Report on WeatherAustralia dataset

Auto2Class

January 12, 2025

Contents

1 Exploratory Data Analysis

1.1 Non-Null Count, Dtype of features

Table 1: Dataset Columns Information

Index	Column	Non-Null Count	Dtype
0	Date	145460	object
1	Location	145460	object
2	MinTemp	143975	float64
3	MaxTemp	144199	float64
4	Rainfall	142199	float64
5	Evaporation	82670	float64
6	Sunshine	75625	float64
7	WindGustDir	135134	object
8	WindGustSpeed	135197	float64
9	WindDir9am	134894	object
10	WindDir3pm	141232	object
11	WindSpeed9am	143693	float64
12	WindSpeed3pm	142398	float64
13	Humidity9am	142806	float64
14	Humidity3pm	140953	float64
15	Pressure9am	130395	float64
16	Pressure3pm	130432	float64
17	Cloud9am	89572	float64
18	Cloud3pm	86102	float64
19	Temp9am	143693	float64
20	Temp3pm	141851	float64
21	RainToday	142199	object
22	RainTomorrow	142193	object

1.2 Descriptive Statistics

Table 2: Dataset Descriptive Statistics

Index	Column Name/Statistic	count	mean	std	min	25%	50%	75%	max
0	MinTemp	143975.0	12.19	6.4	-8.5	7.6	12.0	16.9	33.9
1	MaxTemp	144199.0	23.22	7.12	-4.8	17.9	22.6	28.2	48.1
2	Rainfall	142199.0	2.36	8.48	0.0	0.0	0.0	0.8	371.0
3	Evaporation	82670.0	5.47	4.19	0.0	2.6	4.8	7.4	145.0
4	Sunshine	75625.0	7.61	3.79	0.0	4.8	8.4	10.6	14.5
5	WindGustSpeed	135197.0	40.04	13.61	6.0	31.0	39.0	48.0	135.0
6	WindSpeed9am	143693.0	14.04	8.92	0.0	7.0	13.0	19.0	130.0
7	WindSpeed3pm	142398.0	18.66	8.81	0.0	13.0	19.0	24.0	87.0
8	Humidity9am	142806.0	68.88	19.03	0.0	57.0	70.0	83.0	100.0
9	Humidity3pm	140953.0	51.54	20.8	0.0	37.0	52.0	66.0	100.0
10	Pressure9am	130395.0	1017.65	7.11	980.5	1012.9	1017.6	1022.4	1041.0
11	Pressure3pm	130432.0	1015.26	7.04	977.1	1010.4	1015.2	1020.0	1039.6
12	Cloud9am	89572.0	4.45	2.89	0.0	1.0	5.0	7.0	9.0
13	Cloud3pm	86102.0	4.51	2.72	0.0	2.0	5.0	7.0	9.0
14	Temp9am	143693.0	16.99	6.49	-7.2	12.3	16.7	21.6	40.2
15	Temp3pm	141851.0	21.68	6.94	-5.4	16.6	21.1	26.4	46.7

