

Predicting Rainfall in Australia

...

PROJECT GROUP #3

Hanumasri Bollepalli (801104071)

Madhuri Pawle (801083244)

Pavanitha Jampala (801131462)

Contents

- Introduction
- About the Dataset
- Visualizations
- Data Preprocessing
- Machine Learning Model
- Predictions using the Model
- Evaluating the Model
- Conclusion

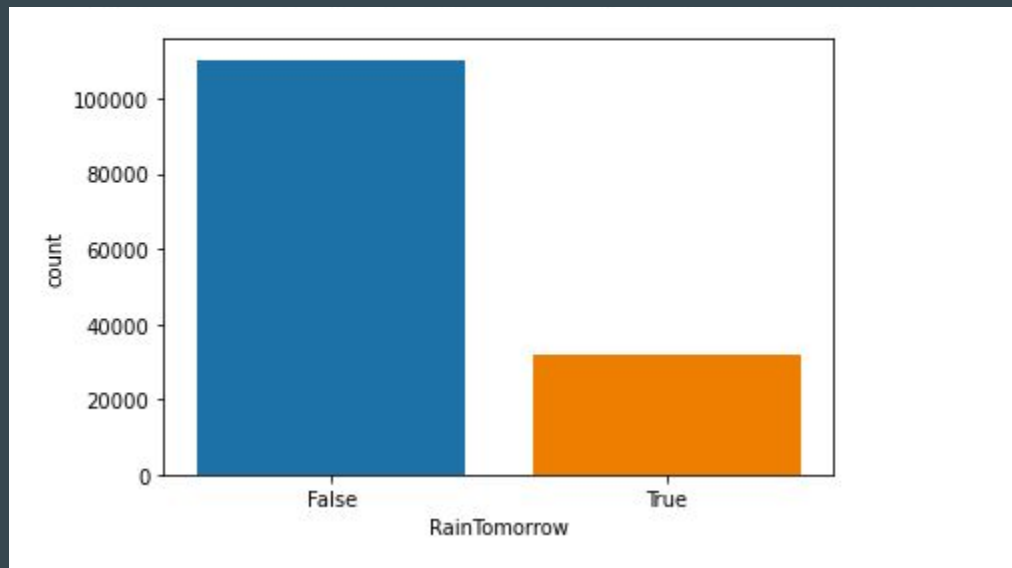
Introduction

- 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.

