Toronto

Final Project Draft

Dhairya Dave[a] (002110382) PRO

Parth Savalia[a] (002982303)

Manan Soni[a] (002982645)

[a] *College of Professional Studies, Master of Professional Studies in Analytics.*

Subject: *ALY6040*

Under the guidance of

Prof. Kamran Hootan

Introduction:

This data is from Kaggle with titles “Rain in Australia”. By gathering quantitative information about the atmosphere's current condition at a specific location. To forecast whether or not it will rain tomorrow, using the Rain Dataset. About 10 years' worth of daily weather measurements from several Australia locales are included in the dataset. Speaking of weather predictions, they might have an impact on daily activities as well as industries like the food sector, tourism, emergency healthcare, etc.

There is a target variable that can be either "Yes" or "No”. Yes, if there was at least 0.1 mm of rain that day. The variables in our dataset that are most likely to cause rain to fall are pressure, humidity, clouds, sunlight and etc.

Research Question:

There are multiple factors to classify rain such are humidity, pressure, temp, cloud, etc. Using this feature is to classify if there will be rain tomorrow. In addition, determination of wind direction and speed do affect in causation of rain.

Unit of Analysis:

Unit of analysis is important for analysis as it also can be called as unit of observation. For our scenario, a sample collected from environment is unit of analysis. Because, observation of environmental phenomenon such pressure, humidity and others at two specific time like 9am and 3pm as well as other observation of clouds, wind speed and direction, rainfall amount are features. Therefore, unit of analysis can be said as observation of atmosphere.

Preprocess of Data:

Initially, retrieved count of nulls from all variables and split into continues and categorically. After determining the distributions, we replaced null values with mean, median and mode. For variables except Rainfall and Evaporation, we replaced null with mean meanwhile for them null was replaced by median as they were totally left skewed. As Rainfall is more than 0.1 mm we needed to replace null of RainToday with yes. For continues variables we just replaced with mode.

After correlation and VIF, we figured out that not a single variable is enough for classification of rain and found out that wind direction has nothing to do with rain where as other do affect on Rainfall.

Finally, variables of interest are as of Fig.1.

```
['MinTemp', 'MaxTemp', 'Rainfall', 'Evaporation', 'Sunshine',
 'WindGustSpeed', 'WindSpeed9am', 'WindSpeed3pm', 'Humidity9am',
 'Humidity3pm', 'Pressure9am', 'Pressure3pm', 'Cloud9am', 'Cloud3pm',
 'Temp9am', 'Temp3pm', 'RainToday'],
```

Fig.1 Variables of Interest

Description of Data:

As illustrated in Fig.2 (Metadata),

- 24-hour format is reshaped from 9 am to 9 am for recording the sample.
- The measure of cloud is fraction of sky covered by cloud, so we can assume that the fraction is in percentage.

```
Date - The date of observation Location -The common name of the location of the weather station
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.
WindGustDi r- The direction of the strongest wind gust in the 24 hours to midnight
WindGustSpeed -The speed (km/h) of the strongest wind gust in the 24 hours to midnight
WindDir9am -Direction of the wind at 9am
WindDir3pm -Direction of the wind at 3pm
WindSpeed9am -Wind speed (km/hr) averaged over 10 minutes prior to 9am
WindSpeed3pm -Wind speed (km/hr) averaged over 10 minutes prior to 3pm
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.
Cloud3pm -Fraction of sky obscured by cloud
Temp9am-Temperature (degrees C) at 9am
Temp3pm -Temperature (degrees C) at 3pm
RainToday -Boolean: 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0
RainTomorrow -The amount of next day rain in mm.
```

Fig 2: Metadata

According to Table 1 (Information of Data types), these are the main 17 feature which eventually can cause a rain. Rain is a natural calamity and we have 145,460 samples. RainTomorrow is out target variable.

