**CAPSTONE PROJECT - FINAL REPORT**



|  |  |
| --- | --- |
| Batch Details | PG-Program in Data Science and Engineering OCT-20 |
| Team members | Sujitha R, Bathula Tejesh, Shantha Reddy, Kumar Suraj, Santhosh Sriram |
| Domain of the project | Weather Forecasting |
| Proposed Project title | Rainfall Prediction |
| Group number | 07 |
| Team Leader | Sujitha R |
| Mentor name | Vikash Chandra |

**Project Group Info**

Date: 07-05-2021

Vikash Chandra Sujitha R

Signature of mentor Signature of Team Lead

**TABLE OF CONTENTS**

Title Page no

1.List of Figures……………………………………………………………………………….4

2.Industry Review

2.1 Industry Review………………………………………………………………………6

2.2 Literature Survey…………………………………………………………………......6

3.Dataset and Domain

3.1 Data Dictionary……………………………………………………………………….7

3.2 Variable Categorization……………………………………………………………….8

3.3 Pre-process data analysis……………………………………………………………...9

3.4 Data Describe………………………………………………………………………...10

4.Project Justification

4.1 Problem Statement…………………………………………………………………...11

5.Data Exploration (EDA)

5.1 Univariate Analysis………………………………………………………………….11

5.2 Bivariate Analysis……………………………………………………………………15

5.3 Multivariate Analysis………………………………………………………………...19

6.Feature Engineering

6.1 Outlier treatment and detection……………………………………………………...21

6.2 Transformation of data………………………………………………………………21

6.3 Null value imputation………………………………………………………………..23

6.4 Iterative Imputer method…………………………………………………………….23

6.5 Statistical Significance variables…………………………………………………….24

7. Approach

7.1 Logistic Model………………………………………………………………………27

7.2 Class imbalance using SMOTE……………………………………………………..30

7.3 Feature Selection using RFE Technique………………………………………….…31

7.4 Random Forest Model……………………………………………………………….33

7.5 Hyper Parameter Tunning for Random Forest Model……………………………….34

7.6 Boosting Models……………………………………………………………………..36

7.7 Hyper Parameter Tunning for LightGBM……………………………………………38

7.8 Model Comparison…………………………………………………………………..39

7.9 Model Stability using KFold…………………………………………………………40

8.Conclusion

8.1 Future work…………………………………………………………………………..41

8.2 Conclusion…………………………………………………………………………...41

9.Reference

9.1 Reference…………………………………………………………………………….41

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| Figure no | Description | Page no |
|  |  |  |
| 1 | Importing the libraries and displaying the dataset | 7 |
| 2 | Shape and count of numerical and categorical features | 8 |
| 3 | Variable categorization | 8 |
| 4 | Distribution of numerical and categorical features | 8 |
| 5 | Dataset info | 9 |
| 6 | Converting Date feature into day, month and year | 10 |
| 7 | Five-point summary for numerical features | 10 |
| 8 | Five-point summary for categorical features | 10 |
| 9 | Univariate distributions for numerical features | 11-13 |
| 10 | Univariate distributions for categorical features | 13-14 |
| 11 | Rainfall based on year and month | 15 |
| 12 | Rainfall based on month and day | 15 |
| 13 | Rainfall based on location | 16 |
| 14 | Visualization of Cities | 17 |
| 15 | Average rainfall in different cities | 17 |
| 16 | Relationship between different variables | 18 |
| 17 | Wind direction at different time spam | 18 |
| 18 | Average rainfall based on wind direction | 19 |
| 19 | Pair plot | 19 |
| 20 | Heat map | 20 |
| 21 | Outlier detection | 21 |
| 22 | Power transformation | 22 |
| 23 | After transformation | 23 |
| 24 | Null value imputation | 23 |
| 25 | Iterative Imputer | 23 |
| 26 | Describing data after imputation | 24 |
| 27 | Statistics Significance | 25 |
| 28 | ML approach | 26-27 |
| 29 | Logistic Regression | 27-28 |
| 30 | Vif | 29 |
| 31 | Applying SMOTE | 30 |
| 32 | Building model After treating imbalance | 31 |
| 33 | Recursive Feature Elimination | 32-33 |
| 34 | Random Forest Model | 33-34 |
| 35 | Hyper parameter tunning Trail 1 | 35 |
| 36 | Hyper parameter tunning Trail 5 | 36 |
| 37 | Boosting Models | 37 |
| 38 | Hyper tunning LightGBM | 39 |
| 39 | Model Comparison | 39 |
| 40 | Model stability using KFold | 40 |
| 41 | Variation of accuracy for light GBM using CV | 41 |

**Industry Review**

Rainfall prediction is significant not only on the micro but also on the macro level. The study is of significance with respect to its vital contribution in the field of agriculture, water reserve management, flood prediction and management with an intention to ease the people by keeping them updated with the weather and rainfall prediction. It is also important to be utilized by the agricultural industries for keeping their crops safe and ensure the production of seasonal fruits and vegetables by updated rainfall prediction. The study will also be significant for the flood management authorities as more precise and accurate prediction for heavy monsoon rains will keep the authorities alert and focused for an upcoming event that of which the destruction could be minimized by taking precautionary measures. The rainfall prediction will impressively help in dealing with the increasing issue of water resource management; as water is a scarce resource and it needs to get saved for the benefit of human beings themselves. Also, it will help the people to manage and plan their social activities accordingly.

**Literature Review**

Rain fall prediction using machine learning algorithms is the common among researchers some of them are:

Rainfall prediction is not an easy job especially when expecting the accurate and precise digits for predicting the rain. The rainfall prediction is commonly used to protect the agriculture and production of seasonal fruits and vegetables and to sustain their production and quality in relation to the amount of rain required by them (Lima & Guedes, 2015).

The rainfall prediction uses several networks and algorithms and obtains the data to be given to the agriculture and production departments. The rainfall prediction is necessary and mandatory especially in the areas where there is heavy rainfall and it’s more often expected (Amoo & Dzwairo, 2016).

The prediction of seasonal rainfall on monthly basis by using the surface data to form annual prediction is also essential for the agricultural activities and therefore the production and supervision of the agriculture and crops. It could be done by recognizing the variations in the supply of moisture in the air. The case of African region illustrates that how this succeeded and how West Africa advantaged from the rainfall prediction in managing their agricultural activities (Omotosho, Balogun, & Ogunjobi, 2000).

The rainfall forecasting is prevailing as a popular research in the scientific areas in the modern world of technology and innovation; as it has a huge impact on just the human life but the economies and the living beings as a whole. Rainfall prediction with several Neural Networks has been analysed previously and the researchers are still trying hard to achieve the more perfect and accurate results in the field of rainfall prediction (Biswas, et al., 2016).

There are huge economies like those of Asia like India and China that that earn a large proportion of their revenue from agriculture and for these economies; rainfall prediction is actually very important (Darji, Dabhi, & Prajapati, 2015).

**Data set and Domain**

**Data Dictionary**

The Dataset which we choose for our Capstone Project is “Rain Prediction in Australia” belongs to the domain Weather Forecasting.

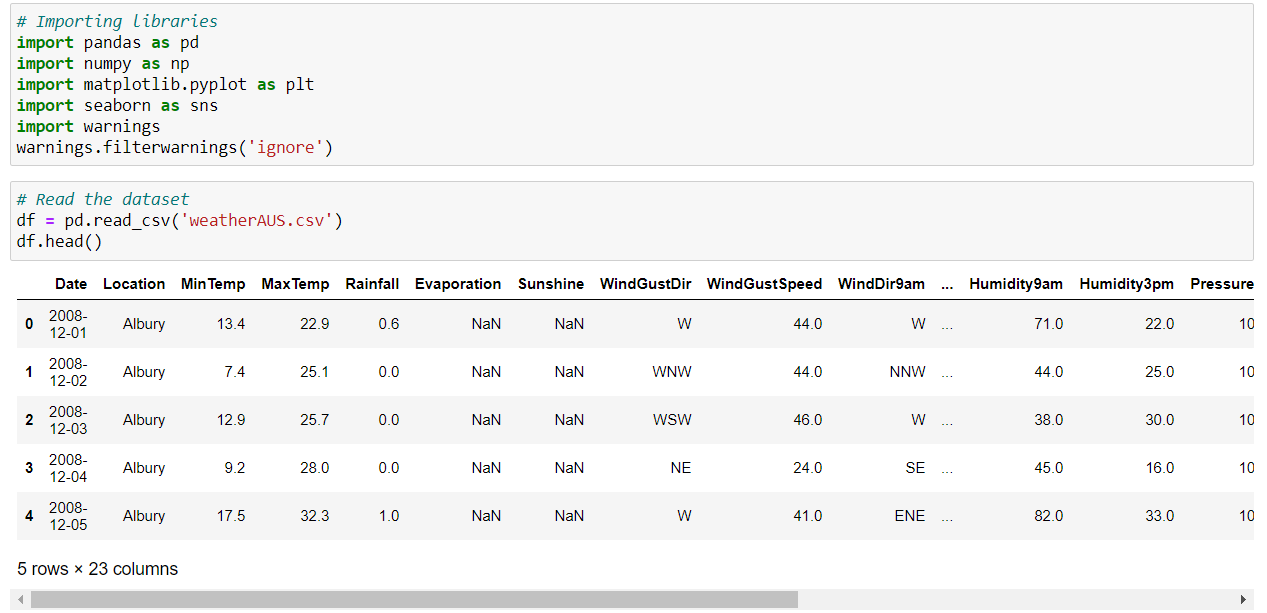


fig (1)

|  |  |  |
| --- | --- | --- |
| Date | Particular day of month Shown in numbers | Object |
| Location | Particular location in Australia | Object |
| Min Temp | Minimum temperature is a physical quantity that measure the coldness in Celsius | Float 64 |
| Max Temp | Maximum temperature is a physical quantity that measures the hotness in Celsius | Float 64 |
| Rainfall | Amount of Rainfall at that particular day | Float 64 |
| Evaporation | The process of turning water into vapor | Float 64 |
| Sunshine | The warmth and light given by sun rays | Float 64 |
| WindGustDir | Wind Gust Direction | Object |
| WindGustSpeed | Wind Gust Speed | Float 64 |
| WindDir9am | Wind Direction at 9 am | Object |
| WindDir3pn | Wind Direction at 3 pm | Object |
| WindSpeed9am | Wind Speed at 9 am | Float 64 |
| WindSpeed3pm | Wind Speed at 3pm | Float 64 |
| Humidity9am | Humidity at 9 am | Float 64 |
| Humidity3pm | Humidity at 3 pm | Float 64 |
| Pressure9am | Pressure at 9 am | Float 64 |
| Pressure3pm | Pressure at 3 pm | Float 64 |
| Cloud9am | Cloud at 9 am | Float 64 |
| Cloud3pm | Cloud at 3 pm | Float 64 |
| Temp9am | Temperature at 9 am | Float 64 |
| Temp3pm | Temperature at 3 pm | Float 64 |
| RainToday | Whether rain will occur today | Object |
| RainTommorow | Whether rain will occur tommorrow | Object |

Firstly, we are importing required libraries.

We are reading the dataset ‘WeatherAUS’ and displaying all the first five rows and columns of the dataset using head()

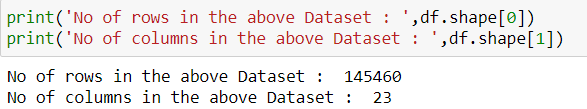
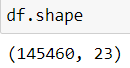


fig (2)

**Variable Categorization**

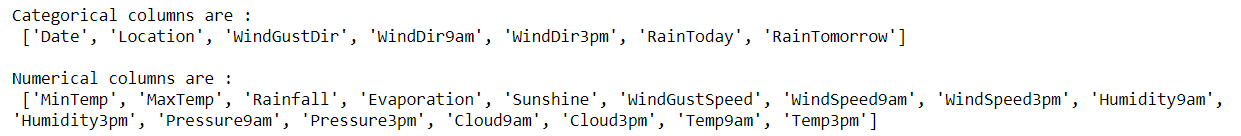


fig (3)

We found that the shape of the dataset is (145460,23)

No of rows in the dataset is 145460 and No of columns in the dataset are 23

and also, we found that there are categorical as well as numerical columns in the dataset with count of 7 Categorical columns and 16 Numerical columns and our target column is rain tomorrow

Let’s have a look at the distribution of the numerical as well as categorical features in the dataset

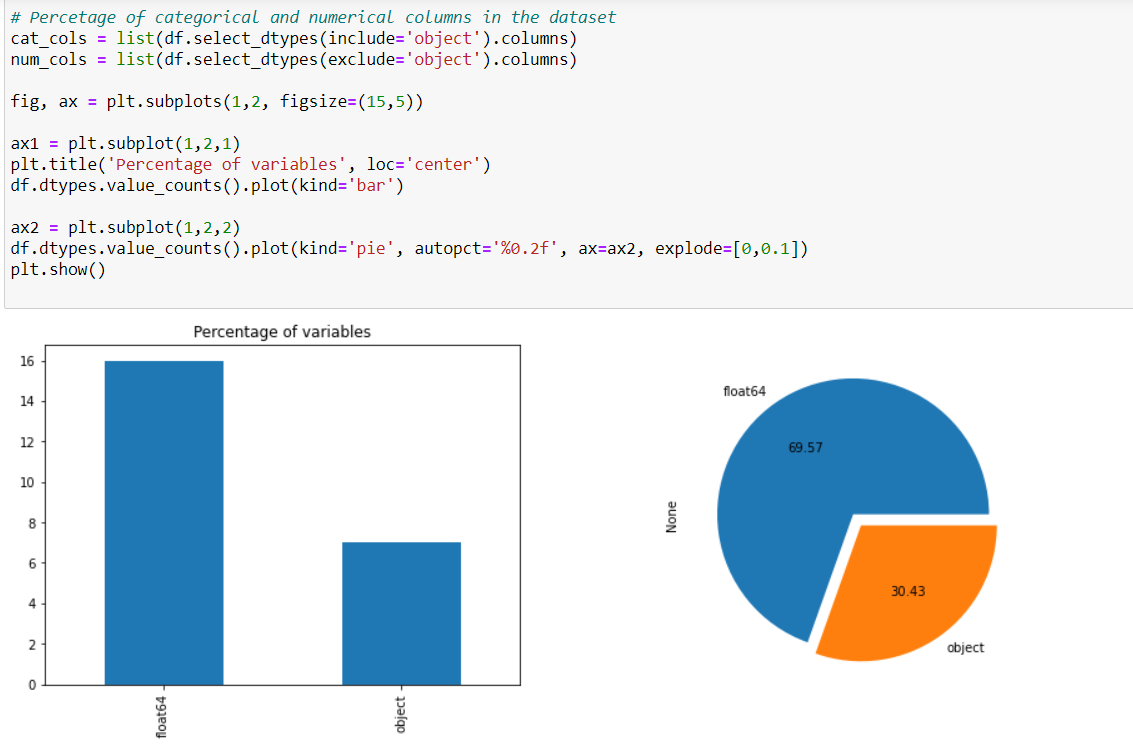


fig (4)

A Bar graph that represents the data in the rectangular bars with heights and lengths proportional to the values that they represent. Pie chart is a circular statistical graph both divides the whole data into slices to illustrate numerical proportion. we can observe that about 70% of the distribution is numerical and only 30% of the distribution is categorical data in the Rainfall Prediction Dataset.

**Pre-Process Data Analysis**

Now, we know the shape of the data and also, regarding type of the features that are present in the data. let’s get into the information regarding the data

There is function called info() in pandas which gives us the concise summary regarding the dataset. It includes the index dtypes, column dtypes, null values and memory usage.

From the below fig (5) we get to know that,

* There are null values/missing values present in the dataset
* We have range of index from 0 to 145459
* Total memory usage is 25.5+ MB
* There are 16 floating datatypes and 7 object.i.e., categorical

By using to-datetime method helps us to convert to object date time into python date time object. Similarly, using to-datetime we have converted our date feature into day, month and year.

