WEATHER FORECAST IN AUSTRALIA

Report date: Jun 10, 2020

INDEX

EXECUTIVE SUMMARY

With an interest in weather changes in Australia and the importance of rain prediction, I was motivated to do a research on Australia's weather.

This research aims to answer two questions with two corresponding analyses. First, I performed clustering analysis to find out locations across Australia that have similar weather conditions. Second, I build a predictive model to forecast the possibility of rain on the following day. As part of the analysis, I used a weather dataset from Australia's Bureau of Meteorology that included 142,193 observations at 49 locations all over Australia, from 2007 to 2017, from which I sampled data in one most recent year between Jul 2016 and Jun 2017.

Regarding clustering analysis, I used k-means clustering with four clusters. The clustering result well exhibits impacts of geographical locations on weather conditions. Particularly, the surveyed locations can be grouped into four clusters. The first group includes coastal and island cities that experience strong wind and high humidity. The second group features inland cities that are not so far from the sea. The cities have moderate temperature with high daily variations and fresh breeze only. The third group is comprised of cities in the middle of Australia that experience high temperature and extremely low rainfall amount. The last group includes the northernmost cities that have highest temperatures and rainfall amounts among the locations of concern. However, the data used for clustering is from 2017; thus, this result may not be representative for Australia right now.

Concerning the rain predictive model, I employed Support Vector Machine with radial kernel. I used recursive feature elimination to reduce the number of variables and got a balanced model with accuracy level being slightly over 83%. Though the model performance is relatively good, it is still a lot lower than other weather predictive models. Typically, one-day forecast should accurately predict over 90%. This large accuracy gap could possibly result from the omittance of important variables like evaporation, sunshine, and cloud that can help a lot to explain rain possibility. Additionally, due to a lack of infrastructure resources, I failed to tune hyperparameters for the model to improve overall performance.

INTRODUCTION

Weather is among the events that create tremendous impacts on human's daily activities. On the one hand, weather supports human's life in some dimensions; for example, being a source of water for food crops, or a resource for wind energy. On the other hand, weather is also a disastrous phenomenon when it brings about flood or drought, causing lots of difficulties to people's lives. This well explains why weather prediction is of a major interest to people. The purpose of forecasting weather changes is to minimize negative impacts of weather hazards and to exploit full potential support from good weather conditions. On a daily basis, the availability and accuracy of weather information also provides a great deal of support. In particular, people can make more appropriate plans when they can tell what the weather is going to be like on the following day; for example, they can either cancel their outdoor activities or make further preparation so that their plans take place smoothly if they know that it is going to rain on that day. However, as a matter of fact, predicting weather has been known as not an easy task with complex data and a lot of noise.

Australia's weather has changed significantly over the years with increasing number of heat events and increasing severity of drought conditions due to below-average rainfall. Among the top ten warmest years on record, eight have occurred since 2005. In late December 2019 - early January 2020, Australia suffered from the most severe bushfires ever that were reported to have burnt over ten million hectares of land in southern regions. Bushfires happen owing to different factors. Strong winds, low humidity, and high temperatures altogether contribute to higher frequency of fire weather days. According to a report by CSIRO, Australia's national science research agency, southern and eastern Australia witnessed record low rainfall and high temperatures in 2019.

OBJECTIVE

Considering the significance of weather prediction and Australia's weather changes, I was motivated to conduct a study on Australia' weather status. The study serves two purposes:

1/ To find out similar weather patterns among different locations all over Australia.

2/ To build a classification model to answer the question "Is it going to rain tomorrow?".

DATA EXPLORATION

Overview

To meet the two mentioned objectives, I employed a weather dataset from Australia's Bureau of Meteorology. The dataset is comprised of 142,193 daily weather observations at 49 different locations all over Australia, gathered from numerous Australian weather stations over a ten-year period between 2007 and 2017. 24 variables are included in this dataset, several of which are categorical variables with Location, WindGustDir, WindDir9am, and WindDir3pm being high cardinality.

Appendix is a brief explanation on the variables.

Initial Analysis

A quick analysis was performed to get some initial understanding about the weather in Australia.

