Practical Data Science Group 2

1. Tran Viet Anh
2. Pham Nguyen Vu
3. Le Duc Nguyen
4. Nguyen Minh Duong

Problem

- The data set for this project is a 10 years collection of daily weather information in Australia.
- The goal is to use this data to predict rain probability for the following day.

Date	Location	MinTemp	MaxTemp	Rainfall
2008-12-01	Albury	13.4	22.9	0.6
2008-12-02	Albury	7.4	25.1	0
2008-12-03	Albury	12.9	25.7	0
2008-12-04	Albury	9.2	28	0
2008-12-05	Albury	17.5	32.3	1
2008-12-06	Albury	14.6	29.7	0.2
2008-12-07	Albury	14.3	25	0
2008-12-08	Albury	7.7	26.7	0
2008-12-09	Albury	9.7	31.9	0
2008-12-10	Albury	13.1	30.1	1.4
2008-12-11	Albury	13.4	30.4	0
2008-12-12	Albury	15.9	21.7	2.2
2008-12-13	Albury	15.9	18.6	15.6
2008-12-14	Albury	12.6	21	3.6
2008-12-15	Albury	8.4	24.6	0
2008-12-16	Albury	9.8	27.7	NA
2008-12-17	Albury	14.1	20.9	0
2008-12-18	Albury	13.5	22.9	16.8

Dataset

1. Date: The date of observation
2. Location: The location of the weather station
3. MinTemp: The minimum temperature in degrees celsius
4. MaxTemp: The minimum 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
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 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 at 3pm. This is measured in "oktas", which are a unit of eigths. It records how many cloud. A 0 measure indicates completely clear sky whilst an 8 indicates that it is completely overcast.
20. Temp9am: Temperature (degrees C) at 9am
21. Temp3pm: Temperature (degrees C) at 3pm
22. RainToday: Boolean: 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0
23. RainTomorrow: The amount of next day rain in mm. Used to create response variable RainTomorrow. A kind of measure of the "ri

Dataset

Data preparation: Issues

EXTRA WHITESPACES.

MISSING VALUES.

Data preparation: Solution

- Drop columns with high percentage of missing values and not contributed to the result of the analysis.
- Object columns filled with "Unknown".
- Numerical columns filled with median to fix the skewness of the set.