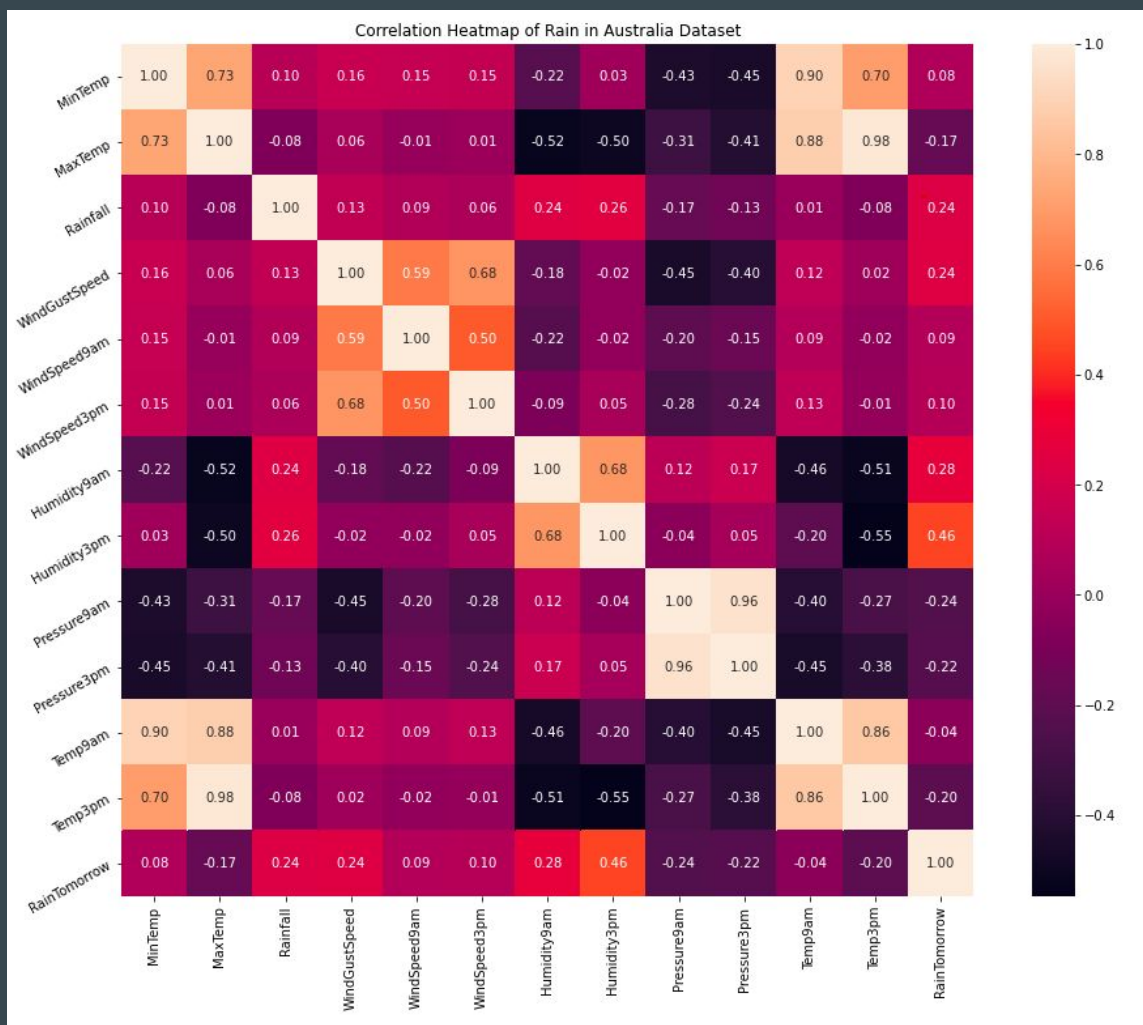
About the Dataset

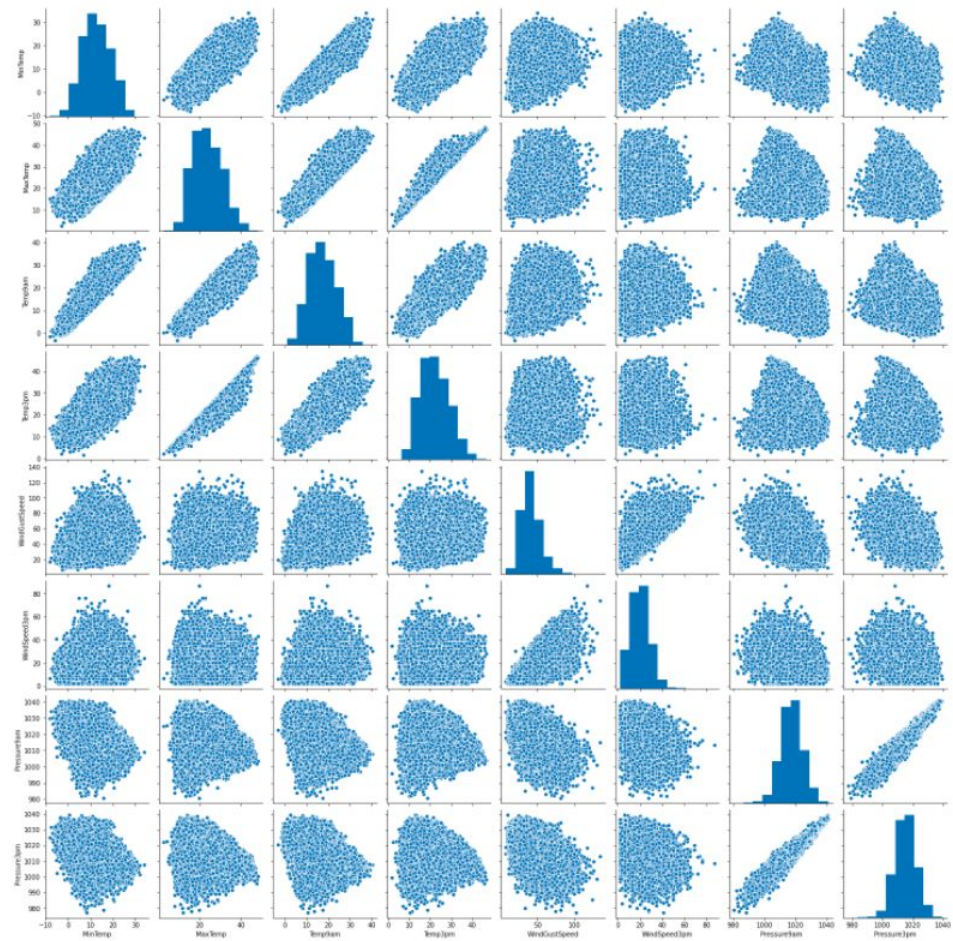
- This dataset contains about 10 years of daily weather observations from numerous Australian weather stations.
- It consists of 24 columns and 142193 rows. It has numerical, categorical and boolean data.
- It is an unbalanced dataset, as the target column RainTomorrow has 110316 "No" entries and 31877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

