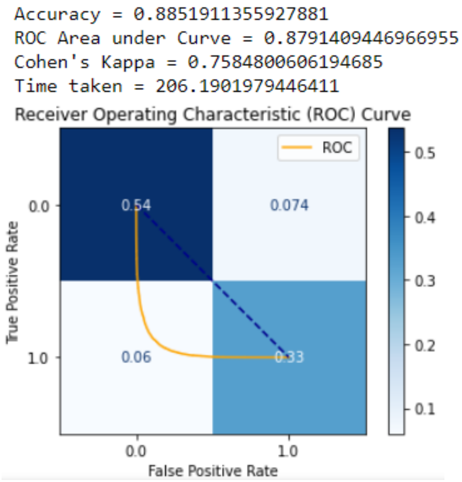


Figure 6

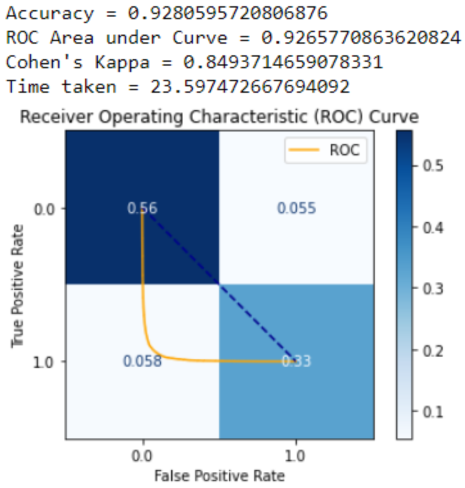
Neural Network is a more complex algorithm which uses complex trees and networks to make prediction. It also employs multiple algorithms. The accuracy improved compared to previous algorithms (Figure 7). It performs well in other tests as well. The time taken is very high so, it requires better processing power.



```
Accuracy = 0.8851911355927881
ROC Area under Curve = 0.8791409446966955
Cohen's Kappa = 0.7584800606194685
Time taken = 206.1901979446411
```

Figure 7

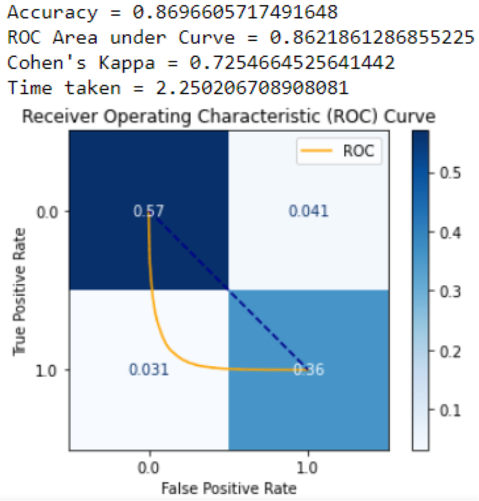
Random foresting is efficient and a very accurate algorithm for large data. It uses multiple decision tree and that's why it's very fast and can even neglect missing data. That's why the accuracy score is very high (Figure 8), and the time taken is also less compared to the gradient boosting algorithms.



```
Accuracy = 0.9280595720806876
ROC Area under Curve = 0.9265770863620824
Cohen's Kappa = 0.8493714659078331
Time taken = 23.597472667694092
```

Figure 8

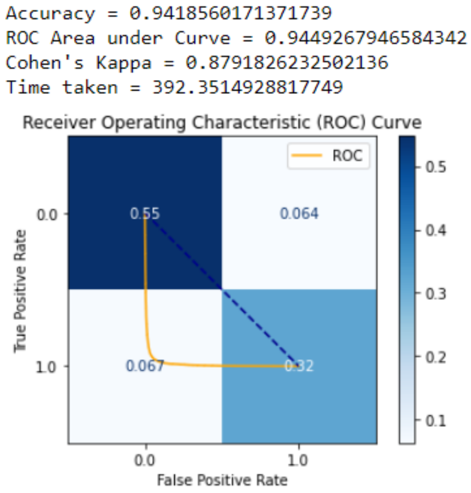
The Light GMB performs good in accuracy score and time. It is a quick and efficient algorithm. (Figure 8)



```
Accuracy = 0.8696605717491648
ROC Area under Curve = 0.8621861286855225
Cohen's Kappa = 0.7254664525641442
Time taken = 2.250206708908081
```