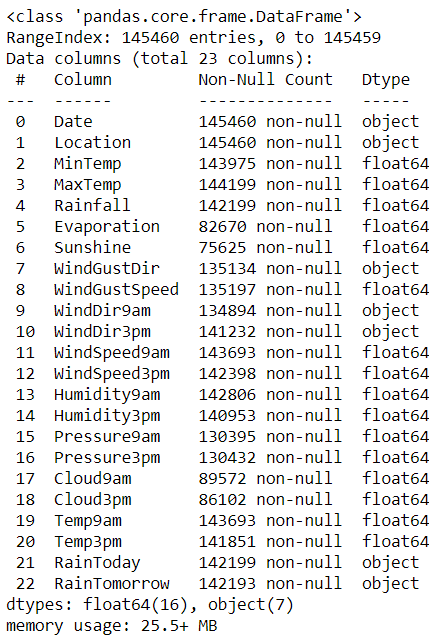


fig (5)

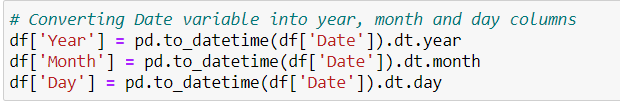


fig (6)

**Data Describe**

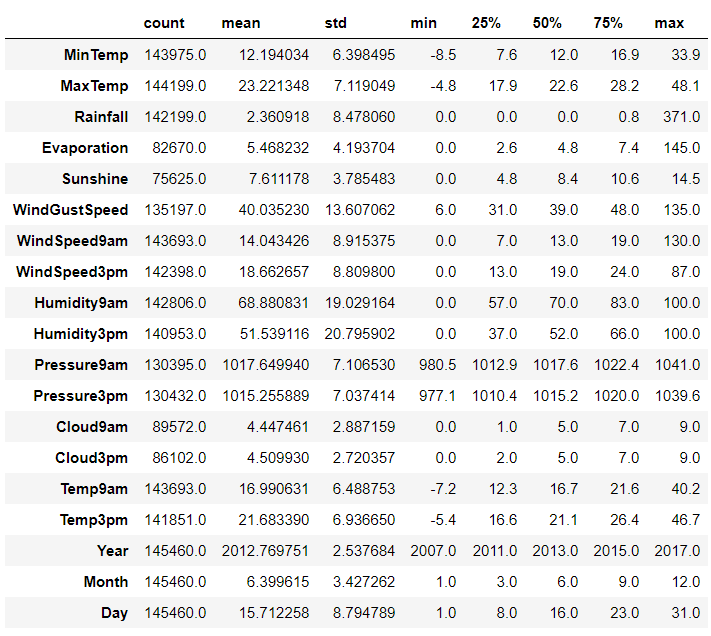


fig (7)

* There are Outliers in the features
* Through five-point summary we get to know the count, mean, std, min, max values
* for min temp feature we have min value as -8.5 and max is of 33.9
* We get to know that there are missing values by looking at the count column

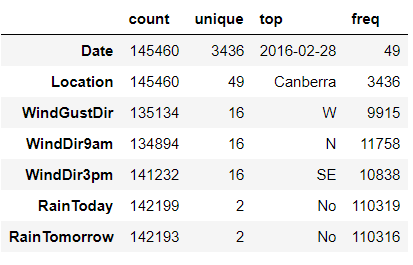


fig (8)

**Project Justification**

**Problem Statement:** -

Hurricanes, Earthquakes, Rainstorm, Flood, Avalanches, Wildfire these events have being happening from millennia and have affected humans throughout every part of the globe. Disaster is natural or man-made event that negatively affects life, property, livelihood and industry often resulting in permanent changes to human societies, ecosystems and environments.

Rainstorm is one of the major issues that should be concentrated as it is highly related with landslides, earthquakes and many more. Having an appropriate approach for rain fall prediction makes it possible to take preventive measures.

Rainfall prediction is about predicting the rain. It is a part of weather forecasting system which is very important in knowing the natural changes of our environment. It has a wide range of applications and serious concerns on the agricultural activity, forestry, power generation, construction, manufacturing units, travel and tourism etc.

Because of the concern, this has successfully gained the attention of Government, Industries, risk management entities, investors and scientific community to work on it and to get maximum benefits out of it.

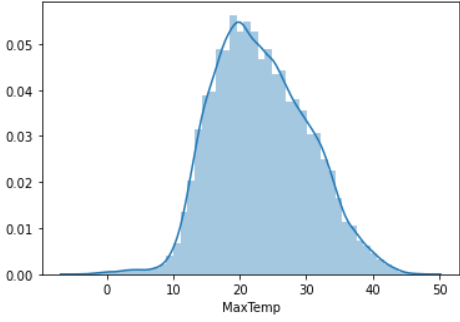
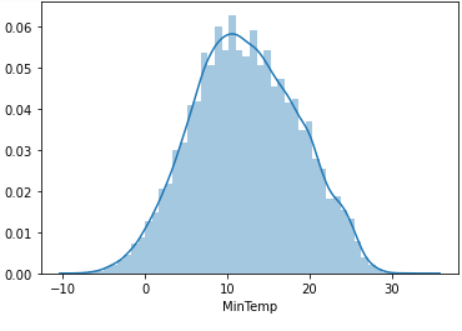
**Data Exploration (EDA)**

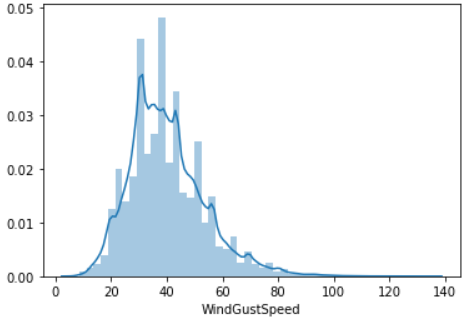
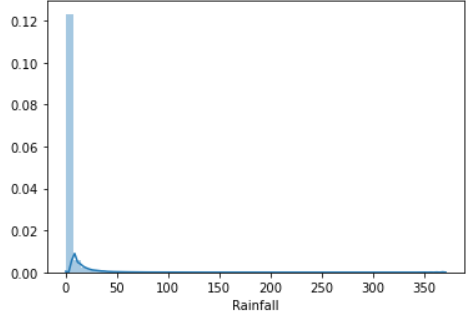
Relation between variables

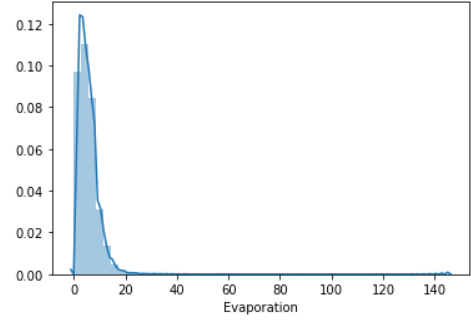
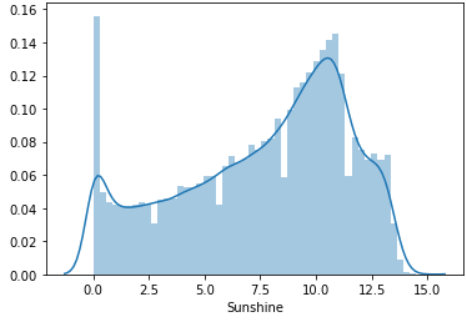
Exploratory Data Analysis is majorly performed using the following methods: -

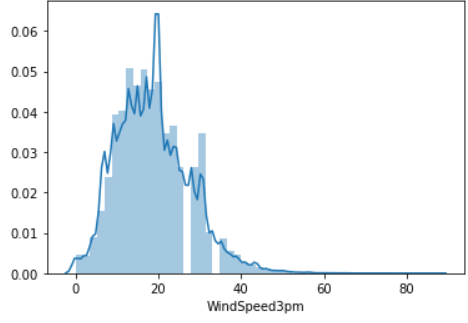
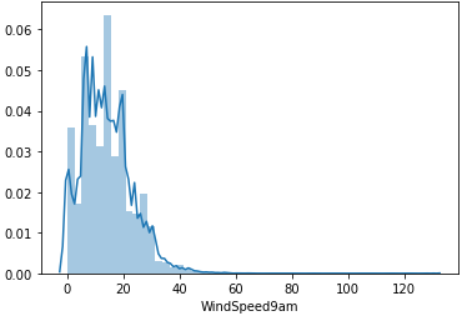
1.Univariate Analysis- provides summary statistics for each field in the raw data set (or) summary only on one variable.

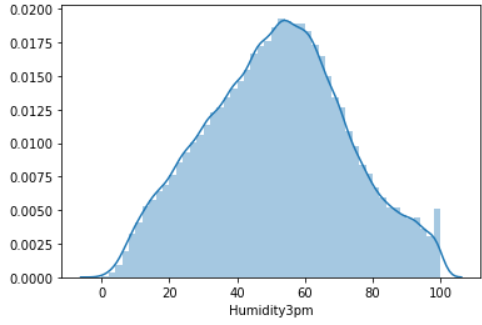
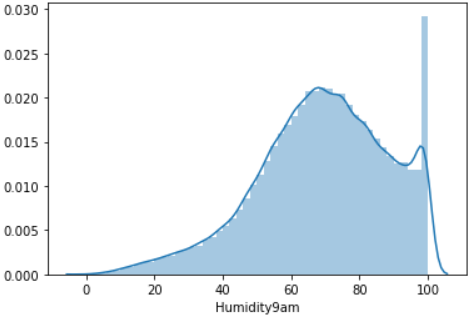
**Univariate Analysis for numerical features**

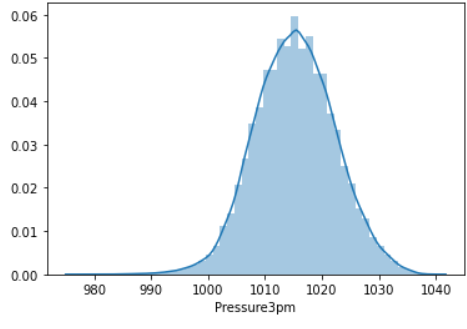
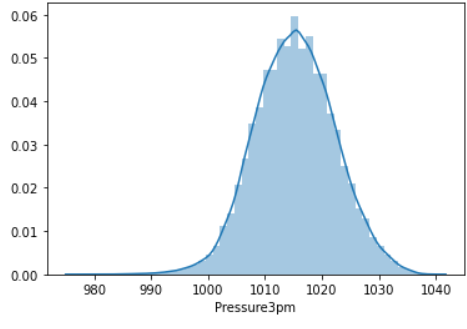


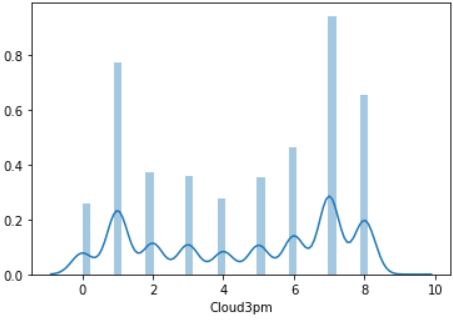
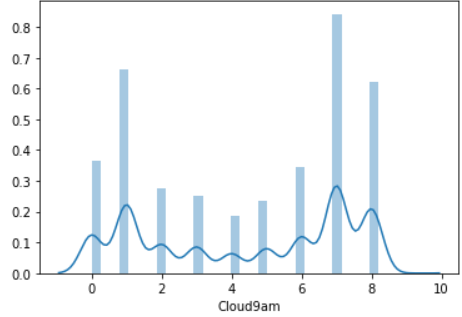












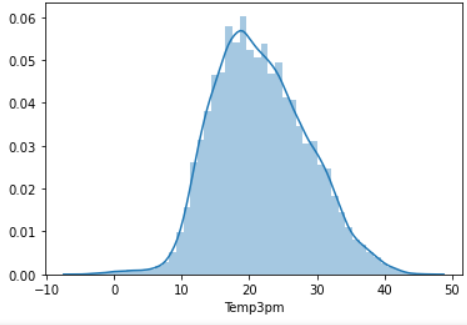
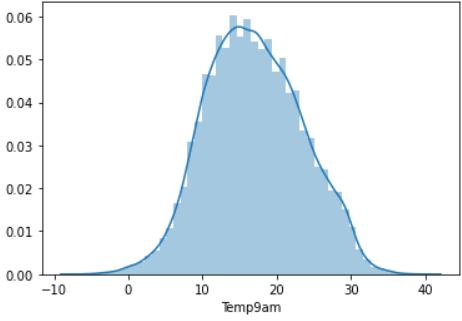
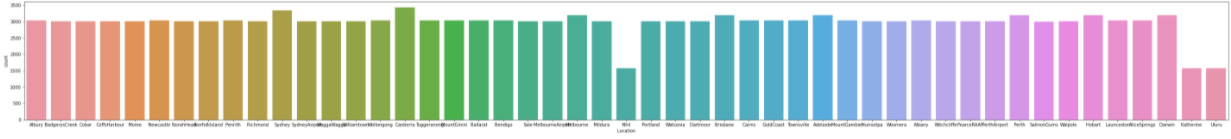
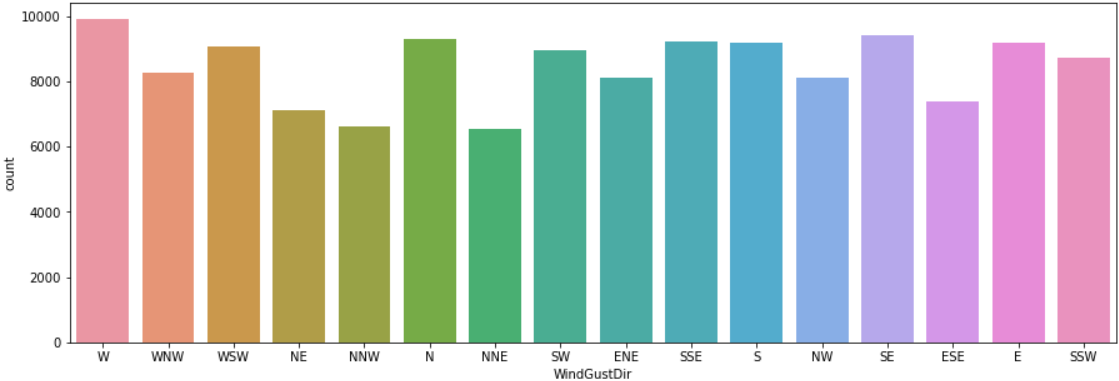


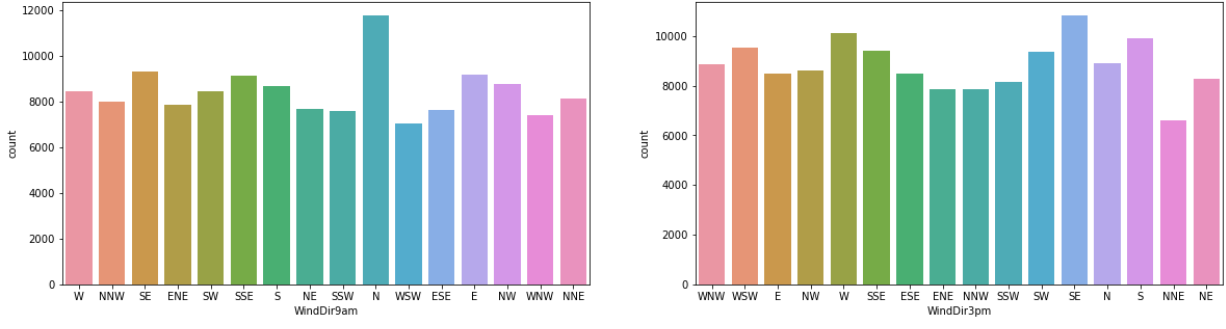
fig (9)

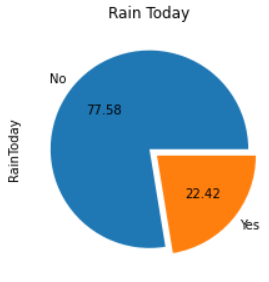
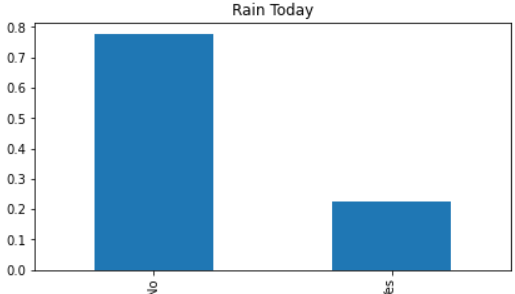
* Mintemp, Maxtemp, pressure3pm, pressure9am, Temp9am, Temp3am variables are almost normally distributed.
* Other variables do not follow Gaussian distribution in the data.
* We can also observe that in most of the days the rainfall is zero.
* In most of the days, minimum temperature is observed between 8-12degree and maximum temperature between 18-22degree.
* Many features are left skewed and many right skewed

**Univariate Analysis for Categorical Features**









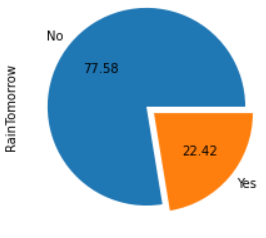
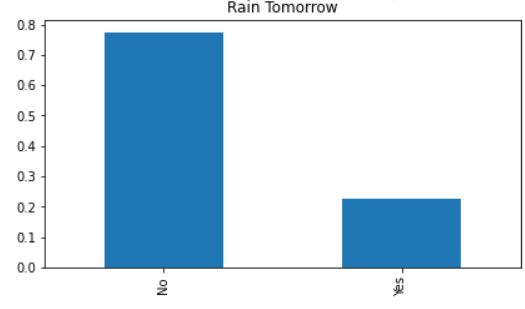


fig (10)

* There are 49 unique locations
* The number of observations for each location is almost same except Nihil, Katherine and Uluru locations.
* At 9AM, direction of wind is more observed in North direction.
* At 3PM, direction of wind is more observed in North-East direction.
* Both variables show data imbalance.
* Also chances of happening rain percentage is same for both today and tomorrow.

