Source: SBS News

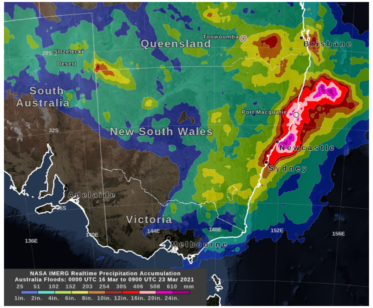
Source: PaymentsJorunal

Source: USA Today

# Rainfall in Australia

Source: Bookmundi

Ravinit Chand, Ryan Cosgrove,
Eric Kye, Revanth Rao




Source: NASA

# Background

Source: National Museum Of Australia

- We got our dataset from Kaggle
  - (link: https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package)
- Various cities in Australia
- Dates ranging from 2008 to 2017
- Weather conditions: temperature, rainfall, wind, pressure, and humidity
- The response variable is if it rains tomorrow

# Rain in Australia

Predict next-day rain in Australia

Data Card    Code (650)    Discussion (21)    Suggestions (0)

## About Dataset

### Context

Predict **next-day rain** by training classification models on the target variable **RainTomorrow**.

### Content

This dataset contains about 10 years of daily weather observations from many locations across Australia.

**RainTomorrow is the target variable to predict. It means -- did it rain the next day, Yes or No? This column is Yes if the rain for that day was 1mm or more.**

### Source & Acknowledgements

Observations were drawn from numerous weather stations. The daily observations are available from http://www.bom.gov.au/climate/data.

An example of latest weather observations in Canberra: http://www.bom.gov.au/climate/dwo/IDCJDW2801.latest.shtml

Definitions adapted from http://www.bom.gov.au/climate/dwo/IDCJDW0000.shtml

Data source: http://www.bom.gov.au/climate/dwo/ and http://www.bom.gov.au/climate/data.



**Usability** ⓘ

10.00

**License**

Other (specified in description)

**Expected update frequency**

Never

**Tags**

Earth and Nature

Classification

Binary Classification

Weather and Climate

**Source: Kaggle**

# Data Description

- The data consists of a combination of quantitative and categorical variables
- There are 145,460 rows being the dates and 23 columns being the variables
- The temperature, rainfall, wind speed, humidity, pressure, and cloud cover variables in the data set are some quantitative variables
- The location, wind gust direction, wind direction are some categorical variables
- Rain today and rain tomorrow are both binary variables taking values of “Yes” or “No”

