Binary Data Prediction

for weather forecasting in Southern China

Statistics Department

The University of Hong Kong

PI: Prof. Jian-feng Yao

CI: LIAO Jiayang Susan, CHAN Kwan Yi, CHEN Jiakai

23/12/2020

Table of Contents

1 Introduction

1.1 Background

Recently, climate change and meteorology have received considerable attention with a burgeoning body of research on how the weather relates to society. Rainfall is one of the climate elements. Insufficient rainfall causes drought that reduces food production and endangers animal survival. A moderate amount of rainfall helps to facilitate the water cycle, agricultural industries, and rural populations. And excessive rainfall brings disasters that floods may destroy farmland, residential houses, industrial factories, or even leads to casualties. From Jan. 26 to Feb. 7 in 2019, widespread flooding hit Australia’s Northern Queensland resulting in total economic losses of at least US$1.2 billion (Insurance Journal, 2019). Therefore， rainfall forecasting is indispensable that we could utilize it to arrange the water resources, prevent harm, and ultimately promote social and economic development.

It is recognized that the atmospheric motion could be fickle and the pattern of rainfall is extremely complex that various factors affect the weather. Moreover, climate extremes such as typhoons and El Nin˜o may interfere with rainfall, making the rainfall prediction more challenging (Evans, Bennett, & Ewenz, 2009). Therefore, a rainfall forecasting tool with high computing capability and great flexibility is required. This project aims to build a predictive model to forecast the rainfall tomorrow using today’s climate data. During the analysis process, different data analytics methods will be used.

1.2 Data analytics methods

Descriptive analysis

Descriptive analysis refers to the method that quantitively describes the data's features, summarizes the information or derives observations during the process (Wikipedia, 2020). In this project, descriptive analysis basically shows the feature of climate data.

Correlation Analysis

Correlation analysis is a statistical method used to evaluate the strength of the relationship between two quantitative variables. In this project, this method is used with the tool of the heatmap to extract significant variables related to the target variable.

Hypothesis Testing

“A statistical hypothesis is a hypothesis that is testable on the basis of observed data modeled as the realized values taken by a collection of random variables (Wikipedia, 2020). In this project, the hypothesis test is used to select data.

2 Current Issues and Objectives

The project goal is to build a model to predict whether it will rain tomorrow or not, given the information of certain weather conditions.
Logistic regression models are implemented to analyze the binary response variable RainTomorrow with two values, “yes” or “no”, in the function of a rich set of explanatory climate variables, including humidity, windspeed, temperature, and so on. Once constructed, it can serve as a prediction model for the probability of “rain” for the response variable for new subjects that are not in the data set used. Hence, the model could be applied to forecast rainfall by inputting climate data.

Project Goals

1. Select appropriate data for analysis from your data set
2. Use descriptive analytics to summarize the data
3. Run a correlation analysis to identify the relationships between the selected response and explanatory variables. Consider possible variable transformation.
4. Build the logistic regression prediction model
5. Visualize the analyzed results and model fitting statistics

3 Data Science Methods

3.1 Data sources, data cleansing, and pre-processing

The original data source consists of 24 variables and 142,193 rows of data with different data types, among which most variables have different portion of N/A values. A summary is generated to have an overall image of the characteristics of the data source.

To get rid of unwanted variables and maintain the concision and accuracy of the data frame, data cleansing is conducted. First, given that the final goal is to predict whether it will rain tomorrow, the variable “RISK_MM” is removed since it stands for the predicted value of the quantity of rain tomorrow. Outliers are dropped by calculating the interquartile range. To deal with the N/A values, variables are divided into categorical ones and continuous ones, and N/A values are filled with mode for categorical ones and mean for continuous ones. By creating a heatmap of correlation, redundant variables that have a high correlation with other variables are removed, as high correlation indicates the two variables have a similar contribution to the final result.

Since some data are not numerical values, data transformation is necessary for

pre-processing for building the logistic regression model. The core of data transformation is simply transforming the non-numerical values into numerical ones. For the variable “Date”, a simple concept of date is useless for the construction of the model. But the position of the date in a year is what really matters. Considering using numbers 1 to 365 to represent the dates, there will be too many categories and only a few samples for each. Therefore, months are used to replace the dates for a trade-off between the number of samples in each category and the accuracy of categorization. Locations are grouped by their climate types and characteristics in reality and numbered. As for other non-numerical variables with categorical values, OridinalEncoder from sklearn module is used to change them into different numbers.

3.2 Models Design

Descriptive analytics

Data frame summaries are generated by “df.info()” and “df.describe()” to get the overall information about the whole data frame.

A heatmap of correlation is utilized to indicate the correlation between each pair of variables to detect the redundant variables.

Confusion matrix, accuracy, precision, recall, F1-score, data visualization, and ROC curve are used to evaluate different models.

Input and output variables

Input variables:

Month: the month of the date.

Location: the location group the place belongs to.

MinTemp: minimum temperature in the day.

MaxTemp: maximum temperature in the day.

Rainfall: the quantity of rainfall of the day.

Evaporation: the so-called Class A pan evaporation in the day.

Sunshine: the number of hours of sunshine in the day.

WindGustSpeed: the speed of the strongest wind gust in the day.