Visualizations

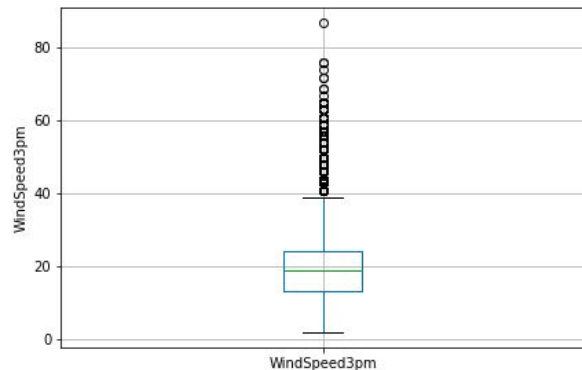
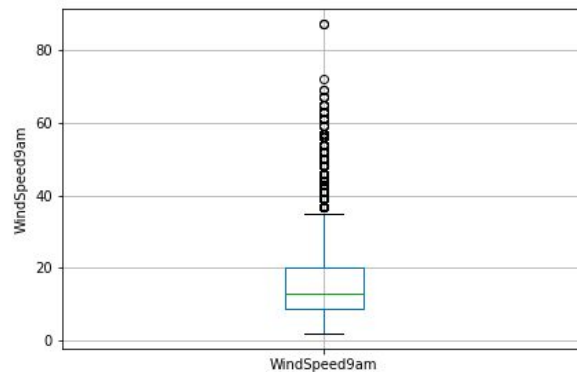
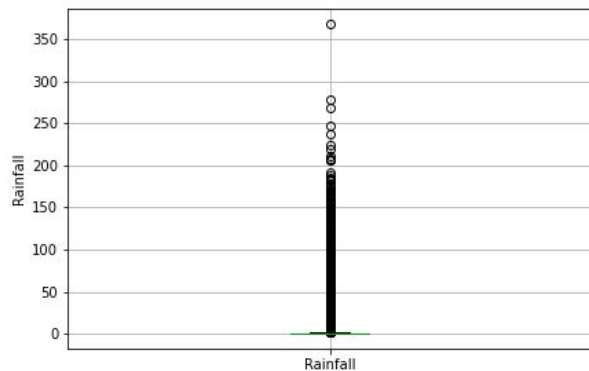


Visualizing the Target Variable





```
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.log_reg_model`
```

```
OPTIONS
```

```
( model_type="logistic_reg",  
  input_label_cols=["RainTomorrow"] ) AS
```

```
SELECT * EXCEPT(YEAR) FROM
```

```
`thematic-flash-266714.australia_weather.preprocessed_new` WHERE YEAR >=2011
```

Aggregate metrics ?

