MaSSP 2021 - Machine Learning & Data Science
Rain in Australia

Nguyen Ngoc Khanh [1] **Nguyen Huy Hai** [1] **Tran Long Hung** [1] **Duong Minh Khoi** [1]
{nguyenngockhanh.pbc, hainguyenhuy2002, hunglongtran2004, minhkhoi102004}@gmail.com

Abstract

The study is conducted to build a machine learning model to predict Australia rain. Adopted from previous works, we used standard techninques in exploratory data analysis and feature engineering to learn from data. We made use of day, month, year features instead of date since they appeared to affect the weather strongly. Moreover, we introduced two new classes of features: direction embedding and time information. Direction embedding is a technique to map direction data into euclidean space that better preserves the distance between close directions. Time information includes past data into current prediction that has its own advantages and disadvantages. In order to tackle a large number of features, we used *XGBoost* to select the best 10 features while maintaining high performance.

1. Introduction

Rain in Australia is a dataset containing ten years of observation in several locations in Australia. This study contains an intensive exploration of data, feature engineering and modelling the dataset to predict whether it will be raining tomorrow given all previous data.

2. Related Work

On *Kaggle*, *Rain in Australia* was studied by many people. Majority of these methods consider this problem as classifying whether it will rain tomorrow based on the collected data today together with feature engineering to build the prediction model.

In *Extensive Analysis - EDA + FE + Modelling* (Banerjee, 2020), the authors did Extreme Value Analysis for variables that have distribution close to normal and Interquartile range for skewed distribution to remove the detectable outliers. Furthermore, to deal with incomplete data, they filled those missing values by the corresponding median.

In *Rain Prediction using seven popular models* (D, 2020), they made use of *MICE* algorithm (van Buuren & Groothuis-Oudshoorn, 2011) to impute missing data points. Although the set of features is not perfectly independent, the authors proceeded to build the model with all features after examining the importance of each feature using the filter method and wrapper method.

Time Series Analysis on Australian rain (das Subeen, 2021) introduced a time-series perspective to the problem. In the findings, one noteworthy point is the very high correlation between rain today and rain tomorrow.

Even though the previous methods successfully produced a very good performance, some techniques might improve the performance of models. We have conducted experiments on two additional feature sets: direction-embedding and time information.

3. Exploratory Data Analysis and Feature Engineering

3.1. Target feature

The target feature is of the categorical type containing Yes, No, and NaN. Since NaN has no actual meaning in our study, we dropped all entries with NaN in the target.

3.2. Categorical features

We have six categorical features in total including *Location*, *Date*, *RainToday*, *WindGustDir*, *WindDir9am*, *WindDir3pm*. Following previous works, we encoded *Date* as a combination of *Day, Month, Year* since the weather in real life typically depends on which month and which it is currently. For direction features (*WindGustDir*, *WindDir9am*, *WindDir3pm*), we used *direction-embedding* that maps direction onto a unit circle.

3.2.1. Location

Different locations might have different rain characteristics. Figure 1 shows the distribution of rain probability over all locations. Some locations have a very low rain probability such as Woomera (202 raining days in 2990 records), while others have a very high rain probability such as Portland (1095 raining days in 2996 records). Hence, *Location* is a key features for our model.

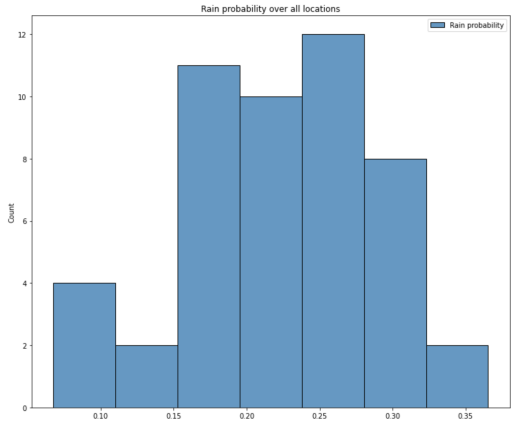

Figure 1. distribution of rain probability in different locations

3.2.2. Day Month Year

Following previous works, we replace *Date* by the combination of *Day*, *Month* and *Year*. In figure 2 and 3, *Year* and *Month* appears to be directly related to the rain probability.

