Ryan Cosgrove
Ravinit Chand
Eric Kye
Revanth Rao

Predicting Rainfall In Australia

Introduction and Background:

Weather is something everyone should account for going into their day. The weather affects much of our lives, ranging from commutes to daily activities. For our project, we will look into the weather in Australia, and build statistical models to predict whether it will rain the next day given various information about a specific date and location.

The dataset we used to predict rainfall in Australia comes from the website Kaggle (linked here), and is fairly comprehensive with over 20 variables describing information from multiple cities within Australia. Additionally, the data covers the years 2008-2017, meaning the sample size is very large. Some of the predictor variables used are the humidity, wind, pressure, rainfall, and temperature for a specific day, and the response variable is whether it will rain the next day or not.

Data Description:

Our data set consists of 23 variables, and features a mix of quantitative, categorical, and binary variables. The quantitative variables are maximum and minimum temperature, rainfall, evaporation, sunshine, wind gust speed, and measurements of wind speed, humidity, pressure, cloud cover, and temperature at both 9 AM and 3 PM. The categorical variables are date, location, wind gust direction, and wind direction at 9 AM and 3 PM. Finally, the binary variables are RainToday and RainTomorrow, both of which take values of “Yes” or “No” to indicate whether it rained on a given day or following day.

When preprocessing the data, we decided to remove columns from the data which we believed were not needed in building a prediction model for rain. These columns were date, location, wind gust direction, and the two wind direction variables. Additionally, after dealing with difficulties in our analysis due to NA values, we decided to remove NA values from the data set. We believed this was a justifiable decision because the overall proportion of data that was null was around 10%, and many of the predictors had fewer than 10% null values, so we felt that we were not losing a substantial amount of data by omitting null values. Using this preprocessed data, we then started our project by performing exploratory data analysis.

Exploratory Data Analysis:

After preprocessing the data set, we began our analysis by exploring the data to spot potential trends or outliers. To start, we obtained summary statistics for the data set, shown below in Table 1:

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
MinTemp	56420	13	6.4	-6.7	8.6	18	31
MaxTemp	56420	24	7	4.1	19	30	48
Rainfall	56420	2.1	7	0	0	0.6	206
Evaporation	56420	5.5	3.7	0	2.8	7.4	81
Sunshine	56420	7.7	3.8	0	5	11	14
WindGustSpeed	56420	41	13	9	31	48	124
WindSpeed9am	56420	16	8.3	2	9	20	67
WindSpeed3pm	56420	20	8.5	2	13	26	76
Humidity9am	56420	66	19	0	55	79	100
Humidity3pm	56420	50	20	0	35	63	100
Pressure9am	56420	1017	6.9	980	1013	1022	1040
Pressure3pm	56420	1015	6.9	977	1010	1019	1039
Cloud9am	56420	4.2	2.8	0	1	7	8
Cloud3pm	56420	4.3	2.6	0	2	7	9
Temp9am	56420	18	6.6	-0.7	13	23	39
Temp3pm	56420	23	6.8	3.7	17	28	46
RainToday	56420						
... No	43958	78%					
... Yes	12462	22%					
RainTomorrow	56420						
... No	43993	78%					
... Yes	12427	22%					

Table 1: Summary Statistics

Looking at the summary statistics, it appears that some variables have large values that should be considered outliers. For example, the Rainfall variable has a mean of 2.1 and a standard deviation of 7, but the maximum value is 206, while the Evaporation variable has a mean of 5.5 and a standard deviation of 3.7, but the maximum value is 81. As a result, we can see that there was a lot of variation in weather patterns throughout Australia. Additionally, the summary statistics show that it did not rain 78% of the days and rained 22% of the days. With this information in mind, we set a baseline accuracy of 78% for our models, as this is the accuracy a naive prediction of no rain every day would achieve.