WindDir9am: wind direction at 9pm transformed into number.

WindSpeed9am: wind speed averaged over 10 minutes prior to 9am.

WindSpeed3pm: wind speed averaged over 10 minutes prior to 3pm.

Humidity3pm: humidity at 3pm.

Pressure3pm: atmospheric pressure reduced to mean sea level at 3pm.

Cloud3pm: fraction of sky obscured by cloud at 3pm.

RainToday: 1 if precipitation in the 24 hours to 9am exceeds 1mm, otherwise 0.

Output variable:

RainTomorrow: the target variable indicating whether it will rain tomorrow.

Predictive Models Design

Exploratory:

Firstly, we will do a correlation analysis to understand the relationship between the input variables and the target variable (RainTomorrow). This is important for us to know how strong the relationship between the factors and whether tomorrow will rain. Besides, by dropping highly-correlated variables, we can lower the biased error of the final model.

Predictive:

Secondly, we will make use of maximum likelihood estimation to build a predictive logistic regression model (Logit) on binary variable RainTomorrow. Then, we do evaluations on different train-test set splitting to prevent overfitting.

Data Visualization

In data science, visualization is extremely crucial to interpreting useful information from plain data. Various charts and graphs are used to show the analytical results and reveal the correlation between the variables.

This project uses several plots to make statistical results more intuitive:

Heatmap: Virtualize the correlation among independent variables and target variable RainTomorrow.

Scatter plot: Virtualize `actual response` vs. `predicted response` in the evaluation stage of the logistic regression model.

Confusion matrices: Virtualize four types of predictive results of binary variables.

Bar chart: Virtualize and compare the “Accuracy”, “Precision”, “Recall”, and “F1-Score” of different regression models in a more intuitive way.

Hypothesis Testing

The hypothesis of the project is all variables have significant relationships under a 5% level of significance in the logistic regression.

4 Result Summary

The coding starts with importing libraries and the data frame. Then, the first five columns are shown to get a first glimpse of the database. As mentioned, the primitive data has 24 columns and 142,193 rows, and their data types are as follows.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 142193 entries, 0 to 142192
Data columns (total 24 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   Date           142193 non-null  object
 1   Location       142193 non-null  object
 2   MinTemp        141556 non-null  float64
 3   MaxTemp        141871 non-null  float64
 4   Rainfall       140787 non-null  float64
 5   Evaporation    81350 non-null   float64
 6   Sunshine       74377 non-null   float64
 7   WindGustDir    132863 non-null  object
 8   WindGustSpeed  132923 non-null  float64
 9   WindDir9am     132180 non-null  object
 10  WindDir3pm     138415 non-null  object
 11  WindSpeed9am   140845 non-null  float64
 12  WindSpeed3pm   139563 non-null  float64
 13  Humidity9am    140419 non-null  float64
 14  Humidity3pm    138583 non-null  float64
 15  Pressure9am    128179 non-null  float64
 16  Pressure3pm    128212 non-null  float64
 17  Cloud9am       88536 non-null   float64
 18  Cloud3pm       85099 non-null   float64
 19  Temp9am        141289 non-null  float64
 20  Temp3pm        139467 non-null  float64
 21  RainToday      140787 non-null  object
 22  RISK_MM        142193 non-null  float64
 23  RainTomorrow   142193 non-null  object
dtypes: float64(17), object(7)
memory usage: 26.0+ MB
```

(summary of data source)

Since variable RISK_MM is a predicted value based on our final prediction RainTomorrow, we drop it here. After the removal of rows with outliers and "RISK_MM", there are 23 variables left, and the number of rows decreases.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 109459 entries, 0 to 142192
Data columns (total 23 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   Date           109459 non-null  object
 1   Location       109459 non-null  object
 2   MinTemp        108929 non-null  float64
 3   MaxTemp        109197 non-null  float64
 4   Rainfall       108169 non-null  float64
 5   Evaporation    62134 non-null   float64
 6   Sunshine       56746 non-null   float64
 7   WindGustDir    102339 non-null  object
 8   WindGustSpeed  102376 non-null  float64
 9   WindDir9am     100997 non-null  object
 10  WindDir3pm     106521 non-null  object
 11  WindSpeed9am   108362 non-null  float64
 12  WindSpeed3pm   107415 non-null  float64
 13  Humidity9am    108160 non-null  float64
 14  Humidity3pm    106722 non-null  float64
 15  Pressure9am    98479 non-null   float64
 16  Pressure3pm    98509 non-null   float64
 17  Cloud9am       66779 non-null   float64
 18  Cloud3pm       64108 non-null   float64
 19  Temp9am        108806 non-null  float64
 20  Temp3pm        107389 non-null  float64
 21  RainToday      108169 non-null  object
 22  RainTomorrow   109459 non-null  object
dtypes: float64(16), object(7)
memory usage: 20.0+ MB
```

