Predicting “Illogical” Weather Condition by Logical Ways

----WeatherAUS Database Research

Group Member: Fangzheng Hu, Kaibin Ye, Min Jiang, Yaoshitu Ma, Yun Lei

Summary

Weather predictions are made by collecting quantitative data about the current state of the atmosphere at a given place and using meteorology to project how the atmosphere will change. Currently, the main predictors of the prediction processes are changes in barometric pressure, current weather conditions, and sky condition or cloud cover, weather forecasting now relies on computer-based models that take many atmospheric factors into account. Researchers will select the best possible forecast model based on model performance and knowledge of model biases. The use of ensembles and model consensus help narrow the error and pick the most likely outcome.*

This project focuses on predicting whether it will rain tomorrow based on the weather indicators today. Logistic regression, KNN, decision tree, and random forest are utilized to perform the prediction. Besides, autoregression is applied to predict the rainfall in a short time period.

1. Data Preprocessing

Our group downloads the data set from kaggle and the description of the data set is in exhibit 1.

1.1. Missing value treatment

We remove column Evaporation, Sunshine, Cloud9am and Cloud3pm because high proportion missing value. Then rows with missing data also are removed.The total sample size decreases from 142,193 to 112,925.

1.2. Check Whether Samples are Balanced

The target of this project is to predict whether it will rain tomorrow. Corresponding to the downloaded dataset, the target variable to be predicted is RainTomorrow, which has 2 values - Yes (means it will rain tomorrow) and No (means it won’t rain tomorrow). First, the dataset should be checked whether it has balanced samples, since fitting an unbalanced dataset makes the fitted models more likely to have bias. Exhibit 2 shows in Australia, the number of not-rainy days is much larger than the number of rainy days. Exhibit 3 shows the same result in each city of Australia. Obviously, the dataset does not have balanced samples. This problem is tackled in the third section of this report.

1.3. Outliers Treatment

According to our observation, we find some observations that appear far away and diverges from an overall pattern which can result in wildly wrong estimations, so we conclude those must be outliers(as shown in exhibit 4). Because there are too many outliers for variable Rainfall and RISK_MM, maybe it is a common situation that numbers deviate from the normal value. So we only remove the outlier of other numerical variables.(Exhibit 5 shows an example of deleting outliers)

1.4. Analysis Covariance

Exhibit 6 shows the covariance of all numerical variables.

2. Autoregression

*Retrived from (https://en.wikipedia.org/wiki/Weather_forecasting)

To predict the rainfall in a short time period, auto-regression Model could be applied to each city. To do that, we first divide the rainfall into 44 different group regarding to different cities. Take the first city Albury as an example. Exhibit 6 is the time series plot for Albury. From the single plot, it is hard to determined the distribution of this series. By applying the autoregressive model, we could test whether rainfall is random walk or related to historical data.

After ran the one step auto-regression model for all 44 cities, only 23 cities fit into AR(1) model at the 10% significance level. However, this does not mean rainfall in other 23 cities follow the random walk. Further research need to be done regarding to the autocorrelation function. In this report, we take Albury as an example here. Based on the autocorrelation and partial autocorrelation function which are shown in Exhibit 7&8, ARMA(2,6) would be the best fit to predict the rainfall in Albury. We run ARMA(2,6) and the model shows as below.

$$Y_t= 0.41Y_{t-1}-0.115Y_{t-2}-0.227\varepsilon^2_{t-1}+0.0289\ \varepsilon^2_{t-2}+0.0642\varepsilon^2_{t-6}+ \sigma^2$$

L-jung Box test could be used to test the efficiency autoregression model. The null hypothesis assume that the error term in this model is white noise and follow the Chi-square distribution. P-value under L-jung Box test is 0.0003516, there is no enough evidence to reject the null hypothesis. By using this model, a short time period rainfall in Albury could be predicted. Other 43 cities could be analysed by this method.

3. Prediction

The target is to predict whether it will rain tomorrow. This is a binary classification problem. Logistic regression, knn, decision tree and random forest are adopted. While predicting, since the dataset is unbalanced, technique bagging is used and the following data split is adopted.