Continuing our exploration of the data, we created a correlation plot, shown in Figure 1:

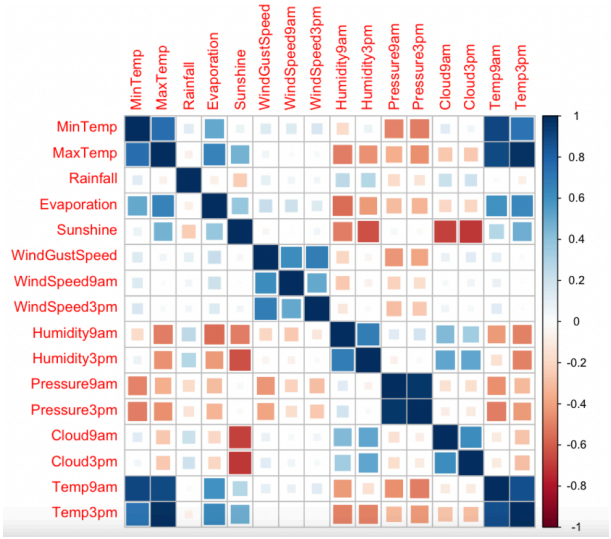

Figure 1: Correlation Plot of Predictor Variables

In this plot, larger boxes indicate higher correlation, both negative and positive, darker blue indicates a more positive correlation, darker red indicates a more negative correlation, and white indicates weak or no correlation. As shown, a lot of variables have somewhat weak correlations with other variables, and many of the strong correlations are intuitive. For example, the correlations between the temperature variables are all near 1, which makes sense as they are all measuring slight variations of temperature on a given day and location. Another example of intuitive correlations are the correlations between Sunshine and the two cloud cover variables, which are both close to -1. This relationship makes logical sense, as you would expect more sunshine on days with fewer clouds, and less sunshine on days with more clouds.

Our final piece of exploratory analysis was creating a histogram of each predictor variable to get a sense of their distributions. These histograms are shown below in Figure 2:

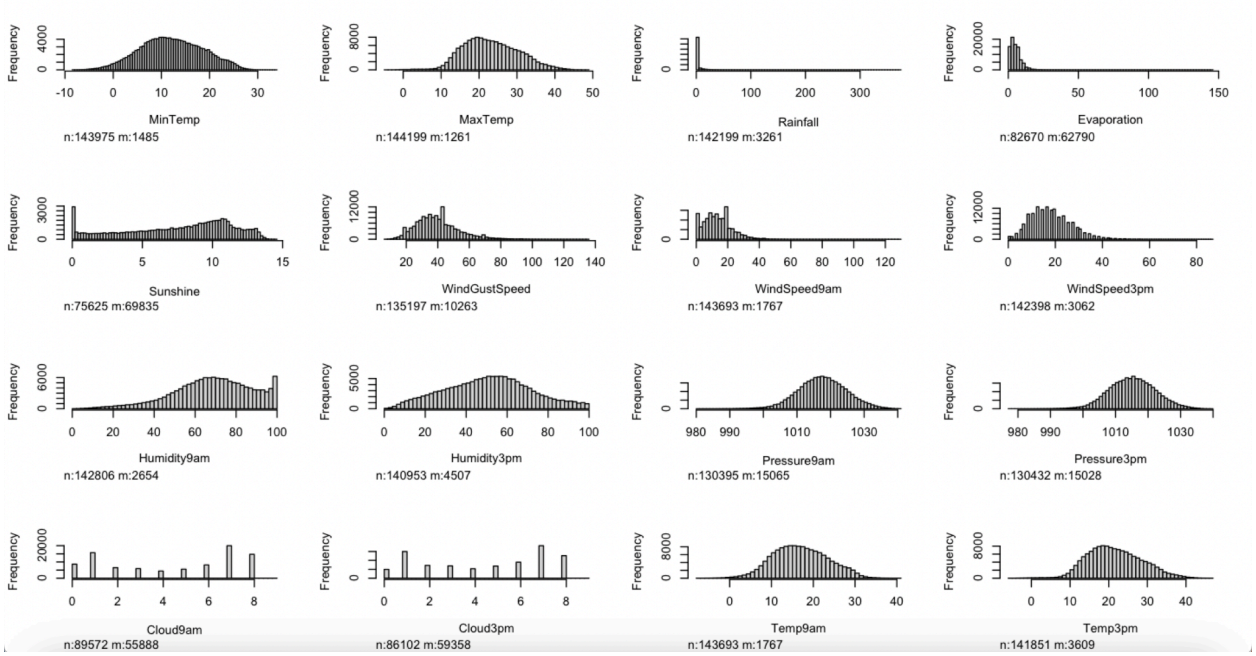

Figure 2: Histograms of Predictor Variables

For many of the variables, we can see that the distribution looks fairly smooth and roughly symmetric. However, certain variables show interesting distributions. For example, as mentioned earlier, rainfall and evaporation have outliers, so those histograms appear to be skewed by the presence of these outliers. Moreover, we can see that the two cloud cover variables are discrete, as those histograms only show frequencies at whole numbers. Finally, another interesting variable to observe is humidity at 9 AM, which appears to be skewed left with many days where the humidity was 100 percent.