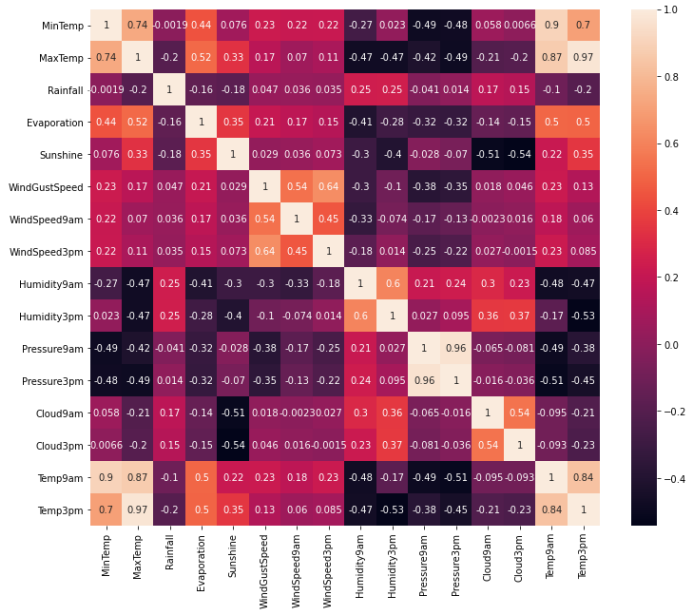
(summary of the data frame after removing outliers)

