NATIONAL UNIVERSITY OF SINGAPORE
Faculty of Science

AY 2021/2022, SEMESTER 2
DSA3101: Data Science in Practice

Rain in Singapore

Student Number: A0205767X

Although Singapore is not fortunate enough to experience the four different seasons as we are located near the equator, we do experience two monsoon seasons: the Northeast Monsoon season occurring from December to early March and the Southwest Monsoon Season from June to September. With Singapore experiencing nearly an all year round of rain, I would like to predict whether it will be raining the next day in different regions of Singapore, and to determine the main indicators which affect whether it will rain the next day.

1 Description of Data

The rainfall dataset from 2018 to 2021 was scrapped from the Meteorological Service Singapore website. It contains information such as the amount of daily rainfall, temperature and wind speed collected from weather stations scattered across Singapore.

2 Data Pre-processing

Firstly, we categorised the various stations into five main regions: North, Northeast, East, Central and West. Next, summarised the data collected daily from each stations into total rainfall, mean rainfall, mean temperature, maximum temperature, minimum temperature, mean wind speed and max windspeed of each regions.

Since we will be predicting a binary output, we created an additional column for the response variable indicating 1: if it rains and 0: if it did not rain. Missing values were also dropped, as some stations did not have any data collected.

The data was also split into both training and testing. Training data comprised of data from 2018 to 2020, while testing data comprised of data from 2021.

3 Exploratory Data Analysis

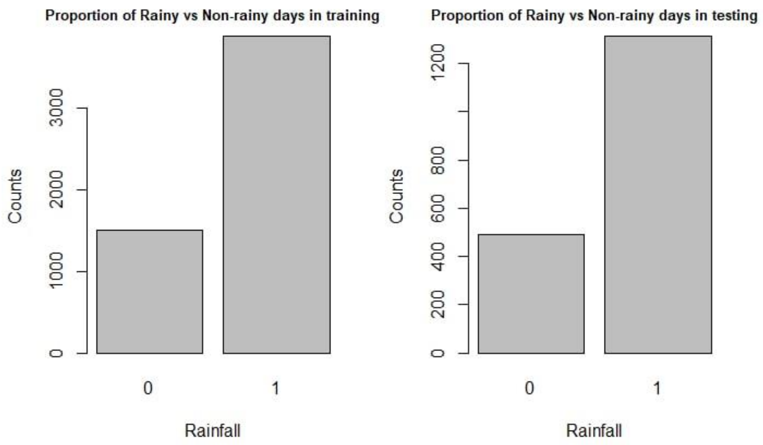


Figure 1: Proportion of rainy vs non-rainy days in training and test dataset

From Figure 1, we observe that there is an imbalanced data. There were more instances of rainy days than non-rainy days, which is mainly due to Singapore experiencing more days of rain. However, both the training and testing set had the same proportion of rainy and non-rainy days. Hence, our test set is still representative of our training set.

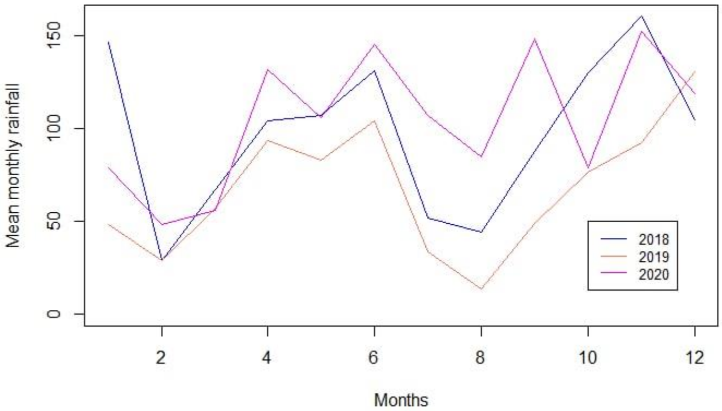


Figure 2: Plot of mean monthly rainfall against months for 2018, 2019, 2020

From Figure 2, we observe that there is a similar trend of period of high rainfall and low rainfall in different years. Singapore experienced more rainfall between periods of March to June, and October to December. Although the data does not seem to correspond to the 2 main monsoon seasons, but it seems to reflect the relatively short inter-monsoon periods which happen between April to May and October to November.

