

## **Section A: Business Understanding**

This part will provide background information on the whole report. Firstly, we will introduce the necessary information about our client. Secondly, the analysis of the main challenge of the client is shown. Thirdly, we describe how our weather prediction service could help the client solve this challenge. Finally, the target variable is defined and potential factors that might have a relatively strong relationship with the target variable to it are briefly discussed.

### ***1. The Client Information for the Test:***

On 1st September, we've got a client who plans to launch a series of open-air theatres for the first time some days during December. The client has no pre-experience and would like to take a cost-saving and successful activity. It needs to finalize the days around early November to prepare for the promotion and other issues. It has to launch the event on sunny nights and knowing the weather condition in advance is important to coordinate other issues. Only in this way, it could provide their consumers with a wonderful experience, and minimize the cost of holding the activity. Under this circumstance, the selection of the dates to hold the open-air theatre activity is the most crucial step.

### ***2. The Main Problem for Our Current Client:***

The client's main challenge is to know the weather beforehand in order to align other issues to hold the event. It is hard for the client to follow the basic weather forecast provided by the Bureau of meteorology because the traditional weather forecast only provides cursory information about the weather condition for the next 15 days. The client needs to know the weather condition of each day in December to choose the proper days to launch the open-air theatres.

### ***3. Why the Prediction is Important for Our Client:***

Our service is to predict whether it will rain at a given period throughout the year in Australia. This is important for our client, as its activities tend to have high requirements on the weather condition. Our prediction might help the client decide the period to launch the activity in order to maximize total profit and reduce the potential risk caused by bad weather condition. In reality, not only does our business help the companies engaged in outdoor tourism to decide the days, period, season to launch the specific activities, but also provide information for the stakeholders such as experiencers and travelers to arrange their time in advance.

### ***4. Target Variable and Other Potential Factors that Might be Related to It:***

The target variable in our test is whether it will rain at a given time period. Our prediction of precipitation is based on the historical data from the past ten years. The result of prediction based on detecting many factors. The factors such as temperature, humidity, sunshine and evaporation could be strongly correlated with the target variable. A more detailed correlation between these factors is shown in Section B.

## Section B: Data Exploration

This section will illustrate the data dictionary (Table 1), explain the correlations between different variables, and explore the distribution of categorical and numerical variables.

### 1. Data dictionary explanation:

Variable	Definition	Type	Unit	Range
Month	The month of the year	categorical	Numbers represent different types	1,2,3...12
Climate	The climate of the Area	categorical	Numbers represent different types	0,1,2,3,4,5,6
MinTemp	Minimum temperature in the 24 hours to 9 am	numerical	Celsius (°C)	-8.5 – 33.9
MaxTemp	Maximum temperature in the 24 hours to 9 am	numerical	Celsius (°C)	-4.8 – 48.1
Rainfall	Precipitation (rainfall) in the 24 hours to 9 am	numerical	Millimetre (mm.)	0.0 – 371.0
Evaporation	"Class A" pan evaporation in the 24 hours to 9 am	numerical	Millimetre (mm.)	0.0 – 145.0
Sunshine	Bright sunshine in the 24 hours to midnight	numerical	Hour	0.0 – 14.5
WindGustDir	Wind direction of strongest wind gust in the 24 hours to midnight	categorical	Numbers represent different types	0,1,2,3...,15
WindGustSpeed	Speed of strongest wind gust in the 24 hours to midnight	numerical	Km/h	6.0 – 135.0
WindDir9am	Wind direction averaged over 10 minutes priorbefore	categorical	Numbers represent different types	0,1,2,3...,15
WinDir3pm	Wind direction averaged over 10 minutes prior to 3 pm	categorical	Numbers represent different types	0,1,2,3...,15
WindSpeed9am	Wind speed averaged over 10 minutes prior to 9 am	numerical	Km/h	0.0 – 130.0
WindSpeed3pm	Wind speed averaged over 10 minutes prior to 3 pm	numerical	Km/h	0.0 – 87.0
Humidity9am	Relative humidity at 9 am	numerical	Percentage (%)	0.0 – 100.0
Humidity3pm	Relative humidity at 3 pm	numerical	Percentage (%)	0.0 – 100.0
Pressure9am	Atmospheric pressure reduced to mean sea level at 9 am	numerical	hPa	980.5 – 1041.0
Pressure3pm	Atmospheric pressure reduced to mean sea level at 3 pm	numerical	hPa	977.1 – 1039.6
Cloud9am	Fraction of sky obscured by cloud at 9 am	numerical	Oktas	0,1...,8
Cloud3pm	Fraction of sky obscured by cloud at 3 pm	numerical	Oktas	0,1...,8
Temp9am	Temperature at 3 am	numerical	Celsius (°C)	-7.2 – 40.2
Temp3pm	Temperature at 3 pm	numerical	Celsius (°C)	-5.4 – 46.7
RainTomorrow	Whether or not it will rain tomorrow	categorical	0 for no rain, 1 for rain	0,1
RainToday	Whether or not it rains today	categorical	0 for no rain, 1 for rain	0,1