However, the number of non-null columns varies from variable to variable. The SimpleImputer in sklearn.impute is used for N/A filling. Given that some variables have categorical values, like wind directions, while others may have continuous values, the categorical data is filled with the mode, and interval data is filled with the mean. By filing all the N/A values, the number of non-null values for each variable is 109,459.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 109459 entries, 0 to 142192
Data columns (total 23 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   Date           109459 non-null  object
 1   Location       109459 non-null  object
 2   MinTemp        109459 non-null  float64
 3   MaxTemp        109459 non-null  float64
 4   Rainfall       109459 non-null  float64
 5   Evaporation    109459 non-null  float64
 6   Sunshine       109459 non-null  float64
 7   WindGustDir    109459 non-null  object
 8   WindGustSpeed  109459 non-null  float64
 9   WindDir9am     109459 non-null  object
 10  WindDir3pm     109459 non-null  object
 11  WindSpeed9am   109459 non-null  float64
 12  WindSpeed3pm   109459 non-null  float64
 13  Humidity9am    109459 non-null  float64
 14  Humidity3pm    109459 non-null  float64
 15  Pressure9am    109459 non-null  float64
 16  Pressure3pm    109459 non-null  float64
 17  Cloud9am       109459 non-null  float64
 18  Cloud3pm       109459 non-null  float64
 19  Temp9am        109459 non-null  float64
 20  Temp3pm        109459 non-null  float64
 21  RainToday      109459 non-null  object
 22  RainTomorrow   109459 non-null  object
dtypes: float64(16), object(7)
memory usage: 20.0+ MB
```

(summary of the data frame after filling N/A values)

Then, a heatmap of correlation is created. It is observed that the correlation between “MaxTemp” and “Temp3pm” is extraordinarily high, which is 0.97. So does the correlation between “Pressure9am” and “Pressure3pm”. This means one of the variables in the pairs must be redundant as the pair contributes similarly to the prediction of whether it will rain tomorrow. Therefore, the ones with originally more N/A values, namely “Temp3pm” and “Pressure9am”, are dropped from the data frame.



(heatmap of correlation)

Till now, all the procedures of data cleansing are accomplished. Non-numerical values are then transformed into numerical ones. Variable “Date” is transformed into “Month”, by extracting the month in the original date information.

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	...
0	2008-12-01	Albury	13.4	22.9	0.6	5.27386	8.120911	W	44.0	W	...
1	2008-12-02	Albury	7.4	25.1	0.0	5.27386	8.120911	WNW	44.0	NNW	...
2	2008-12-03	Albury	12.9	25.7	0.0	5.27386	8.120911	WSW	46.0	W	...
3	2008-12-04	Albury	9.2	28.0	0.0	5.27386	8.120911	NE	24.0	SE	...
4	2008-12-05	Albury	17.5	32.3	1.0	5.27386	8.120911	W	41.0	ENE	...

(part of the data frame before the transformation of “Date”)

	Month	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	...
0	12	Albury	13.4	22.9	0.6	5.27386	8.120911	W	44.0	W	...
1	12	Albury	7.4	25.1	0.0	5.27386	8.120911	WNW	44.0	NNW	...
2	12	Albury	12.9	25.7	0.0	5.27386	8.120911	WSW	46.0	W	...
3	12	Albury	9.2	28.0	0.0	5.27386	8.120911	NE	24.0	SE	...
4	12	Albury	17.5	32.3	1.0	5.27386	8.120911	W	41.0	ENE	...

