Rainfall Prediction using Australia Dataset

Alexender Kissiedu
IT Training and Support Section
Directorate of ICT Services
Cape Coast, Ghana
PS/MCS/21/0007

Evans Ankomah
IT Training and Support Section
Directorate of ICT Services
Cape Coast, Ghana
PS/MCS/21/0009

Abstract— **Particularly, human beings make decisions in the present depending on a variety of variables, such as short-term rainfall and temperature projections. Because it depends on so many different factors or features, forecasting rainfall can be difficult. Typically, probabilistic models are utilized, which provide forecasts with a margin of error, making them often not very accurate. As a result, applying machine learning techniques can greatly enhance these forecasts. In order to determine whether rainfall will occur in Australia or not, we explained and implemented how to construct a Logistic Regression model to make such prediction. At the end our model was able to achieve accuracy score of 84%.**

Keywords—Logistic Regression, rainfall prediction, Machine learning

I. Introduction

Weather, and humankind's ability to accurately predict it, plays a critical role in many aspects of life. From farmers growing crops to a family planning a weekend vacation to logistical decision making within airlines, rain in particular is highly influential regarding plans. In some instances, the impact of rain can have large financial consequences. As a result, there is a strong interest from stakeholders in the ability to accurately forecast rain. Given the high uncertainty of the information that is sometimes available, probabilistic models are utilized to make the forecast, and errors frequently occur (Valipour, 2016).

The big question is, if it rains today, will it rain again tomorrow? Making prediction of real-time and accurate rainfall remains challenging for many decades due to its non-linear nature. Traditionally, to address this big question, numerical predictions(Wu, 2008) that is centered on complex mathematical models of thermodynamics and fluid dynamics were used. The use of machine learning techniques for modeling huge amount of data has become readily available recently (Murphy & Winkler, 1984). This paper evaluates the effectiveness of outlier detection, classification, and Logistic Regression machine learning technique to predict rainfall based on the Australian dataset. There are various factors or features that can influence the weather patterns and it affects different parts of the world.

This paper however seeks to use the available Australian data to create a next-day prediction model for whether or not it will rain. This model could be utilized in a weather app for the benefit of the public at large. We will focus on finding out the features from the dataset that contribute significantly in making raining prediction in Australia. By training a logistic regression model with the target variable "RainTomorrow", we will ultimately determine whether or not it will rain tomorrow.

II. Data Set

The data used in this project was downloaded from the Kaggle dataset titled(*Rain in Australia | Kaggle*, n.d.), which in itself was also originally sourced from the Australian Bureau of Meteorology's(*Daily Weather Observations*, n.d.). Additional weather metrics for Australia can be found within the bureau's (Climate Data Online - Map Search, n.d.) web app. The dataset contains about 10years of daily weather observations from different Australian weather stations. Observations were drawn from numerous weather stations. The dataset contains 145460 records under 23 different categories or features. Out of the 23 features 6 of them are categorical variables, 16 are numerical and one target variable. Below is a short description of each feature according the author of the Kaggle dataset and the (*Notes to Accompany Daily Weather Observations*, n.d.)published by the Australian Bureau of Meteorology.

A. Categorical Features

- Date: The date of observation
- Location: The common name of the location of the weather station
- RainToday: Boolean: 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0
- WindDir3pm: Direction of the wind at 3pm
- WindDir9am: Direction of the wind at 9am
- WindGustDir: The direction of the strongest wind gust in the 24 hours to midnight

B. Numerical Features

- 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.
- WindGustSpeed: The speed (km/h) of the strongest wind gust in the 24 hours to midnight
- WindSpeed9am: Wind speed (km/hr) averaged over 10 minutes prior to 9am
- WindSpeed3pm: Wind speed (km/hr) averaged over 10 minutes prior to 3am
- Humidity9am: Humidity (percent) at 9am
- Humidity3pm: Humidity (percent) at 3pm
- Pressure9am: Atmospheric pressure (hpa) reduced to mean sea level at 9am
- Pressure3pm: Atmospheric pressure (hpa) reduced to mean sea level at 3pm
- Cloud9am: Fraction of sky obscured by cloud at 9am. This is measured in "oktas", which are a unit of eigths. It records how many eigths of the sky are obscured by cloud. A 0 measure indicates completely clear sky whilst an 8 indicates that it is completely overcast
- Cloud3pm: Fraction of sky obscured by cloud (in "oktas": eighths) at 3pm. See Cload9am for a description of the values
- Temp9am: Temperature (degrees C) at 9am
- Temp3pm: Temperature (degrees C) at 3pm

