Statistical analysis and modelling of weather data in Melbourne

Zane Hoogendoorn, Sai Kumar Murali Krishnan, Cody Lewis, and Priyom Sarkar

Faculty of Science, Monash University

ADS2001: Data challenges 3

28 May, 2021

Table of contents

Executive summary

Weather forecasting is vital to many aspects of modern society, including tourism, agriculture, mining, airport administration, power generation, and general public knowledge, among other things. The process has undergone many changes over time; current forecasting uses recorded weather data from specialised instruments and advanced modelling techniques performed by computers. In Australia, the Bureau of Meteorology (BOM) is the Government agency that is responsible for providing this public service.

The data used in this project was sourced from the BOM instruments in Melbourne, Australia. The data was 'scraped' from the live feed on the BOM website. Many issues were found in the data, which negatively impacted the data quality and could thereby have affected the models. Various techniques were used in Python to process the data before building any models to ensure the data was logically correct and complete without gaps.

Before developing any models, the data was explored thoroughly using various analytical techniques. It was found that the trend of temperature and rainfall was decreasing since 2011, despite the popular opinion in Melbourne that both of these have been increasing. Additionally, the correlations between the weather features were found using a correlation matrix which revealed that "Dew Point Temperature" and "Dry Bulb Temperature" along with "Relative Humidity" and "Mean Sea Level Pressure" were very highly positively correlated while both "Relative Humidity" and "Mean Sea Level Pressure" were largely negatively correlated to "Dry Bulb Temperature".

After the processing, various irregularities were found to be present in the data during the exploratory analysis. These irregularities were not indicative of an error that was missed during processing; instead, they were anomalous because they did not follow the expected pattern. These deviations from the norm were theorised to be a result of external climatological phenomena. Simply put, the weather in a particular region is not a closed system; many external factors can impact the weather of that region in an asynchronous pattern, hence causing anomalous data.

In particular, the El Niño–Southern Oscillation (ENSO) and Indian Ocean Dipole (IOD)are two such climatological phenomena that influence the weather experienced in Melbourne. For the ENSO, when the ENSO index is above 0, Melbourne was expected to have an increased likelihood of warmer weather and reduced rainfall. In conjunction, when the index fell below 0, it was expected that Melbourne experienced a month cooler and wetter than what was historically expected. Statistical analysis found that the processed data supported this; however, rainfall was influenced much more than temperature.

During a positive phase of IOD, it was expected that Melbourne would experience higher temperatures and less rainfall due to the western Indian Ocean waters having a higher sea surface temperature and the eastern waters having a lower sea surface temperature. The opposite was expected for the negative phase. These suppositions were supported by the data for rainfall and temperature during the positive phase of IOD, although only rainfall appeared to be influenced during the negative phase.

The broad scope of the project allowed for freedom in selecting which weather conditions to forecast and with what models. The features which were decided to be modelled were temperature and rainfall. These features were chosen because they are the most popular weather conditions in weather forecasts and the most useful for most people.

Modelling rainfall was accomplished using Random Forest Classification. The decision to use a classification model rather than a regression model was based on how, generally speaking, knowing the exact amount of rainfall is less valuable than knowing the intensity of rainfall which is generally more applicable to more scenarios. The optimal RFC model was produced on a specially balanced dataset and achieved an accuracy of 87.70% ± 0.07%, a precision of 87.64% ± 0.07%, and a recall of 87.70% ± 0.07%.

Contrariwise, modelling temperature was conducted using time series forecasting. An attempt was made to model the temperature on the assumption that Melbourne's weather could be modelled with a sinusoid, which meant ignoring long term climate effects—employing an ARIMA model, which meant looking at historical data then taking a moving average then performing regression which yielded promising results, with a mean absolute error of fewer than 2 degrees among other similar metrics.

Introduction

Weather forecasting provides vital information that impacts society in various ways, including tourism, agriculture, mining, airport administration, and power generation, not to mention for the general population. The process of forecasting weather has changed over time. In the modern world, sophisticated tools have been developed that greatly help the collection of data as well as forecasting itself. The Bureau of Meteorology (BOM) is the Australian Government agency that is responsible for providing this public service in Australia and its surrounding regions. BOM uses various recording instruments spread all across the country to measure a range of climatological features and makes some of this information available to the public via the internet.

This project utilises data 'scraped' directly from the live feed of recordings from the BOM's weather instruments located inside Melbourne. The data collected included the: dry-bulb temperature (°C), dew-point temperature (°C), apparent temperature (°C), relative humidity (%), wind direction, average wind speed over the past 10 minutes (km/h), maximum wind gust over the past 10 minutes (km/h), mean sea-level pressure (hPa), and amount of rainfall since 9 AM (mm).

Data quality

Various artefacts were introduced into the dataset during the ‘scrape’ from the internet. A first scan reveals placeholder numerical and string values, duplicate entries for the same timestamp, and the format given as a mix of space and tab-separated. In addition, anomalous data entries were present, such as -9999 entries in wind-related columns, which were fixed by taking the average of surrounding values. A similar process was applied to the pressure column, where it was found that the column entries for pressure and wind direction had been swapped for a short time. A secondary check was done to ensure each column type was the appropriate type and within the expected range, thereby passing the preliminary preprocessing stage.

A more detailed inspection of the partially cleaned dataset reveals oddities, such as a non-zero mean wind speed but zero maximum wind speed, a mean wind speed of 0 but a non-calm wind direction and vice versa. If the wind direction disagreed with the wind gust and mean wind speed (where both are 0), then the value was set to ‘calm’. Otherwise, the wind gust was determined to typically be 40% more than the mean wind speed, which was in turn used to replace missing wind gust/mean wind speed values. Following this, a regular DateTime column was created and used to resample the data into half-hour intervals. The original dataset had no consistent frequency and varied over time; it was unusable for time series forecasting. Finally, any missing data revealed by resampling was interpolated, yielding the processed data frame.

Expected relationships

Given the nature of the data, it was expected that it would be very seasonal and would, in a broad sense, repeat itself from year to year. This was expected due to typical weather fluctuations and how weather usually repeats itself every year more-or-less the same.

Furthermore, weather features are typically not isolated from other weather features, given that they are all affected by the same system of forces and have overlapping areas of impact. It was therefore expected that some of the features of the data would be correlated. For example, it was expected that the dry-bulb temperature, dew-point temperature, and apparent temperature would be highly correlated simply because

of their units of measurement and respective definitions. Similarly, the relative humidity was expected to be correlated with the amount of rainfall and dew-point temperature and the average wind speed with the wind gust.

Modelling challenges and applied solutions

The project prompt was broad, which allowed for freedom in choosing which features to model. Temperature and rainfall were decided to be the project's focus, given they are two weather features that are the most popular for weather forecasts and typically the most impactful for general society. These two features were the 'target' variables for the models, and all other weather features were the 'feature' variables.

Decent results were expected, given that the weather typically acts in repetitive and systematic ways. For example, when a day is cold, some other weather features may reflect this with atypical values also if correlated to temperature. However, many inconsistencies in the data made optimisation and achieving a near-perfect model very challenging. These anomalies were evidence of external factors altering the observable weather in Melbourne.

The temperature was forecasted over a range of years using an additive ARIMA (AutoRegressive Integrated Moving Average) model, which is the sum of non-linear components that model seasonality. The forecasting was done with the FBProphet library, which uses machine learning techniques. Due to the inherent randomness of temperature prediction, uncertainty accumulates with more extended forecasts, meaning that forecasts were kept to a year at most. ARIMA was applied to rainfall and other columns, which was much more inaccurate in comparison.

Random Forest Classification was used for forecasting the rainfall weather feature. It was decided to use classification on the rainfall data since people generally care more about rainfall intensity on a given day and not so much about the exact amount. Using Random Forests was inspired by the size of the dataset, the desire to run many iterations of training and testing, the variables were varying in units and dimension, and it also makes it simple to order variables based on importance.

Statistical analysis

In order to gain a better understanding of the trends in Melbourne's climate, a surface level inspection into the data from the Bureau of Meteorology was conducted. To better understand the way that variables related to each other, a correlation matrix containing each variable from the dataset was created, as seen in Figure 1. It was clear that "Dew Point Temperature" and "Dry Bulb Temperature" along with "Relative Humidity" and "Mean Sea Level Pressure" were very highly positively correlated, as expected. Conversely, "Relative Humidity" and "Mean Sea Level Pressure" were largely negatively correlated to "Dry Bulb Temperature". These relationships were logical given the contrasting nature of air temperature and humidity.