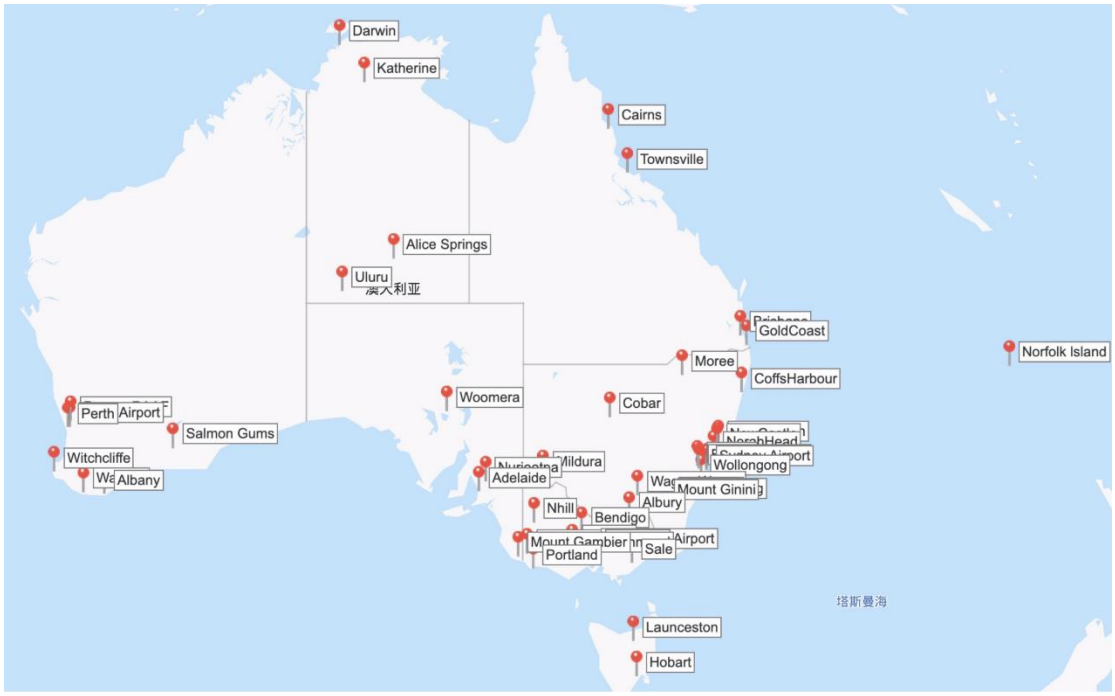
(part of the data frame after the transformation of “Date”)

All other categorical variables except for locations, for example, wind direction, whether it rained on that day, are encoded into numerical symbols with the help of OrdinalEncoder from sklearn module.

	Location	WindGustDir	WindGustSpeed	WindDir9am	RainToday	RainTomorrow
0	Albury	13.0	44.0	13.0	0.0	0.0
1	Albury	14.0	44.0	6.0	0.0	0.0
2	Albury	15.0	46.0	13.0	0.0	0.0
3	Albury	4.0	24.0	9.0	0.0	0.0
4	Albury	13.0	41.0	1.0	0.0	0.0

(categorical data after transformation)

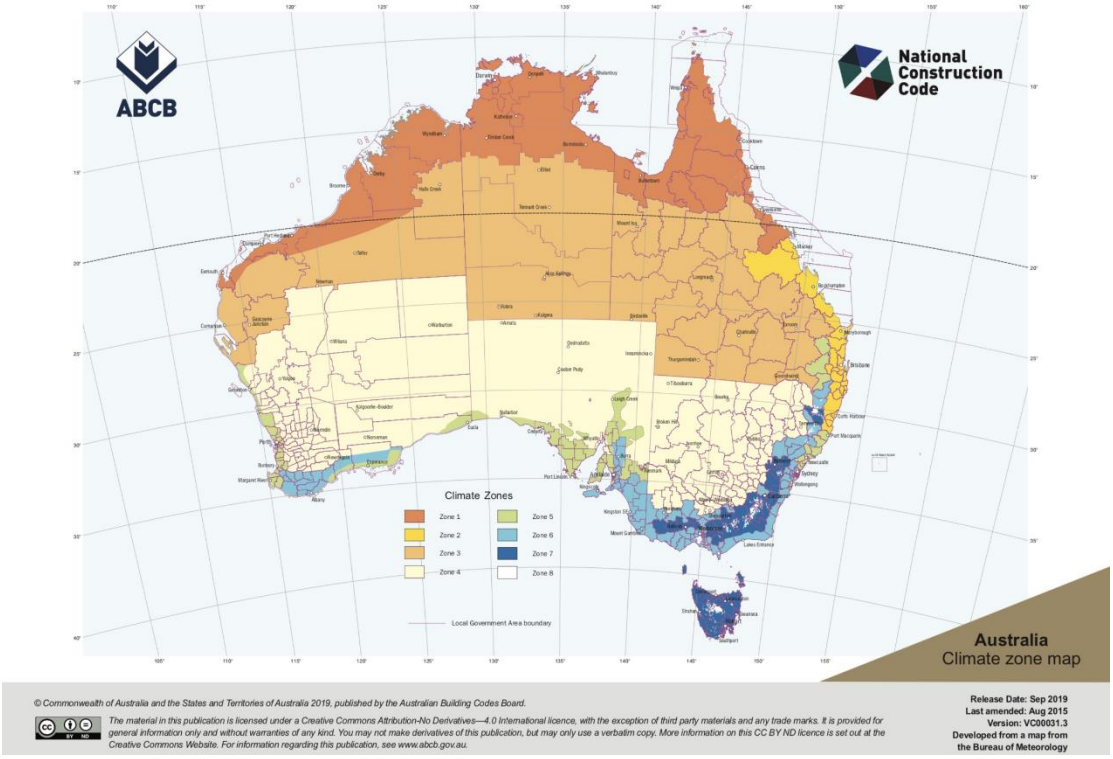
As for the transformation of the variable “Location”, the locations are first divided into different groups. After counting down the “Location” column, there are 142,193 values in total. Some values are from the same location. Therefore, these data are actually from 49 weather stations in Australia. Different weather stations are located in different climate conditions. It is known that the climate condition will affect the rainfall, which is a relatively stable factor. For example, coastal cities like Shanghai have more rainfall than desert places like Xinjiang. The situation is also similar in this Australian case.



(Australian map with 49 weather stations marked)

All 49 weather stations are marked on one map for observing directly. It is found that some stations are definitely located in one similar climate condition (e.g., Sydney and Sydney Airport), while some are totally different (e.g., Sydney and Uluru). To

simplify the model, the 49 locations of weather stations can be grouped by their weather types and features.



(Australia climate zone map)

In consideration of the rationality of this method, according to Australia Climate Zone Map (ABCB, 2019), 48 locations are grouped into eight climate zones, and Norfolk Island, an offshore island, is a special case, which is set to be in Zone 9. Thus, all the locations are transformed into zone numbers from 1 to 9. For example, Sydney and Perth are all in Warm Temperate (Zone 5); while Uluru is in Hot Dry Summer and Cool Winter (Zone 3). This can eliminate the interference of geographical environment differences and improve the accuracy of the rainfall prediction.

	Month	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	...
0	12	4	13.4	22.9	0.6	5.27386	8.120911	13.0	44.0	13.0	...
1	12	4	7.4	25.1	0.0	5.27386	8.120911	14.0	44.0	6.0	...
2	12	4	12.9	25.7	0.0	5.27386	8.120911	15.0	46.0	13.0	...
3	12	4	9.2	28.0	0.0	5.27386	8.120911	4.0	24.0	9.0	...
4	12	4	17.5	32.3	1.0	5.27386	8.120911	13.0	41.0	1.0	...

(part of the data frame after transformation of locations)

After all data processing work, we separate the labels (target variable: RainTomorrow) and variables. And for convenience in model fitting, we transform all data types to

“float64” based on the above work. Next, according to later optimization method, we split the train-test set by test_size=0.05, which is neither overfitted nor underfitted comparing to 0.01 and 0.1. Then, we reshape the data frame.

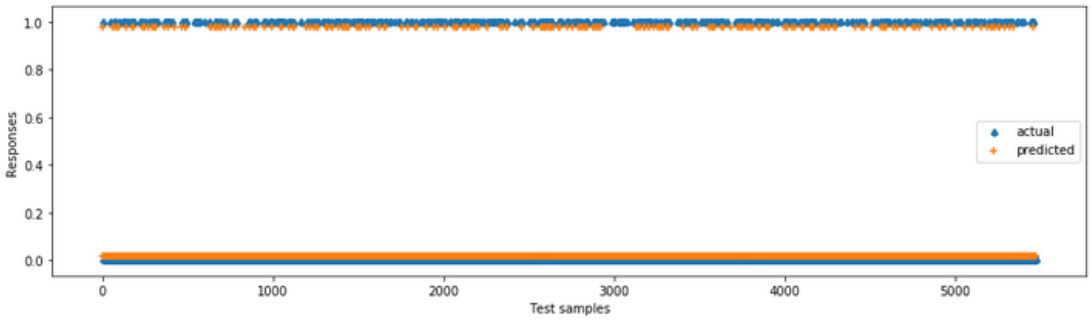
```
# Reshape the dataframe
for i in [Xtrain, Xtest, Ytrain, Ytest]:
    i.index = range(i.shape[0])
display(Xtrain.shape, Ytrain.shape, Xtest.shape, Ytest.shape)
```