| | Date | Location | MinTemp | MaxTemp | Rainfall | Evaporation | Sunshine | WindGustDir | WindGustSpeed | WindDir9am | WindDir3pm | WindSpeed9am | WindSpeed3pm | Humidity9am |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2008-12-01 | Albury | 13.4 | 22.9 | 0.6 | NA | NA | W | 44 | W | WNW | 20 | 24 | 71 |
| 2 | 2008-12-02 | Albury | 7.4 | 25.1 | 0.0 | NA | NA | WNW | 44 | NNW | WSW | 4 | 22 | 44 |
| 3 | 2008-12-03 | Albury | 12.9 | 25.7 | 0.0 | NA | NA | WSW | 46 | W | WSW | 19 | 26 | 38 |
| 4 | 2008-12-04 | Albury | 9.2 | 28.0 | 0.0 | NA | NA | NE | 24 | SE | E | 11 | 9 | 45 |
| 5 | 2008-12-05 | Albury | 17.5 | 32.3 | 1.0 | NA | NA | W | 41 | ENE | NW | 7 | 20 | 82 |
| 6 | 2008-12-06 | Albury | 14.6 | 29.7 | 0.2 | NA | NA | WNW | 56 | W | W | 19 | 24 | 55 |
| 7 | 2008-12-07 | Albury | 14.3 | 25.0 | 0.0 | NA | NA | W | 50 | SW | W | 20 | 24 | 49 |
| 8 | 2008-12-08 | Albury | 7.7 | 26.7 | 0.0 | NA | NA | W | 35 | SSE | W | 6 | 17 | 48 |
| 9 | 2008-12-09 | Albury | 9.7 | 31.9 | 0.0 | NA | NA | NNW | 80 | SE | NW | 7 | 28 | 42 |
| 10 | 2008-12-10 | Albury | 13.1 | 30.1 | 1.4 | NA | NA | W | 28 | S | SSE | 15 | 11 | 58 |
| 11 | 2008-12-11 | Albury | 13.4 | 30.4 | 0.0 | NA | NA | N | 30 | SSE | ESE | 17 | 6 | 48 |
| 12 | 2008-12-12 | Albury | 15.9 | 21.7 | 2.2 | NA | NA | NNE | 31 | NE | ENE | 15 | 13 | 89 |
| 13 | 2008-12-13 | Albury | 15.9 | 18.6 | 15.6 | NA | NA | W | 61 | NNW | NNW | 28 | 28 | 76 |
| 14 | 2008-12-14 | Albury | 12.6 | 21.0 | 3.6 | NA | NA | SW | 44 | W | SSW | 24 | 20 | 65 |
| 15 | 2008-12-15 | Albury | 8.4 | 24.6 | 0.0 | NA | NA | NA | NA | S | WNW | 4 | 30 | 57 |
| 16 | 2008-12-16 | Albury | 9.8 | 27.7 | NA | NA | NA | WNW | 50 | NA | WNW | NA | 22 | 50 |
| 17 | 2008-12-17 | Albury | 14.1 | 20.9 | 0.0 | NA | NA | ENE | 22 | SSW | E | 11 | 9 | 69 |
| 18 | 2008-12-18 | Albury | 13.5 | 22.9 | 16.8 | NA | NA | W | 63 | N | WNW | 6 | 20 | 80 |
| 19 | 2008-12-19 | Albury | 11.2 | 22.5 | 10.6 | NA | NA | SSE | 43 | WSW | SW | 24 | 17 | 47 |
| 20 | 2008-12-20 | Albury | 9.8 | 25.6 | 0.0 | NA | NA | SSE | 26 | SE | NNW | 17 | 6 | 45 |
| 21 | 2008-12-21 | Albury | 11.5 | 29.3 | 0.0 | NA | NA | S | 24 | SE | SE | 9 | 9 | 56 |
| 22 | 2008-12-22 | Albury | 17.1 | 33.0 | 0.0 | NA | NA | NE | 43 | NE | N | 17 | 22 | 38 |
| 23 | 2008-12-23 | Albury | 20.5 | 31.8 | 0.0 | NA | NA | WNW | 41 | W | W | 19 | 20 | 54 |
| 24 | 2008-12-24 | Albury | 15.3 | 30.9 | 0.0 | NA | NA | N | 33 | ESE | NW | 6 | 13 | 55 |
| 25 | 2008-12-25 | Albury | 12.6 | 32.4 | 0.0 | NA | NA | W | 43 | E | W | 4 | 19 | 49 |
| 26 | 2008-12-26 | Albury | 16.2 | 33.9 | 0.0 | NA | NA | WSW | 35 | SE | WSW | 9 | 13 | 45 |
| 27 | 2008-12-27 | Albury | 16.9 | 33.0 | 0.0 | NA | NA | WSW | 57 | NA | W | 0 | 26 | 41 |
| 28 | 2008-12-28 | Albury | 20.1 | 32.7 | 0.0 | NA | NA | WNW | 48 | N | WNW | 13 | 30 | 56 |
| 29 | 2008-12-29 | Albury | 19.7 | 27.2 | 0.0 | NA | NA | WNW | 46 | NW | WSW | 19 | 30 | 49 |
| 30 | 2008-12-30 | Albury | 12.5 | 24.2 | 1.2 | NA | NA | WNW | 50 | WSW | SW | 11 | 22 | 78 |
| 31 | 2008-12-31 | Albury | 12.0 | 24.4 | 0.8 | NA | NA | W | 39 | WNW | WNW | 17 | 17 | 48 |
| 32 | 2009-01-01 | Albury | 11.3 | 26.5 | 0.0 | NA | NA | WNW | 56 | W | WNW | 19 | 31 | 46 |