1.3 Distribution of features

1.3.1 Histograms of Numerical columns

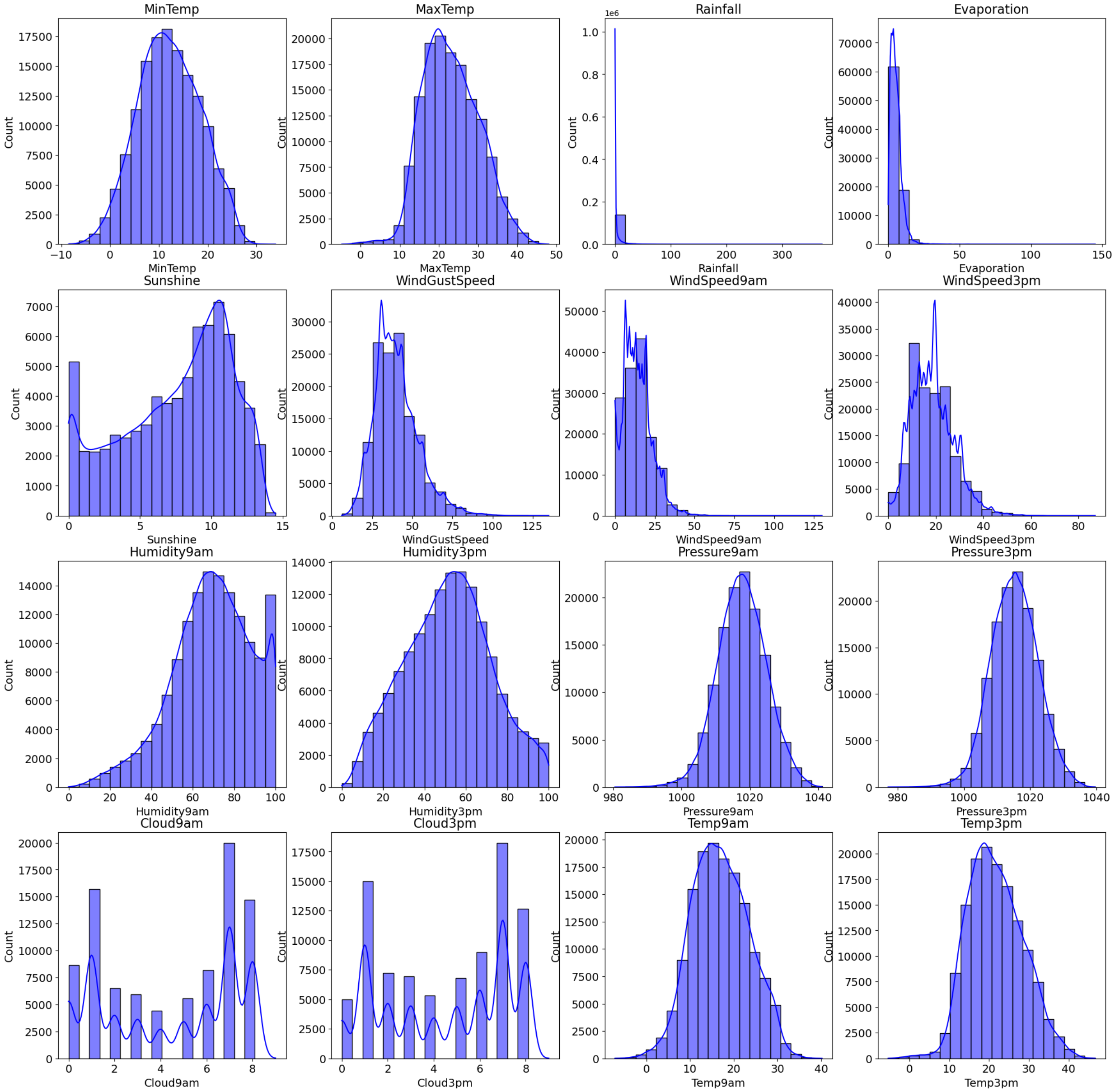

Figure 1: Histograms of Numerical columns

1.3.2 Bar Charts of Categorical columns

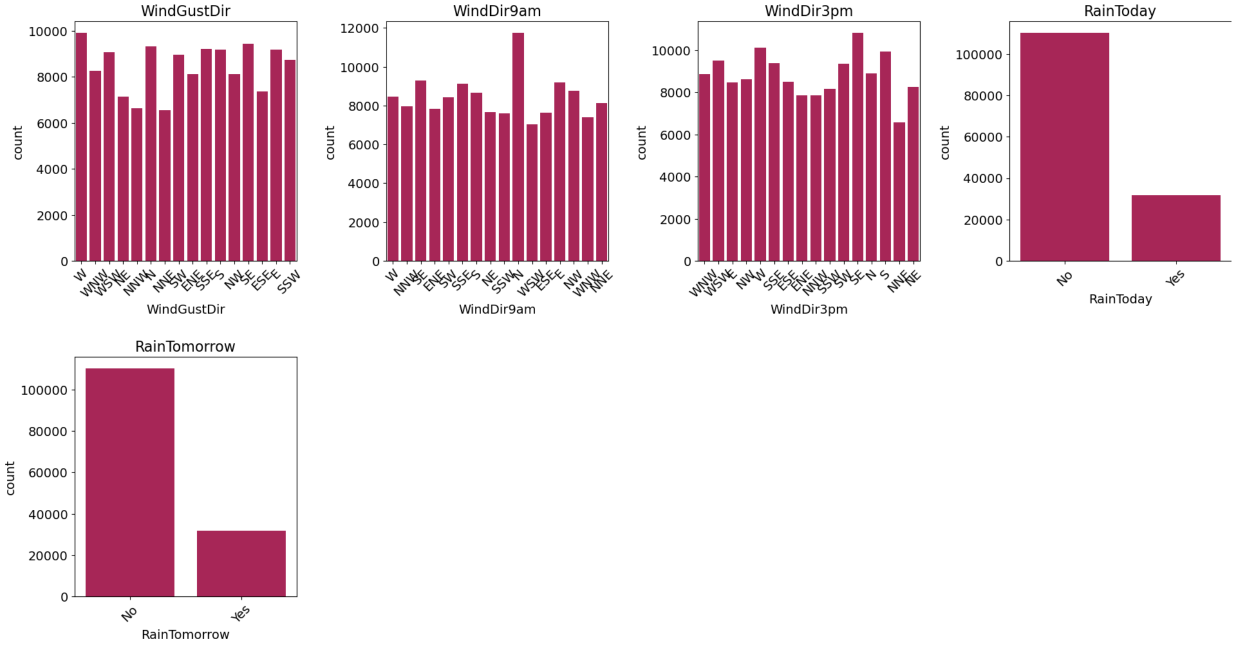

Figure 2: Bar Charts of Categorical columns

2 Evaluation Metrics

2.1 Accuracy

Accuracy is one of the simplest evaluation metrics for classification models. It is defined as the ratio of correctly predicted observations to the total number of observations:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

While accuracy is intuitive and easy to understand, it may not be suitable for imbalanced datasets. For example, in a dataset where 95% of the samples belong to one class, predicting the majority class for every instance would result in high accuracy but poor performance on the minority class.

2.2 F1 Score

The **F1 Score** is the harmonic mean of Precision and Recall, providing a balance between the two. It is particularly useful when dealing with imbalanced datasets. Precision and Recall are defined as follows:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

The F1 Score combines these metrics:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

A high F1 Score indicates a good balance between Precision and Recall, making it a valuable metric in scenarios where false positives and false negatives have significant costs.

2.3 ROC AUC

The Receiver Operating Characteristic (ROC) curve plots the True Positive Rate (Recall) against the False Positive Rate at various threshold settings. The **Area Under the Curve (AUC) of the ROC curve** measures the overall ability of the model to distinguish between classes.

