Predicting Rainfall in Australia

...

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

Contents

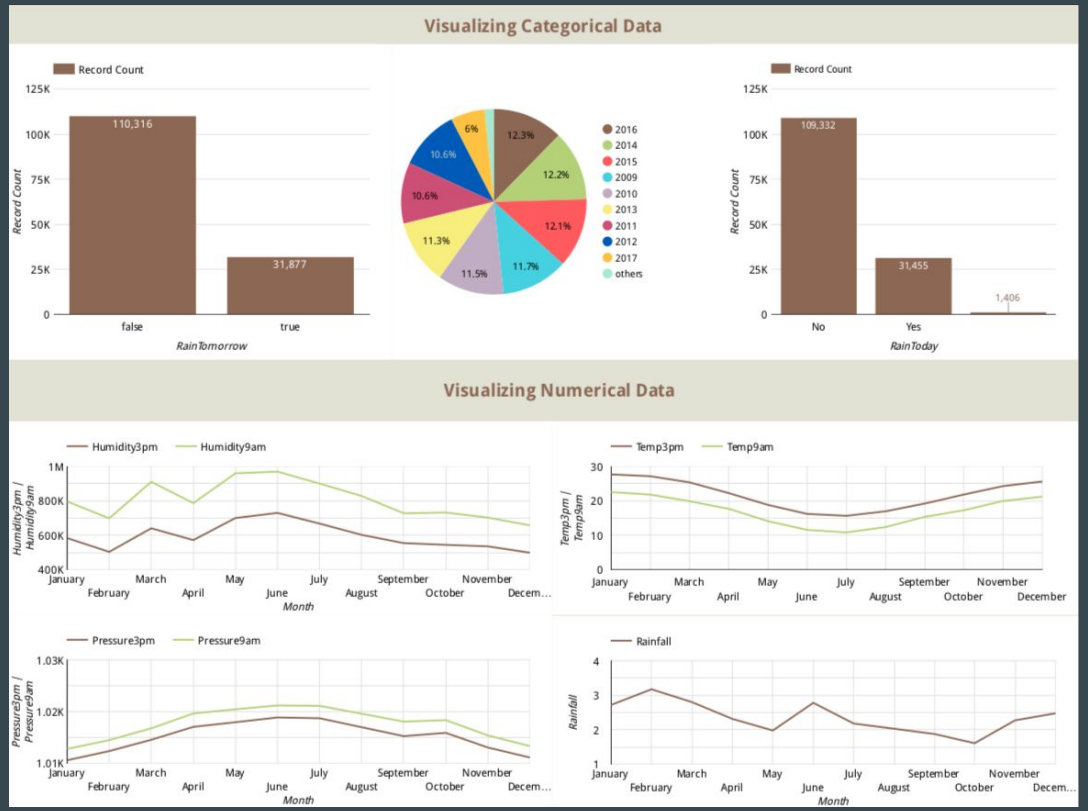
Introduction

- 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.

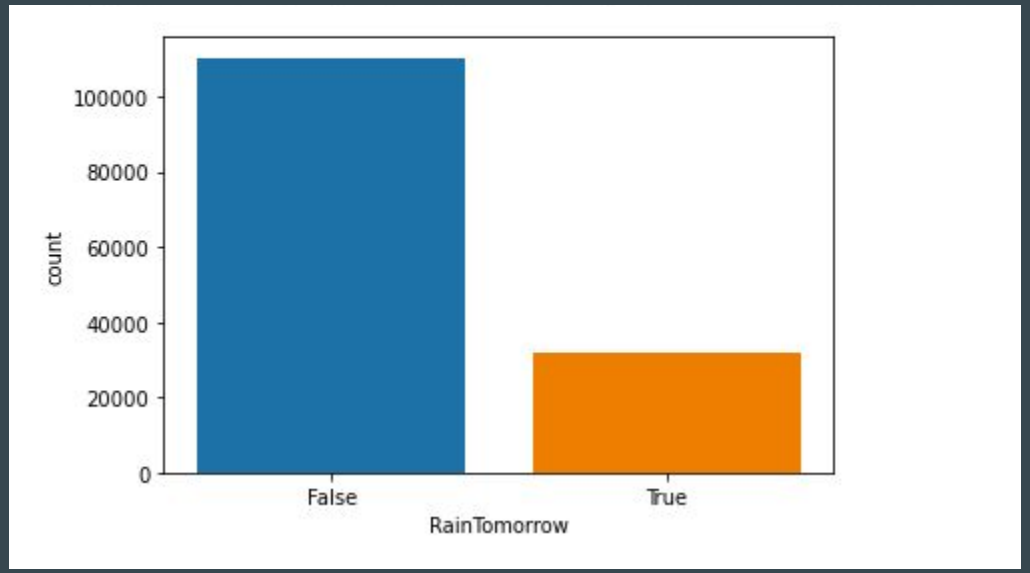
About the Dataset

- 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.