```
(103986, 20)
(103986,)
(5473, 20)
(5473,)
```

(reshaping the data frame of train-test set)

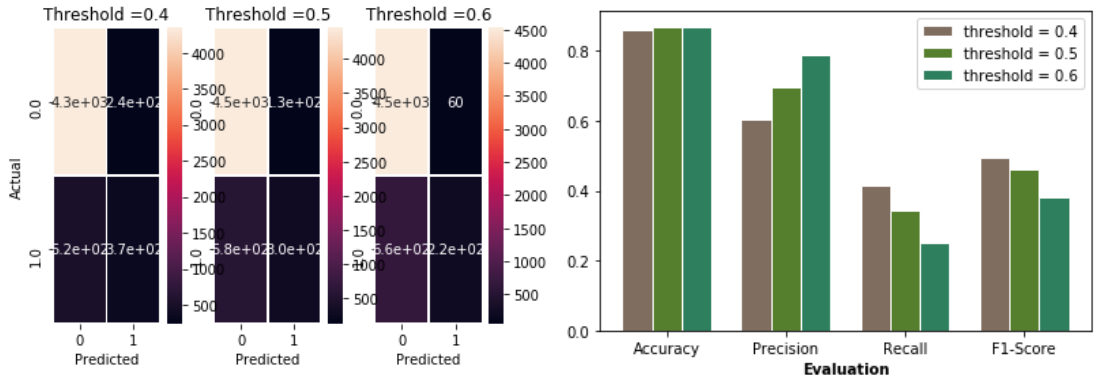
The Maximum Likelihood method is applied to fit the logit model under the hypothesis of a 0.05 level of significance. The process of dropping insignificant variables under the null hypothesis will be further elaborated in Chapter 5.

After model fitting, predictions are made on the test dataset based on the fitted logistic regression model. Here, three different predictive lasses are generated with thresholds equal to 0.4, 0.5, and 0.6. A scatter plot of actual response versus predicted response is created to roughly display the accuracy of the regression model.



(comparison of actual and predicted result)

Then we move on to detailed evaluation from accuracy, precision, recall, and F1_score through their confusion matrices.



(confusion matrices and evaluation of thresholds)

5 Data Interpretation and Discussion

5.1 Hypothesis Testing

Hypothesis testing is applied in the process of MLE model fitting of Logit. In model fitting, for each variable, the Null Hypothesis H_0 is "the variable has a significant relationship with the MLE model" at 5% level of significance.

1st Model fitting All 21 variables