$$\text{AUC} = \int_{\text{FPR}=0}^{1} \text{TPR}(\text{FPR})\, d(\text{FPR})$$

Key points about ROC AUC:

- An AUC of 0.5 indicates random guessing.
- An AUC of 1.0 indicates perfect classification.
- It is a threshold-independent metric, providing an aggregate measure of performance across all classification thresholds.

ROC AUC is particularly useful for binary classification tasks and provides insights into the trade-off between sensitivity and specificity.

3 Model Optimization Results

3.1 Optimization Results Tables

Table 3: Random Forest Hyperparameters and achivied metrics

Index	Metric/Hyperp.\ Iteration	0
0	f1	0.9472
1	accuracy	0.9473
2	roc_auc	0.9912
3	n_estimators	100
4	criterion	gini
5	max_depth	None
6	min_samples_split	2
7	min_samples_leaf	1
8	min_weight_fraction_leaf	0.0
9	max_features	sqrt
10	bootstrap	1

Table 4: Decision Tree Hyperparameters and achivied metrics

Index	Metric/Hyperp. \ Iteration	0	1	2	3	4	5	6	7
0	f1	0.9082	0.704	0.8557	0.7101	0.6383	0.7847	0.879	0.4955
1	accuracy	0.9085	0.7103	0.8558	0.7144	0.6521	0.7847	0.8793	0.5005
2	roc_auc	0.9085	0.7133	0.9108	0.7759	0.6506	0.8695	0.907	0.5008
3	criterion	gini	log_loss	log_loss	gini	gini	entropy	entropy	entropy
4	splitter	best	best	best	best	random	best	random	best
5	max_depth	None	None	40	10	40	10	40	40
6	min_samples_split	2	10	2	10	5	5	5	5
7	min_samples_leaf	1	2	4	4	1	1	1	4
8	max_features	None	None	sqrt	None	None	None	log2	log2
9	class_weight	None	None	None	None	balanced	balanced	balanced	balanced
10	min_impurity_decrease	0.0	0.1	0.0	0.01	0.05	0.0	0.0	0.1

