Capstone Project

Names: GIACOMO BUGLI, LUIGI NOTO, GUILHERME ALBERTINI

Introduction

Analysis of weather data is traditionally performed using complex physical models with a considerable number of variables that nonetheless provide fairly inaccurate predictions due to instability of the weather conditions. Thus, in this project we focused on the analysis of weather data to demonstrate that data science methods can be used to complement such complex models by testing hypotheses and training simple models on historical weather data that can possibly provide sufficiently accurate insights and forecasts over a short time period.
The first dataset used consists of about 10 years of daily weather data collected from multiple weather stations located across all Australia [1]. Each row represents an observation of a weather station on a given day and location (features *Date* and *Location*) where the rainfall, evaporation, sunshine, strongest wind gust direction and speed, minimum and maximum temperature on the 24 hours have been recorded; moreover, two measurements at 9am and 3pm respectively, have been recorded for wind direction, wind speed, humidity, pressure, temperature and fraction of sky obscured by cloud. Finally, two dummy variables indicate whether it rained that day (*RainToday*) and rained the day after (*RainTomorrow*). After processing the data to take care of null values and outliers, it is important to highlight the distribution of the target variable for Question 3, which is *RainTomorrow*. From the countplot, Figure 1, it is possible to see that there's a great imbalance between the two classes, where instances equal to 0 account for 77.62% of the total observations. This is consistent with expectations since observing rain is quite anomalous in large parts of Australia, after all. It would be then desirable for us to avoid metrics such as accuracy when evaluating the predictive models' performances largely due to the fact that deterministic predictors (classifying the datapoints as 0 with probability 1) will have such a preponderance. An alternative to accuracy would then be the average precision since this metric takes into account both positive and negative classes. Simply put, precision is the ratio of true positives to total predicted positives: we want almost all our forecasted days of rain to be days where it was actually observed.
Moreover, in light of what Question 1 will address, it is interesting to visualize the difference in average max temperature across all locations between the year of 2009 and 2017, Figure 2. It is possible to see, even on such small time range, that overall temperatures have increased, meaning that we should expect a significant difference over a larger time range.
The second dataset used contains 51 years of daily maximum temperature measurements, from 1971 to 2021, collected from a weather station located near Melbourne Airport. The data are available on the official website of the Australian Government Bureau of Meteorology[2].

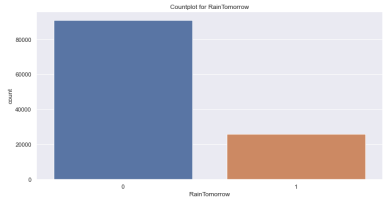

Figure 1: Countplot of the feature RainTomorrow.

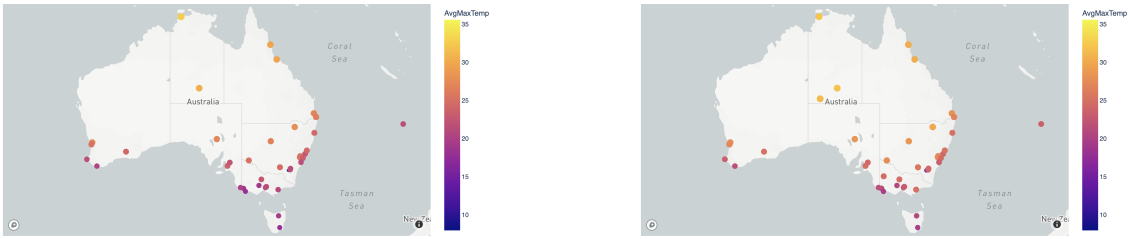

Figure 2: Map of the average max temperatures in the different locations between 2009 (left) and 2017 (right).

[1] https://www.kaggle.com/jsphyg/weather-dataset-rattle-package
[2] http://www.bom.gov.au/jsp/ncc/cdio/weatherData/av?p_nccObsCode=122&p_display_type=dailyDataFile&p_startYear=&p_c=&p_stn_num=086282

Question 1: is Australia getting hotter?

The first question that we want to answer is about investigating climate change in Australia, i.e. using Australian weather data to determine whether temperatures significantly increased over the years through hypothesis testing. Since combining the temperatures of multiple cities located all over Australia by averaging would not make much sense and just add noise to the observations, we decided to focus on just one location, i.e. the weather station located near Melbourne Airport.

As anticipated in the introduction, the dataset contains 51 years of daily maximum temperature measurements, from 1971 to 2021. We assessed whether there is a significant difference in temperatures over the period 1971-1999 vs the period 2000-2021 by running a permutation test. The null hypothesis is that the overall yearly temperature of the most recent period is higher than the one of the other period just due to random chance, i.e. it is not significantly higher. We adopt 0.005 as significance level.