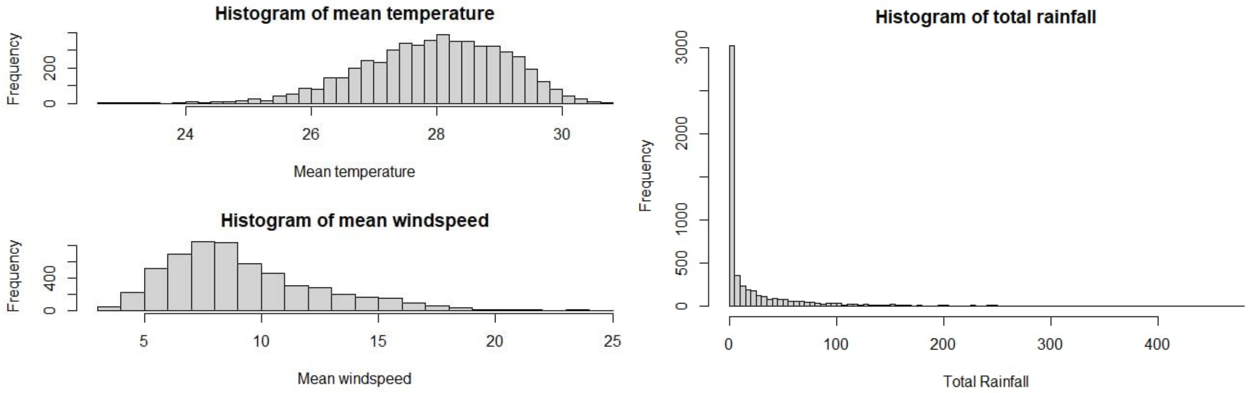


Figure 3: Distribution of mean temperature, mean windspeed and total rainfall

From Figure 3, outliers were observed. There seems to be extreme weather events recorded such as low mean temperatures of around 20 °C, high total rainfall of around 1175mm and high mean windspeed of approximately 24km/h. However, these were all valid results and will not be removed from our observations.

There were also high correlation among the predictors, which may indicate multicollinearity. Mean temperature and maximum temperature have a correlation of 0.734. Mean temperature and minimum temperature have a correlation of 0.686. However, since their Variance Inflation Factors were less than 10 as shown in Table 1, the predictors were not removed.

Mean Temperature	Maximum Temperature	Minimum Temperature
3.816	2.340	2.038

Table 1: VIF values of mean, maximum and minimum temperature

4 Methodology

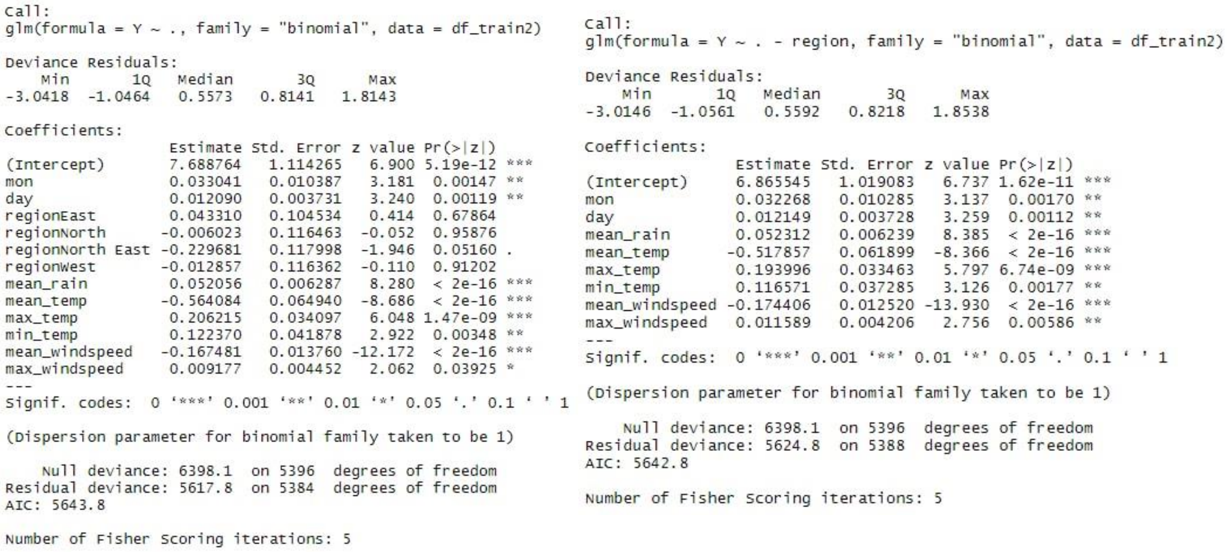
To predict whether it will rain tomorrow, today's data was used as tomorrow's data is still not available yet. The predictors used are month, day of month, region, mean rainfall, mean temperature, maximum temperature, minimum temperature, mean windspeed and maximum windspeed.