Table 5: XGBoost Hyperparameters and achivied metrics

Index	Metric/Hyperp. \ Iteration	0
0	f1	0.8379
1	accuracy	0.8379
2	roc_auc	0.9189
3	eval_metric	logloss
4	n_estimators	100
5	max_depth	6
6	learning_rate	0.3
7	subsample	1.0
8	colsample_bytree	1.0
9	min_child_weight	1
10	gamma	0
11	reg_alpha	0
12	reg_lambda	1

3.2 Boxplots of accuracy, f1, roc_auc

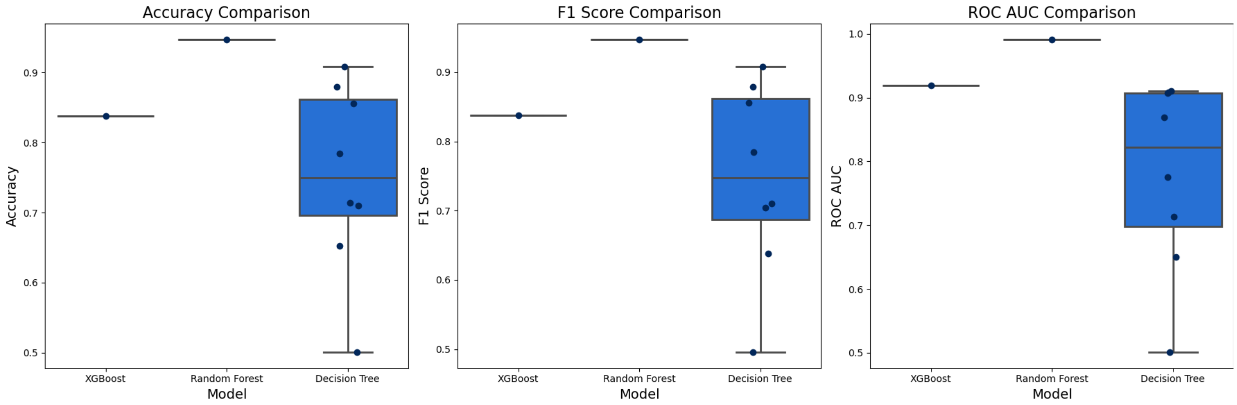

Figure 3: Boxplots of accuracy, f1, roc_auc

3.3 Barplots of maximum values of metrics achievied by model

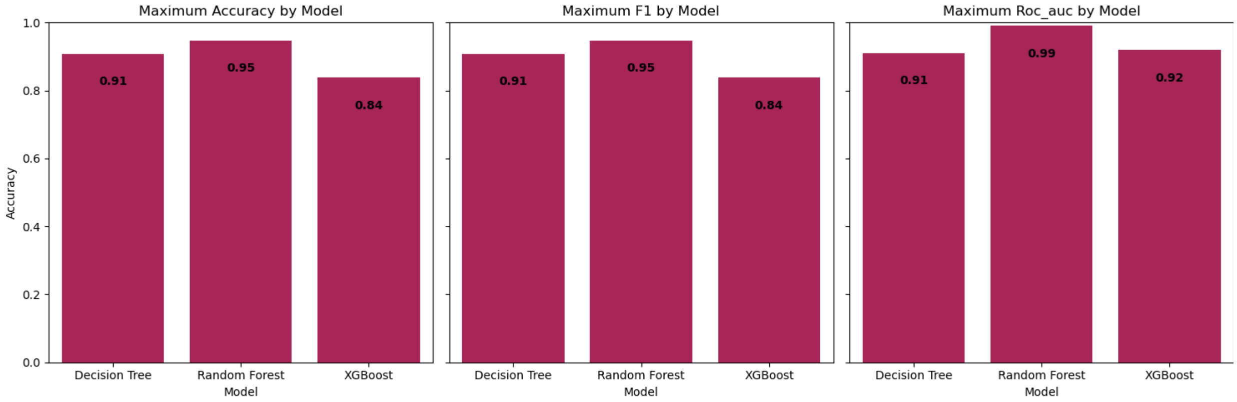

Figure 4: Barplots of maximum values of metrics achievied by model