Log loss ? 0.3638

ROC AUC ? 0.8663

Score threshold

Positive class threshold ?  0.5542

Positive class true

Negative class false

Precision 0.7913

Recall 0.3882

Accuracy ? 0.8400

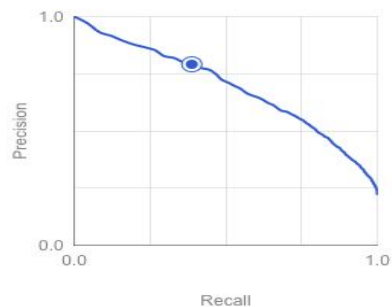
F1 score ? 0.5208

Confusion matrix

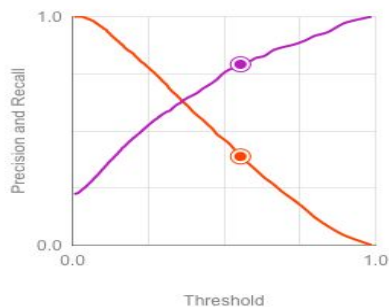
Actual labels	Predicted labels	
	true	false
true	38.82%	61.18%
false	2.96%	97.04%

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

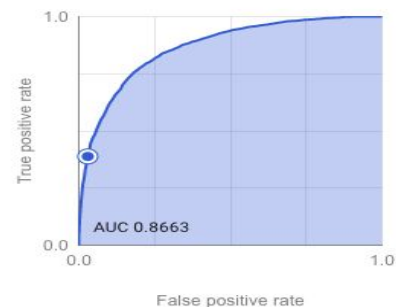
Precision-Recall curve



Precision and Recall vs Threshold



ROC curve



Predictions Using the Model

- We use the following query to evaluate the model

```
SELECT * FROM ML.PREDICT (MODEL `thematic-flash-266714.australia_weather.log_reg_model`,  
  
SELECT * EXCEPT(YEAR) FROM  
  
`thematic-flash-266714.australia_weather.preprocessed_new`  
  
WHERE YEAR < 2011 ))
```

```

1 SELECT * FROM
2 ML.PREDICT(MODEL `thematic-flash-266714.australia_weather.log_reg_model`,
3   (SELECT * EXCEPT(YEAR) FROM
4     `thematic-flash-266714.australia_weather.preprocessed_new` WHERE YEAR < 201

```



Run



Save query



Save view



Schedule query



More

Query results

SAVE RESULTS

EXPLORE DATA

Query complete (3.4 sec elapsed, 92.3 MB processed)

Job information

Results

JSON

Execution details

Row	predicted_RainTomorrow	predicted_RainTomorrow_probs.label	predicted_RainTomorrow_probs.prob
1	false	true	0.12808793434238738
		false	0.8719120656576126
2	true	true	0.5297993431692399
		false	0.47020065683076007
3	false	true	0.48164343216357963
		false	0.5183565678364204
4	false	true	0.2119057304407178
		false	0.7880942695592822
5	false	true	0.369439295879749
		false	0.630560704120251
6	false	true	0.37659448765017495
		false	0.623405512349825

Evaluating the Model

- We use the following query to evaluate the model

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

1 <code>SELECT * FROM ML.EVALUATE (MODEL `thematic-flash-266714.australia_weather.log_reg_model`)</code>						
<div><div>Run</div><div>Save query</div><div>Save view</div><div>Schedule query</div><div>More</div></div>						
<div>Query results</div> <div>SAVE RESULTSEXPLORE DATA</div>						
Query complete (0.2 sec elapsed, 0 B processed)						
Job informationResultsJSONExecution details						
Row	precision	recall	accuracy	f1_score	log_loss	roc_auc
1	0.7544738725841088	0.4612691466083151	0.8456711442298265	0.5725149375339489	0.36384063927514204	0.8663436563436564

Conclusion

- The accuracy of the model is 0.85 and precision is 0.75 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!