Predicting Rainfall in Australia

...

PROJECT GROUP #3
Hanumasri Bollepalli (801104071)
Madhuri Pawle (801083244)
Pavanitha Jampala (801131462)

Contents

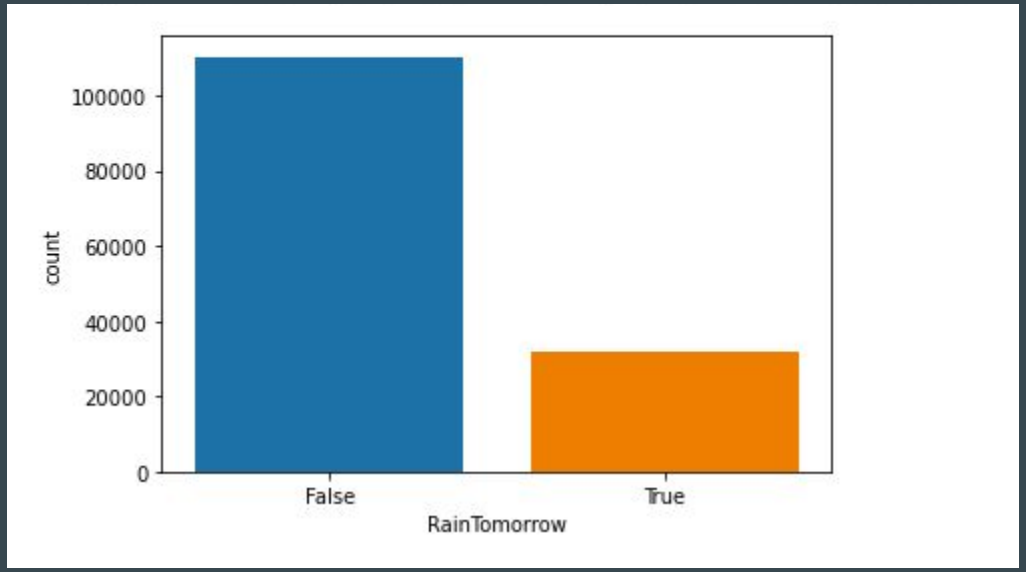
Introduction

- The aim of the project is to predict if it will rain in Australia.
- The target audience is the general public of Australia.
- EDA (Exploratory Data Analysis) was performed to visualize various important trends.
- A logistic regression model was trained to predict rain.
- The model was evaluated using various classification metrics.

