Rain Prediction Using Machine Learning

Tools

Table of Contents

01 Business Background and Objectives

02 Data Preparation

03 Exploratory Data Analysis (EDA)

04 Feature Engineering

05 Modelling and Evaluation

06 Conclusion and Recommendation

1

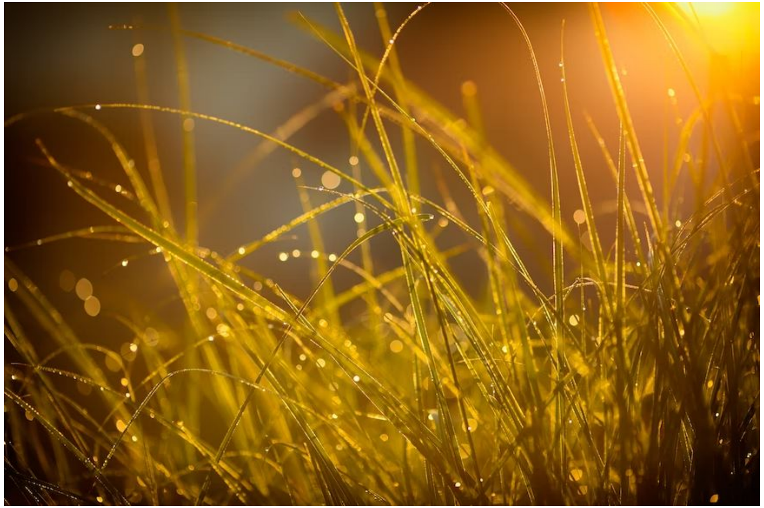
Business Background and Objectives

Introduction and Problems

Weather has a significant impact on many life aspects, one of which is agricultural industry and because of that, being able to predict it helps farmers in their day-to-day decisions such as how to plan efficiently, minimize costs and maximize yields.

A major agricultural company needs to have an accurate **rain prediction algorithm** that will improve their decision-making on typical farming activities such as planting and irrigating.

Using historical rain information from Australia regions in 10 years as research data, it is necessary to **predict** weather(**rain**) in the **next day**.



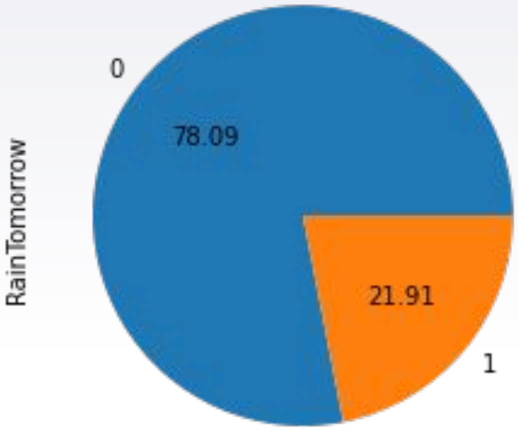
Objectives

- What factors and conditions in current day that will cause rain in the next day?
- What machine learning algorithms are suitable for predicting rain in the next day?
- Where is the location that has the highest frequency of rain?
- What is the impact of the predictive model for business problems that operating in the agricultural sector?

2

Data Preparation

Dataset Overview



Target distribution is **imbalance**, I will use **ROC** curve and **AUC** score as model metric evaluation.

ROC-AUC is performance measurement for the classification problems at various threshold settings.

Data contains **145460 rows** with **22 features** and **1 binary column** with **RainTomorrow** as **target**.

All features are the weather and climate elements that occur on that day in certain location that are used to predict rain in the next day.

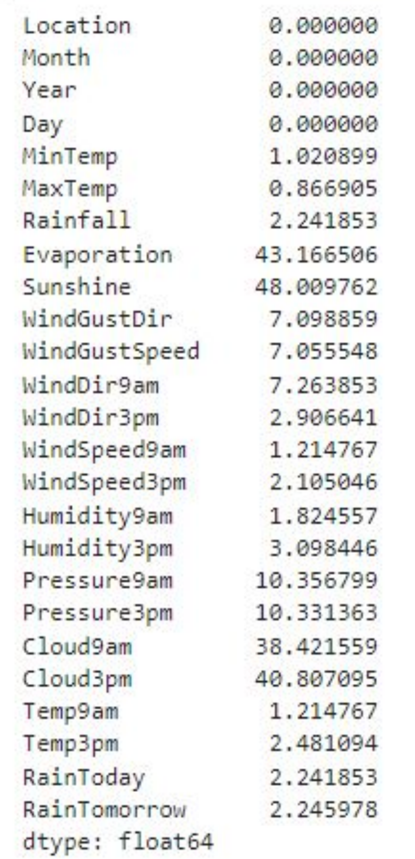
Features consists **16 numerical features** and **6 categorical features** (Date will be transform to datetime format).

Numerical Features		Categorical Features
● MinTemp	● Humidity9am	
● MaxTemp	● Humidity3pm	● Date
● Rainfall	● Pressure9am	● Location
● Evaporation	● Pressure3pm	● WindGustDir
● Sunshine	● Cloud9am	● WindDir9am
● WindGustSpeed	● Cloud3pm	● WindDir3pm
● WindSpeed9am	● Temp9am	● RainToday
● WindSpeed3pm	● Temp3pm	

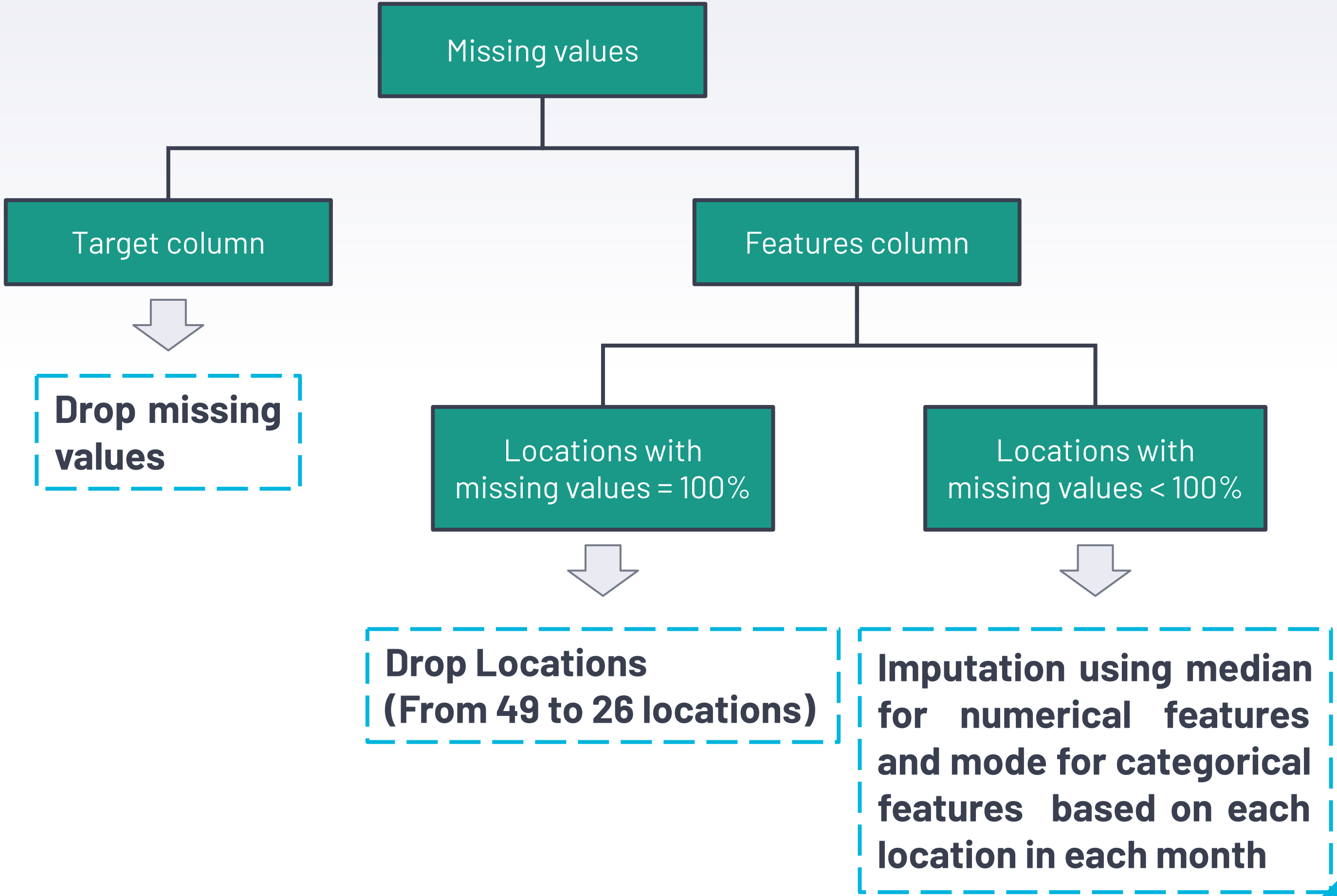
Dataset obtained from **kaggle**.