Showing 1 to 32 of 145,460 entries, 23 total columns

| WindSpeed3pm | Humidity9am | Humidity3pm | Pressure9am | Pressure3pm | Cloud9am | Cloud3pm | Temp9am | Temp3pm | RainToday | RainTomorrow |
|---|---|---|---|---|---|---|---|---|---|---|
| 24 | 71 | 22 | 1007.7 | 1007.1 | 8 | NA | 16.9 | 21.8 | No | No |
| 22 | 44 | 25 | 1010.6 | 1007.8 | NA | NA | 17.2 | 24.3 | No | No |
| 26 | 38 | 30 | 1007.6 | 1008.7 | NA | 2 | 21.0 | 23.2 | No | No |
| 9 | 45 | 16 | 1017.6 | 1012.8 | NA | NA | 18.1 | 26.5 | No | No |
| 20 | 82 | 33 | 1010.8 | 1006.0 | 7 | 8 | 17.8 | 29.7 | No | No |
| 24 | 55 | 23 | 1009.2 | 1005.4 | NA | NA | 20.6 | 28.9 | No | No |
| 24 | 49 | 19 | 1009.6 | 1008.2 | 1 | NA | 18.1 | 24.6 | No | No |
| 17 | 48 | 19 | 1013.4 | 1010.1 | NA | NA | 16.3 | 25.5 | No | No |
| 28 | 42 | 9 | 1008.9 | 1003.6 | NA | NA | 18.3 | 30.2 | No | Yes |
| 11 | 58 | 27 | 1007.0 | 1005.7 | NA | NA | 20.1 | 28.2 | Yes | No |
| 6 | 48 | 22 | 1011.8 | 1008.7 | NA | NA | 20.4 | 28.8 | No | Yes |
| 13 | 89 | 91 | 1010.5 | 1004.2 | 8 | 8 | 15.9 | 17.0 | Yes | Yes |
| 28 | 76 | 93 | 994.3 | 993.0 | 8 | 8 | 17.4 | 15.8 | Yes | Yes |
| 20 | 65 | 43 | 1001.2 | 1001.8 | NA | 7 | 15.8 | 19.8 | Yes | No |
| 30 | 57 | 32 | 1009.7 | 1008.7 | NA | NA | 15.9 | 23.5 | No | NA |
| 22 | 50 | 28 | 1013.4 | 1010.3 | 0 | NA | 17.3 | 26.2 | NA | No |
| 9 | 69 | 82 | 1012.2 | 1010.4 | 8 | 1 | 17.2 | 18.1 | No | Yes |
| 20 | 80 | 65 | 1005.8 | 1002.2 | 8 | 1 | 18.0 | 21.5 | Yes | Yes |
| 17 | 47 | 32 | 1009.4 | 1009.7 | NA | 2 | 15.5 | 21.0 | Yes | No |
| 6 | 45 | 26 | 1019.2 | 1017.1 | NA | NA | 15.8 | 23.2 | No | No |
| 9 | 56 | 28 | 1019.3 | 1014.8 | NA | NA | 19.1 | 27.3 | No | No |
| 22 | 38 | 28 | 1013.6 | 1008.1 | NA | 1 | 24.5 | 31.6 | No | No |
| 20 | 54 | 24 | 1007.8 | 1005.7 | NA | NA | 23.8 | 30.8 | No | No |
| 13 | 55 | 23 | 1011.0 | 1008.2 | 5 | NA | 20.9 | 29.0 | No | No |
| 19 | 49 | 17 | 1012.9 | 1010.1 | NA | NA | 21.5 | 31.2 | No | No |
| 13 | 45 | 19 | 1010.9 | 1007.6 | NA | 1 | 23.2 | 33.0 | No | No |
| 26 | 41 | 28 | 1006.8 | 1003.6 | NA | 1 | 26.6 | 31.2 | No | No |
| 30 | 56 | 15 | 1005.2 | 1001.7 | NA | NA | 24.6 | 32.1 | No | No |
| 30 | 49 | 22 | 1004.8 | 1004.2 | NA | NA | 21.6 | 26.1 | No | Yes |
| 22 | 78 | 70 | 1005.6 | 1003.4 | 8 | 8 | 12.5 | 18.2 | Yes | No |
| 17 | 48 | 28 | 1006.1 | 1005.1 | 1 | NA | 16.9 | 22.7 | No | No |
| 31 | 46 | 26 | 1004.5 | 1003.2 | NA | NA | 19.7 | 25.7 | No | No |

# Pre-Processing Steps and Project Focus

- Pre-processing steps:
  - Our data included NA values within multiple predictor variables which caused for more difficulty in analysis, leading us to remove NA values from the data
  - We also removed columns that were found to be not useful to our analysis. This included the date, location, wind gust direction, and wind direction

- Our project attempts to predict whether it will rain tomorrow based on the other variables remaining in the data set