Figure 1 gives an overview on the mean values of continuous variables concerning two groups of rain and no rain on the following day. Some highlights are:

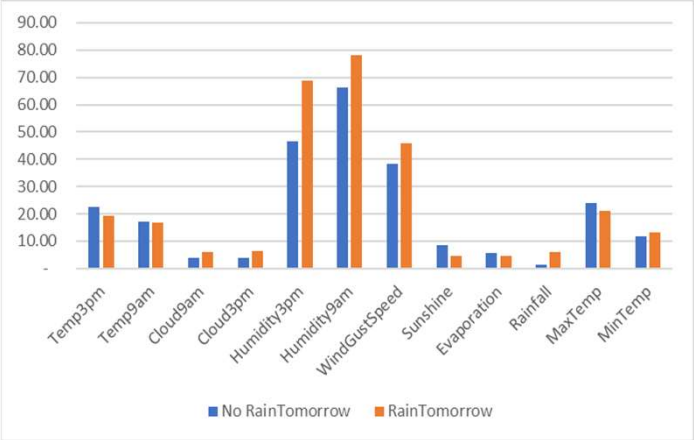

Figure 1. Continuous variables vs RainTomorrow

- Temperature seems not to be an indicator of rain. There is no clear difference between temperature before days that have rain and those that do not.

- Cloud and sunshine are among potential indicators of rain. It is more likely to rain on the next day if there is more cloud and less sunshine today.

- Before days when it rains, average humidity level and wind speed level are a lot higher than days before which it does not rain.

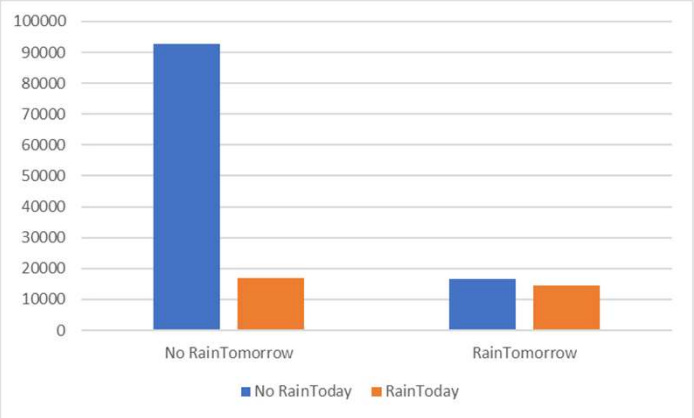

Figure 2. RainToday vs RainTomorrow

The chart above shows the likelihood that rain occurs on consecutive days. As can be inferred from the chart, it is not likely that it will rain tomorrow if it rains today.

Figure 3 summarizes several high correlation values in the dataset. Based on the correlation matrix among numeric variables in the dataset, several variables are highly positively correlated. High correlation values are highlighted in yellow.

	MinTemp	MaxTemp	Pressure9am	Temp9am
Pressure3pm	-0.502	-0.458	0.962	-0.507
Temp9am	0.907	0.893	-0.453	1.000
Temp3pm	0.727	0.985	0.319	0.871

Figure 3. High correlation values

The two figures 4 & 5 describe changes in the minimum and maximum temperature and average rainfall recorded over years and over months.