Table 1 - Data dictionary

During the data pre-processing stage, we conducted label encoding to convert the types of variables into different numbers and used the numbers to represent different types. Also, during this stage, the condition for judging whether it rains is whether the rainfall is larger than or equal to 1mm. If yes, then we consider it as 1 for RainToday. Table 2 illustrates the meaning for the categorical variables at different numbers:

Variable	Explanation
<b>Month</b>	1,2...,12 represents January, February..., December
<b>Climate</b>	0 Cool temperate 1 High humidity summer, warm winter 2 Hot dry summer, cool winter 3 Hot dry summer, warm winter 4 Mild temperate 5 Warm humid summer, mild winter 6 Warm temperate
<b>WindGustDir &amp; WindDir9am &amp; WinDir3pm</b>	0 E (East) 1 ENE (East-northeast) 2 ESE (East-southeast) 3 N (North) 4 NE (Northeast) 5 NNE (North-northeast) 6 NNW (North-northwest) 7 NW (Northwest) 8 S (South) 9 SE (Southeast) 10 NNW (North-northwest) 11 SSW (South-southwest) 12 N (North) 13 W (West) 14 WNW (West-northwest) 15 WSW (West-southwest)
<b>RainTomorrow &amp; RainToday</b>	0 for no rain, 1 for rain

Table 2 - Explanation table for categorical variables

Oktas in the Table 1 is a measurement of the extent to which clouds cover the sky, and 0 represents an absolute clear sky and 8 represents an absolute overcast. Table 3 illustrates the meaning of the unit of Cloud9am and Cloud3pm.

Oktas	Definition	Category
0	Sky clear	Fine
1	1/8 of sky covered or less, but not zero	Fine
2	2/8 of sky covered	Fine
3	3/8 of sky covered	Partly cloudy
4	4/8 of sky covered	Partly cloudy
5	5/8 of sky covered	Partly cloudy
6	6/8 of sky covered	Cloudy
7	7/8 of sky covered or more, but not 8/8	Cloudy
8	8/8 of sky completely covered, no breaks	Overcast

*Table 3 - Explanation table for unit of Cloud9am and Cloud3pm*

## **2. Summary Statistical Analysis:**

	count	mean	std	min	1%	5%	10%	25%	50%	75%	90%	99%	max
Month	142193	6.4	3.43	1	1	1	2	3	6	9	11	12	12
Climate	142193	3.37	2.29	0	0	0	0	1	4	6	6	6	6
MinTemp	142193	12.19	6.39	-8.5	-1.8	1.8	4	7.6	12	16.8	20.8	25.8	33.9
MaxTemp	142193	23.23	7.11	-4.8	9.1	12.8	14.5	17.9	22.7	28.2	32.9	40.1	48.1
Rainfall	142193	2.35	8.42	0	0	0	0	0	0	0.8	6	37.2	371
Evaporation	142193	5.47	3.17	0	0.4	1.2	2	4	5.47	5.47	8.2	15.8	145
Sunshine	142193	7.62	2.73	0	0	1.4	3.7	7.62	7.62	8.7	11.1	13.2	14.5
WindGustDir	142193	8.06	4.69	0	0	0	1	4	9	13	14	15	15
WindGustSpeed	142193	39.98	13.14	6	15	20	24	31	39	46	57	80	135
WindDir9am	142193	7	4.51	0	0	0	1	3	7	11	13	15	15
WindDir3pm	142193	7.79	4.55	0	0	0	1	4	8	12	14	15	15
WindSpeed9am	142193	14	8.85	0	0	0	4	7	13	19	26	39	130
WindSpeed3pm	142193	18.64	8.72	0	2	6	9	13	18.64	24	30	43	87
Humidity9am	142193	68.84	18.93	0	18	34	44	57	70	83	94	100	100
Humidity3pm	142193	51.48	20.53	0	9	17	23	37	51.48	65	79	97	100
Pressure9am	142193	1018	6.75	981	1001	1007	1009	1014	1018	1022	1026	1034	1041
Pressure3pm	142193	1016	6.68	977	998	1004	1006	1011	1015	1019	1024	1031	1040
Cloud9am	142193	5.4	2.59	0	0	0	1	3	7	7	8	8	8
Cloud3pm	142193	5.51	2.43	0	0	1	1	4	7	7	7	8	8
Temp9am	142193	16.99	6.47	-7.2	2.6	7	8.9	12.3	16.8	21.5	25.8	31.2	40.2
Temp3pm	142193	21.69	6.87	-5.4	7.7	11.6	13.3	16.7	21.3	26.3	31.1	38.5	46.7
RainToday	142193	0.24	0.43	0	0	0	0	0	0	0	1	1	1
RainTomorrow	142193	0.22	0.42	0	0	0	0	0	0	0	1	1	1