Methodology:

Given that the goal of our project was to build prediction models for rain, we decided to split our data into a training data set, which was used to train each prediction model, and a test

data set, which was used to evaluate the accuracy of the models. We decided to use 60% of the data to train the model and 40% of the data to test the model, and chose data from the years 2008–2013 as our training set and data from 2014–2017 as our test set. The reason we split the data this way is because the data set involved time series data, and we wanted to avoid using observations from future years to predict rain on a given day, so we opted to use the first six years for training and the last four for testing.

After splitting the data set, we used four methods to build prediction models for rain. The first was linear and quadratic discriminant analysis, or LDA and QDA. Both of these models are commonly used for classification problems because they use decision boundaries to separate the data into classes. The main difference between LDA and QDA is that LDA has the same covariance matrix in each class while QDA has a different covariance matrix in each class. The second method we used was logistic regression. This is a popular classification model for binary response variables, as is the case with our response variable RainTomorrow, because it provides probabilities between 0 and 1 for the predictions. The third method we used was lasso and ridge regression models, which use shrinkage penalties to constrain the coefficients on the predictors in an effort to improve prediction accuracy as compared to a linear regression model. For both the lasso and ridge regression models, we used cross-validation to select the optimal tuning parameter λ. Finally, the fourth method we used was a Random Forest, which is a model that grows many decision trees on the training data set, then combines them to make predictions. The Random Forest method is particularly useful because it is a more advanced version of decision trees and typically yields higher prediction accuracy, which would allow us to make better predictions on our test data. For each method utilized, we used RainTomorrow as the response variable and every other variable in the preprocessed data set as predictors in order to have consistency across the different models.

Main Results:

The QDA and LDA models displayed fairly similar results, as the QDA model had a prediction accuracy of 85.4% and the LDA model had a prediction accuracy of 83.7%. Figure 3 examines the LDA model more closely:

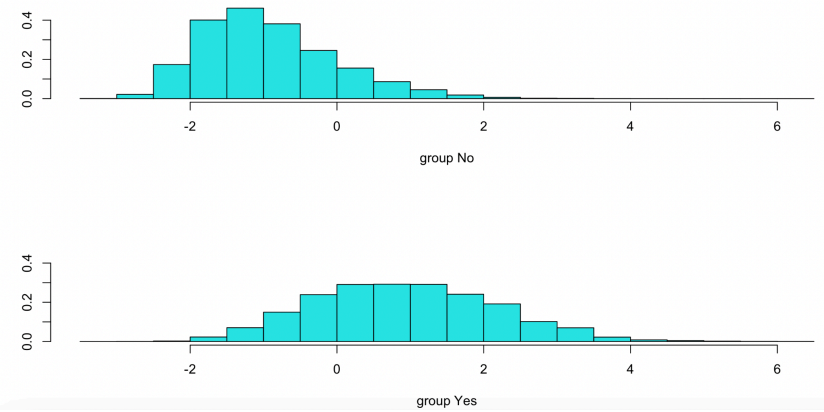

Figure 3: Histogram of Linear Discriminant Scores

Figure 3 shows histograms for both classes, labeled Group No and Group Yes, with the linear discriminant scores on the x-axis. As shown, the two graphs don't exhibit a substantial overlap, as most of the linear discriminant scores are below 0 in the Group No histogram, while most of the linear discriminant scores are above 0 in the Group Yes graph. This implies that the LDA model did a fairly good job of discerning between "Yes" and "No" for whether it would rain tomorrow based on the predictors in the data set.