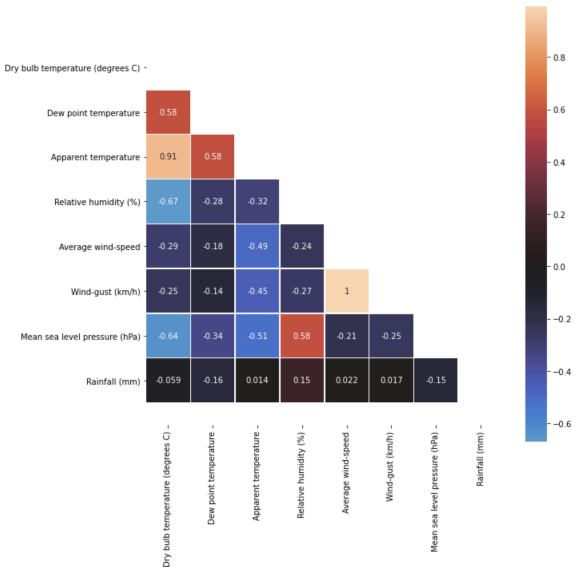


Figure 1. Correlation matrix of the weather features in the dataset.

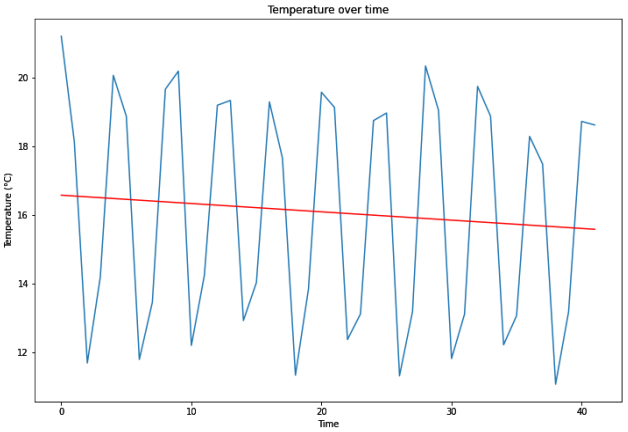


Figure 2. Trend of temperature in Melbourne since 2011.

A topic which has caused great debate and discussion in recent years is how the average temperature and average rainfall has been increasing over time. However, after looking into the provided dataset containing Melbourne climate data over the past decade, it can be concluded that average temperature and rainfall have both been declining. Seen through the negative slant of the line of best fit on the temperature and rainfall graphs in Figure 2 and Figure 3, the average temperature and rainfall appears to have fallen by approximately 1 degree and 0.2mm respectively. Given the data supporting global warming, these investigations can be deemed as out of the ordinary. These obscurities may indicate a lack of quality within the dataset which could increase the difficulty in modelling this data, and could negatively impact the overall accuracy of the models.

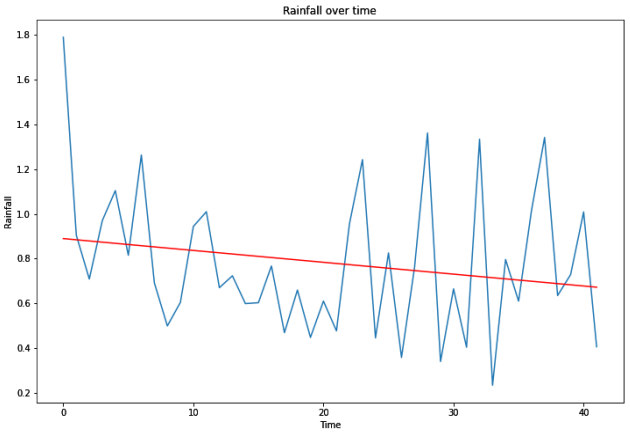


Figure 3. Trend of rainfall in Melbourne since 2011.

Comparison to Bureau of Meteorology data

Before delving into the comparisons made with the data available on the Bureau of Meteorology, an important aspect that needs to be discussed is the conversion of our data's observations into daily and monthly statistics. We did this by simply resampling the data into daily and monthly intervals, finding the mean, minimum and maximum for each interval. These specific statistics were chosen since the Bureau of Meteorology mainly contained these types of statistics, making it easier to compare. Despite being a relatively straightforward process, some complications were faced when converting the data into daily and monthly statistics. Specifically, when converting rainfall into a statistical value, there was an issue due to its observation being recorded at 9 am and issues when resampling for the statistical values. Hence, we decided to use the observations at 8:30 to replace the data containing the maximum daily statistics to remedy this. It would allow for easier comparisons with the Bureau of Meteorology's monthly statistics. In addition, another minor issue that occurred was that there was no statistical data available on the Bureau of Meteorology's website from the Melbourne Olympic Park (where our data's observations primarily came from). Thus, when comparing the BOM data we used, statistics came from the nearest station, Essendon Airport.

After completing this preliminary process, comparisons between our statistics and the data on BOM could be made. The comparison made mainly focused on how Melbourne's weather over the past decade fared against the climate statistics available on the BOM website in the past seventy years, seeing which periods experienced cooler or hotter temperatures—also identifying which periods experienced more or less rainfall.

We observed any anomalies with temperature and rainfall to make comparisons between the mean maximum temperatures and the mean rainfall.

