Comparison of three Prediction Algorithms on Raining Dataset: RFC, LR and NBC

Three prediction algorithms namely Random Forest Classifier, Logistic Regression and Naive Bayes Classifier have been used to train the available rainfall dataset. This dataset contains about 10 years of daily weather observations from numerous Australian weather stations. RainTomorrow is the target variable to predict.

Date The date of observation	Location The common name of the location of the weather station	MinTemp The minimum temperature in degrees celsius	MaxTemp The maximum temperature in degrees celsius	Rainfall The amount of rainfall recorded for the day in mm	Evaporation The so-called Class A pan evaporation (mm) in the 24 hours to 9am	Sunshine The number of hours of bright sunshine in the day.	WindGustDir The direction of the strongest wind gust in the 24 hours to midnight
1Nov07 25Jun17	Canberra 2% Sydney 2% Other (138680) 95%	NA 1% 11 1% Other (143076) 98%	506 unique values	0 63% 0.2 6% Other (45619) 31%	NA 43% 4 2% Other (79331) 55%	NA 48% 0 2% Other (73266) 50%	NA 7% W 7% Other (125219) 86%
2008-12-01	Albury	13.4	22.9	0.6	NA	NA	W
2008-12-02	Albury	7.4	25.1	0	NA	NA	WNW
2008-12-03	Albury	12.9	25.7	0	NA	NA	WSW
2008-12-04	Albury	9.2	28	0	NA	NA	NE
2008-12-05	Albury	17.5	32.3	1	NA	NA	W
2008-12-06	Albury	14.6	29.7	0.2	NA	NA	WNW
2008-12-07	Albury	14.3	25	0	NA	NA	W
2008-12-08	Albury	7.7	26.7	0	NA	NA	W
2008-12-09	Albury	9.7	31.9	0	NA	NA	NNW
2008-12-10	Albury	13.1	30.1	1.4	NA	NA	W
2008-12-11	Albury	13.4	30.4	0	NA	NA	N
2008-12-12	Albury	15.9	21.7	2.2	NA	NA	NNE
2008-12-13	Albury	15.9	18.6	15.6	NA	NA	W
2008-12-14	Albury	12.6	21	3.6	NA	NA	SW

Fig-1: Dataset containing Rainfall Data With 23 different attribues

There are various other attributes in the dataset like date, location, minTemp, maxTemp, rainfall(in mm), evaporation, sunshine, wind gust direction, wind gust speed and wind direction. The shape of the dataset is (145460, 23) where range index ranges from 0 to 145459 entries and total number of columns are 23.

We can also calculate the target count of column, ‘RainTomorrow’, for our reference.

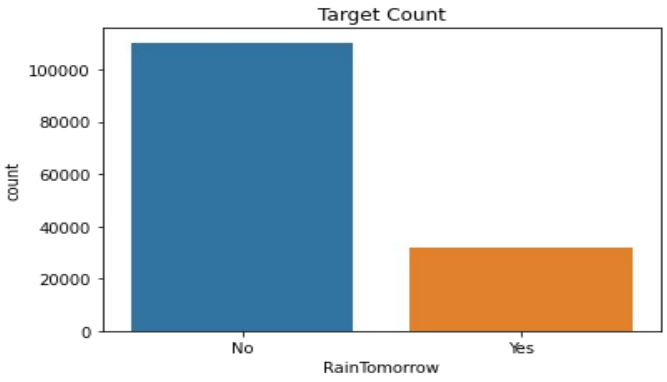

Fig-2: Target Count

Here, firstly, we will perform the data preprocessing on the weatherAUS.csv file containing the input dataset. We will drop the null values from the target column and then we will split our dataset into categorical and integer data-types. Categorical columns will consists of columns like date, location, wind gust direction, RainToday and RainTomorrow, while integer data types will consists of columns like minimum/maximum temperature, rainfall, sunshine, pressure and such similar attributes.

After that, outliers need to be analyzed for numerical datatype columns, which can be done using box-plot.

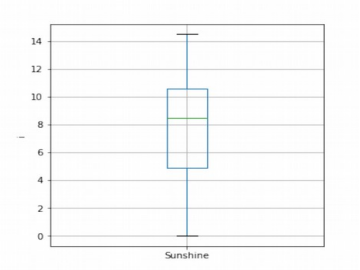

Fig-3: Boxplot of Sunshine

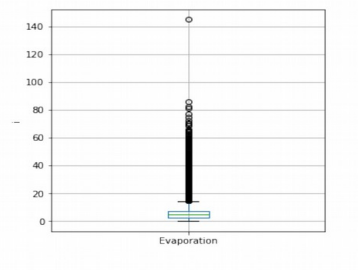

Fig-4: Boxplot of Evaporation

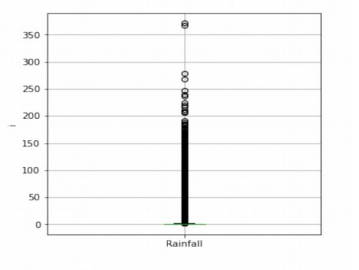

Fig-5: Boxplot of Rainfall

Then, we will remove those outliers using IQR (Interquartile Range) method. IQR is used to measure variability by dividing a data set into quartiles. IQR is the range between the first and the third quartiles namely Q1 and Q3: IQR = Q3 – Q1. The data points which fall below Q1 – 1.5 IQR or above Q3 + 1.5 IQR are outliers.

Then, we will split the data into training and testing dataset having shape (85016, 24) and (21255, 24) respectively. For numerical data, we will replace all the null values with their respective column’s median value and for categorical data, we will use mode to replace null values. Then, we will use LabelEncoder() for encoding the categorical variables.

Before training our model, we will perform Min-max normalization on the dataset. Now, finally we are all set to train our model using Logistic Regression, Naive Bayes Classifier and Random Forest Classifier available in sklearn library of Python. The accuracy score for each are:

Logistic Regression Performance Score = 0.8711
Naive Bayes Performance Score = 0.6317
Random Forest Performance Score = 0.8673

As, observed LR and RFC almost give similar scores and perform much better compared to NBC. Since, LR gives the best score, we will further calculate confusion matrix, sensitivity and specificity using it.
The confusion matrix is obtained as : [[17466 531] [2209 1049]]
The sensitivity and specificity scores are 0.8877 and 0.6639 respectively.

Though, here LR performed better than the other two, but the performance of any algorithm always entirely depends on the type of our dataset. Logistic regression is easier to implement, interpret, and very efficient to train, but if the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to overfitting.

Courtesy:

https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package
http://www.bom.gov.au/climate/dwo/
http://www.bom.gov.au/climate/data
https://www.kaggle.com/code/sghanvir/rain-in-australia-prediction-via-lrc-nbc-rfc