2.Bivariate Analysis- is performed to find the relationship between each variable in the dataset and the target variable of interest or using 2 variables and finding the relationship between them.

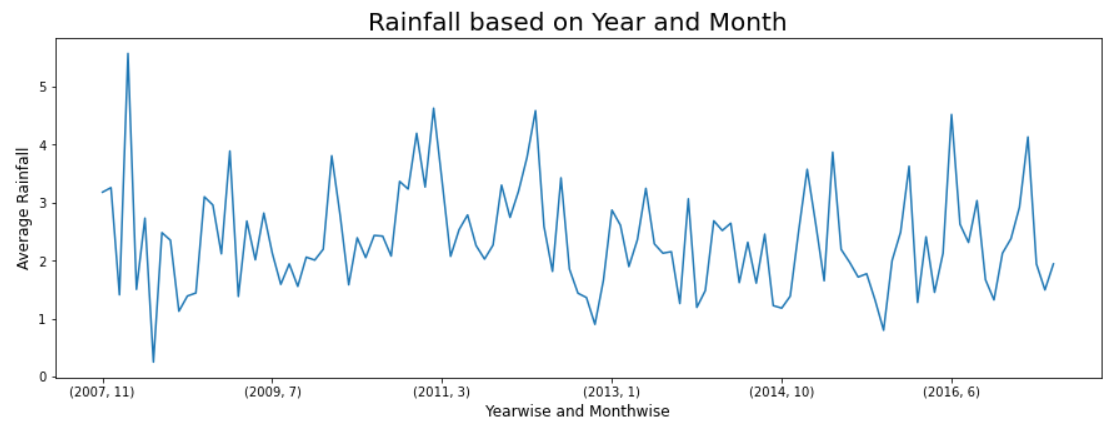


fig (11)

* Rainfall data is available for 10years
* Both maximum and minimum average rainfall is observed in 2008 for the months of Feb and April.

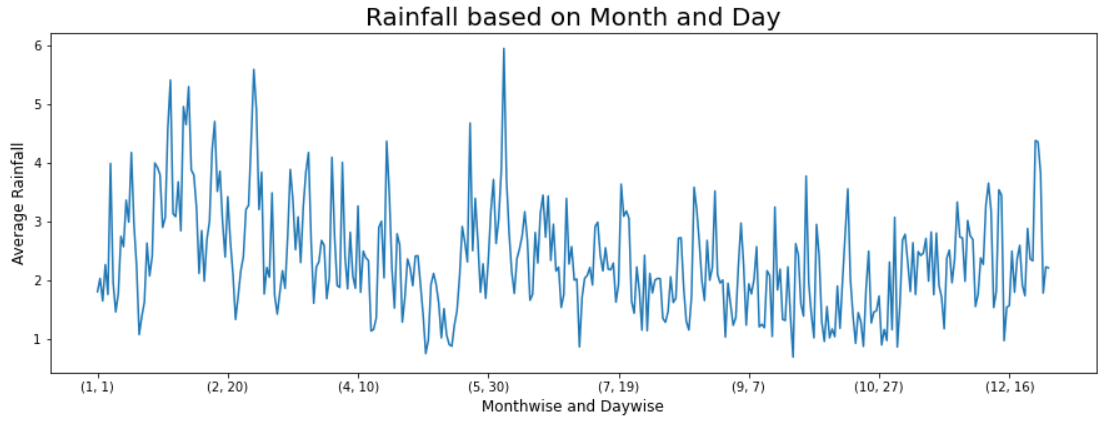
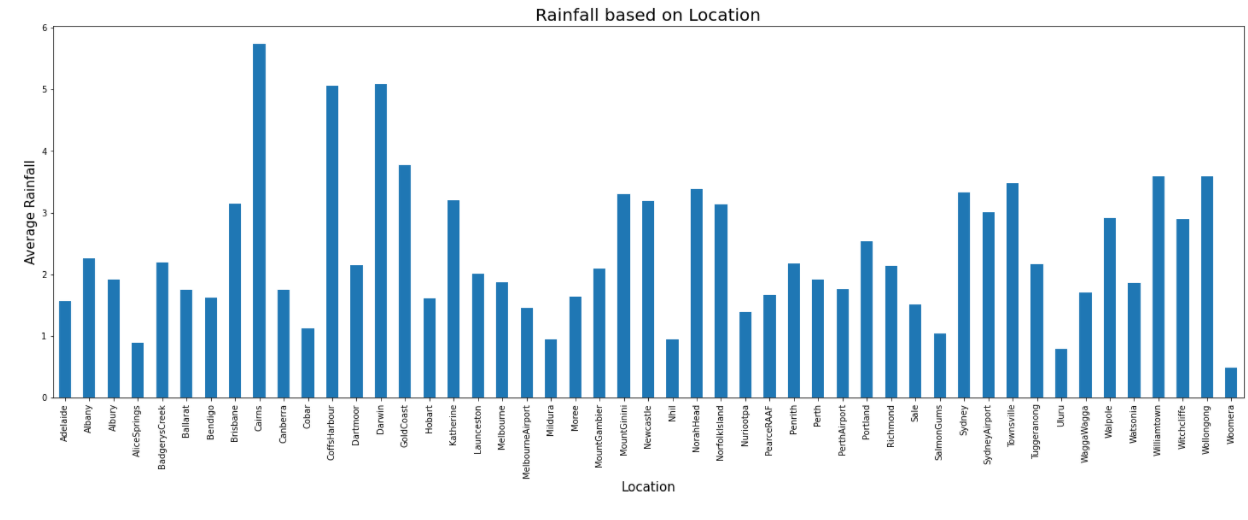


fig (12)

* Maximum rainfall is observed in month of June.
* Minimum rainfall is observed in month of September.



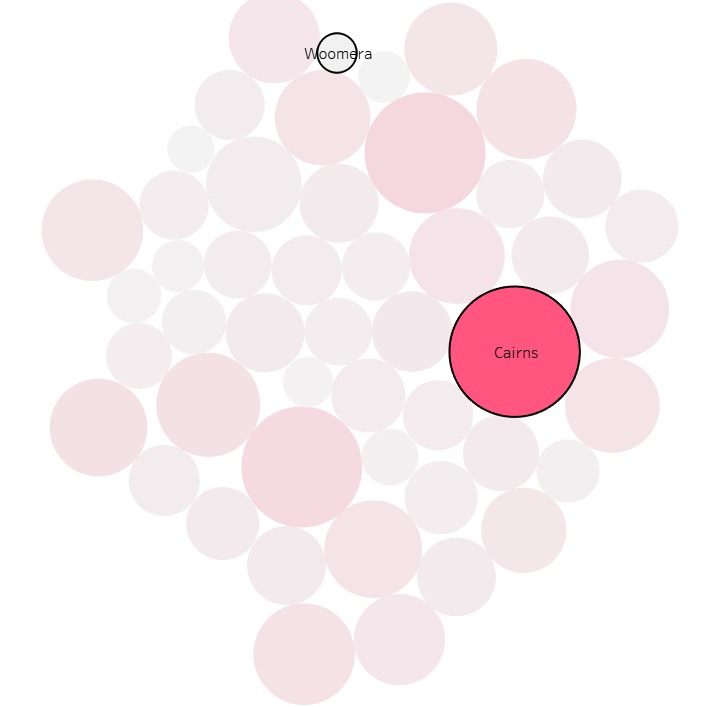
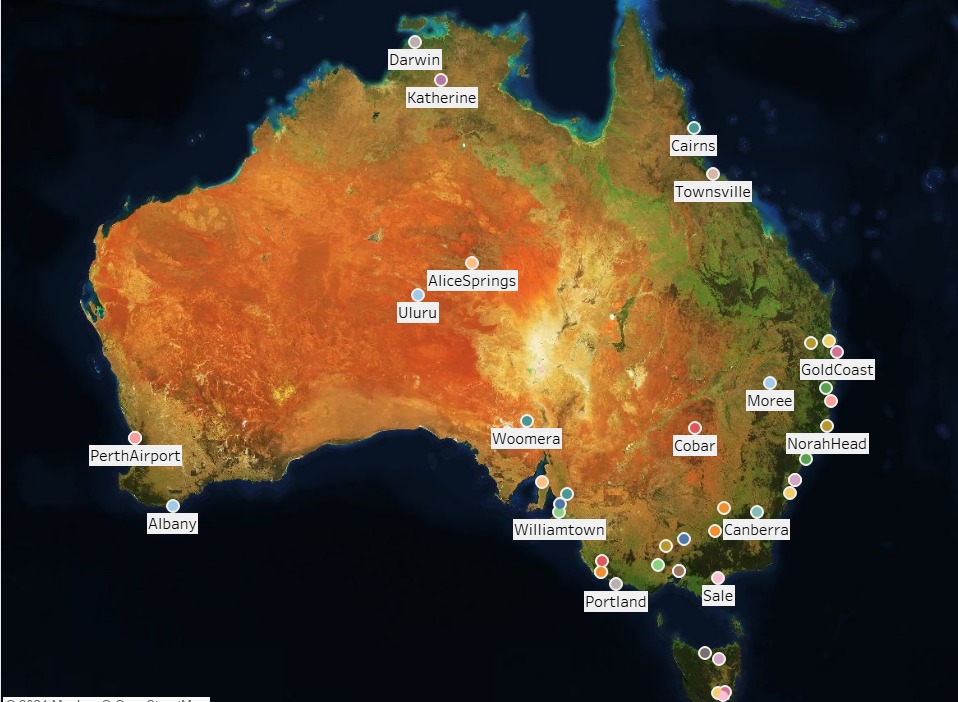
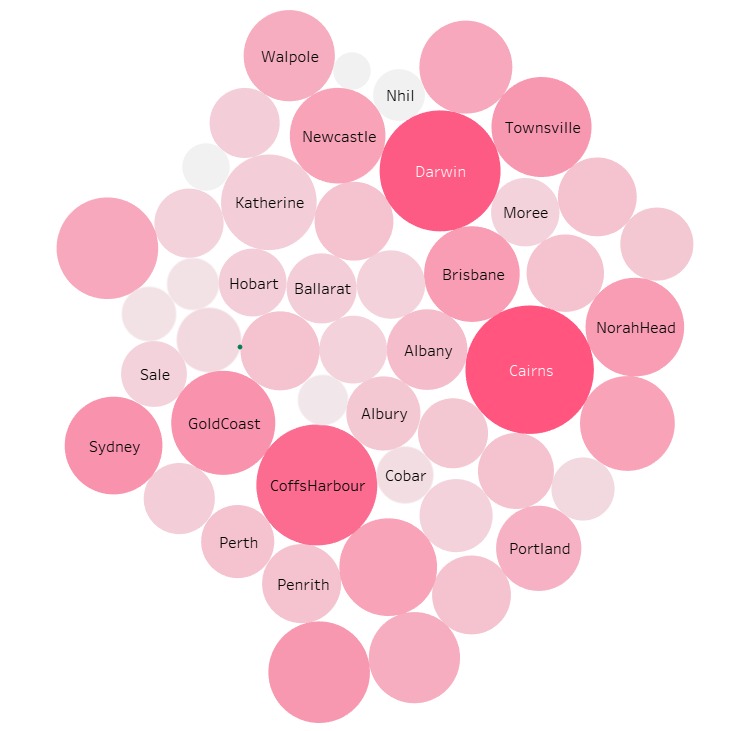


fig (13)

* Location Woomera have minimum rainfall
* Location Cairns have maximum rainfall.

 fig (14)

  
 fig (15)

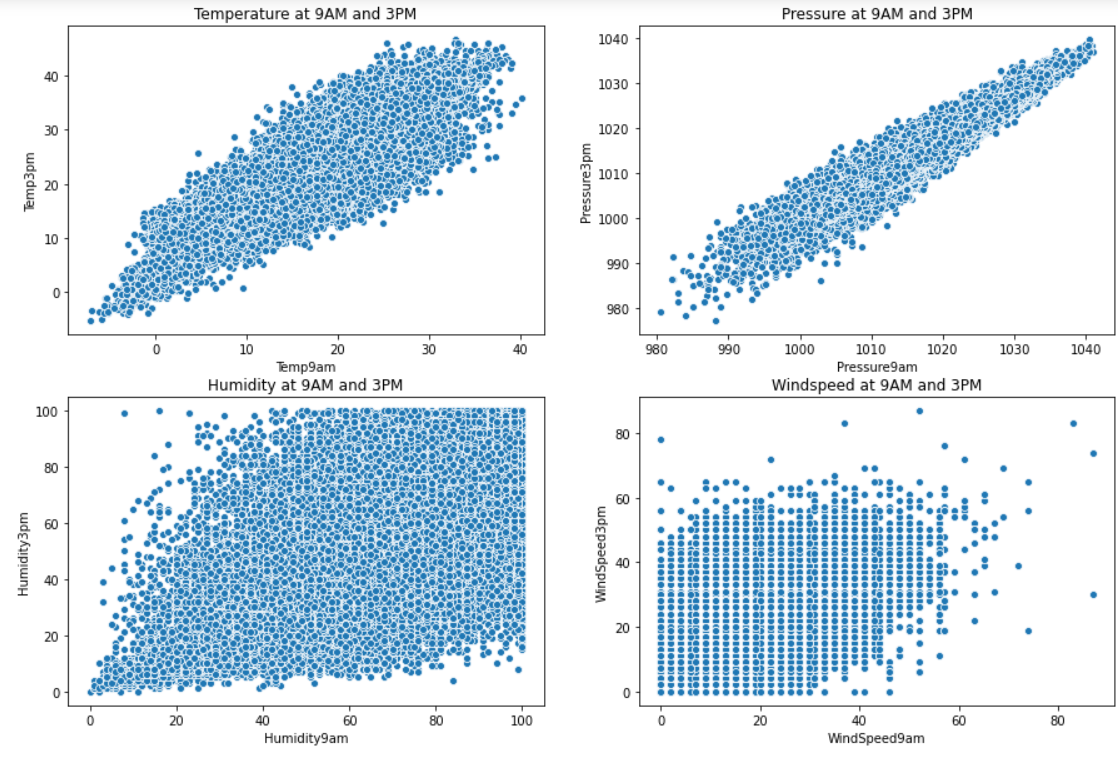


fig (16)

* Both Temperature and Pressure shows linear relationship at different timespan.
* For Humidity and Windspeed, scatter plot not showing linear relationship at different timespan.

