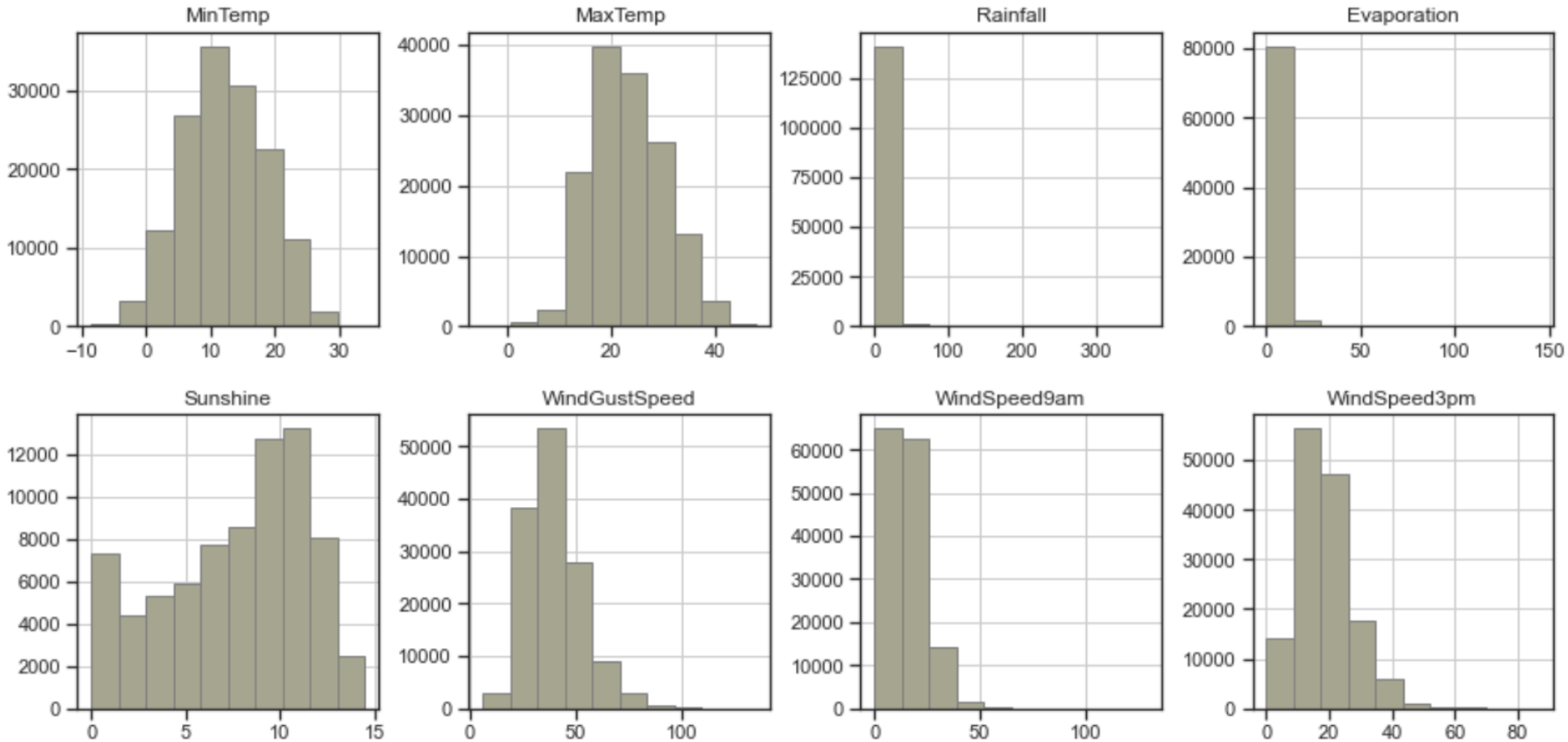
Rainfall Prediction

DATA	145460 rows and 23 features
OBJECTIVE	Build a machine learning model to predict if there will be rain the next day using weather conditions of the previous day
BUSINESS CASE	Reduce incorrect prediction of False Negatives for Rain Tomorrow

FEATURES OF DATASET

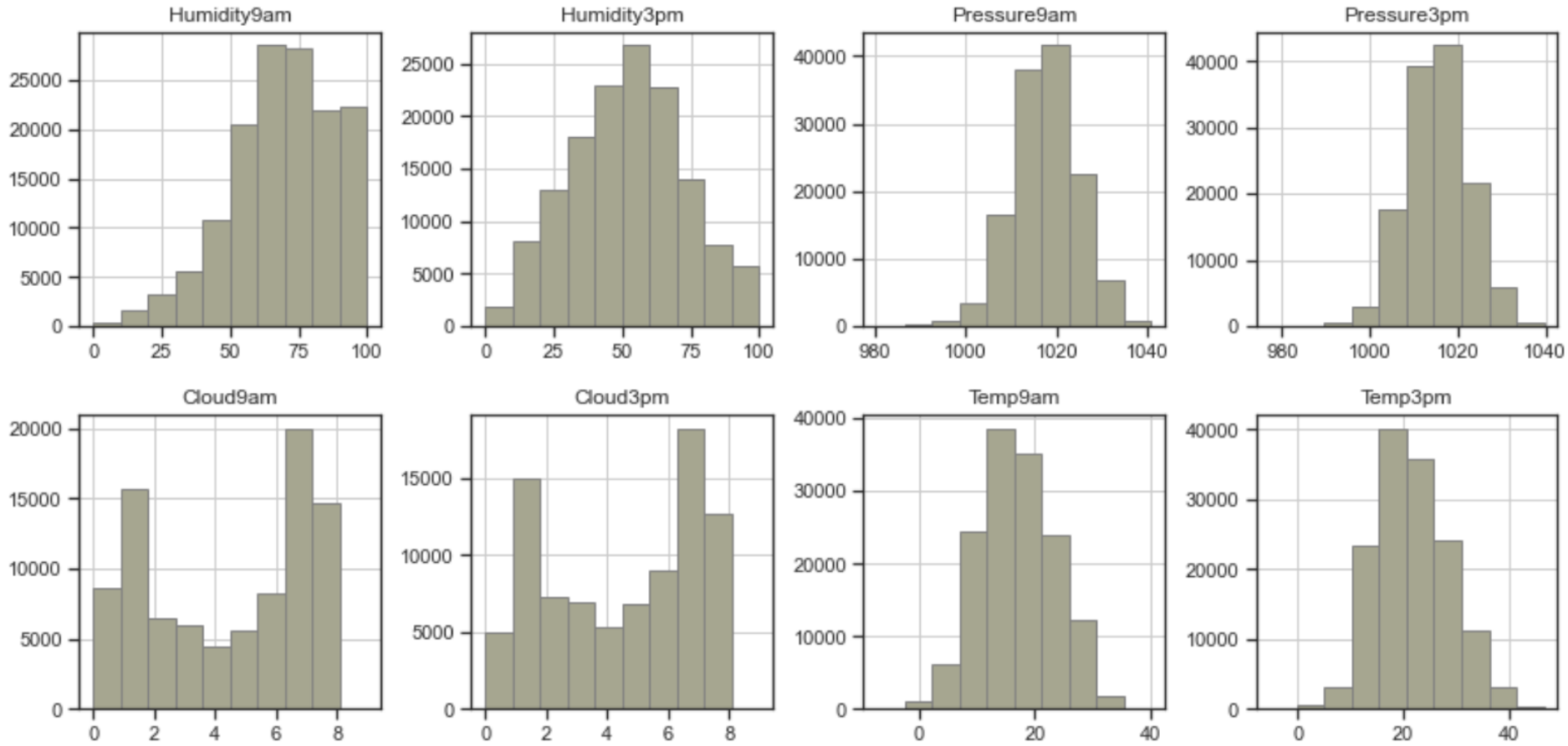
- Date
- Location
- Daily Minimum Temperature
- Daily Maximum Temperature
- Daily Rainfall
- Daily Evaporation Rate
- Daily Amount of Sunshine
- Daily Wind Gust Direction
- Daily Wind Speed
- Daily Morning Wind Direction
- Daily Afternoon Wind Direction
- Daily Morning Wind Speed
- Daily Afternoon Wind Speed
- Daily Morning Humidity
- Daily Afternoon Humidity
- Daily Morning Atmospheric
- Pressure
- Daily Afternoon Atmospheric
- Pressure
- Daily Morning Cloud Cover
- Daily Afternoon Cloud Cover
- Daily Morning Temperature
- Daily Afternoon Temperature
- Rain Today