*Table 4 - Summary statistical table*

After data cleaning and pre-processing, there are no missing values for each variable, the mean, standard deviation, extremum, and the typical value at different levels (1%, 5%... 99% of data) are showed above. Take MinTemp as an instance, 90% of the values of minimum temperate are below or equal to 20.8 °C. Additionally, the standard deviation for each variable is between 0.22 and 20.53, and most variables have a low standard deviation, which indicates now the data quality is not poor.

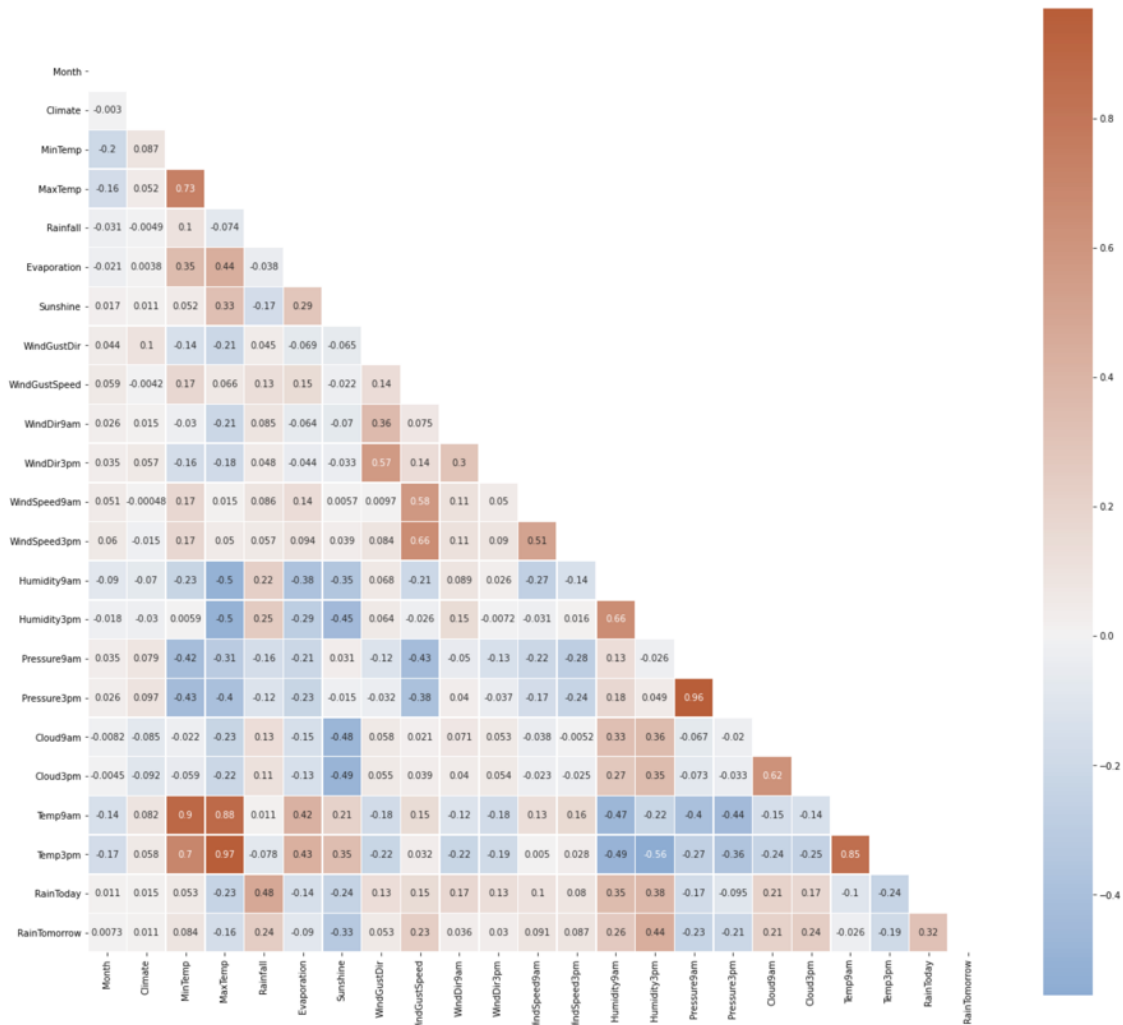


Chart 1 - Correlation Matrix

## 2.1 Correlation Between Target Variable (RainTomorrow) and Input Variables:

According to NurseKillam (2014), two variables are correlated when the correlation coefficient is greater than or equal to 0.20. Based on the correlation matrix, the variables that are positively correlated to the target variable (correlation coefficient  $\geq 0.2$ ) are: Humidity3pm (0.44), RainToday (0.32), Humidity9am (0.26), Rainfall (0.24), Cloud3pm (0.24), WindGustSpeed (0.23), Cloud9am (0.21). Variables that are negatively correlated to the target variable (correlation coefficient  $\leq -0.2$ ) are: Sunshine (-0.33), Pressure9am (-0.23), Pressure3pm (-0.21).