Table 1 Information of Data type

#	Column	Non-Null Count	Dtype
0	MinTemp	145460 non-null	float64
1	MaxTemp	145460 non-null	float64
2	Rainfall	145460 non-null	float64
3	Evaporation	145460 non-null	float64
4	Sunshine	145460 non-null	float64
5	WindGustSpeed	145460 non-null	float64
6	WindSpeed9am	145460 non-null	float64
7	WindSpeed3pm	145460 non-null	float64
8	Humidity9am	145460 non-null	float64
9	Humidity3pm	145460 non-null	float64
10	Pressure9am	145460 non-null	float64
11	Pressure3pm	145460 non-null	float64
12	Cloud9am	145460 non-null	float64
13	Cloud3pm	145460 non-null	float64
14	Temp9am	145460 non-null	float64
15	Temp3pm	145460 non-null	float64
16	RainToday	145460 non-null	int64

As demonstrated in the Table 2 (Description of Data),

- the difference in humidity at 9 am and 3 pm is significant as humidity reduce from 69 to 52 while both recorded maximum 100.
- Whereas for pressure, there is slight difference for mean pressure at 9 am to 3pm.
- It has recorded that maximum 9% fraction of sky was covered by cloud at specific time frame.
- 371 mm rainfall has been registered. And this amount of rain can cause flood like disaster.

Table 2 Description of Data

	count	mean	std	min	25%	50%	75%	max
MinTemp	145460.0	12.194317	6.364469	-5.950000	7.700000	12.100000	16.800000	30.450000
MaxTemp	145460.0	23.225145	7.067566	2.700000	18.000000	22.700000	28.200000	43.500000
Rainfall	145460.0	0.381674	0.608638	0.000000	0.000000	0.000000	0.600000	1.500000
Evaporation	145460.0	5.095891	1.709594	1.797653	4.000000	5.468232	5.468232	7.670579
Sunshine	145460.0	7.922535	1.386787	5.977944	7.611178	7.611178	8.700000	10.333234
WindGustSpeed	145460.0	39.716321	12.174937	8.500000	31.000000	39.000000	46.000000	68.500000
WindSpeed9am	145460.0	13.952432	8.555347	0.000000	7.000000	13.000000	19.000000	37.000000
WindSpeed3pm	145460.0	18.576025	8.442192	0.000000	13.000000	18.662657	24.000000	40.500000
Humidity9am	145460.0	68.932605	18.703608	18.000000	57.000000	69.000000	83.000000	100.000000
Humidity3pm	145460.0	51.539116	20.471189	0.000000	37.000000	51.539116	65.000000	100.000000
Pressure9am	145460.0	1017.676878	6.568430	1001.050000	1013.500000	1017.649940	1021.800000	1034.250000
Pressure3pm	145460.0	1015.274311	6.528871	998.650000	1011.100000	1015.255889	1019.400000	1031.850000
Cloud9am	145460.0	4.447461	2.265604	0.000000	3.000000	4.447461	6.000000	9.000000
Cloud3pm	145460.0	4.544125	2.026092	1.000000	4.000000	4.509930	6.000000	9.000000
Temp9am	145460.0	16.991738	6.440803	-1.500000	12.300000	16.800000	21.500000	35.300000
Temp3pm	145460.0	21.685669	6.812734	2.450000	16.700000	21.400000	26.200000	40.450000
RainToday	145460.0	0.351430	0.477419	0.000000	0.000000	0.000000	1.000000	1.000000