EXPLORATORY DATA ANALYSIS – Numerical Features

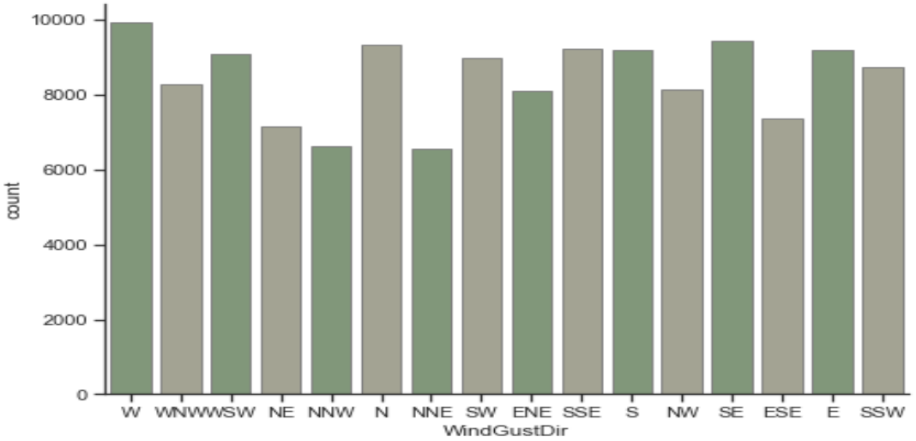


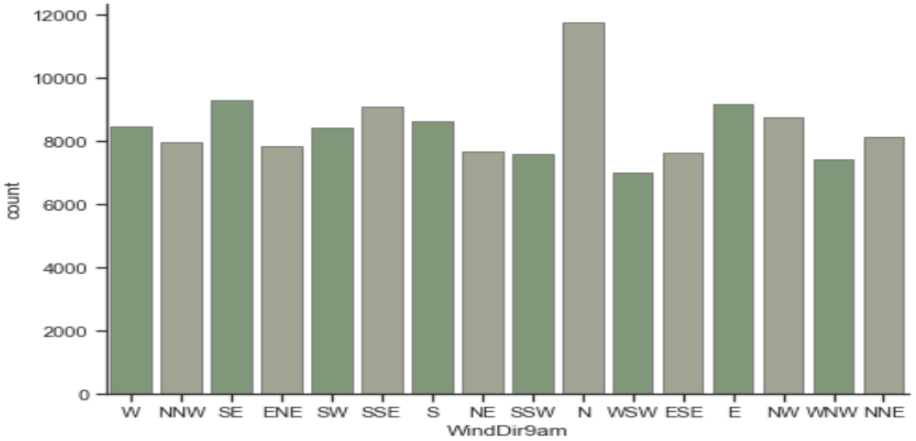
Skewness and Outliers observed

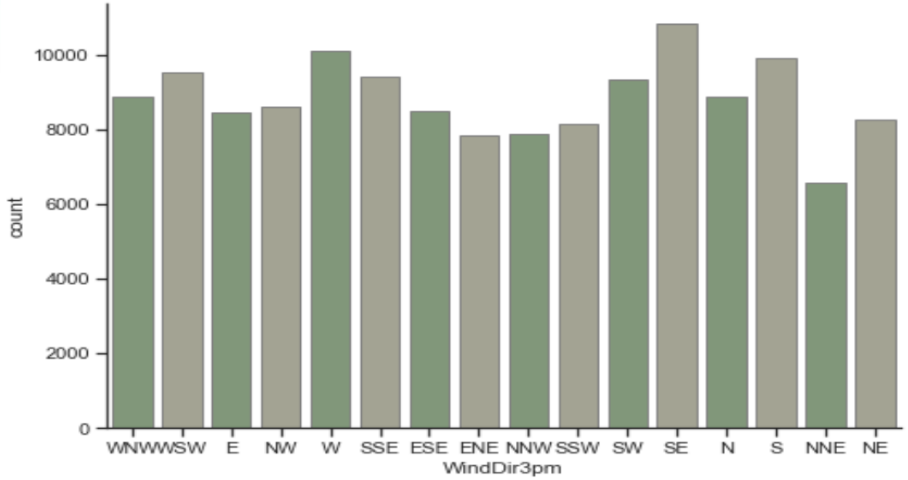
EXPLORATORY DATA ANALYSIS – Numerical Features