Missing Values Handling

Missing Values Ratio



```
Location          0.000000
Month             0.000000
Year              0.000000
Day               0.000000
MinTemp           1.020899
MaxTemp           0.866905
Rainfall          2.241853
Evaporation      43.166506
Sunshine         48.009762
WindGustDir       7.098859
WindGustSpeed     7.055548
WindDir9am        7.263853
WindDir3pm        2.906641
WindSpeed9am      1.214767
WindSpeed3pm      2.105046
Humidity9am       1.824557
Humidity3pm       3.098446
Pressure9am      10.356799
Pressure3pm      10.331363
Cloud9am         38.421559
Cloud3pm         40.807095
Temp9am           1.214767
Temp3pm           2.481094
RainToday         2.241853
RainTomorrow      2.245978
dtype: float64
```



Missing values

Target column

Features column

Drop missing values

Locations with missing values = 100%

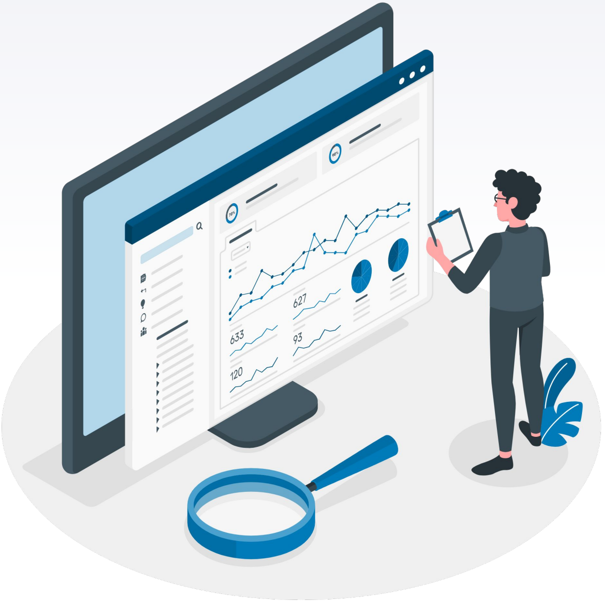
Locations with missing values < 100%

Drop Locations (From 49 to 26 locations)

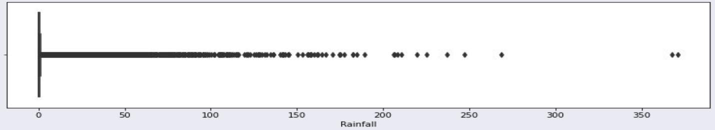
Imputation using median for numerical features and mode for categorical features based on each location in each month

3

Exploratory Data Analysis (EDA)