Figure 9

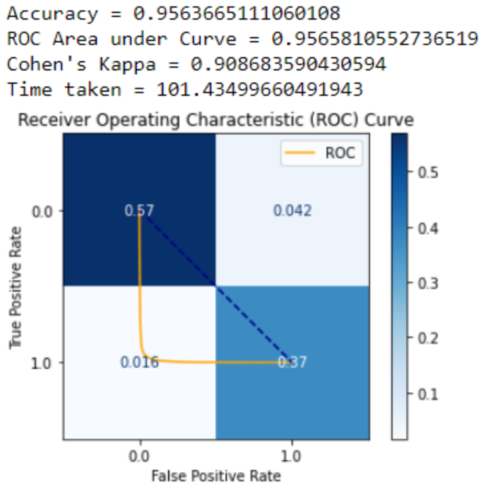
CatBoost is an advanced algorithm and performs well. The accuracy is very high and so is the Roc are under the curve. Cohen’s Kappa is also high which means more reliability. It is slow to process.



```
Accuracy = 0.9418560171371739
ROC Area under Curve = 0.9449267946584342
Cohen's Kappa = 0.8791826232502136
Time taken = 392.3514928817749
```

Figure 10

XGBoost is an efficient algorithm designed for speed and performance and hence it performs the best at accuracy and comparatively better at time taken than other complex algorithms. The other scores the best too. (Figure 11, 12)



```
Accuracy = 0.9563665111060108
ROC Area under Curve = 0.9565810552736519
Cohen's Kappa = 0.908683590430594
Time taken = 101.43499660491943
```

Figure 11

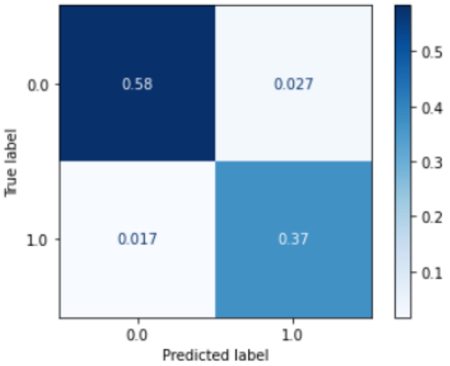
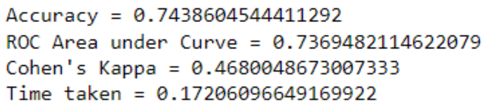
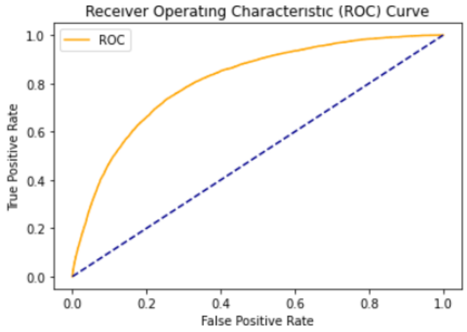


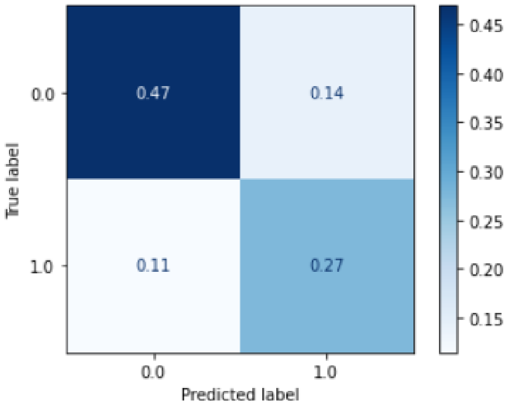
Figure 12

Naïve Bayes is a group of supervised machine learning algorithms which uses bayes theorem for prediction. We used Gaussian Naïve Bayes algorithm which is a quicker alternative.



```
Accuracy = 0.7438604544411292
ROC Area under Curve = 0.7369482114622079
Cohen's Kappa = 0.4680048673007333
Time taken = 0.17206096649169922
```





B. Comparison

The below decision region plots compare the different algorithms. It shows the class limits between the datapoints and the regional boundaries. CatBoost shows a definite regional boundary and XGBoost and Random Forest also has a smaller number of misclassified data points.