```
# Logistic regression model from 1015
Xtrainc = sm.add_constant(Xtrain)
log_reg = sm.Logit(Ytrain, Xtrainc).fit()
log_reg.summary()
```

```
Optimization terminated successfully.
         Current function value: 0.326691
         Iterations 7
```

Logit Regression Results

Dep. Variable:	RainTomorrow	No. Observations:	103986
Model:	Logit	Df Residuals:	103965
Method:	MLE	Df Model:	20
Date:	Thu, 03 Dec 2020	Pseudo R-squ.:	0.2599
Time:	20:02:29	Log-Likelihood:	-33971.
converged:	True	LL-Null:	-45899.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	79.4344	1.971	40.299	0.000	75.571	83.298
Month	0.0184	0.003	6.127	0.000	0.013	0.024
Location	-0.0119	0.006	-1.872	0.061	-0.024	0.001
MinTemp	-0.0067	0.005	-1.414	0.157	-0.016	0.003
MaxTemp	0.0107	0.005	2.270	0.023	0.001	0.020
Rainfall	0.4409	0.042	10.424	0.000	0.358	0.524
Evaporation	-0.0199	0.006	-3.611	0.000	-0.031	-0.009
Sunshine	-0.1130	0.005	-24.766	0.000	-0.122	-0.104
WindGustDir	-0.0005	0.002	-0.191	0.848	-0.005	0.004
WindGustSpeed	0.0547	0.001	43.752	0.000	0.052	0.057
WindDir9am	-0.0175	0.003	-6.963	0.000	-0.022	-0.013
WindDir3pm	-0.0023	0.003	-0.873	0.383	-0.008	0.003
WindSpeed9am	-0.0138	0.002	-8.652	0.000	-0.017	-0.011
WindSpeed3pm	-0.0301	0.002	-18.570	0.000	-0.033	-0.027
Humidity9am	0.0006	0.001	0.586	0.558	-0.001	0.003
Humidity3pm	0.0521	0.001	56.585	0.000	0.050	0.054
Pressure3pm	-0.0838	0.002	-43.943	0.000	-0.088	-0.080
Cloud9am	-0.0038	0.006	-0.656	0.512	-0.015	0.008
Cloud3pm	0.1391	0.006	22.168	0.000	0.127	0.151
Temp9am	-0.0064	0.007	-0.943	0.346	-0.020	0.007
RainToday	-0.3388	0.069	-4.876	0.000	-0.475	-0.203

2nd Model fitting All 16 variables

```
Optimization terminated successfully.
         Current function value: 0.326709
         Iterations 7
```

Logit Regression Results

Dep. Variable:	RainTomorrow	No. Observations:	103986
Model:	Logit	Df Residuals:	103970
Method:	MLE	Df Model:	15
Date:	Thu, 03 Dec 2020	Pseudo R-squ.:	0.2598
Time:	20:02:30	Log-Likelihood:	-33973.
converged:	True	LL-Null:	-45899.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	79.2563	1.959	40.448	0.000	75.416	83.097
Month	0.0177	0.003	5.965	0.000	0.012	0.023
Location	-0.0114	0.006	-1.839	0.066	-0.024	0.001
MinTemp	-0.0104	0.003	-3.126	0.002	-0.017	-0.004
MaxTemp	0.0082	0.004	2.282	0.022	0.001	0.015
Rainfall	0.4454	0.042	10.685	0.000	0.364	0.527
Evaporation	-0.0209	0.005	-3.835	0.000	-0.032	-0.010
Sunshine	-0.1125	0.004	-25.679	0.000	-0.121	-0.104
WindGustSpeed	0.0545	0.001	44.255	0.000	0.052	0.057
WindDir9am	-0.0185	0.002	-7.953	0.000	-0.023	-0.014
WindSpeed9am	-0.0140	0.002	-9.021	0.000	-0.017	-0.011
WindSpeed3pm	-0.0303	0.002	-18.928	0.000	-0.033	-0.027
Humidity3pm	0.0523	0.001	67.415	0.000	0.051	0.054
Pressure3pm	-0.0836	0.002	-44.032	0.000	-0.087	-0.080
Cloud3pm	0.1376	0.006	23.185	0.000	0.126	0.149
RainToday	-0.3423	0.069	-4.941	0.000	-0.478	-0.207

3rd Model fitting All 15 variables