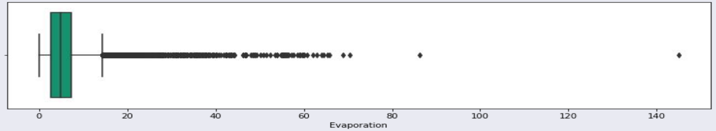


Outliers Check



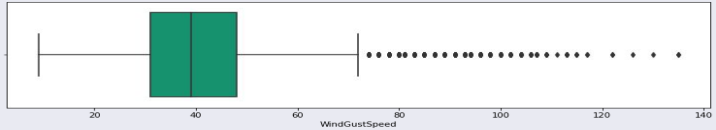
Rainfall column outliers

Values > 2.4



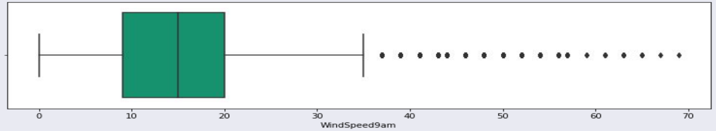
Evaporation column outliers

Values > 21.2



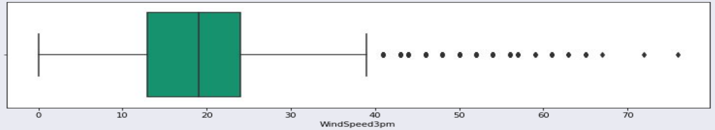
WindGustSpeed column outliers

Values > 99.0



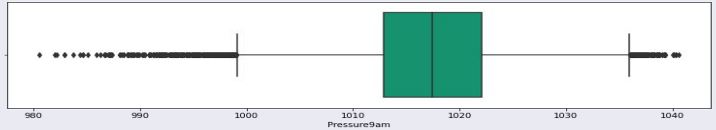
WindSpeed9am column outliers

Values > 53.0



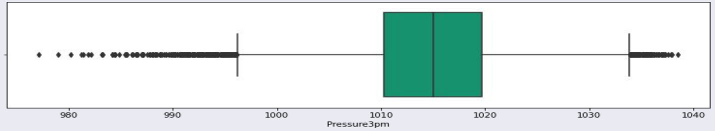
WindSpeed3pm column outliers

Values > 57.0



Pressure9am column outliers

Values < 985.3

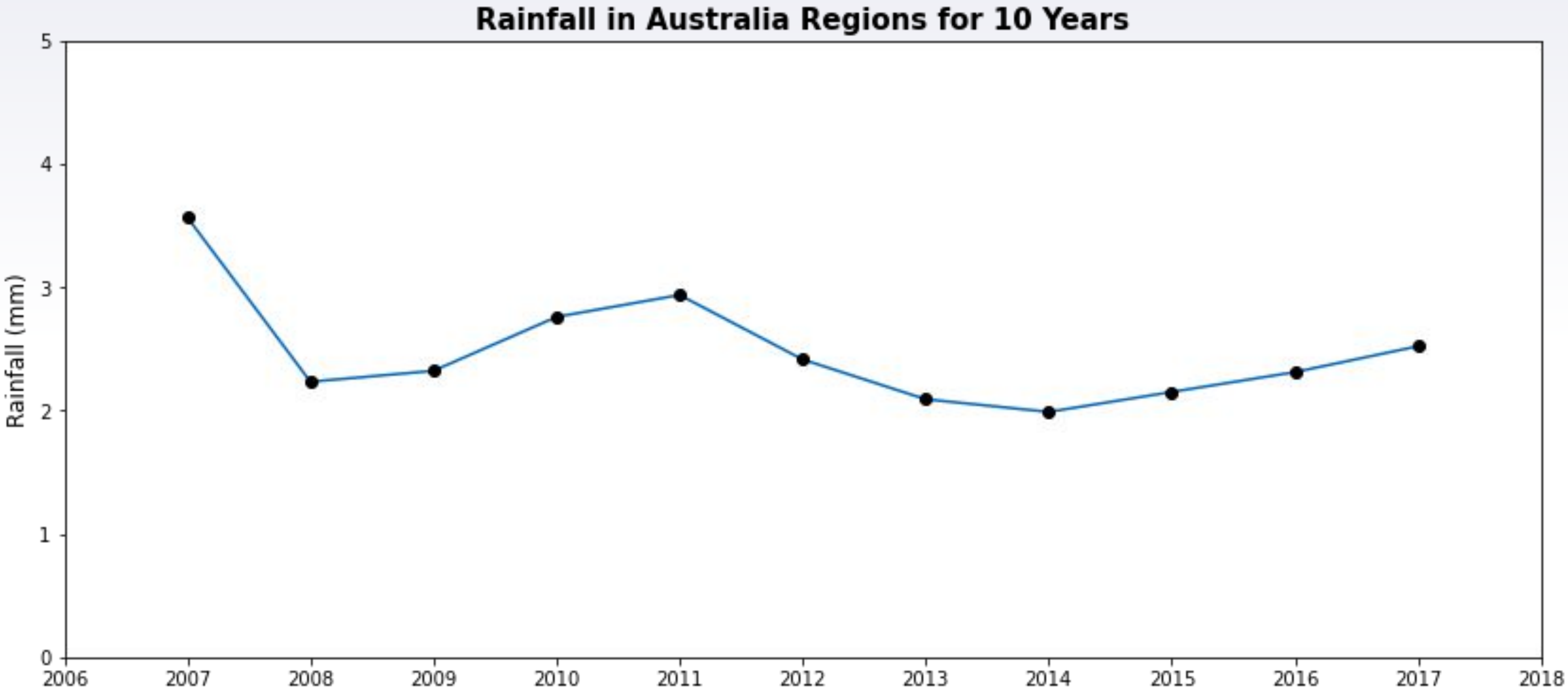


Pressure3pm column outliers

Values < 982.1

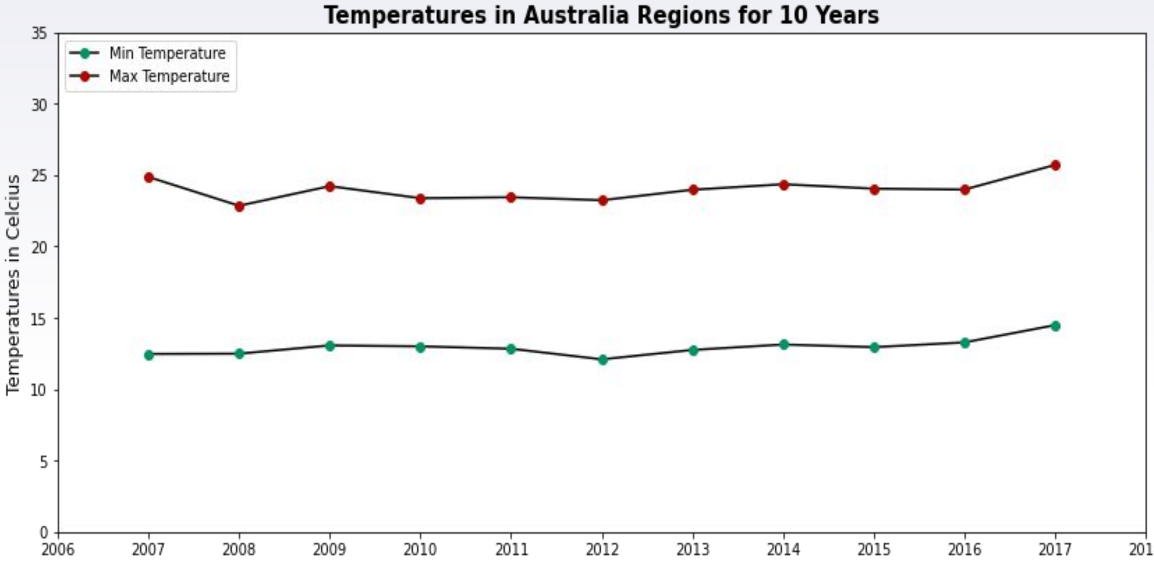
Seven features have extreme **outliers** and need to removed based **IQR** (Interquartile Range) **upper** and **lower** limit.

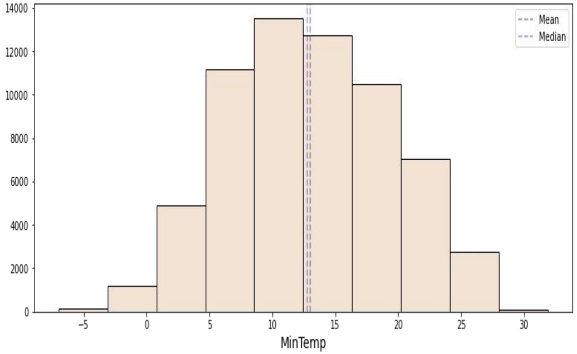
Rainfall in Australia Regions

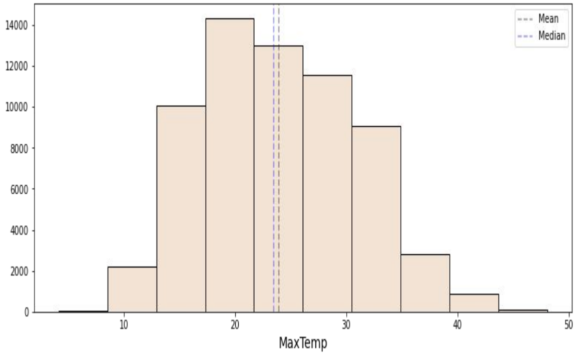


Average rainfall in Australia regions in 10 years relatively **decreased** over years. It seems the decreased rainfall caused by global warming issue.