C. *Target Variable*

- RainTomorrow: The target variable. Did it rain tomorrow? (1 = yes, 0 = no)

III. Data Preprocessing

A. *Handling missing values*

The primary preprocessing used for treating this dataset is handling the missing values. We dropped all columns with more than 30% of null values, because we have quiet a hunge dataset and dropping them might not reduce the dataset much but rather help us come out with a more robust model. The columns or features that were affected by this decision are: Sunshine, Evaporation, Cloud3pm, and Cloud9am.

Similarly, even though the Location and Date columns or features were void of null values we dropped them since we were interested in predicting whether it will rain in a particular location in Australia.

For the categorical missing values, we decided to replace them with their respective modes. We decided to go this way after carefully visualizing and understanding the variables. From all the plots visualized we observed that the data the categorical features looked somehow normally distributed and hence replacing null values for these categorical values will the ideal way to go.

Again, after performing a descriptive statistic on the numerical data and visually them using hist and boxplot. We observed that most of the numerical variables have their mean less than the 75th percentile. And all of them seems to be normally distributed with most values falling closer to the mean except the rainfall feature that looks quite skewed extremely. This means that deciding to replace these missing values with their respective mean will not introduce additional outliers into the dataset except for the rainfall feature which is positively skewed. Hence, we decided to replace the missing values in rainfall with the median and the rest of the features with their respective mean.

B. *Detecting and removing outliers*

We used boxplot for visualizing and checking for outliers using in the features. From the boxplot we observed that all the features had outliers except Humidity3pm. Since most are skewed we decided to treat the outliers in the features using the interquartile range (IQR) approach(Vinutha et al., 2018). We did so by setting up a “fence” outside of Q1 and Q3. And we consider any value outside any of this fence as outliers. To build this fence we took 1.5 times the IQR and then subtract this value from Q1 and added this value to Q3.

IV. Visualizations

Fig. 1 shows that the target is imbalanced and the number of No's is far less than the Yes's. This might affect the model's performance. But we left it as it is so since we don't want to alter the true reflection of the model’s prediction.

Fig. 2 also shows scatter plot that shows that there is a linear relationship between the max and min temperature. That’s as the “minTemp” increase the “Maxtemp” increases as well. And this linear relationship increases the tendency of raining tomorrow

Similarly, from Fig. 3 we can tell that as humility increase around 9am the probability of “rain tomorrow” also increase and as the temperature also increase the probability of “rain tomorrow” decreases.

In fig. 4 we see a heatmap showing how the features correlate with each other and their is a very strong positive correlation between pressure3pm and pressure9am that's 0.96 as well a strong correlation between temp3pm and maxTemp with 0.97.