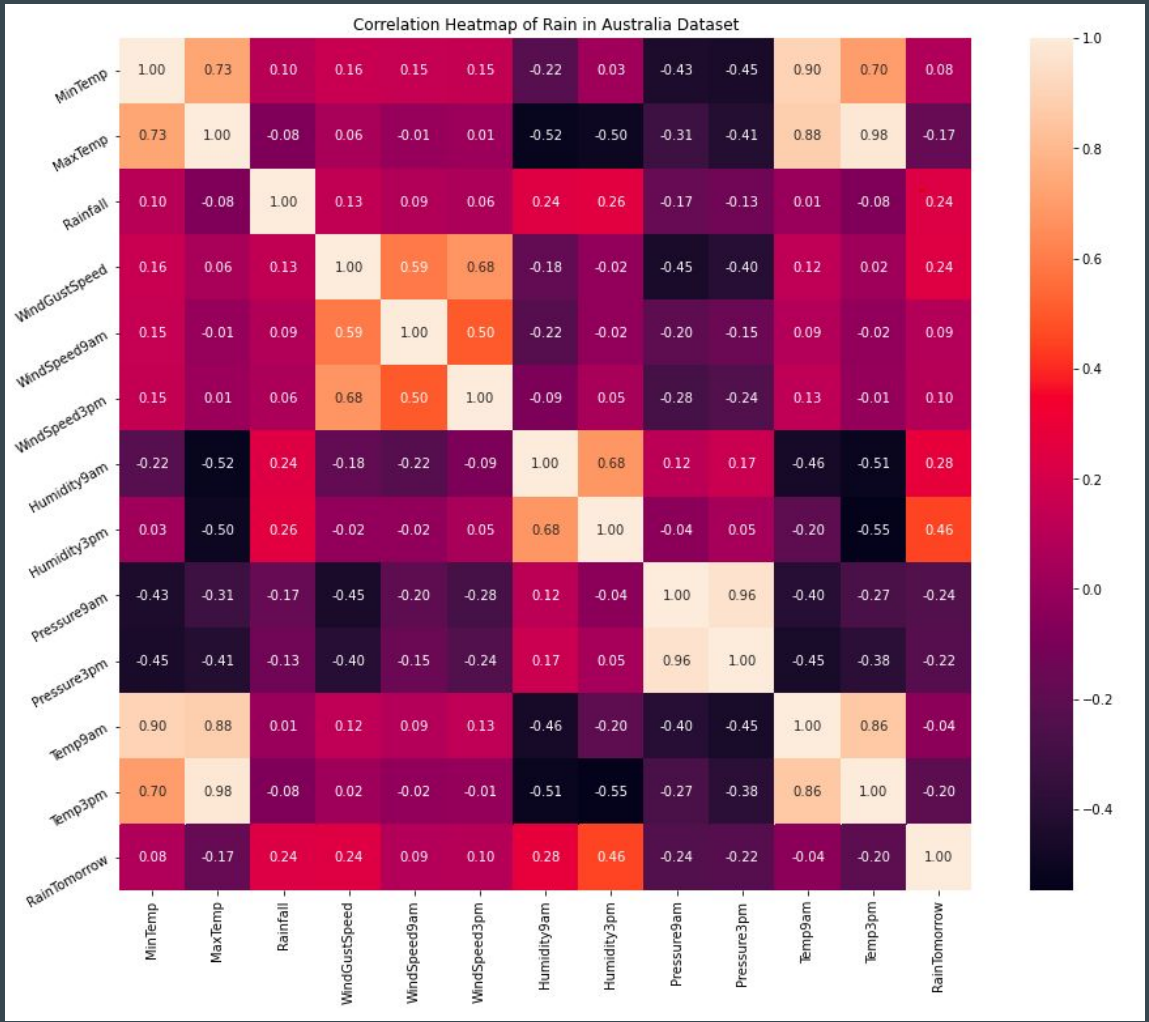
Dashboard

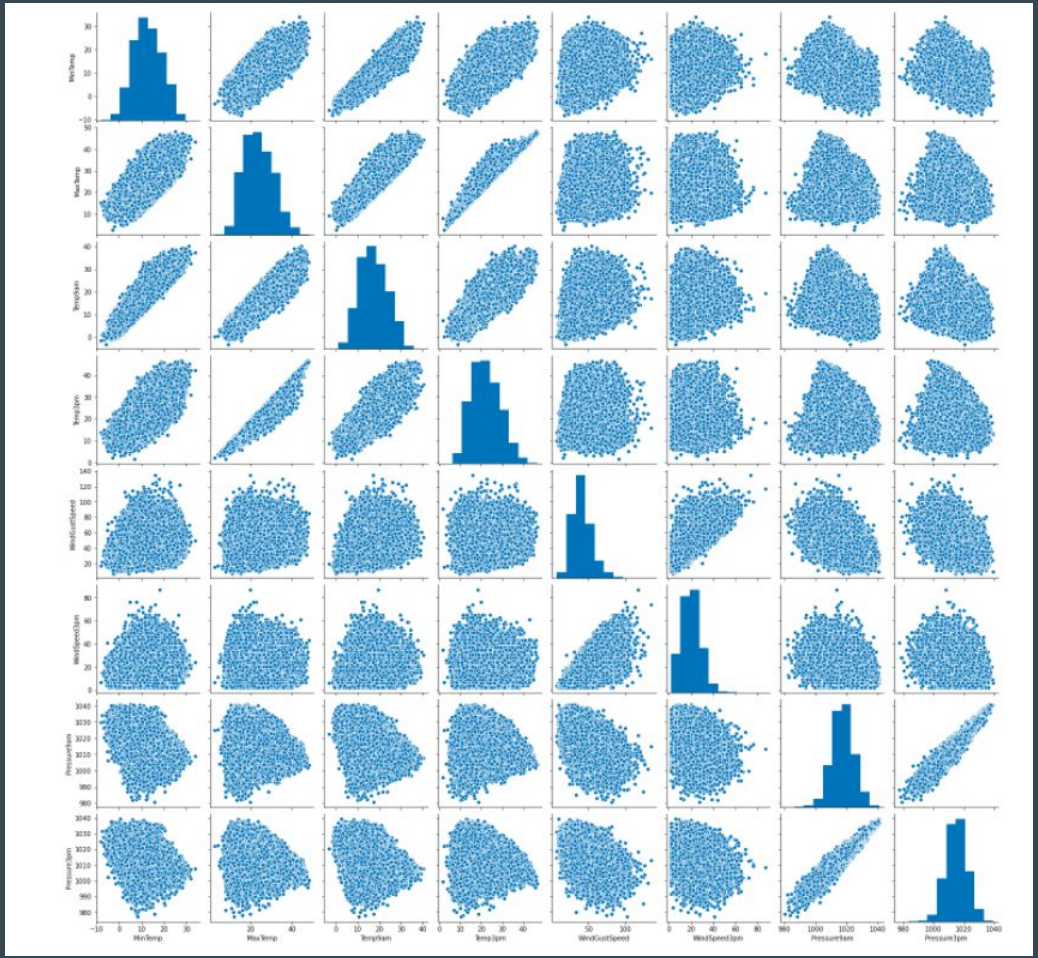

Visualizing Categorical Data
Record Count
125K
100K
75K
50K
25K
0
Record Count
110,316
31,877
false
true
RainTomorrow
6%
12.3%
10.6%
12.2%
10.6%
12.1%
11.3%
11.7%
11.5%
2016
2014
2015
2009
2010
2013
2011
2012
2017
others
Record Count
125K
100K
75K
50K
25K
0
Record Count
109,332
31,455
1,406
No
Yes
RainToday
Visualizing Numerical Data
Humidity3pm
Humidity9am
1M
800K
600K
400K
Humidity3pm / Humidity9am
January
February
March
April
May
June
July
August
September
October
November
Decem...
Month
Temp3pm
Temp9am
30
20
10
0
Temp3pm / Temp9am
January
February
March
April
May
June
July
August
September
October
November
December
Pressure3pm
Pressure9am
1.03K
1.02K
1.01K
Pressure3pm / Pressure9am
January
February
March
April
May
June
July
August
September
October
November
Decem...
Month
Rainfall
4
3
2
1
Rainfall
January
February
March
April
May
June
July
August
September
October
November
Decem...
Month