The whole dataset is first splitted into 2 subsets, with one only containing no-rain-tomorrow samples and the other have-rain-tomorrow, labelled as No-Dataset and Yes-Dataset respectively. In Yes-Dataset, 30% are randomly picked for testing. In No-Dataset, we randomly pick as many samples as 30% of No-Dataset for testing. In this way, the testing dataset can be balanced. After the testing dataset are picked, 501 balanced training datasets are randomly generated from the samples remained. Models are trained on these balanced training datasets. Then technique bagging is applied on these fitted models: every model votes and the final prediction is the one gets most votes.

3.1. Linear Regression

In the whole dataset, all numeric variables cannot have negative values because of their physics meanings. Thus, linear regression is not applicable since linear regression may output a negative value.

3.2. Logistic Regression

501 training datasets are generated, and 501 logistic regressions are performed. The accuracy of these 501 logistic regressions is shown in Exhibit 9. The accuracy is between 74% and 76%. Bagging with all these logistic regressions, the accuracy is 75.32% (the red line in Exhibit 9). It seems that bagging achieves the average accuracy of all these logistic regression.

3.3. KNN

*Retrived from (https://en.wikipedia.org/wiki/Weather_forecasting)

Knn has a parameter called k, meaning how many nearest samples in the training dataset to be adopted for predicting. In this project, k is assigned as 101, 201, and 301 respectively. Since training knn is time consuming, we only generate 11 different training datasets and thus only fitting 11 different knns. Exhibit 10 shows that the accuracy of all these 33 knns are very close to 77%. After bagging, the accuracy for different k are also very close to 77%, which means these knns are very similar to each other. See Exhibit 11.

3.4. Decision Tree and Random Forest

Similar to what has been done in the former two parts, 501 training datasets are generated and 501 trees are fitted. The accuracy of these 501 trees are shown in Exhibit 12. From the histogram, the prediction accuracy of all these trees are between 71% and 75%. The majority is below 73%. Bagging with all these 501 trees, the accuracy is 74.37%. See Exhibit 13, the red line represents the accuracy after bagging. From these results, bagging seems to achieve an average prediction performance of all the trees.

Random forest is generated from decision trees and has already used bagging. The difference from the technique used above is that random forest not only uses different training datasets, but also different subsets of indicators. So fitting 501 random forests is time consuming and only 10 random forests are fitted. The accuracy of these 10 random forests are shown in Exhibit 14. The accuracy is very stable, centering at 78.8%.

4. Conclusions

This project focuses on predicting whether it will rain tomorrow. The results of auto regression show that weather conditions in the past may be effective to predict whether it will rain tomorrow. We further confirm that using machine learning techniques, such as logistic regression, knn, decision tree and random forest, the weather condition can be predicted. In this project, random forest achieves the high prediction accuracy, roughly 78.8%. All the models used in this project can achieve a prediction accuracy higher than 74%. Besides, from the results of knn, logistic regression and decision tree, bagging can achieve a more stable prediction accuracy than using one model since bagging indeed achieves the average accuracy of all the models used for bagging.

*Retrived from (https://en.wikipedia.org/wiki/Weather_forecasting)

Appendix

The dataset is from kaggle: https://www.kaggle.com/jsphyg/weather-dataset-rattle-package

The meaning of each variable in this dataset:

1. Date: The date of the observation

2. Location: The common name of the location of the weather station

3. MinTemp: The minimum temperature in degrees celsius

4. MaxTemp: The maximum temperature in degrees celsius

5. Rainfall: The amount of rainfall recorded for the day in mm

6. Evaporation: The so-called Class A pan evaporation (mm) in the 24 hours to 9am

7. Sunshine: The number of hours of bright sunshine in the day.

8. WindGustDir: The direction of the strongest wind gust in the 24 hours to midnight

9. WindGustSpeed: The speed (km/h) of the strongest wind gust in the 24 hours to midnight

10. WindDir9am: Direction of the wind at 9am

11. WindDir3pm: Direction of the wind at 3pm

12. WindSpeed9am: Wind speed (km/hr) averaged over 10 minutes prior to 9am

13. WindSpeed3pm: Wind speed (km/hr) averaged over 10 minutes prior to 3pm

14. Humidity9am: Humidity (percent) at 9am

15. Humidity3pm: Humidity (percent) at 3pm

16. Pressure9am: Atmospheric pressure (hpa) reduced to mean sea level at 9am

17. Pressure3pm: Atmospheric pressure (hpa) reduced to mean sea level at 3pm

*Retrived from (https://en.wikipedia.org/wiki/Weather_forecasting)

18. 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.