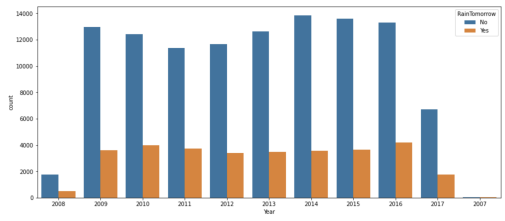

Figure 2. number of raining days in different years

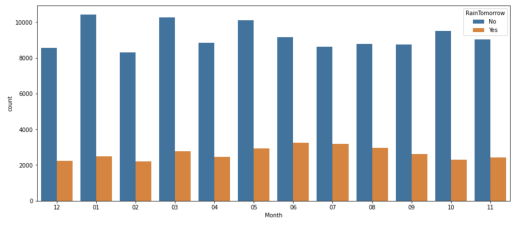

Figure 3. number of raining days on different months

Despite the fact that *Day* is not directly related to the probability of rain, we still added *Day* to the feature set since the independence can be explained by the high frequency. *Day* alone might not be a factor affecting the rain on the next day, but some combination of *Day* and *Month* might.

3.2.3. Rain Today

RainToday is an important feature to our prediction in an obvious way. As rain usually lasts for several days. So that most of the data points lie on sequences of rainy days and sequences of sunny days. Therefore, rain is a state that is not likely to change switch. The figure 4 demonstrated our intuition.

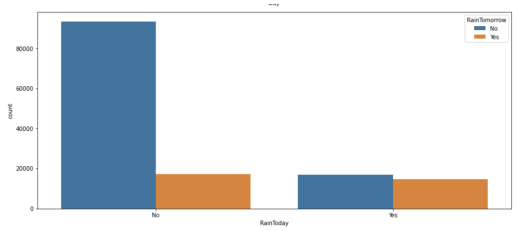

Figure 4. number of raining days given the previous day

3.2.4. Wind Direction

For the same reason as *Location*, different wind directions have their different effects on rain tomorrow, as shown below. The count of rainy days appears to be a sine wave along the direction, especially *WindGustDir* and *WindDir3pm*. Therefore, we imposed the use of *direction-embedding* for wind direction as the following.

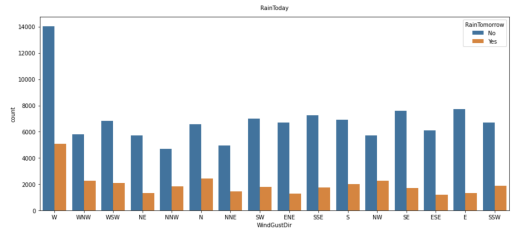

Figure 5. number of raining days given wind gust direction

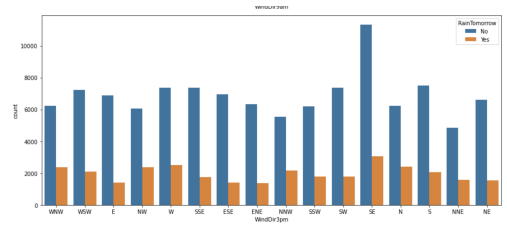

Figure 6. number of raining days given wind diretion at 3pm

Direction-Embedding Our goal is to embed direction into Euclidean space that better preserves the closeness between directions. We mapped each direction into a point on the unit circle and introduced two new features associated with the angle formed by that direction and a base direction (we used East as our base direction), namely DirSin and DirCos.

A direction feature is now embedded into a two-dimensional space and treated as numerical features.

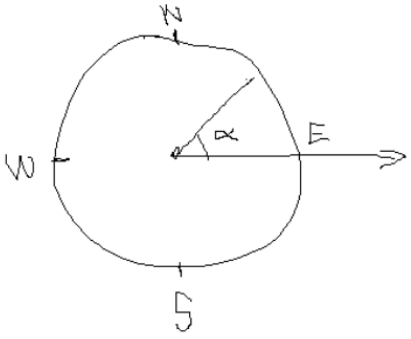

Figure 7. direction-embedding

3.2.5. Missing data issue

To deal with missing data in categorical features, we imputed those records with missing data by their respective most frequent value (mod).

3.3. Numerical features

We have 16 numerical features [1] together with 6 new direction-embedding features corresponding to 3 direction features [2].