Hypothesis

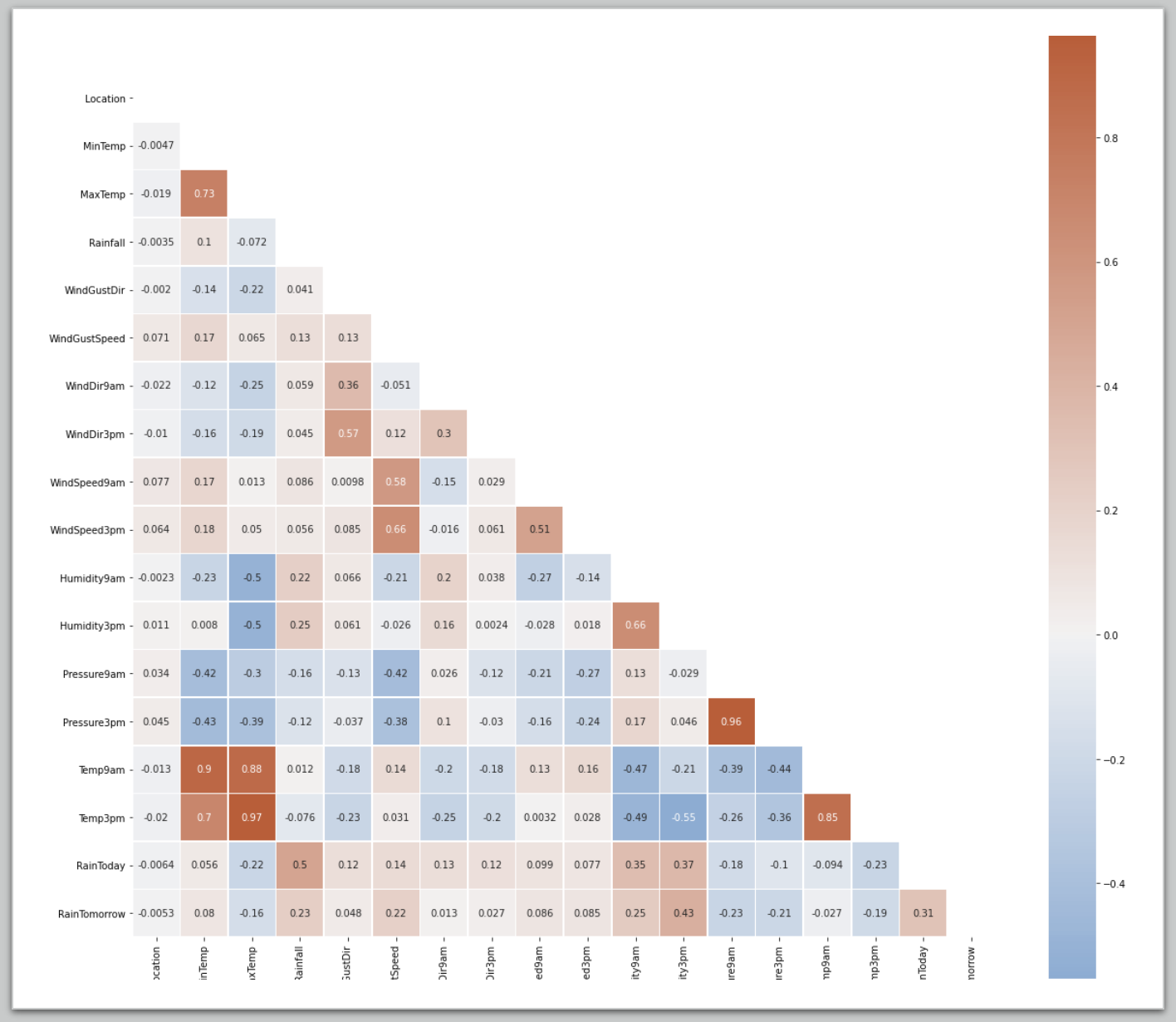
People tends to have their own judgement