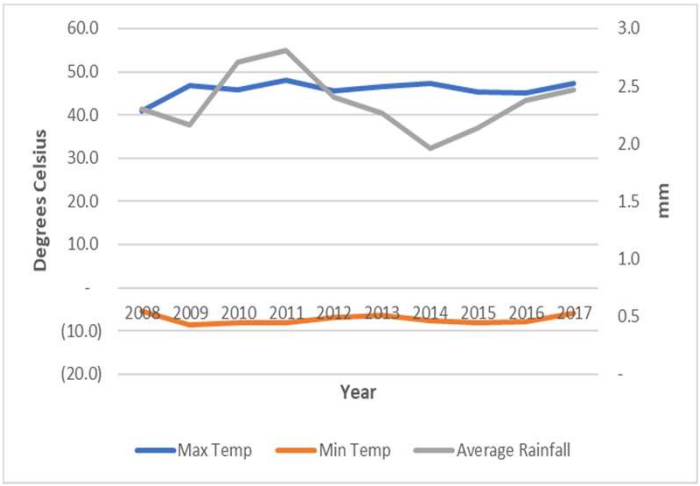

Figure 4. Fluctuation over years

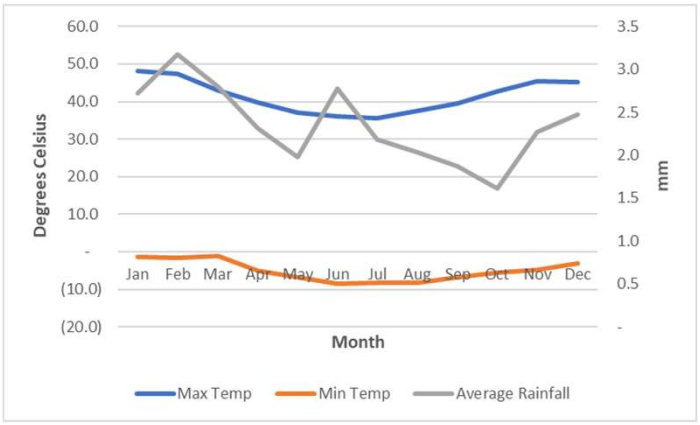

Figure 5. Fluctuation over months

- Overall, Australia has a wide daily temperature variation, about 40 degrees Celsius in all the seasons throughout the year. Temperature drops below zero even in the summer.

- There was much variation in minimum and maximum temperatures in the country over the

selected period, but there was no clear upwards nor downwards trend in the 10 years. Three years of 2011, 2014, and 2017 experienced record high temperatures at 48.1, 47.3, and 47.3 respectively.

- Australia received extremely low rainfall throughout the year over the surveyed 10-year span. Average daily rainfall is classified as light rain (0-5mm). Noticeably, summer in Australia is known as extremely dry, but summer rainfall was still recorded to be higher than other seasons.

CLUSTERING ANALYSIS

Clustering analysis was performed to find out locations with similar weather patterns.

Data

Sampling

Considering the purpose of clustering analysis, which is to find out locations with similar weather patterns based on mean values per year, I decided to use weather data in one year only instead of the whole 10-year dataset. The most recent data was selected, which is data between July 2016 and June 2017. The sample has 17,286 observations.

Irrelevant/Redundant features

As the first step, I checked to keep only meaningful variables.

- As per the initial analysis, some numeric variables are highly correlated; therefore, I removed three variables *Temp9am, Temp3pm*, and *Pressure9am* from the analysis.

- I performed Chi-squared tests on three categorical variables *WindGustDir*, *WindGust9am*, and *WindGust3pm* and found that the variables are not independent. Therefore, I removed two variables *WindGust9am*, and *WindGust3pm*.

Figure 6 summarizes Chi-squared test results.

Variable 1	Variable 2	p-value
WindGustDir	WindDir9am	2.20E-16
WindGustDir	WindDir3pm	2.20E-16
WindDir9am	WindDir3pm	2.20E-16

Figure 6. Chi-squared test results

- I removed several additional variables that are not expected to add much contribution to explaining the clustering result.

o *Date, RainToday, RainTomorrow*: The analysis focuses on mean values of each location in 1 selected year, so the three variables do not make much sense.

o *WindGustDir*: This variable tells the direction of the strongest wind during the day, so it does not help explain weather conditions at the place.

o *Risk_mm:* This tells the rain amount on the next day. This variable is actually providing the same informational meaning as Rainfall.

Missing values

Four variables below have quite high number of missing values. Thus, I decided to remove the four variables from the analysis.

Variable	Missing values	Percentage
Evaporation	10,742	62%
Sunshine	12,714	74%
Cloud9am	7,805	45%
Cloud3pm	9,167	53%

Figure 7. Variables with many missing values

After removing the irrelevant and incomplete variables, the dataset is left with 3,027 rows with missing values, accounting for 17.5%. Even though this is not a small percentage, in case the incomplete observations are removed, there is still a large number of observations for the analysis. Importantly, there is low possibility that important data will be omitted because the remaining observations still cover almost all areas of Australia in one year. Also, in an analysis using

unsupervised learning like clustering, I prefer using complete observations to imputing missing value. The main reason is that by imputing missing values, we are inputting our best guesses of information into the model, not the true values; however, there is no proper way to assess accuracy levels of the final analysis result. This means that there is a risk that the result may mislead.