It can be stated that the relative humidity at both 9 am and 3 pm, whether it rains today, precipitation (rainfall) in the 24 hours to 9 am, fraction of sky obscured at both 9 am and 3pm, and speed of strongest wind gust during the day have positive effects on the weather tomorrow (rain or not). The bright sunshine in the 24 hours to midnight, atmospheric pressure reduced to mean sea level at both 9 am and 3 pm, temperature at 3 pm, the maximum temperature in the 24 hours to 9 am have negative effects on rainfall tomorrow.

## 2.2 Correlation Between Input Variables:

As the correlation matrix shows, the following variables have positive correlation: MaxTemp and Temp3pm (0.97), Pressure9am and Pressure3pm (0.96), MinTemp and Temp9am (0.9), MaxTemp and Temp9am (0.88), MinTemp and MaxTemp (0.73), MinTemp and Temp3pm (0.7), Humidity9am and Humidity3pm (0.66), Cloud9am and Cloud3pm (0.62), WindGustDir and WindDir3pm (0.57), WindSpeed9am and WindSpeed3pm (0.51). The following variables have a negative correlation: Humidity3pm and Temp3pm (-0.56). Also, Humidity9am and Temp3pm (-0.49), Sunshine and Cloud3pm (-0.49), Sunshine and Cloud9am (-0.48) have a nearly strong negative correlation.

## 2.3 Correlation Summary

As the indexes in one day will remain relatively stable, it is normal to see the correlation of MaxTemp, MinTemp, Temp9am and Temp3pm are strong, and so does the correlation between Pressure9am and Pressure3pm, between Humidity9am and Humidity3pm, between Cloud9am and Cloud3pm, and of WindGustDir, WindDir9am and WindDir3pm. Furthermore, the temperature usually decreases as humidity increases; thus, it is normal to see their strong negative correlation. For the number of hours of bright sunshine in the 24 hours to midnight (Sunshine) and the fraction of sky obscured by cloud (Cloud9am and Cloud3pm), this is normal relationship of strong negative correlation as the more number of hours of bright sunshine throughout the day, the less sky will be obscured by cloud.