# Exploratory Analysis

- Summary statistics for the variables used in the models are shown on the right

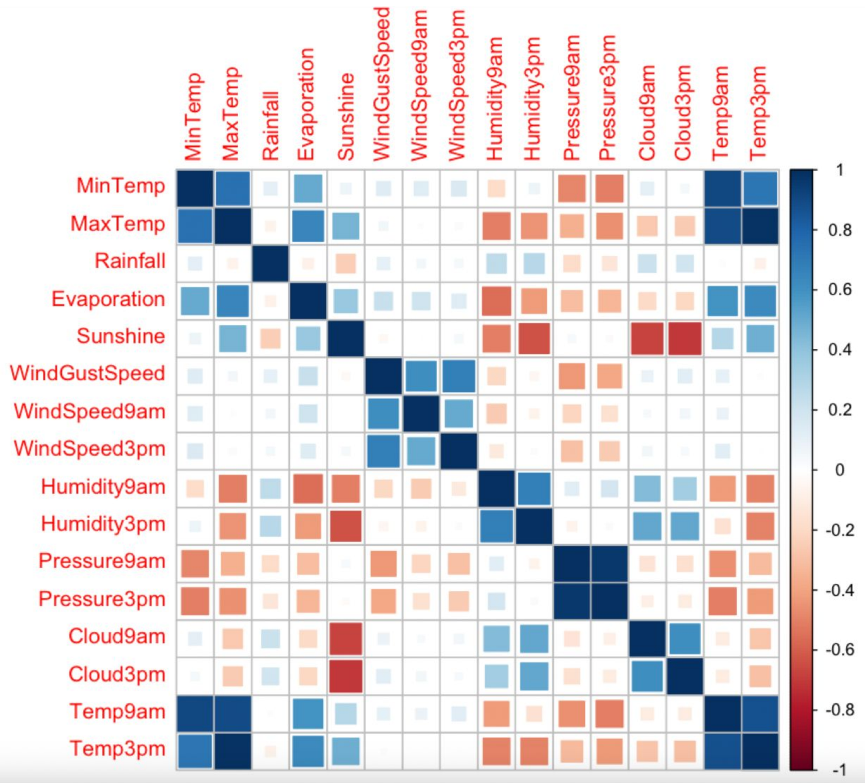
- It appears that some variables have large values that should be considered outliers (ex: Rainfall, Evaporation)

- In the data, we can see that it didn't rain 78% of the days and rained 22% of the days

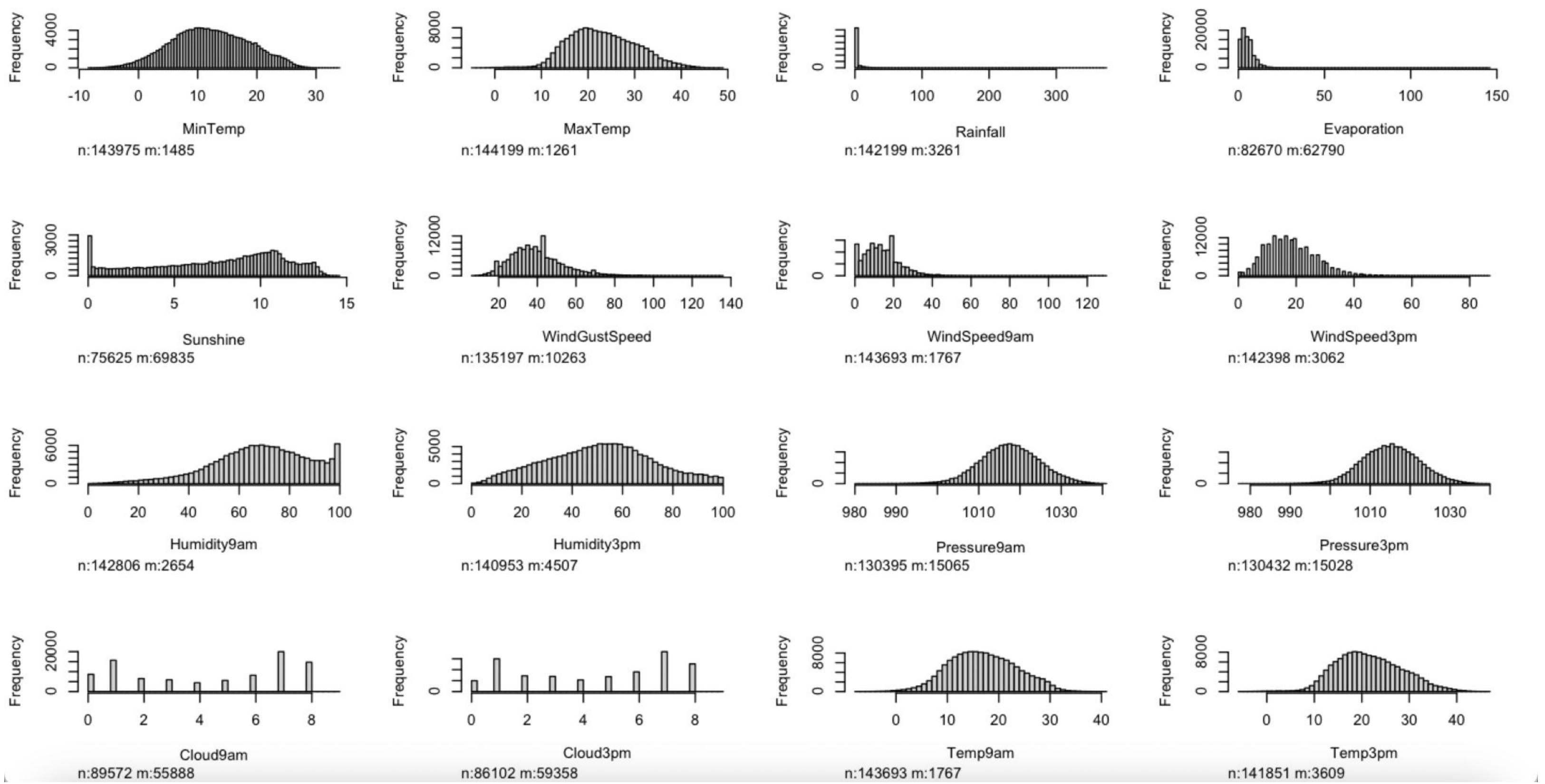
| Variable | N | Mean | Std. Dev. | Min | Pctl. 25 | Pctl. 75 | Max |
|---|---|---|---|---|---|---|---|
| MinTemp | 56420 | 13 | 6.4 | -6.7 | 8.6 | 18 | 31 |
| MaxTemp | 56420 | 24 | 7 | 4.1 | 19 | 30 | 48 |
| Rainfall | 56420 | 2.1 | 7 | 0 | 0 | 0.6 | 206 |
| Evaporation | 56420 | 5.5 | 3.7 | 0 | 2.8 | 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% | | | | | |