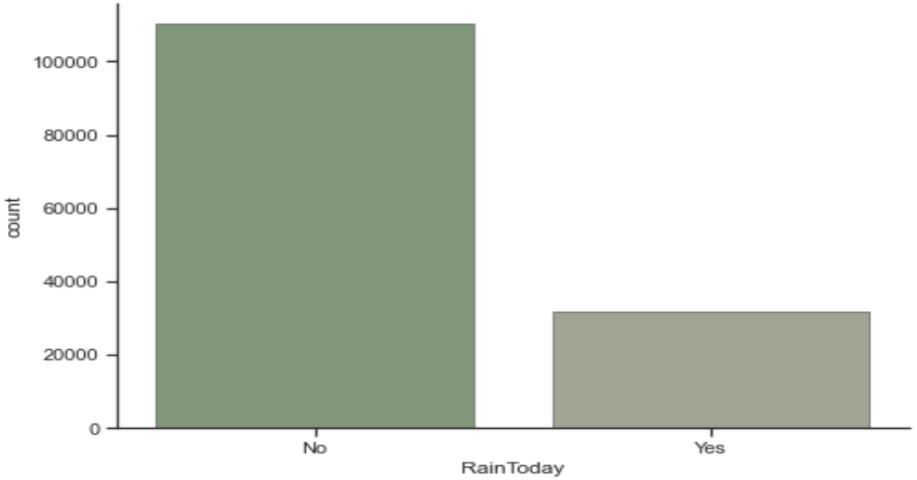


EXPLORATORY DATA ANALYSIS – Categorical Features

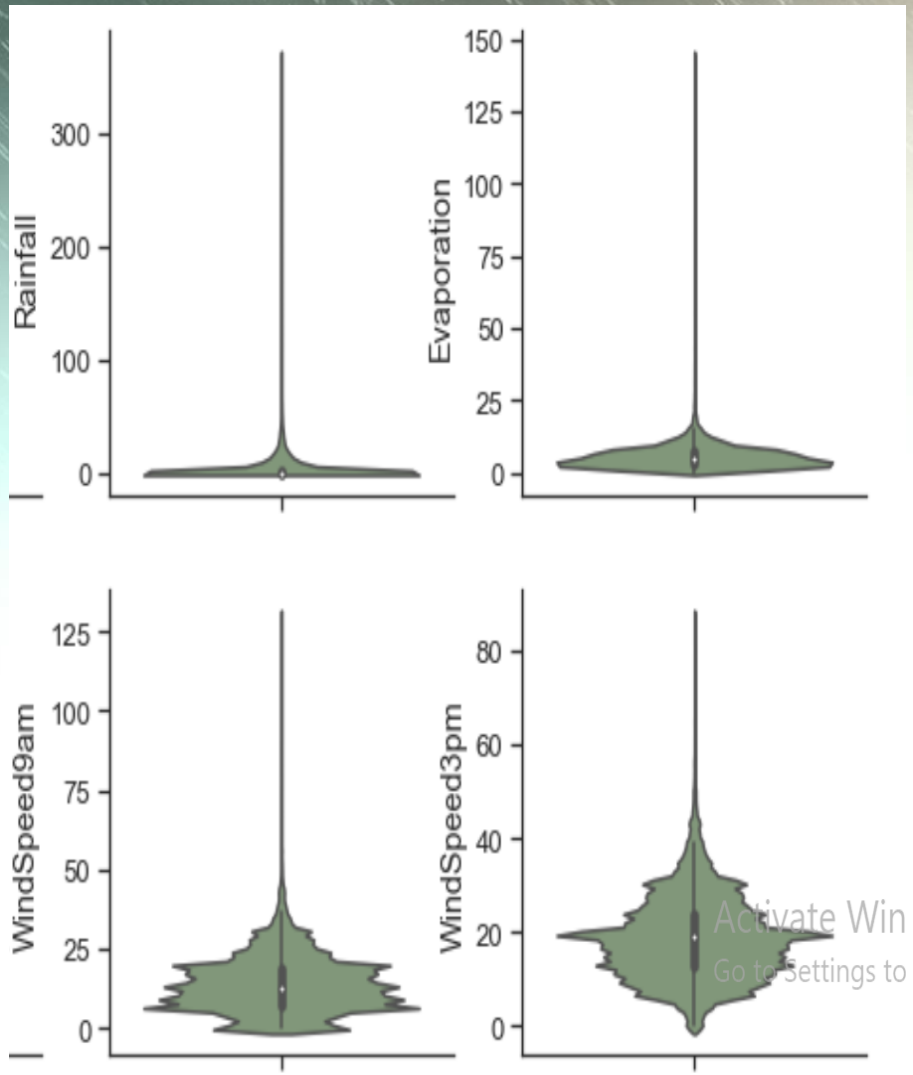


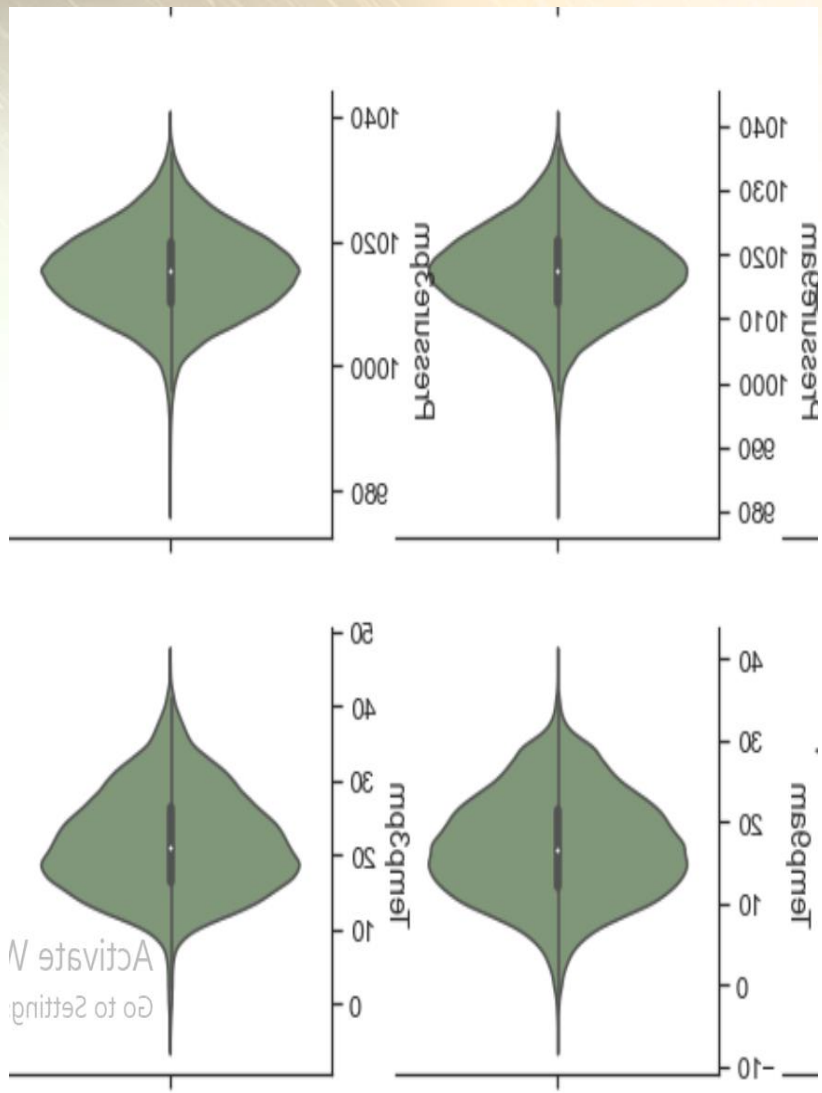




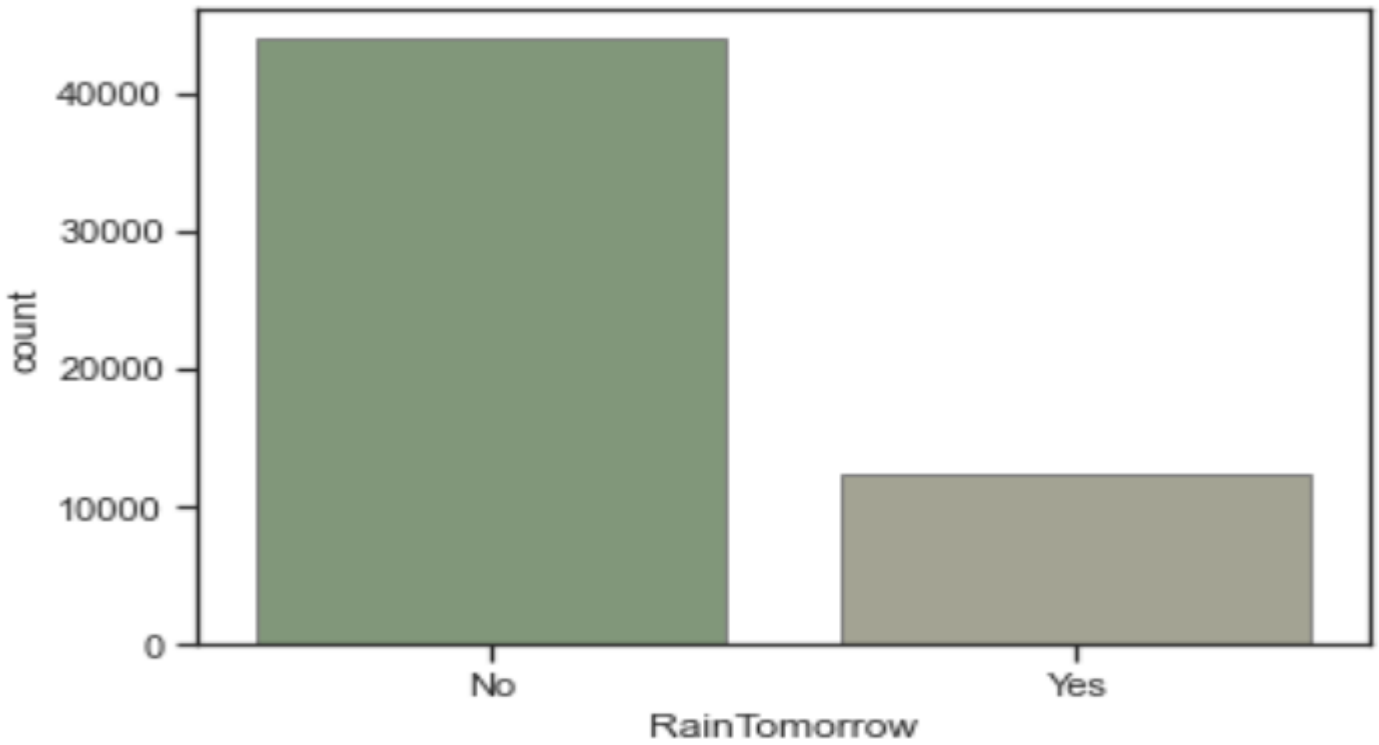


EXPLORATORY DATA ANALYSIS – Outliers





EXPLORATORY DATA ANALYSIS