Visualizations

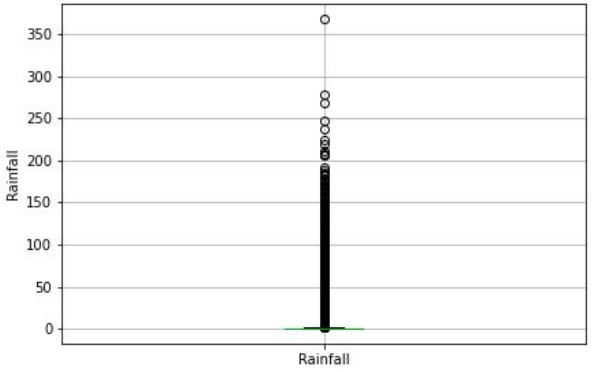


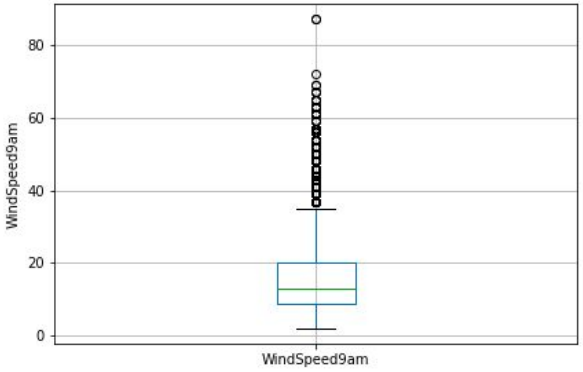
Visualizing the Target Variable

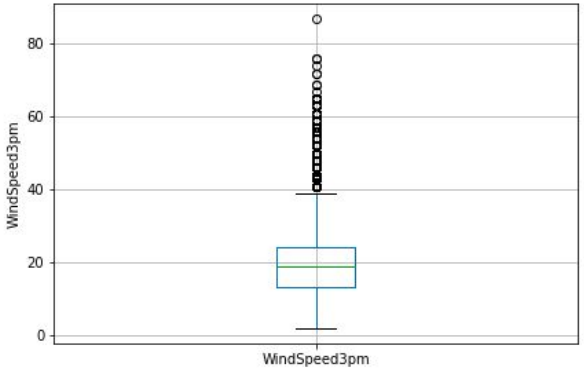

Correlation Heatmap of Rain in Australia Dataset
MinTemp 1.00 0.73 0.10 0.16 0.15 0.15 -0.22 0.03 -0.43 -0.45 0.90 0.70 0.08
MaxTemp 0.73 1.00 -0.08 0.06 -0.01 0.01 -0.52 -0.50 -0.31 -0.41 0.88 0.98 -0.17
Rainfall 0.10 -0.08 1.00 0.13 0.09 0.06 0.24 0.26 -0.17 -0.13 0.01 -0.08 0.24
WindGustSpeed 0.16 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.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

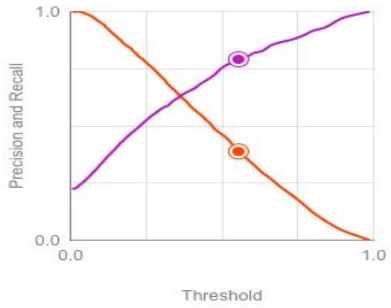
Use this slider above to see which score threshold works best for your model.

Confusion matrix

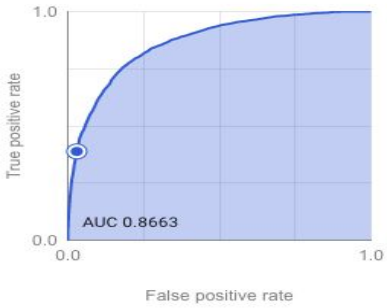
Actual labels \ Predicted labels	true	false
true	38.82%	61.18%
false	2.96%	97.04%

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 ))
```

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

Run | Save query | Save view | Schedule query | More

Query results | SAVE RESULTS | EXPLORE DATA

Query complete (0.2 sec elapsed, 0 B processed)

Job information | Results | JSON | Execution 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!