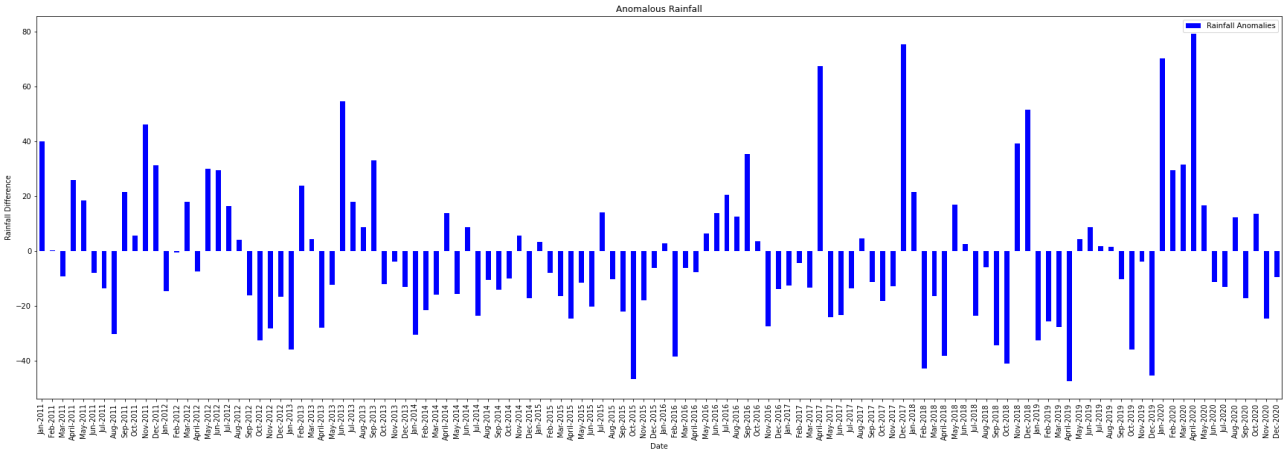


Figure 4. Anomalous Rainfall Data

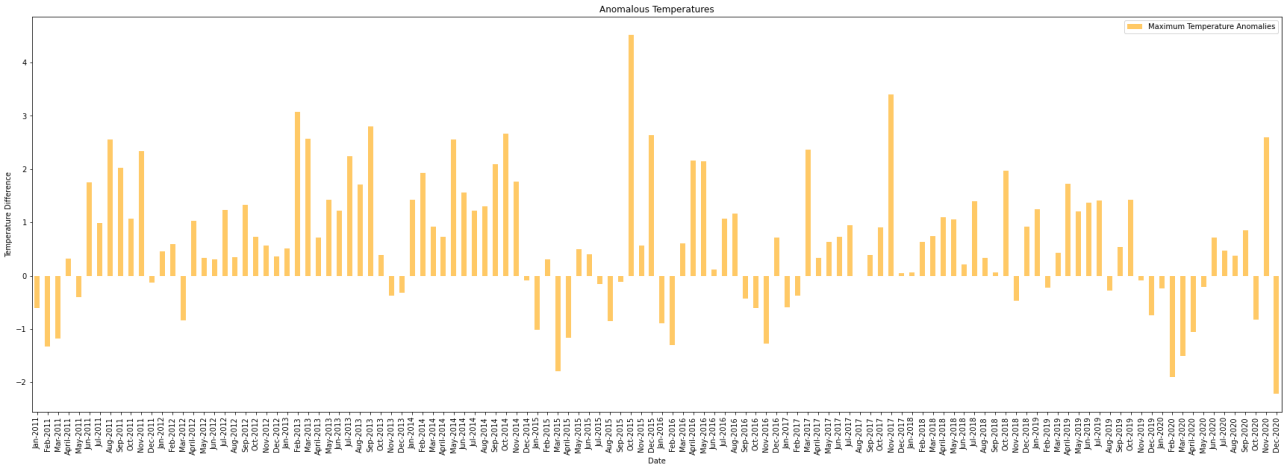


Figure 5. Anomalous Temperature Data

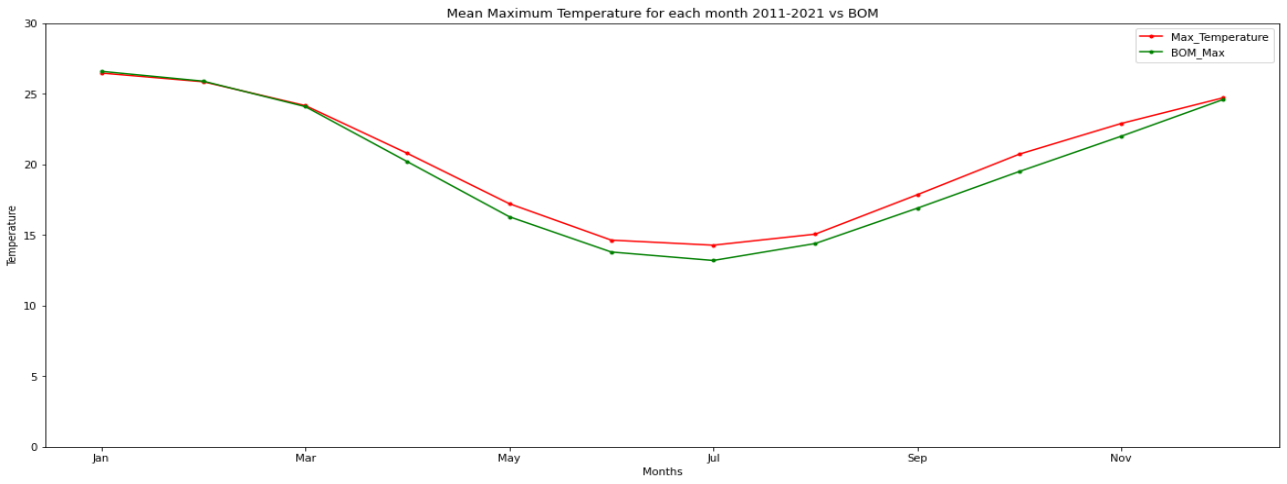


Figure 6. Comparison of Mean Maximum Temperatures

So from Figure 4, which represents the anomalous rainfall between our statistics and BOM's statistics, we can see that within the decade that there were more instances of less rainfall occurring than the average value (66 periods out of 120). In addition, we can also see from Figure 5 that there seem to be more instances of Melbourne experiencing hotter weather over the past decade. Hence, this may lead to the belief that Melbourne has been experiencing much hotter and drier weather over the past decade. Which in turn, may suggest that there is a trend that within the next couple of decades, Melbourne may be experiencing hotter and drier weather. To further demonstrate, we can see in Figure 6 that indeed Melbourne's weather within this decade is higher than the BOM statistics. In essence, what is being observed is that when comparing our data to BOM, there seems to be a trend suggesting that Melbourne may be experiencing hotter and drier weather over the next decade. In addition, there is also a possibility that the anomalies that have occurred may have been attributed to some of the climatological phenomena that occur. Henceforth, we investigated further into the effects that specific climatological phenomenon may have on Melbourne's Weather.

Influence of climatological phenomena

El Nino-Southern Oscillation

While the provided variables such as 'dew-point temperature', 'relative humidity' and 'wind direction' are all critical for forecasting Melbourne's weather, various climatological phenomena disturb the typical weather patterns across Australia and make predicting features much more difficult. The 'El Niño–Southern Oscillation (ENSO) is the natural oscillation of winds and sea surface temperatures over the tropical eastern Pacific Ocean. The period where the sea surface temperature of the Pacific Ocean is above average is referred to as the 'El Niño' period, and the period in which the temperature falls below average is referred to as the 'La Niña' period.

The 'ENSO index' is the primary indicator for monitoring El Niño and La Niña. The index can be plotted on a scale ranging from approximately -3 to 3 with the high values on the Oceanic Niño Index (ONI˚) scale corresponding to an intense El Niño period and negative values corresponding to a substantial La Niña period. According to the BOM, when the oceans are in a state of El Niño, Australia is likely to experience reduced rainfall, warmer temperatures and increased fire danger (Australian Bureau of Meteorology, 2014). This research aimed to quantifiably measure the impact that ENSO has on Melbourne's temperature.

ENSO values are concerned with the difference between the oceans current temperature and average temperature over a significant period. For this reason, the anomaly of Melbourne's daily air temperature was calculated by subtracting each month's average maximum temperature since 2011 by the monthly average temperature for the previous seventy years (Australian Bureau of Meteorology, 2015). The exact process was followed to find anomalies in the rainfall data. This revealed which months were hotter or wetter than usual and could then be compared against the ENSO data to see if a correlation exists between Melbourne's weather and the state of the El Niño–Southern Oscillation.

After declimitizing the values, the Melbourne air temperature data was reduced to a range of approximately -2 to 4; the range of the ENSO data was between -3 and 3. Figure 7 depicts these two datasets plotted against each other. As the variance of monthly rainfall is much larger than that of temperature, the range of the declimatized rainfall data was between -60 and 40. For the rainfall data to be plotted against a parallel axis with the ENSO data, the rainfall plot was inverted over the y-axis.

Figure 7 shows the relationship between the ENSO index and Melbourne weather anomaly since 2011. As Melbourne's temperature is expected to increase during the stages of El Niño, the trends seen in the blue and orange lines should follow a similar trajectory. From initial observations, it can be inferred that a positive correlation exists between the two variables. This is particularly evident in 2015 and 2016, where the trends share very similar peaks and troughs. Additionally, the negative effect that the La Niña period has on the temperature is prominent from the years 2019 to 2021 as both trends simultaneously descend below an ENSO value of 0.

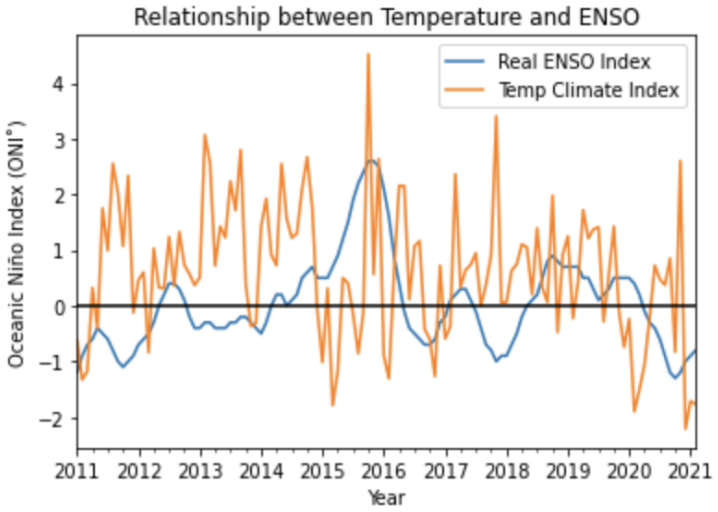


Figure 7. Comparison of anomalous temperature data and ENSO.

Despite signs of correlation, multiple sectors within the graph oppose the expected relationship. In particular, there is a large discrepancy between values from 2011 to 2015. While the ENSO Index suggests that the temperature should be below average in 2011, Melbourne's climate was warmer than usual. This anomaly recurred again in the summer of 2018 and the winter months of 2013 and 2020.

In order to see a more consistent correlation between the variables, the 'Temp Climate Index' could have been converted into a three-month moving average instead of being plotted with each monthly value. This simplification would have reduced the magnitude and variance between the different data points and resulted in a more continuous graph.

Many inferences can be made from observing the Figure 7, but a singular quantitative value representing the relationship between these two datasets was found using a correlation matrix. This matrix found that the correlation between the ENSO data and declimatized temperature data was 3.9%. The positive nature of this value indicates a relationship between the state of the ENSO and Melbourne's temperature. However, if some of the variations were taken from the 'Temp Climate Index' by creating a moving average, a more significant correlation could be seen.

While the relationship between temperature and the state of the ENSO was minimal but present, the correlation between the ENSO and rainfall in Melbourne was much more prevalent. Figure 8 shows this relationship.

