Name: Jann Moises Nyll B. De los Reyes

Section: CPE22S3

Date: March 22, 2024

Submitted to: Engr. Roman M. Richard

Logistic Regression Classifier Tutorial with Python

1. Introduction to Logistic Regression

When data scientist may come across a new classification problem, the first algorithm that may come across their mind is **Logistic Regression**. It is a supervised learning classification algorithm which is used to predict observations to a discrete set of classes. Practically, it is used to classify observations into different categories. Hence, its output is discrete in nature. **Logistic Regression** is also called **Logit Regression**. It is one of the most simple, straightforward and versatile classification algorithms which is used to solve classification problems.

2. Logistic Regression intuition

In statistics, the **Logistic Regression Model** is a widely used statistical model which is primarily used for classification purposes. It means that given a set of observations, Logistic Regression algorithm help us to classify these observations into two or more discrete classes. So, the target variable is discrete in nature.

The Logistic Regression algorithm works as follows -

Implement linear equation

Logistic Regression Algorithm works by implementing a linear equation with independent or explanatory variables to predict a response value. For example, we consider the example of number of hours studied and probability of passing the exam. Here, number of hour studied is the explanatory variable and it is denoted by x1. Probability of passing the exam is the response or target variable and it is denoted by z.

If we have one explanatory variable(x1) and one response variable(z), then linear equation would be given mathematically with the following equation-

$$z = \beta_0 + \beta_1 x_1$$

Here, the coefficient $\beta 0$ and $\beta 1$ are the parameters of the model.

If there are multiple explanatory variables, then the above equation can be extended to

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_n x_n$$

Here, the coefficient β_0, β_1, β_2 and β_n are the parameters of the model.

So, the predicted response value is given by the above equations and is denoted by z.

Sigmoid Function

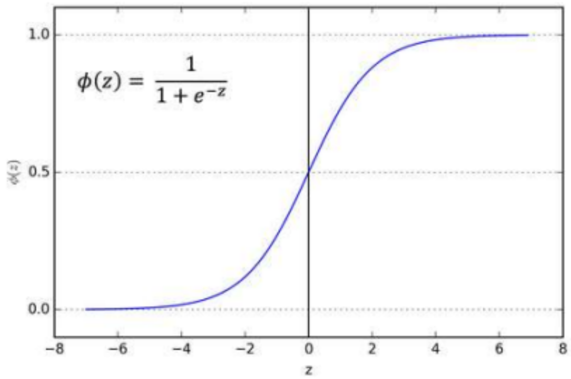
This predicted response value, denoted by z is then converted into a probability value that lie between 0 and 1. We use the sigmoid function in order to map predicted values to probability values. This sigmoid function then maps any real value into a probability value between 0 and 1.

In machine learning, sigmoid function is used to map prediction to probabilities. The sigmoid function has an S shaped curve it is also called sigmoid curve.

A Sigmoid function is a special case of the Logistic function. It is given by the following mathematical formula.

Graphically, we can represent sigmoid function with the following graph.

Sigmoid Function



Decision Boundary

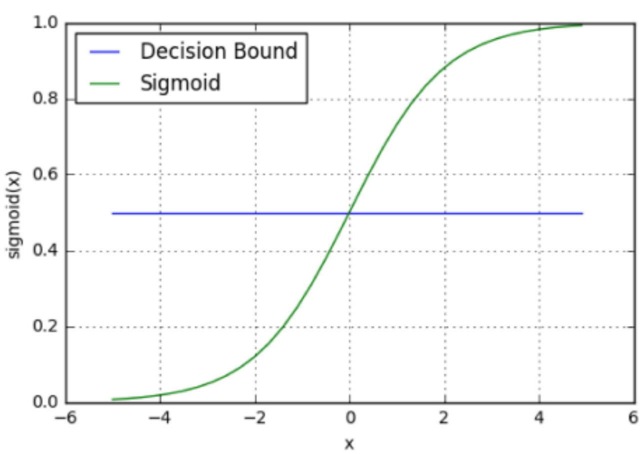
The Sigmoid Function returns a probability between 0 and 1. This probability value is then mapped to a discrete class which is either "0" or "1". In order to map this probability value to a discrete class (pass/fail, yes/no, true/false), we select an threshold value. This threshold value is called Decision Boundary. Above this threshold value, we will map the probability values into class 1 and below which we will map values into class 0.

Mathematically, it can be expressed as follows:-

$$p \geq 0.5 => class = 1$$

$$p \leq 0.5 => class = 1$$

Generally, the decision boundary is set to 0.5, So, if the probability value is 0.8(> 0.5), we will map this observation to class 1.Similarly, if the probability value is 0.2(< 0.5), we will map this observation to class 0. This represent in the graph below -



Making Predictions

Now, we know about sigoid function and decision boundary in logistic regression. We can use our knowledge of sigmoid function and decision boundary to write a prediction function. A prediction function in logistic regression returns the probability of the observation being positive, YES or True. We call this as class 1 and it is denoted as P(class = 1). If the probability inches closert to one, then we will be more confident about our model that the observation is in class 1, otherwise it is in class 0.

3 Assumptions of Logistic Regression

The Logistic Regression model requires several key assumptions. These are as follows:-

1. Logistic Regression model requires the dependent variable to be binary, multinomial or ordinal in nature.

2. It requires the observation to be independent to each other. So, the observation should not come from the repeated measurements.

3. Logistic Regression algorithm requires little or no.multicollinearity among the independent variables. It means that the independent variables should not be too highly correlated with each other.

4. Logistic Regression algorithm requires model assumes linearity of independent variables and log odds.

5. The success of Logistic Regression model depends on the sample size. Typically requires a large sample size to achieve the high accuracy.

4 Types of Logistic Regression

Logistic Regression model can be classified into three groups based on the target variabl categories. These three groups are describe below:-

1. Binary Logistic Regression

In Binary Logistic Regression, the target variable has two possible categories. The common examples of categories are yes or no, good or bad,true or false, spam or no spam and pass or fail.

2. Multinomial Logistic Regression

In Multinomial Logistic Regression, the target variable has three or more categories which are not in any particular order. So, there are three or more nominal categories. The exampples include type of categories of fruits - apple, mango,orange and banana.

3. Ordinal Logistic Regression

In Ordinal Logistic Regression, the target variable has three or more ordinal categories. So, there is intrinsic order involved with the categories. For example, the student performance can be categories as poor, average, good and excellent.

5. Import libraries

In [111…

```
#This python 3 environment comes with many helpful analytics libraries installed
#It is defined by the kaggle/python docker image: https://github.com/kaggle/docker-python
#For example, here's several helpful packages to Load in

import numpy as np #linear algebra
import pandas as pd #data processing,csv file i/o (e.g. pd.read_csv)
import matplotlib.pyplot as plt #data visualization
import seaborn as sns #statistical data visualization
%matplotlib inline

#Input data files are available in the ".../input/"directory
#For example, running this(by clicking run or pressing Shift + Enter) will list all files under the input directory

import os

for dirname, _, filenames in os.walk('/kaggle/input'):
  for filename in filenames:
    print(os.path.join(dirname,filename))

#Any result you write to the current directory are saved as output
```

In [112…

```
import warnings
warnings.filterwarnings('ignore')
```

6.Import dataset

In [113…

```
data = '/content/drive/MyDrive/Module 11/weatherAUS.csv'
df = pd.read_csv(data)
```

7. Explanatory Data Analysis

Now, we will explore the data to gain insights about the data

In [114…

```
#view dimension of the dataset

df.shape
```

Out[114…

```
(142193, 24)
```

We can see that there are 142193 instances and 24 variables in the data set.