Therefore, I decided to remove 3,207 incomplete observations from the dataset. As a result, seven cities are also omitted from the dataset.

The final dataset used for clustering analysis has 14,259 observations with nine variables, namely *MinTemp, MaxTemp, Rainfall, WindGustSpeed, WindSpeed9am, WindSpeed3pm, Humidity9am, Humidity3pm,* and *Pressure3pm.* The dataset was aggregated with mean values for each location and transformed into a dataset with 42 observations.

Outliers

I found two outliers in the dataset, Darwin with outliers in Rainfall and Pressure3pm, and Alice Springs in Humidity9am.

The treatment of outliers will be further discussed when clustering method is mentioned.

Method

In this clustering analysis, I chose to proceed with k-means clustering. In fact, the dataset is not a large one with 42 locations only; thus, both hierarchical and k-means clustering can perform well. However, considering that k-means is known to give particularly good performance with naturally occurring, clear groupings while the analysis is conducted on location-weather data, I believe k-means clustering is a better method in this case.

Because k-means clustering is sensitive to outliers, outlier removal is recommended. However, I wanted to see the impact of outliers in this case to clustering result; thus, I performed clustering on both original datasets and datasets with no outliers. The two results were the same with non-outlier observations.

The analysis described below employed the full dataset with outliers.

Scaling

The dataset contains numeric data on different scales. Temperatures vary between 6 and 32 degrees Celsius. Wind speed (km/h) falls into a similar range of values with temperatures. Rainfall ranges from 0 to 7mm. Atmospheric pressure values are over 1000. Also, the variances of the variables scatter a lot, ranging from 1.39 to 121.9. Thus, when clustering is performed on this dataset to find groups of locations with similar patterns, the result might be biased; when distances between any points are calculated, for example. Therefore, I standardized the data before proceeding with clustering analysis.

Optimal k

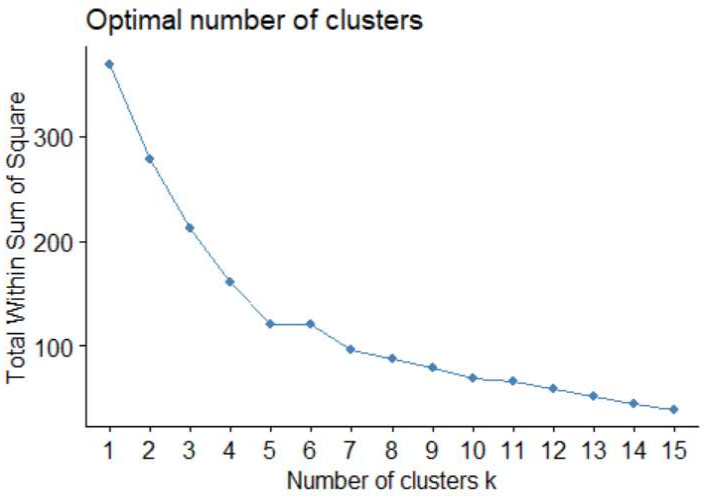

Figure 8. Plot of total within sum of squares

In order to find the optimal k for clustering, I plotted total within sum of squares against k

values of up to 15 for a better overview.

It can easily be seen from the graph that the more clusters we choose to go with, the better we can categorize the locations; however, using too many clusters is not quite a good idea because it would be complicated for interpretation. I followed the Elbow method and chose 4 clusters to go with.

Result

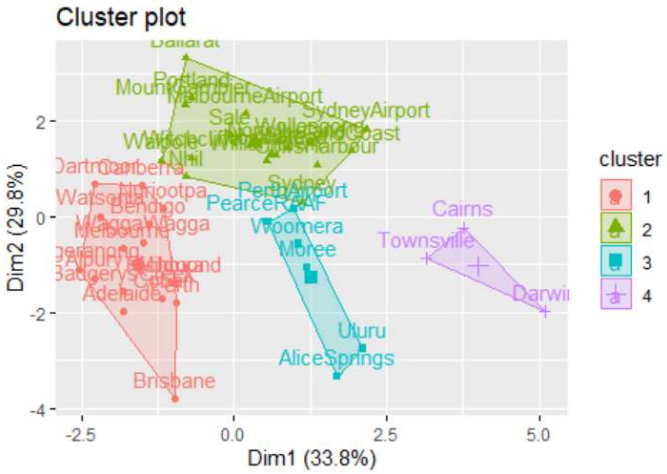

Figure 9. Clustering result

I visualized the clusters on Australia's map for a better understanding.

