
Given Today's Weather,
Will It Rain Tomorrow?
GitHub
EricB10
samraykhman

The Process

Acquire Dataset

Kaggle Dataset

Fill Missing Values

Dark Sky API

Interpret Data

Exploratory Data Analysis

Prediction Modeling

Machine Learning

Machine Learning Study: Rain Forecast Predictions

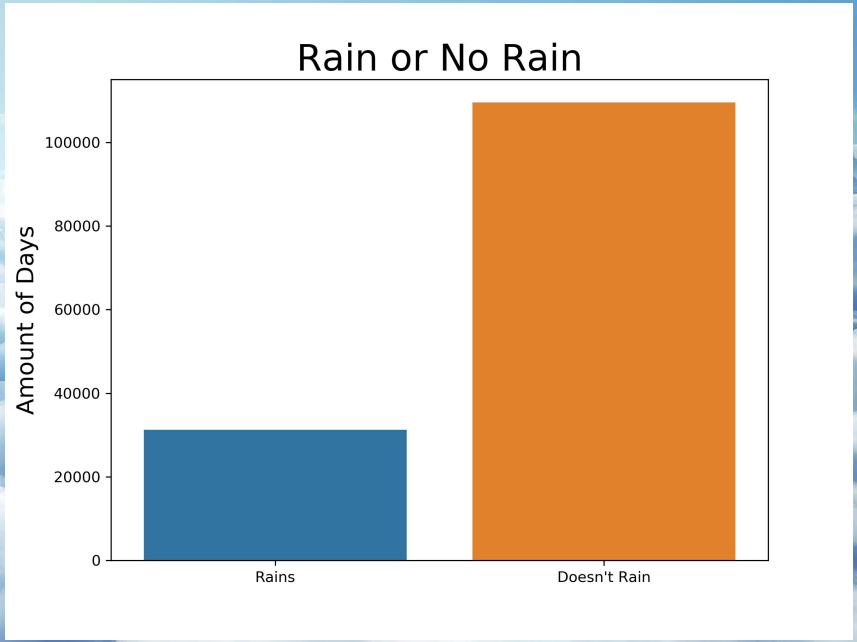
50 Cities in Australia

10 Years

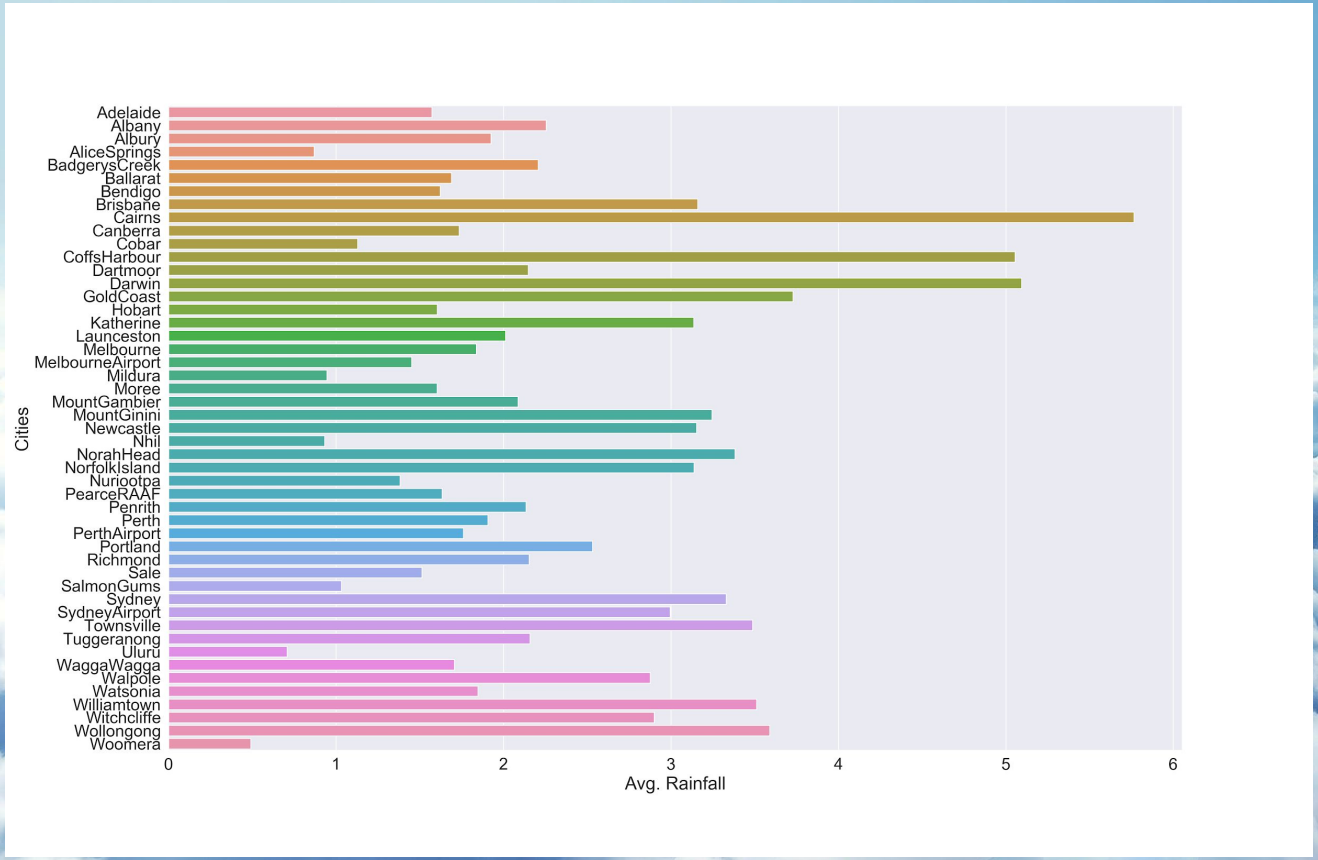
Daily Weather Data

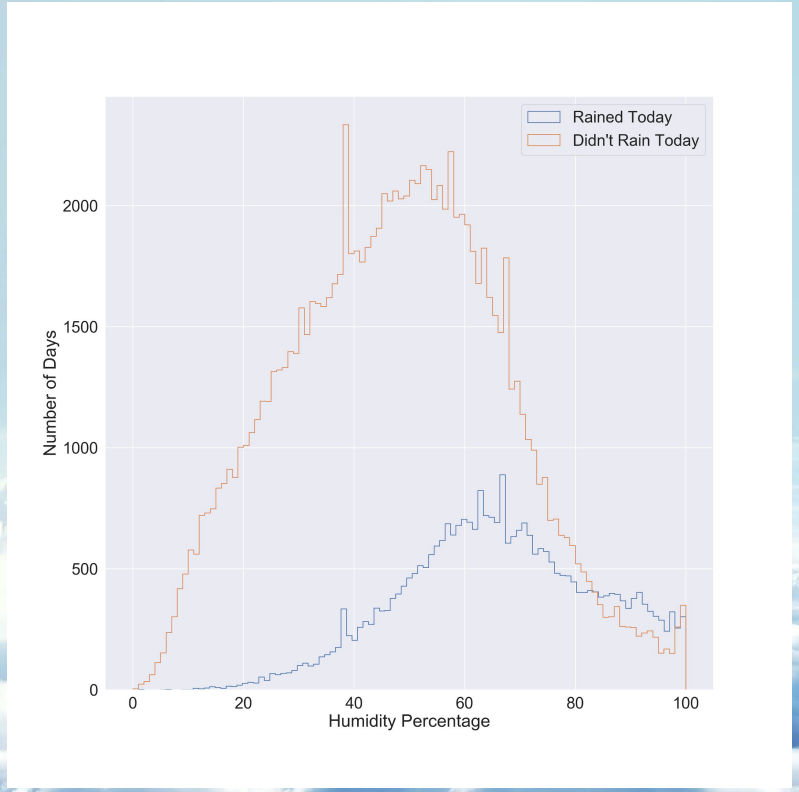
Data Used for Predicting Rain the Following Day:

