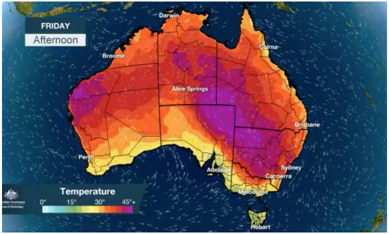
Predict Rain in Australia.

by Y. Kostrov

Contents

Overview



The purpose of this project is to use weather data set from Kaggle to predict rainfall for the next day, based on the data about today's weather.

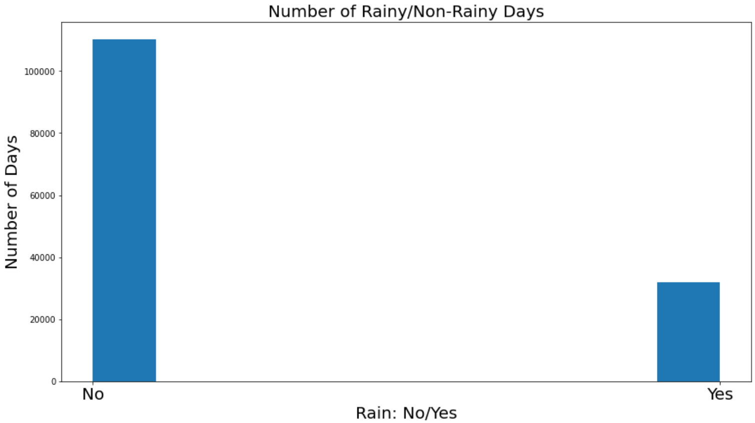
Business Problem

- Predicting rainy weather for the next day is a very important task.
- Usually weather is predicted by using complicated deterministic models involving partial differential equations.
- I will suggest a model that predicts weather by using Machine Learning.

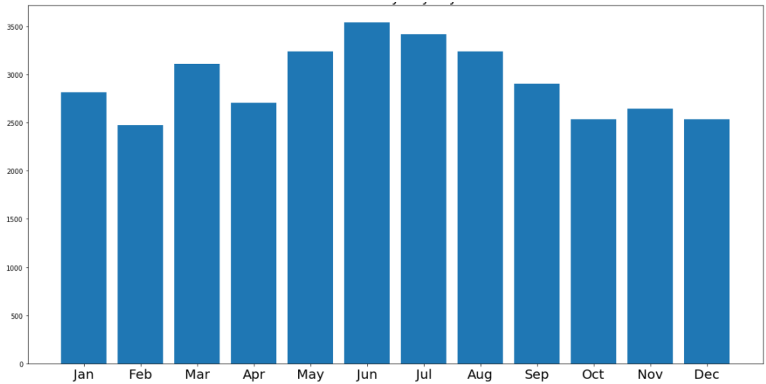
Data Used in the Project

- This data set contains about 10 years of daily weather observations from many locations across Australia.
- RainTomorrow is the target variable to predict. It means – did it rain the next day, Yes or No? This column is Yes if the rain for that day was 1mm or more.

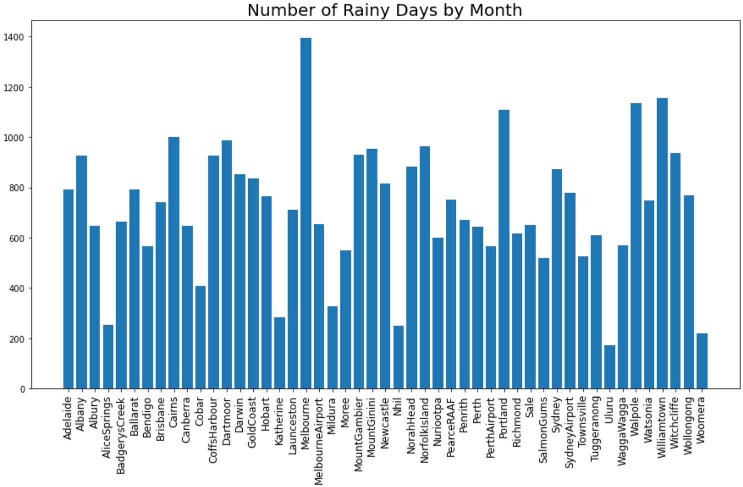
Number of Rainy and Sunny days.



Rainy and Sunny days by Month.



Rainy and Sunny days by City.



Modeling: Creating Models

I have built the following two classifiers

- Random Forest Classifier
- XG Boost Classifier

I used F1 metric to assess the models.

Explanation of Recall

- Recall is defined as:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} = \frac{\text{True Positive}}{\text{Total Actual Positive}}$$

Explanation of Recall

- Recall calculates how many of the Actual Positives our model captures by marking it as Positive (True Positive).
- Thus Recall is a better model metric when there is a high cost associated with False Negative.
- In our case False Negative is predicting "No Rain" when there is a "Rain Tomorrow".

Explanation of Recall

- For instance, in rain prediction.
- If it rains tomorrow (Actual Positive) is predicted as no rain tomorrow (Predicted Negative), then the person who relies on the prediction will be really upset since being unprepared for bad weather.

Explanation of Precision

- There is another metric we have to watch for, called "precision".
- Precision is defined as:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} = \frac{\text{True Positive}}{\text{Total Predicted Positive}}$$

Explanation of Precision

- Precision describes how precise/accurate your model is out of those predicted positive, how many of them are actual positive.
- Precision is a good measure when we worry about the costs of False Positive.

Explanation of Precision

- In our rain prediction, a false positive means "No Rain" tomorrow (actual negative) has been identified as "Rain" tomorrow.
- It is not that bad, since a person will carry an umbrella or rain coat for nothing.

Explanation of F1

- I use F1 metric in my analysis.
- F1 is a function of Precision and Recall.

Explanation of F1

- Looking at Wikipedia, the formula is given as follows:

$$F1 = 2 \times \frac{\text{Precision} \times \mathit{Recall}}{\text{Precision} + \text{Recall}}$$

- F1 is used when you look for the balance between Precision and Recall and there is imbalance in class distribution.

How Well Models Performed

- Logistic Regression achieved 89% on the F1 metric and it is balanced on the precision and recall at 89%
- XGBoost achieved 88.8% on the F1 metric and it is, also, balanced on the precision and recall at 89% and 88% respectively.

Business Suggestion

Based on my analysis,

- I suggest to use XGBoost model for the prediction of rain tomorrow based on the data about today's weather.

THE END

THANK YOU!