4.1 Baseline Model

Given our imbalanced dataset, we implemented a naïve model as our baseline model which will predict every day as a rainy day. It achieves a test accuracy of 72.7%, high sensitivity of 100% and low specificity of 0%. Hence, our subsequent models will aim to achieve a better performance than the current model.

4.2 Logistic Regression Model

Logistic Regression was first attempted to explore the performance of a linear relationship model on the data.



```
Call:
glm(formula = Y ~ ., family = "binomial", data = df_train2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0418  -1.0464   0.5573   0.8141   1.8143

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        7.688764   1.114265   6.900 5.19e-12 ***
mon                0.033041   0.010387   3.181  0.00147 **
day                0.012090   0.003731   3.240  0.00119 **
regionEast         0.043310   0.104534   0.414  0.67864
regionNorth       -0.006023   0.116463  -0.052  0.95876
regionNorth East -0.229681   0.117998  -1.946  0.05160 .
regionWest        -0.012857   0.116362  -0.110  0.91202
mean_rain          0.052056   0.006287   8.280  < 2e-16 ***
mean_temp         -0.564084   0.064940  -8.686  < 2e-16 ***
max_temp           0.206215   0.034097   6.048 1.47e-09 ***
min_temp           0.122370   0.041878   2.922  0.00348 **
mean_windspeed    -0.167481   0.013760 -12.172  < 2e-16 ***
max_windspeed      0.009177   0.004452   2.062  0.03925 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6398.1  on 5396  degrees of freedom
Residual deviance: 5617.8  on 5384  degrees of freedom
AIC: 5643.8

Number of Fisher Scoring iterations: 5
```

```
Call:
glm(formula = Y ~ . - region, family = "binomial", data = df_train2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0146  -1.0561   0.5592   0.8218   1.8538

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      6.865545   1.019083   6.737 1.62e-11 ***
mon              0.032268   0.010285   3.137  0.00170 **
day              0.012149   0.003728   3.259  0.00112 **
mean_rain        0.052312   0.006239   8.385  < 2e-16 ***
mean_temp       -0.517857   0.061899  -8.366  < 2e-16 ***
max_temp         0.193996   0.033463   5.797 6.74e-09 ***
min_temp         0.116571   0.037285   3.126  0.00177 **
mean_windspeed -0.174406   0.012520 -13.930  < 2e-16 ***
max_windspeed    0.011589   0.004206   2.756  0.00586 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6398.1  on 5396  degrees of freedom
Residual deviance: 5624.8  on 5388  degrees of freedom
AIC: 5642.8

Number of Fisher Scoring iterations: 5
```

Figure 4: Summary table of logistic regression models

Performing logistic regression model on all the predictors, from Figure 4 on the left, we observe that region was statistically insignificant, as it has a p-value greater than 0.05. A second logistic regression model without the region predictor was performed to observe if there will be any improvements in the model's performance.

		Truth	
		0	1
Predicted	0	44	21
	1	449	1293

		Truth	
		0	1
Predicted	0	48	23
	1	445	1291

Table 2: Confusion matrix of logistic regression models

From Table 2, the first logistic regression model achieved a test accuracy of 73.9%, specificity of 8.92 % and sensitivity of 98.4%. However, the second logistic regression model sees a slight improvement with a test accuracy of 74.1%, specificity of 9.74% and sensitivity of 98.3%.