19. Cloud3pm: Fraction of sky obscured by cloud (in "oktas": eighths) at 3pm. See Cload9am for a description of the values

20. Temp9am: Temperature (degrees C) at 9am

21. Temp3pm: Temperature (degrees C) at 3pm

22. RainTodayBoolean: 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0

23. RISK_MM: The amount of next day rain in mm. Used to create response variable RainTomorrow. A kind of measure of the "risk".

24. RainTomorrow is the target variable. Did it rain tomorrow?

Exhibit 1

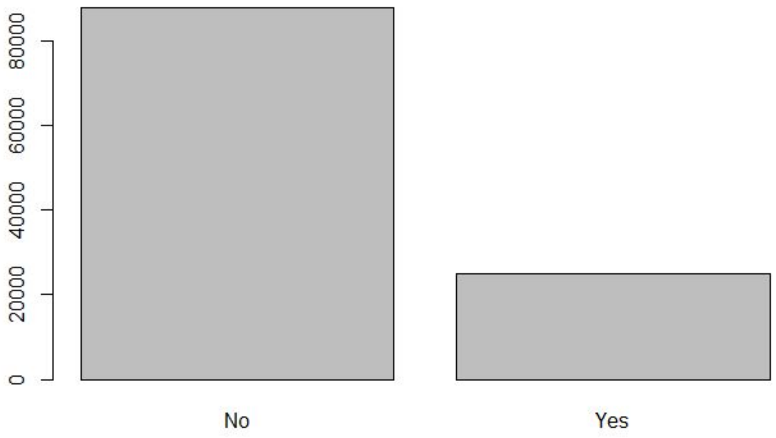


Exhibit 2

*Retrived from (https://en.wikipedia.org/wiki/Weather_forecasting)

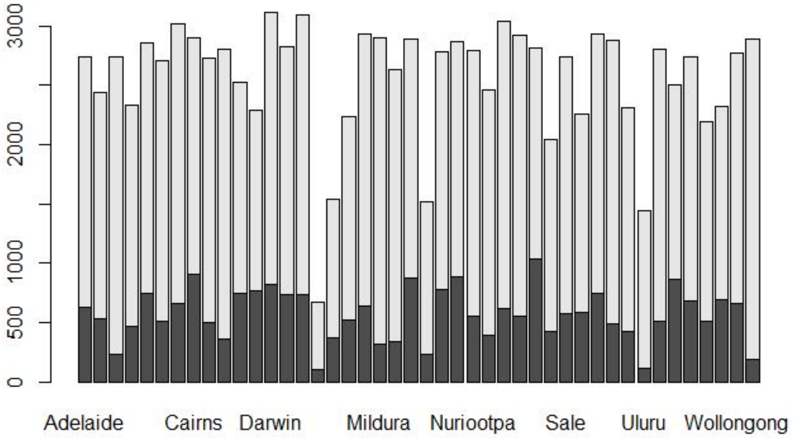
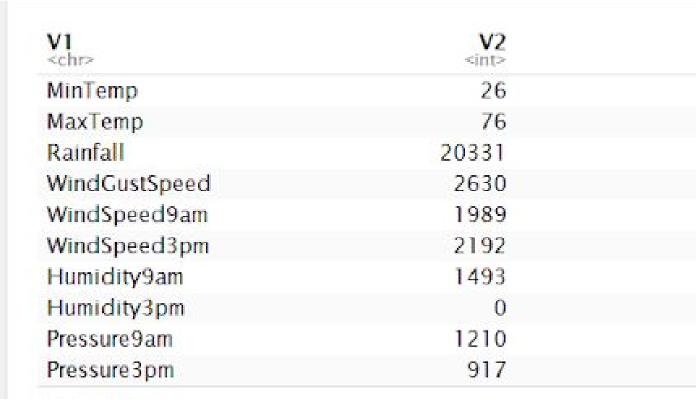


Exhibit 3



V1 <chr>	V2 <int>
MinTemp	26
MaxTemp	76
Rainfall	20331
WindGustSpeed	2630
WindSpeed9am	1989
WindSpeed3pm	2192
Humidity9am	1493
Humidity3pm	0
Pressure9am	1210
Pressure3pm	917