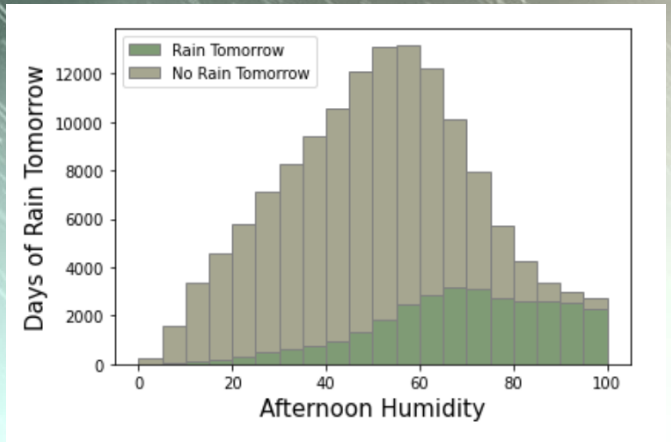
No Rain Tomorrow :

43993 rows 78%

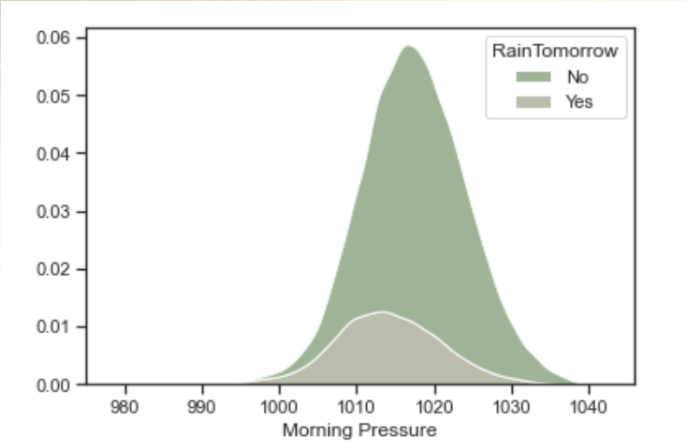
Rain Tomorrow :

12427 rows 22%

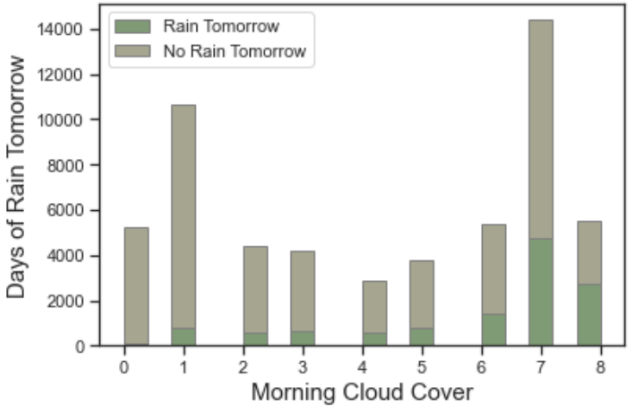
EXPLORATORY DATA ANALYSIS



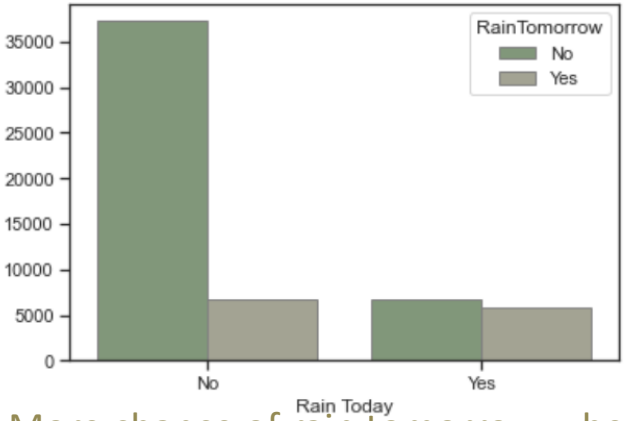
More chance of rain tomorrow when humidity is high today