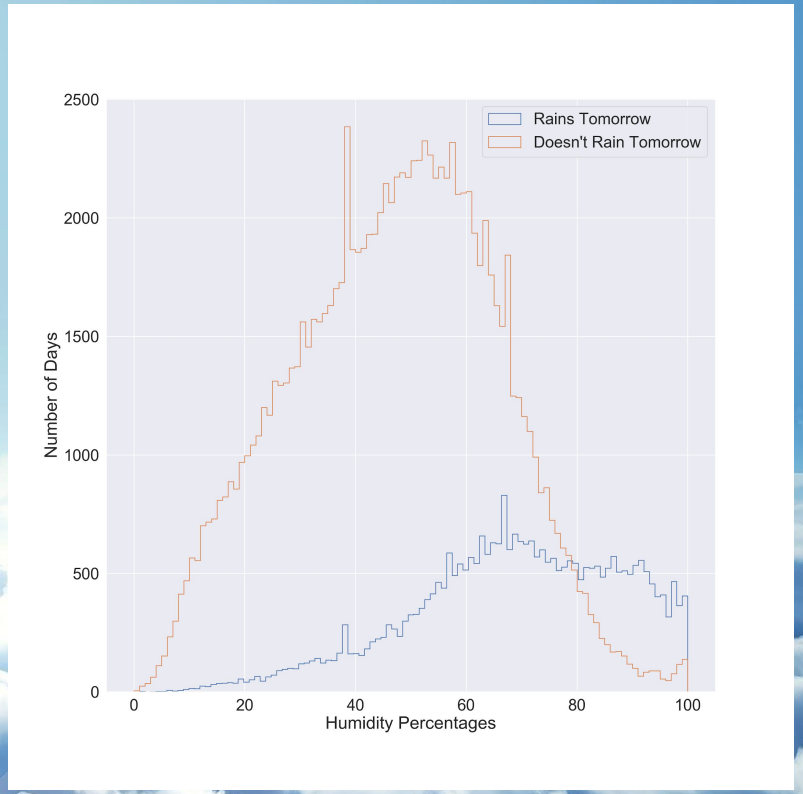
- Date
- Rainfall & Evaporation
- Mean Rainfall per City
- Temperature, Pressure & Humidity
- Sunlight & Cloud Coverage
- Wind Direction & Speed