## 3. Exploratory Data Analysis:

### 3.1 Analyzing Categorical Variables:

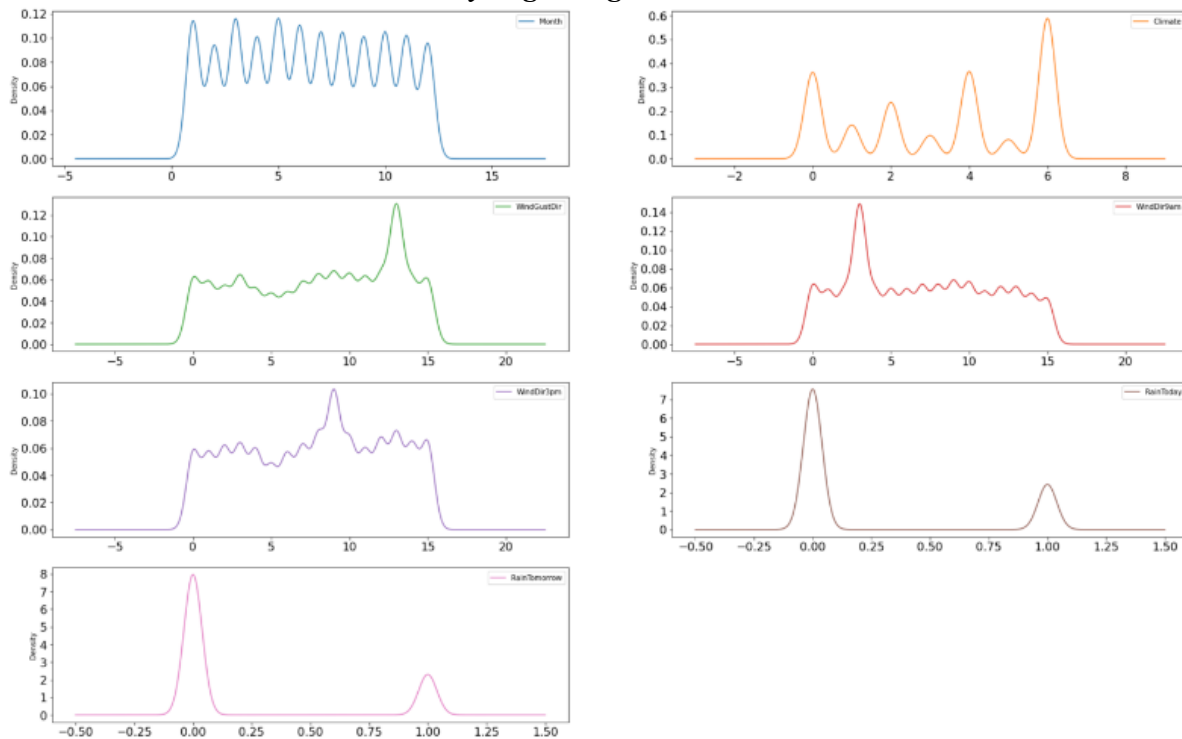


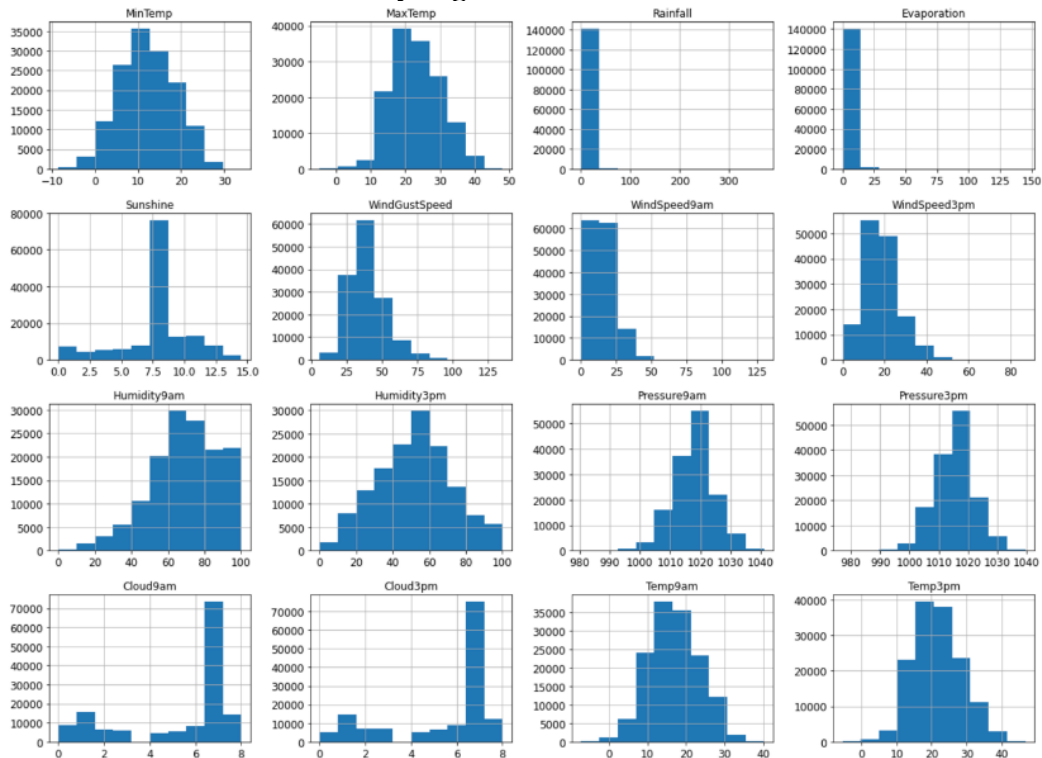
Chart 2 - Categorical variable distribution chart

5	13055	3	21406
3	13036	9	9162
1	12921	0	9024
6	12389	10	8966
10	11804	7	8552
7	11779	8	8493
8	11762	13	8260
11	11461	12	8237
9	11345	5	7948
4	11306	6	7840
12	10810	1	7735
2	10525	2	7558
Name: Month, dtype: int64		4	7527
6	44687	11	7448
4	27768	14	7194
0	27510	15	6843
2	17936	Name: WindDir9am, dtype: int64	
1	10772	9	14441
3	7406	13	9911
5	6114	8	9598
Name: Climate, dtype: int64		15	9329
13	19110	12	9182
9	9309	10	9142
0	9071	3	8667
3	9033	14	8656
10	8993	7	8468
8	8949	2	8382
15	8901	0	8342
12	8797	4	8164
11	8610	11	8010
14	8066	6	7733
7	8003	1	7724
1	7992	5	6444
2	7305	Name: WindDir3pm, dtype: int64	
4	7060	0	107585
6	6561	1	34608
5	6433	Name: RainToday, dtype: int64	
Name: WindGustDir, dtype: int64		0	110316
		1	31877
		Name: RainTomorrow, dtype: int64	

*Chart 3 - Categorical variable distribution map*

The most common type of Climate in the dataset is relatively type 6 (Warm temperate). For WindGustDir, WindDir9am, WindDir3pm, the most common type is respectively type 13 (West), type 3 (North), and type 9 (Southeast).