In [115…

```
# preview the dataset

df.head()
```

Out[115…

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	...	Humidity3pm
0	2008-12-01	Albury	13.4	22.9	0.6	NaN	NaN	W	44.0	W	...	22.0
1	2008-12-02	Albury	7.4	25.1	0.0	NaN	NaN	WNW	44.0	NNW	...	25.0
2	2008-12-03	Albury	12.9	25.7	0.0	NaN	NaN	WSW	46.0	W	...	30.0
3	2008-12-04	Albury	9.2	28.0	0.0	NaN	NaN	NE	24.0	SE	...	16.0
4	2008-12-05	Albury	17.5	32.3	1.0	NaN	NaN	W	41.0	ENE	...	33.0

5 rows × 24 columns

In [116…

```
col_names = df.columns

col_names
```

Out[116…

```
Index(['Date', 'Location', 'MinTemp', 'MaxTemp', 'Rainfall', 'Evaporation',
       'Sunshine', 'WindGustDir', 'WindGustSpeed', 'WindDir9am', 'WindDir3pm',
       'WindSpeed9am', 'WindSpeed3pm', 'Humidity9am', 'Humidity3pm',
       'Pressure9am', 'Pressure3pm', 'Cloud9am', 'Cloud3pm', 'Temp9am',
       'Temp3pm', 'RainToday', 'RISK_MM', 'RainTomorrow'],
      dtype='object')
```

Drop RISK_MM variable

It is given in the dataset description, that we shoudl drop the `RISK_MM` feature variable from the dataset description. So, we should drop it as follows-

In [117…

```
df.drop(['RISK_MM'],axis =1, inplace = True)
```

In [118…

```
#view summary of dataset

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 142193 entries, 0 to 142192
Data columns (total 23 columns):
 #   Column         Non-Null Count   Dtype  
---  ------         --------------   -----  
 0   Date           142193 non-null  object 
 1   Location       142193 non-null  object 
 2   MinTemp        141556 non-null  float64
 3   MaxTemp        141871 non-null  float64
 4   Rainfall       140787 non-null  float64
 5   Evaporation    81350 non-null   float64
 6   Sunshine       74377 non-null   float64
 7   WindGustDir    132863 non-null  object 
 8   WindGustSpeed  132923 non-null  float64
 9   WindDir9am     132180 non-null  object 
 10  WindDir3pm     138415 non-null  object 
 11  WindSpeed9am   140845 non-null  float64
 12  WindSpeed3pm   139563 non-null  float64
 13  Humidity9am    140419 non-null  float64
 14  Humidity3pm    138583 non-null  float64
 15  Pressure9am    128179 non-null  float64
 16  Pressure3pm    128212 non-null  float64
 17  Cloud9am       88536 non-null   float64
 18  Cloud3pm       85099 non-null   float64
 19  Temp9am        141289 non-null  float64
 20  Temp3pm        139467 non-null  float64
 21  RainToday      140787 non-null  object 
 22  RainTomorrow   142193 non-null  object 
dtypes: float64(16), object(7)
memory usage: 25.0+ MB
```

Types of variables

In this section, I segregate the dataset into categorical and numerical variables. There are a mixture of categorical and numerical variables in the dataset. Categorical variables have data type object. Numerical variables have data type float64.

first of all, we will find categorical variables

In [119…

```
# find categorical variables

categorical = [var for var in df.columns if df[var].dtype =='O']

print('There are {} categorical variables\n'.format(len(categorical)))
print('The categorical variables are:', categorical)
```

```
There are 7 categorical variables

The categorical variables are: ['Date', 'Location', 'WindGustDir', 'WindDir9am', 'WindDir3pm', 'RainToday', 'RainTomorrow']
```

In [120…

```
#view categorical variables

df[categorical].head()
```

Out[120…

	Date	Location	WindGustDir	WindDir9am	WindDir3pm	RainToday	RainTomorrow
0	2008-12-01	Albury	W	W	WNW	No	No
1	2008-12-02	Albury	WNW	NNW	WSW	No	No
2	2008-12-03	Albury	WSW	W	WSW	No	No
3	2008-12-04	Albury	NE	SE	E	No	No
4	2008-12-05	Albury	W	ENE	NW	No	No

Summary of categorical variables

- There is a date variable. It is denoted by `Date` column.
- There are 6 categorical variables. These are given by `Location`, `WindGustDir`, `WindDir9am`, `WindDir3pm`, `RainToday` and `RainTommorow`.
- There are two binary categorical variables - `RainToday` and `RainTomorrow`.
- `RainTomorrow` is the target variable.

Explore problems within categorical variables

First, I will explore the categorical variables.

Missing Values in categorical variables

In [121…

```
#Check missing values in categorical variables

df[categorical].isnull().sum()
```

Out[121…

```
Date                0
Location            0
WindGustDir      9330
WindDir9am      10013
WindDir3pm       3778
RainToday        1406
RainTomorrow        0
dtype: int64
```

In [122…

```
#print categorical variables containing missing values

cat1 = [var for var in categorical if df[var].isnull().sum() != 0]

print(df[cat1].isnull().sum())
```

```
WindGustDir     9330
WindDir9am     10013
WindDir3pm      3778
RainToday       1406
dtype: int64
```

We can see that there are only 4 categorical variables in dataset which contains missing values. These are `WindGustDir`, `WindDir9am`, `WindDir3pm` and `RainToday`.

Frequency counts of categorical variables

Now, I will check the frequency counts of categorical variables.