Mean Rainfall per City







Distribution of Humidity Percentages

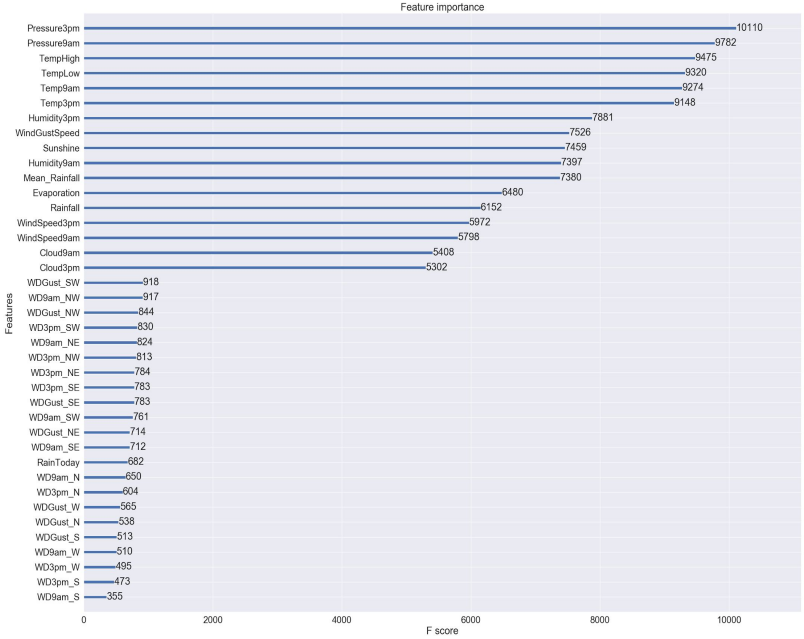
Prediction Model Comparison

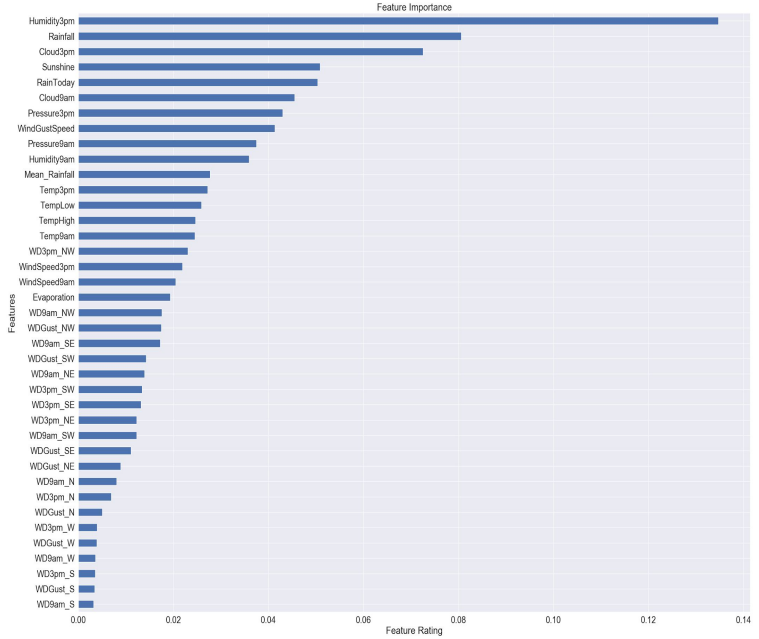
	Accuracy	Precision	F1
Logistic Regression	0.8487	0.5007	0.5914
Random Forest	0.8573	0.6182	0.6555
Naive Bayes	0.7953	0.6556	0.5843
Ensemble by Voting	0.8561	0.5473	0.6254
XGBoost	0.8588	0.8436	0.5766

Expectations of Feature Importance

- We expected Today's Rainfall would heavily influence Tomorrow's, but other features such as Temperature, Pressure and Humidity had a greater impact.
- We expect Cloud Coverage *would* have been important; however, we were unable to obtain many of the values. We decided to replace the missing values with the median value per city and as a result the variable was not impactful.

Feature Importance of Best Models





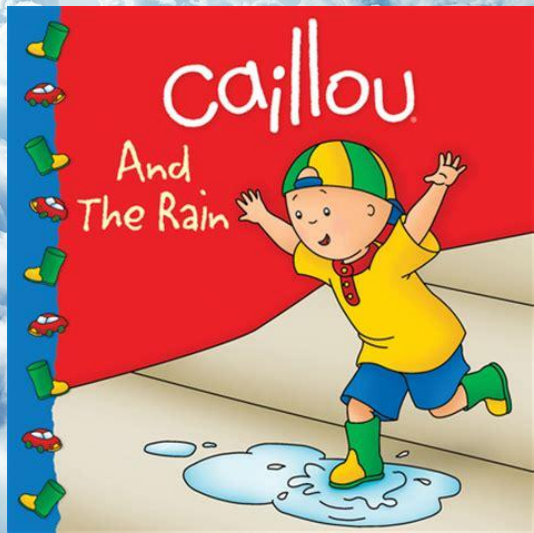
XGBoost

Random Forest

Takeaways

- Higher humidity and lower pressure levels are a great indicator for predicting rain. Temperature is also an important indicator but has a nonlinear relationship. Wind speed and direction were not very helpful.

- Most models had a fairly high degree of Accuracy. XGBoost had the highest Precision & Accuracy but Random Forest had the highest F1 Score. It was difficult to choose one over the other since we would rather predict false positives than false negatives, but also achieve a reasonable balance.



Given More Time...

- We would have liked to explore more methods of data collection such as web scraping to fill in missing cloud coverage and evaporation data.
- We would have liked to experiment with and compare more prediction models in attempt to improve Accuracy, Precision and F1 Scores.
- Finally, after fine tuning the best model we would like to create a Pipeline workflow to easily plug new data into.


Thank You!
Special thanks to Matt Wasserman
for lending us his Dark Sky API key
GitHub
Eric Blander
Sam Raykhman