As an inverse relationship between a high ENSO index and rainfall is expected, the blue line graph representing Melbourne's rainfall was inverted. At a glance, it is evident that a strong correlation exists between the two variables. Unlike the temperature, monthly rainfall follows a similar trend to the ENSO index from 2011 to the present. The anomalies in the weather patterns are accounted for, such as in 2015, when there is less rainfall than usual and in 2017 and 2020 where there is more rainfall than what is historically expected.

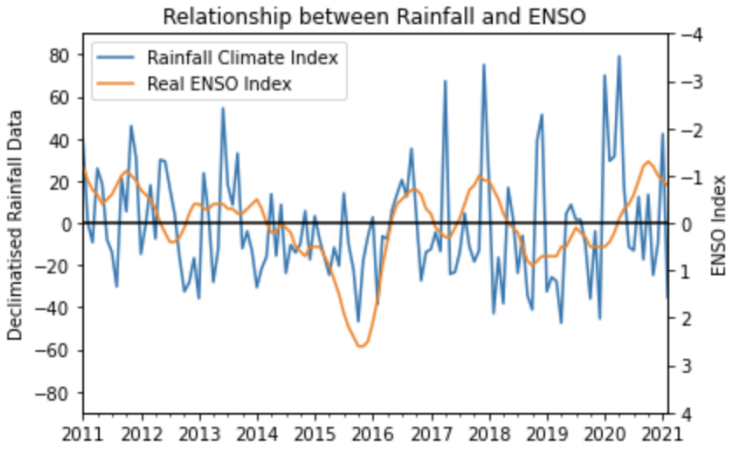


Figure 8. Comparison of anomalous rainfall data and ENSO.

Despite the plots in Figure 8 predominantly following the same pattern, there are some inconsistencies. Late 2019 and early 2020 are two of the most significant outliers in the graph. Each period shows increased and decreased amounts of rainfall given the El Niño–Southern Oscillation state. This inconsistency could occur for many reasons, such as conflicting natural phenomena. Nonetheless, it does show that there is not a perfect relationship between rainfall and the ENSO index.

Just as is the case for Figure 7, a moving average of monthly rainfall in Figure 8 would have been incredibly beneficial in finding a correlation. As the general trends of each graph are highly correlated, a reduction in the amount of variance in the 'Rainfall Climate Index' would have made a relationship even more evident.

Just as before, the singular quantitative value for the relationship between the data was found using a correlation matrix. The correlation was found to be -21.5%, indicating a higher correlation of ENSO to rainfall than to temperature in Melbourne, as expected from the visual comparison of Figure 7 to Figure 8.

Australia is likely to experience increased rainfall and cooler temperatures when in a state of El Nino. As a result, the correlation between the ENSO index and rainfall is expected to be negative as there is an inverse relationship between a high ENSO index and high amounts of rainfall. This relationship was found for rainfall but not so much for temperature, given that the rainfall correlation was 5.5 times larger than that of temperature. This enables the conclusion to be drawn that although only slightly, the El Niño–Southern Oscillation state directly affects Melbourne's temperature. However, the El Niño–Southern Oscillation state significantly affects the amount of rainfall experienced in Melbourne, which proves the initial hypothesis.

Indian Dipole

Before discussing how the Indian Ocean dipole impacts Melbourne's weather, it is essential to note that the data obtained relating to the Indian Ocean Dipole may differ from the data that is on the Bureau of

Meteorology. This is because the way the Bureau of Meteorology obtained their IOD indexes was using NOAA ERSST V5 gridded data. In contrast, the IOD data used in the project came from the Working Group on Surface Pressure is “calculated at NOAA/PSL using the HadISST1.1 SST dataset. SST has 1981-2010 climatology removed for each month, and the two regions are area-averaged and subtracted from each other.” (Working Group on Surface Pressure, 2020). The reason as to why this data was used over the data that is presented on the Bureau of Meteorology was due to how easily accessible the data was. For instance, the Bureau of Meteorology did not contain any time series data spanned over many decades; hence the IOD data from the Working Group on Surface Pressure was used to conduct the analysis.

The Indian Ocean Dipole, or IOD for short, is “defined by the difference in sea surface temperature between the western pole in the Arabian Sea (western Indian Ocean one) and an eastern pole in the eastern Indian Ocean south of Indonesia.” (Australian Bureau of Meteorology, 2021) It is one of the climatological phenomena which influences Australian weather depending on which phase it is in. These phases are described to be either positive, negative or neutral. During a positive phase, the Indian Ocean Dipole would result in the western waters having a higher sea surface temperature and the eastern waters having a lower sea surface temperature due to the western winds weakening along the equator. This generally results in our Australian weather to experience temperatures which would be higher than usual and that the rainfall to be significantly lower.

On the other hand, a negative phase occurs when the eastern sea surface temperature is warmer. In contrast, the western sea surface temperature would be cooler due to the western winds along the equator intensifying. Hence, the Australian weather would result in more rainfall during the winter and spring times. In addition, when IOD is in a neutral phase, the Australian weather would be as usual. So, there is a generalization that the Indian Ocean Dipole greatly influences Australian Weather but does this generalization of the Indian Ocean Dipole extend towards the weather that occurs in Melbourne, meaning that when the positive and negative phases occur would the related weather events occur?

To assess whether the IOD has any significant impact on Melbourne weather, we would require a standardized way of determining what constitutes either a positive phase or a negative phase. According to the Bureau of Meteorology, a way of determining a positive or negative phase is by observing the Dipole Mode Index (the indicator that measures the strength of the IOD) for observations that are above 0.4 degrees which is a positive IOD or below 0.4 degrees which is a negative IOD. These observations would then serve as a potential indicator for the potential weather Melbourne may be experiencing.

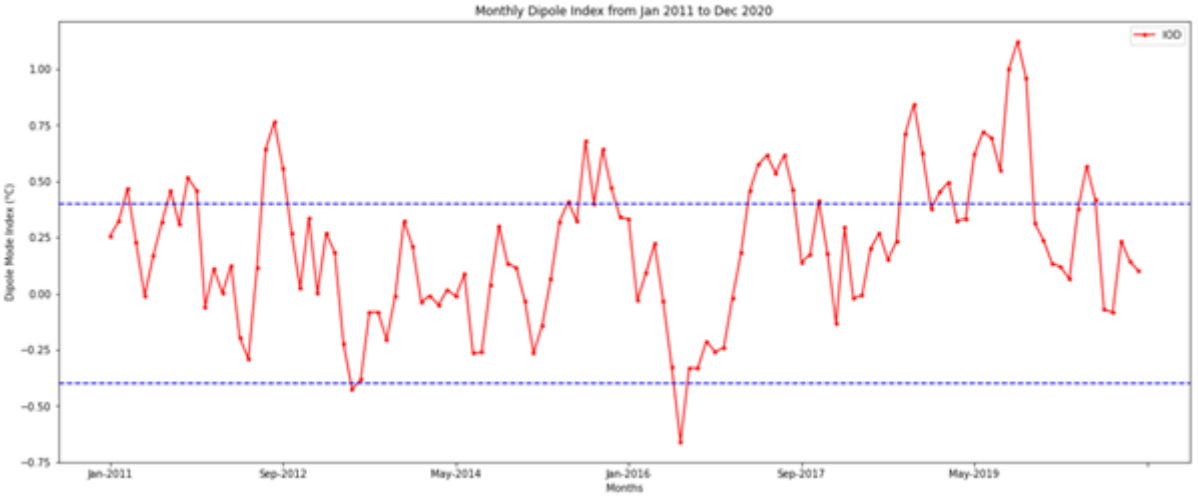


Figure 9. Graphical representation of the Indian Ocean Dipole Index from January 2011 to December 2020.

After identifying which periods would constitute as positive or negative, we would utilize the data extracted from the comparisons that were made to the BOM data (the rainfall and mean maximum temperature) and try to identify any patterns that may be present, seeing whether the IOD would impact the Melbourne weather.