# Exploratory Analysis (cont.)

- Correlation plot between quantitative explanatory variables shown on the right
- Large boxes indicate higher correlation (both negative and positive)
- Blue indicates positive correlation, red indicates negative correlation, white indicates weak or no correlation
- Many of the correlations are intuitive (for example, the correlations between the Sunshine and Cloud variables are near -1)



# Exploratory Analysis (cont.)



# Methodology

- Split data into two parts: 2008-2013 (training) and 2014-2017 (test):
    - Because we’re working with time series data, we don’t want to predict rain using data from days in the future that haven’t occurred

**Methods used:**

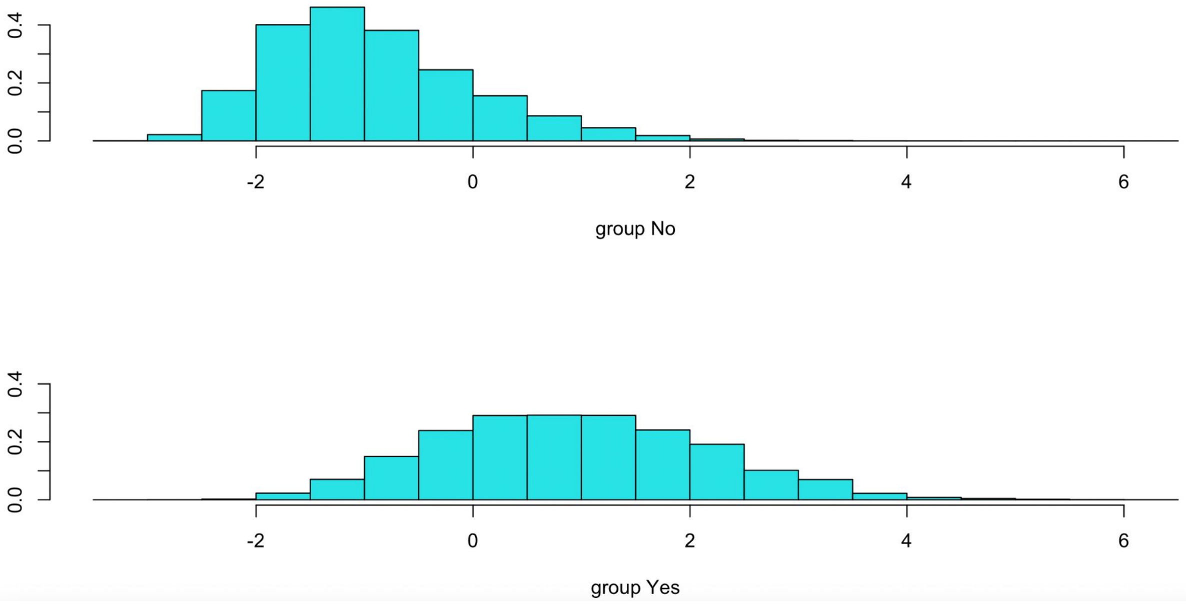
1. LDA/QDA:
    - Both methods are commonly used for classification because they use decision boundaries to separate the data. LDA has the same covariance matrix in each class while QDA has a different covariance matrix in each class
2. Logistic Regression:
    - This is a very popular classification model when there are 2 classes, as is the case with our response variable RainTomorrow, which has classes “Yes” and “No”.

