Predicting Rain for Tomorrow
Jaidaa Alidrisi

Abstract

The goal of this project is to use classification models to predict whether it is going to rain the next day or not based on weather observations from the previous day. I worked with dataset from Kaggle that was provided by Bureau of Meteorology in Australia, and by leveraging what we learned in data cleaning and exploratory data analysis then apply machine learning classification models I found that logistic regression is the model with the most promising results as it predicts rain with 83% accuracy and 64% f1 score.

Design

The aim of this project is predicting rain using machine learning models and analyzing weather features that could affect the possibility of rain, and it is beneficial for various fields. As a software developer in the National Center for Meteorology, I can help weather forecasters and meteorologists with issuing weather alerts to the public or to other governmental sectors in case of severe weather conditions. In addition, predicting rain assists farmers managing and improving their produce and it also helps event organizers plan their outdoor events better.

Data

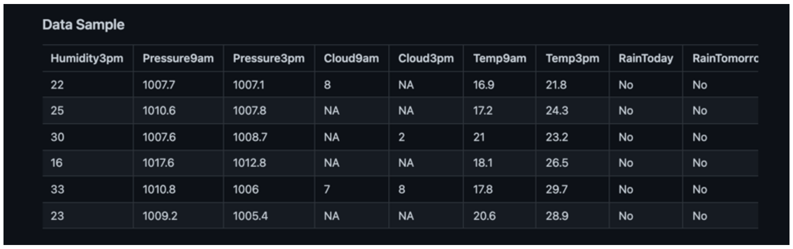
The dataset I used can be obtained from Kaggle and it was provided by Bureau of Meteorology in Australia, and it can be downloaded through this link. It contains 10 years of weather observations including 145461 records and 23 features some of them are numerical and some are categorical. The highlighted features that will help predict rain are:

- MinTemp: The minimum temperature registered in the day
- MaxTemp: The maximum temperature registered in the day
- Humidity9am: Humidity at 9AM
- Humidity3pm: Humidity at 3PM
- Pressure9am: Atmospheric pressure at 9AM
- Pressure3pm: Atmospheric pressure at 3pm
- Temp9am : temperature at 9AM
- Temp3pm: temperature at 3PM
- Rainfall: The amount of rain
- Location (city name)
- Wind Direction
- Wind Speed
- RainToday: whether there is rain today or not

And the Target column is:

- RainTomorrow: weather there is rain tomorrow or not

Data Sample



Data Sample

Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RainTomorrc
22	1007.7	1007.1	8	NA	16.9	21.8	No	No
25	1010.6	1007.8	NA	NA	17.2	24.3	No	No
30	1007.6	1008.7	NA	2	21	23.2	No	No
16	1017.6	1012.8	NA	NA	18.1	26.5	No	No
33	1010.8	1006	7	8	17.8	29.7	No	No
23	1009.2	1005.4	NA	NA	20.6	28.9	No	No

Algorithms

- **Feature Engineering**
    - Converting categorical features (Location,wind gust direction and wind direction) to binary dummy variables to make them more useful and expressive.
    - Create new features (year,month,day) of Date feature
    - Imputing missing values using SimpleImputer
    - Scaling numerical features to ensure that no feature has a disproportionate impact on the model's loss
- **Modeling and Evaluation**
    - Logistic Regression, Random Forest Classifier, and XGBClassifier were applied to the dataset. Since we have imbalanced class distribution, F1 score was the metric I used for evaluating each model because it evaluates the mean between precision and recall, therefore it will give us better prediction for both majority and minority classes.
    - Evaluation :

Model Name	Accuracy	Precision	Recall	F1 Score
Logistic Regression	83.51%	62.15%	66.12%	64.08%
Random Forest Classifier	77.05%	48.95%	74.75%	59.16%
XGBClassifier	77.35%	49.34%	69.13%	57.58%

Tools

- Numpy and Pandas for data manipulation
- Seaborn and Plotly for data visualization
- Scikit-learn for modeling

Communication

- PowerPoint slides for displaying the data analysis and modeling results
- Project Writeup.
- Jupyter notebook for EDA and modeling details.

Future Enhancements

- Test the models with different dataset such as weather stations from Saudi Arabia.
- Hyperparameter tuning for all models and reevaluating them.