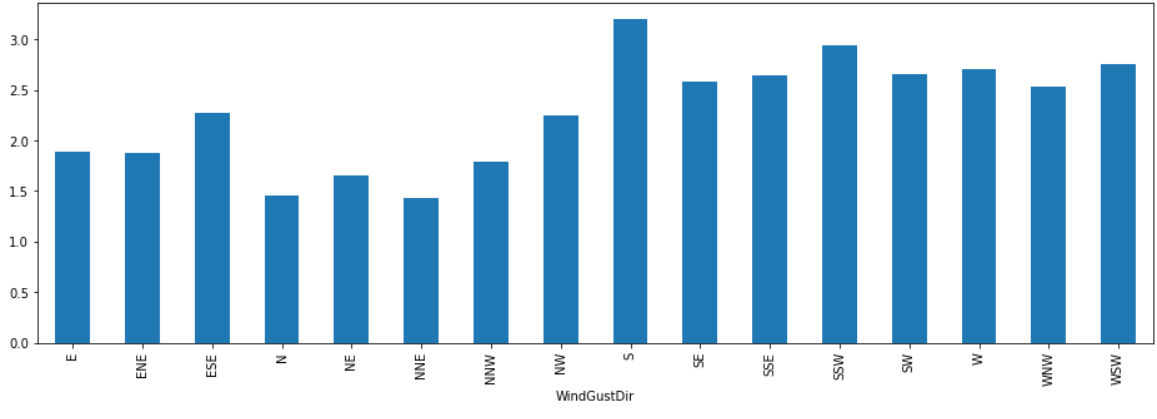


fig (17)

* In south direction, rainfall is more.

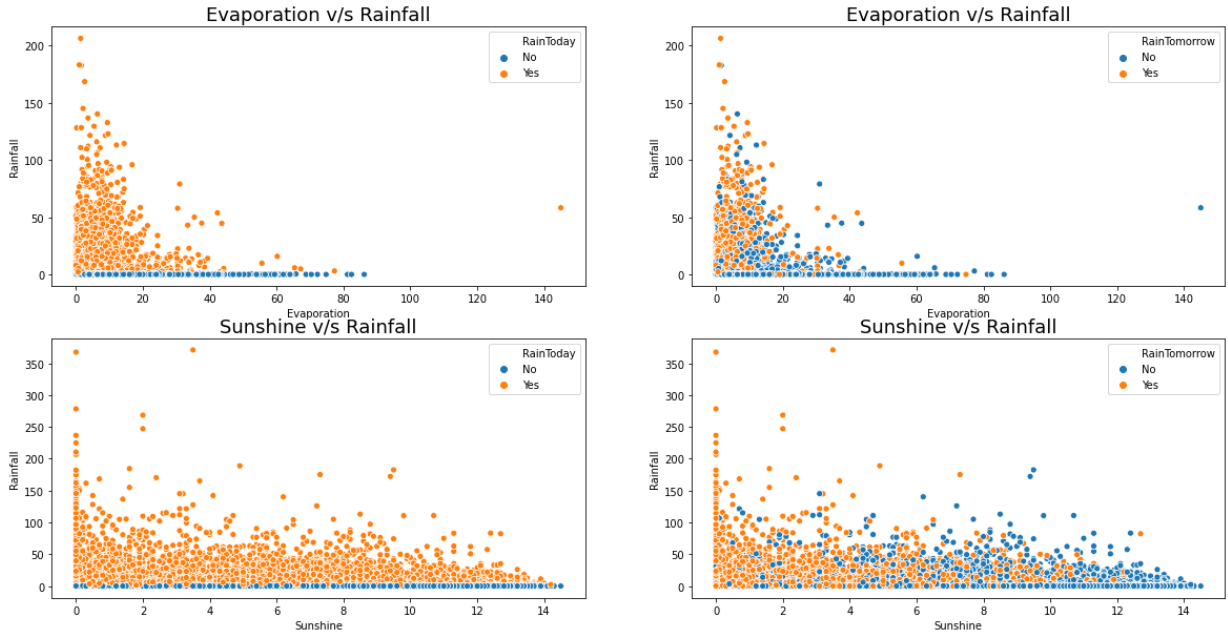


fig (18)

* As Evaporation increases, chances of rainfall happening is very less.
* As Sunshine increases, chances of rainfall are there but rainfall rate is less.

3.Multivariate Analysis- is performed to understand interactions between different fields in the dataset (or) finding interactions between variables more than 2.

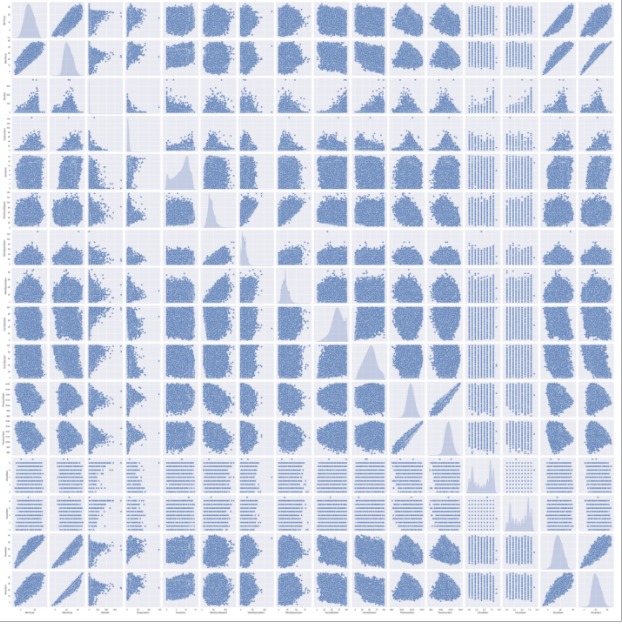


fig (19)

* Few variables are normally distributed and shows linear relationship with other variables.

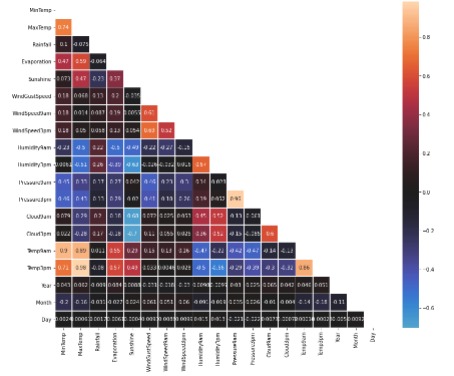


fig (20)

Few pairs of variables show strong correlation between them,

\* MinTemp and MaxTemp (r=0.74)

\* WindSpeed3pm and WindGustSpeed (r=0.69)

\* Temp9am and MinTemp (r=0.90)

\* Pressure3pm and Pressure9am (r=0.96)

\* Temp9am and MaxTemp (r=0.89)

\* Temp3pm and MinTemp (r=0.71)

\* Temp3pm and MaxTemp (r=0.98)

Above these variables shows multicollinearity each other.

Multicollinearity occurs when two or more independent variables are highly correlated with one another in a regression model. This means that an independent variable can be predicted from another independent variable in a regression model.

**Feature Engineering**

**Outliers Detection and Treatment**

Outlier is a [data point](https://en.wikipedia.org/wiki/Data_point) that differs significantly from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error. An outlier can cause serious problems in statistical analyses.

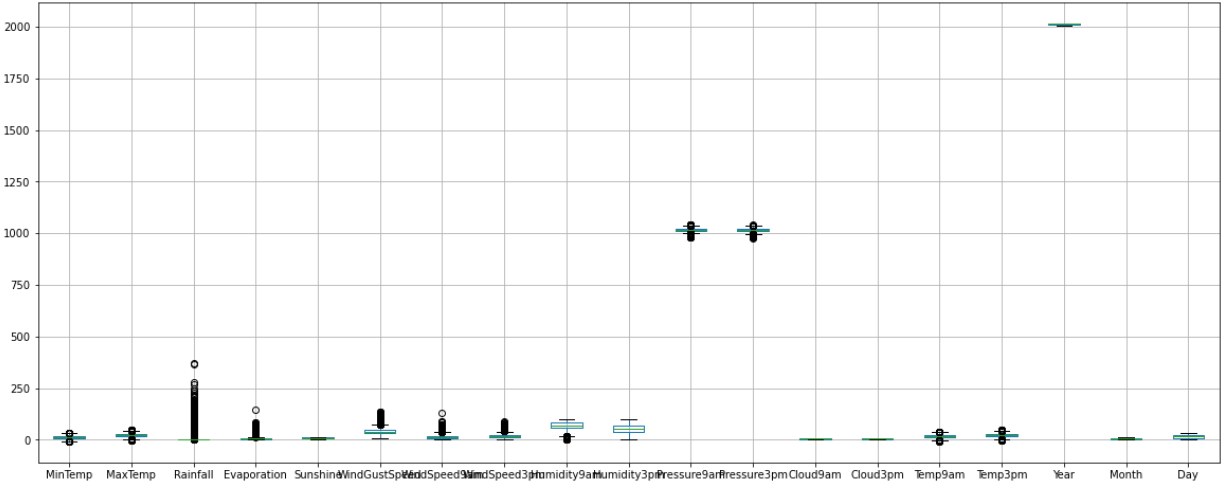


fig (21)

* Presence of outliers seen in all continuous variables except Sunshine, Humidity3pm, Cloud9am and Cloud3pm.
* These outliers are transformed using Power transformer instead of removing it.

**Transformation of the data**

Power transforms are a technique for transforming numerical input or output variables to have a Gaussian or more-Gaussian-like probability distribution.

When the normality of the data is denied to make it normal, we use power transform.

Since, we have converted the date column into date, month and year so we are dropping the date column.

* All the variables are converted into continuous data form.
* Before transforming data, remove all ordinal or label encoded data. Because power transformer won't have effect on it.
* Transformation techniques are used because removal outliers lead to data loss of around 38.8%.

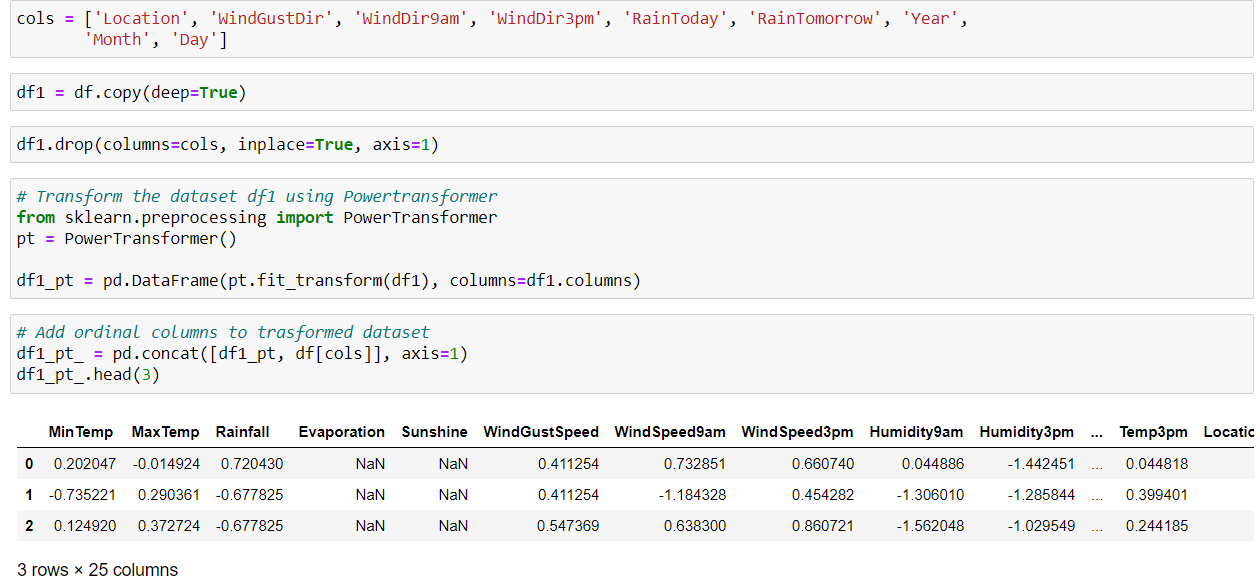


fig (22)

After power transform

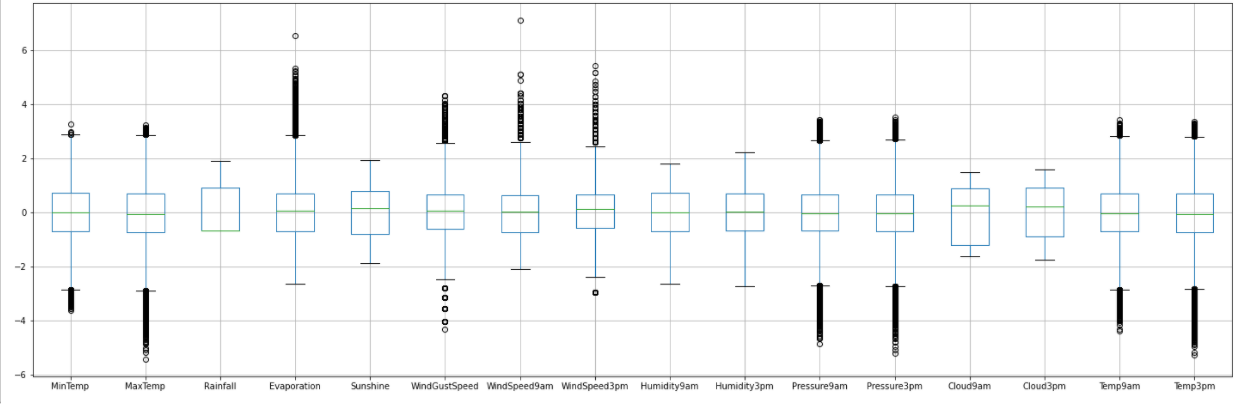


fig (23)

* The data is better than previous model
* We can notice the transformation of the data

**Null value Imputation**

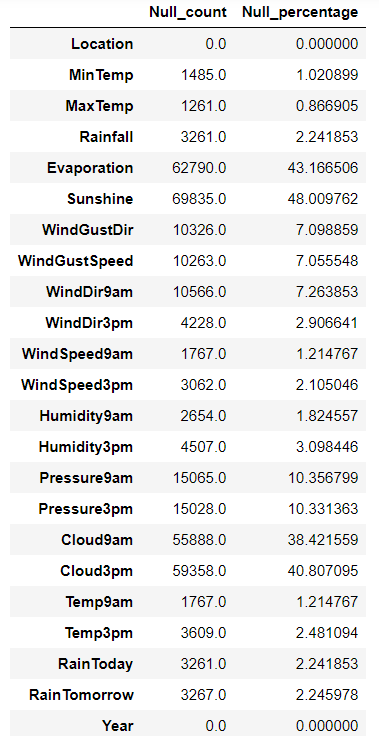


fig (24)

* Null values are imputed using Iterative imputer method.658
* Rows are removed if target column have null values in it.

**Iterative Imputer Method**

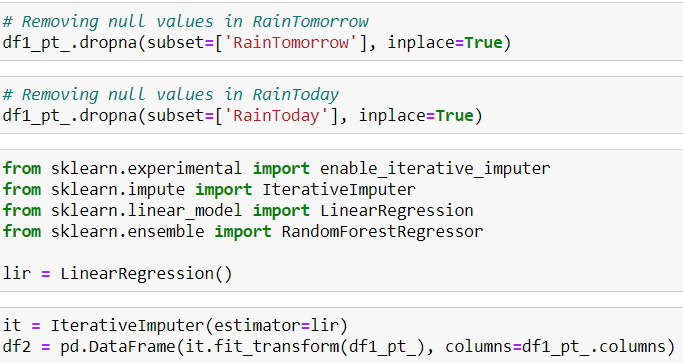


fig (25)

Iterative imputation refers to a process where each feature is modelled as a function of the other features. Each feature is imputed sequentially, one after the other, allowing prior imputed values to be used as part of a model in predicting subsequent features.

It is iterative because this process is repeated multiple times, allowing ever improved estimates of missing values to be calculated as missing values across all features are estimated.