# Methodology (continued)

3. Lasso/Ridge Regression
   - Lasso and Ridge shrink the coefficients close to zero to improve the prediction accuracy
   - Used cross-validation to select the optimal tuning parameter $\lambda$
4. Random Forest
   - Random Forest grows many decision trees on the training data, then combines them in order to make predictions
   - This method is useful because it is an advanced version of decision trees with higher prediction accuracy, allowing us to make better predictions on our test data

# Overall Results

- QDA/LDA:
  - QDA had an accuracy of 85.4%
  - LDA had an accuracy of 83.7%
  - Both models were highly accurate

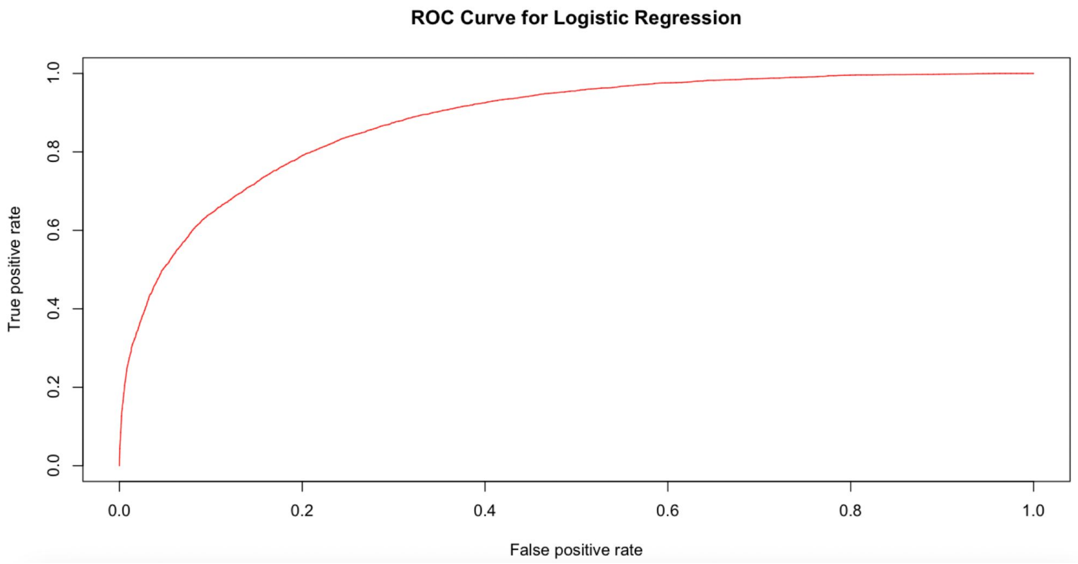



LDA

# Overall Results (continued)

Logistic Regression

- We used a cutoff point of 0.5 for our predictions
  - Predictions above 0.5 were classified as “Yes” and predictions below 0.5 were classified as “No” for RainTomorrow
  - Trying other cutoff points resulted in very similar accuracy as well
- Overall accuracy was 85.5%, implying that the logistic regression model was pretty accurate
- Very similar to QDA/LDA results