In [123…

```
# view frequency of categorical variables

for var in categorical:

  print(df[var].value_counts()/float(len(df)))
```

```
Date
2013-12-01    0.000345
2014-01-09    0.000345
2014-01-11    0.000345
2014-01-12    0.000345
2014-01-13    0.000345
                ...
2007-11-29    0.000007
2007-11-28    0.000007
2007-11-27    0.000007
2007-11-26    0.000007
2008-01-31    0.000007
Name: count, Length: 3436, dtype: float64
Location
Canberra            0.024038
Sydney              0.023468
Perth               0.022455
Darwin              0.022448
Hobart              0.022420
Brisbane            0.022230
Adelaide            0.021731
Bendigo             0.021337
Townsville          0.021330
AliceSprings        0.021316
MountGambier        0.021309
Launceston          0.021295
Ballarat            0.021295
Albany              0.021211
Albury              0.021175
PerthAirport        0.021161
MelbourneAirport    0.021161
Mildura             0.021147
SydneyAirport       0.021133
Nuriootpa           0.021112
Sale                0.021098
Watsonia            0.021091
Tuggeranong         0.021084
Portland            0.021070
Woomera             0.021028
Cairns              0.021014
Cobar               0.021014
Wollongong          0.020979
GoldCoast           0.020957
WaggaWagga          0.020929
Penrith             0.020845
NorfolkIsland       0.020845
SalmonGums          0.020782
Newcastle           0.020782
CoffsHarbour        0.020768
Witchcliffe         0.020761
Richmond            0.020753
Dartmoor            0.020697
NorahHead           0.020599
BadgerysCreek       0.020592
MountGinini         0.020444
Moree               0.020071
Walpole             0.019825
PearceRAAF          0.019424
Williamtown         0.017954
Melbourne           0.017125
Nhil                0.011034
Katherine           0.010964
Uluru               0.010697
Name: count, dtype: float64
WindGustDir
W      0.068780
SE     0.065467
E      0.063794
N      0.063526
SSE    0.063245
S      0.062936
WSW    0.062598
SW     0.061867
SSW    0.060552
WNW    0.056726
NW     0.056283
ENE    0.056205
ESE    0.051374
NE     0.049651
NNW    0.046142
```

```
NNE    0.045241
Name: count, dtype: float64
WindDir9am
N      0.080123
SE     0.064434
E      0.063463
SSE    0.063055
NW     0.060144
S      0.059729
W      0.058090
SW     0.057928
NNE    0.055896
NNW    0.055136
ENE    0.054398
ESE    0.053153
NE     0.052935
SSW    0.052380
WNW    0.050593
WSW    0.048125
Name: count, dtype: float64
WindDir3pm
SE     0.074990
W      0.069701
S      0.067500
WSW    0.065608
SW     0.064574
SSE    0.064293
N      0.060952
WNW    0.060875
NW     0.059553
ESE    0.058948
E      0.058667
NE     0.057415
SSW    0.056332
NNW    0.054384
ENE    0.054321
NNE    0.045319
Name: count, dtype: float64
RainToday
No     0.768899
Yes    0.221213
Name: count, dtype: float64
RainTomorrow
No     0.775819
Yes    0.224181
Name: count, dtype: float64
```

Number of labels: cardinality

The number of labels within a categorical variables is known as **cardinality**. A high number of labels within a variable is known as **high cardinality**. High cardinality may pose some serious problems in the machine learning model.So, we will check for high cardinality.

In [124…

```
#check for cardinality in categorical variables

for var in categorical:

  print(var,' contains ', len(df[var].unique()), ' labels')
```

```
Date  contains  3436  labels
Location  contains  49  labels
WindGustDir  contains  17  labels
WindDir9am  contains  17  labels
WindDir3pm  contains  17  labels
RainToday  contains  3  labels
RainTomorrow  contains  2  labels
```

We can see that there is a `Date` variable which needs to be preprocessed. We will do preprocessing in the following section.

All the other variables contain relatively smaller number of variables.

Feature Engineering of Date Variable

In [125…

```
df['Date'].dtypes
```

Out[125…

```
dtype('O')
```

We can see that the data type of `Date` variable is object. We will parse the date currently coded as object into datetime format.

In [126…
```
# parse the dates, currently coded as string, into datetime format

df['Date'] = pd.to_datetime(df['Date'])
```

In [127…
```
#extract year from date

df['Year'] =  df['Date'].dt.year

df['Year'].head()
```

Out[127…
```
0    2008
1    2008
2    2008
3    2008
4    2008
Name: Year, dtype: int32
```

In [128…
```
#extract year from month

df['Month'] =  df['Date'].dt.month

df['Month'].head()
```

Out[128…
```
0    12
1    12
2    12
3    12
4    12
Name: Month, dtype: int32
```

In [129…
```
#extract year from day

df['Day'] =  df['Date'].dt.day

df['Day'].head()
```

Out[129…
```
0    1
1    2
2    3
3    4
4    5
Name: Day, dtype: int32
```

In [130…
```
#again view the summary of Dataset

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 142193 entries, 0 to 142192
Data columns (total 26 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   Date           142193 non-null  datetime64[ns]
 1   Location       142193 non-null  object
 2   MinTemp        141556 non-null  float64
 3   MaxTemp        141871 non-null  float64
 4   Rainfall       140787 non-null  float64
 5   Evaporation    81350 non-null   float64
 6   Sunshine       74377 non-null   float64
 7   WindGustDir    132863 non-null  object
 8   WindGustSpeed  132923 non-null  float64
 9   WindDir9am     132180 non-null  object
 10  WindDir3pm     138415 non-null  object
 11  WindSpeed9am   140845 non-null  float64
 12  WindSpeed3pm   139563 non-null  float64
 13  Humidity9am    140419 non-null  float64
 14  Humidity3pm    138583 non-null  float64
 15  Pressure9am    128179 non-null  float64
 16  Pressure3pm    128212 non-null  float64
 17  Cloud9am       88536 non-null   float64
 18  Cloud3pm       85099 non-null   float64
 19  Temp9am        141289 non-null  float64
 20  Temp3pm        139467 non-null  float64
 21  RainToday      140787 non-null  object
 22  RainTomorrow   142193 non-null  object
 23  Year           142193 non-null  int32
 24  Month          142193 non-null  int32
 25  Day            142193 non-null  int32
dtypes: datetime64[ns](1), float64(16), int32(3), object(6)
memory usage: 26.6+ MB
```

We can see that there are three additional columns created from `Date` variable. Now, I will drop the original `Date` variable from the dataset.

In [131…

```
#drop the original Date variable

df.drop('Date',axis = 1, inplace= True)
```

In [132…

```
#preview the dataset again

df.head()
```

Out[132…

	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	...	Pressure
0	Albury	13.4	22.9	0.6	NaN	NaN	W	44.0	W	WNW	...	1
1	Albury	7.4	25.1	0.0	NaN	NaN	WNW	44.0	NNW	WSW	...	1
2	Albury	12.9	25.7	0.0	NaN	NaN	WSW	46.0	W	WSW	...	1
3	Albury	9.2	28.0	0.0	NaN	NaN	NE	24.0	SE	E	...	1
4	Albury	17.5	32.3	1.0	NaN	NaN	W	41.0	ENE	NW	...	1

5 rows × 25 columns

Now, we can see that the `Date` variable has been removed to the dataset.

Explore Categorical Variables

Now, we will explore the categorical variables one by one.

In [133…

```
#find categorical variables

categorical = [var for var in df.columns if df[var].dtypes == 'O']

print('There are {} categorical variables\n'.format(len(categorical)))

print('The categorical variables are:', categorical)
```

```
There are 6 categorical variables

The categorical variables are: ['Location', 'WindGustDir', 'WindDir9am', 'WindDir3pm', 'RainToday', 'RainTomorrow']
```

We can see that there are 6 categorical variables in the dataset. The `Date` variable has been removed. First, I will check missing values in categorical variables.

In [134…

```
#check for missing values in categorical variables

df[categorical].isnull().sum()
```

Out[134…

```
Location            0
WindGustDir      9330
WindDir9am      10013
WindDir3pm       3778
RainToday        1406
RainTomorrow        0
dtype: int64
```

We can see that `WindGustDir`, `WindDir9am`, `WindDir3pm`, `RainToday` variables contain missing values. We will explore these variables one by one.