Temperature

Humidity

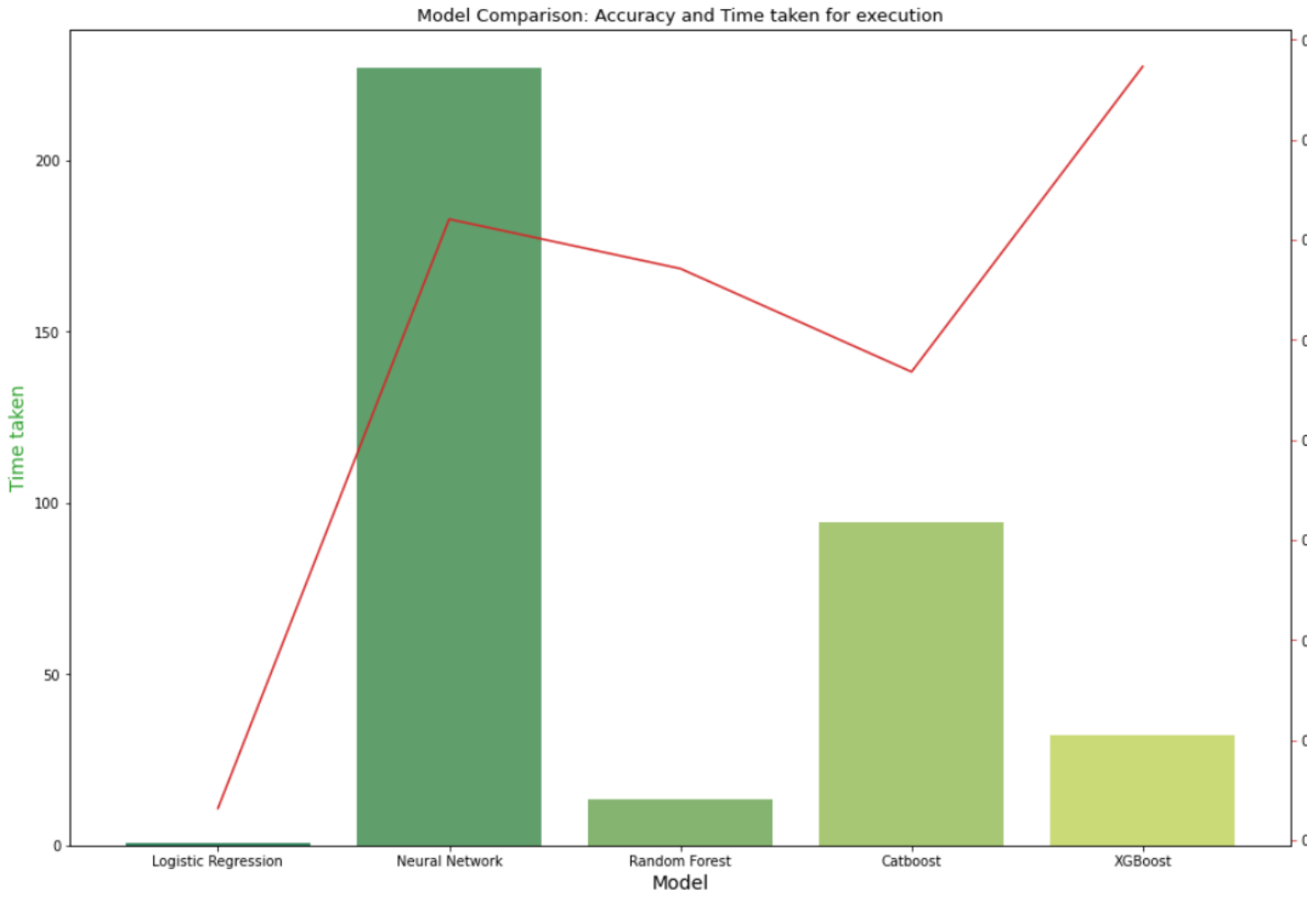
Wind

Has it rain today?



Modelling: Feature engineering

- Correlation analysis
- The following pairs of features are having high correlation between them:
- **MaxTemp** and **MinTemp**
- **Pressure9am** and **Pressure3pm**
- **Temp9am** and **Temp3pm**
- **MaxTemp** and **Temp3pm**
- But there is no case that the correlation value is equal to a perfect "1". As a result, we will not discard any feature.
-