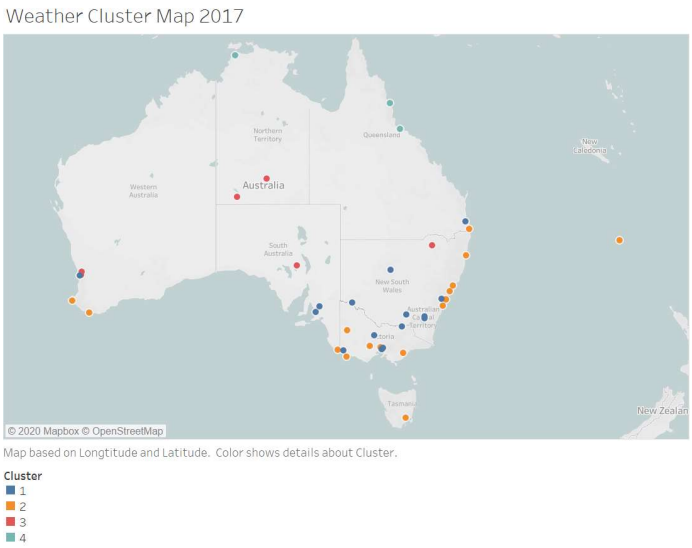

Figure 10. Weather cluster map 2017

Based on the above clustering analysis, I labelled the four clusters as four groups of towns and cities with associated features as below:

- **Cluster 1:** *inland cities, but not too far from sea.* Cities in this group have moderate temperatures and large daily temperature variations. Rainfall amount is in the lower level of the range between 0 and 5mm per day. The range from 0 to 5mm per day is generally regarded as light rain. According to Beaufort Wind Force Scale, daily strongest wind speed in the cities fall in the range between 29 and 38 km/h, which is classified as fresh breeze. Humidity levels in the cities are moderate in the morning but fall low in the afternoon. Atmospheric pressure is high.

- **Cluster 2:** *coastal & island cities.* This group includes cities with moderate temperature with low daily temperature variation. Temperature of the cities in cluster 2 is generally lower than that of cluster 1 cities. Rainfall level is in the middle of the light rain range. These cities experience daily strong wind over 40km/h per day. Humidity level is high in the morning and moderate in the afternoon.

- **Cluster 3:** *cities in the middle of Australia*. Cities in this group have high temperature with large daily temperature variation. This group has the lowest rainfall in those areas. Extremely low rainfall amount is recorded. The cities experience strong wind and low humidity.

- **Cluster 4:** *northernmost cities.* This group includes only 3 cities, namely Cairns, Darwin, and Townsville. These cities have highest temperatures among the surveyed locations. However, daily temperature variation is low. The cities also have light rain per day, but the amount is in the upper level of the range. Wind speed levels in these cities are at the intersection of fresh and strong breeze, around 38-39km/h. Moderate humidity (59-70%) and low atmospheric pressure are recorded in the cities.

Overall, weather conditions are much defined by geographical locations. The clusters could well exhibit this impact. However, there are still some exceptions. For example, even though Perth Airport and Pearce RAAF are close to the sea, the two places are placed in cluster 3 with cities in the middle of Australia. Interestingly, Perth and Perth Airport are close to each other, but weather conditions in two places bear some differences. In detail, rainfall amount in Perth and Perth Airport is lower and wind speed levels are recorded higher than in Perth and locations nearby.

Further Analysis

Australia is known to have changed significantly in response to a warming global climate. Therefore, I did a quick clustering analysis in the year of 2009 to see if there is any change over almost 10 years in Australia's weather patterns.