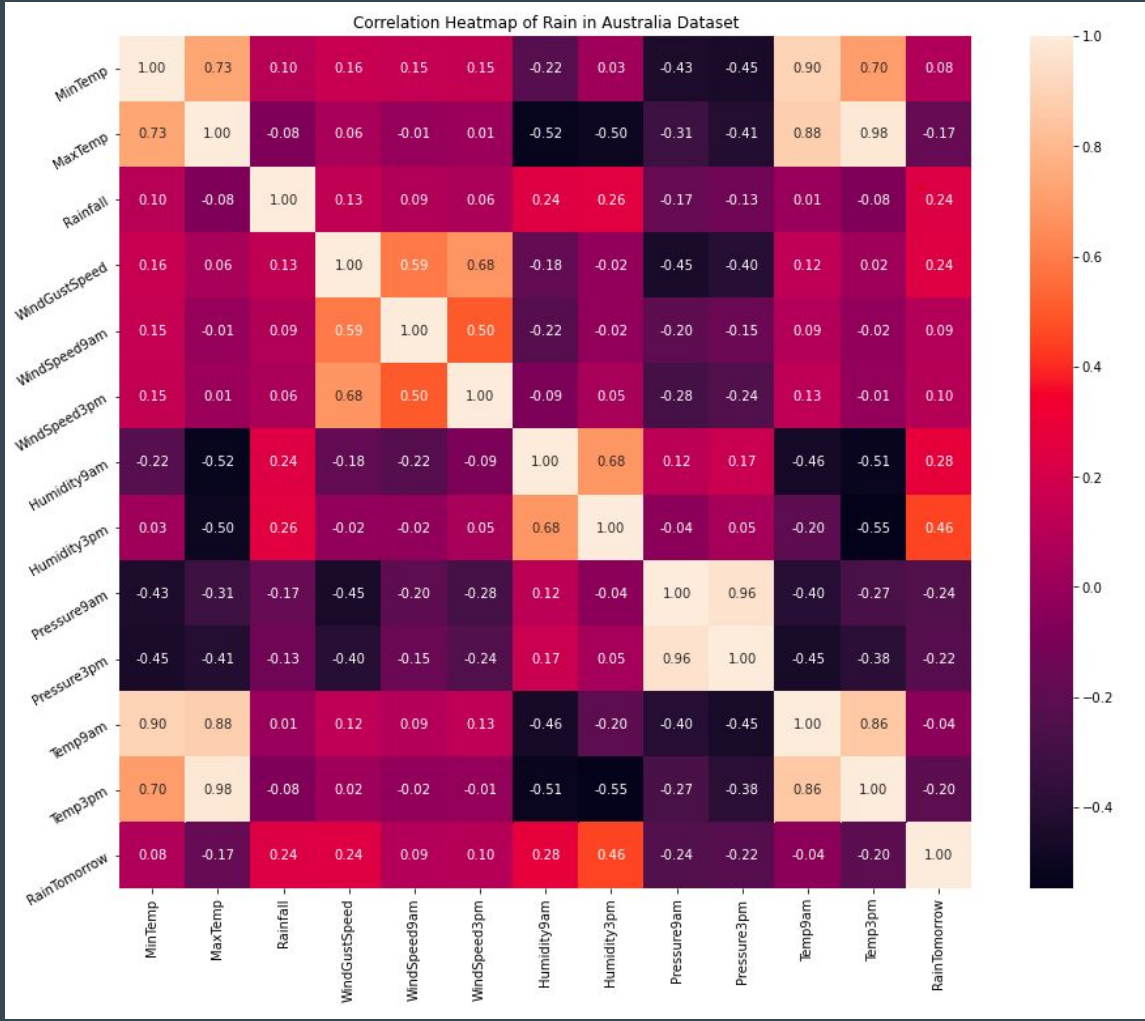
About the Dataset

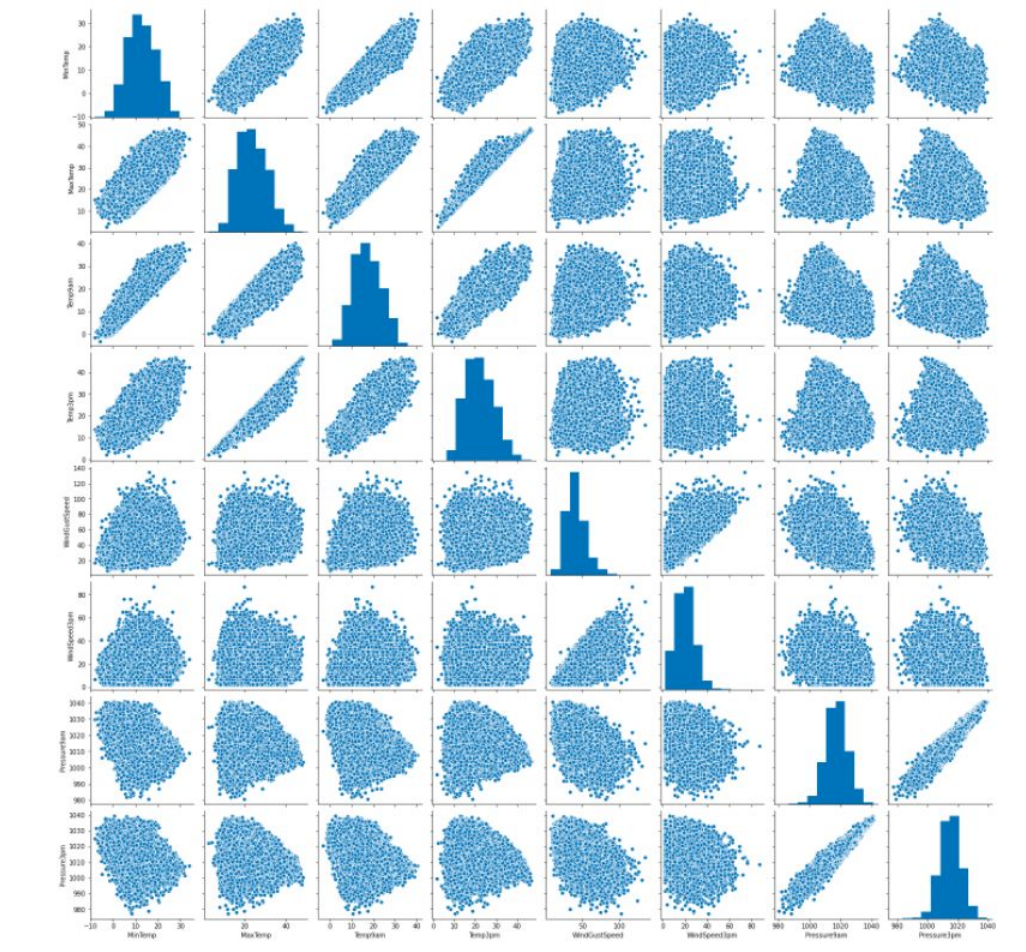
- This dataset contains about 10 years of daily weather observations from numerous Australian weather stations.
- It consists of 24 columns and 142193 rows. It has numerical, categorical and boolean data.
- It is an unbalanced dataset, as the target column RainTomorrow has 110316 "No" entries and 31877 "Yes" entries.