More chance of rain tomorrow when atmospheric pressure is ~1015 today

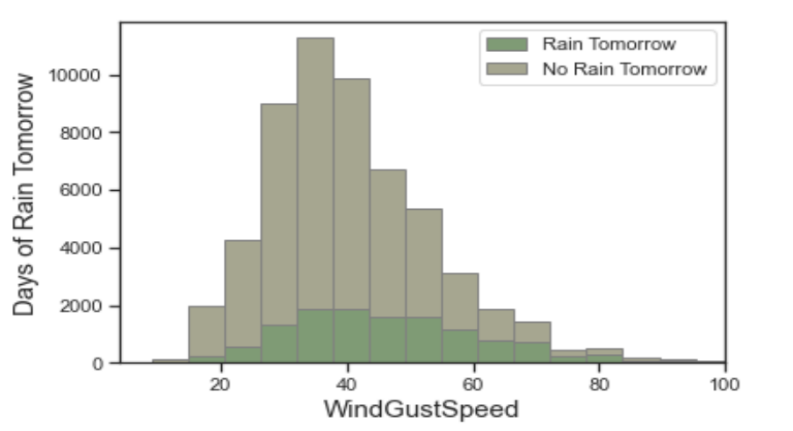


More chance of rain tomorrow when cloud cover is 7 or 8 oktas today

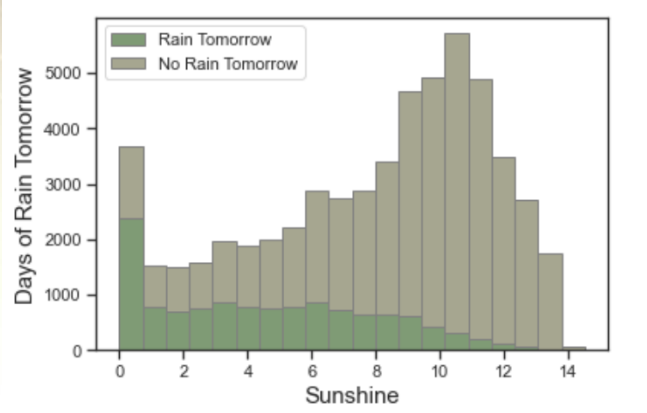


More chance of rain tomorrow when it has rained today

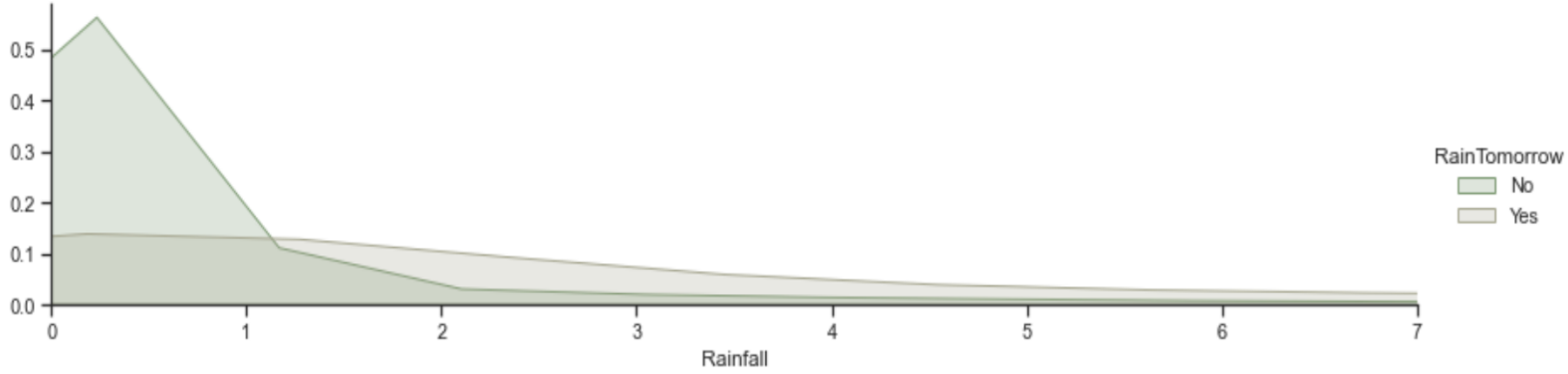
EXPLORATORY DATA ANALYSIS



More chance of rain tomorrow when wind speed is higher today



More chance of rain tomorrow when there is little sunshine today



More chance of rain tomorrow when it has rained today and the rainfall today is higher

DATA PREPROCESSING

01 Feature Transformation

- Scaling all Numerical Columns using RobustScaler

02 Feature Extraction

- Label Encoding of Target
- One Hot Encoding of Categorical Features

03 Feature Engineering

- Added a month column from the Date column as there are certain months that have more rain than others

04 Feature Selection

- Removed 2 features due to eliminate multicollinearity

BASELINE MODEL

Logistic Regression

Accuracy	: 85%
F1 Score	: 61%
Precision (Major Class)	:87%
Precision (Minor Class)	:72 %
Recall (Major Class)	:94%
Recall (Minor Class)	:52%
Training Accuracy	:85%
Test Accuracy	:85%

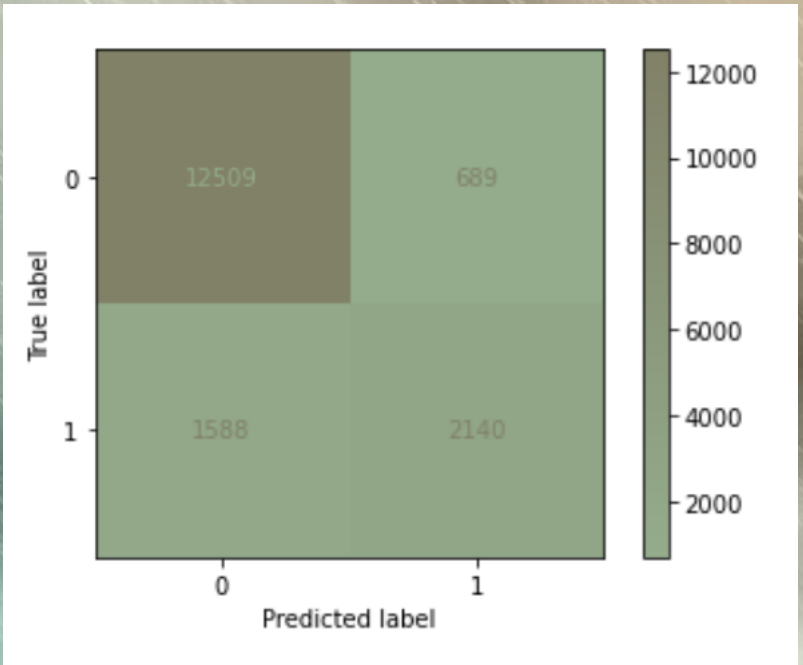
MULTIPLE MODELS

	Logistic	KNN	SVM	RF	XGBoost	MLP
Accuracy	: 85 %	83%	83%	86%	86%	82%
F1 Score	: 61 %	58%	43%	62%	65%	60%
Precision (Major Class)	:89 %	87%	83%	87%	89%	89%
Precision (Minor Class)	:76 %	65%	86%	77%	74%	67%
Recall (Major Class)	:95 %	92%	98%	95%	94%	89%
Recall (Minor Class)	:57 %	53%	28%	51%	58%	61%