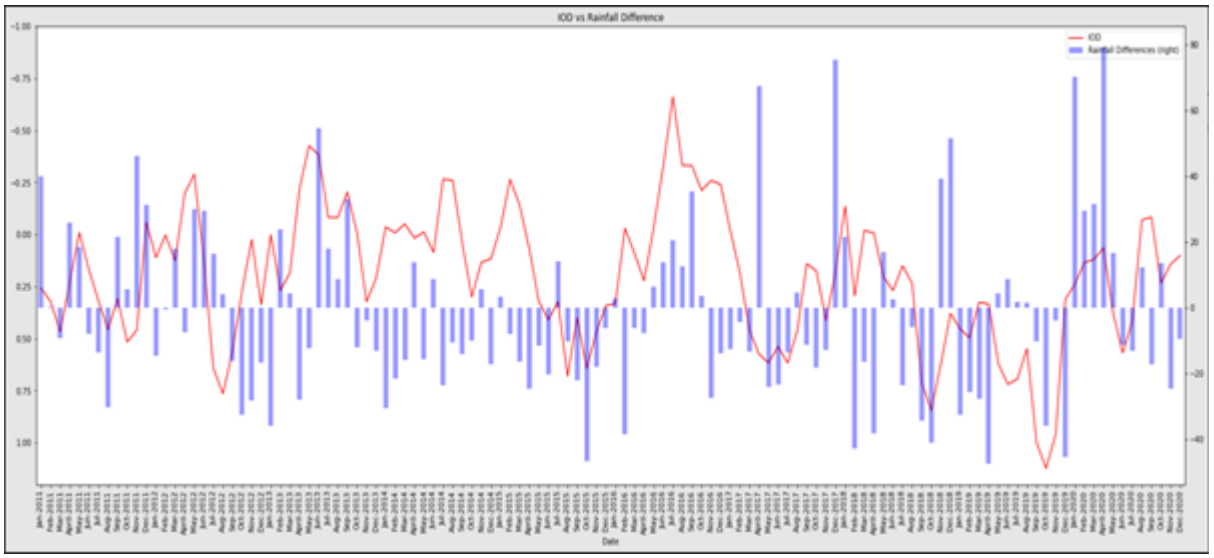


Figure 10. Comparison of anomalous rainfall data and IOD.

Figure 10 compares the pattern between the IOD and the anomaly rainfall. The line represents the IOD index, with its y-axis values being on the left side of the axis and being inverted. In contrast, the bars represent the anomaly rainfall with its y-axis values being on the right side. We can observe from this plot that the IOD can result in less rain for a specific period when in a positive phase. For instance, the period between September 2019 and December 2019, where the largest IOD value was observed, coincides with Melbourne having significantly less rainfall. In addition, there is also evidence of the rainfall increasing while the IOD is in a negative phase. For instance, during the 2016 winter-spring period, there was an increase in rainfall for that period and was undergoing a positive phase.

Conversely, many instances contradict the expected outcome of certain IOD indexes. For example, during the 2014 period, the IOD was mainly in a neutral phase nearing a negative phase. So, what would be

expected was that the rainfall during that period would be close to the average rainfall or even possibly increase. However, this was not the case, as it is evident with the rainfall being below average for that entire period. This may suggest that the Melbourne rainfall is not solely impacted by the Indian Ocean Dipole but also other potential factors such as climatological phenomenon. Overall, there are instances where the IOD would impact the rainfall which Melbourne experiences during a period. However, there may be more vital factors that would influence the rainfall for Melbourne.

Figure 11 compares the IOD index against the anomaly weather temperatures for Melbourne. Like the previous plot, the line represents the IOD index with its y-values on the left side of the axis. The bars represent the anomaly weather temperature, with its y-values being on the right side of the plot. We can see from this plot that there is some correlation between the IOD indexes and the anomalous temperature. This is evident as, throughout the decade, there are instances where the IOD was in a positive phase, and the temperature for that period was higher than the average. This may suggest that Melbourne's temperature for those periods tends to be hotter during a positive phase. However, unlike the positive phase, the negative phase does not necessarily result in cooler weather, as evident from the plot above. Overall, the temperature of specific periods can be impacted by a positive phase of the IOD, resulting in hotter weather. However, we can also state that it does not result in cooler weather when the IOD is in a negative phase.

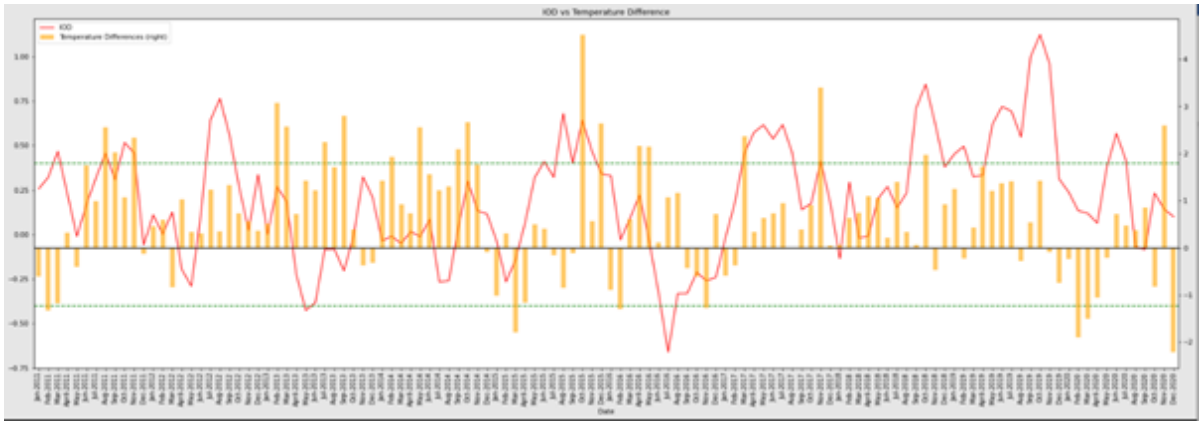


Figure 11. Comparison of anomalous temperature data and IOD.

In essence, the IOD is a climatological phenomenon that has shown evidence to impact Melbourne's Weather. During a positive phase, the IOD has resulted in Melbourne's weather becoming hotter than usual and producing less rainfall. On the other hand, the IOD during a negative phase is shown to have less of an impact on Melbourne's weather since there are still instances of having a negative phase but not having the weather that is to be expected. This may suggest that although this phenomenon may have a more profound impact on other areas of Australia, Melbourne's weather may not be as heavily impacted by the IOD and could be impacted by other factors.

Machine learning models

Machine learning techniques are the modern method used for forecasting weather. These models can be more accurate, faster, and more extensive than the traditional methods before machine learning. The use of machine learning is also encouraged by the greater volume of weather data being collected. Automation is also a key factor. The entire process of reading measurements, cleaning outlier data samples, forecasting the weather, and making the forecasts publicly available can be made automatic with only human interaction required for maintenance and oversight.

A possible and gratifying result of using machine learning is for the models to learn subtle patterns, which may have been invisible to the traditional forecasting method, that can signify an upcoming natural disaster such as lightning storms, tsunamis, etc. This would greatly benefit communities by allowing for advanced planning and evacuation which can not only reduce costs of damages, but also save lives.

Time Series Forecasting

Regression involving time series datasets is complex, primarily since our process (the temperature as a function of time) is non-deterministic in the sense that we may not have all the information to model the weather accurately, or the process itself is inherently random. A primary feature of our time series is that it contains a seasonal component, which corresponds directly to the year's seasons. The model used under the hood by FBProphet, a time series forecasting library, is additive. Seasonality is incorporated as a periodic function done with the Fourier series to allow for easier fitting. A fundamental aspect of time series regression is uncertainty. As our model is univariate, it does not consider other variables. In addition, if the forecast period is set to be longer, uncertainty accumulates and has a noticeable effect on the model. For this reason, shorter periods, such as a year, are considered when measuring the accuracy of a fit. The error metrics used to judge a time series model are mean and median absolute error, root means square error and the coefficient of determination.

While it may seem reasonable to optimise any hyperparameters for the model, FBProphet achieves a close to optimal result with its default settings when a preliminary grid search of parameters is done. A brief synopsis of the tweakable hyperparameters involves the degree of the Fourier series used in seasonality, with a balance between more terms that may introduce overfitting, or fewer terms, which may fail to capture proper seasonality.

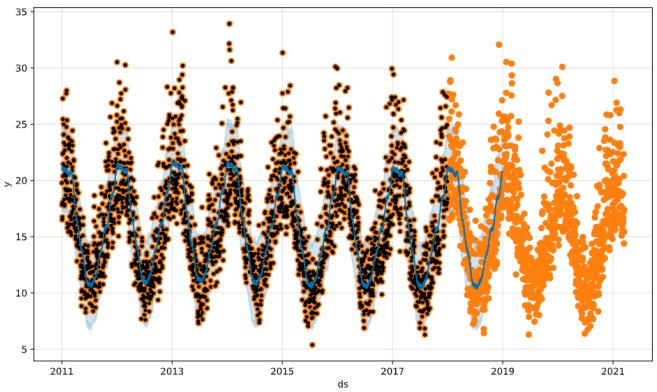


Figure 12: 1 year forecast of dry bulb temperature, given data from 2011 to 2018

With the default Prophet settings, various data sets are passed in, such as forecasting from 2011 to 2015 or from 2015 onwards for the dry-bulb temperature. Figure 12 details a one-year forecast of the year 2019, which yields an R2 value of 0.62, a root mean square error of 2.88 degrees, along similar metrics for mean and median absolute error. These metrics qualify as an excellent fit for a process that involves randomness and may be dependent on other unknown variables.