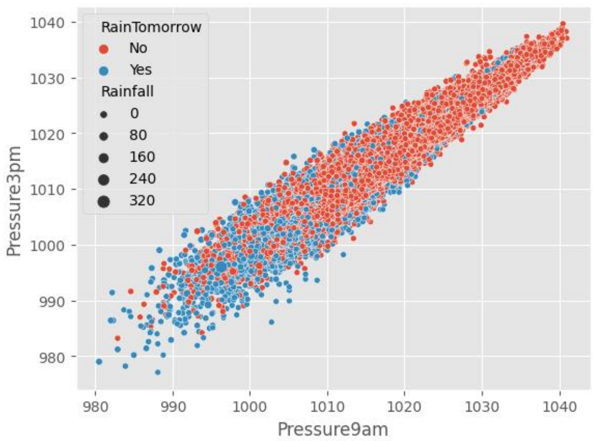

Fig.3 Pressure at 9am vs 3pm

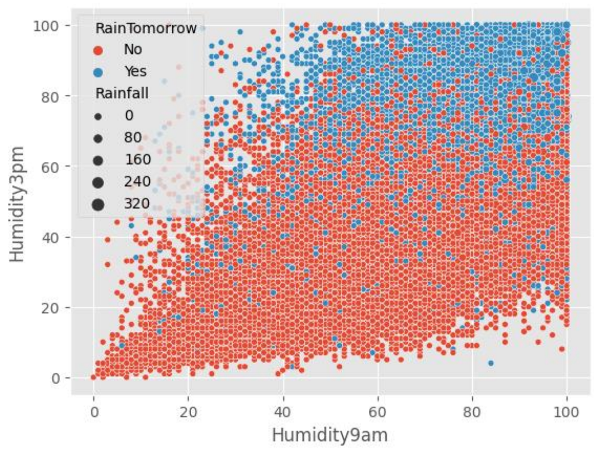

Fig.4 Humidity at 9am vs 3pm

Fig.3 and Fig.4 demonstrates some patterns such as pressure less than 1015 at 9am and 3pm can cause a good chance of rain while there are high chances of rain when humidity is greater than 60 at 3pm.

Methodology for Research Problems:

Dependent variables are as illustrated in Fig.1 which were mentioned earlier.

As our independent variable is RainTomorrow which has only 2 unique values such as "yes" and *"no"* and achieving our research question classification method such as Logistic Regression, Random Forest and Gradient boosting are suitable. All of them are classification algorithm.

Logistic Regression because it is simple model when there is binary classification. Whereas, Gradient boosting and Random Forest are advanced approaches to the decision tree algorithm. There is significant difference in both algo which can help to achieve greater accuracy.

Predictive Analysis:

Initially, we split the data of 145460 samples into train and test split with randomly with 0.3 threshold which means for index it selects randomly from dataframe and it selects 30% of dataframe for testing. Therefore, 70% of samples will be used for training the models.

We will use this data to input 3 different models such as Logistic Regression, Random Forest Classifier and Gradient Boosting Classifier.

As we can interpret from Table 3, accuracy of Random Forest is highest among 3 where as Gradient Boosting scored highest F1. F1 score should be prefer as its stands for overall model performance by evaluating precision and recall of the model. In addition, precision and recall are score for false positives and negatives.

F1 score has value in range of 0 to 1 and higher is better for model.

Table 3 Models with Accuracy and F1 Score

	Model	Accuracy	F1 Score
0	Logistic Regression	84.19	0.548748
1	Random Forest	85.65	0.601350
2	Gradient Boosting	85.60	0.612836

Confusion matrixes:

As showing in the Fig.5, Fig.6 and Fig.7, there is approx. 10-12% false negatives and 3-4% false positives which are considerable amount as there are many external factors can affect the causation of rain such as global warming, deforestation, etc.

There is very less prediction difference in Random Forest and Gradient Boosting.

