Predicting Rain in Australia

Mohammed Abduldaim

Background

Data provided by Australian Bureau of Meteorology and obtained from Kaggle, this dataset contains about 10 years of daily weather observations from many locations across Australia.

The goal is to explore the data and predict if tomorrow there will be rain or not.

Dataset

The data contains over 145,000 row and 23 columns.

Some of the notable columns are:

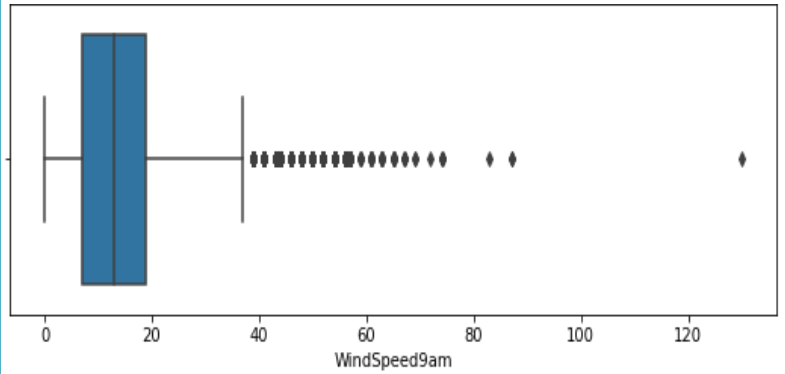
- Date
- Location
- Raint Today
- Rain Tomorrow

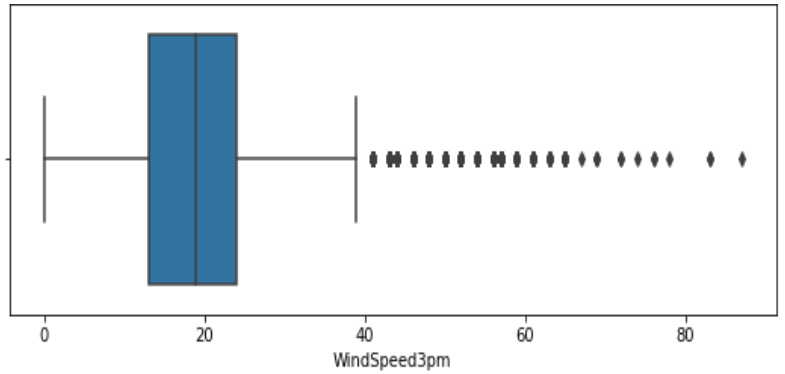
The rest of the columns describe state of weather each day, such as temperature humidity, sunshine, etc. Which should be very useful in building a model

