Will it rain tommorrow?
(Predict next day rain in Australia)

Ruoheng Yuan

Data Science Institute
Brown University
Github link: https://github.com/DaveYuan23/Data1030_rain_au

December 9, 2024

Table of Contents

Data Recap

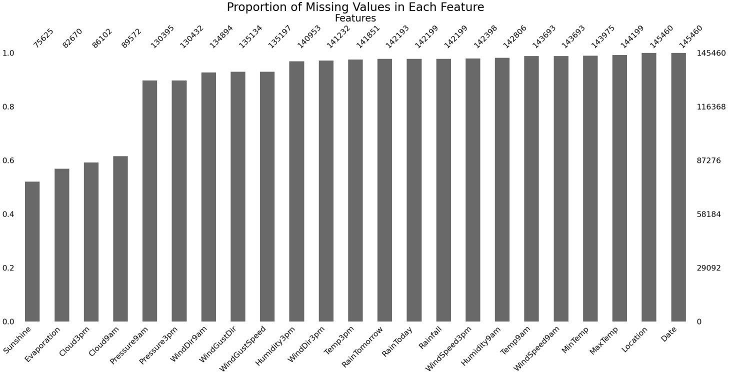
Rain plays an essential role in our lives, as it provides water for plants, animals, and humans. The weather department works to forecast when it will rain, and similarly, I aim to predict whether it will rain in Australia tomorrow.

- **Problem Type:** Binary classification task determining whether it will rain tomorrow.
- **Data source:** Kaggle Rain in Australia dataset https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package/data.
- **Data collection method:** The observations were gathered from a multitude of weather stations from Austrail goverment. You can access daily observations from http://www.bom.gov.au/climate/data.

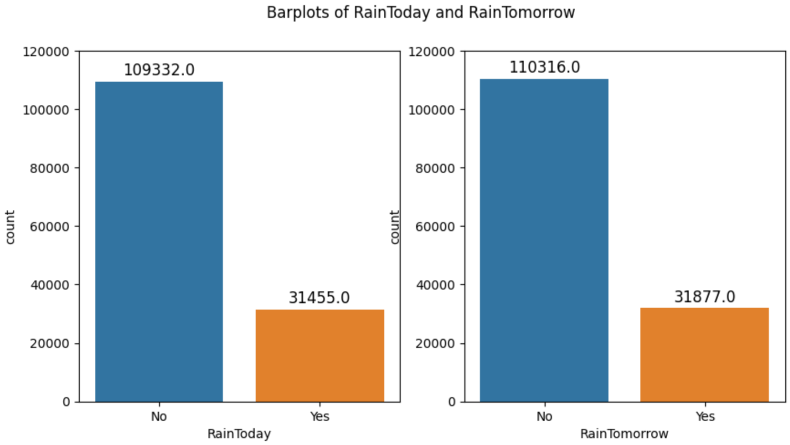
Missing Values

This dataset contains 10 years of daily weather observations from 49 different Australian weather stations. It includes 16 continuous variables and 6 categorical variables.

- **Dataframe Shape:** The dataframe has 145,460 rows and 23 columns.
- **Missing Values:**

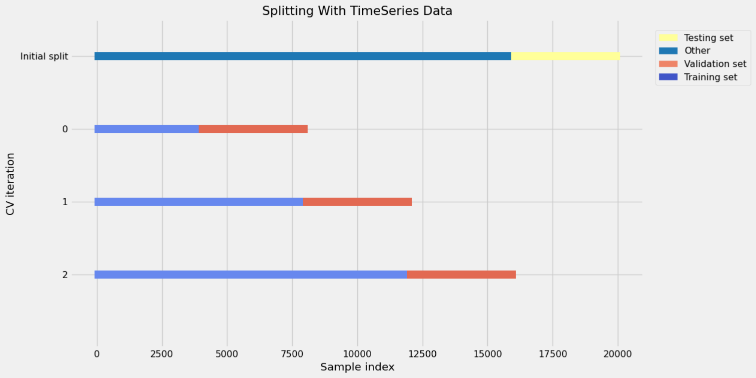


Rain Tommorrow?



- RainToday is a feature, and RainTomorrow is the target variable.
- **Time Series Property:** Data points have a temporal dependency, so we consider it as a time series problem.

Splitting



- We first use train_test_split without shuffling the data to separate the test set from our dataset.
- We then use TimeSeriesSplit from Scikit-learn to split the remaining dataset.
- This approach ensures that we do not use future information to train our model!

CV Pipeline

Input: Set of models $\mathcal{M}$, random states $\mathcal{R}$, unique rows $\mathcal{U}$ (reduced feature method), parameter grid $\mathcal{P}$, number of folds k

Output: Comparison of test scores for different models

```
foreach model m ∈ M do
    foreach random_state r ∈ R do
        foreach unique_rows u ∈ U do
            foreach parameter p ∈ P do
                foreach fold f in k-fold do
                    Compute validation score val_score;
                end
                average_val_score across k-folds;
            end
            Find the best parameter p* based on average_val_score;
            Obtain test score using the best estimator;
        end
        Repeat the pipeline and compute test scores over all r random
         states;
    end
    Compare different models using their test scores;
end
```