For our logistic regression model, we used a cutoff point of 0.5 for predictions, meaning predictions above 0.5 were considered "Yes" and predictions of 0.5 or below were considered "No". We experimented with varying cutoffs, but each produced extremely similar prediction accuracies, so we opted to use 0.5, which is the midpoint for probabilities. This approach produced a prediction accuracy of 85.5%, an accuracy rate that is quite close to the QDA and LDA accuracy rates. To further examine the accuracy, we plotted a Receiver Operating Characteristic (ROC) Curve in Figure 4:

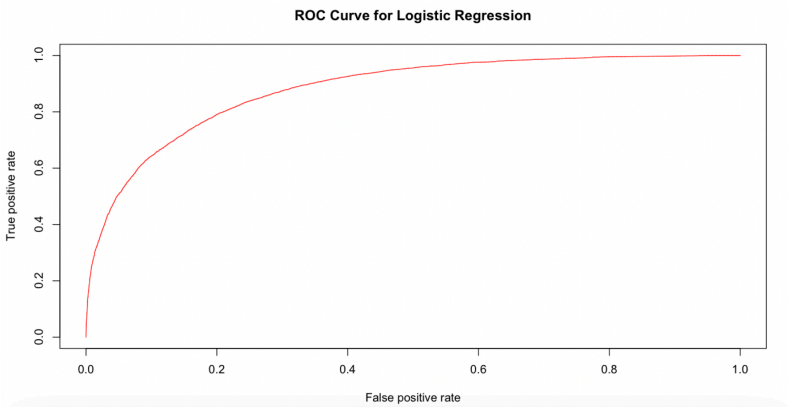

Figure 4: Receiver Operating Characteristic Curve for Logistic Regression Model

The ROC curve shows an initial steep slope before flattening out, suggesting that there is a high true positive rate for predictions compared to false positive rate and that the model does a good job of distinguishing between the two classes. Additionally, this shape leads to larger areas under the curve, meaning that the model is fairly accurate in predicting rain tomorrow.

For our lasso and ridge regression models, we first used cross validation to produce the value of λ for our models. By doing this, we obtained the value $\lambda = 0.00018$ for the lasso model and the value $\lambda = 0.019$ for the ridge model. Afterward, we ran the regressions and obtained the coefficients and accuracies. Both models had very close accuracies, as the lasso regression model had a prediction accuracy of 85.5%, while the ridge regression model had a prediction accuracy of 85.1%. Additionally, for each model, while many coefficients were shrunk to very small values, none of the coefficients were shrunk down to zero. This would indicate that all of the predictors chosen had relevance and were needed for creating predictions. To look into the chosen λ values for these models, we plotted the binomial deviance for varying $\log(\lambda)$ values for both the lasso and ridge models in Figure 5:

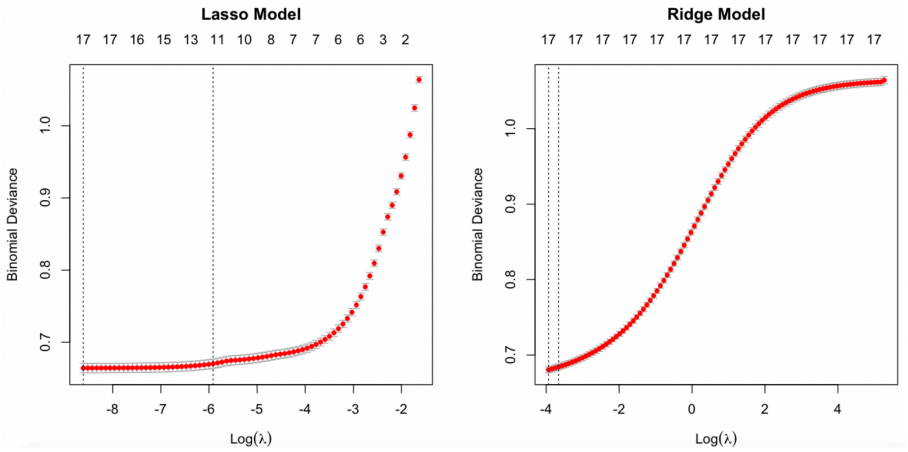

Figure 5: Lasso and Ridge Model Plots

Both plots above show that the binomial deviance, which measures how well the model fits to the data, is minimized for small values of λ, which are depicted by the far left dashed line in each plot. Given this information and the knowledge that none of the coefficients were shrunk to zero in either model, we could conclude that the optimal models for lasso and ridge do not have large shrinkage penalties.