Explore `Location` variable

In [135…

```
#print number of labels in Location variable

print('Location contains', len(df.Location.unique()),'labels')
```

```
Location contains 49 labels
```

In [136…

```
#check labels in Location variable

df.Location.unique()
```

Out[136…

```
array(['Albury', 'BadgerysCreek', 'Cobar', 'CoffsHarbour', 'Moree',
       'Newcastle', 'NorahHead', 'NorfolkIsland', 'Penrith', 'Richmond',
       'Sydney', 'SydneyAirport', 'WaggaWagga', 'Williamtown',
       'Wollongong', 'Canberra', 'Tuggeranong', 'MountGinini', 'Ballarat',
       'Bendigo', 'Sale', 'MelbourneAirport', 'Melbourne', 'Mildura',
       'Nhil', 'Portland', 'Watsonia', 'Dartmoor', 'Brisbane', 'Cairns',
       'GoldCoast', 'Townsville', 'Adelaide', 'MountGambier', 'Nuriootpa',
       'Woomera', 'Albany', 'Witchcliffe', 'PearceRAAF', 'PerthAirport',
       'Perth', 'SalmonGums', 'Walpole', 'Hobart', 'Launceston',
       'AliceSprings', 'Darwin', 'Katherine', 'Uluru'], dtype=object)
```

In [137…

```
#check the frequency distribution of values in Location variable

df.Location.value_counts()
```

Out[137…

```
Location
Canberra            3418
Sydney              3337
Perth               3193
Darwin              3192
Hobart              3188
Brisbane            3161
Adelaide            3090
Bendigo             3034
Townsville          3033
AliceSprings        3031
MountGambier        3030
Launceston          3028
Ballarat            3028
Albany              3016
Albury              3011
PerthAirport        3009
MelbourneAirport    3009
Mildura             3007
SydneyAirport       3005
Nuriootpa           3002
Sale                3000
Watsonia            2999
Tuggeranong         2998
Portland            2996
Woomera             2990
Cairns              2988
Cobar               2988
Wollongong          2983
GoldCoast           2980
WaggaWagga          2976
Penrith             2964
NorfolkIsland       2964
SalmonGums          2955
Newcastle           2955
CoffsHarbour        2953
Witchcliffe         2952
Richmond            2951
Dartmoor            2943
NorahHead           2929
BadgerysCreek       2928
MountGinini         2907
Moree               2854
Walpole             2819
PearceRAAF          2762
Williamtown         2553
Melbourne           2435
Nhil                1569
Katherine           1559
Uluru               1521
Name: count, dtype: int64
```

In [138…

```
#Let's Do One Hot Encoding of Location variable
#get k-1 dummy variables after One Hot Encoding
#preview the dataset with head() method

pd.get_dummies(df.Location, drop_first = True).astype(int).head()
```

Out[138…

	Albany	Albury	AliceSprings	BadgerysCreek	Ballarat	Bendigo	Brisbane	Cairns	Canberra	Cobar	...	Townsville	Tuggeranong	Ulur
0	0	1	0	0	0	0	0	0	0	0	...	0	0	
1	0	1	0	0	0	0	0	0	0	0	...	0	0	
2	0	1	0	0	0	0	0	0	0	0	...	0	0	
3	0	1	0	0	0	0	0	0	0	0	...	0	0	
4	0	1	0	0	0	0	0	0	0	0	...	0	0	

5 rows × 48 columns

Explore `WindGustDir` variable

In [139…

```
#print number of labels in WindGustDir variable

print('WindGustDir contains', len(df.WindGustDir.unique()),'labels')
```

```
WindGustDir contains 17 labels
```

In [140…

```
#check labels in WindGustDir variable

df['WindGustDir'].unique()
```

Out[140…

```
array(['W', 'WNW', 'WSW', 'NE', 'NNW', 'N', 'NNE', 'SW', 'ENE', 'SSE',
       'S', 'NW', 'SE', 'ESE', nan, 'E', 'SSW'], dtype=object)
```

In [141…

```
#check frequecy distribution of values in WindGustDir variable

df.WindGustDir.value_counts()
```

Out[141…

```
WindGustDir
W      9780
SE     9309
E      9071
N      9033
SSE    8993
S      8949
WSW    8901
SW     8797
SSW    8610
WNW    8066
NW     8003
ENE    7992
ESE    7305
NE     7060
NNW    6561
NNE    6433
Name: count, dtype: int64
```

In [142…

```
#Let's Do One Hot Encoding of WindGustDir variable
#get k-1 dummy variables after One Hot Encoding
#also add an additional dummy variable to indicate there was missing data
#preview the dataset with head() method

pd.get_dummies(df.Location, drop_first = True, dummy_na = True).astype(int).head()
```

Out[142…

	Albany	Albury	AliceSprings	BadgerysCreek	Ballarat	Bendigo	Brisbane	Cairns	Canberra	Cobar	...	Tuggeranong	Uluru	WaggaW
0	0	1	0	0	0	0	0	0	0	0	...	0	0	
1	0	1	0	0	0	0	0	0	0	0	...	0	0	
2	0	1	0	0	0	0	0	0	0	0	...	0	0	
3	0	1	0	0	0	0	0	0	0	0	...	0	0	
4	0	1	0	0	0	0	0	0	0	0	...	0	0	

5 rows × 49 columns

In [144…

```
#sum the number for 1s per boolean variable over the rows of the dataset
#it will tell us how many observation we have for each category
```

```
pd.get_dummies(df.WindGustDir, drop_first = True, dummy_na = True).astype(int).sum(axis = 0)
```

Out[144…

```
ENE    7992
ESE    7305
N      9033
NE     7060
NNE    6433
NNW    6561
NW     8003
S      8949
SE     9309
SSE    8993
SSW    8610
SW     8797
W      9780
WNW    8066
WSW    8901
NaN    9330
dtype: int64
```

We can see that there are 9330 missing values in WindGustDir variable.

Explore `WindDir9am` variable

In [145…

```
# print number of labels in WindDir9am variable

print('WindDir9am contains', len(df['WindDir9am'].unique()), 'labels')
```

```
WindDir9am contains 17 labels
```

In [146…

```
#check labels in WindDir9am variable

df['WindDir9am'].unique()
```

Out[146…

```
array(['W', 'NNW', 'SE', 'ENE', 'SW', 'SSE', 'S', 'NE', nan, 'SSW', 'N',
       'WSW', 'ESE', 'E', 'NW', 'WNW', 'NNE'], dtype=object)
```

In [147…

```
#check frequency distribution of values  in WindDir9am variable

df['WindDir9am'].value_counts()
```

Out[147…

```
WindDir9am
N      11393
SE      9162
E       9024
SSE     8966
NW      8552
S       8493
W       8260
SW      8237
NNE     7948
NNW     7840
ENE     7735
ESE     7558
NE      7527
SSW     7448
WNW     7194
WSW     6843
Name: count, dtype: int64
```

In [148…

```
#Let's Do One Hot Encoding of WindDir9am variable
#get k-1 dummy variables after One Hot Encoding
#also add an additional dummy variable to indicate there was missing data
#preview the dataset with head() method

pd.get_dummies(df.WindDir9am, drop_first = True, dummy_na = True).astype(int).head()
```

Out[148...