Algorithm 1: Pipeline with Reduced Feature Method

Supervised ML Algorithms

Table: ML Algorithms and Parameter Grid

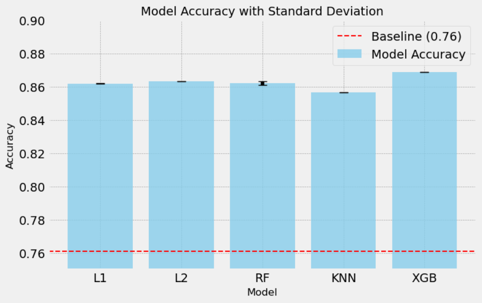
ML Algorithms	Reduced Feature	Regularization Term	Tuning Parameters	Parameter Grid
Logistic Regression	Yes	L_1	C	$[0.01, 0.1, 1, 10, 100]$
Logistic Regression	Yes	L_2	C	$[0.01, 0.1, 1, 10, 100]$
Random Forest	Yes	–	$max\_features$ max_depth	$max\_features : ['log2', 'sqrt']$ $max\_depth : [2, 4, 8, 16, 32]$
KNN	Yes	–	n-neighbors	$[2, 4, 6, 8, 16, 32, 64]$
XGBoost	No	–	α, λ	$\alpha : [0, 0.01, 0.1, 1, 10, 100]$ $\lambda : [0, 0.01, 0.1, 1, 10, 100]$

- **ML Models:** Logistic Regression with L1, L2 regularization terms, Random Forest, KNN, and XGBoost.
- **Handling Missing Values:** Only XGBoost can be trained directly on datasets with missing values. For the other models, we applied a reduced feature method to address the missing values.

Test Scores Comparison

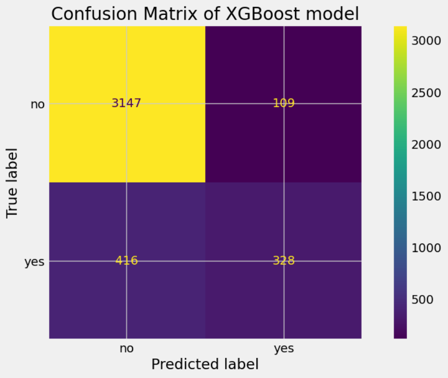
Table: Model Performance Comparison

	Baseline	L1	L2	Random Forest	KNN	XGBoost
Accuracy	0.761	0.862	0.858	0.864	0.857	0.869



- The baseline accuracy for this problem is 0.761.
- XGBoost achieves the highest test accuracy at 0.869, followed by Random Forest and L1.

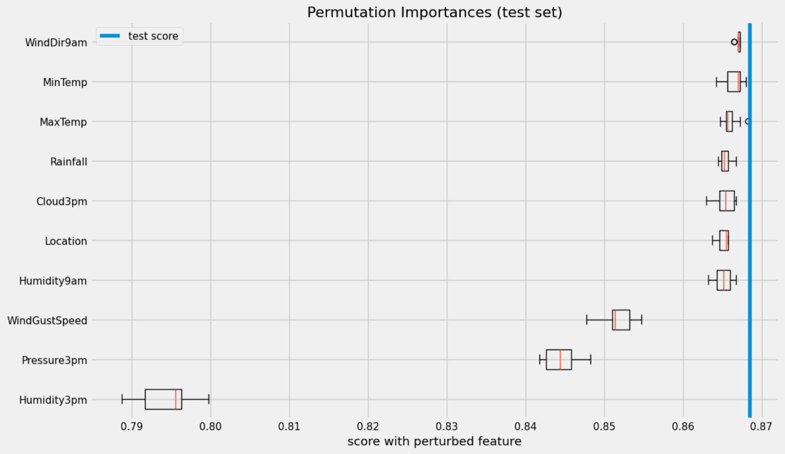
Confusion Matrix



	Precision	Recall	Accuracy	F1
Value	0.751	0.441	0.869	0.555

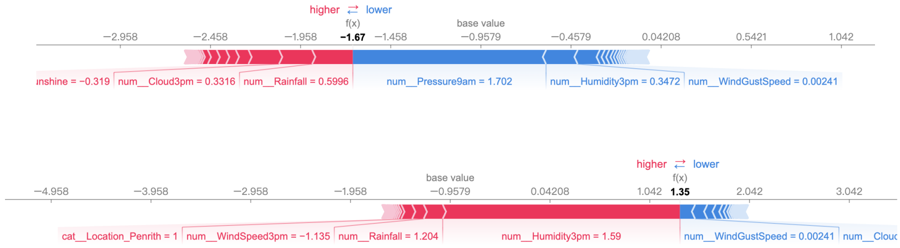
Table: Summary of XGBoost Performance Metrics

Permutation Importance



- Humidity3pm has the highest Permutation Importance.
- Pressure9am and WindGustSpeed are the second and third highest features.

Shap Force Plot for two observations in the testing set



- For the first data point, *Pressure9am* and *Humidity3pm* push the baseline value to -1.67.
- For the second data point, *Rainfall*, *Windspeed*, and *Humidity* push the baseline value to 1.35, suggesting that it will probably rain tomorrow.

Outlook

- **Improve predictive power:**
  - Use more data points to train the ML pipeline.
  - Experiment with advanced deep neural networks, such as MLP and LSTM.
  - For continuous variables, explore different feature engineering techniques, such as log transformation, Box-Cox transformation, or adding polynomial features.
- **Improve interpretability:**
  - Calculate various global feature importance metrics and compare the differences among them.
  - Utilize LIME to explain individual predictions with local approximations.

Question?