Temperatures in Australia Regions

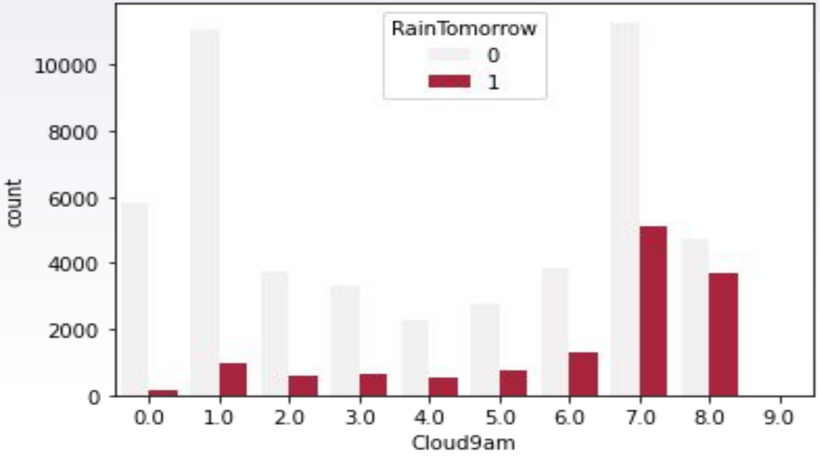


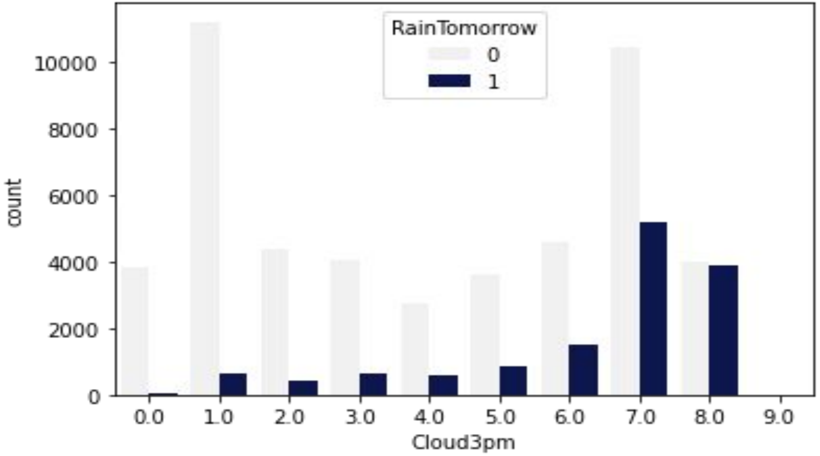




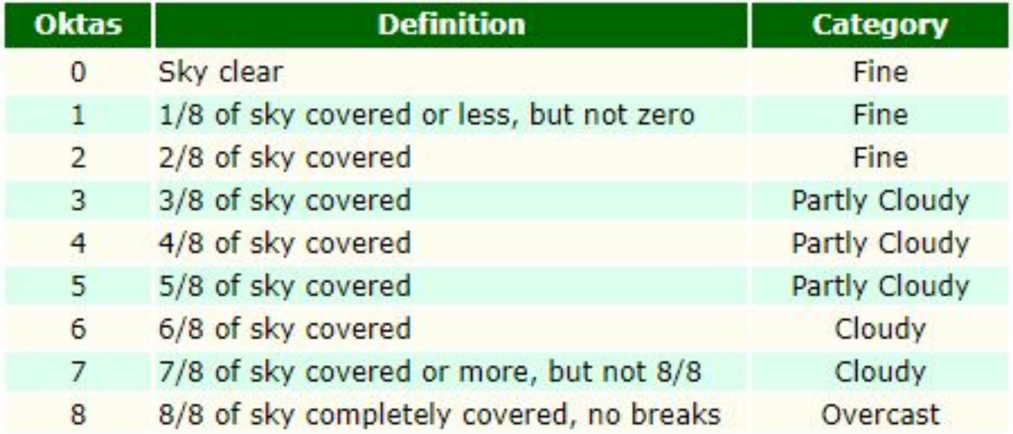
Average minimum and **maximum temperatures** for 10 years **slightly increase**, due to global warming issue. For temperatures **data distribution**, relatively **normal distributed**. It can be concluded that Australia regions have a **relatively stable minimum** and **maximum temperatures**.

Fraction of Sky by Cloud (in Oktas)





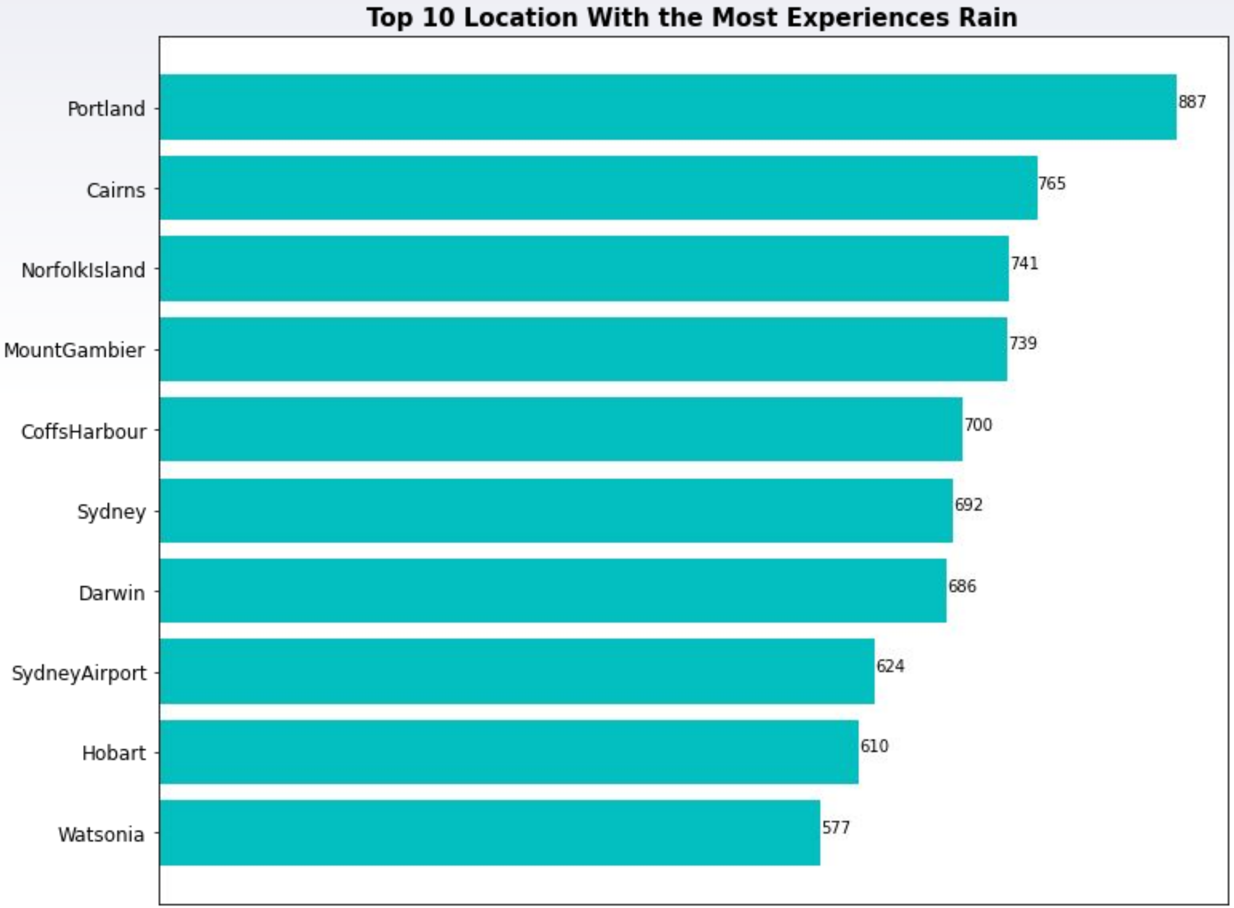
Sky in Australia **mostly cloudy** (**Fraction** of **sky** by **cloud** is **7 oktas**). **Rain** in the **next day** tend to **happen** when **cloud** in **9am** and **3pm** is **7** or **8 oktas**.



Oktas	Definition	Category
0	Sky clear	Fine
1	1/8 of sky covered or less, but not zero	Fine
2	2/8 of sky covered	Fine
3	3/8 of sky covered	Partly Cloudy
4	4/8 of sky covered	Partly Cloudy
5	5/8 of sky covered	Partly Cloudy
6	6/8 of sky covered	Cloudy
7	7/8 of sky covered or more, but not 8/8	Cloudy
8	8/8 of sky completely covered, no breaks	Overcast

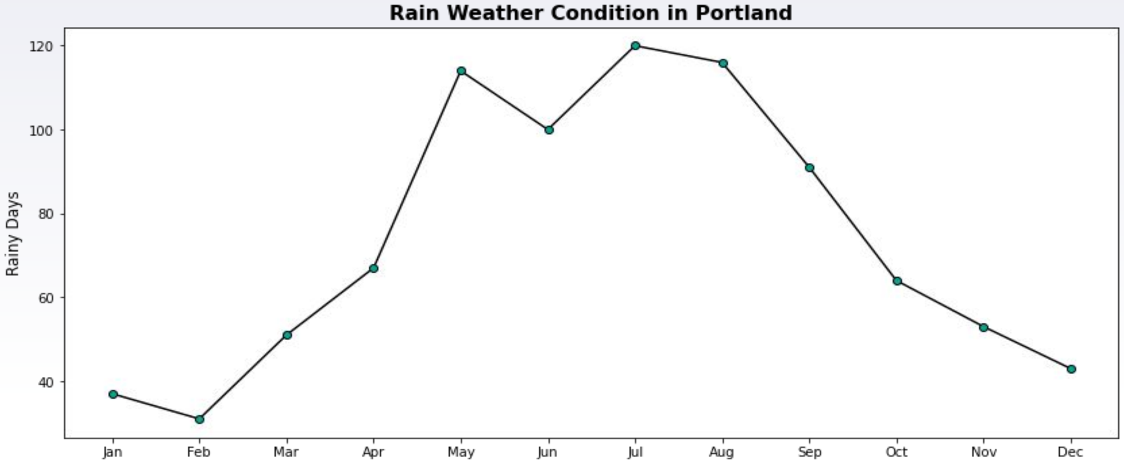
(Reference: Worldweather)