	ENE	ESE	N	NE	NNE	NNW	NW	S	SE	SSE	SSW	SW	W	WNW	WSW	NaN
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
3	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

In [149...

```
#sum the number for 1s per boolean variable over the rows of the dataset
#it will tell us how many observation we have for each category

pd.get_dummies(df.WindDir9am, drop_first = True, dummy_na = True).astype(int).sum(axis = 0)
```

Out[149...

```
ENE     7735
ESE     7558
N      11393
NE      7527
NNE     7948
NNW     7840
NW      8552
S       8493
SE      9162
SSE     8966
SSW     7448
SW      8237
W       8260
WNW     7194
WSW     6843
NaN    10013
dtype: int64
```

Explore `WindDir3pm` variable

In [150...

```
# print number of labels in WindDir3pm variable

print('WindDir3pm contains', len(df['WindDir3pm'].unique()), 'labels')
```

```
WindDir3pm contains 17 labels
```

In [151...

```
#check labels in WindDir3pm variable

df['WindDir3pm'].unique()
```

Out[151...

```
array(['WNW', 'WSW', 'E', 'NW', 'W', 'SSE', 'ESE', 'ENE', 'NNW', 'SSW',
       'SW', 'SE', 'N', 'S', 'NNE', nan, 'NE'], dtype=object)
```

In [152...

```
#check frequency distribution of values  in WindDir3pm variable

df['WindDir3pm'].value_counts()
```

Out[152...

```
WindDir3pm
SE     10663
W       9911
S       9598
WSW     9329
SW      9182
SSE     9142
N       8667
WNW     8656
NW      8468
ESE     8382
E       8342
NE      8164
SSW     8010
NNW     7733
ENE     7724
NNE     6444
Name: count, dtype: int64
```

In [153...

```
#Let's Do One Hot Encoding of WindDir3pm variable
#get k-1 dummy variables after One Hot Encoding
#also add an additional dummy variable to indicate there was missing data
#preview the dataset with head() method
```

```
pd.get_dummies(df.WindDir3pm, drop_first = True, dummy_na = True).astype(int).head()
```

Out[153…

	ENE	ESE	N	NE	NNE	NNW	NW	S	SE	SSE	SSW	SW	W	WNW	WSW	NaN
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0

In [154…

```
#sum the number for 1s per boolean variable over the rows of the dataset
#it will tell us how many observation we have for each category

pd.get_dummies(df.WindDir3pm, drop_first = True, dummy_na = True).astype(int).sum(axis = 0)
```

Out[154…

```
ENE     7724
ESE     8382
N       8667
NE      8164
NNE     6444
NNW     7733
NW      8468
S       9598
SE     10663
SSE     9142
SSW     8010
SW      9182
W       9911
WNW     8656
WSW     9329
NaN     3778
dtype: int64
```

There are 3778 missing values in the `WindDir3pm` variable.

Explore `RainToday` variable

In [155…

```
# print number of labels in RainToday variable

print('RainToday contains', len(df['RainToday'].unique()), 'labels')
```

```
RainToday contains 3 labels
```

In [156…

```
#check labels in RainToday variable

df['RainToday'].unique()
```

Out[156…

```
array(['No', 'Yes', nan], dtype=object)
```

In [157…

```
#check frequency distribution of values  in RainToday variable

df['RainToday'].value_counts()
```

Out[157…

```
RainToday
No     109332
Yes     31455
Name: count, dtype: int64
```

In [158…

```
#Let's Do One Hot Encoding of  RainToday variable
#get k-1 dummy variables after One Hot Encoding
#also add an additional dummy variable to indicate there was missing data
#preview the dataset with head() method

pd.get_dummies(df. RainToday, drop_first = True, dummy_na = True).astype(int).head()
```

Out[158...

	Yes	NaN
0	0	0
1	0	0
2	0	0
3	0	0
4	0	0

In [159...

```
#sum the number for 1s per boolean variable over the rows of the dataset
#it will tell us how many observation we have for each category

pd.get_dummies(df.RainToday, drop_first = True, dummy_na = True).astype(int).sum(axis = 0)
```

Out[159...

```
Yes    31455
NaN     1406
dtype: int64
```

There are 1406 missing values in the `RainToday` variable.

Explore Numerical Variables

In [160...

```
#find numerical variables

numerical = [var for var in df.columns if df[var].dtypes != 'O']

print( 'There are {} numerical variables\n'.format(len(numerical)))

print('The numerical variables are: ',numerical)
```

```
There are 19 numerical variables

The numerical variables are:  ['MinTemp', 'MaxTemp', 'Rainfall', 'Evaporation', 'Sunshine', 'WindGustSpeed', 'WindSpeed9am', 'W
indSpeed3pm', 'Humidity9am', 'Humidity3pm', 'Pressure9am', 'Pressure3pm', 'Cloud9am', 'Cloud3pm', 'Temp9am', 'Temp3pm', 'Year',
'Month', 'Day']
```

In [161...

```
#view the numerical values

df[numerical].head()
```

Out[161...

	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm
0	13.4	22.9	0.6	NaN	NaN	44.0	20.0	24.0	71.0	22.0
1	7.4	25.1	0.0	NaN	NaN	44.0	4.0	22.0	44.0	25.0
2	12.9	25.7	0.0	NaN	NaN	46.0	19.0	26.0	38.0	30.0
3	9.2	28.0	0.0	NaN	NaN	24.0	11.0	9.0	45.0	16.0
4	17.5	32.3	1.0	NaN	NaN	41.0	7.0	20.0	82.0	33.0

Summary of numerical variables

- There are 16 numerical variables.
- These are given by `MinTemp`, `MaxTemp`, `Rainfall`, `Evaporation`, `Sunshine`, `WindGustSpeed`, `WindSpeed9am`, `WindSpeed3pm`, `Humidity9am`, `Humidity3pm`, `Pressure9am`, `Pressure3pm`, `Cloud9am`, `Cloud3pm`, `Temp9am`, and `Temp3pm`.
- All of the numerical variable are of continuous type.

Explore problems within numerical variables

Now, We will explore the numerical variables

Missing values in numerical variables

In [162...

```
#check missing values in numerical variables

df[numerical].isnull().sum()
```

Out[162…
```
MinTemp            637
MaxTemp            322
Rainfall          1406
Evaporation      60843
Sunshine         67816
WindGustSpeed     9270
WindSpeed9am      1348
WindSpeed3pm      2630
Humidity9am       1774
Humidity3pm       3610
Pressure9am      14014
Pressure3pm      13981
Cloud9am         53657
Cloud3pm         57094
Temp9am            904
Temp3pm           2726
Year                 0
Month                0
Day                  0
dtype: int64
```

We can see that all the 16 numerical variables contain missing values.

Outliers in numerical variables

In [163…
```python
# view summary statistics in numerical variables

print(round(df[numerical].describe()),2)
```

```
        MinTemp   MaxTemp  Rainfall  Evaporation  Sunshine  WindGustSpeed  \
count  141556.0  141871.0  140787.0      81350.0   74377.0       132923.0
mean       12.0      23.0       2.0          5.0       8.0           40.0
std         6.0       7.0       8.0          4.0       4.0           14.0
min        -8.0      -5.0       0.0          0.0       0.0            6.0
25%         8.0      18.0       0.0          3.0       5.0           31.0
50%        12.0      23.0       0.0          5.0       8.0           39.0
75%        17.0      28.0       1.0          7.0      11.0           48.0
max        34.0      48.0     371.0        145.0      14.0          135.0

       WindSpeed9am  WindSpeed3pm  Humidity9am  Humidity3pm  Pressure9am  \
count      140845.0      139563.0     140419.0     138583.0     128179.0
mean           14.0          19.0         69.0         51.0       1018.0
std             9.0           9.0         19.0         21.0          7.0
min             0.0           0.0          0.0          0.0        980.0
25%             7.0          13.0         57.0         37.0       1013.0
50%            13.0          19.0         70.0         52.0       1018.0
75%            19.0          24.0         83.0         66.0       1022.0
max           130.0          87.0        100.0        100.0       1041.0

       Pressure3pm  Cloud9am  Cloud3pm   Temp9am   Temp3pm      Year  \
count     128212.0   88536.0   85099.0  141289.0  139467.0  142193.0
mean        1015.0       4.0       5.0      17.0      22.0    2013.0
std            7.0       3.0       3.0       6.0       7.0       3.0
min          977.0       0.0       0.0      -7.0      -5.0    2007.0
25%         1010.0       1.0       2.0      12.0      17.0    2011.0
50%         1015.0       5.0       5.0      17.0      21.0    2013.0
75%         1020.0       7.0       7.0      22.0      26.0    2015.0
max         1040.0       9.0       9.0      40.0      47.0    2017.0

          Month       Day
count  142193.0  142193.0
mean        6.0      16.0
std         3.0       9.0
min         1.0       1.0
25%         3.0       8.0
50%         6.0      16.0
75%         9.0      23.0
max        12.0      31.0   2
```