Describing the Data after imputation

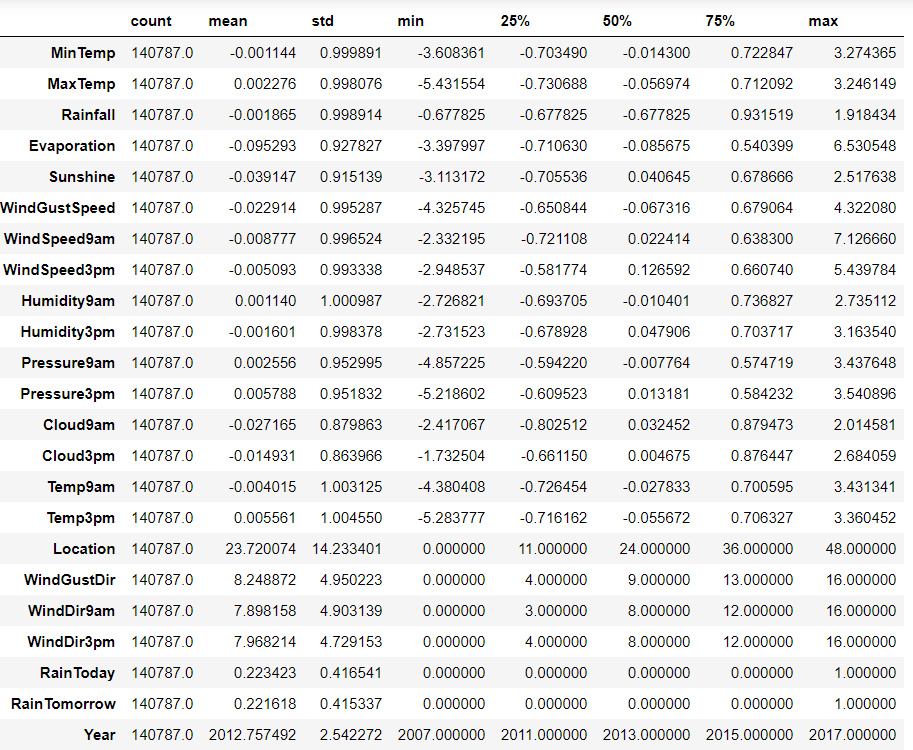


fig (26)

* After imputation, variation in the data is very minute or less.
* Hence Imputation using Linear regression algorithm is found to be satisfactory.

**Statistical Significance of variables**

The Shapiro test is the test for normality. The test rejects the hypothesis of normality when the p-value is less than or equal to 0.05. p**-**value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct. ... A smaller p-value means that there is stronger evidence in favour of the alternative hypothesis.



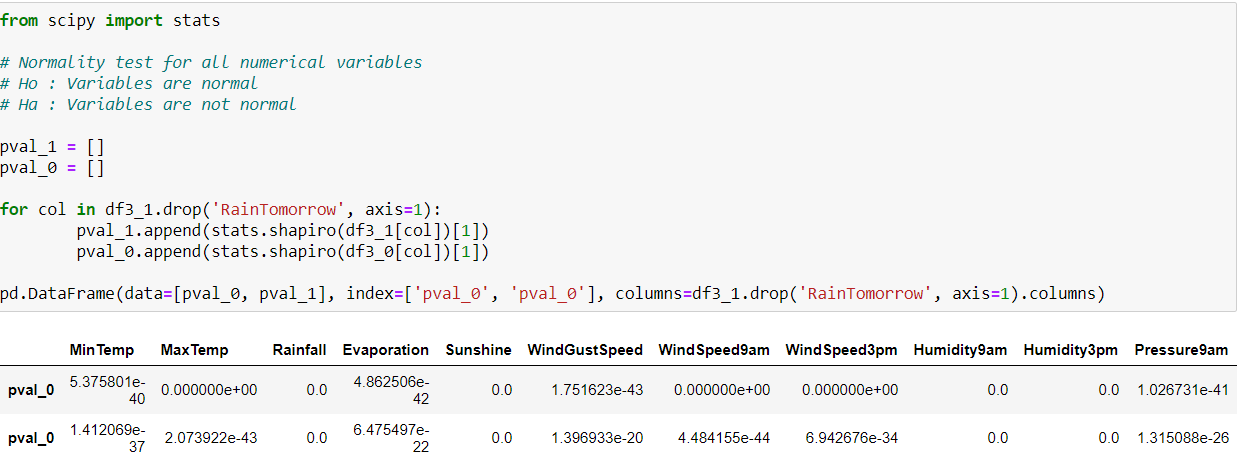
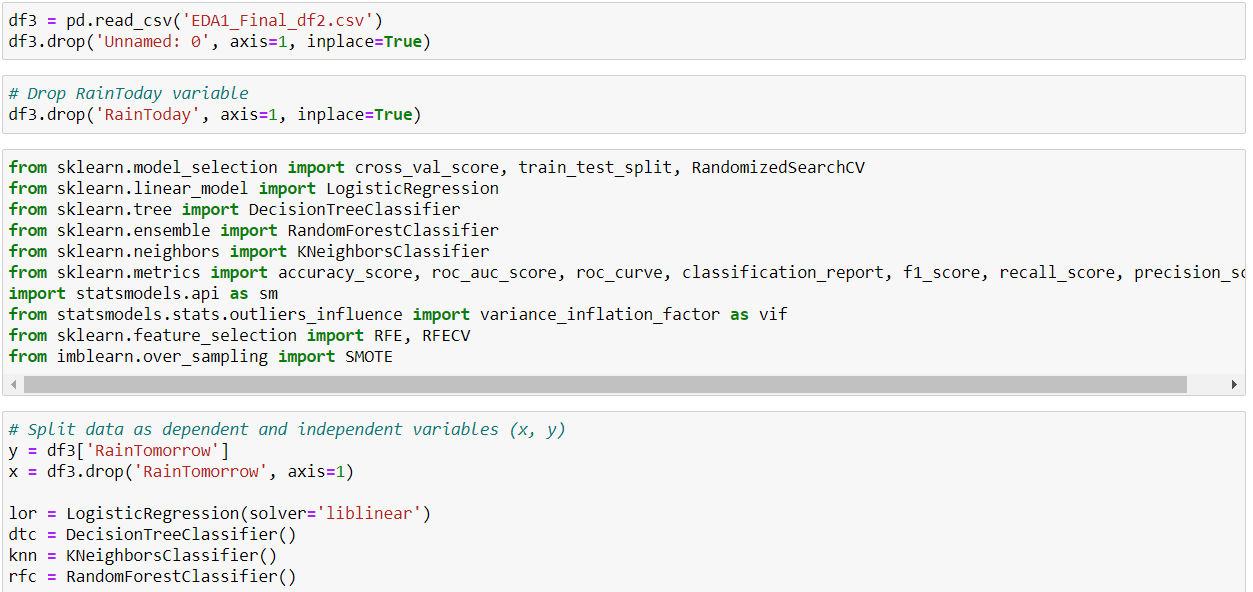


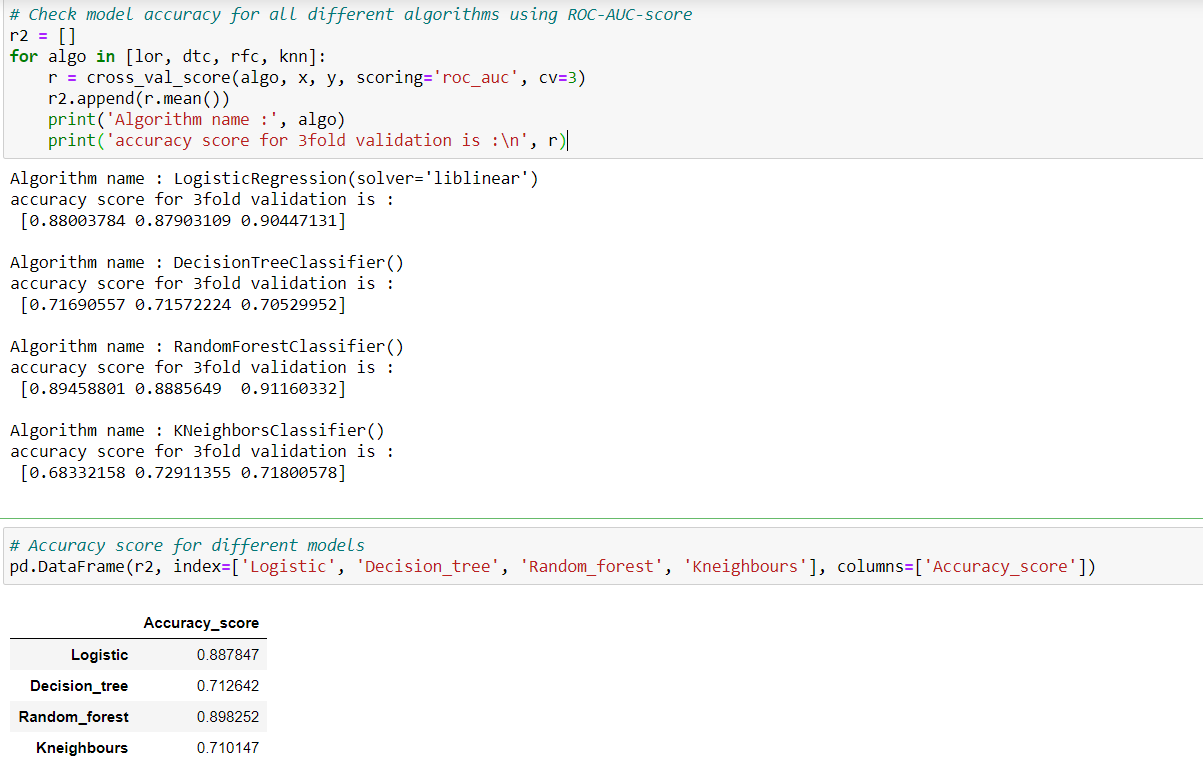


fig (27)

* For all the variables, p-value is less 0.05, hence reject null hypothesis.
* Hence all the numerical and categorical variables have relation with target variable (Rain Tomorrow).
* There are no insignificant variables are found using statistical analysis

**Approach**





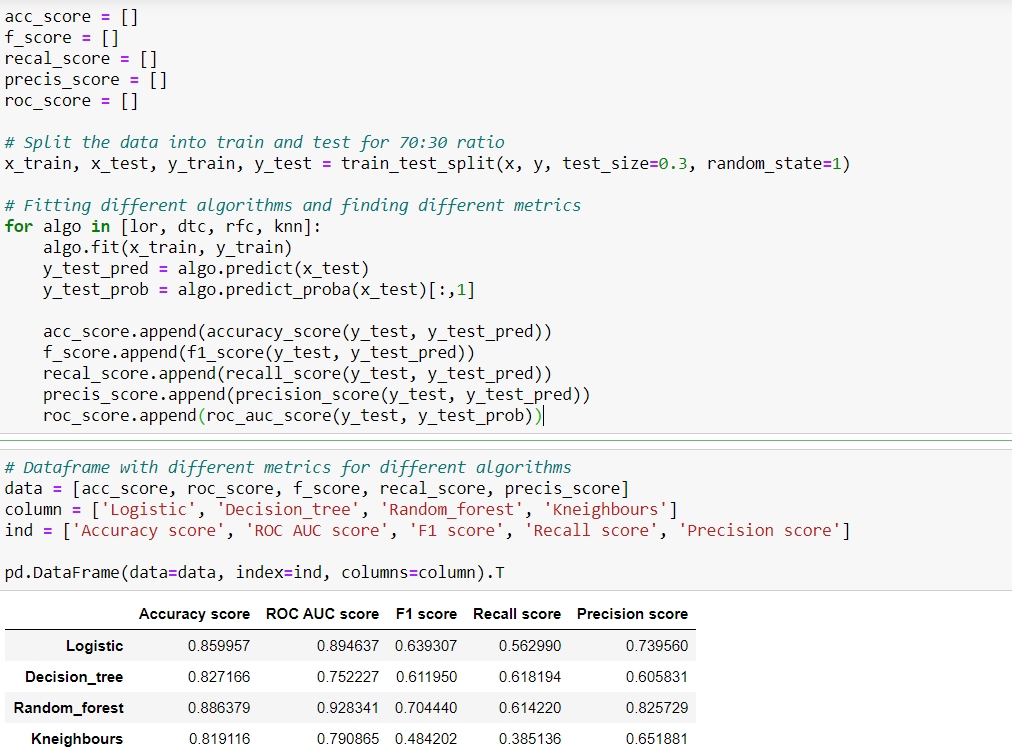
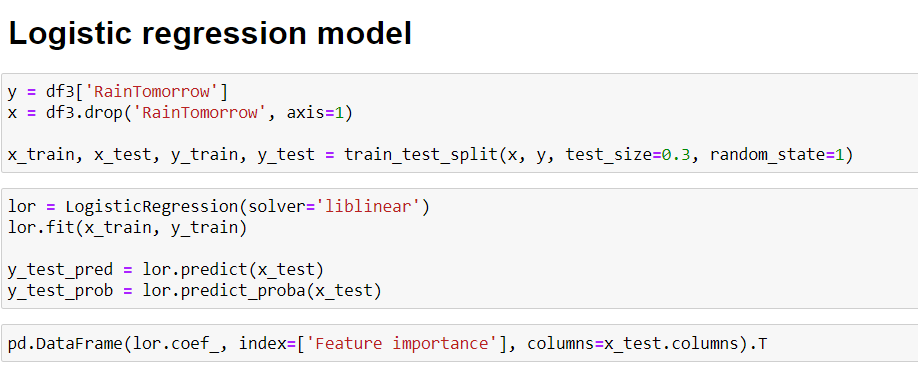
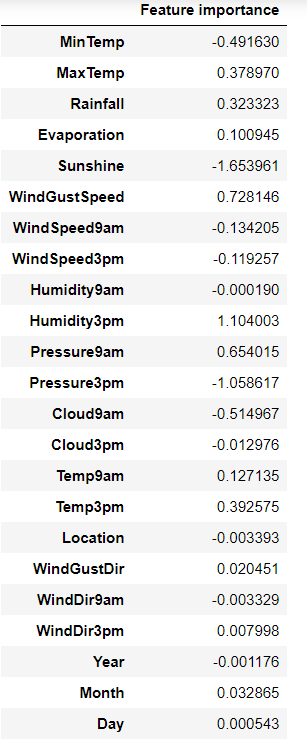


fig (28)

* From above 3fold validation method and with different metric comparison, both Logistic regression and Random forest is giving good results.
* Let’s consider Logistic regression as base model.





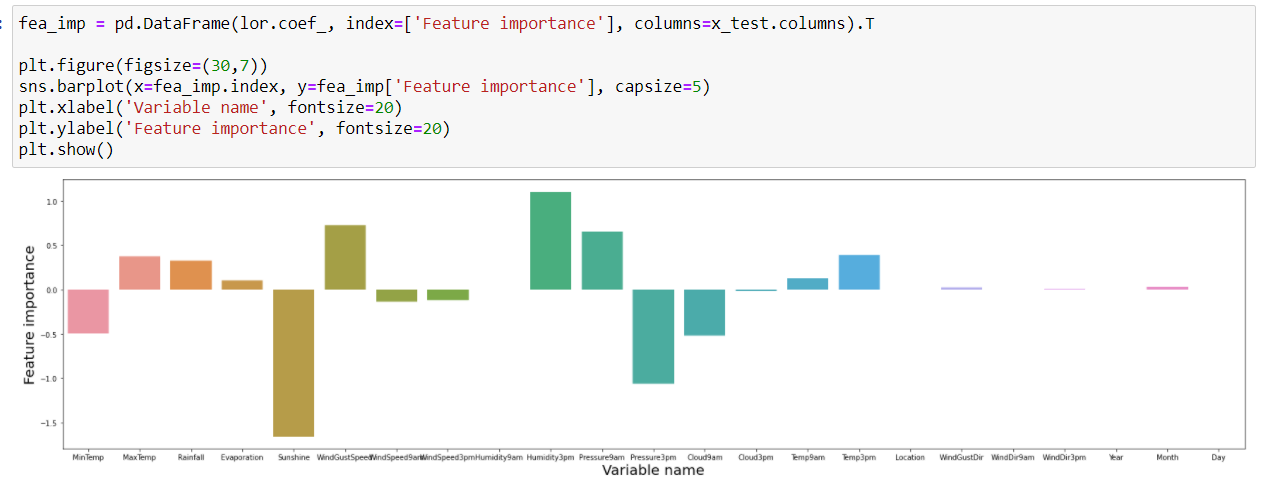




fig (29)

* Year, Month, WindDir9am, WindDir3pm, WindDir9am, Day and Location are categorical variables, which are not performing well for improving model accuracy.
* Variables such as Sunshine, Pressure3pm, Cloud9am, MinTemp shows strong negative correlation with target variable.
* Variables such as Rainfall, MaxTemp, Pressure9am, WindGustSpeed and Humidity3pm shows strong positive correlation with target variable.