Rain Frequency Based on Locations



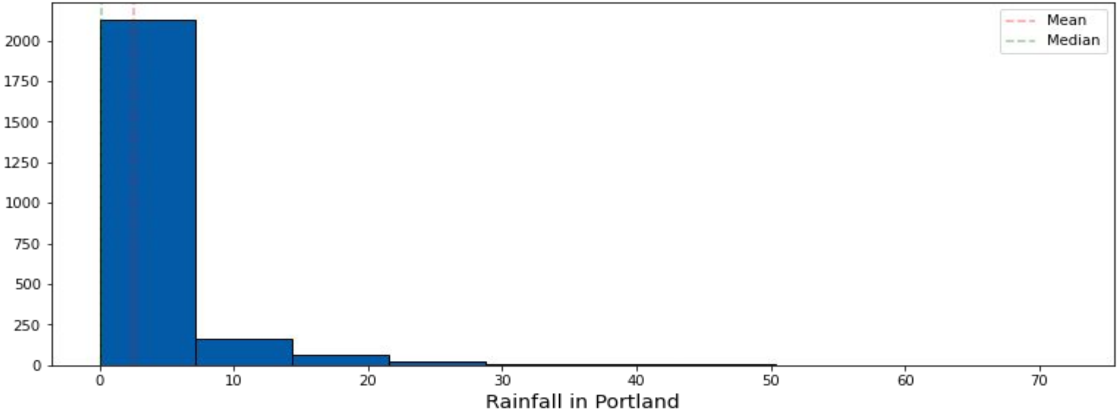
Portland is the **location** in Australia that **experiences** the **most rain**. It means Portland is the location with the enough water source.

Rain Condition in Portland



Rain Period in Portland

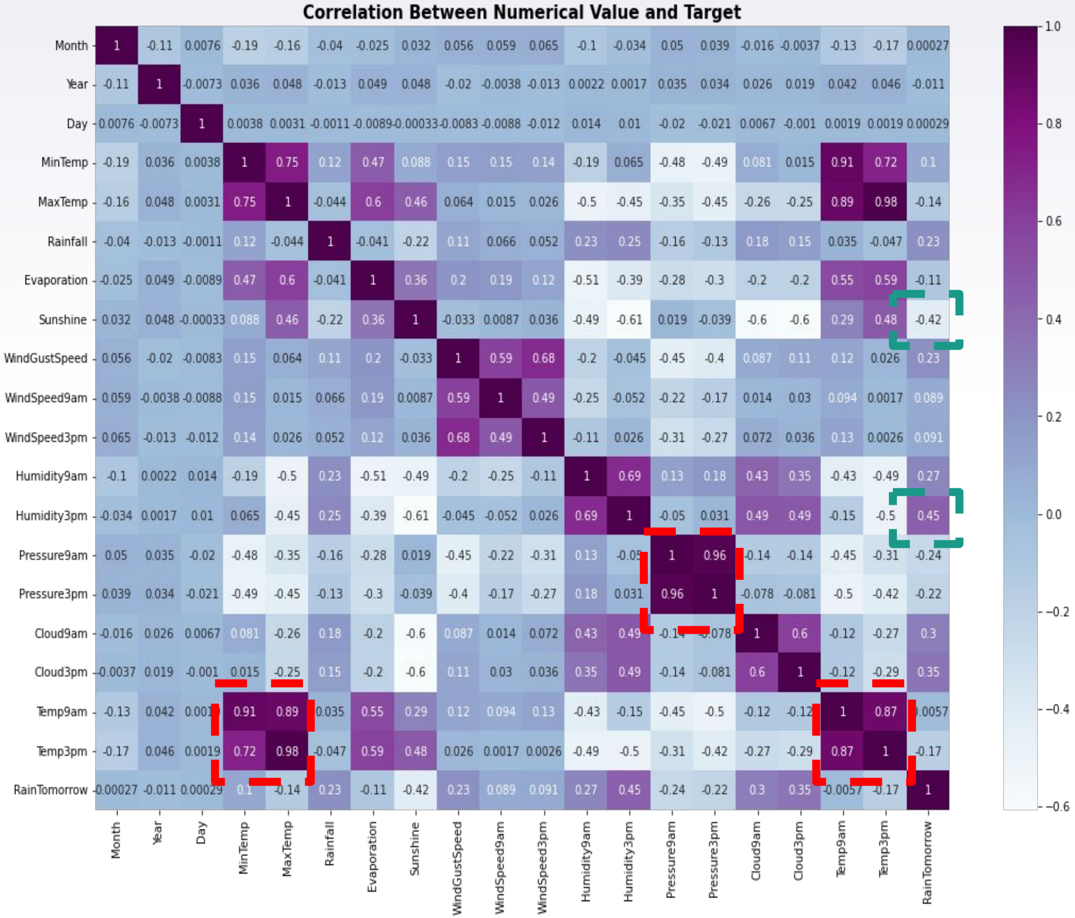
Rain in Portland is most common to **happen** from **period May** to **August**. We can conclude that these period is the rainy season in Portland.



Rainfall in Portland

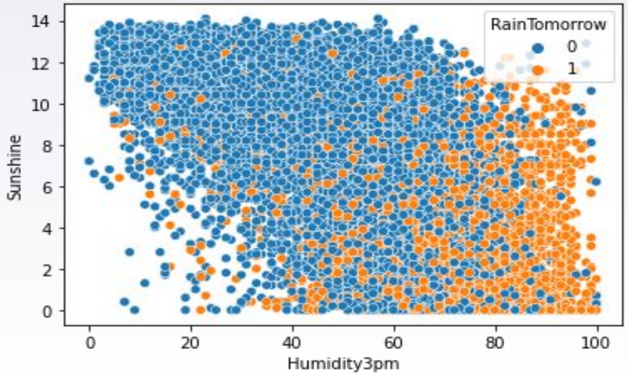
Rainfall in Portland majority in **range** around **0 - 8 mm** with **average** of **rainfall** is **2.49 mm**.