The seasonality components may similarly be examined for any trends. Figure 13 describes the weekly and yearly trends for our time series. In the latter case, the trends seem to line up with seasonal temperature changes, such as temperature drops in the middle of winter, a spike in temperature closer to summer and a gradual increase in temperature transitioning from winter to spring and then to summer. As for the smaller scale trend decomposition, there does not seem to be any signs of periodicity, as the primary assumption is that trend components are driven by the climate, which occurs on a larger scale of time. Observing the long-term trend seemingly implies that the temperature decreases over time, contrary to global warming; however, this is just localised data that does not account for historical data spanning decades or even centuries.

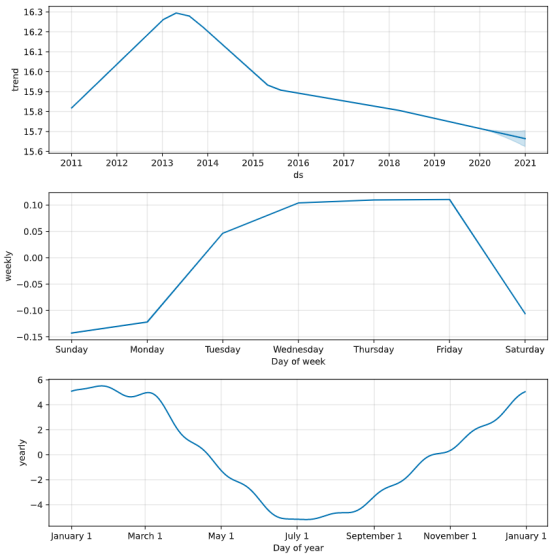


Figure 13: Seasonality components of the model in Figure 12 over a year, week and the decade,

Time Series Forecasting Addendum - Rainfall

In the process of experimenting, it was found that rainfall was remarkably harder to model, as early indicators showed that rainfall contained other non-seasonal components that were outside the scope of the data provided. In addition, rainfall contained many non-seasonal outlier days, which hinder the model's ability to forecast. Nonetheless, it is of interest to reason the effectiveness of the temperature model by juxtaposing the inefficiency of a forecasting rainfall model. The results are shown in Figure 14, and it is evident that rainfall outliers are more extreme in comparison to Figure 12, a forecast of the dry bulb temperature.

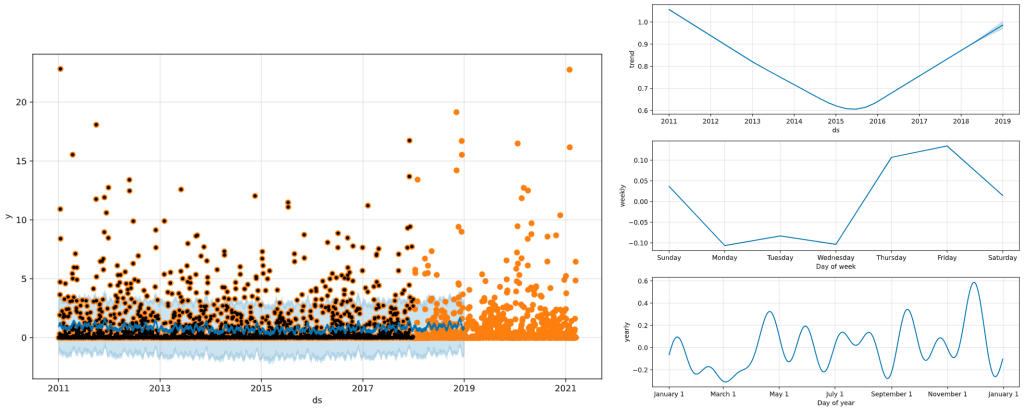


Figure 14: 1 year forecast of rainfall, given data from 2011 to 2017 along with trend components

Future outcomes

Had there been more data available for different weather phenomena, multivariate time series forecasting could have been implemented. The multivariate time series forecasting process involves using additional regressors, such as humidity, rainfall, or other data provided to facilitate forecasting of the temperature. Of course, an immediate issue is present; future regressor data must be available, defeating the purpose of forecasting. A possible method to forecast is to perform univariate forecasting with each regressor then use the predicted data for advanced temperature prediction. Additionally, implementing custom seasonalities for phenomena prediction allows for even more intricate models, as the degree of the Fourier series can be tweaked until there is an increase in accuracy forecasting. On top of the current library being used, NeuralProphet could be used instead as the library is a successor to FbProphet. It uses PyTorch to power deep learning-based forecasting, which extends beyond the basic AutoRegressive model. If decades of data was supplied, long term trends could also be detected, and changes in climate could be detected, which could be used to study the effects of global warming.

Rainfall classification forecasting

Additional processing of data

While the preliminary processing of the data was sufficient for previous aspects of the project, various additional modifications were required to ensure the rainfall data was as accurate as possible for modelling. Various problems were observed, including:

- Values at 9 AM were prematurely reset, resulting from inconsistent sample timestamps in the raw data and the method of resampling for the initial processing.
- Values decreased at a time other than 9:00 - 9:30 AM. This was considered an error in the data, given that the amount of rainfall caught by the recording device should not decrease.
- Rainfall was reported as a cumulative sum that would hinder comparison to literature definitions of rainfall intensity and was more confusing than rainfall per hour.

Once resolved, consideration was made to building a model that considers a short history of weather features instead of just a single day of weather data. Quite simply, the data features were duplicated and then added as extra features to the following day samples of the dataset. This meant that each sample had twice

the number of observations, one set for the current day and another for the previous day. This was repeated such that each sample consisted of three consecutive days of weather data.

Finally, the target variable was constructed from the rainfall amount using the following categories.

- No rain: 0 mm of rainfall in the past hour.
- Light rain: less than or equal to 1.5 mm of rainfall in the past hour.
- Moderate rain: greater than 1.5 mm but less than or equal to 4 mm of rainfall in the past hour.
- Heavy rain: greater than 4 mm of rainfall in the past hour.

The values were sourced from the American Meteorological Society but decreased to ensure enough samples existed for each category (American Meteorological Society, 2000). Each sample had its target variable calculated and then used as the target for the previous three-day sample so that the model would predict future classifications based on previous data.

Dataset balancing

The initial exploration of using a Random Forest Classifier (RFC) used default model parameters for the RandomForestClassifier from the sklearn Python module. This model had an accuracy of 63.90% ± 0.09%, a precision of 35.09% ± 0.84%, a recall of 31.58% ± 0.08%, and took 1.12 s ± 4.5 ms to train the model (n = 400, 95% confidence). Figure 15 depicts a confusion matrix for this model. This demonstrates how the model was predominately classifying as the major class[1], which paradoxically gave the model a high accuracy. The reason for this was that the dataset contained 60% of samples belonging to the major class, thus if the model classified samples only as the major class, it would get 60% correct. This was further supported by the low precision and recall of the model.

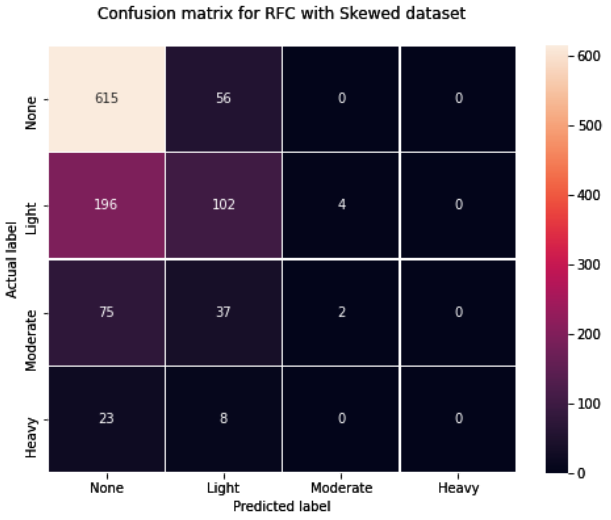


Figure 15. Confusion matrix of the default RFC model trained on the original (skewed) dataset.