Outliers

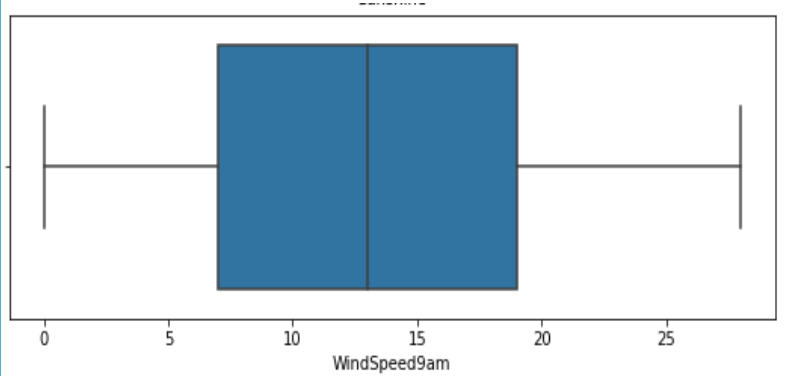
Some of the outliers before and after

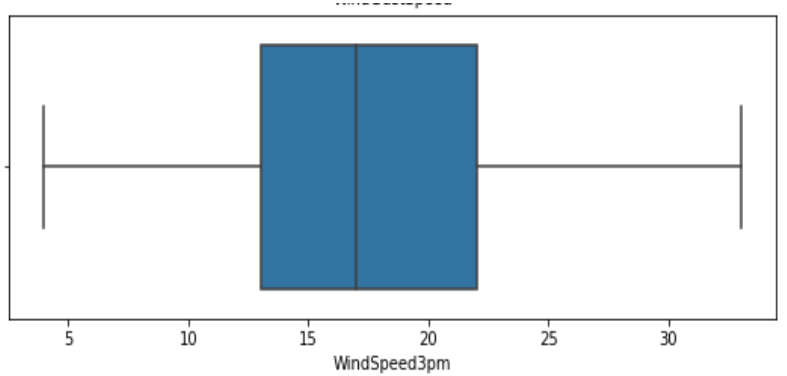
Before