In order to define the test statistic, we first computed the average temperature of each year, thereby obtaining 51 average temperatures, and organized them in two arrays, one for the period 1971-1999 and the other for the period 2000-2021. Since the value of a test statistic is usually large when the null hypothesis is false, the test statistic is defined as the difference between the average yearly temperature in 2000-2021 and the average yearly temperature in 1971-1999. We then used the 51 yearly average temperatures to determine the distribution of the test statistic under the null hypothesis. We formed two synthetic arrays for the two periods (of the same size of the original ones) by sampling without replacement from the 51 measurements, and computed the "shuffled" test statistic using such arrays. This process has been done 10 million times, obtaining 10 million observations of the shuffled test statistic. By plotting a histogram of these values, we get an approximation of the null distribution of the test statistic.

The resulting "exact p-value", i.e. the proportion of shuffled test statistics greater than or equal to the empirical test statistic, is $1 \cdot 10^{-7}$. We conclude that yearly average temperatures in 2000 or after are significantly higher than yearly average temperatures before 2000.

As a result, our analysis suggests that Melbourne got hotter over the last few decades. This reasoning most probably would extend to most or all of Australia, and can be verified by applying our analysis to data collected by other weather stations.

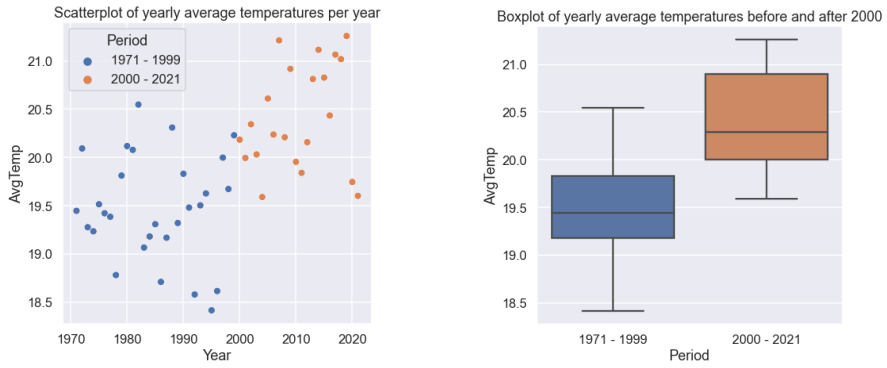

Figure 3: Scatterplot and boxplot of yearly average measurements from 1971 to 2021.

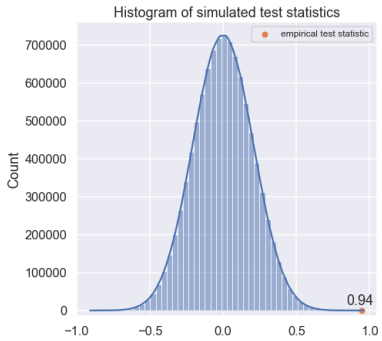

Figure 4: Histogram and kernel density estimation of the shuffled test statistics.

Question 2: do locations differ based on frequency of precipitation?

The second question that we wanted to answer regards the difference in frequency of precipitation between Australia's major geographical areas across the 10 years. In order to do so, we first proceed to cluster the 43 different locations on the dataset based on latitude and longitude. Relying on the Elbow rule (Figure 5), we find that the optimal number of clusters is 4. From Figure 6, it is possible to see that the 4 clusters are effectively corresponding to North, South, West and East of Australia. Then, based on this result, we can substitute the 43 different locations with their corresponding cluster label so that we would have less categories to encode in the prediction part of Question 3.