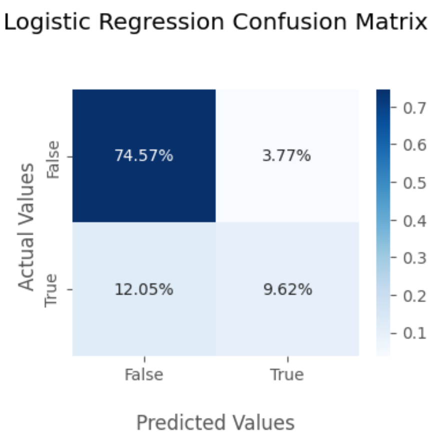

Fig.5 CF of Logistic Regression

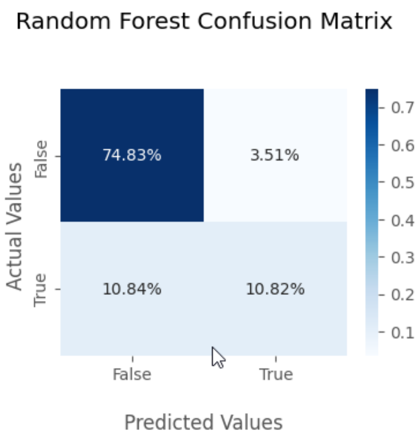

Fig.6 CF of Random Forest

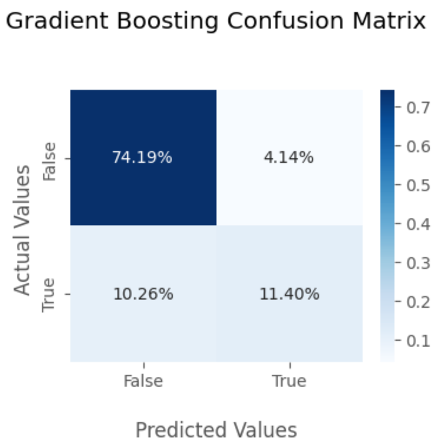

Fig.7 CF of Gradient Boosting

AUC curves:

As demonstrated AUC curves in Fig.8, Fig.9 and Fig.10, they are almost same as minor change can be seen but AUC of gradient boosting is 0.73 which is significant.

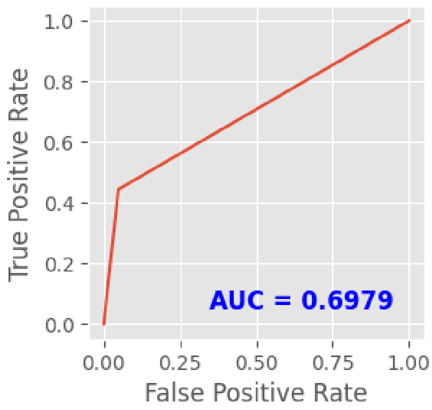

Fig.8 AUC of Logistic Regression

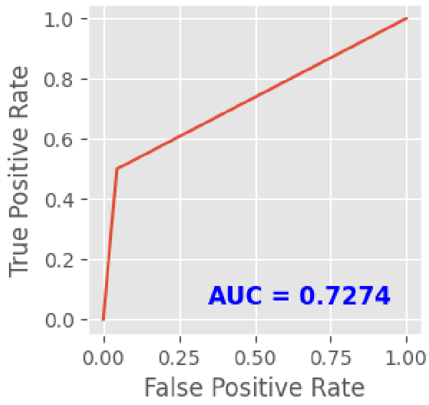

Fig.9 AUC of Random Forest

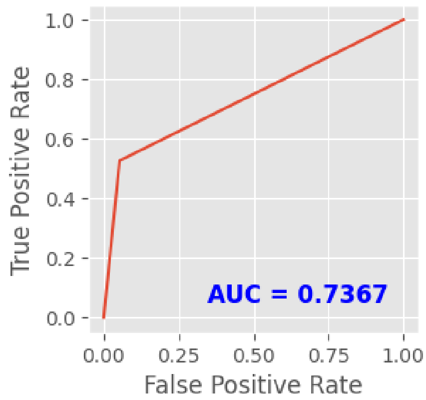

Fig.10 AUC of Gradient Boosting

Conclusion

As we evaluate 3 different model who have different statistical approaches, we conclude that Gradient Boosting has better performance in comparison to Random Forest and Logistic Regression. To support this, we have confusion matrix and AUC curve which proves that Gradient Boosting is slightly better than Random Forest.

We tried modifying variables such addition of Wind Directions but it doesn't affect our model and from non-data driven perspective wind doesn't affect rain.

Furthermore, in research questions:

- Reduce the errors in classification to make models more perfect.
- Changing the hyperparameters such number of trees, it's depth, learning rate, etc to gain achieve better F1 score and AUC score.
- After Gradient Boosting, we can try out XGBoost which is extreme gradient boosting.
- Evaluation of Date and Location features for clustering or further classification.

References

[1] M. (2022, October 9). *GitHub - Mxnxn/Rain_Forecasting_Australia*. GitHub. Retrieved October 9, 2022, from https://github.com/Mxnxn/Rain_Forecasting_Australia

[2] Z., & posts by Zach, V. A. (2020, July 20). *How to Calculate VIF in Python - Statology*. Statology. Retrieved October 9, 2022, from https://www.statology.org/how-to-calculate-vif-in-python/