Exhibit 4

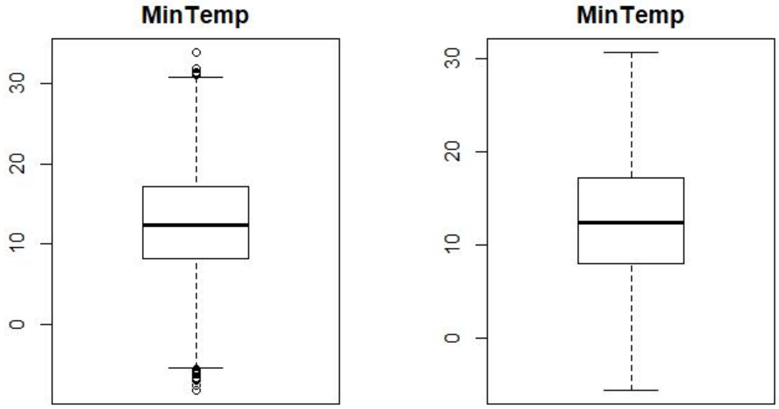


Exhibit 5

*Retrived from (https://en.wikipedia.org/wiki/Weather_forecasting)

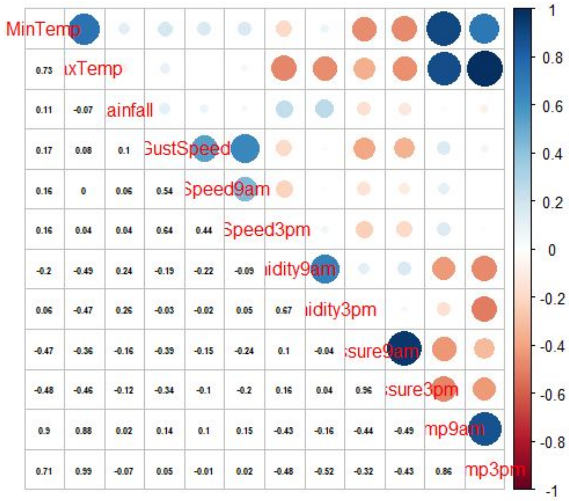


Exhibit 6

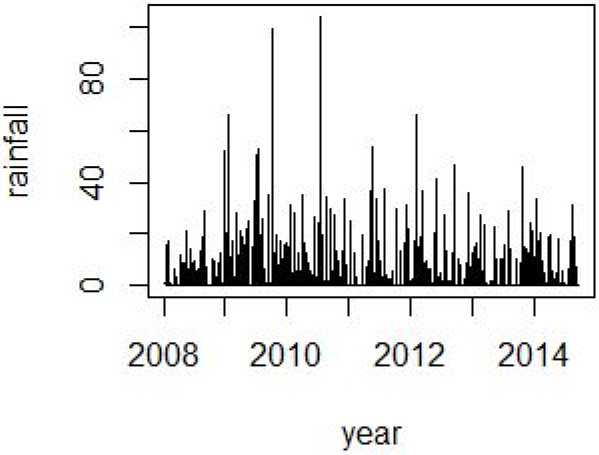


Exhibit 7

*Retrived from (https://en.wikipedia.org/wiki/Weather_forecasting)