Visualizations

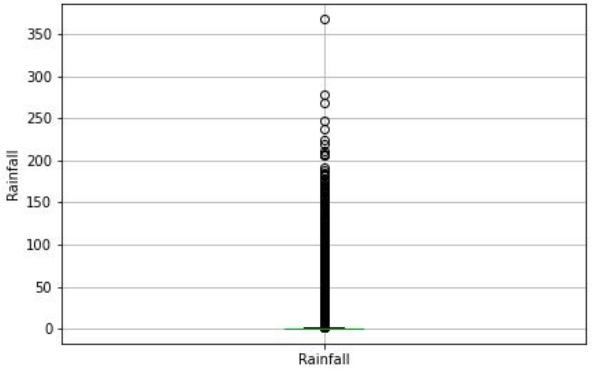


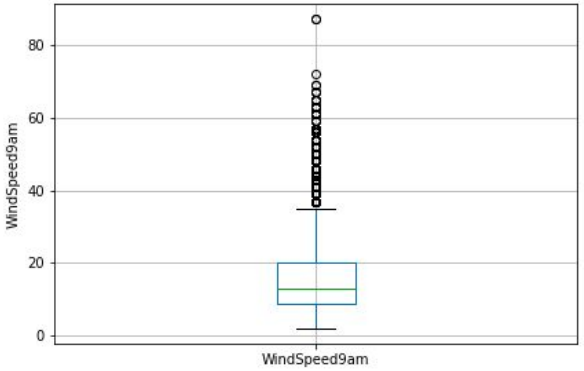
Visualizing the Target Variable

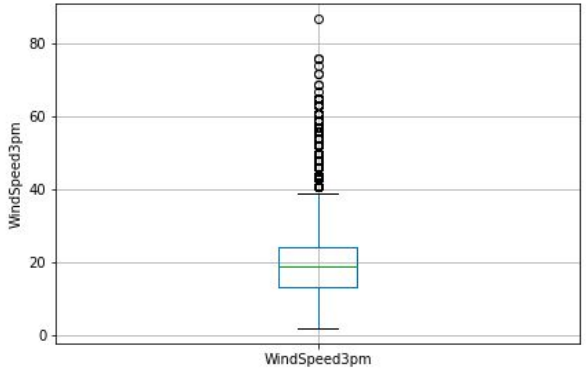

Correlation Heatmap of Rain in Australia Dataset
MinTemp 1.00 0.73 0.10 0.16 0.15 0.15 -0.22 0.03 -0.43 -0.45 0.90 0.70 0.08
MaxTemp 0.73 1.00 -0.08 0.06 -0.01 0.01 -0.52 -0.50 -0.31 -0.41 0.88 0.98 -0.17
Rainfall 0.10 -0.08 1.00 0.13 0.09 0.06 0.24 0.26 -0.17 -0.13 0.01 -0.08 0.24
WindGustSpeed 0.16 0.06 0.13 1.00 0.59 0.68 -0.18 -0.02 -0.45 -0.40 0.12 0.02 0.24
WindSpeed9am 0.15 -0.01 0.09 0.59 1.00 0.50 -0.22 -0.02 -0.20 -0.15 0.09 -0.02 0.09
WindSpeed3pm 0.15 0.01 0.06 0.68 0.50 1.00 -0.09 0.05 -0.28 -0.24 0.13 -0.01 0.10
Humidity9am -0.22 -0.52 0.24 -0.18 -0.22 -0.09 1.00 0.68 0.12 0.17 -0.46 -0.51 0.28
Humidity3pm 0.03 -0.50 0.26 -0.02 -0.02 0.05 0.68 1.00 -0.04 0.05 -0.20 -0.55 0.46
Pressure9am -0.43 -0.31 -0.17 -0.45 -0.20 -0.28 0.12 -0.04 1.00 0.96 -0.40 -0.27 -0.24
Pressure3pm -0.45 -0.41 -0.13 -0.40 -0.15 -0.24 0.17 0.05 0.96 1.00 -0.45 -0.38 -0.22
Temp9am 0.90 0.88 0.01 0.12 0.09 0.13 -0.46 -0.20 -0.40 -0.45 1.00 0.86 -0.04
Temp3pm 0.70 0.98 -0.08 0.02 -0.02 -0.01 -0.51 -0.55 -0.27 -0.38 0.86 1.00 -0.20
RainTomorrow 0.08 -0.17 0.24 0.24 0.09 0.10 0.28 0.46 -0.24 -0.22 -0.04 -0.20 1.00
MinTemp MaxTemp Rainfall WindGustSpeed WindSpeed9am WindSpeed3pm Humidity9am Humidity3pm Pressure9am Pressure3pm Temp9am Temp3pm RainTomorrow
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4


MinTemp
MaxTemp
Temp9am
Temp3pm
WindGustSpeed
WindSpeed3pm
Pressure9am
Pressure3pm

```
Text(0, 0.5, 'WindSpeed3pm')
```







Checking for Outliers

Data Preprocessing

- The dataset is stored in BigQuery.
- The preprocessing is done using Jupyter Notebooks in AI Platform
- The datatypes of the columns are corrected.
- The columns that have less than 60% of data are dropped.
- The rows that consist of missing values in any of the columns are dropped.
- Outliers are handled using top-coding approach.
- Categorical data is encoded using One Hot Encoding as Logistic Regression cannot handle categorical data.

Machine Learning Model

- A Logistic Regression model is trained using BigQuery ML.
- The query used to create the model is :

```
CREATE OR REPLACE MODEL
 `thematic-flash-266714.australia_weather.model`
OPTIONS
 ( model_type="logistic_reg",
  input_label_cols=["RainTomorrow"] ) AS
SELECT * FROM
  `thematic-flash-266714.australia_weather.preprocessed`
```

Aggregate metrics

Log loss	0.4087
ROC AUC	0.8432

Score threshold

Positive class threshold 0.5063

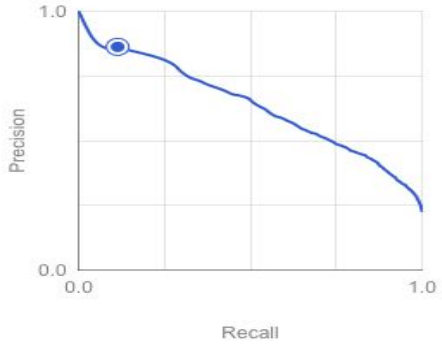
Positive class	true
Negative class	false
Precision	0.8118
Recall	0.2501
Accuracy	0.8156
F1 score	0.3824

Use this slider above to see which score threshold works best for your model.