Multivariate Analysis



Correlation Between Numerical Value and Target

	Month	Year	Day	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainTomorrow
Month	1	-0.11	0.0076	-0.19	-0.16	-0.04	-0.025	0.032	0.056	0.059	0.065	-0.1	-0.034	0.05	0.039	-0.016	-0.0037	-0.13	-0.17	0.00027
Year	-0.11	1	-0.0073	0.036	0.048	-0.013	0.049	0.048	-0.02	-0.0038	-0.013	0.0022	0.0017	0.035	0.034	0.026	0.019	0.042	0.046	-0.011
Day	0.0076	-0.0073	1	0.0038	0.0031	-0.0011	-0.0089	-0.00033	-0.0083	-0.0088	-0.012	0.014	0.01	-0.02	-0.021	0.0067	-0.001	0.0019	0.0019	0.00029
MinTemp	-0.19	0.036	0.0038	1	0.75	0.12	0.47	0.088	0.15	0.15	0.14	-0.19	0.065	-0.48	-0.49	0.081	0.015	0.91	0.72	0.1
MaxTemp	-0.16	0.048	0.0031	0.75	1	-0.044	0.6	0.46	0.064	0.015	0.026	-0.5	-0.45	-0.35	-0.45	-0.26	-0.25	0.89	0.98	-0.14
Rainfall	-0.04	-0.013	-0.0011	0.12	-0.044	1	-0.041	-0.22	0.11	0.066	0.052	0.23	0.25	-0.16	-0.13	0.18	0.15	0.035	-0.047	0.23
Evaporation	-0.025	0.049	-0.0089	0.47	0.6	-0.041	1	0.36	0.2	0.19	0.12	-0.51	-0.39	-0.28	-0.3	-0.2	-0.2	0.55	0.59	-0.11
Sunshine	0.032	0.048	-0.00033	0.088	0.46	-0.22	0.36	1	-0.033	0.0087	0.036	-0.49	-0.61	0.019	-0.039	-0.6	-0.6	0.29	0.48	-0.42
WindGustSpeed	0.056	-0.02	-0.0083	0.15	0.064	0.11	0.2	-0.033	1	0.59	0.68	-0.2	-0.045	-0.45	-0.4	0.087	0.11	0.12	0.026	0.23
WindSpeed9am	0.059	-0.0038	-0.0088	0.15	0.015	0.066	0.19	0.0087	0.59	1	0.49	-0.25	-0.052	-0.22	-0.17	0.014	0.03	0.094	0.0017	0.089
WindSpeed3pm	0.065	-0.013	-0.012	0.14	0.026	0.052	0.12	0.036	0.68	0.49	1	-0.11	0.026	-0.31	-0.27	0.072	0.036	0.13	0.0026	0.091
Humidity9am	-0.1	0.0022	0.014	-0.19	-0.5	0.23	-0.51	-0.49	-0.2	-0.25	-0.11	1	0.69	0.13	0.18	0.43	0.35	-0.43	-0.49	0.27
Humidity3pm	-0.034	0.0017	0.01	0.065	-0.45	0.25	-0.39	-0.61	-0.045	-0.052	0.026	0.69	1	-0.05	0.031	0.49	0.49	-0.15	-0.5	0.45
Pressure9am	0.05	0.035	-0.02	-0.48	-0.35	-0.16	-0.28	0.019	-0.45	-0.22	-0.31	0.13	-0.05	1	0.96	-0.14	-0.14	-0.45	-0.31	-0.24
Pressure3pm	0.039	0.034	-0.021	-0.49	-0.45	-0.13	-0.3	-0.039	-0.4	-0.17	-0.27	0.18	0.031	0.96	1	-0.078	-0.081	-0.5	-0.42	-0.22
Cloud9am	-0.016	0.026	0.0067	0.081	-0.26	0.18	-0.2	-0.6	0.087	0.014	0.072	0.43	0.49	-0.14	-0.078	1	0.6	-0.12	-0.27	0.3
Cloud3pm	-0.0037	0.019	-0.001	0.015	-0.25	0.15	-0.2	-0.6	0.11	0.03	0.036	0.35	0.49	-0.14	-0.081	0.6	1	-0.12	-0.29	0.35
Temp9am	-0.13	0.042	0.001	0.91	0.89	0.035	0.55	0.29	0.12	0.094	0.13	-0.43	-0.15	-0.45	-0.5	-0.12	-0.12	1	0.87	0.0057
Temp3pm	-0.17	0.046	0.0019	0.72	0.98	-0.047	0.59	0.48	0.026	0.0017	0.0026	-0.49	-0.5	-0.31	-0.42	-0.27	-0.29	0.87	1	-0.17
RainTomorrow	0.00027	-0.011	0.00029	0.1	-0.14	0.23	-0.11	-0.42	0.23	0.089	0.091	0.27	0.45	-0.24	-0.22	0.3	0.35	-0.0057	-0.17	1

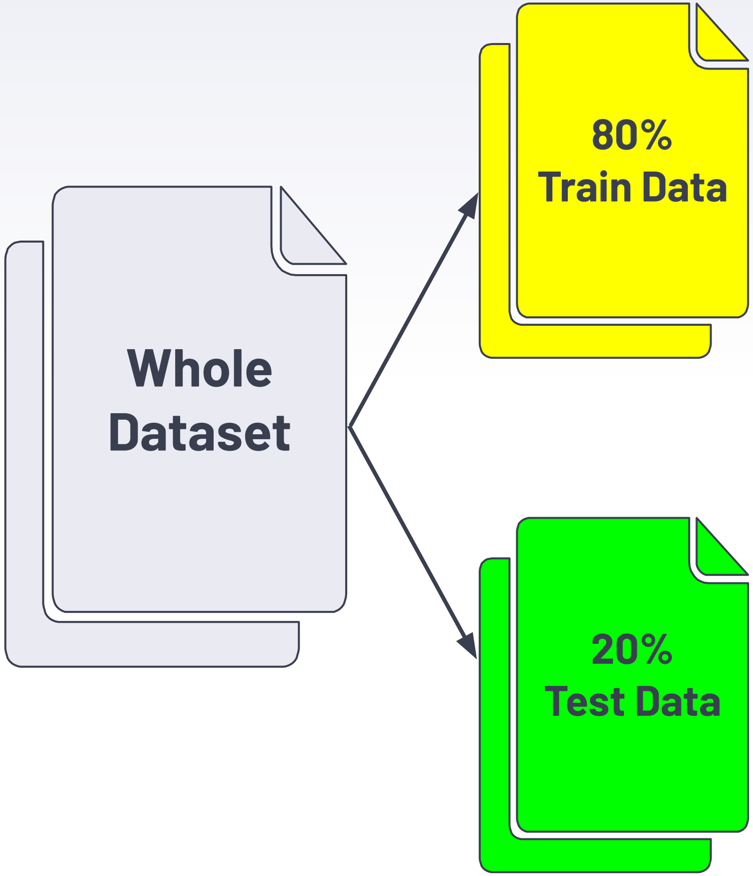


- **Humidity3pm** (**Positive**) and **Sunshine** (**Negative**) features have the **highest correlation** with **target**.
- **Pressure9am** and **Pressure3pm** then **MinTemp**, **MaxTemp**, **Temp9am**, and **Temp3pm** features have **high correlate each other**. **Pressure9am** and **Temp3pm** have **higher correlation with target** and will be kept for modelling.
- When **Humidity3pm ratio** is **high** and number of **hours** of **bright Sunshine** is **low**, **rain** in the **next day** tend to **happen**.

4

Feature Engineering

Dataset Split

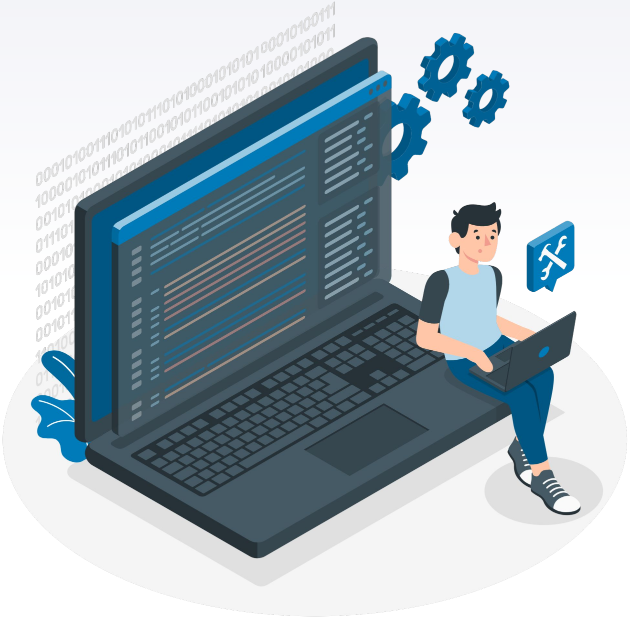


Feature Engineering

Aspects	Action	Treatment
Categorical features	Categorical feature with 2 distinct values -> Binary Encoding Categorical feature with more than 2 distinct values -> One Hot Encoding	Train Data Test Data
Features Drop	Multicollinearity -> Pressure3pm, MinTemp, MaxTemp, and Temp9am Don't contribute for modelling -> Year	Train Data Test Data
Imbalance Data	Resampling use Undersampling	Train Data
Scaling	Normalization using Min-Max Scaler	Train Data Test Data