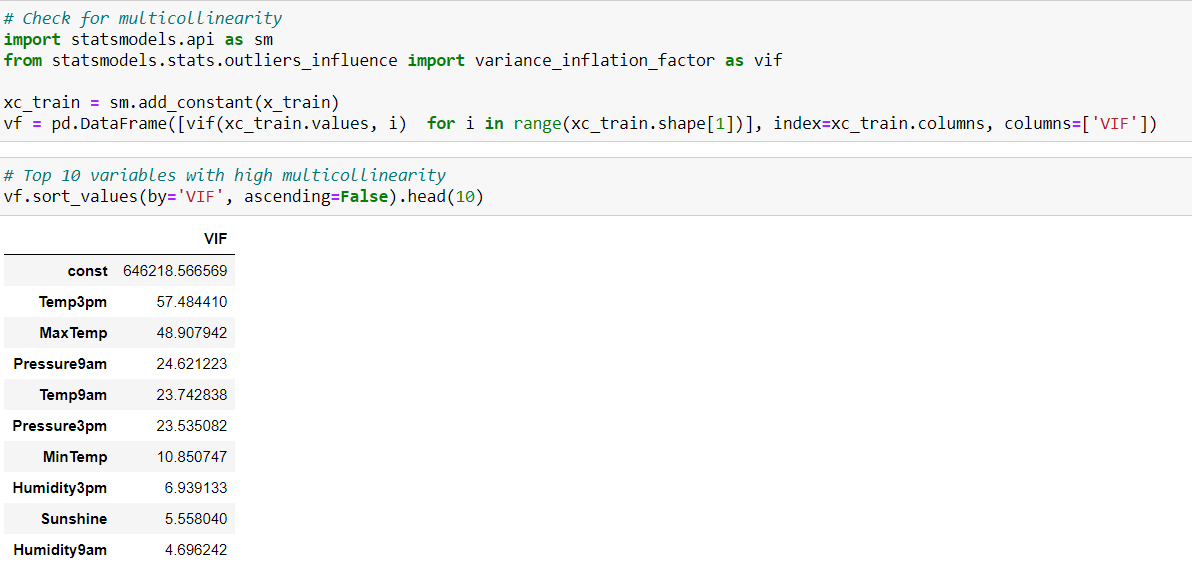


fig (30)

* Temp3pm, MaxTemp, Pressure9am, Temp9am, Pressure3pm and MinTemp variables have more multicollinearity, these can be removed.
* Same variables show high correlation in descriptive statistics; hence it confirms that these variables can be removed.

**Inference from above approach**

* Few common insignificant variables are seen in both EDA and ML approach (Based on VIF)

Example: - Temp3pm, MaxTemp, Pressure9am, Temp9am, Pressure3pm, MinTemp, WindSpeed9am, WindGustSpeed

* The categorical variables are not helping to predict rain.

Example: - Year, Month, WindDir9am, WindDir3pm, Day and Location

* The variables like Sunshine, Evaporation, Humidity3pm, Cloud3pm, Rainfall, Pressure3pm, Temp3pm are top features contributing towards prediction of rainfall.
* In ML approach, both Logistics regression and Random forest are giving good model result, we chosen Logistics regression as base model and inferences are on top of it.
* From EDA and Statistical approach, it confirms that variables are not normally distributed. It is resolved using Power transformation technique.
* Null values are imputed using multivariant imputation technique (Iterative Imputer with Linear regression model).
* Removal of outliers leads to huge data loss of 38.8%, hence transformation techniques are performed to resolve this issue.
* From statistical and ML approach, it confirms that most of variables have relation with target for rain prediction.
* Further class imbalance, feature elimination, hyper parameter tuning methods are performed to increases model accuracy as well as prediction rate.

**Improvisation of model based on the above insights**

**Class Imbalance Treatment using SMOTE**

In Machine Learning and Data Science we often come across a term called **Imbalanced Data Distribution**, generally happens when observations in one of the class are much higher or lower than the other classes

Imbalanced Data Handling Techniques: There are mainly 2 mainly algorithms that are widely used for handling imbalanced class distribution.

1. SMOTE
2. Near Miss Algorithm

SMOTE (Synthetic Minority Oversampling Technique) – Oversampling

SMOTE (synthetic minority oversampling technique) is one of the most commonly used oversampling methods to solve the imbalance problem.  
It aims to balance class distribution by randomly increasing minority class examples by replicating them.



fig (31)

* Class imbalance is treated using SMOTE technique and made it balance.

Building model after treating imbalance

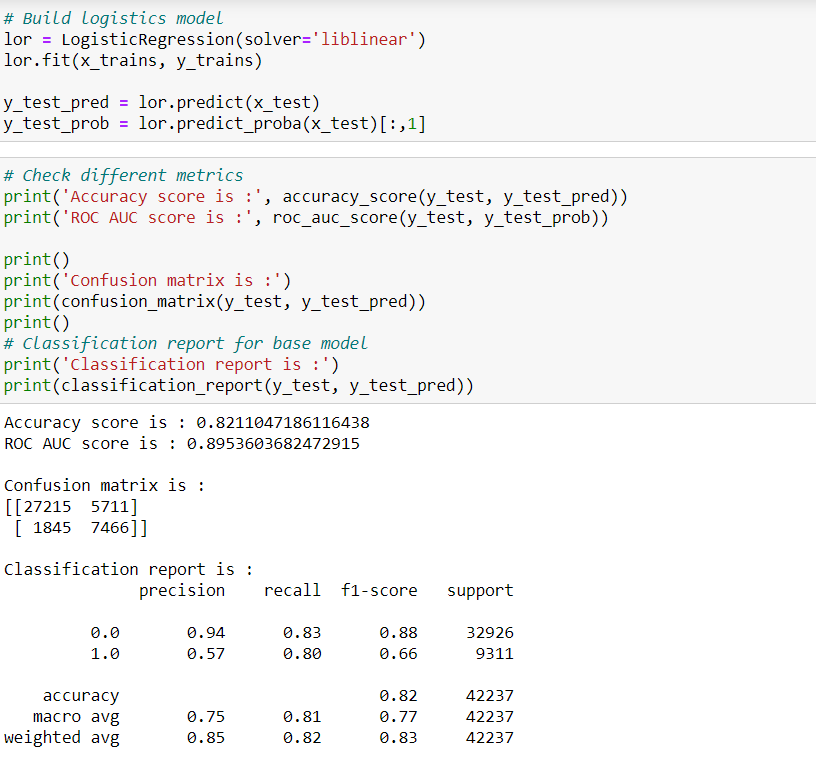


fig (32)

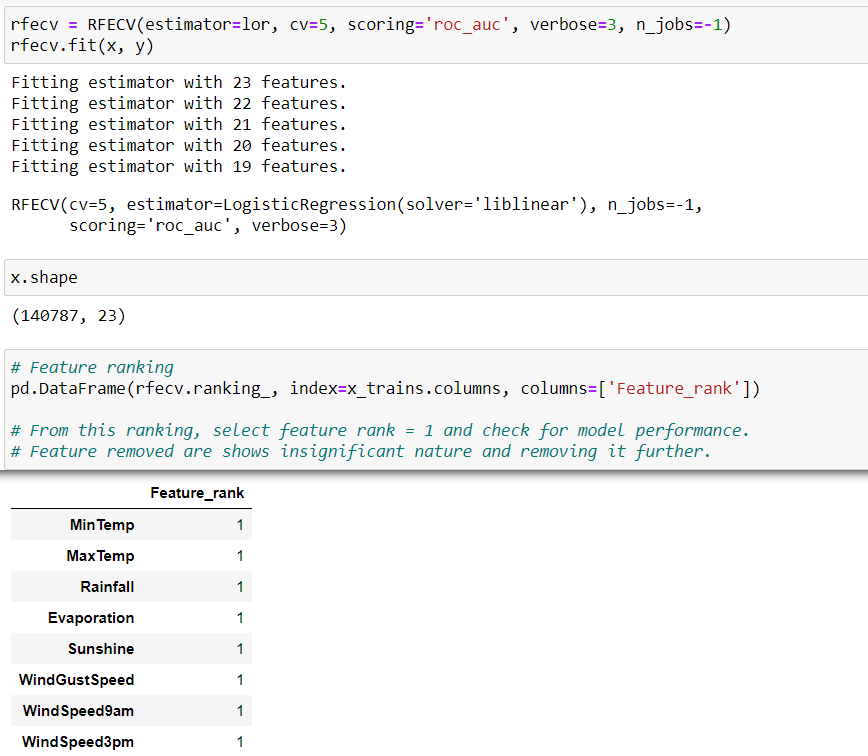
* Overall accuracy of the model is decreased from 85.99% to 82.10%
* Recall score is increased from 56% to 80%. So able to predict rain in better way compared to base model.
* But precision score is dropped from 74% to 57%, so chances of predicting non rainy days as rainy is more.
* There is significant increase in F1 score is observed from 64% to 66%.

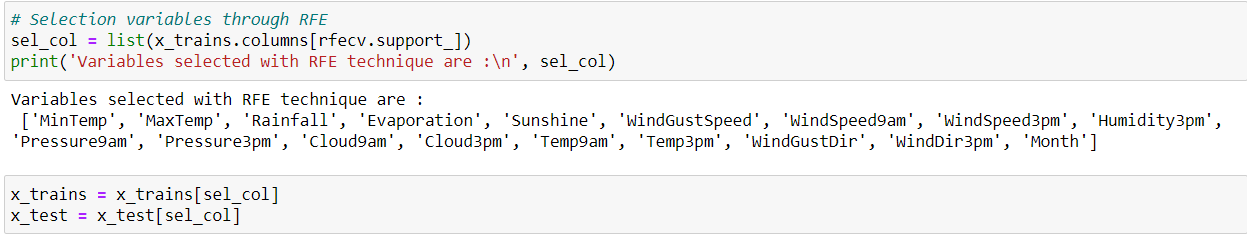
**Feature Selection using RFE (Recursive Feature Elimination) Technique**

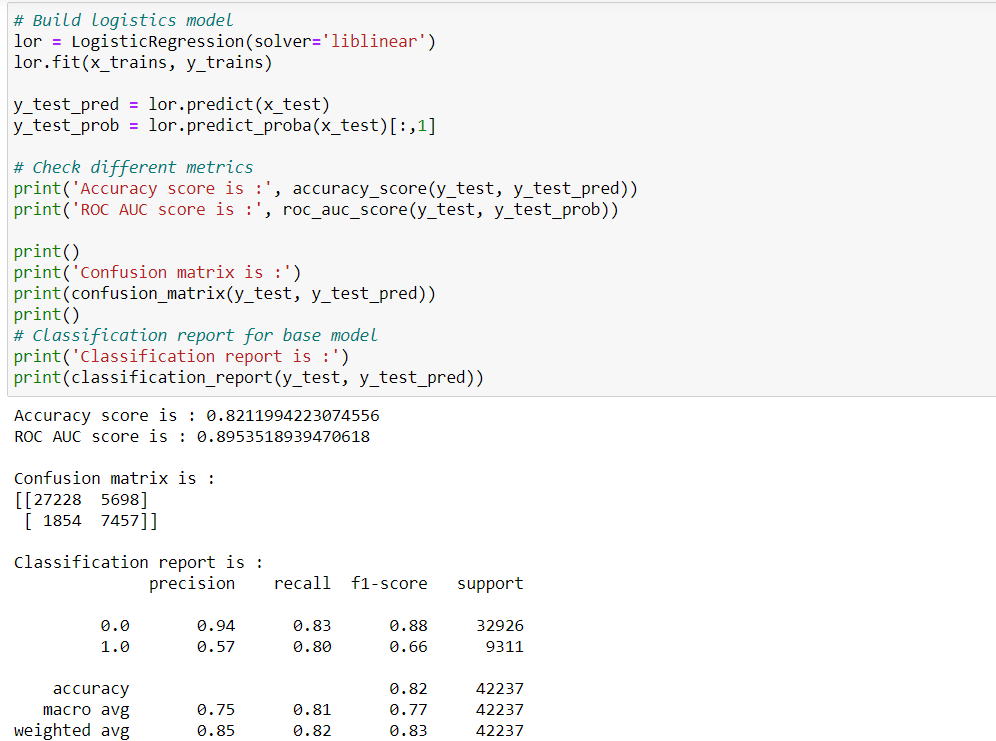
Recursive feature elimination (RFE) is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached. Features are ranked by the model’s coef or feature importances attributes, and by recursively eliminating a small number of features per loop, RFE attempts to eliminate dependencies and collinearity that may exist in the model.

RFE requires a specified number of features to keep, however it is often not known in advance how many features are valid. To find the optimal number of features cross-validation is used with RFE to score different feature subsets and select the best scoring collection of features.

The RFECV visualizer plots the number of features in the model along with their cross-validated test score and variability and visualizes the selected number of features.







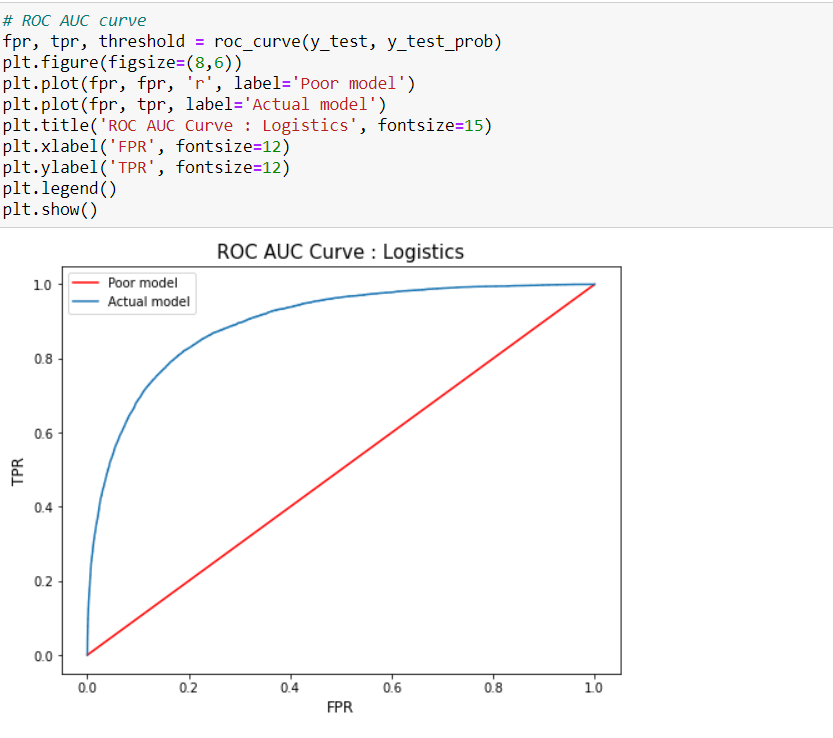
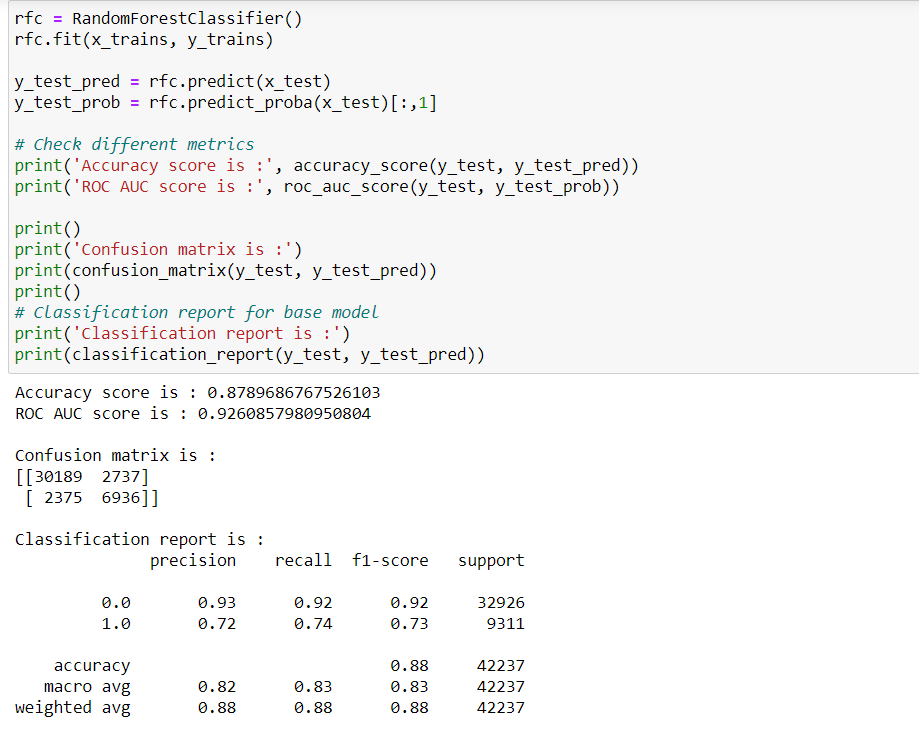


fig (33)

* Even after feature selection, no improvement in model accuracy and other metrics is observed.