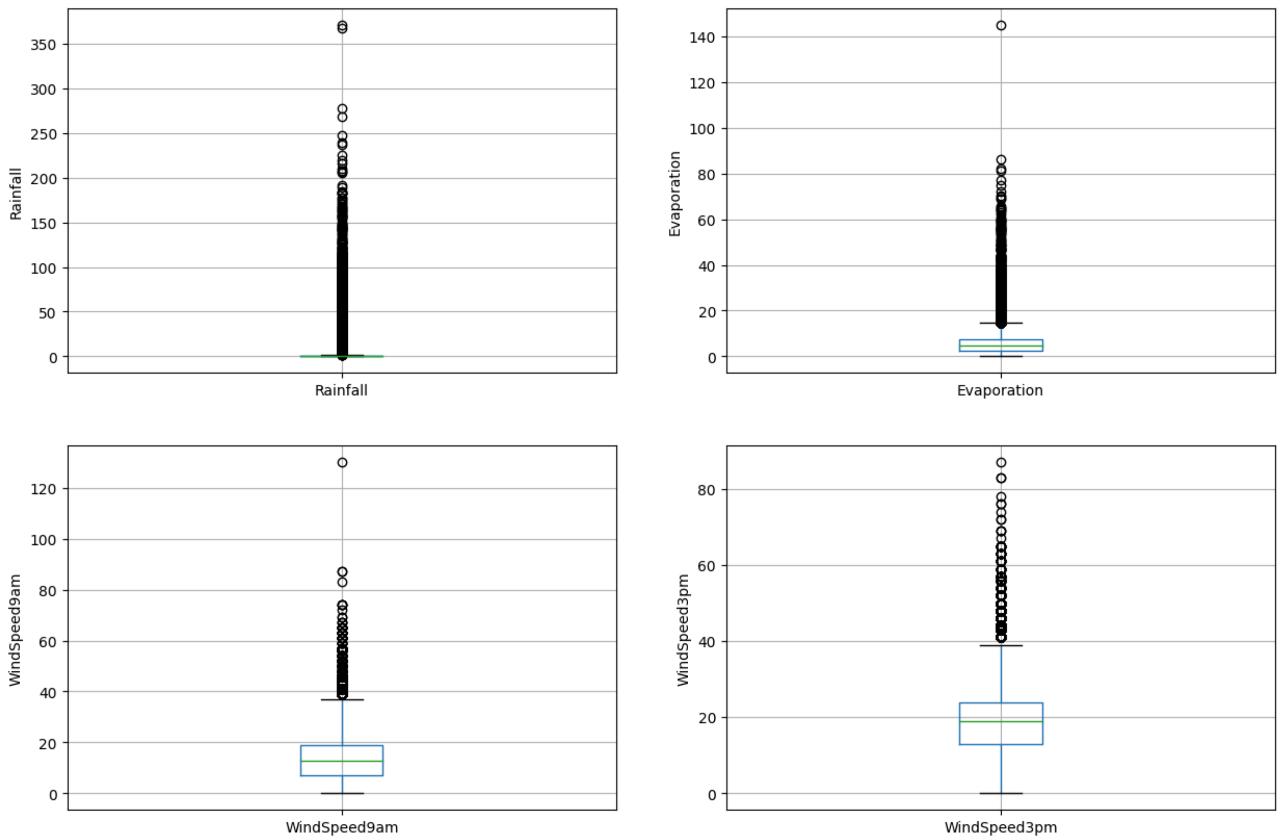
On closer inspection, we can see that the `Rainfall`, `Evaporation`, `WindSpeed9am` and `WindSpeed3pm` columns may contain outliers.

We will draw boxplot to visualize outliers in above variables

In [164…
```python
# draw boxplot to visualize outliers

plt.figure(figsize=(15,10))

plt.subplot(2,2,1)
fig = df.boxplot(column = 'Rainfall')
```

```
fig.set_title('')
fig.set_ylabel('Rainfall')

plt.subplot(2,2,2)
fig = df.boxplot(column = 'Evaporation')
fig.set_title('')
fig.set_ylabel('Evaporation')

plt.subplot(2,2,3)
fig = df.boxplot(column = 'WindSpeed9am')
fig.set_title('')
fig.set_ylabel('WindSpeed9am')

plt.subplot(2,2,4)
fig = df.boxplot(column = 'WindSpeed3pm')
fig.set_title('')
fig.set_ylabel('WindSpeed3pm')
```

Out[164… Text(0, 0.5, 'WindSpeed3pm')



The above boxplot confirm that there are lot of outliers in these variables.

Check the distribution of variables

Now, I will plot the histogram to check distribution to find out if they are normal of skewed. If the variable follows normal distribution, then we will do `Extreme Value Analysis` otherwise if they are skewed, We will find IQR (Interquartile range)