### 3.2 Analyzing Numerical Variables:



*Chart 4 - Numerical distribution chart*

For most of numerical variables, the distribution is center in a typical value. Many numerical variables have approximately normal distributions, while others are skewed to left or right.

### ***3.3 Summary of Key Analytic Points:***

On the one hand, Humidity, RainToday, Rainfall, Cloud, WindGustSpeed have positive correlations to RainTomorrow, while Sunshine, Pressure, and Temp3pm have negative correlations to RainTomorrow. On the other hand, Evaporation and 5 out of 6 categorical variables (Month, Climate, WindGustDir, WindDir9am, WindDir3pm) don't have significant impacts on tomorrow's raining.

## **Section C: Model Building and Evaluation**

This part will briefly introduce 5 types of machine learning models: Random Forest, KNN, Logistic Regression and Naïve Bayes. Then, we will explain all the parameters related to these models. Furthermore, three types of evaluation strategy will be introduced. The target class of FDR is 0, which means that it will not rain as the client wants to know which day will not rain and choose among these days for the series of open-air theaters.

### ***1. Random Forest***

Random forest is based on decision tree model. It consists of a large number of trees which make prediction independently. Then, the model will select the best one. There are several advantages of the random forest model. Firstly, the random forest can process high dimensionality of features (Shilin, 2019). Our data set includes 23 features and can be considered as relatively high dimensional. Secondly, when creating a random forest, the unbiased estimation is used for the generalization error, and the model has strong generalization ability (Shilin, 2019). However, the drawback of random forest is a black-box model, which means people can only try between different parameters and random seeds and it is difficult to interpret (Shilin, 2019).

Jianping (2016) suggests that the appropriate number of trees is 100, because if the number is too small, it is easy to underfit. Moreover, if number of trees is too large, the calculation amount tends to be huge. After the trees reach a certain amount, the model improvement obtained by increasing number of trees will be very small.

The most intuitive manifestation of over-fitting is that the accuracy of training set is extremely high, while the testing set is remarkable low, that is, the difference between the two is extremely large (Foster and Tom, 2013). The accuracy of testing set is 0.849, which is almost similar to training set (0.856). Therefore, this model has not problem of overfitting. This method is also used to evaluate whether the model is overfitting in the following models.

### ***2. KNN***

K Nearest Neighbor (KNN) means dividing the value into different clusters, and a new example will be classified according to the nearest neighbor rule. One of the advantages is it can be used for nonlinear classification. In addition, KNN is an online model, new data can be added directly to the data set without retraining (Yahia and Ibrahim, 2003). However, the k value has not theoretical optimal value, and it is often combined with K-fold cross-validation to obtain the k value.

Having attempted for several times, we found that when  $k=90$ , which means there are 90 clusters, the model can achieve the best performance. The accuracy of testing set of KNN is 0.850, which has no big difference with training set (0.840). In other words, KNN is not overfitting.

### **3. Logistic Regression**

This model is usually used to predict the probability of a new example belonging to the classification. This model has a good level of interpretability and we can see the impact of different features on the final result from the weight of the feature (Tansey R and et al, 1996). Also, this model fits binary classification problems, because no need to scale input features.

L1 type of regularization is chosen to prevent overfitting. C, the strength of regularization, equals 8, which leads a better performance of Recall. Its testing set's accuracy is 0.845, which almost equals to training set's accuracy (0.835). Hence, logistic regression is not overfitting.

### **4. SVM**

Support vector machines, also called SVM, are linear discriminants, which try to maximize the two class's boundaries. One reason for choosing this model is that SVM can process high-dimensional problems. Moreover, it has strong generalization ability.

Cost is one of its parameters, which is penalty coefficient. The bigger the cost, the easier the model is overfitting. Having tried for several times, cost is 1, which causes the best results. Kernel function includes RBF and Sigmoid, and the latter one has better performance. The testing set's accuracy is 0.729, which is similar to training set (0.769). therefore, SVM is not overfitting.

### **5. Naïve Bayes**

Naïve bayes is also one of the classification methods. The equation of this methods is

$$P\left(\frac{A}{B}\right) \times P(B) = P\left(\frac{B}{A}\right) \times P(A)$$

which means to calculate the probability of C when evidence (E) occurs. The advantage of Naïve Bayes is that saves storage space and computation time efficiently (Foster and Tom, 2013). Moreover, the logic of the algorithm is very simple, and the algorithm is relatively stable, when the data presents different characteristics, the classification performance of Naive Bayes will not have much difference (Kevin, 2006).