XGBOOST

Metrics

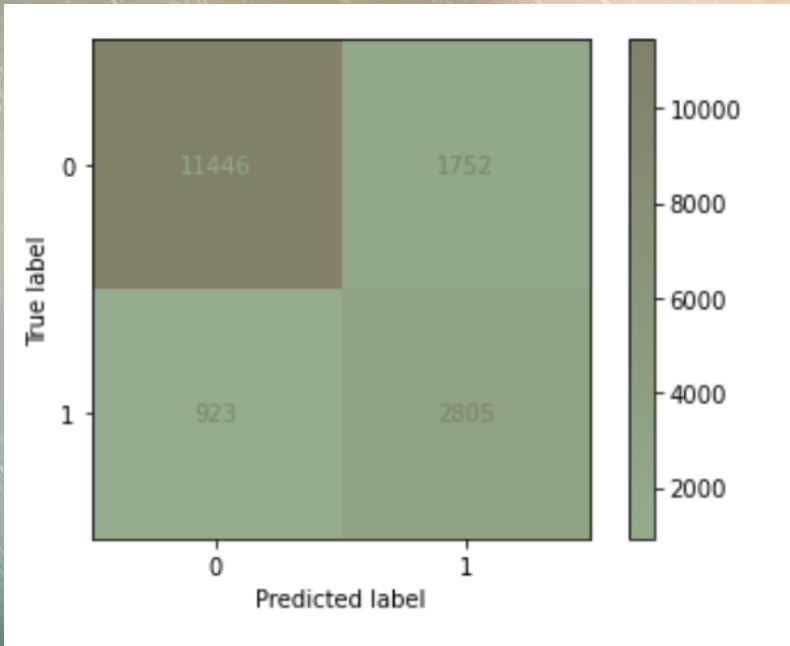
Accuracy	: 85%
F1 Score	: 61%
Precision (Major Class)	:89%
Precision (Minor Class)	:76 %
Recall (Major Class)	:95%
Recall (Minor Class)	:57%
Training Accuracy	:92%
Test Accuracy	:87%



HYPERPARAMETER TUNING FOR XGBOOST

Metrics

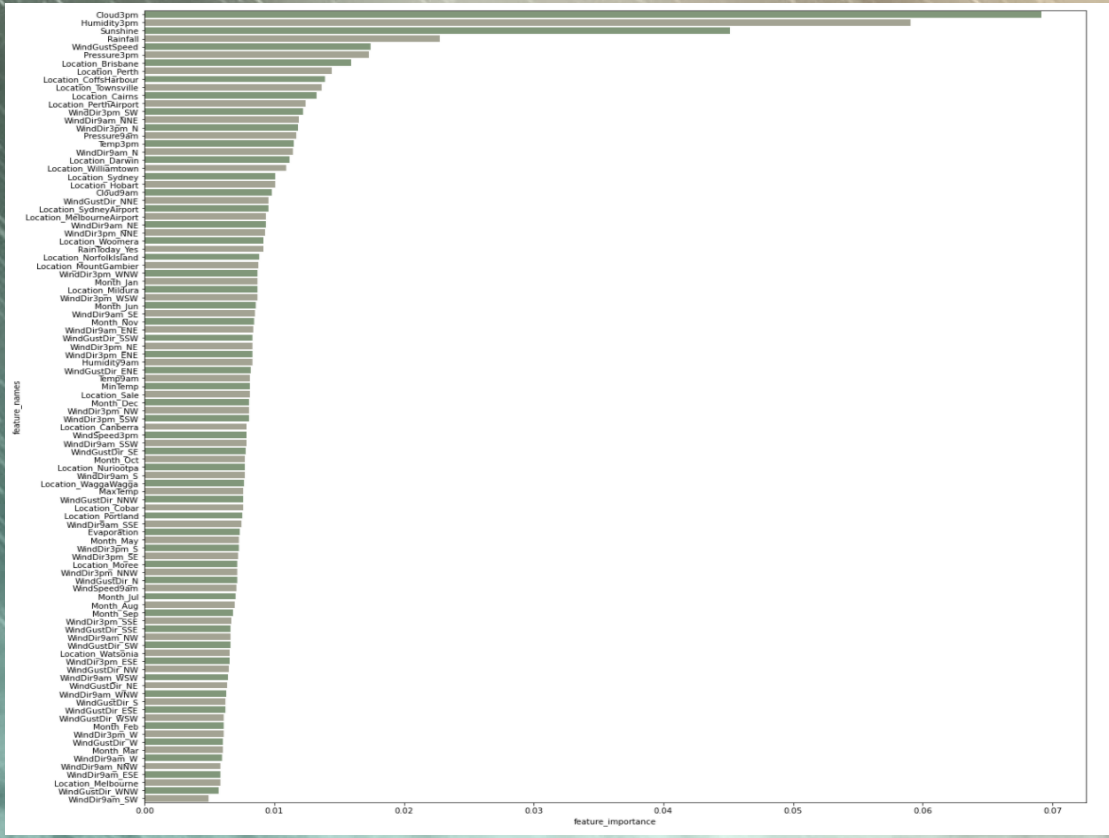
Accuracy : 85%
F1 Score : 62%
Precision :94%
(Major Class)
Precision :62 % ↓ 14%
(Minor Class)
Recall :75%
(Major Class)
Recall :62% ↑ 5%
(Minor Class)
Training Accuracy :92%
Test Accuracy :87%



learning_rate : 0.1
max_depth : 8
reg_lambda : 1.0
gamma : 1.5
scale_pos_weight : 3