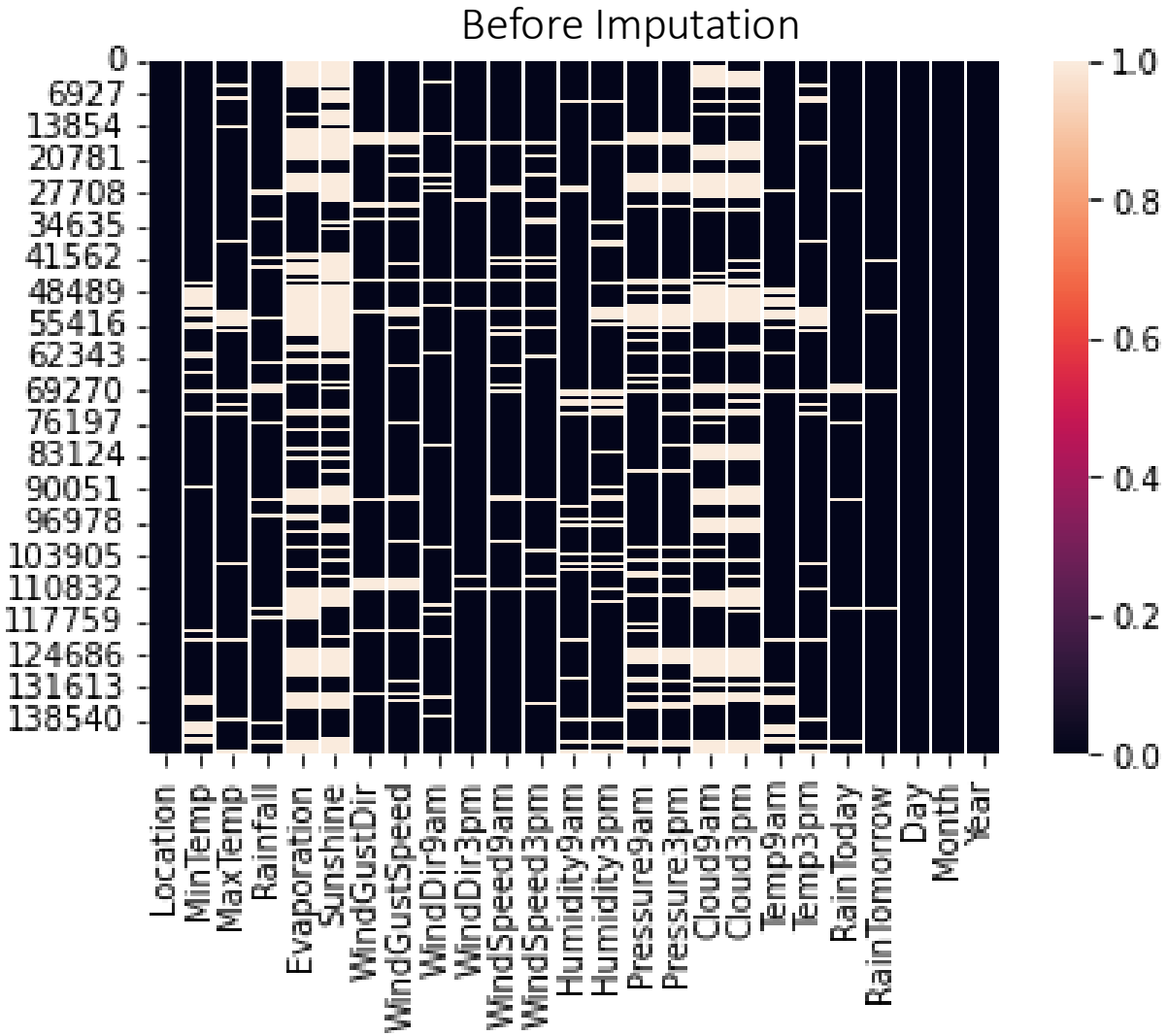
After



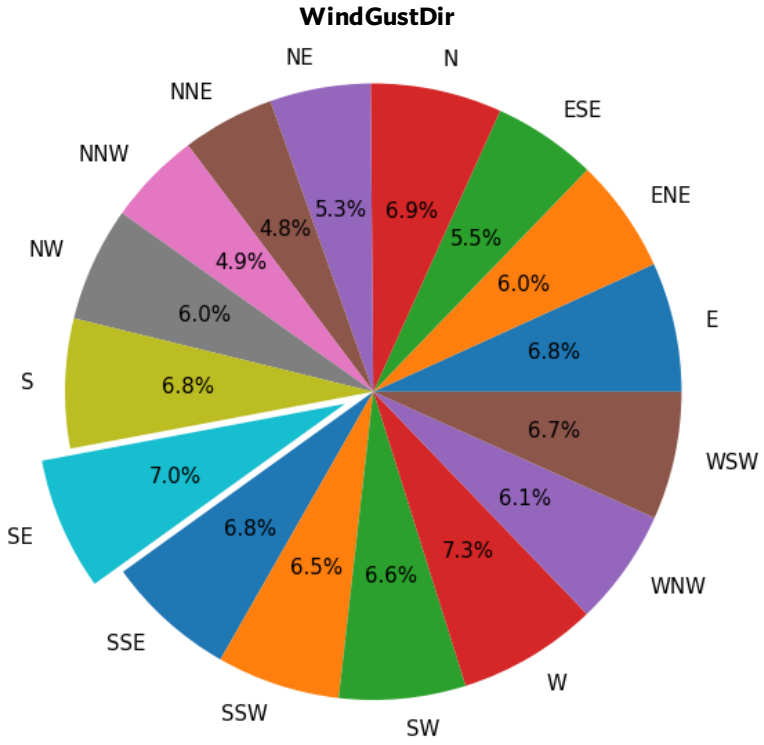


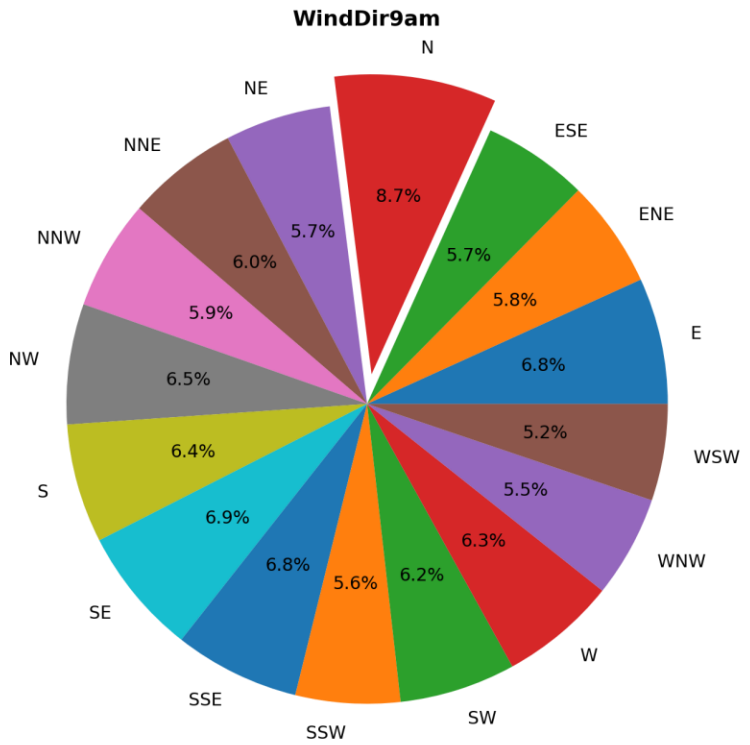
Missing Values