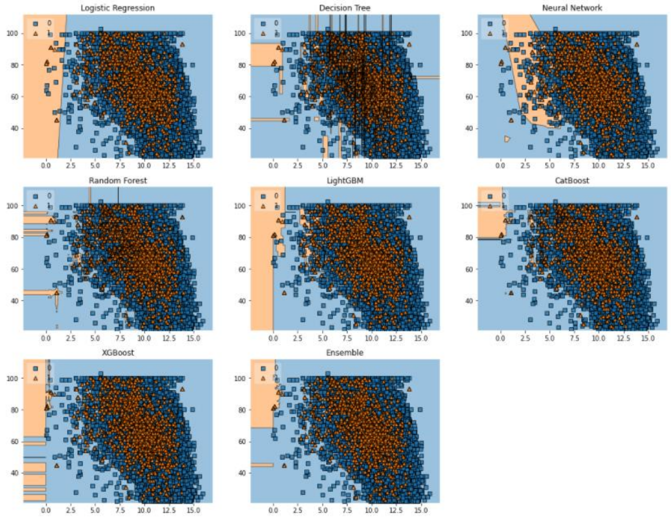
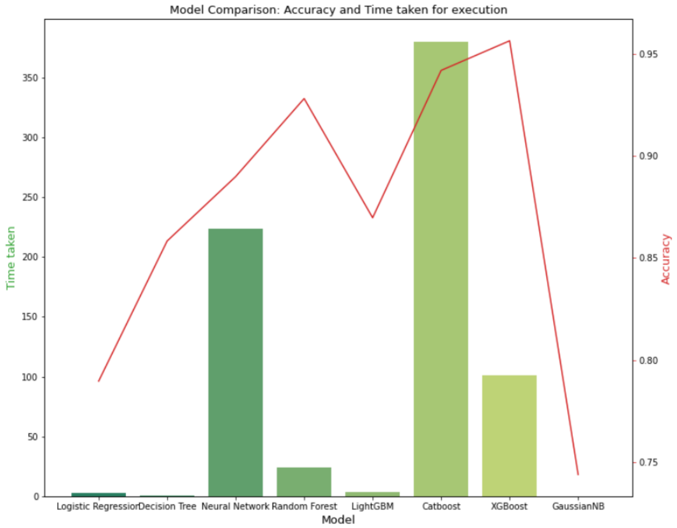
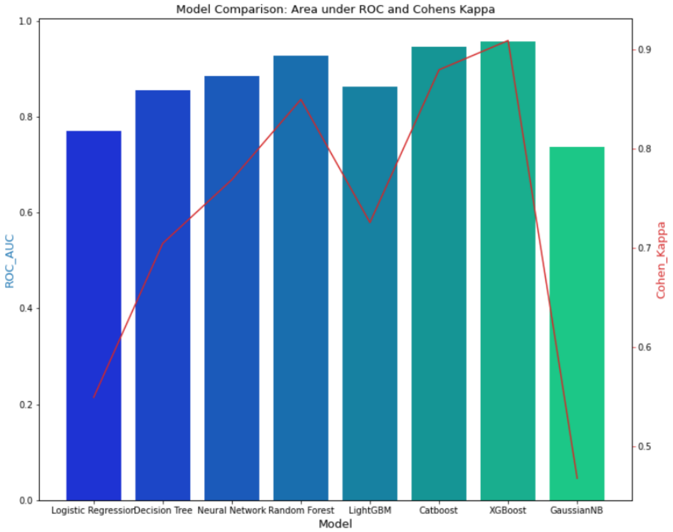


Figure 13

The Graph below directly compares the accuracy and Time taken for all the algorithms. The accuracy is the most for XGBoost algorithm. CatBoost and Random Forest performed very well too. Logistic Regression is the fastest of all the algorithms.



Below is the comparison between the Cohen’s Kappa score and Roc area under curve. XGBoost performed the best.



V. Conclusion

Rainfall prediction is a simple yet complicated task. Many supervised models can be applied for the prediction. This research provides the analysis of seven different algorithms. The data preprocessing steps are also explained which are important for the understanding of the research.

Rainfall prediction technology continue to evolve and improve with time. With more processing power more complex algorithms can be used over larger data and hence it is important to employ new algorithms for this problem to serve the population.

References

[1] Ridwan, Wanie M., et al. "Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia." *Ain Shams Engineering Journal* 12.2 (2021): 1651-1663. https://www.sciencedirect.com/science/article/pii/S2090447920302069

[2] Gaikwad, Ganesh P., and V. B. Nikam. "Different rainfall prediction models and general data mining rainfall prediction model." *Int. J. Eng. Res. Technol* 2 (2013): 115-123.

[3] Hirani, Dhawal, and Nitin Mishra. "A survey on rainfall prediction techniques." *International Journal of Computer Application* 6.2 (2016): 28-42.