4.3 Random Forest Model

Random Forest Model was attempted to observe if a more flexible model would perform better. Since it only considers a subset of predictors at each split, it reduces variances and will also be more robust to outliers. Furthermore, random forest would also provide insights on the variable importance of the predictor. We used cross-validation to determine the optimal number of predictors to consider at each split. Training on 2018 and 2019 data, and validating on 2020 data.

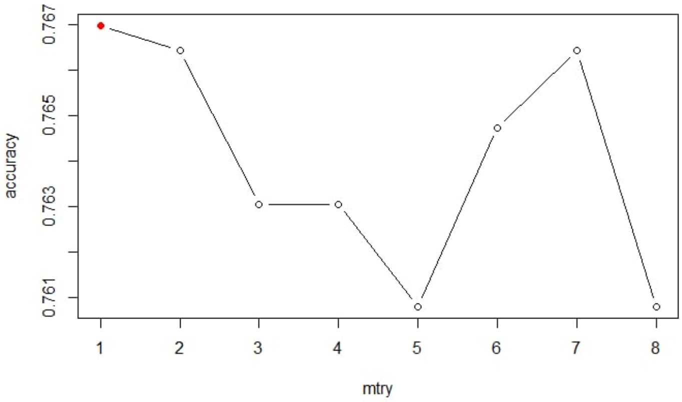


Figure 5: Cross-validation plot of accuracy against mtry

		Truth	
		0	1
Predicted	0	108	66
	1	385	1248

Table 3: Confusion matrix of random forest model

From Figure 5, the optimal number of predictors to consider at each split is 1. From Table 3, the test accuracy achieved is 75%, specificity is 21.9% and sensitivity is 95%. This is an improvement from both the baseline model and logistic regression model.

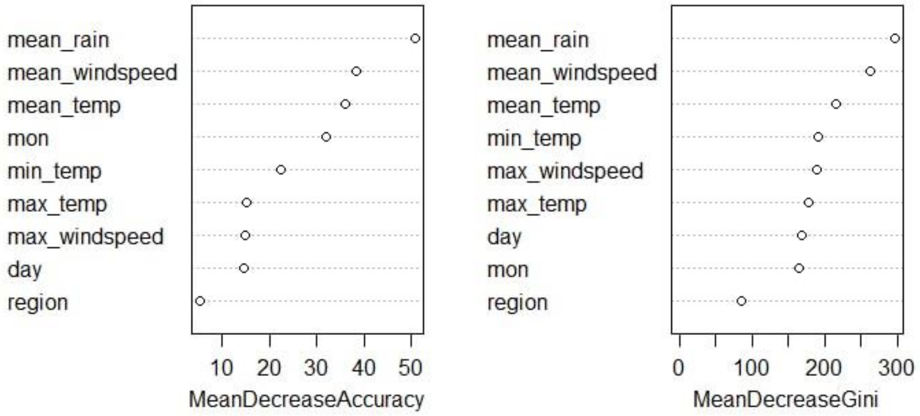


Figure 6: Variable importance plot of random forest

From Figure 6, the previous day's mean rainfall, mean windspeed and mean temperature were the most important variables the model uses to make correct predictions of whether it will rain today.

4.4 XGBoost Model

Lastly, XGBoost model was attempted as the model may perform better in cases where there is an imbalanced data compared to Random Forest. This is because when XGBoost is unable to predict the response correctly, it gives more weightage to the minority classes in future iterations. Performing cross-validation on various combinations of parameters such as `eta` which controls the learning rate, `nrounds` which is the number of boosting iterations, `max_depth` of each boosted tree and `gamma` a regularization parameter. The optimal values of the parameters are `eta` = 0.2, `gamma` = 0.01, `nrounds` = 8 and `max_depth` = 5.

Predicted		Truth	
		0	**1**
	0	157	92
	1	336	1222

Table 4: Confusion matrix of XGBoost model

It achieves the highest test accuracy of 76.3%, specificity of 31.8% and sensitivity of 93%. The improvement in accuracy and specificity may suggest XGBoost is more robust against imbalanced data.

5 Conclusion

In conclusion, XGBoost was the best performing model for this data. Previous day's mean rainfall, mean windspeed and mean temperature were also important indicators of whether it will rain today. Further improvements such as increasing lag or using additional data such as humidity could be made.