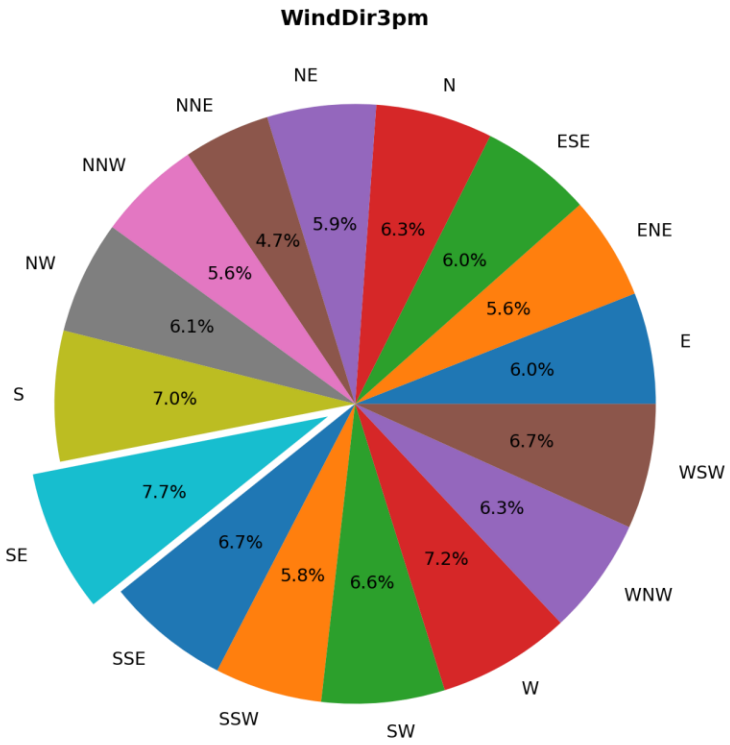
Sunshine has the highest count of missing values out of the numerical columns And WindDir9am out of the categorical columns