```
Optimization terminated successfully.
         Current function value: 0.326726
         Iterations 7
```

Logit Regression Results

Dep. Variable:	RainTomorrow	No. Observations:	103986
Model:	Logit	Df Residuals:	103971
Method:	MLE	Df Model:	14
Date:	Thu, 03 Dec 2020	Pseudo R-squ.:	0.2598
Time:	20:02:31	Log-Likelihood:	-33975.
converged:	True	LL-Null:	-45899.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	79.1478	1.958	40.419	0.000	75.310	82.986
Month	0.0183	0.003	6.241	0.000	0.013	0.024
MinTemp	-0.0096	0.003	-2.912	0.004	-0.016	-0.003
MaxTemp	0.0094	0.004	2.669	0.008	0.002	0.016
Rainfall	0.4452	0.042	10.682	0.000	0.364	0.527
Evaporation	-0.0212	0.005	-3.896	0.000	-0.032	-0.011
Sunshine	-0.1126	0.004	-25.711	0.000	-0.121	-0.104
WindGustSpeed	0.0541	0.001	44.668	0.000	0.052	0.057
WindDir9am	-0.0181	0.002	-7.825	0.000	-0.023	-0.014
WindSpeed9am	-0.0139	0.002	-8.967	0.000	-0.017	-0.011
WindSpeed3pm	-0.0300	0.002	-18.839	0.000	-0.033	-0.027
Humidity3pm	0.0523	0.001	67.440	0.000	0.051	0.054
Pressure3pm	-0.0836	0.002	-44.030	0.000	-0.087	-0.080
Cloud3pm	0.1375	0.006	23.151	0.000	0.126	0.149
RainToday	-0.3418	0.069	-4.934	0.000	-0.478	-0.206

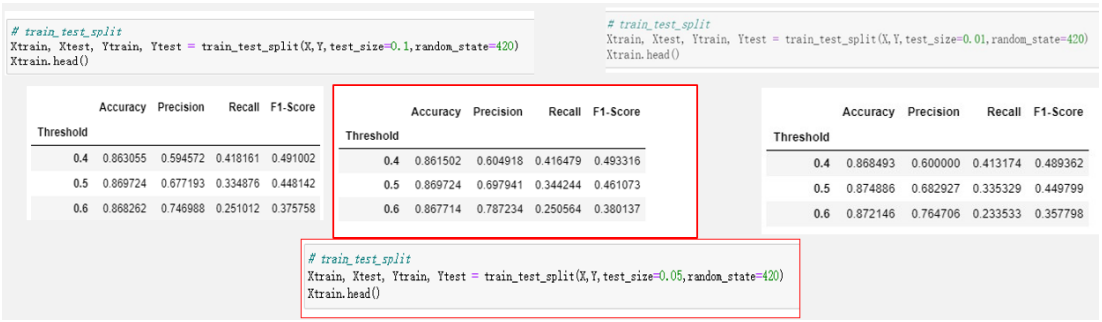
(summary of three regression models)

In three model fitting, we gradually drop all insignificant variables under the α=0.05 null hypothesis. The third model is the finalized Logit model by MLE model fitting.

5.2 Regression

Optimized Logistic regression is constructed for RainTomorrow. The optimization strategies include:

1. Split train-test set under three different proportions (0.01, 0.05, 0.1). Find the split method that neither overfitting nor underfitting, which turns out to be 0.05 as follow.



Threshold	Accuracy	Precision	Recall	F1-Score
0.4	0.863055	0.594572	0.418161	0.491002
0.5	0.869724	0.677193	0.334876	0.448142
0.6	0.868262	0.746988	0.251012	0.375758

Threshold	Accuracy	Precision	Recall	F1-Score
0.4	0.861502	0.604918	0.416479	0.493316
0.5	0.869724	0.697941	0.344244	0.461073
0.6	0.867714	0.787234	0.250564	0.380137

Threshold	Accuracy	Precision	Recall	F1-Score
0.4	0.868493	0.600000	0.413174	0.489362
0.5	0.874886	0.682927	0.335329	0.449799
0.6	0.872146	0.764706	0.233533	0.357798

(effects of different train/test proportions)

2. Fit the logistic regression model by MLE method under 5% significant level as above. Find the relations between the left-over variables and the target variable.

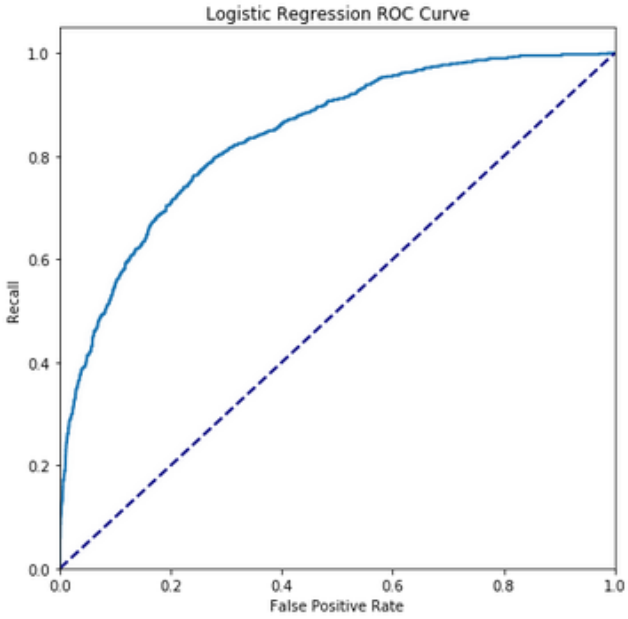
$$RainTomorrow = 79.1478 + 0.0183 * Month - 0.0096 * MinTemp + 0.0094 * MaxTemp + 0.4452 * Rainfall - 0.0212 * Evaporation - 0.1126 * Sunshine + 0.0541 * WindGustSpeed - 0.0181 * WindDir9am - 0.0139 * WindSpeed9am - 0.03 * WindSpeed3pm + 0.0523 * Humidity3pm - 0.0836 * Pressure3pm + 0.1375 * Cloud3pm - 0.3418 * RainToday$$

3. Use different thresholds (0.4,0.5,0.6) to achieve the better recall with the optimized Logit equation in step 2, which is 0.4 here.

Threshold	Accuracy	Precision	Recall	F1-Score
0.4	0.861502	0.604918	0.416479	0.493316
0.5	0.869724	0.697941	0.344244	0.461073
0.6	0.867714	0.787234	0.250564	0.380137

(effects of different thresholds)

4. Use the ROC curve to evaluate the model again. It is observed that with a threshold of 0.4, the AUC is quite large in the graph, which indicates that the model is quite efficient.



(ROC curve)

5.3 Further reflections

Outliers

In the process of data cleansing, columns with data out of the range (Q1-1.5*IQR, Q3-1.5*IQR) are all deleted from the data frame. Removing outliers is a common procedure in the analysis of data. However, it may not be necessary to remove all the outliers in this specific project. Since the weather data are recorded by authoritative institutions, the values are most credible. Although there will be some errors among the massive data, the main bias of the outliers comes from some extreme weather conditions in reality. From this point of view, if the correctly recorded extraordinary values are recognized as outliers and removed from the training data, the model will actually lose the capability to predict the probability of raining tomorrow under some abnormal weather conditions. Also, removing the outliers leads to a loss of almost

one-third of the original data. The dramatic decrease in the quantity of data may hinder the effectiveness of model fitting and then decreasing the accuracy. On the other hand, retaining these “outliers” will probably jeopardize the performance of the model when the input variables are under common weather conditions. Thus, further investigations on how to define the “outliers” in this project are needed to reach a balance of model performance between normal and abnormal conditions.

Recall

With the evaluation and improvements of the model, all kinds of parameters are optimized to maximize the recall. Although the reliability of the final model is already improved under the current structure, the final recall is still not ideal. Further research could investigate how to adjust other parameters to make a qualitative improvement ensuring a higher recall .

6 Conclusion

Through the rational use of different analysis methods, the final model is built and evaluated by its accuracy and rationality.

In the descriptive analysis, a conclusion is reached that the rainfall tomorrow can be predicted by analyzing the influencing factors today. In the correlation analysis, by analyzing the positive and negative correlation of each variable to the prediction result and the strength of the correlation, there are 15 effective variables that correlated with the researching goals. In the hypothesis testing, all the variables used in this project have significant relationships under a 5% level of significance in the logistic regression. In the prescriptive analysis, after splitting the train-test set and fitting the logistic regression model by MLE method under 5% significant level, the final model equation is concluded and evaluated to find the better recall so that to achieve the most accurate and effective model.

For the project objectives, pre-set objectives are all accomplished. Descriptive

analysis is used to evaluate the way and extent of different influencing factors. After that, hypotheses are proposed based on the correlation of different variables. In the end, a more reasonable prediction model is built and visualized using a heatmap, scatter plot, confusion matrices, and bar chart.

In conclusion, the project built a model with the variables of today's atmospheric and geographical influencing factors that can predict whether it will rain tomorrow. Based on the analysis, the parameters are optimized to maximize the recall, which enhances the reliability of the model. Even though there are limitations like the dilemma of removing outliers and the recall of the model, the project is still hoped to promote the further development and deepening of meteorological prediction.

7 References

Australian Building Codes Board, (2019), Climate zone map: Australia wide, Australian Building Codes Board, https://www.abcb.gov.au/Resources/Tools-Calculators/Climate-Zone-Map-Australia-Wide

Evans, Alexander D., Bennett, John M., & Ewenz, Caecilia M. (2009). South Australian rainfall variability and climate extremes. Climate Dynamics,33(4), 477-493.

Insurance Journal. (2019, March 8). Australia Floods Cost Insurers US$635M in February: Aon's Catastrophe Recap. Retrieved from https://www.insurancejournal.com/news/international/2019/03/08/520022.htm

Wikipedia. (2020, November 29). Statistical hypothesis testing. Retrieved from https://en.wikipedia.org/wiki/Statistical_hypothesis_testing

Wikipedia. (2020, October 31). Descriptive statistics. Retrieved fromhttps://en.wikipedia.org/wiki/Descriptive_statistics