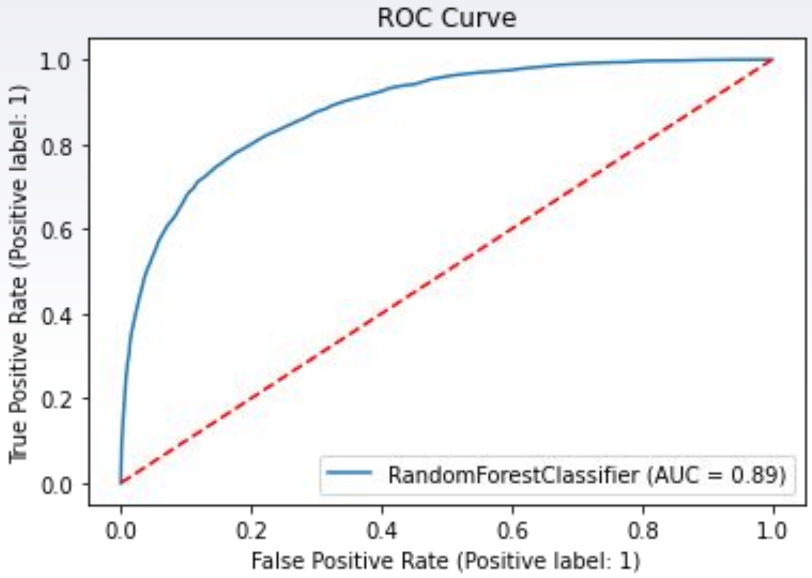
5

Modelling and Evaluation



Modelling

Model	AUC Score
K-Nearest Neighbors	79%
Logistic Regression	88%
Decision Tree	72%
Random Forest	**89%**
XG-Boost	88%



I **trained** data using **5 classification algorithms**. Since the **data** is **imbalance**, I use **ROC** curve and **AUC** score as **metric evaluation**. **Random Forest** have the **best AUC score** compared than other algorithms.

Hyperparameter Tuning

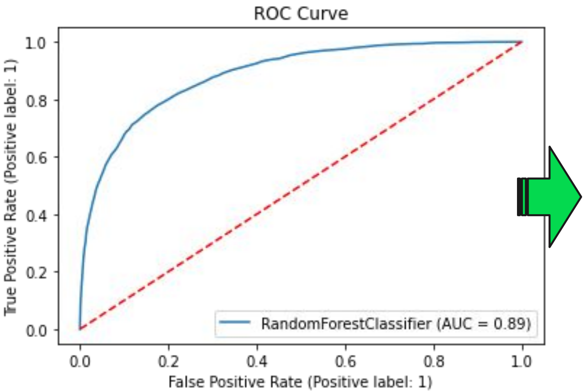
Parameter	Value
n_estimators (Number of trees)	600, 700, 800, 900, 1000
max_depth (Longest path between the root node and the leaf node)	60, 70, 80, 90, 100
criterion (Function to measure the quality of a split)	'gini', 'entropy'

Tuning

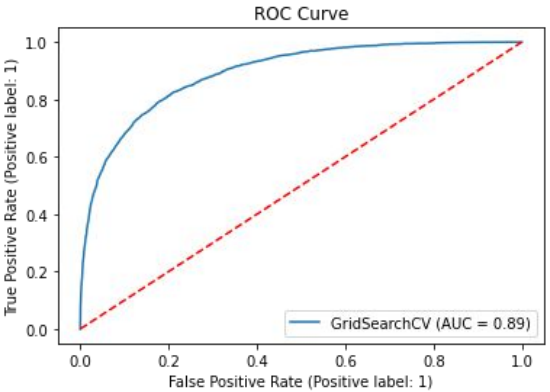
GridSearch CV

Parameter	Best Value
n_estimators	800
max_depth	60
criterion	'entropy'