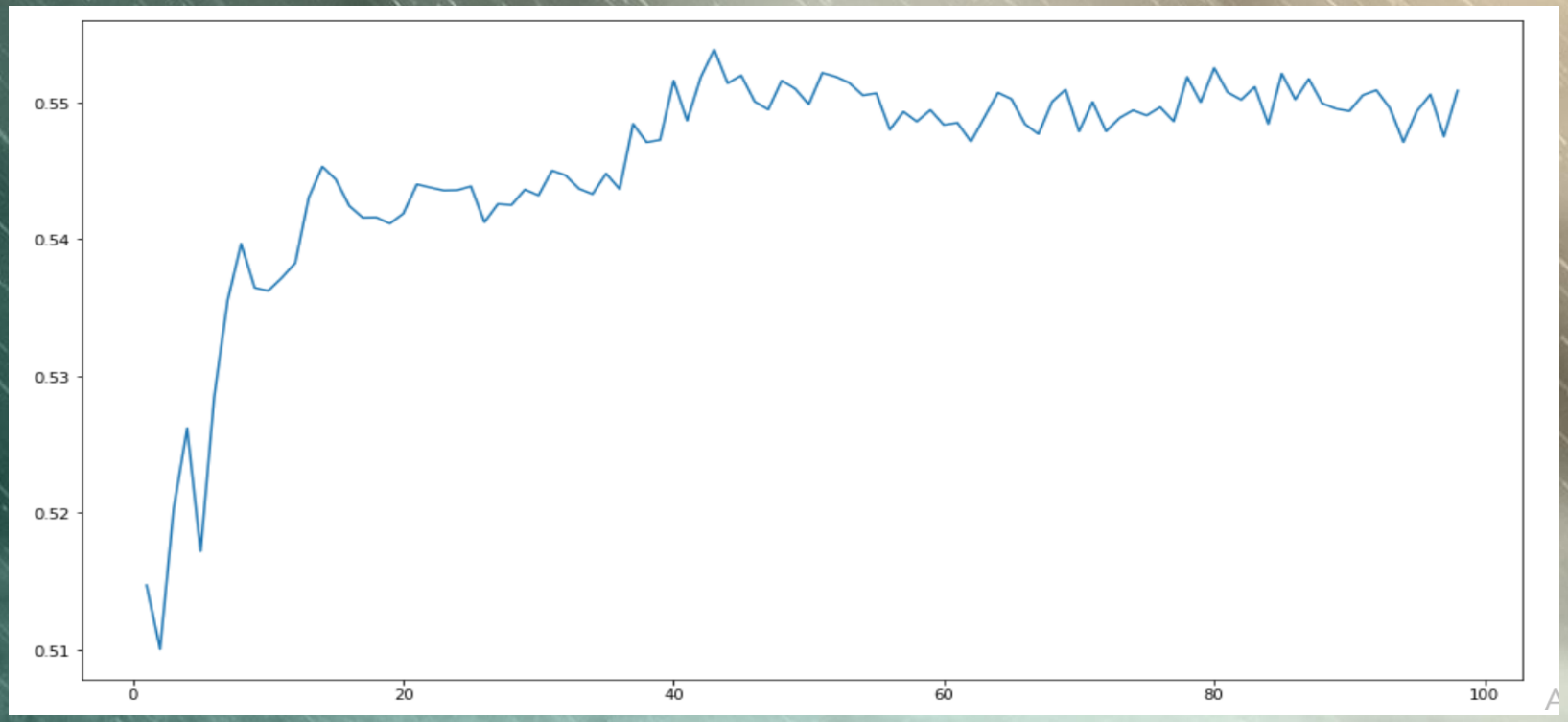
FEATURE SELECTION

Feature Importance



FEATURE SELECTION

Recursive Feature Elimination with Cross Validation

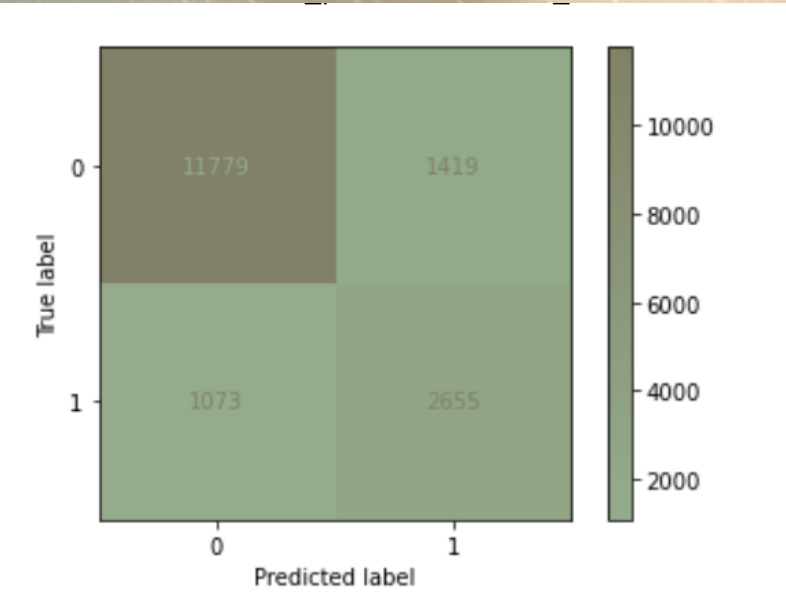


rfecv.n_features_ = 43

XGBOOST WITH FEATURE SELECTION

Metrics

Accuracy	: 85%	
F1 Score	: 68%	
Precision (Major Class)	:92%	
Precision (Minor Class)	:65%	↑ 3%
Recall (Major Class)	:89%	
Recall (Minor Class)	:71%	↑ 9%
Training Accuracy	:96%	
Test Accuracy	:85%	



learning_rate : 0.1
max_depth : 11
reg_lambda : 5.0
gamma : 0.5
scale_pos_weight : 3

HANDLING IMBALANCED DATASET - SMOTE

SMOTE Strategy – Over and Under Sampling

Oversample Minor Class and Undersample Major Class

- Increase Minor Class to 70% of Major Class
- Decrease Major Class to 10 % more than Minor Class

Major Class : Minor Class

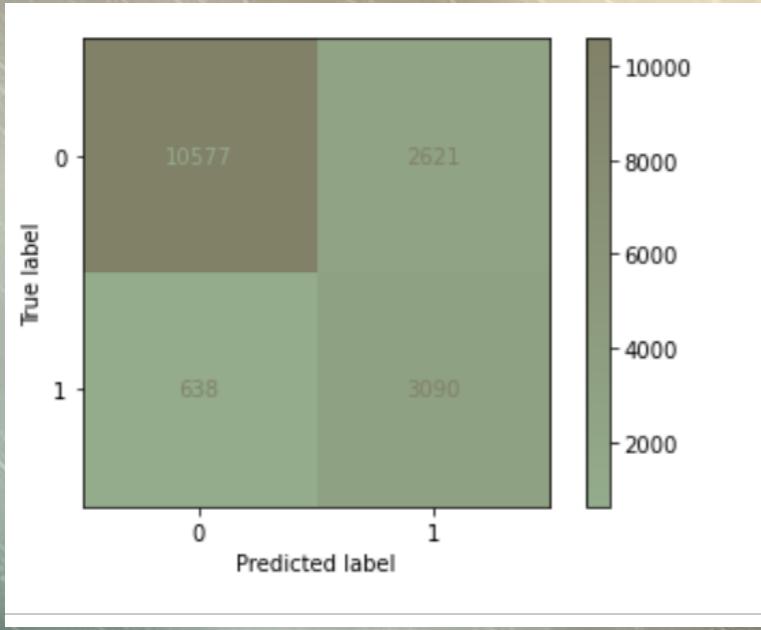
1 : 0.9

	Rain Tomorrow	No Rain Tomorrow
Before SMOTE	12427 (22%)	43993 (78%)
After SMOTE	21556 (47%)	23951 (53%)