APPENDIX

The dataset was obtained from: http://www.weather.gov.sg/climate-historical-daily/

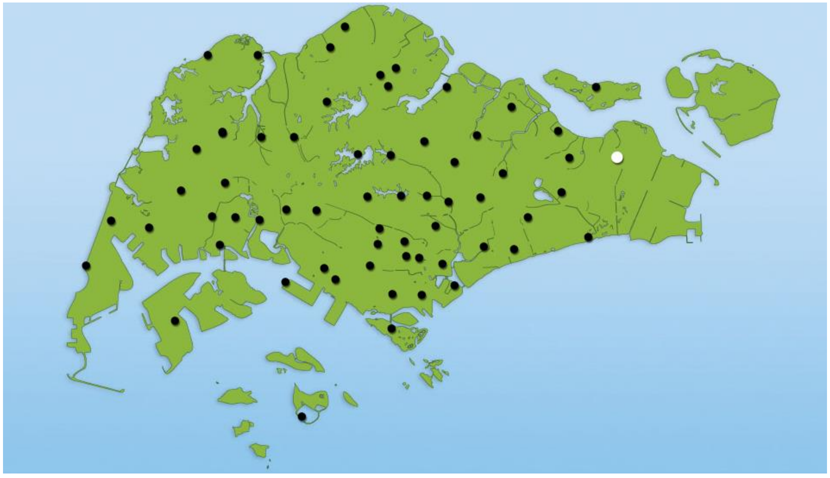


Figure 7: Weather stations across Singapore

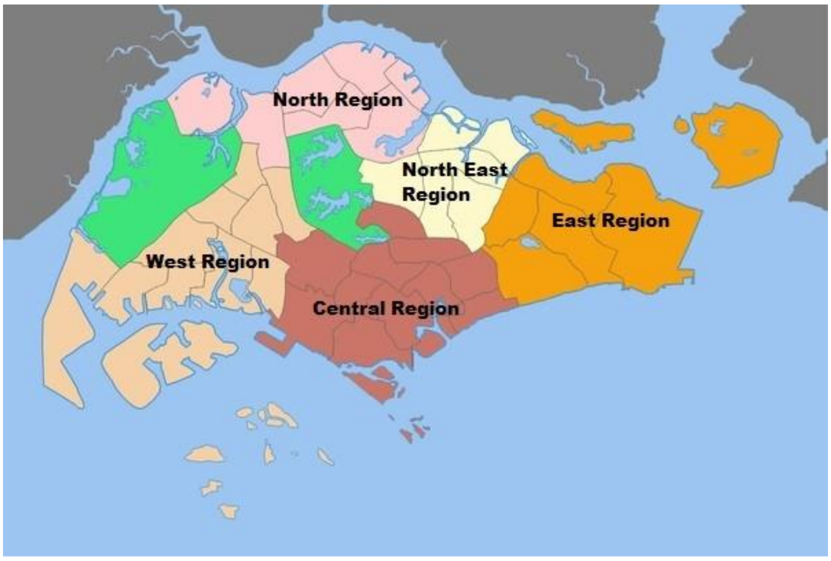


Figure 8: The five regions in Singapore

Variable Name	Description
mon	Month
day	Day of the month
region	Regions in Singapore: North, North East, East, Central, West
mean_rain	Yesterday's mean rainfall at a particular region
mean_temp	Yesterday's mean temperature at a particular region
max_temp	Yesterday's maximum temperature at a particular region
min_temp	Yesterday's minimum temperature at a particular region
mean_windspeed	Yesterday's mean windspeed at a particular region
max_windspeed	Yesterday's maximum windspeed at a particular region
Y	1: Rained today; 0: Did not rain today

Table 5: Table of descriptions of variables

Models	Accuracy	Specificity	Sensitivity
Baseline model	72.7%	0%	100%
Logistic Regression with all predictors	73.9%	8.92%	98.4%
Logistic Regression with all predictors except region	74.1%	9.74%	98.3%
Random Forest	75%	21.9%	95%
XGBoost	76.3%	31.8%	93%

Table 6: Summary of models attempted and its accuracy, specificity & sensitivity