Before Tuned

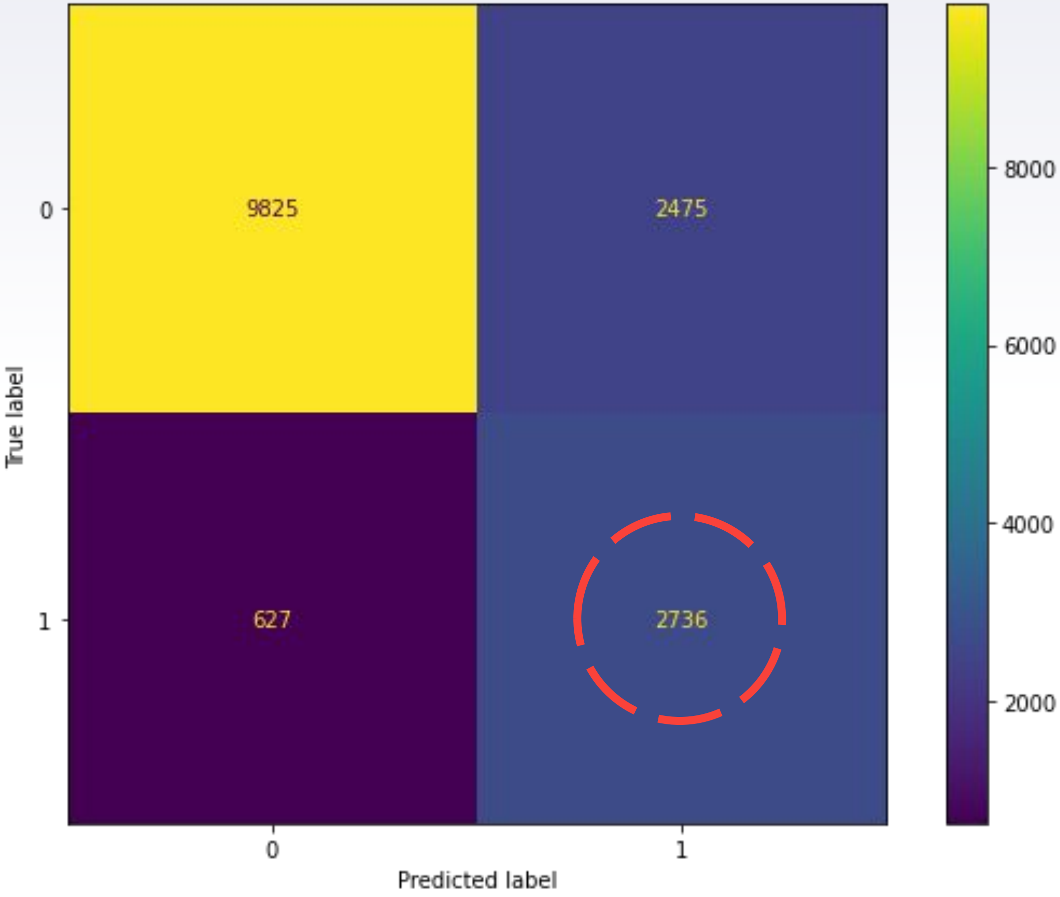


After Tuned



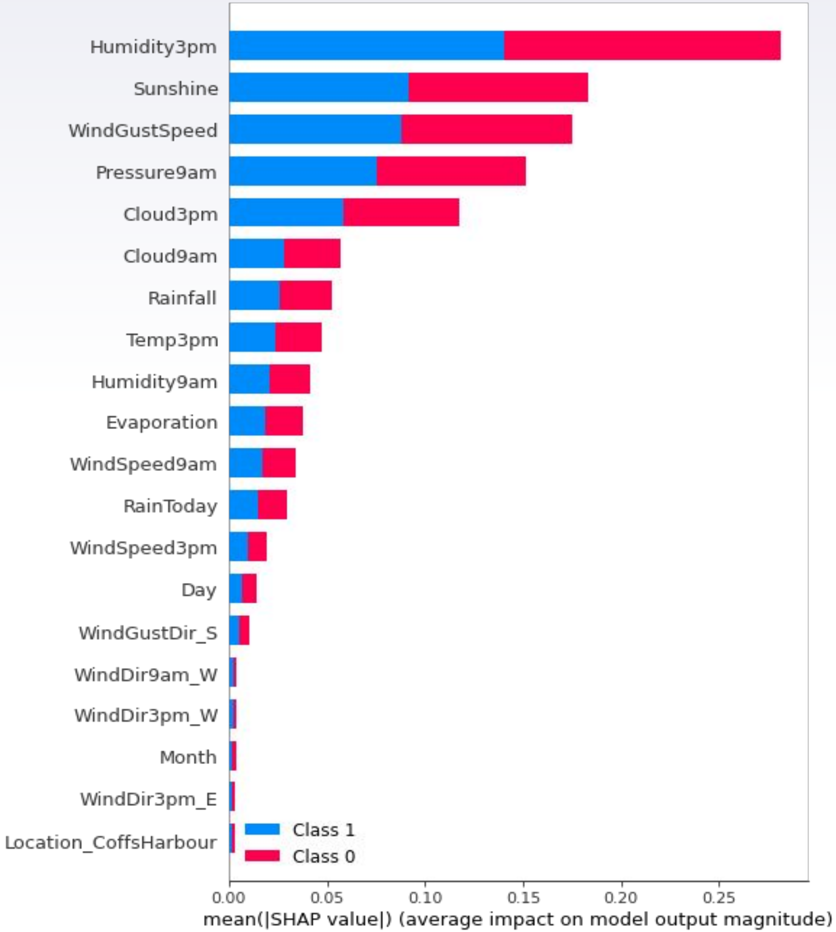
Overall, **ROC** curve and **AUC** score **before** and **after tuned** is **relatively same**. It means **before tuned**, the **model** already **have optimum performance**. But still, **hyperparameter tuning** is **essentials** part for **controlling model behavior**. It's also very important to **avoid** model **overfitting**.

Confusion Matrix



From confusion matrix we can analyze model have **True Positive Rate** (**TPR**) by **81%**. It means, from 3363 days when our model predict will be rain, 2736 days it really rained.

Feature Importances



Feature values in **pink** cause to **increase** the **prediction**. **Size** of the **bar** shows the **magnitude** of the **feature's effect**. **Feature** values in **blue** cause to **decrease** the **prediction**. Based on **SHAP** value, **Humidity3pm** and **Sunshine** feature have the **highest effect** on the **prediction**. It's same with heatmap correlation.

Potential Impact in Business

We can do simulation our predictive model for business in the agricultural sector in Australia. For example, water supply costs for irrigation 1000 hectares rice field in dry season.

Before

Water needs	= 1917 Liter/month
Price of water	= 0.29 USD/m^3
Cost for water	= 555,930 USD

After

Water needs	= 1312 Liter/month
Price of water	= 0.29 USD/m^3
Cost for water	= 380,460 USD

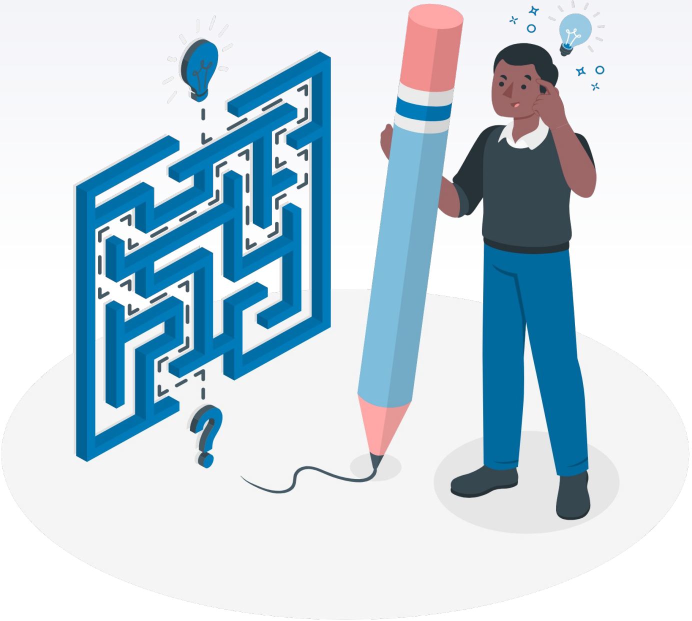
SAVING

$ 175,470 or 32%

on water supply for irrigation per month

6

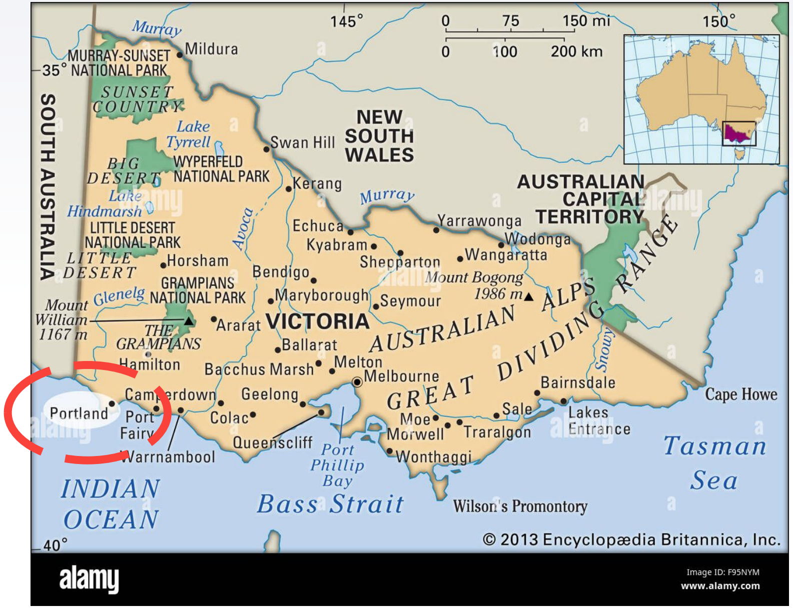
Conclusion and Recommendation



Conclusion

- Based on Heatmap and SHAP values for feature importances, Humidity at 3pm and Sunshine has the big impact to cause rain in the next day.
- Random Forest classifier is the best model algorithm for predicting rain the next day because have the highest AUC score than other classifier algorithm.
- The location with the most rain frequency is Portland with rain season tend to happen in May until August.
- Based on simulation, model performance can help save company cost for water supply by 32%.

Recommendation



© 2013 Encyclopædia Britannica, Inc.

If we want to start running the company that operating in agricultural sector, I think Australia regions is become one of the good choice since the temperature is relatively stable. Many agricultural products that can grown well based on those temperatures range. Also, I think Portland is the best location due to high rain frequency in a year, because it's very helpful for doing farming activities, such as planting and irrigation.

The best time for harvesting and make as much water stock is May until August. So, when the dry season comes, we won't too worried about lack of water and still can do activities like irrigation and farm will not easily to drought.

As a model evaluation, model have good performance and as performance trial, it would be better if the model was applied in Portland as the location with the highest rainfall frequency.

THANK YOU