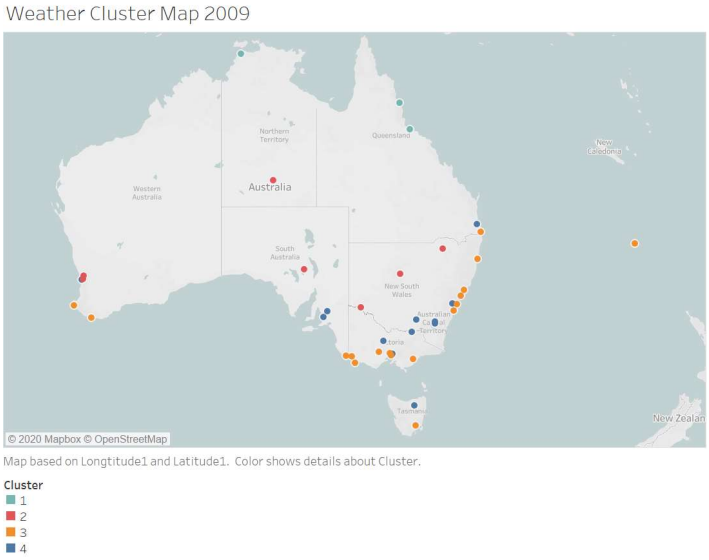

Figure 11. Weather cluster map 2009

Basically, the clustering analysis gave similar results, except re-classification of several cities after 10 years. Particularly, in 2009, Cobar and Mildura were classified in the same weather group with cities in the middle of Australia. These two are inland cities and far from the sea, but not really in the middle of Australia. Meanwhile, Dartmoor and Melbourne are near the sea and had weather of coastal cities. However, almost 10 years later, the four cities are re-classified to the group of inland cities that are not far from sea. The difference mainly lies in increasing wind speed in the cities.

Limitations

I would love to spend more time to develop this analysis which is still relatively simple. Some limitations are as follows:

- The data used for analysis is from 2016-2017. Australia continues to change significantly in response to global climate change. Thus, the weather data in the selected year may not be representative for Australia now.

- A number of observations with missing values were removed. The removal may lead to omittance of important information and thus, the result may be misleading.

PREDICTIVE MODELLING

A rain predictive model was built to answer the question "Is it going to rain tomorrow?".

Data

Sampling

To build a model to predict rain, I used the sample of observations between Jul 2016 and Jun 2017, considering that the sample size is large enough.

Irrelevant features

I removed *Date, Location*, and *Risk_mm* from the dataset because they do not help explain rain possibility. It may be tempting to think that it is more likely to rain at some points of time throughout the year and at some locations; however, rain is actually forecasted from weather

conditions, not Date & Location. Also, Risk_mm tells the rain amount on the following day. Thus, it needs to be removed from the prediction because it has revealed the rain outcome already.

Missing values

Similar with clustering analysis, I removed four variables *Evaporation, Sunshine, Cloud9am*, and *Cloud3pm* that had too many missing values. I also removed all rows with missing values. The final dataset has 13,570 observations with 17 variables.

Model Development

With the aim of choosing a model with highest accuracy, I tried to build model using different techniques. I tried three techniques, including k Nearest Neighbors, Support Vector Machine with radial kernel, and Artificial Neural Networks.

I did not use Naïve Bayes and Decision Trees because weather variables in this dataset are somewhat correlated. For example, today's rain leads to a higher humidity level, or an increase in temperature will decrease the relative humidity. However, Naïve Bayes assumes that all variables are equally important and independent, and Decision Trees do not take into account interactions between attributes. Regarding Support Vector Machine, the model with radial kernel is known as the one that usually produces the best prediction and can reduce overfitting; thus, I chose to proceed with this method only instead of trying all four kernel functions.

Out of the three methods, SVM produces the highest accuracy of 80% without being tuned, while models using kNN and ANN accurately predicts around 70% even when hyperparameters are tuned. Therefore, I proceeded with SVM.

SVM model development is described with further details as follows:

Data Cleaning

SVM can handle irrelevant and redundant variables. Missing values need to be corrected. In this case, missing values were removed. Categorical variables were converted to dummy variables. Rescaling was done while the model was being created.