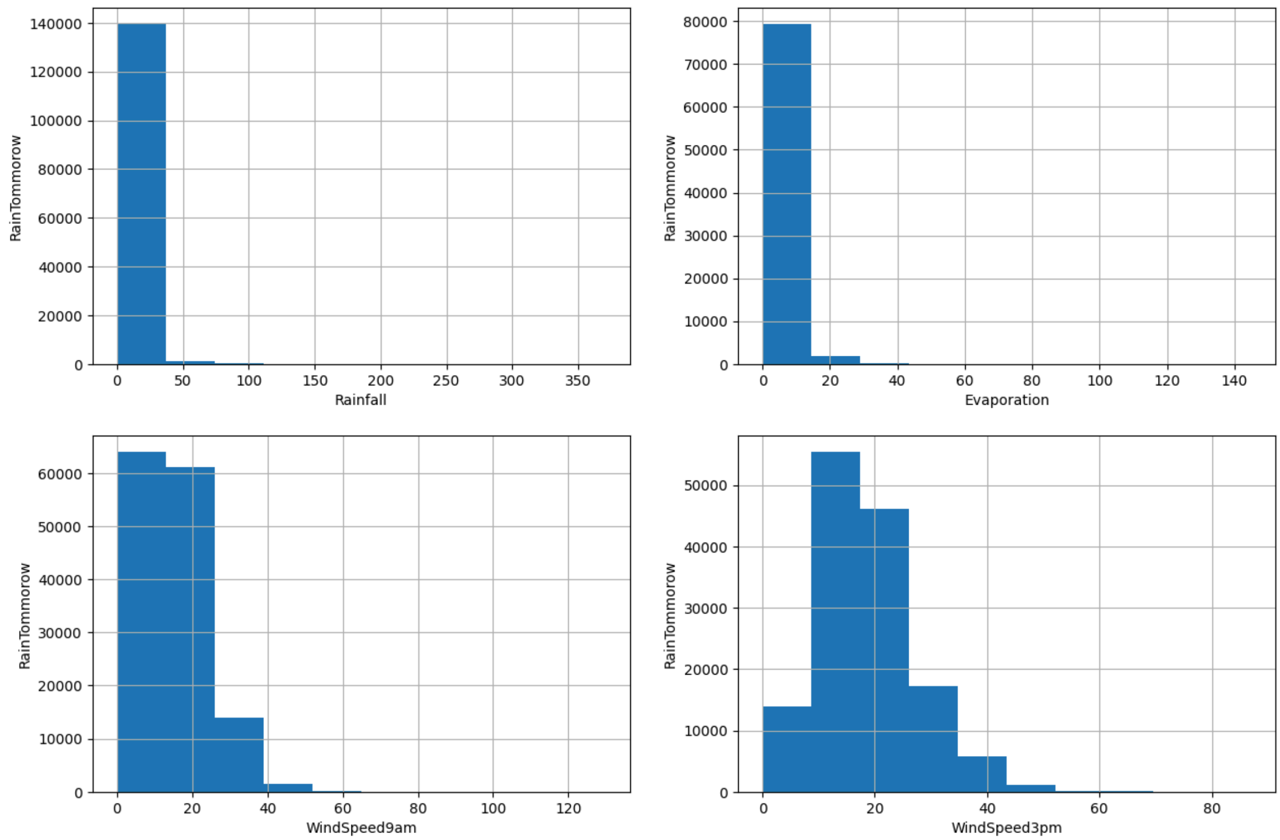
In [165…

```
#plot histogram to check distribution

plt.figure(figsize=(15,10))

plt.subplot(2,2,1)
fig = df.Rainfall.hist(bins=10)
fig.set_xlabel('Rainfall')
fig.set_ylabel('RainTommorow')

plt.subplot(2,2,2)
fig = df.Evaporation.hist(bins=10)
fig.set_xlabel('Evaporation')
fig.set_ylabel('RainTommorow')

plt.subplot(2,2,3)
fig = df.WindSpeed9am.hist(bins=10)
```

```
fig.set_xlabel('WindSpeed9am')
fig.set_ylabel('RainTommorow')

plt.subplot(2,2,4)
fig = df.WindSpeed3pm.hist(bins=10)
fig.set_xlabel('WindSpeed3pm')
fig.set_ylabel('RainTommorow')
```

Out[165…

```
Text(0, 0.5, 'RainTommorow')
```



We can see that all the four variables are skewed. So, we will use interquartile range to find outliers.

In [166…

```
#find outliers for Rainfall variable

IQR = df.Rainfall.quantile(0.75)- df.Rainfall.quantile(0.25)
Lower_fence = df.Rainfall.quantile(0.25)-(IQR*3)
Upper_fence = df.Rainfall.quantile(0.75)+(IQR*3)

print('Rainfall outliers are values < {lowerboundary} or > {upperboundary}'.format(lowerboundary=Lower_fence, upperboundary =U
```

```
Rainfall outliers are values < -2.4000000000000004 or > 3.2
```

For `Rainfall`, the minimum and maximum values are 0.0 and 371.0. So, the outliers are values > 3.2.

In [167…

```
#find outliers for Evaporation variable

IQR = df.Evaporation.quantile(0.75)- df.Evaporation.quantile(0.25)
Lower_fence = df.Evaporation.quantile(0.25)-(IQR*3)
Upper_fence = df.Evaporation.quantile(0.75)+(IQR*3)

print('Evaporation outliers are values < {lowerboundary} or > {upperboundary}'.format(lowerboundary=Lower_fence, upperboundary
```

```
Evaporation outliers are values < -11.800000000000002 or > 21.800000000000004
```

For `Evaporation`, the minimum and maximum values are 0.0 and 145.0. So, the outliers are values > 21.8.

In [168…

```
#find outliers for WindSpeed9am variable

IQR = df.WindSpeed9am.quantile(0.75)- df.WindSpeed9am.quantile(0.25)
Lower_fence = df.WindSpeed9am.quantile(0.25)-(IQR*3)
Upper_fence = df.WindSpeed9am.quantile(0.75)+(IQR*3)

print('WindSpeed9am outliers are values < {lowerboundary} or > {upperboundary}'.format(lowerboundary=Lower_fence, upperboundar
```

```
WindSpeed9am outliers are values < -29.0 or > 55.0
```

For `WindSpeed9am`, the minimum and maximum values are 0.0 and 130.0. So, the outliers are values > 55.0.

In [169…

```
#find outliers for WindSpeed3pm variable

IQR = df.WindSpeed3pm.quantile(0.75)- df.WindSpeed3pm.quantile(0.25)
Lower_fence = df.WindSpeed3pm.quantile(0.25)-(IQR*3)
Upper_fence = df.WindSpeed3pm.quantile(0.75)+(IQR*3)

print('WindSpeed3pm outliers are values < {lowerboundary} or > {upperboundary}'.format(lowerboundary=Lower_fence, upperboundar
```

```
WindSpeed3pm outliers are values < -20.0 or > 57.0
```

For `WindSpeed3pm`, the minimum and maximum values are 0.0 and 87.0. So, the outliers are values > 57.0.

8.Declare feature vector and target variable

In [170…

```
X = df.drop(['RainTomorrow'],axis = 1)
y = df['RainTomorrow']
```

9.Split data into separate training and test set

In [171…

```
#split X and y into training and testing sets

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)
```

In [172…

```
#check the shape of  X_train and X_test

X_train.shape, X_test.shape
```

Out[172…

```
((113754, 24), (28439, 24))
```

10. Feature Engineering

Feature Engineering is the process of transforming raw data into useful features that help us to understand our model better and increase its predictive power. We will carry out feature engineering on different types of variables.

First, we will display the categorical and numerical variables again separately.