Modelling: Algorithm

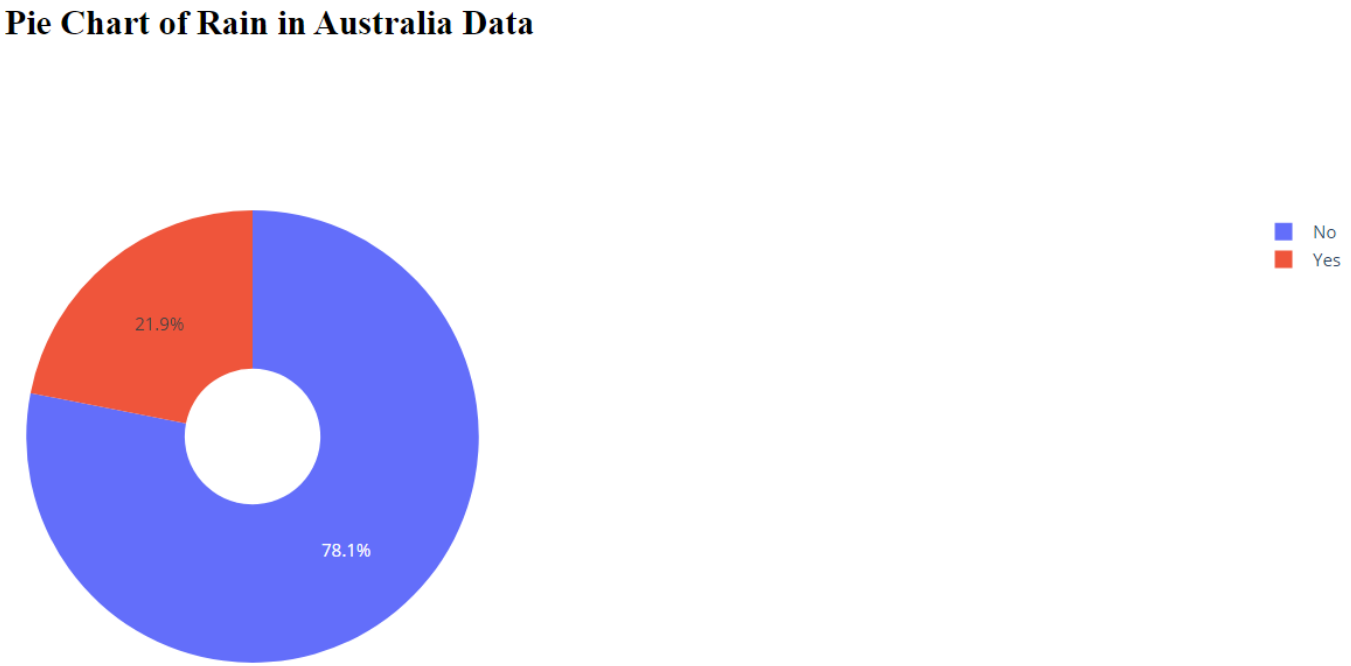
Modelling: Result

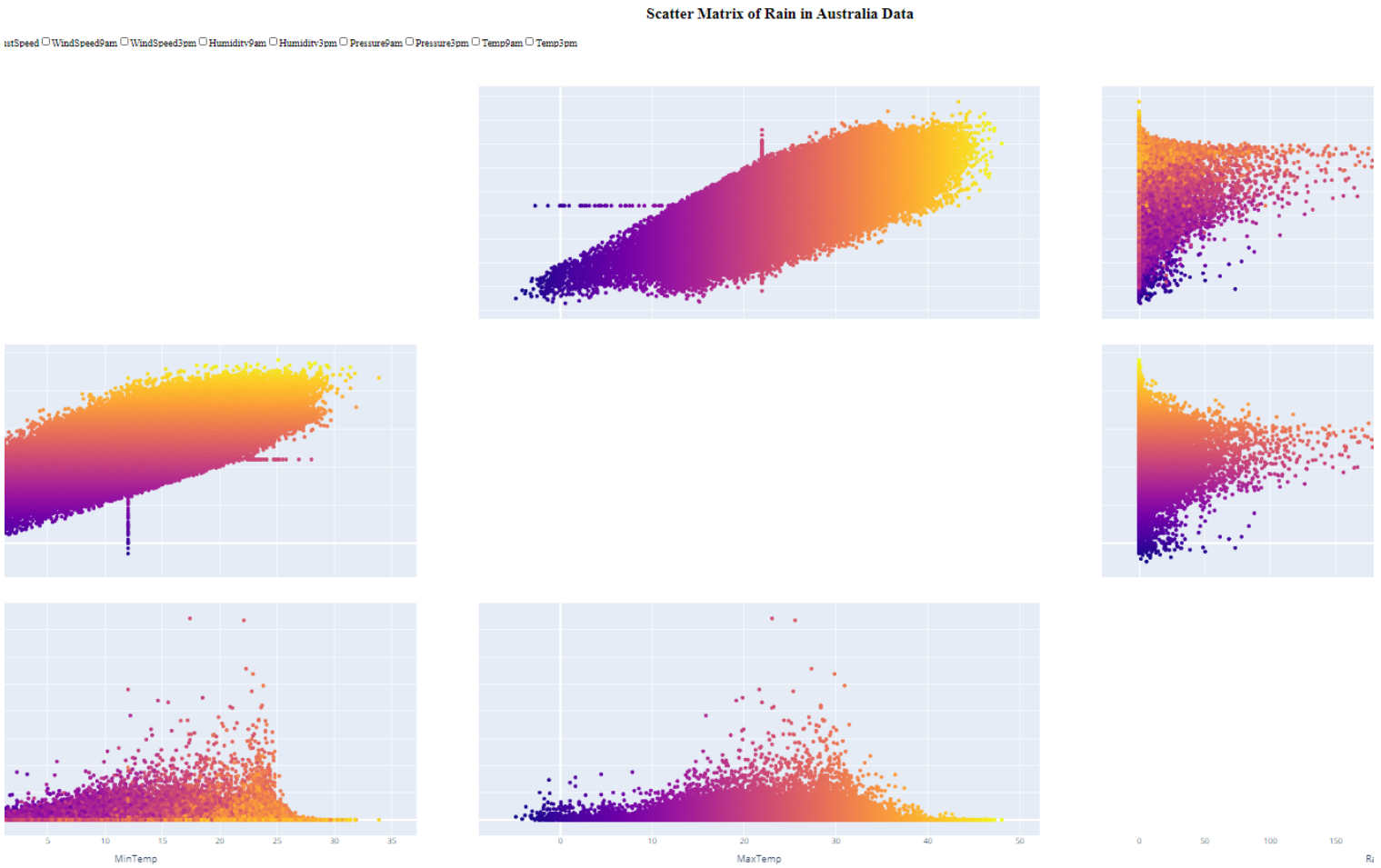
Model	Accuracy	ROC - AUC	Cohen's Kappa	Time Taken
Model 1 - Logistic Regression	0.8406	0.7016	0.4629	0.7379
Model 2 - Neural Network	0.8524	0.7296	0.5147	227.006
Model 3 - Random Forest	0.8514	0.7180	0.5000	13.3136
Model 4 - Catboost	0.8493	0.7310	0.5110	94.5475
Model 5 - XGBoost	0.8554	0.7446	0.5356	32.2469