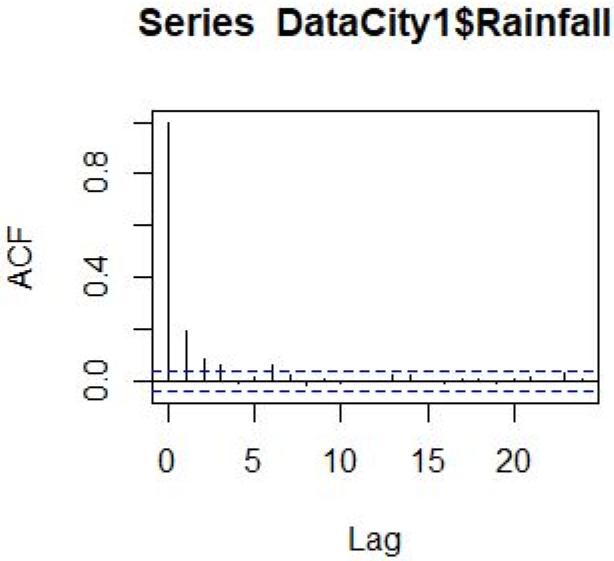


Exhibit 8

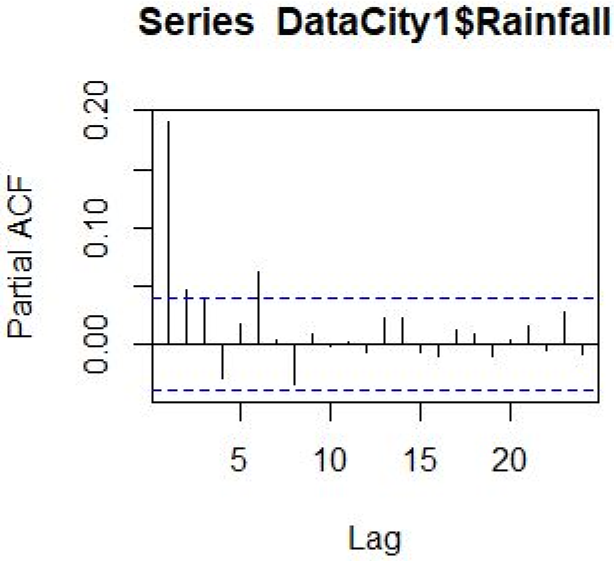


Exhibit 9

*Retrived from (https://en.wikipedia.org/wiki/Weather_forecasting)