We plotted the distribution of features with respective to different target variables. Humidity at 3pm appears to be the feature changes the most if the target feature switches between raining and sunny.

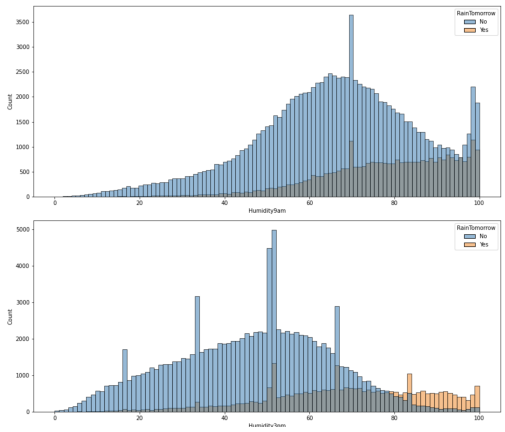

Figure 8. Humidity at 9am and 3pm

All numerical features follow a normal distribution loosely so that we did not perform any transformation on any field.

3.3.1. Correlation between numerical features

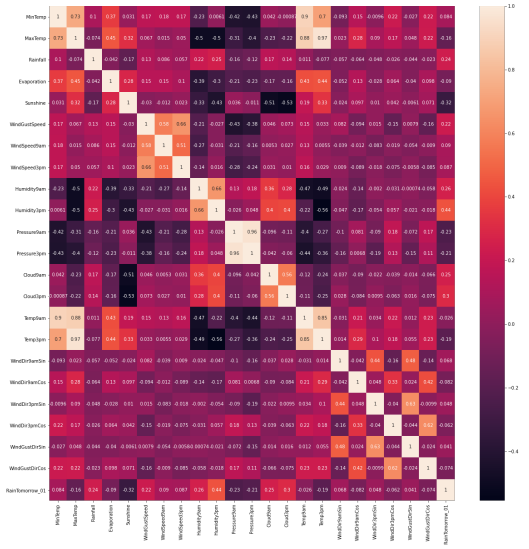

Figure 9. Correlation between numerical features

In the correlation matrix, we did not observe any highly correlated pair of features, and all of them have some effects on the target feature. Hence, we decided not to discard any features.

3.3.2. Missing data issue

To deal with missing data in numerical features, we imputed those records with missing data by their respective median.

3.3.3. Outlier removal

For outlier detection and removal, we have used IQR and capped those values outside the chosen range.

3.4. Time Information

Similar to the effects of analyzed features, the features in one day before the observation day somewhat contributed to our target variable. As shown in the figure 10 below, the rain on Wednesday is not only affected by the observation on Tuesday but Monday.

Therefore, we included yesterday observations into our feature set. This change made number of features doubled, so in the next section, by using XGBoost, we overcame the issue by select only a small fraction of feature set.

[1] MinTemp, MaxTemp, Rainfall, Evaporation, Sunshine, WindGustSpeed, WindSpeed9am, WindSpeed3pm, Humidity9am, Humidity3pm, Pressure9am, Pressure3pm, Cloud9am, Cloud3pm, Temp9am, Temp3pm

[2] WindGustDir, WindDir9am, WindDir3pm

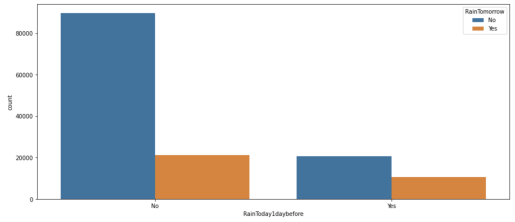

Figure 10. number of raining days given the previous day

4. Models and Evaluation

4.0.1. Evaluation Metrics

Since the problem is binary classification, we chose AUC as the performance metric to evaluate the model. This metric eliminates the need for fine-tuning classification threshold.

4.0.2. Model and Hypothesis Testing

As mentioned earlier, the two hypotheses of the study are (1) the effect of direction embedding and (2) the effect of time information. Additionally, we want to select a small fraction of feature set due to the large number of features after adding time information.