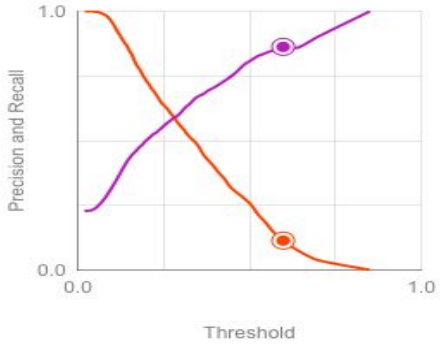
Confusion matrix

Actual labels \ Predicted labels	true	false
true	25.01%	74.99%
false	1.71%	98.29%

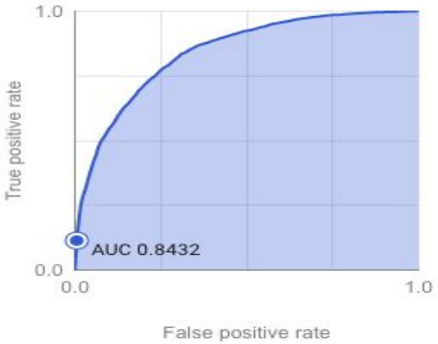
Precision-Recall curve



Precision and Recall vs Threshold



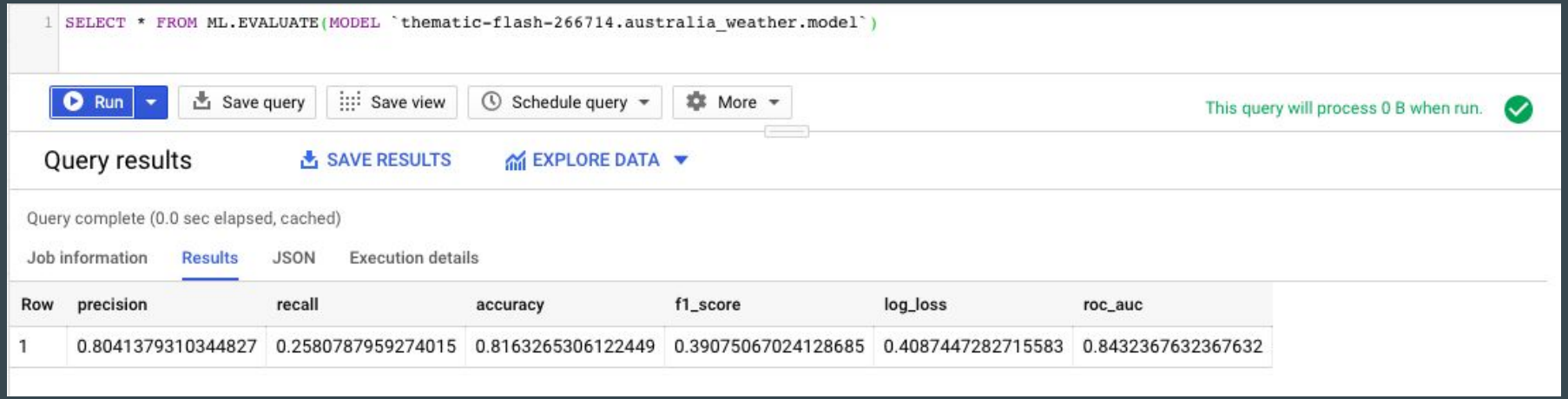
ROC curve



Evaluating the Model

- We use the following query to evaluate the model

SELECT * FROM ML.EVALUATE(MODEL `thematic-flash-266714.australia_weather.model`)



Row	precision	recall	accuracy	f1_score	log_loss	roc_auc
1	0.8041379310344827	0.2580787959274015	0.8163265306122449	0.39075067024128685	0.4087447282715583	0.8432367632367632

Conclusion

- The accuracy of the model is 0.82 and precision is 0.80 which is good for a classification model.
- But the recall and f1-score is low, which implies that the model needs improvement.
- The performance of the model suffers due to the unbalanced dataset.

THANK YOU!