For this model, I chose to split the dataset into training data and testing data by 80%-20%. I chose 80% for training set because I believe the dataset with 13,570 observations is big enough and 80% of the dataset is good enough to train the model.

Imbalanced Data

This dataset exhibits issues with class imbalance. The number of days with rain on the following day is only about one third of those without rain. Therefore, I improved the issue using SMOTE.

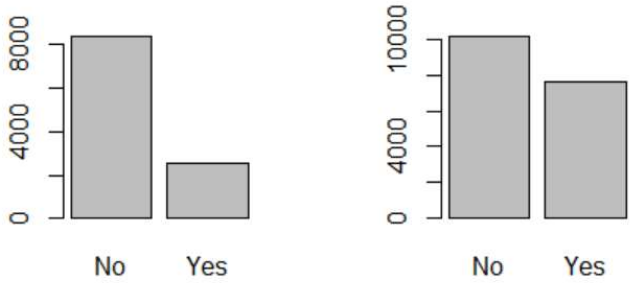

Figure 12. Class distribution of training data before and after resampling

Result

By running an SVM with radial kernel model, I got Model 1 with performance on training data and testing data being summarized as below:

	Train	Test
Accuracy	0.9085	0.8164
Kappa	0.8128	0.5226
F-measure	0.9206	0.8762

Figure 13. Performance by Model 1

Overall, the model performs better on training data set than testing set, which implies that the model has an overfitting issue. Accuracy level is 90.85% when the models is applied on the training data set, but only 81.64% when it comes to the testing set. The model performs moderately. Kappa statistic is 0.8164 for training data, which is interpreted as a very good value. However, the corresponding figure for testing data is a lot lower - just 0.5226, interpreted as a moderate value. Possibility of a correct prediction by chance is not good. F-measure represents a goodness of fit assessment for the classification model. The values for training data & testing data are quite good, being 0.9206 and 0.8762 respectively. In general, this is not a good model, and overfitting issue should be fixed.

I tried to tune gamma and cost values of SVM model to improve the performance; however, R kept showing processing status and did not return any result. Therefore, I tried other methods to fix overfitting issue of the model.

Feature Selection

I used recursive feature elimination to perform feature selection and reduced the number of independent variables from 58 to 16.

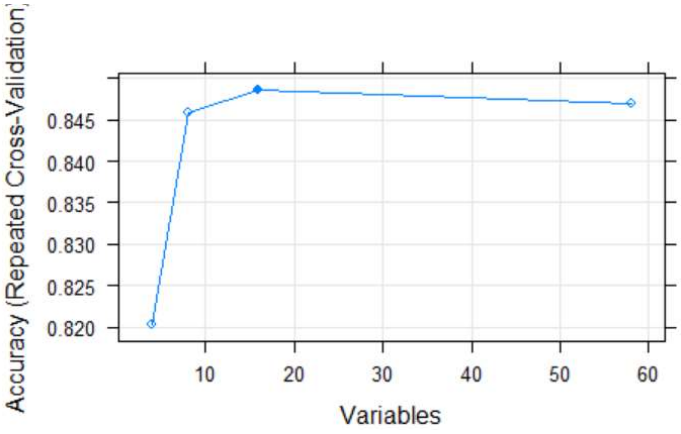

Figure 14. Accuracy against number of variables

The optimal variables include *Humidity9am, Humidity3pm, Pressure9am, Pressure3pm, WindGustSpeed, WindSpeed9am, WindSpeed3pm, Temp9am, Temp3pm, Rainfall, MinTemp, MaxTemp, RainToday_Yes, WindDir9am_NNE, WindGustDir_N, and WindDir3pm_N*. The result also indicates five most important variables as *Humidity3pm, Pressure3pm, WindGustSpeed, Humidity9am, and Temp3pm.*

By using feature selection, overfitting issue was eliminated, and I got Model 2 with performance being improved slightly and no overfitting issue.

	Train	Test
Accuracy	0.8379	0.8319
Kappa	0.6643	0.5607
F-measure	0.864	0.887

Figure 15. Performance by Model 2

This model is balanced with similar performance on both training and testing data. Both training and testing data predict with accuracy of over 83%, which is a relatively good level. Kappa statistic on training and testing data implies moderate performance. High F-measure on both training and testing data at over 0.86 represents a goodness of fit assessment for a classification model. Overall, this model is better than the first one with moderate performance.