Dashboard

1. Pie Chart

RainToday

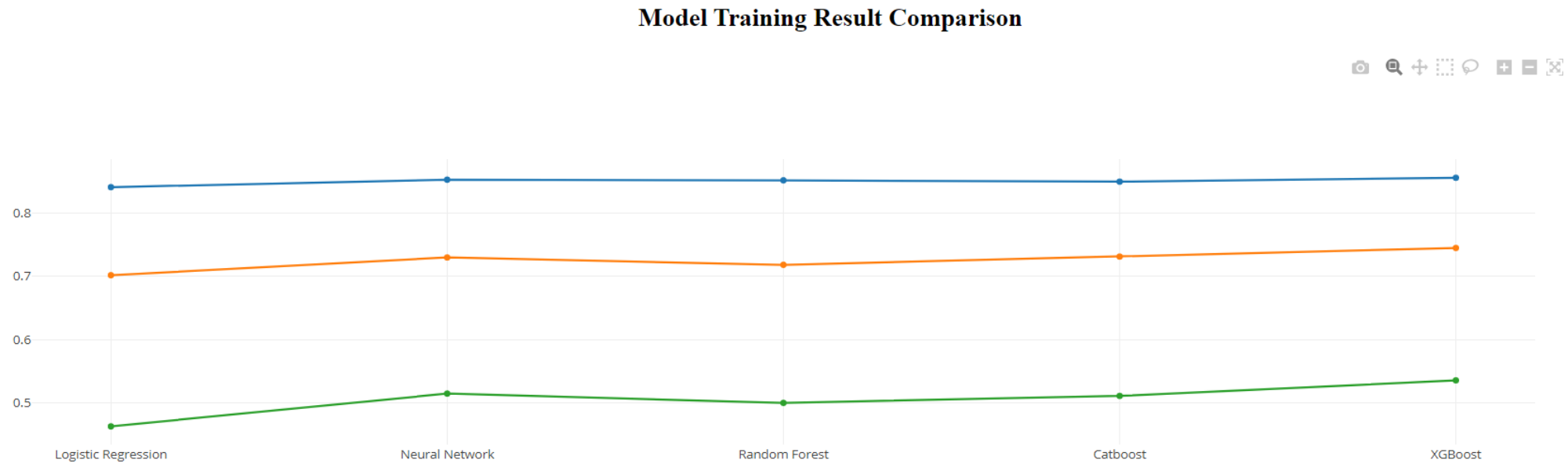



Scatter Matrix of Rain in Australia Data
WindSpeed9am
WindSpeed3pm
Humidity9am
Humidity3pm
Pressure9am
Pressure3pm
Temp9am
Temp3pm
MinTemp
MaxTemp

Dashboard

Dashboard

3. Multiple Line Chart



Conclusion

The result supports initial hypothesis

The possibility of rain according to data analysis relies on humidity, temperature, wind speed and wind direction