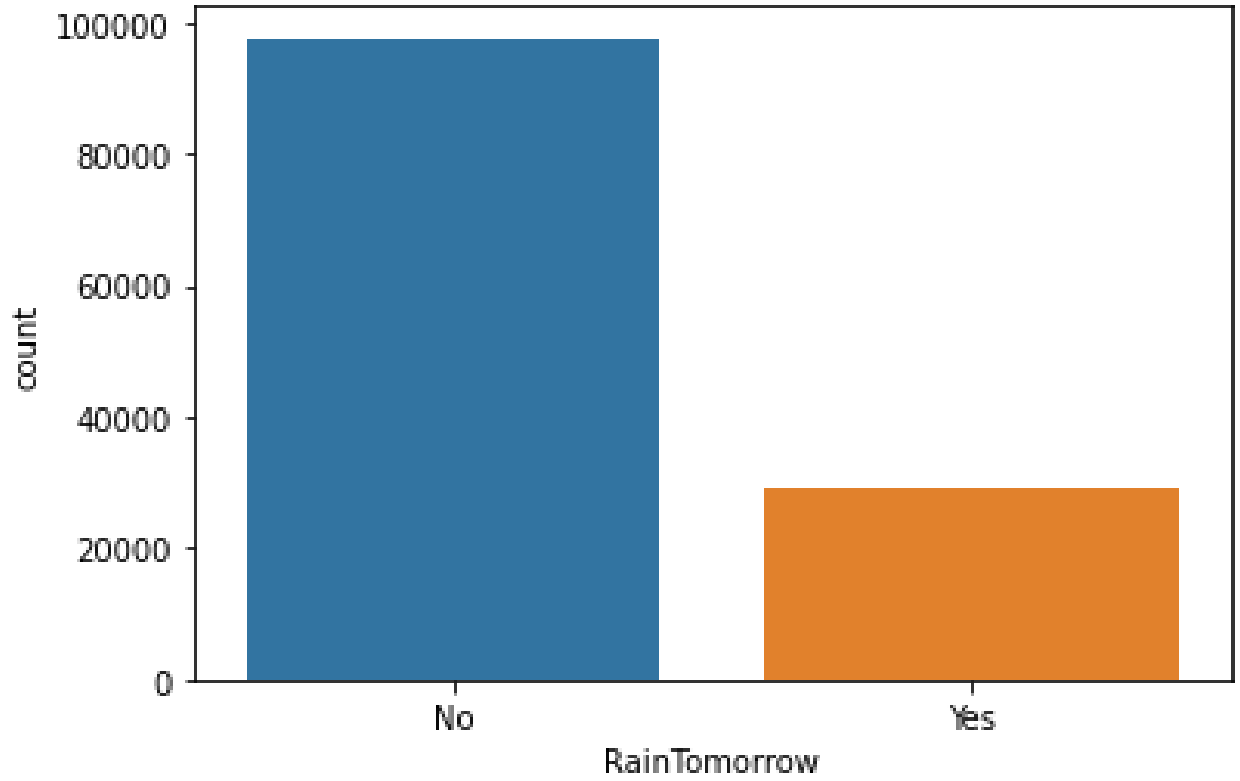
Categorical Columns







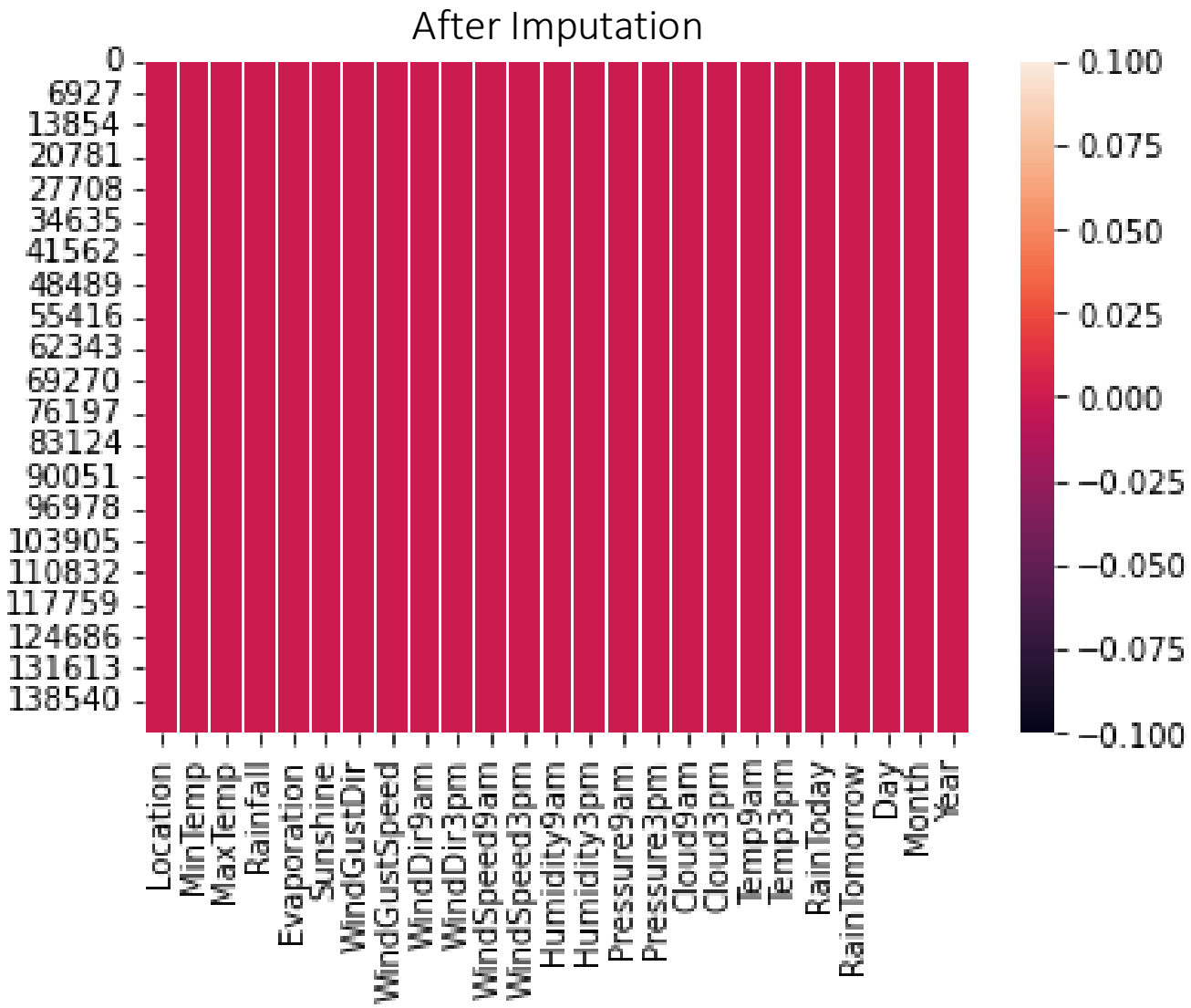
Target



Filling the missing values with the highest outcome

Missing Values

After imputation of the categorical columns with highest occurring value and numerical columns with the mean value