The accuracy of testing set is 0.809, and the accuracy of training set is 0.7876. The percentage of difference is 16.45%. This model has chance to be overfitting.

## **6. Evaluation Strategy**

### **6.1 FDR:**

False Discovery Rate (FDR) is false positive divided by (false positive plus true positive). This rate will show the probability that when the weather predicted not to rain tomorrow, the wrong predictions, which is raining tomorrow. Hence, the lower this ratio is, the better the model is. Each model's FDR will illustrate as following:



		Predicted		
		0	1	Σ
Actual	0	28575	1285	29860
	1	4231	4301	8532
	Σ	32806	5586	38392

Chart 5 – confusion matrix of random forest

		Predicted		
		0	1	Σ
Actual	0	28784	1076	29860
	1	4763	3769	8532
	Σ	33547	4845	38392

Chart 6 – confusion matrix of KNN

		Predicted		
		0	1	Σ
Actual	0	28245	1615	29860
	1	4336	4196	8532
	Σ	32581	5811	38392

Chart 7 – confusion matrix of logistic regression

		Predicted		
		0	1	Σ
Actual	0	27634	2226	29860
	1	6650	1882	8532
	Σ	34284	4108	38392

Chart 8 – confusion matrix of SVM

		Predicted		
		0	1	Σ
Actual	0	24679	5181	29860
	1	2972	5560	8532
	Σ	27651	10741	38392

Chart 9 – confusion matrix of Naïve Bayes

Model	Random forest	KNN	Logistic Regression	SVM	Naïve Bayes
FDR	0.1331	0.1419	0.1331	0.1940	0.1075

Table 5 – FDR score of five models

## 6.2 AUC:

AUC is Area Under the ROC curvy. It is the capability of distinguishing between classifications and range of AUC is between 0 and 1. The closer the number is to 1, the higher the separability of the model. For example, random forest's AUC is 0.883, which means there is 88.3% chance the model can recognize positive class and negative one.

Model	Random Forest	KNN	Logistic Regression	SVM	Naïve Bayes
AUC	0.882	0.867	0.859	0.799	0.825

Table 6 – AUC score of five models

Table 6 demonstrates that Random Forest performs best in AUC evaluation.

## 6.3 Accuracy:

Accuracy is the number of all the right decision divided by the whole instances. The higher the value is, the better the model is. For example, random forest's accuracy should equal  $(28575+4301)/38392=0.8563$ . The same method for other models.

Model	Random Forest	KNN	Logistic Regression	SVM	Naïve Bayes
Accuracy	0.8563	0.8479	0.8450	0.7688	0.7876

Table 7 – Accuracy score of five models

Random forest gets the highest score. However, this evaluation method is less to consider, except to evaluate if this model is overfitting or not, because it only calculates the number of correct decisions but cannot reflect the two types of error.

## Section D: Implementation and Business Case Recommendations

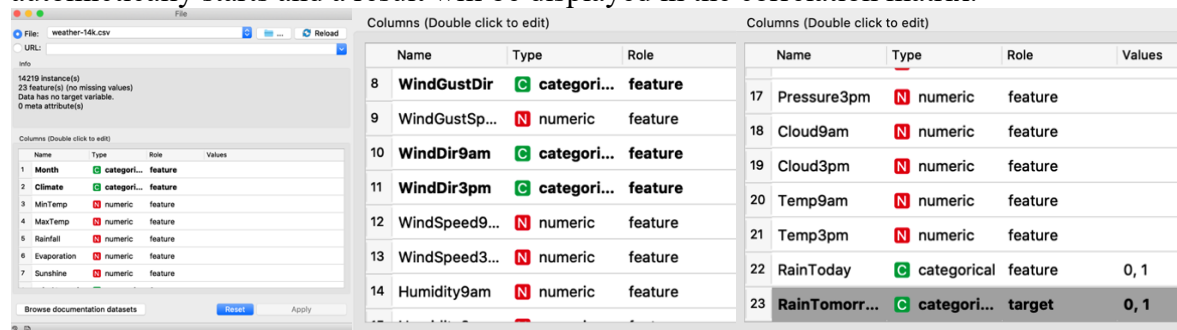
### 1. The 'Winning' Classifier:

Overall, the best model is Random Forest, because its False Discovery Rate (FDR) gets the second-best score (0.1331) and the AUC score and accuracy are respectively 0.882 and 0.8563, which are both ranked as No.1.