[1] The *major class* is the class which contains the most number of samples. In this particular dataset, the major class is the 'no rainfall' class. Conversely, the *minor class* is the class which contains the least number of samples; in this case the 'heavy rainfall' class.

Evidently, the imbalance of the dataset was causing issues of overfitting in the model, thus, the dataset required balancing. There were many ways which this could have been done, so various techniques were compared to find the optimal strategy for balancing. The methods used were as follows:

1. **Down-sampling** (also known as under-sampling)
   This technique randomly sampled within each of the classes, without replacement, until each class contained the same number of samples as the minor class of the original dataset.
   A potential issue with using down-sampling was that it could lead to significant information loss if the difference between the size of the major and minor classes was quite large. The given dataset had a ratio of major class to minor class of approximately 16:1, hence this problem was a likely occurrence.
   The model trained on a down-sampled dataset had an accuracy of 36.95% ± 0.33%, the precision was 36.42% ± 0.35%, the recall was 36.95% ± 0.33%, and the time taken to train the model was 0.26 s ± 8.0 ms (n = 400, 95% confidence). Appendix A1 shows the confusion matrix for this model.

2. **Up-sampling** (also known as over-sampling)
   This technique was quite similar to down-sampling; however, it used random sampling with replacement to equalise the number of samples in each class to be the same number of samples as the major class.
   An issue associated with this technique was the potential for too many samples to be duplicated; quite obviously an issue which grows with the difference in the number of samples in the major and minor classes. This would lead to a model trained to expect exact samples. Testing would also have unexpectedly good results (depending on the method used to test) since the model would test on samples which it had trained on.
   The model trained on an up-sampled dataset had an accuracy of 94.9% ± 0.05%, the precision was 94.9% ± 0.05%, the recall was 94.9% ± 0.05%, and the time taken to train the model was 2.0268 ± 0.0194 (n = 400, 95% confidence). Appendix A2 shows the confusion matrix for this model.

3. **Mean-sampling**
   This technique was much like the previous two; each class was randomly resampled (with a replacement only if required). Each class contained the same number of samples as the average number of samples of each class in the original skewed dataset.
   This method was used to minimise the issues of the previous two techniques hopefully, but by doing so could also introduce both problems rather than having only one.
   The model trained on an mean-sampled dataset had an accuracy of 86.0% ± 0.09%, the precision was 85.86% ± 0.1%, the recall was 86.0% ± 0.09%, and the time taken to train the model was 1.1914 ± 0.0043 (n = 400, 95% confidence). Appendix A3 shows the confusion matrix for this model.

4. **Synthetic Minority Over-sampling Technique** (SMOTE)
   This technique resembles up-sampling, but instead of simply randomly resampling with replacement to duplicate samples, some form of interpolation was applied to the existing samples of the non-major class. This created new, unique samples within the bounds of the existing samples, ideally reducing the over-fitting of the model compared to up-sampling since the model would not have seen each of the samples multiple times.
   The model trained on a SMOTE dataset had an accuracy of 87.05% ± 0.06%, the precision was 86.96%

± 0.06%, the recall was 87.05% ± 0.06%, and the time taken to train the model was 3.2169 ± 0.0103 (n = 400, 95% confidence). Appendix A4 shows the confusion matrix for this model.

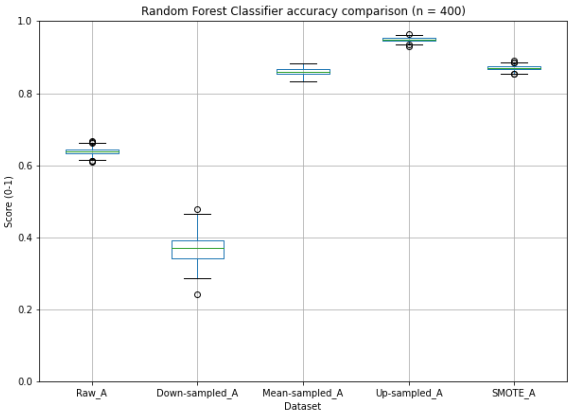


Figure 16. Box plot comparison of the accuracy of each of the five default RFC models trained on resampled datasets.

Figure 16 shows the accuracy of each of the default RFC models trained on the five differently resampled datasets. Interestingly, the best dataset in regards to accuracy was the up-sampled dataset. However, this is an example of the issue discussed prior; the model was training and testing on duplicates of various other samples. Consequently, this dataset was not used in the final model. Instead, the SMOTE dataset was used, given that it showed less evidence of over-fitting. It maintained good precision and recall in addition to accuracy (see Appendix B1 and B2 for box plot comparisons of precision and recall, respectively).

Model optimisation

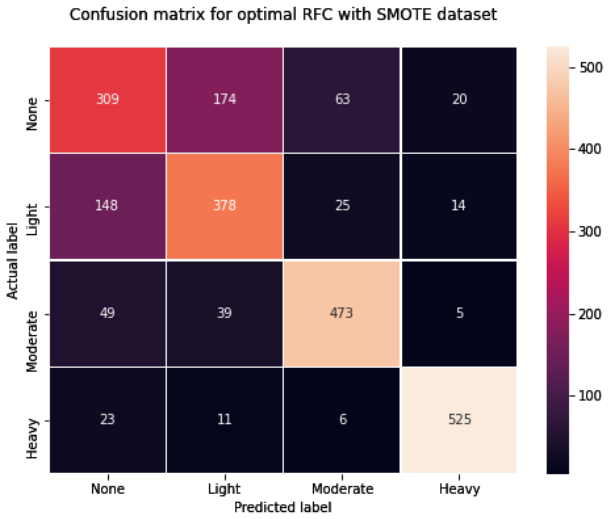


Figure 17. Confusion matrix of the optimised RFC model trained on the SMOTE dataset.

Since balancing the dataset was solved, the next step to developing the model was to optimise the parameters. This was primarily done using another sklearn function called GridSearchCV which

cross-validates models using every combination of RFC parameters given as a parameter to GridSearchCV. The range and variation of the RFC parameters that were given to the GridSearchCV function were determined manually.

The optimal parameters were found and used to finalise the RFC model to classify rainfall. The final model had an accuracy of 87.70% ± 0.07%, a precision of 87.64% ± 0.07%, a recall of 87.70% ± 0.07%, and took 7.6 s ± 0.03 s to train (n = 400, 95% confidence). Figure 17 depicts the confusion matrix for this model.

Model significance

The potential applications of a model which can accurately forecast rainfall are pretty expansive. Rainfall is most important for airport management, although this particular example would require models trained in more regions than just one city. This could be a potential avenue of exploration, developing a model that considers region or measurement. Rainfall forecasting is also widely used by the general public to aid decision making, such as what clothes to wear, when to have events and predict traffic conditions, among other things.

In future, it may be beneficial to attempt synthetic creation of samples using expected correlations with features to introduce a new class that represents storm-like rainfall. If possible, this would significantly improve the usability of the model for emergencies. It could also be applied to existing classes to replace SMOTE and may lead to improved accuracy.

It may also be beneficial to stack forecasts and their respective margins of error to obtain a forecast spanning multiple days. This would be more in line with the type of forecasts typical from professional meteorology bureaus worldwide.

Conclusion

The quality of the provided data calls into question the overall accuracy of the produced models. The data required significant processing to remove logical impossibilities and missing data. Therefore, the quality of the models produced in this project can only be said to be as good as the data quality. This is to say that the accuracy of these models should not be considered absolute until they can be tested on data of verifiably high quality.

Apropos the climatological phenomena, these external factors have an impact on Melbourne's weather. In line with the statements made by the Bureau of Meteorology, when the ENSO index is above 0, Melbourne has an increased likelihood of warmer weather and reduced rainfall. In conjunction, when the index falls below 0, Melbourne likely experiences a month cooler and wetter than what is historically expected. While both the temperature and rainfall in Melbourne are affected depending on the state of the El Niño–Southern Oscillation, it is seen that the amount of rainfall in Melbourne is altered the most.

Similarly, the positive phase of the IOD has been shown to influence both rainfall and temperature in Melbourne by causing higher temperatures and less rainfall due to the western Indian Ocean waters having a higher sea surface temperature and the eastern waters having a lower sea surface temperature. However, the negative phase of the IOD has only been shown to impact the rainfall in Melbourne, causing more significant amounts of rainfall than expected.

Inaccuracies in the time series and rainfall models could be a direct result of external climatological phenomena shown to cause irregularities in Melbourne's weather. Further investigation could include attempting to account for the various effects of these external climatological phenomena to produce a more accurate model. Employing custom seasonality for time series forecasting could have led to greater accuracy of the model by allowing the climatological phenomena to be accounted for with time series forecasting models.