# Overall Results (cont.)

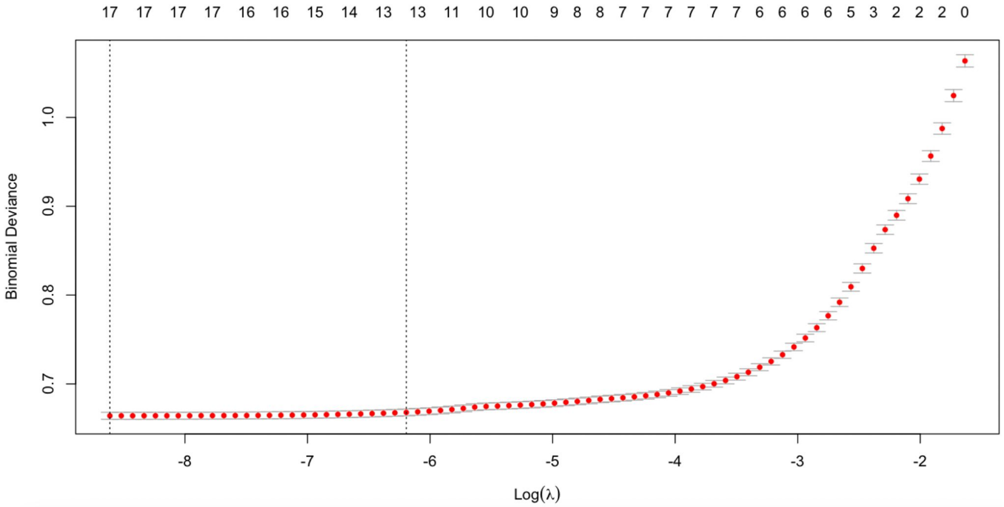
- Lasso/Ridge Regression:
  - Lasso Accuracy: 85.5%
  - Ridge Accuracy:  85.1%
- Lasso had λ = 0.00018, Ridge had λ = 0.019
- No coefficients were shrunk to 0

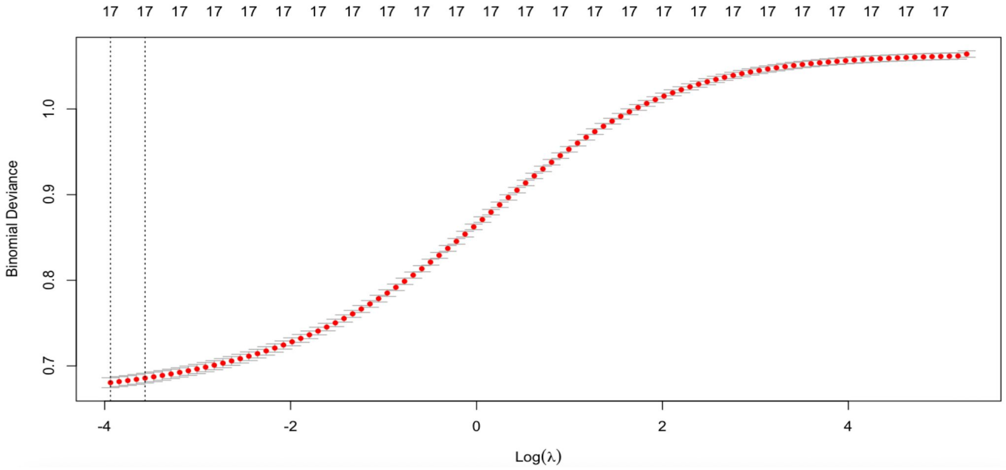
**Lasso**

| | lambda.min |
|---|---|
| (Intercept) | 56.672900914 |
| MinTemp | -0.039524262 |
| MaxTemp | 0.002861825 |
| Rainfall | 0.012524012 |
| Evaporation | -0.002833204 |
| Sunshine | -0.139631993 |
| WindGustSpeed | 0.059845053 |
| WindSpeed9am | -0.010288875 |
| WindSpeed3pm | -0.027363712 |
| Humidity9am | 0.001074917 |
| Humidity3pm | 0.057336990 |
| Pressure9am | 0.141957640 |
| Pressure3pm | -0.204532156 |
| Cloud9am | -0.014934135 |
| Cloud3pm | 0.125236085 |
| Temp9am | 0.037511549 |
| Temp3pm | 0.005217262 |
| RainTodayYes | 0.418399522 |