Our target is 0 (tomorrow will not rain). Among these evaluations, FDR is the most important criteria to consider, because the cost of False Positive (FR) is relatively high, which is predicting it will not rain while the reality is rainy. In other words, if it the days we predict will not rain, the sponsor will has taken a lot of cost to prepare for the open-air theater. However, the reality is that it rains in these days, which cause the work to become a sink cost and it will bring unpleasant experience for customers. In this evaluation strategy, the lower the FDR, the better the model. Although the Naïve Bayes model has the lowest FDR scores, its performance evaluated by AUC and accuracy is normal. Therefore, we should look at the second-best ones, Random Forest and Logistic Regression. Compared to Logistic Regression, Random Forest perform better in AUC and accuracy, so we choose Random Forest as our best model.

### 2. Instruction on Processing on Test Set using the Model File:

First, open the file of "final model workflow.ows" and load the model "final model.pkcls". Then, select the file "test dataset.csv", and change the type of variable for "Month", "Climate", "WindGustDir", "WindDir9am", "WindDir3pm", "RainToday", and "RainTomorrow" to categorical, and set the role of variable for "RainTomorrow" to target. The predictions will automatically starts and a result will be displayed in the correlation matrix.



The screenshot shows the Orange3 data mining software interface. On the left, the 'Info' panel displays dataset statistics: 14219 instances, 23 features (no missing values), and 0 meta attributes. The 'Columns' panel is open, showing a list of variables with their types and roles. The 'RainTomorrow' variable is highlighted in grey, indicating it is the target variable.

Name	Type	Role	Values
1 Month	categorical	feature	
2 Climate	categorical	feature	
3 MinTemp	numeric	feature	
4 MaxTemp	numeric	feature	
5 Rainfall	numeric	feature	
6 Evaporation	numeric	feature	
7 Sunshine	numeric	feature	
8 WindGustDir	categorical	feature	
9 WindGustSp...	numeric	feature	
10 WindDir9am	categorical	feature	
11 WindDir3pm	categorical	feature	
12 WindSpeed9...	numeric	feature	
13 WindSpeed3...	numeric	feature	
14 Humidity9am	numeric	feature	
17 Pressure3pm	numeric	feature	
18 Cloud9am	numeric	feature	
19 Cloud3pm	numeric	feature	
20 Temp9am	numeric	feature	
21 Temp3pm	numeric	feature	
22 RainToday	categorical	feature	0, 1
23 RainTomorr...	categorical	target	0, 1

### 3. Summary of the Business Case:

Our client is planning to launch a series of open-air theaters for the first time, it needs to know which days will not rain as if the day is rainy then the event will be closed and another day should be chosen, which will produce idle cost and unpleasant experience of customers. The goal is to accurately predict which days will not rain, and make less mistakes by achieving lower

FDR. Our best model, Random Forest has the second lowest FDR of 0.1331, while achieve the highest AUC score of 0.882 and accuracy rate of 0.8563. Therefore, we successfully provide the days will not rain with low chance of making misstates to our client and the event could be scheduled successfully.

#### ***4. Recommendations for Further Potential Analysis:***

We can use more approaches to conduct parameter adjustment of each model like adopting some algorithms to seek the best parameter instead of adjusting manually. We can also predict the range of rainfall and divide the rainy days into different degrees like “light rain”, “medium rain”, and “heavy rain”, so that our client can know more information of the weather condition and make better decision.

The limitations of the dataset and data collection could be the limited information of some variables. For example, there are Temp9am and Temp3pm, but there is not temperate data of other times during a day. Another limitation is that the data is collected in Australia, and the climate types only have 5 types, which means that the clients outside Australia and the climate type of the place not belonged to the 5 types, maybe cannot use this prediction model. Therefore, it is necessary to develop diverse market and provide service to more customers from other countries in future. The dimension of variables is not enough as there are many factors that can influence the weather and we can gain more variables to increase the performance of our models.

## References

Foster, P. and Tom, F. (2013). *Data science for business*, 1st edn. pp 113-114.

Interpreting correlation coefficients in a correlation matrix (2014) YouTube video, added by NurseKillam [Online]. Available at: <https://www.youtube.com/watch?v=qUmmATEJdgM> (Accessed: 9 November, 2020).

Liu, J. (2016). Summary of adjustment of random forest's parameters, available at: <https://www.cnblogs.com/pinard/p/6160412.html> (Accessed: December 6, 2020)

Murphy, K. P. (2006). Naive bayes classifiers. University of British Columbia, 18, 60.

Shilin, J. (2019). The advantages and disadvantages of random forest, available at: [https://blog.csdn.net/qq\\_35290785/article/details/100561148](https://blog.csdn.net/qq_35290785/article/details/100561148) (Accessed: December 6, 2020)

Tansey, R., White, M., Long, R. G., & Smith, M. (1996). A comparison of loglinear modeling and logistic regression in management research. *Journal of management*, 22(2), 339-358.

Yahia, M. E., & Ibrahim, B. A. (2003). K-nearest neighbor and C4. 5 algorithms as data mining methods: advantages and difficulties. *Computer Systems and Applications*, 103.