Indeed, having more data than was provided could have led to better forecasting regarding the phenomena. This would be the case since the model would have then been able to learn the patterns of the phenomena, whereas currently, there is not enough data to establish many concrete patterns.

In future, other sources of data should be explored to bolster the scraped dataset. One particular method which could be used is sourcing data from personal weather station networks. These are networks of weather stations that people own on their properties. If possible to gain access to this network data, then a much broader area of Melbourne can be explored for developing a generic location model. It may also lead to data spanning a more extended period depending on the age of the network and local users of the network that existed during that time.

References

American Meteorological Society. (2000, June). *Rain*. Glossary of Meteorology. https://web.archive.org/web/20100725142506/http://amsglossary.allenpress.com/glossary/search?id=rain1

Australian Bureau of Meteorology. (2014, June). *What is El Niño and what might it mean for Australia?* Bureau of Meteorology. http://www.bom.gov.au/climate/updates/articles/a008-el-nino-and-australia.shtml

Australian Bureau of Meteorology. (2015, January). *Climate statistics for Australian locations*. Bureau of Meteorology. http://www.bom.gov.au/climate/averages/tables/cw_086071.shtml

Australian Bureau of Meteorology. (2021, April 19). *The Indian Ocean Dipole (IOD)*. Bureau of Meteorology. http://www.bom.gov.au/climate/enso/history/ln-2010-12/IOD-what.shtml

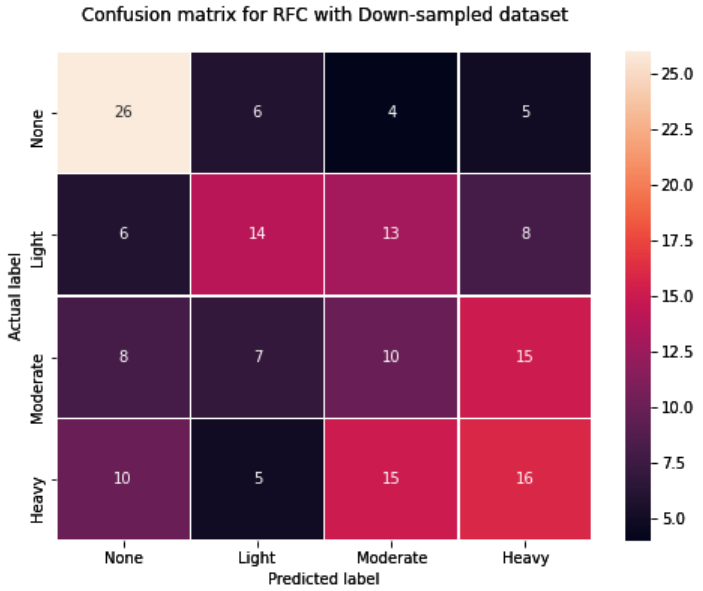
Working Group on Surface Pressure. (2020, May 1). *Dipole Mode Index (DMI)*. Global Climate Observing System. https://psl.noaa.gov/gcos_wgsp/Timeseries/DMI/

Appendices

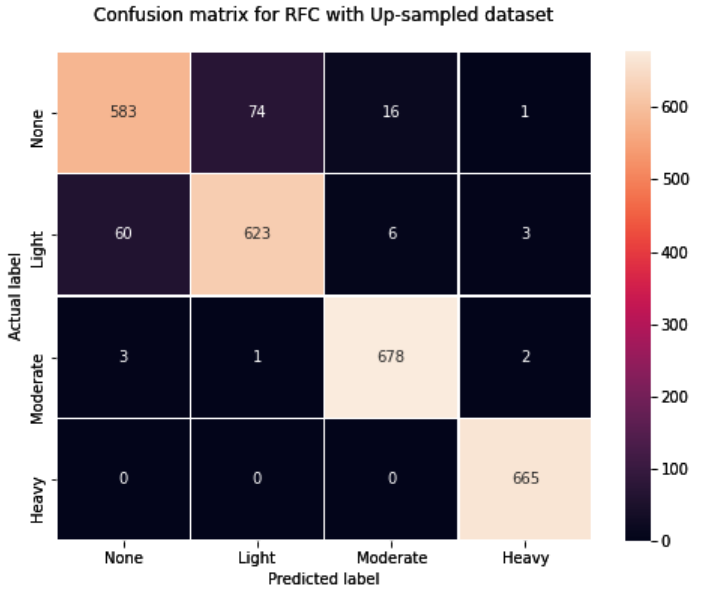
Appendix A

Confusion matrices for default Random Forest Classifier models using various forms of resampling for dataset balancing.

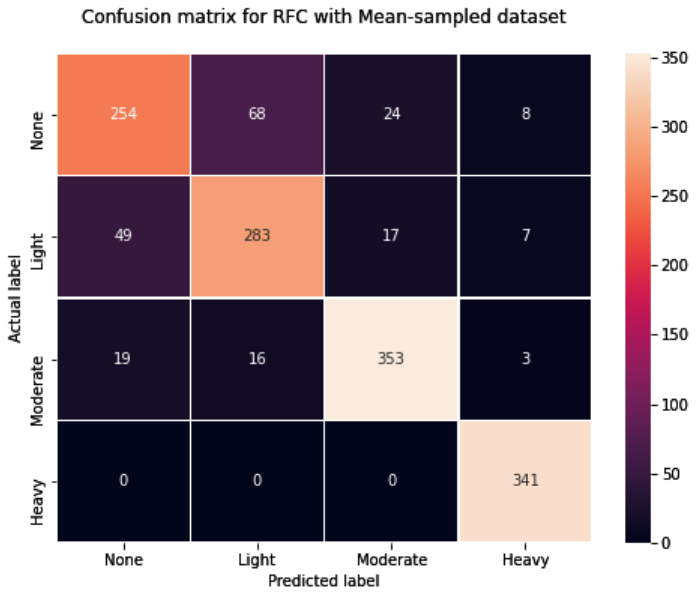
A1 : Confusion matrix of Down-sampled dataset default RFC model



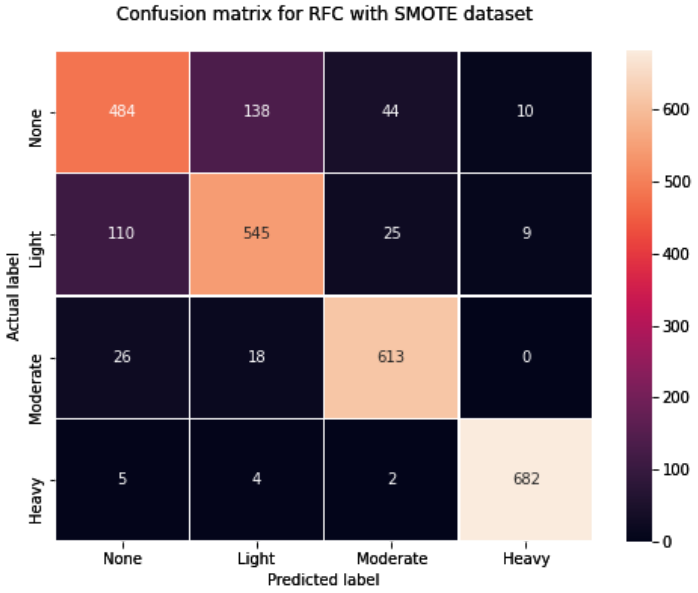
A2 : Confusion matrix of Up-sampled dataset default RFC model



A3 : Confusion matrix of Mean-sampled dataset default RFC model



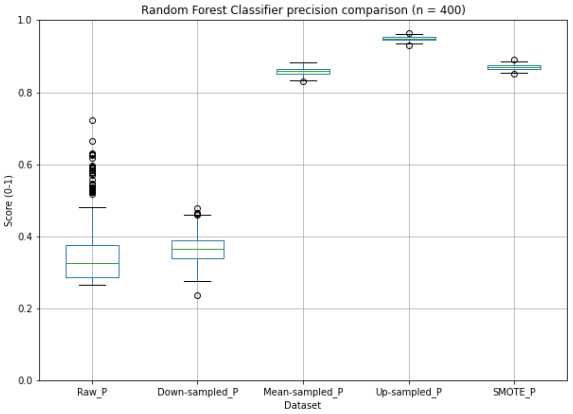
A4 : Confusion matrix of SMOTE dataset default RFC model



Appendix B

Box plot comparison of precision and recall for each of the five dataset default RFC models.

B1 : Comparison of precision



B2 : Comparison of recall