The final model tested was the Random Forest model, which produced a prediction accuracy of 85.7%. This was the highest figure for prediction accuracy that we obtained, but it was still very close to the accuracy produced by the other models. In addition to accuracy, we also created Random Forest variable importance plots, shown below in Figure 6:

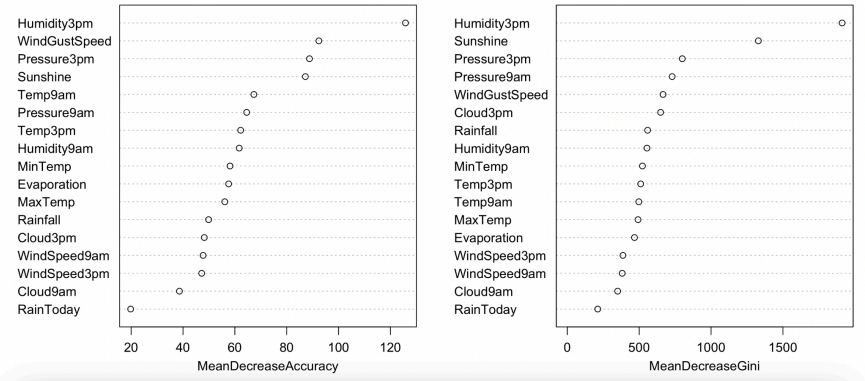

Figure 6: Variable Importance Plots for Random Forest

The two plots above show the mean decrease in model accuracy and the mean decrease in Gini. These measures aim to measure how significant a predictor variable is for model performance. For each plot, the variables humidity, sunshine, pressure at 3 PM, and wind gust speed are among the top predictors in MeanDecreaseAccuracy and MeanDecreaseGini, which indicates that these variables appear to be most critical for our Random Forest model in predicting rain.

Discussion (Strengths and Weaknesses):

While working on this project, we found some strengths in our analysis that made our conclusions more accurate, one being that the large dataset allowed us to have robust training and test sets, enabling the models to be more precise due to the bigger sample sizes. Also, model performance was fairly consistent across every method as prediction accuracy ranged from 83–86%, and since it didn't rain 78% of the days, every model comfortably beat a naive prediction of no rain every day. Finally, using a variety of methods allowed us to get a good sense of a reasonable prediction accuracy for the data. The limitations that we came across were that the dataset had many NA values, leading to some holes in the data, and some of the methods, such as the Random Forest, are somewhat computationally intensive and expensive.

We also discovered some improvements and next steps to continue adding to this project which would improve our analysis and conclusions. One addition that we could make is to use other methods such as polynomial regression, naive Bayes, or more advanced decision tree methods to see if differing results are found or if our conclusions stay the same. We could also replace the NA values found throughout the data set with the mean, median, or other values rather than simply removing them from the data. In addition, to refine our model, we could use best subset selection or forward and backward stepwise selection in an effort to potentially remove unnecessary predictors and improve our models. Finally, we could also experiment with different sizes for test and training sets to see if the results are replicable.

Conclusion:

In conclusion, all of our methods had fairly similar accuracy in predicting rain, and we believe that all of them were reliable. From the results of the models, we conclude that this data allows for good classification models due to the high accuracy. The consistency in performance across different models reinforces the robustness of our dataset and the effectiveness of our selected predictor variables.

Ultimately, even though the Random Forest method had the highest accuracy rate, we may want to use other simpler methods because the Random Forest model is computationally expensive and takes a lot of computer power to implement. This makes logistic regression and LDA particularly attractive options, as they offer ease of interpretation and require significantly less computational resources.

Given the similar accuracy in model performance, it may be preferable to use models like logistic regression or LDA as they are simple and easy to interpret. Additionally, exploring further refinements and incorporating more advanced data handling techniques, such as imputation for missing values, could enhance the overall predictive performance and robustness of these models. Implementing best subset selection or stepwise selection methods could also streamline the model by identifying the most significant predictors.

This study lays a solid foundation for future research and practical applications in weather prediction. By leveraging the insights gained from our analysis, future work can continue to improve the accuracy and efficiency of rainfall prediction models, contributing to better weather forecasting and planning.

Dataset Used:

https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package