**Random Forest Model**

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.



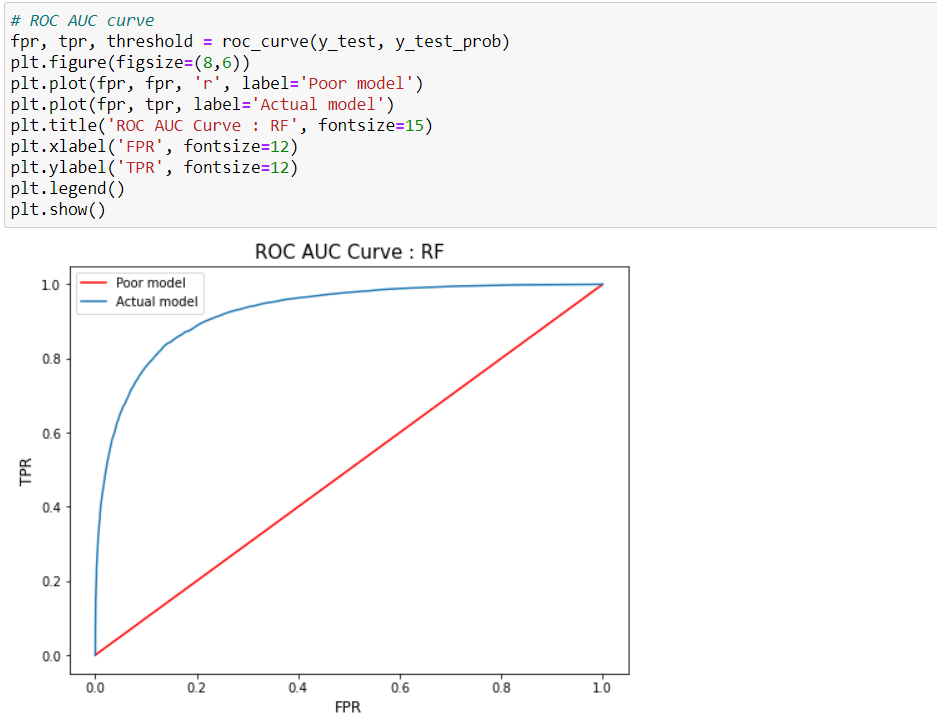


fig (34)

* Overall accuracy of model is increased from 82.11% to 87.92%.
* Recall score for class 0 is good compared to class 1.
* For class 1, recall score again reduced from 80% to 74% but better than base model.
* Precision score is also improved compared to previous model.
* F1 score is increased from 66% to 73% and shows better prediction rate compared to above models.
* From ROC curve, significant increase in TPR and decrease in FPR is observed.

**Hyper Parameter tuning for Random Forest Model**

A Machine Learning model is defined as a mathematical model with a number of parameters that need to be learned from the data. By training a model with existing data, we are able to fi the model parameters.

However, there is another kind of parameters, known as Hyperparameters, that cannot be directly learned from the regular training process. They are usually fixed before the actual training process begins. These parameters express important properties of the model such as its complexity or how fast it should learn.

Models can have many hyperparameters and finding the best combination of parameters can be treated as a search problem. Two best strategies for Hyperparameter tuning are:

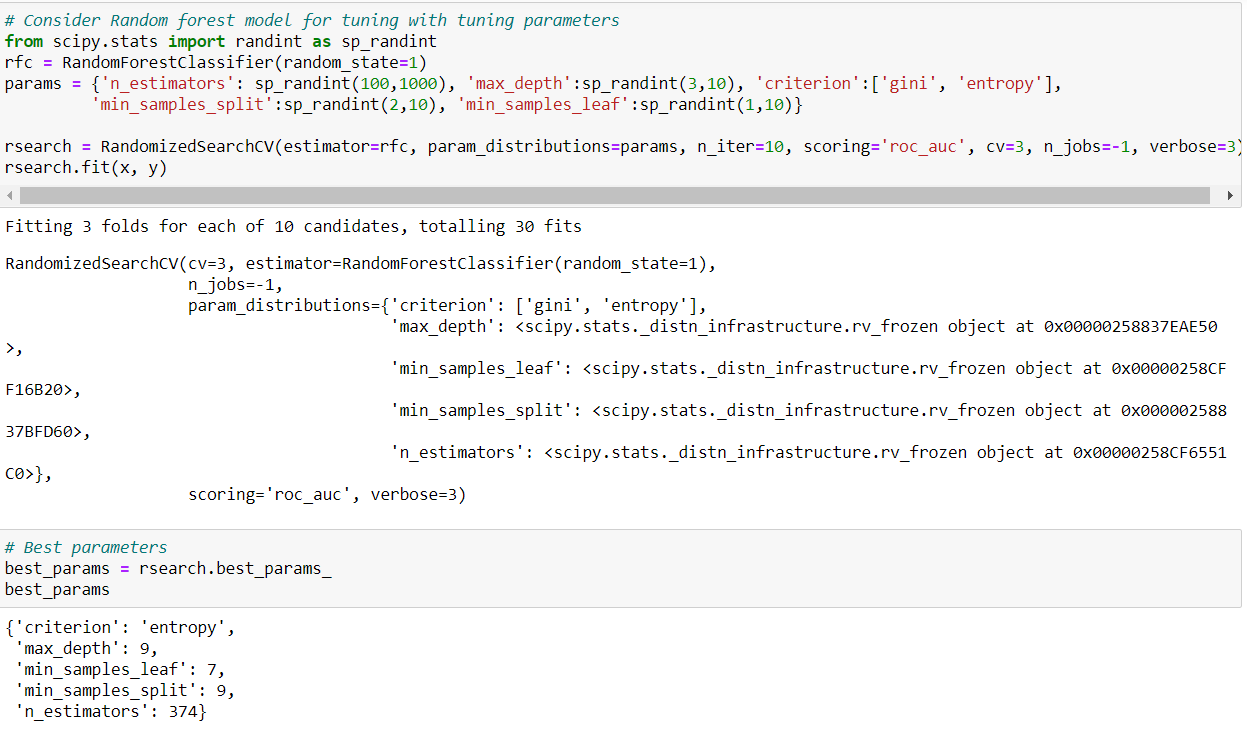
* [GridSearchCV](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html)
* [RandomizedSearchCV](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html)

In GridSearchCV approach, machine learning model is evaluated for a range of hyperparameter values. This approach is called GridSearchCV, because it searches for best set of hyperparameters from a grid of hyperparameters values.

RandomizedSearchCV solves the drawbacks of GridSearchCV, as it goes through only a fixed number of hyperparameter settings. It moves within the grid in random fashion to find the best set hyperparameters. This approach reduces unnecessary computation.

The Random forest or Random Decision Forest is a supervised Machine learning algorithm used for classification, regression, and other tasks using decision trees.  
The Random forest classifier creates a set of decision trees from a randomly selected subset of the training set. It is basically a set of decision trees (DT) from a randomly selected subset of the training set and then It collects the votes from different decision trees to decide the final prediction.

**TRAIL 1:**



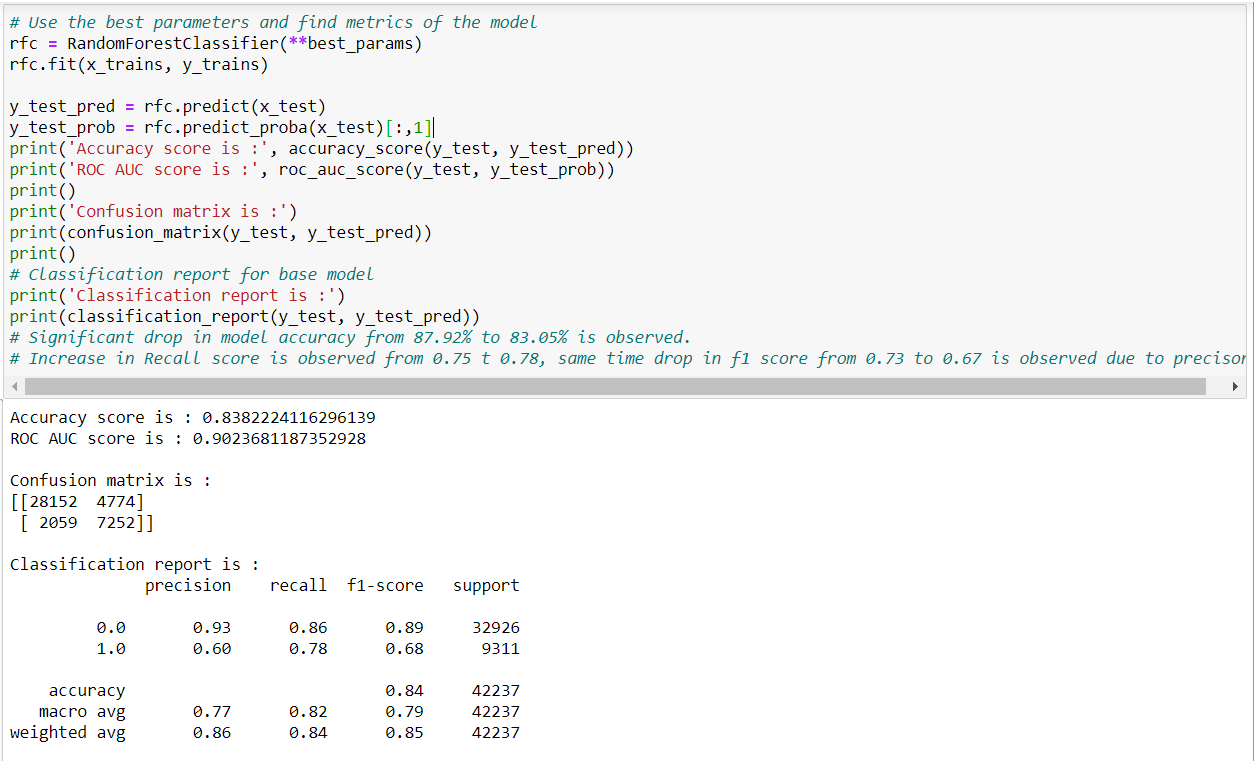
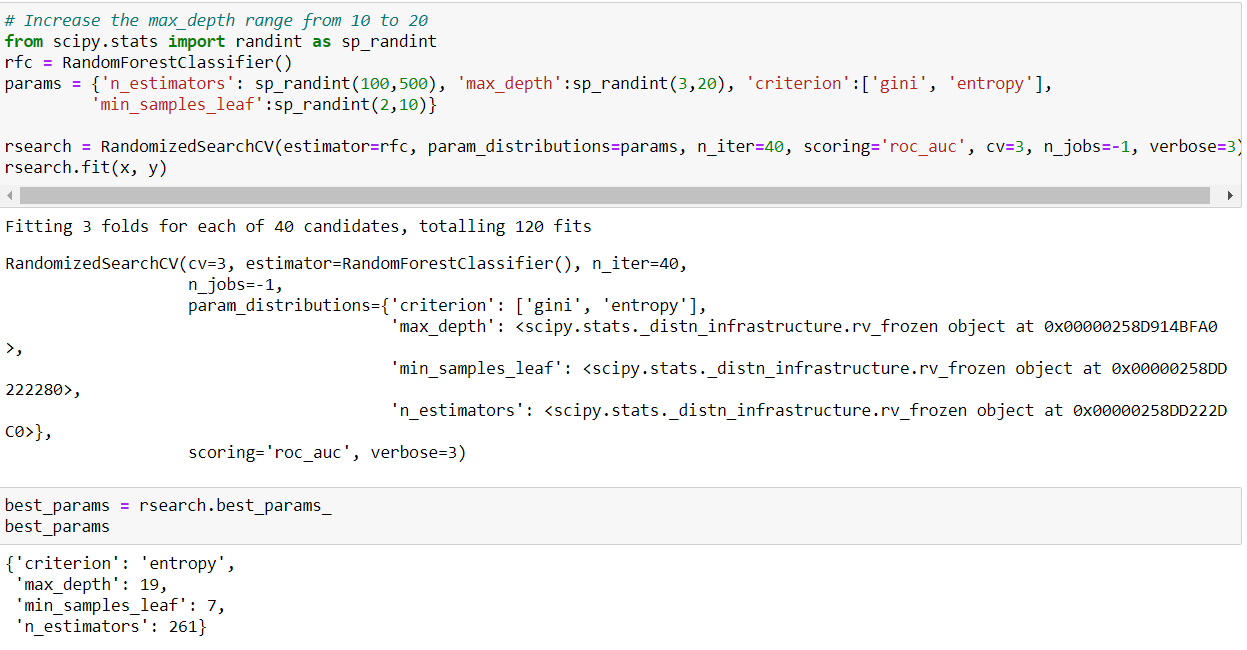


fig (35)

**TRAIL 5:**



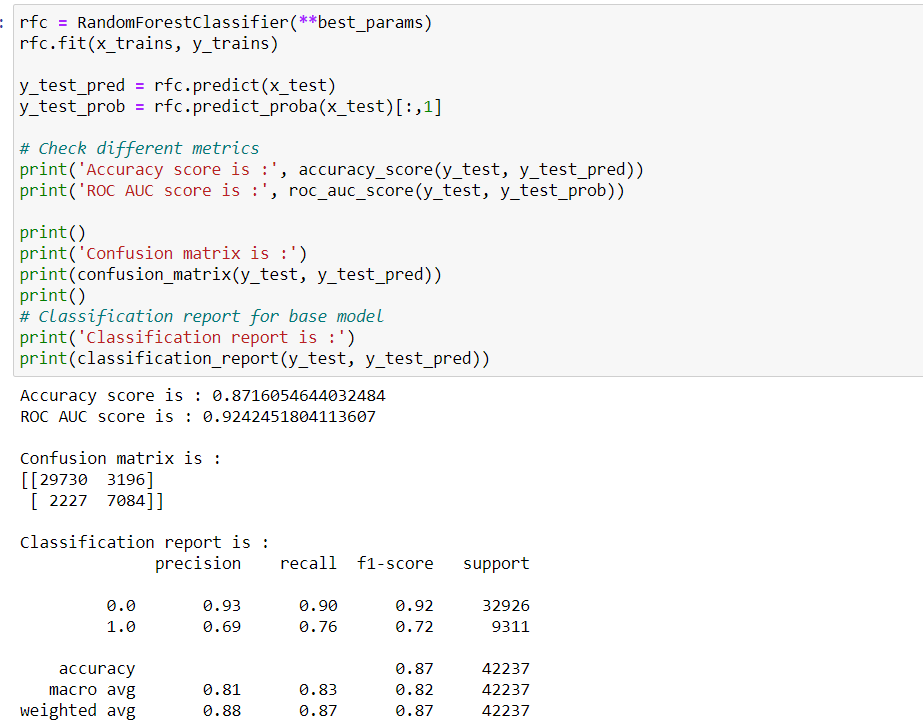


fig (36)

We almost performed 5 trails to check whether there is any improvisation in the model. But we couldn’t find any improvement in the model.

**Boosting Models**

**Boosting** is an ensemble modelling technique which attempts to build a strong classifier from the number of weak classifiers. It is done building a model by using weak models in series. Firstly, a model is built from the training data. Then the second model is built which tries to correct the errors present in the first model. This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models are added.