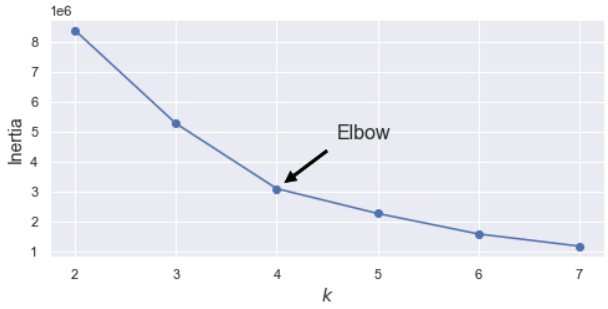

Figure 5: Plot of inertia for each k in 2 to 7

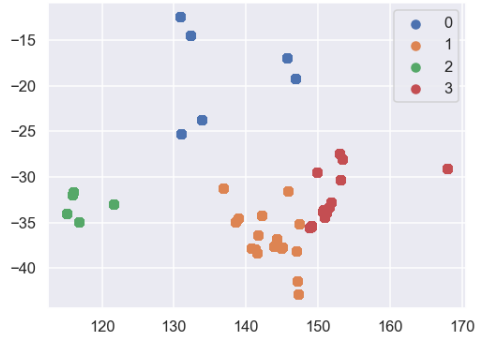

Figure 6: Scatter plot of the locations and corresponding cluster

We can then perform hypothesis testing in order to determine if the differences between precipitations on the 4 geographical areas are significant. Given that the dependent variable (*RainToday*) is nominal, we can apply the Kruskal-Wallis H test. The null hypothesis is that the samples originate from the same distribution, while the alternative hypothesis is that at least one distribution of one group is different from the distribution of at least one other group. By performing the test on the variable *RainToday* for each geographical region the associated p-value is $2.3084 \cdot 10^{-32}$, which is significant at $\alpha = 0.005$. We can then conclude that there is a difference in the frequency of precipitation between the North, South, West and East regions. Given this result, we can investigate more and test comparisons between groups to determine which groups are different. This can be done by performing the Kruskal-Wallis test (which for only two groups is equivalent to a Mann-Whitney U test) for each possible combination of the four regions obtaining the following p-values:

- *South vs East:* $9.7105 \cdot 10^{-16}$
- *South vs West:* 0.6842
- *South vs North:* $4.3980 \cdot 10^{-10}$
- *East vs West:* $6.0299 \cdot 10^{-8}$
- *East vs North:* $2.7424 \cdot 10^{-30}$
- *West vs North:* $2.5180 \cdot 10^{-8}$

From these results we can see that the difference in the frequency of precipitation is significant for all combinations of the four regions, except between South and West. Since the locations labeled as West are only 5, the reason behind the non-significance of the test could be low amount of samples that happen to register a frequency of precipitation similar to the ones in the south. In order to address this issue and see if the frequency of precipitation is similar we should increase the number of locations in the West region.

Question 3: Can we effectively predict rain tomorrow using the data?

About ten years' worth of weather data (first cleaned, then encoded and scaled) were ingested by three classifier models: Extreme Gradient Boosting, Random Forest Classification, and Logistic Regression. In attempting to predict whether the model at hand would effectively predict the occurrence of rain on the following day the data was logged, we see that we run into the immediate issue of data imbalance. Australia rarely experiences days of rain, and when it does, precipitation is minimal. We stratify with respect to the occurrence of rain on the next day. As there are too few instances of the minority class (*RainTomorrow*) for a model to effectively learn the decision boundary, stratification helps with balancing the imbalanced data set partitions when it comes to cross validation. With these approaches, we set out to assess the performance of the model chiefly in regards to the average *precision* due to the aforementioned; we seek the fraction of relevant instances among the retrieved instances due to the significant class disparity.

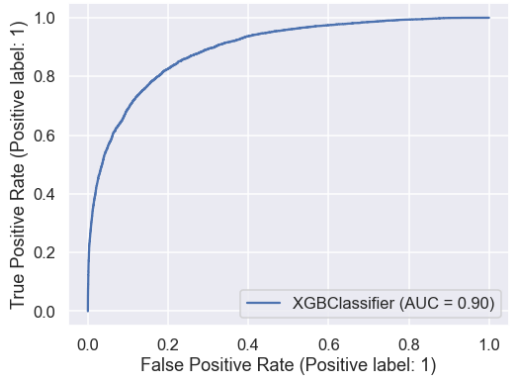

Figure 7: ROC plot of XGBoost Classifier with AUC of 0.90

```
Estimator: Random Forest
Fitting 7 folds for each of 12 candidates, totalling 84 fits
Best params are : {'criterion': 'entropy', 'max_depth': 50, 'n_estimators': 200}
Best training average precision: 0.7415
Test accuracy score for best params: 0.8618
Test recall score for best params: 0.5124
Test precision score for best params: 0.7890
Test average precision score for best params: 0.7463
Test ROC AUC for best params: 0.8935

Estimator: Logistic Regression
Fitting 7 folds for each of 6 candidates, totalling 42 fits
Best params are : {'C': 0.2, 'penalty': 'l1'}
Best training average precision: 0.7013
Test accuracy score for best params: 0.8515
Test recall score for best params: 0.5155
Test precision score for best params: 0.7343
Test average precision score for best params: 0.7037
Test ROC AUC for best params: 0.8739

Estimator: Extreme Gradient Boosting
Fitting 7 folds for each of 1 candidates, totalling 7 fits
Best params are : {'learning_rate': 0.2, 'max_depth': 12, 'n_estimators': 300}
Best training average precision: 0.7500
Test accuracy score for best params: 0.8638
Test recall score for best params: 0.5752
Test precision score for best params: 0.7511
Test average precision score for best params: 0.7549
Test ROC AUC for best params: 0.8954
```