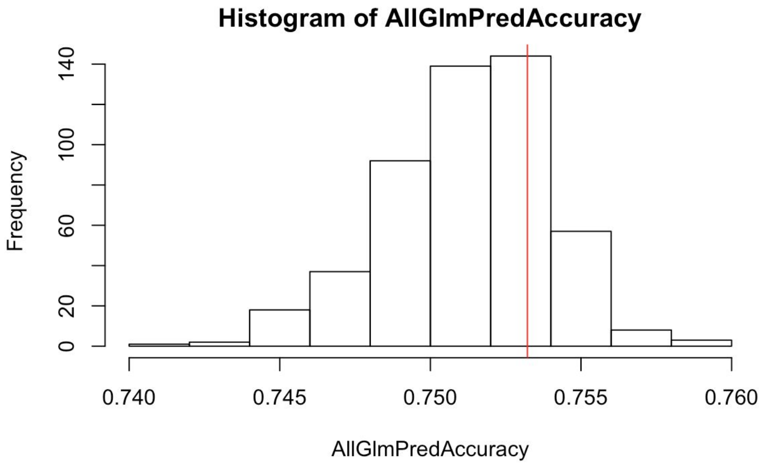


Exhibit 10

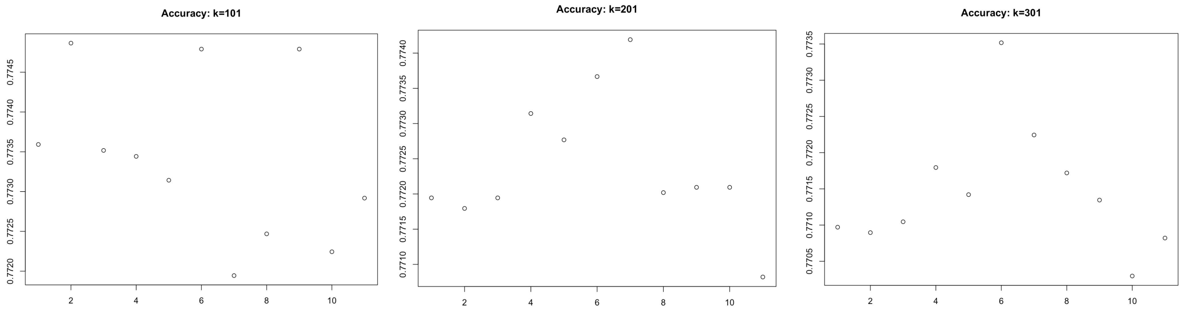


Exhibit 11

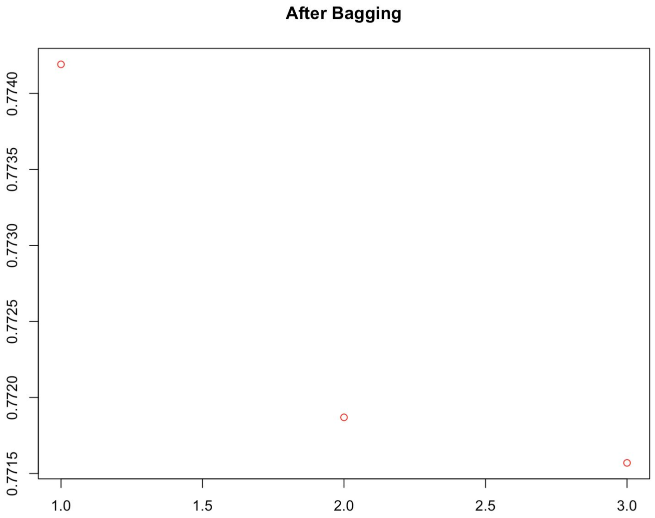


Exhibit 12

*Retrived from (https://en.wikipedia.org/wiki/Weather_forecasting)

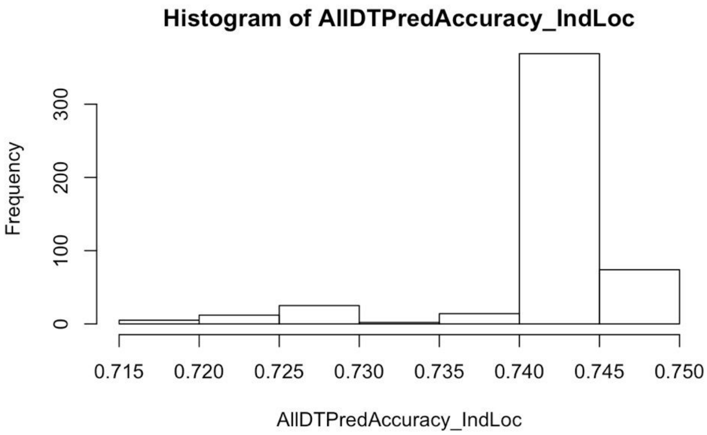


Exhibit 13

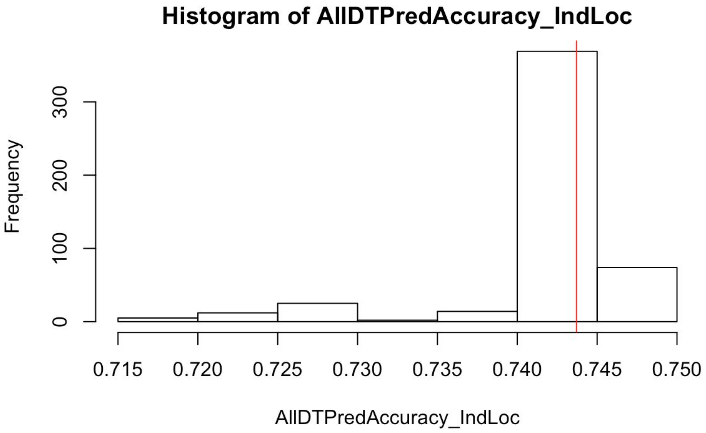
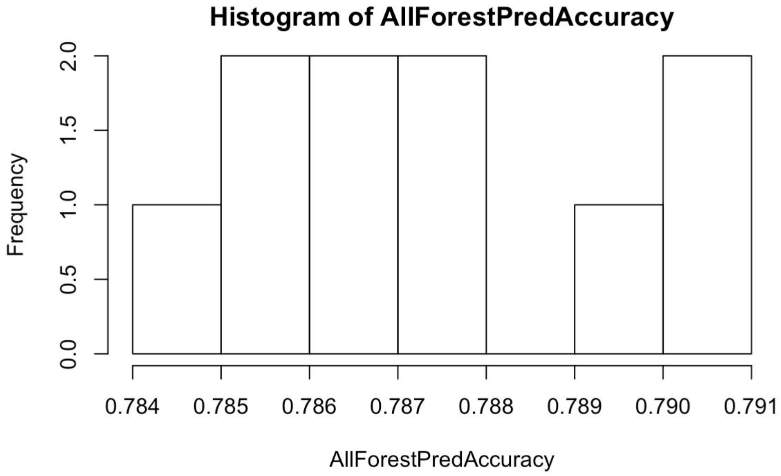


Exhibit 14



*Retrived from (https://en.wikipedia.org/wiki/Weather_forecasting)