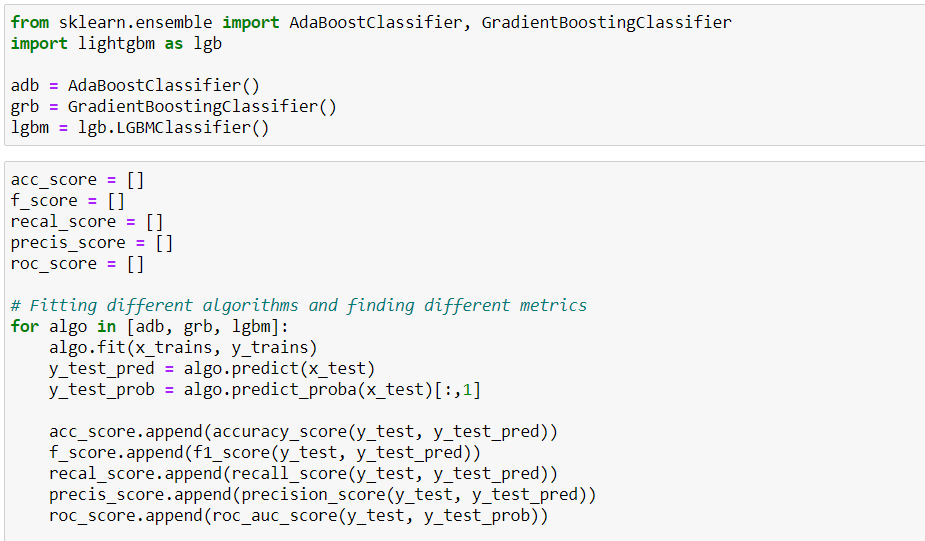
**AdaBoost** was the first really successful boosting algorithm developed for the purpose of binary classification. AdaBoost is short for Adaptive Boosting and is a very popular boosting technique which combines multiple “weak classifiers” into a single “strong classifier”.

Gradient Boosting is a popular boosting algorithm. In gradient boosting, each predictor corrects its predecessor’s error. In contrast to Adaboost, the weights of the training instances are not tweaked, instead, each predictor is trained using the residual errors of predecessor as labels.

There is a technique called the Gradient Boosted Trees whose base learner is CART (Classification and Regression Trees).

**LightGBM** is a gradient boosting framework based on decision trees to increases the efficiency of the model and reduces memory usage.

It uses two novel techniques: **Gradient-based One Side Sampling** and **Exclusive Feature Bundling (EFB)** which fulfils the limitations of histogram-based algorithm that is primarily used in all GBDT (Gradient Boosting Decision Tree) frameworks.



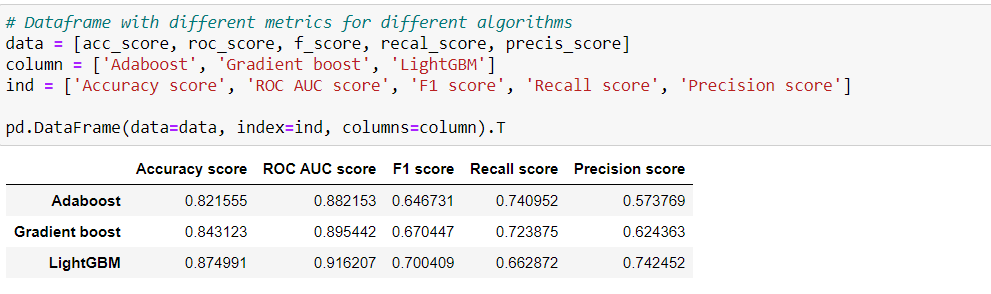
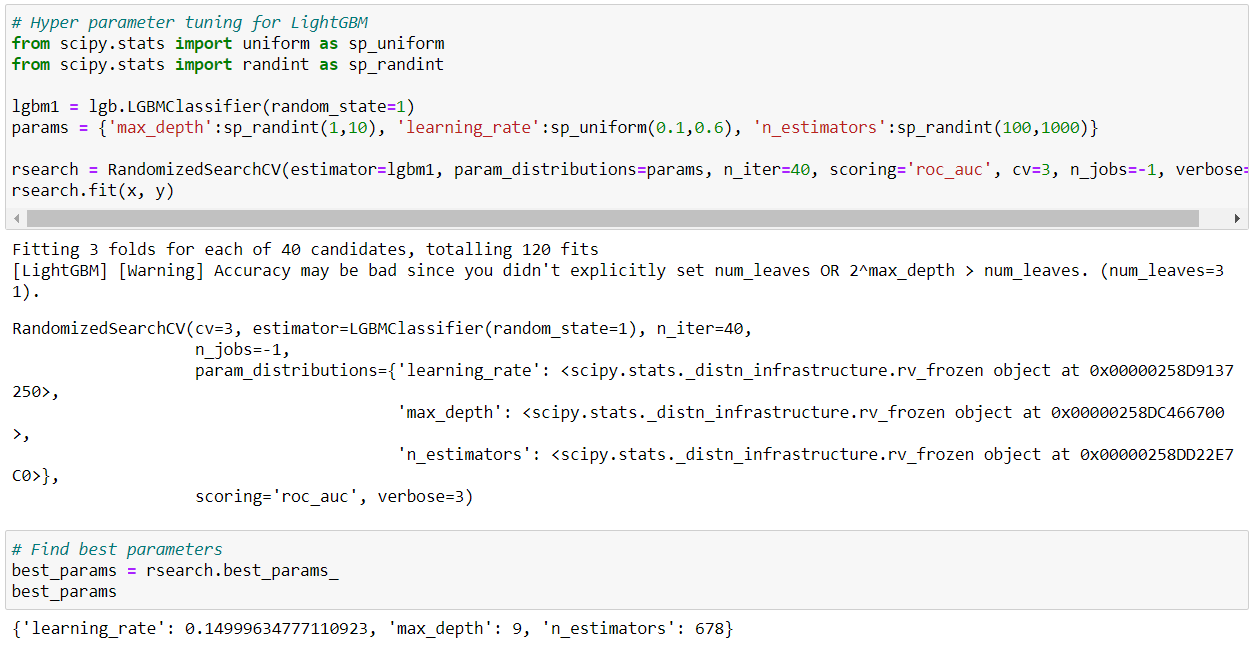
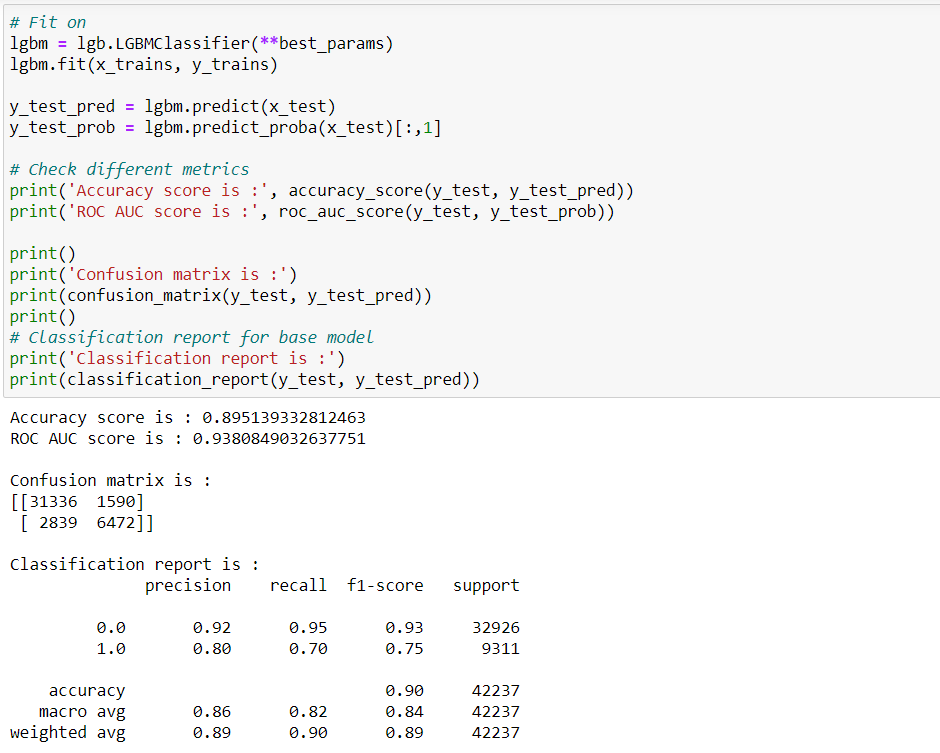


fig (37)

* From above boosting models, LightGBM is giving good model accuracy as well as ROC AUC score.
* Precision score and F1 score obtained is good for LightGBM model.
* Chosen this model further for our analysis.

**Hyper Parameter Tunning for LightGBM**





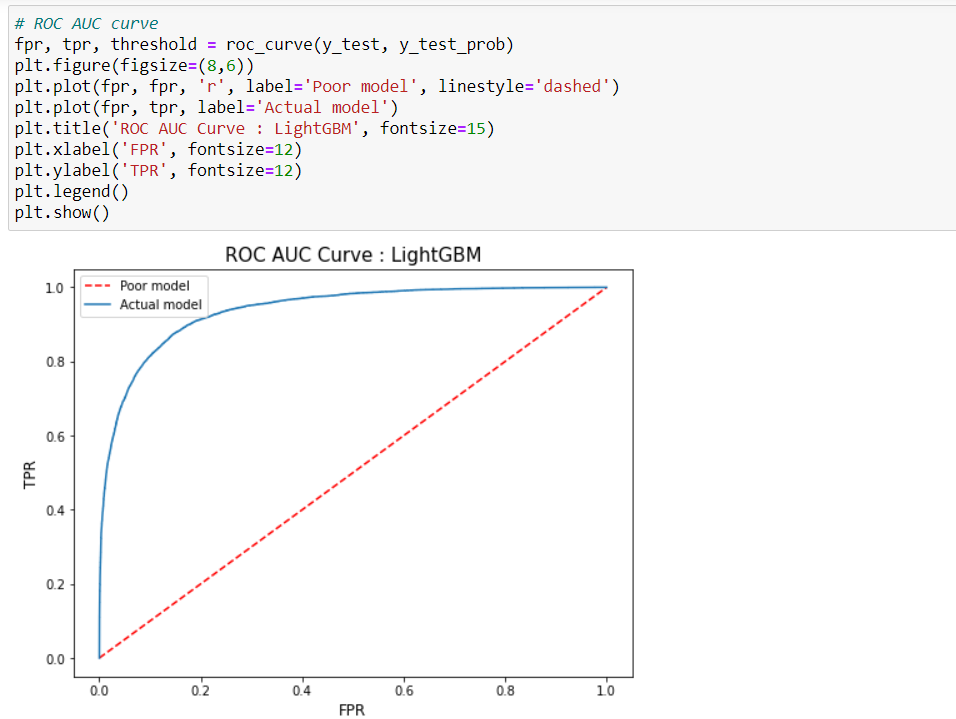


fig (38)

**Model Comparison**

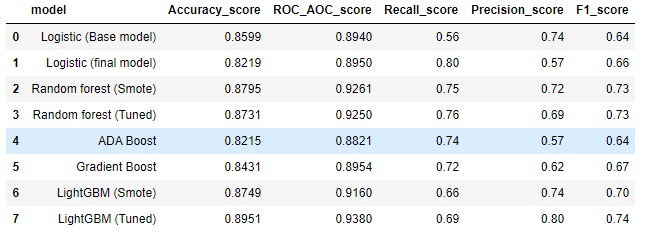


fig (39)

We went through various model analysis like Logistic model, Random Forest, Hyper Tunning,

and also Boosting the models.

Finally, we are considering LightGBM (Tunned) to be our final model.

LightGBM (Tuned) is giving good model accuracy as well as ROC AUC score.

Also, the Precision score and F1 score obtained is good for LightGBM (Tuned) model.

From the above, data frame we are selecting LightGBM (Tuned) as the best method for predicting whether tomorrow is the rainy day or not

**Model Stability using KFold**

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

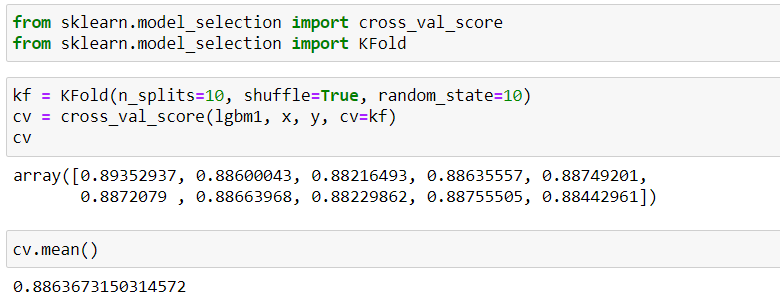


fig (40)

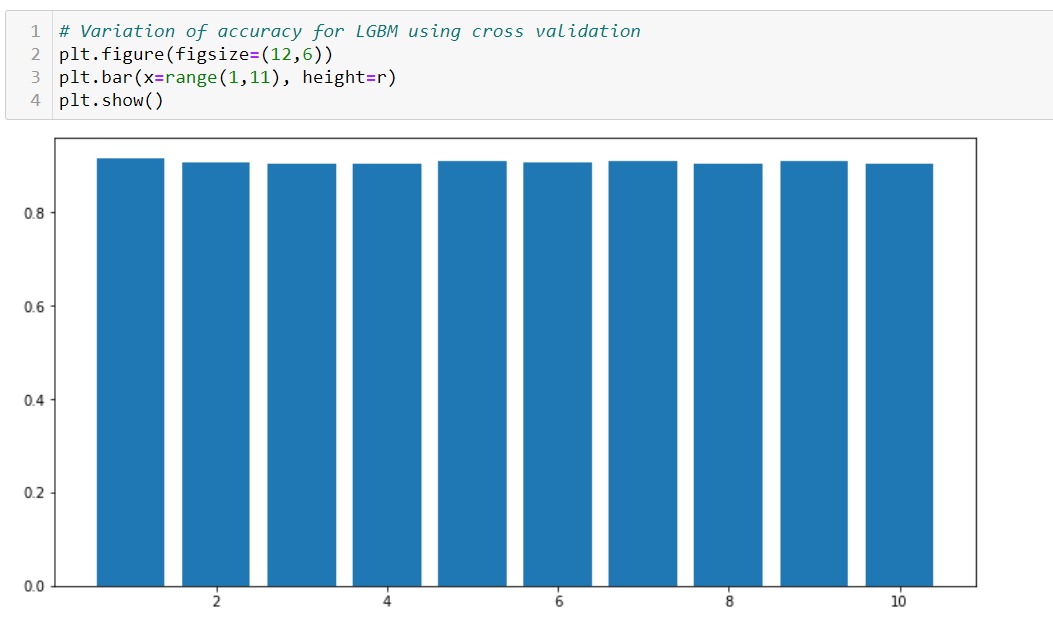


fig (41)

We have checked the stability of our final model (LGBM) with K-Fold cross validation with n\_splits = 10.

From the bar plot, we can infer that for each split, accuracy of around 90% is achieved which matches with our LGBM (Tuned) model accuracy. Hence, we consider this model to be most stable.

**Future Scope**

The fact that Rainfall nowadays is not happening seasonally, there are many reasons such as Deforestation, Industrial revolutions, Urbanization or Modernization, hazardous gases from vehicles and many more.

Usage of older techniques such as Persistence, usage of barometer, looking the sky, nowcasting won’t help now to predict the rainfall effectively.

The Objective or Solution is to Predict the occurrences of rainfall in future using analytical or Machine learning techniques using various Machine Learning models, with the help of various features.

**Conclusion**

Currently machine learning is used in many industries. As the data increases the complexity of that data will increase and for that we are using machine for the better understanding of that data. In Weather predictions it’s pretty helpful with good accuracy score and in rainfall also its gives pretty good predictions. In future we are planning to increase our work in Storm predictions and Crop prediction with the rainfall prediction.

**References: -**

* https://www.kaggle.com/jsphyg/weather-dataset-rattle- package/tasks?taskId=278
* <https://en.wikipedia.org/wiki/Weather_forecasting#Ancient_forecastin>g
* https://ieeexplore.ieee.org/document/8938211/references#references