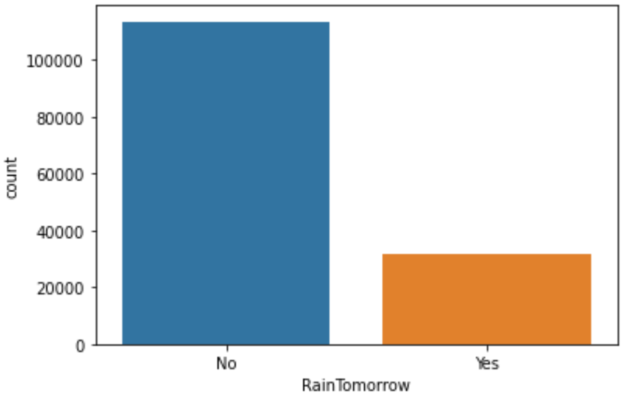

Fig. 1. Distribution of the target variable

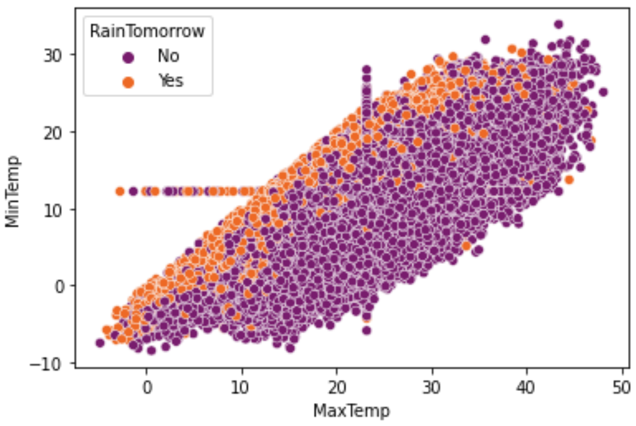

Fig. 2. Scatterplot showing the relationship between max and min temperature

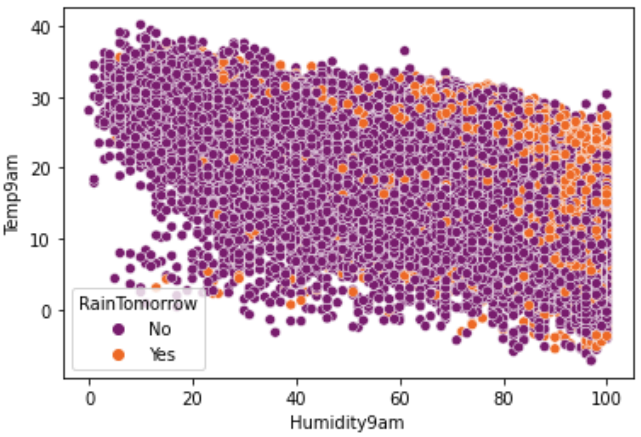

Fig. 3. Scatterplot showing the relationship between humidity9am and tem9m

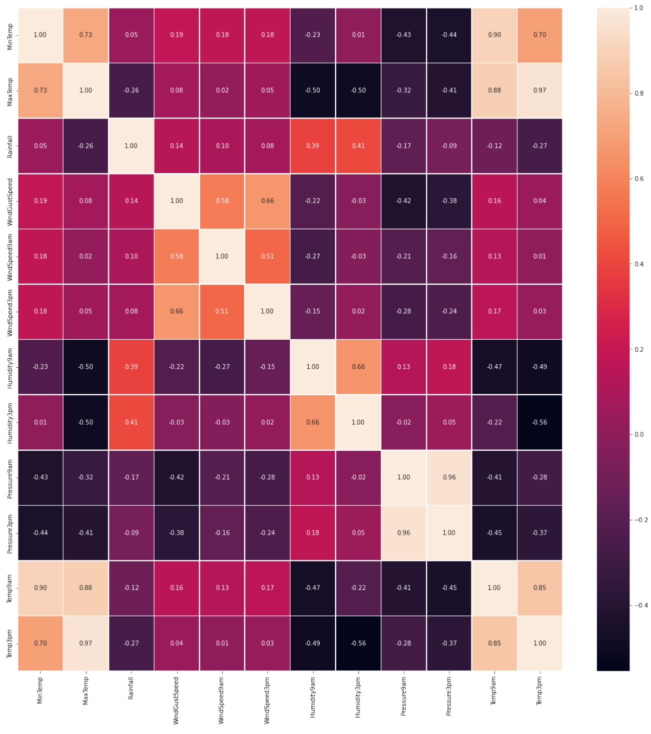

Fig. 4 Heatmap of the various features

V. Feature selection and Model building

A. Feature Enoding

Before proceeding to build the logistic regression model, we first encoded the categorical features using the one hot encoding. There are other encoding schemes such as label encoding, but we decided to go with the one hot encoding technique. The one hot encoding is the preferred as compared to label encoding which makes the data seem like there is a rank in the values(Srinidhi, 2020). The purpose of this is to help our model not to assume that higher numbers are more important.

B. Feature Scaling

Since the features are measured from different scales, we applied feature scaling to the dataset. We resulted to use the standard scaler in in standardizing the dataset because our dataset looks normally distributed. And the purpose of doing this is to removes the mean and scales each feature to unit variance. It involves the estimation of the empirical mean and standard deviation of each feature.

C. Model Building

As stated in the introduction, we will build, train, and forecast whether or not it will rain tomorrow in Australia logistic regression algorithm on the dataset. This algorithm simply uses statistics to solve classification issues. It enables us to forecast the likelihood that an input will fall into a particular category. The core of this is the logit function or sigmoid function. The data science community claims that logistic regression can resolve 60% of the current categorization issues (Dinesh, 2021).