**Ridge**

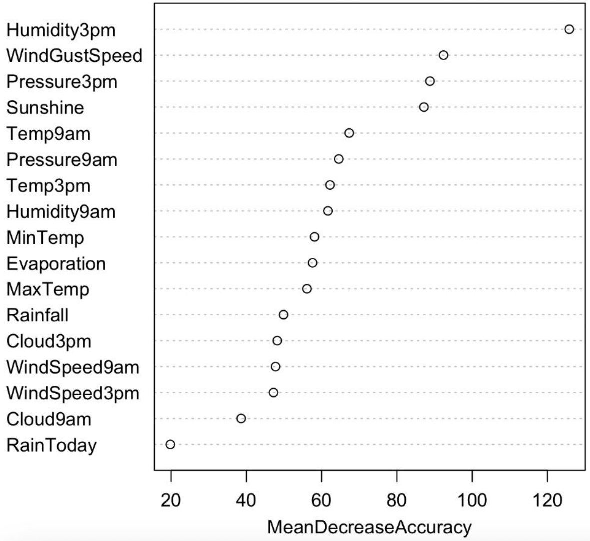
| | lambda.min |
|---|---|
| (Intercept) | 57.1883694152 |
| MinTemp | -0.0004860718 |
| MaxTemp | 0.0049484487 |
| Rainfall | 0.0129704448 |
| Evaporation | -0.0199356224 |
| Sunshine | -0.1231279365 |
| WindGustSpeed | 0.0383400223 |
| WindSpeed9am | -0.0060877086 |
| WindSpeed3pm | -0.0125301733 |
| Humidity9am | 0.0050073728 |
| Humidity3pm | 0.0363669856 |
| Pressure9am | -0.0047604579 |
| Pressure3pm | -0.0565809903 |
| Cloud9am | 0.0033221501 |
| Cloud3pm | 0.1264372069 |
| Temp9am | 0.0174424096 |
| Temp3pm | -0.0075681767 |
| RainTodayYes | 0.2988933497 |

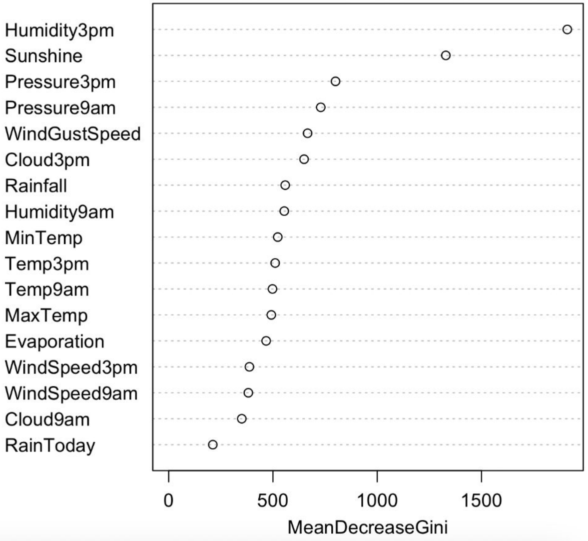


# Overall Results (continued)

- Random Forest
  - Accuracy of 85.7%
- Points on the right side of the plot show the importance of certain variables for improving model predictions
- Humidity3pm, Pressure3pm, Sunshine, and WindGustSpeed appear to be the most important variables for model performance



# Strengths/Limitations

- Strengths:
    - Large dataset allowed us to have large training and test sets
    - Model performance was fairly consistent across every method
        - Accuracy ranged from 83–86%
    - Given that it didn't rain 78% of the days, every model comfortably beat a naive prediction of no rain every day
    - Using a variety of methods allowed us to get a good sense of a reasonable prediction accuracy for the data
- Limits:
    - Dataset had many NA values, leading to some holes in the data
    - Some of the methods, such as the Random Forest, are somewhat computationally intensive

# Areas for Improvement / Next Steps

- Use other methods such as polynomial regression, naive Bayes, or more advanced decision tree methods
- Replace NA values with mean, median, or other value rather than removing them from the data
- Use best subset selection or forward/backward stepwise selection to refine the model by removing predictors
- Experiment with different sizes for test and training sets to see if the results are replicable

# Conclusion

- All of our methods had fairly similar accuracy in predicting rain
- We believe that all of our models were reliable, and that this data allows for good classification models due to the high accuracy
- Given the similar accuracy in model performance, it may be preferable to use models like logistic regression or LDA which are simple and easy to interpret

| Method | Accuracy |
|---|---|
| LDA | 85.4% |
| QDA | 83.7% |
| Logistic Regression | 85.5% |
| Lasso Regression | 85.5% |
| Ridge Regression | 85.1% |
| Random Forest | 85.7% |