Figure 8: Evaluation metrics for the three models

Additional simplifications were made along the way, with the regional mapping of cities being mentioned prior and the focus on month as opposed to the more granular features of latitude, longitude, and day that were dropped. After preprocessing, GridSearchCV analyzed the best pairings of parameters under the average precision metric. The results, along with the recall and accuracy are given in Figure 8. Notice immediately that Extreme Gradient Boosting bested the other two classifier models under the average precision metric; its AUC of 0.90 is depicted on the ROC plot found in Figure 7. The best pre-selected parameter pairings for every model are given in Figure 8.

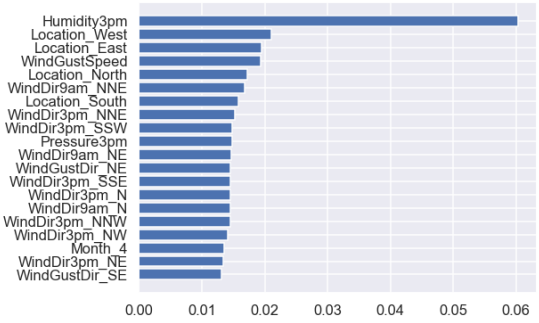

Figure 9: Bar plot of the 20 most important features for XGBoost prediction

As we can see from Figure 9 the level of humidity measured at 3pm plays an important role in the prediction of rain tomorrow. This is consistent with intuition, since humidity can be generally thought both as a precursor and result of rain. Geographical indicators weigh more heavily on the presence (or absence) of rain on the following day, whereas wind direction generally has less of an influence on this outcome. Wind gust speed does also factor more heavily into our predictions but is shadowed by the role that afternoon humidity plays in determining future rain outcomes.

Overall Summary and Conclusions

In this project, we presented an analysis on weather data to address different type of questions related to temperature and precipitation. Overall, we have learnt that Australia is a country characterized by low precipitations and increasingly high temperatures, with an heterogeneous distribution in terms of weather conditions over its extent.
In particular, we were able to assess with high significance that climate change on the city of Melbourne has brought substantial increases in temperatures between the periods of 1971-1999 and 2000-2021, which is likely to be reflective of the nation's overall climate trend. Moreover, a significant difference in the frequency of precipitation is found between North, South, East and West regions across the 10 year period from 2008 to 2017. This is indeed confirmed in Question 3 where the geographical location of the city plays an important role in predicting future precipitations.
The classification models (random forest, logistic regression, and extreme gradient boosting) set out to predict the presence of rain tomorrow given a data set that logged the presence of rain on the day alongside other factors (chiefly temporospatial in nature). The Extreme Gradient Booster model, though hardly optimized, produced the best model under the metric of average precision (producing an AUC of 0.90).
It's important to underscore a few aspects about these last results. First of all, we made the decision to use a feature location that is significantly less granular than the different 45 locations from before, reducing the important categories to 4. This necessarily reduces the variance for the single observations and hence the ability of the model to accurately predict rain in the following day due to less precise data points. However, this is an effective when predicting the weather of specific locations for which we don't have previously logged data, allowing the model to extrapolate to every city of Australia, beyond the locations for which we already have observations. Moreover, this decision allows us to have less parameters to optimize in the models which in terms means less running time and computational power required, crucial when having limited resources as in our case. Finally, we obtain a precision of .75 — the model is pretty effective in classifying positive occurrences; when it predicts rain it usually rains, while a recall of .58 means that it is more conservative in its predictions, which is preferable since rain is not very frequent in Australia.
As regards limitations and further analysis, obviously a richer dataset in terms of time range and number of locations would have helped in obtaining more precise results. In particular, as concerns Question 2, data related to more locations would have allowed for more precise clustering and probably a significant difference in the frequency of precipitations between South and West regions given that the latter are underrepresented. Consequently, a better clustering would, without a shadow of doubt, improve the models' predictions in Question 3, providing for more representative locations.
With regard to Question 1, as anticipated before, it would be interesting to verify whether there is a significant increase in temperatures in Australian locations other than Melbourne Airport, by carrying out a similar hypothesis test to the one performed in this project. Moreover, further analysis can be carried out by adding precipitation and CO_2 emissions data from 1971 to 2021 for Melbourne, in order to see whether there are also significant variations in rainfall and frequency of rain and investigate the role of CO_2 in such climate changes.