Logistic Regression We chose logistic regression due to its deterministic nature. In logistic regression, we selectively added/removed some features and measured the performance (AUC) of the model with K-fold cross-validation and oversampling for imbalance labels. Firstly, we chose the baseline as the model with current day data. Then, we removed direction features and direction-embedding features. Finally, we added yesterday data into input features.

model	AUC
baseline	0.8647
baseline - direction	0.8648
baseline - direction embedding	0.8637
baseline + yesterday data	0.8697

In the results from logistic regression, removing direction-embedding reduced the AUC score and adding yesterday data improved the AUC score. Surprisingly, removing categorical direction actually increase the AUC score. Categorical direction seems to be very bad for logistic regression.

XGBoost with hyperparameter search In the second experiment, we firstly used XGBoost (Chen & Guestrin, 2016) on the whole dataset to find the feature importance [3]

[3]Top 15 features: Humidity3pm, RainToday, WindGustSpeed, Sunshine, Cloud3pm, Pressure3pm, Rainfall, WindDir3pmSin, Location, WindDir9amSin, WindDir3pmCos, Humidity3pm1daybefore, Month, WindGustDirCos, WindGustDirSin

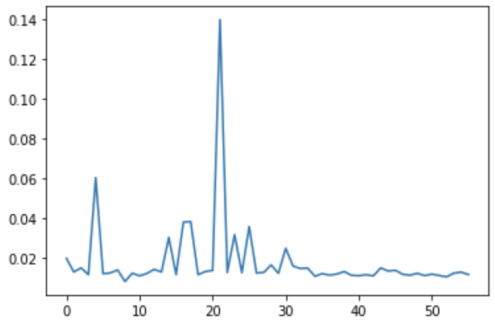

Figure 11. feature importance (unsorted)

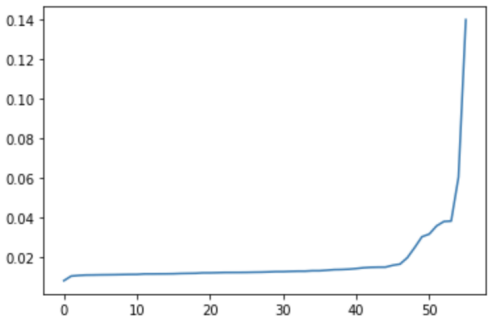

Figure 12. feature importance (sorted)

In the top 15 features, there is no occurrence of any direction feature. Additionally, for all 3 direction features [4], direction embedding features are more likely to have higher feature importance.

After that, we perform search on the hyperparameter space as follow: number of features [1, 10], number of trees [50, 100], tree max depth [2, 5], L1 norm regularization [0, 1], L2 norm regularization [0, 1]. We ran Bayesian Optimization for 50 iterations and finally got the AUC of 0.8820 where the hyperparameters at number of features 10, number of trees 95, tree max depth 5, L1-norm regularization 1.0 and L2-regularization 1.0. Comparing with the performance on all features of XGBoost, the result of using only 10 features was very close to the actual all features model of AUC 0.8922.

5. Conclusion and Future Work

In conclusion, we have demonstrated a slightly better performance using direction embedding and time information from the dataset. Furthermore, we selected the ten best features and achieved a performance very close to all features performance using XGBoost, which reveals those most important features related to rain in Australia. One might

[4]WindGustDir, WindDir9am, WindDir3pm

predict the rain tomorrow with relatively high accuracy using some simple features such as humidity at 3 pm and whether it is raining today.

For future work, we have planned to solve the problem with other methods such as frequency-based methods and neural network methods, which can be better in handling time-series data since the current day, month, year using a fixed frequency that might not be true in the real world

References

Banerjee, P. Extensive analysis - eda + fe + modelling, 2020. URL https://www.kaggle.com/prashant111/extensive-analysis-eda-fe-modelling.

Chen, T. and Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL http://doi.acm.org/10.1145/2939672.2939785.

D, C. Extensive analysis - eda + fe + modelling, 2020. URL https://www.kaggle.com/chandrimad31/rainfall-prediction-using-7-popular-models.

das Subeen, S. Extensive analysis - eda + fe + modelling, 2021. URL https://www.kaggle.com/sudhakordassubeen/time-series-analysis-on-australia-rain.

van Buuren, S. and Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011. URL https://www.jstatsoft.org/v45/i03/.