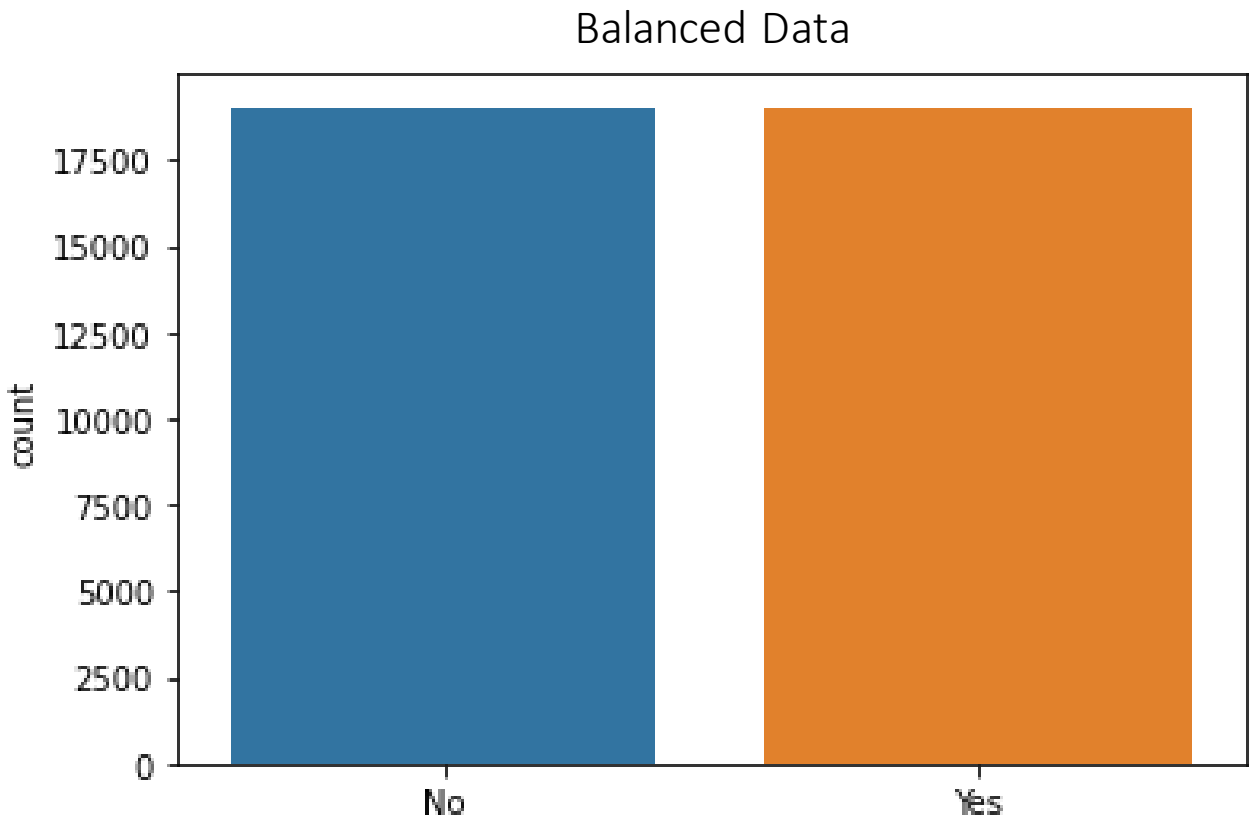
Metrics	Train Scores	Validation Scores	Test Scores
Accuracy	81.4	81.3	81
F1	44.5	45	43.9
Precision	64	64	63
Recall	34	34.8	33.7
Fbeta of 2	37.6	38.3	36

Results

Result of the logistic regression model with imbalanced data

Balancing

Balancing data by under sampling



Metrics	Train Scores	Validation Scores	Test Scores
Accuracy	74	73.7	73
F1	74.1	56	56
Precision	73.9	44.4	44
Recall	74.3	75.8	75.5
Fbeta of 2	74.3	66	64

Results

Result of the logistic regression model with balanced data (undersampled)

Thank You