External Evaluation

To have a better evaluation of this model, I made a quick comparison with other weather prediction models. According to an article by NOAA SciJinks, a website by the NASA Space Place team, a seven-day forecast can produce prediction with accuracy level of about 80% and five-day forecast 90%. The shorter the forecast is, the more accurate the prediction result is. Accordingly, a one-day forecast should accurately predict rain over 90% of the time. This means that the model built from this dataset is not really a good one.

Limitations

This model still has some limitations as below:

- The removed variables, *evaporation, sunshine,* and *cloud* are important factors in predicting rain. Overall, rain cycle can be described briefly as follows. Water evaporates from the earth surface, rises into the atmosphere, and becomes part of a cloud, and then rain falls. Thus, evaporation, sunshine, and cloud play an important part in explaining the possibility that rain occurs. However, these important features are omitted from the model due to missing value issue, reducing the model's accuracy.

- Due to a lack of computational capabilities and infrastructure resource, I failed to tune hyperparameters for SVM model to get the best values of gamma and cost to improve model accuracy.

CONCLUSION

Based on the analysis, below is the summary of my findings from answering the two research questions:

- Weather conditions at cities across Australia can be divided into four different groups, corresponding with their geographical locations. Australia's weather continues to change in response to climate warming, leading to a little change in weather clustering after nearly ten years.

- According to Feature Selection result for Model 2, wind direction does not show clearly significant impact in forecasting rain on the following day, except the case when prevailing winds are north or north-north east. Meanwhile, weather conditions after 3pm (Humidity, Temperature, and Pressure) and wind speed are significant in determining rain possibility on the following day.

- Evaporation, the number of bright sunshine hours, and fraction of sky obscured by cloud seem to be among important indicators of rain possibility on the following day. The exclusion of these variables from rain predictive model can possibly be a reason for reduced accuracy.

Besides, this analysis may have several implications as follows:

- In a such a dry country as Australia, the ability of accurately predicting rain over a long period can help the government take appropriate measures to minimize negative impacts in the dry season, widespread bushfires, for example.

- By having different locations with similar weather patterns grouped together, authorities in those locations can work together to make plans or can share good practices to deal with severe weather conditions.

- Concerning businesses that are located at places where weather forecasts are not published and cannot be easily extracted, for example, mining or farming locations in the most remote areas, they can collect relevant information and build a weather predictive model to get important information to serve their operations.

However, in general, accurate weather prediction is still a challenge. Particularly, weather conditions continue to change and are becoming more complicated; therefore, forecasting models need updating frequently to incorporate those changes and need to be built to deal with increasing complexity of the weather. In addition, weather predictive models need to be improved in such a way to increase accuracy levels of long-term forecasts.

APPENDIX – Variable Definitions

Variables	Definition
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
WindGustDir	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. This is measured in "oktas", which are a unit of eighths. It records how many eighths of the sky are obscured by cloud. A 0 measure indicates completely clear sky whilst an 8 indicates that it is completely overcast.
Cloud3pm	Fraction of sky obscured by cloud (in eighths) at 3pm
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
RISK_MM	The amount of next day rain in mm. Used to create response variable RainTomorrow. A kind of measure of the "risk".
RainTomorrow	The target variable. Did it rain tomorrow?

REFERENCES

1/ Bureau of Meteorology, CSIRO (2018). *State of the Climate 2018*. Retrieved from https://www.csiro.au/en/Showcase/state-of-the-climate

2/ *How Reliable Are Weather Forecasts?*. Retrieved from https://scijinks.gov/forecast-reliability/

3/ *JetStream Max: Beaufort Wind Force Scale*. Retrieved from https://www.weather.gov/jetstream/beaufort_max

4/ Joe Young (2017, Dec 04). *Rain in Australia.* Retrieved from https://www.kaggle.com/jsphyg/weather-dataset-rattle-package

(The primary data source is Commonwealth of Australia 2010, Bureau of Meteorology. However, I am not the one to collect and consolidate the data, so I refer to Kaggle link as the secondary data source.)