VI. Results

Our model produced an accuracy score 0.8415 on our testing data. This means that our model has the tendency of predicting 84.15% correctly on a new dataset. After evaluating our model however, we had an accuracy score of 83.75% (which if approximately 84%). This score although not ultimate is a good score for this model.

A. Overfitting and Under Fitting:

For checking the concept of overfitting or underfitting we compared the accuracy score on our training dataset with that of the testing dataset. The model predicted 84.11% correctly on the training set and 84.15% correctly on the testing dataset. This indicates and reduces the concerns of the model being overfitted or underfitted. So, we can say that our model can generally perform well for new unseen data since the accuracy score were both similar.

B. Evaluating our Model

To evaluate the performance of our model we used the K_fold cross validation. This is a more accurate method of evaluating our model since it uses all of our data to build and test the model. We spliced our data into 5 folds. For each fold we produce an evaluation metric, in this case we used the accuracy to evaluate our model, which is essentially the percentage of correct predictions. We then took the mean of the 5 scores to evaluate our model.

At the end we got an overall Accuracy of 0.8375, that is approximately 84%. Hence, we can establish that our model has a good accuracy of 84%, that is our logistic regression model can get approximately 84% of predictions correct, which is a good prediction score.

Similarly, we can say deduce from our classification report that, from the recall we saw that only only 45% of the data were predicted to be positive which is no so good and only 72% were correct according to the precision. we also observed that the metrics for ”rain tommorrow(1)” was calculated from 8038 data and a total of 28327 were used to calculate for the ”not rain tommorrow(0)”. We also observed an f1 score that gave us a harmonic mean of 56% and 90% for both positive and negative prediction respectively. Again, we achieved a precision score 84% that is we were able to predict 84% as correct according the classification report.

Again, we plotted the ROC curve and realized a ROC AUC score of 70%. This means the prediction has 70% probability of distinguishing between positive and negative classes. This although is not the best score is considered as an acceptable score according to (Mandrekar, 2010)

C. *Finding whether model performance could be improved using Cross Validation Score*

After running a cross validation check on the training dataset, we had a validation score of 0.8372. That is the mean accuracy score of cross validation was almost the same as the original model’s accuracy score of 0.8415. So, we are certain that the accuracy of our model may not be improved massively using Cross-validation.

VII. Conclusion

In conclusion we were able to build, train and test a logistic regression that could predict 84% correctly on an unseen dataset.

While this model is a good starting point for rain prediction in Australia, there are several ways in which the model could be improved upon: Further hyperparameter tuning and engineering new features such as trailing amounts of rain or sunshine can be done to improve the model. Similarly, collecting additional data from nearby countries (for example, does rain originating in USA or Ghana have predictive power?). Again, another area to look at in the future works is to by attempting to predict the amount of rainfall.

References

Climate Data Online - Map search. (n.d.). Retrieved December 5, 2022, from http://www.bom.gov.au/climate/data/

Daily Weather Observations. (n.d.). Retrieved December 5, 2022, from http://www.bom.gov.au/climate/dwo/

Dinesh, K. (2021, August 21). *Introduction to Logistic Regression - Sigmoid Function, Code Explanation*. https://www.analyticssteps.com/blogs/introduction-logistic-regression-sigmoid-function-code-explanation

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316.

Murphy, A. H., & Winkler, R. L. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association*, *79*(387), 489–500.

Notes to accompany Daily Weather Observations. (n.d.). Retrieved December 5, 2022, from http://www.bom.gov.au/climate/dwo/IDCJDW0000.shtml

Rain in Australia | Kaggle. (n.d.). Retrieved December 5, 2022, from https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package

Srinidhi, S. (2020). *Label Encoder vs. One Hot Encoder in Machine Learning*. Medium.

Valipour, M. (2016). How Much Meteorological Information Is Necessary to Achieve Reliable Accuracy for Rainfall Estimations? *Agriculture 2016, Vol. 6, Page 53*, *6*(4), 53. https://doi.org/10.3390/AGRICULTURE6040053

Vinutha, H. P., Poornima, B., & Sagar, B. M. (2018). Detection of outliers using interquartile range technique from intrusion dataset. In *Information and decision sciences* (pp. 511–518). Springer.

Wu, J. (2008). A novel nonlinear ensemble rainfall forecasting model incorporating linear and nonlinear regression. *Proceedings - 4th International Conference on Natural Computation, ICNC 2008*, *3*, 34–38. https://doi.org/10.1109/ICNC.2008.586