XGBOOST WITH FEATURE SELECTION AND SMOTE

Metrics

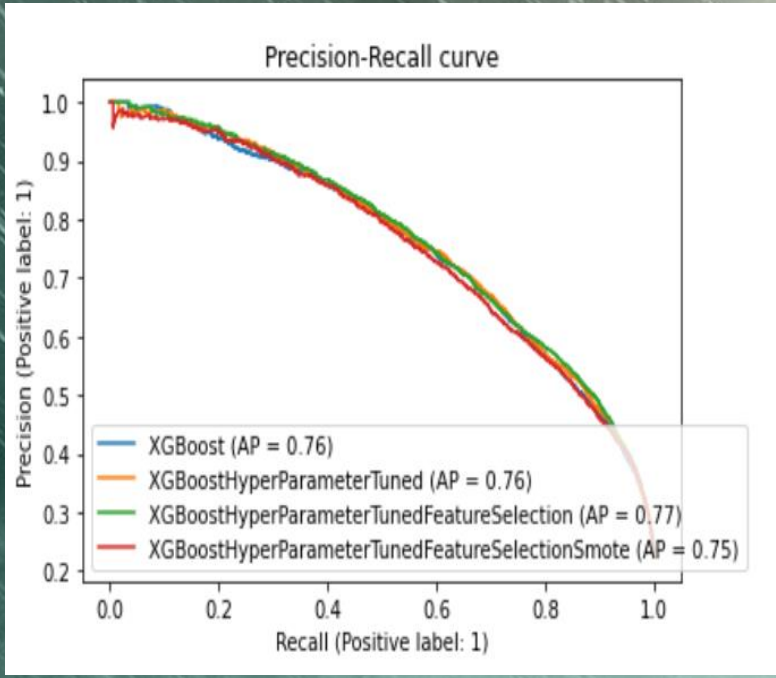
Accuracy : 85%
F1 Score : 69%
Precision :94%
(Major Class)
Precision :54% 11%
(Minor Class)
Recall :80%
(Major Class)
Recall :83% 11%
(Minor Class)
Training Accuracy :91%
Test Accuracy :81%



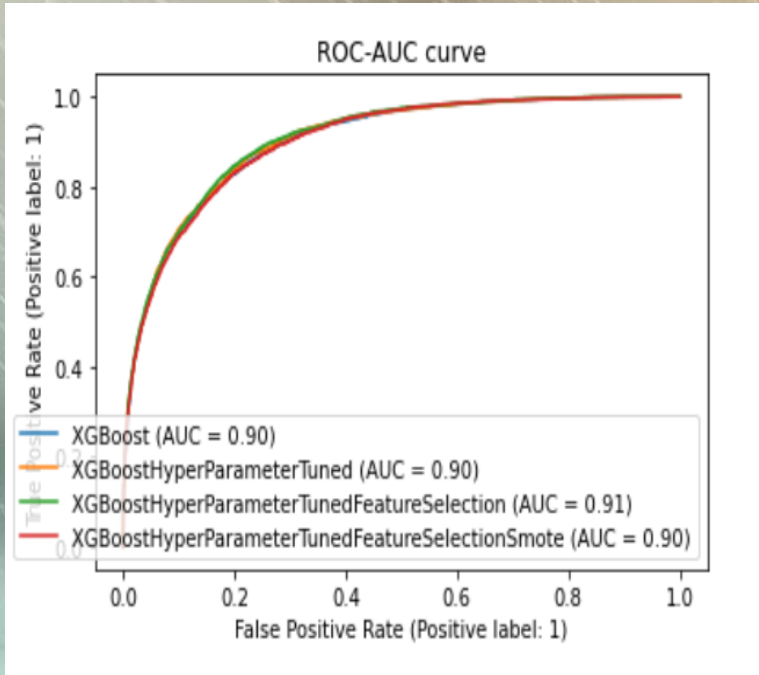
learning_rate : 0.1
max_depth : 8
reg_lambda : 1.0
gamma : 1.5
scale_pos_weight : 3

PR CURVE AND ROC-AUC CURVE

Precision-Recall



ROC-AUC



MORE FEATURE ENGINEERING

- Used ExtraTreesClassifier and Information Gain for Feature Selection – reduced from 98 features to 15
- Imputed Null Values using corresponding Day and Month of other years
- Increased Number of rows from 56+K to 120+K
- Added a new column to reflect high and low rainfall based on EDA
- Applied the SMOTE over and under sampling to balance the dataset

SMOTE

	Rain Tomorrow	No Rain Tomorrow
Before SMOTE	12427 (22%)	43993 (78%)
After SMOTE	21556 (47%)	23951 (53%)
Before SMOTE (After Feature Engineering)	24961 (22%)	87668 (78%)
After SMOTE (After Feature Engineering)	61367 (47%)	68185 (53%)

MULTIPLE MODELS

	Logistic	KNN	SVM	RF	XGBoost	MLP
Accuracy	: 78 %	84%	79%	87%	83%	81%
F1 Score	: 76 %	84%	77%	86%	79%	79%
Precision (Major Class)	:78 %	93%	79%	88%	84%	81%
Precision (Minor Class)	:78 %	78%	78%	86%	88%	80%
Recall (Major Class)	:81 %	76%	81%	87%	90%	83%
Recall (Minor Class)93	:75 %	93%	76%	86%	76%	78%

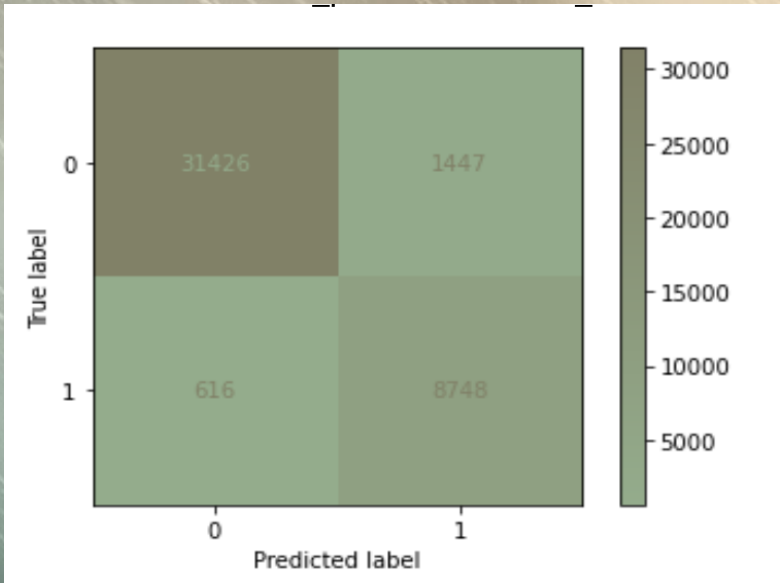
MULTIPLE MODELS

	Logistic	KNN	SVM	RF	XGBoost	MLP
Accuracy	: 85 %	83%	83%	86%	86%	82%
F1 Score	: 61 %	58%	43%	62%	65%	60%
Precision (Major Class)	:89 %	87%	83%	87%	89%	89%
Precision (Minor Class)	:76 %	65%	86%	77%	74%	67%
Recall (Major Class)	:95 %	92%	98%	95%	94%	89%
Recall (Minor Class)	:57 %	53%	28%	51%	58%	61%

HYPERPARAMETER TUNING FOR RANDOM FOREST

Metrics

Accuracy	: 88%	
F1 Score	: 87%	
Precision (Major Class)	:89%	
Precision (Minor Class)	:87 %	↑ 22%
Recall (Major Class)	:88%	
Recall (Minor Class)	:87%	↑ 16%
Training Accuracy	:95%	
Test Accuracy	:95%	



n_estimators : 500
max_features : 'auto'
min_samples_split :2
min_samples_leaf :4
max_depth: 35
criterion: 'entropy

IMPROVEMENTS

- Handle Outliers (True outliers vs Errors, IQR)
- More feature engineering e.g. Seasons, combining features
- Time series forecasting (e.g. predict rainfall over a longer period of time and predicting rain next week instead of tomorrow)

THANK YOU