In [173…

```
#check data types in X_train

X_train.dtypes
```

Out[173…

```
Location          object
MinTemp          float64
MaxTemp          float64
Rainfall         float64
Evaporation      float64
Sunshine         float64
WindGustDir       object
WindGustSpeed    float64
WindDir9am        object
WindDir3pm        object
WindSpeed9am     float64
WindSpeed3pm     float64
Humidity9am      float64
Humidity3pm      float64
Pressure9am      float64
Pressure3pm      float64
Cloud9am         float64
Cloud3pm         float64
Temp9am          float64
Temp3pm          float64
RainToday         object
Year               int32
Month              int32
Day                int32
dtype: object
```

In [174…

```
#display categorical variables

categorical = [ col for col in X_train.columns if X_train[col].dtypes == 'O']
```

```
categorical
```

Out[174…

```
['Location', 'WindGustDir', 'WindDir9am', 'WindDir3pm', 'RainToday']
```

In [175…

```
#display numerical variables

numerical = [ col for col in X_train.columns if X_train[col].dtypes != 'O']

numerical
```

Out[175…

```
['MinTemp',
 'MaxTemp',
 'Rainfall',
 'Evaporation',
 'Sunshine',
 'WindGustSpeed',
 'WindSpeed9am',
 'WindSpeed3pm',
 'Humidity9am',
 'Humidity3pm',
 'Pressure9am',
 'Pressure3pm',
 'Cloud9am',
 'Cloud3pm',
 'Temp9am',
 'Temp3pm',
 'Year',
 'Month',
 'Day']
```

Engineering missing values in numerical variables

In [176…

```
#check missing values in numerical variables in X_train

X_train[numerical].isnull().sum()
```

Out[176…

```
MinTemp            495
MaxTemp            264
Rainfall          1139
Evaporation      48718
Sunshine         54314
WindGustSpeed     7367
WindSpeed9am      1086
WindSpeed3pm      2094
Humidity9am       1449
Humidity3pm       2890
Pressure9am      11212
Pressure3pm      11186
Cloud9am         43137
Cloud3pm         45768
Temp9am            740
Temp3pm           2171
Year                 0
Month                0
Day                  0
dtype: int64
```

In [177…

```
#check missing values in numerical variables in  X_test

X_test[numerical].isnull().sum()
```

Out[177…
```
MinTemp            142
MaxTemp             58
Rainfall           267
Evaporation      12125
Sunshine         13502
WindGustSpeed     1903
WindSpeed9am       262
WindSpeed3pm       536
Humidity9am        325
Humidity3pm        720
Pressure9am       2802
Pressure3pm       2795
Cloud9am         10520
Cloud3pm         11326
Temp9am            164
Temp3pm            555
Year                 0
Month                0
Day                  0
dtype: int64
```

In [178…
```
#print percentage of missing values in numerical variables in training set

for col in numerical:
  if X_train[col].isnull().mean() > 0:
    print(col, round(X_train[col].isnull().mean(),4))
```

```
MinTemp 0.0044
MaxTemp 0.0023
Rainfall 0.01
Evaporation 0.4283
Sunshine 0.4775
WindGustSpeed 0.0648
WindSpeed9am 0.0095
WindSpeed3pm 0.0184
Humidity9am 0.0127
Humidity3pm 0.0254
Pressure9am 0.0986
Pressure3pm 0.0983
Cloud9am 0.3792
Cloud3pm 0.4023
Temp9am 0.0065
Temp3pm 0.0191
```

Assumption

We assume that the data are missing completely at random (MCAR). There are two method which can be used to impute missing values. One is mean or median imputation and other on is random sample imputation. When there are outliers in the dataset, we should use median imputation. So, we will use median imputation because median imputation is robust to outliers.

We will impute missing values with appropriate statistical measures of the data, in this case median. Imputation should be done over the training set, and then propagated to the test set.It means that the statistical measures to be used to fill missing values both train and test set, should be extracted from the train set only. This is to avoid overfitting.

In [179…
```
# impute missing values in X_train and X_test with respective column median in X_train

for df1 in [X_train, X_test] :
  for col in numerical:
    col_median = X_train[col].median()
    df1[col].fillna(col_median, inplace = True)
```

In [180…
```
# check again missing values in numerical variables in X_train

X_train[numerical].isnull().sum()
```

Out[180…

```
MinTemp          0
MaxTemp          0
Rainfall         0
Evaporation      0
Sunshine         0
WindGustSpeed    0
WindSpeed9am     0
WindSpeed3pm     0
Humidity9am      0
Humidity3pm      0
Pressure9am      0
Pressure3pm      0
Cloud9am         0
Cloud3pm         0
Temp9am          0
Temp3pm          0
Year             0
Month            0
Day              0
dtype: int64
```

In [181…

```
# check missing values in numerical variables in X_test

X_test[numerical].isnull().sum()
```

Out[181…

```
MinTemp          0
MaxTemp          0
Rainfall         0
Evaporation      0
Sunshine         0
WindGustSpeed    0
WindSpeed9am     0
WindSpeed3pm     0
Humidity9am      0
Humidity3pm      0
Pressure9am      0
Pressure3pm      0
Cloud9am         0
Cloud3pm         0
Temp9am          0
Temp3pm          0
Year             0
Month            0
Day              0
dtype: int64
```

Now, we can see that there are no missing values in numerical columns of training and test set.

Engineering missing values in categorical variables

In [182…

```
# print percentage of missing values in categorical variables in training set

X_train[categorical].isnull().mean()
```

Out[182…

```
Location       0.000000
WindGustDir    0.065114
WindDir9am     0.070134
WindDir3pm     0.026443
RainToday      0.010013
dtype: float64
```

In [183…

```
# print categorical variables with missing data

for col in categorical:
  if X_train[col].isnull().mean() > 0:
    print(col, (X_train[col].isnull().mean()))
```

```
WindGustDir 0.06511419378659213
WindDir9am 0.07013379749283542
WindDir3pm 0.026443026179299188
RainToday 0.01001283471350458
```

In [184…

```
# impute missing categorical variables with most frequent value

for df2 in [X_train, X_test]:
  df2['WindGustDir'].fillna(X_train['WindGustDir'].mode()[0],inplace =True)
  df2['WindDir9am'].fillna(X_train['WindDir9am'].mode()[0],inplace =True)
  df2['WindDir3pm'].fillna(X_train['WindDir3pm'].mode()[0],inplace =True)
  df2['RainToday'].fillna(X_train['RainToday'].mode()[0],inplace =True)
```

In [185…

```
# check missing values in categorical variables in X_train

X_train[categorical].isnull().sum()
```

Out[185…

```
Location       0
WindGustDir    0
WindDir9am     0
WindDir3pm     0
RainToday      0
dtype: int64
```

In [186…

```
#check missing values in categorical variables in x_test

X_test[categorical].isnull().sum()
```

Out[186…

```
Location       0
WindGustDir    0
WindDir9am     0
WindDir3pm     0
RainToday      0
dtype: int64
```

As a final check, I will checl for missing values in X_train and X_test.

In [187…

```
# check missing values in X_train
X_train.isnull().sum()
```

Out[187…

```
Location         0
MinTemp          0
MaxTemp          0
Rainfall         0
Evaporation      0
Sunshine         0
WindGustDir      0
WindGustSpeed    0
WindDir9am       0
WindDir3pm       0
WindSpeed9am     0
WindSpeed3pm     0
Humidity9am      0
Humidity3pm      0
Pressure9am      0
Pressure3pm      0
Cloud9am         0
Cloud3pm         0
Temp9am          0
Temp3pm          0
RainToday        0
Year             0
Month            0
Day              0
dtype: int64
```

In [188…

```
# check missing values in X_test

X_test.isnull().sum()
```

Out[188…
```
Location         0
MinTemp          0
MaxTemp          0
Rainfall         0
Evaporation      0
Sunshine         0
WindGustDir      0
WindGustSpeed    0
WindDir9am       0
WindDir3pm       0
WindSpeed9am     0
WindSpeed3pm     0
Humidity9am      0
Humidity3pm      0
Pressure9am      0
Pressure3pm      0
Cloud9am         0
Cloud3pm         0
Temp9am          0
Temp3pm          0
RainToday        0
Year             0
Month            0
Day              0
dtype: int64
```

We can see that there are no missing values in X_train and X_test.

Engineering outliers in numerical variables

We have seen that the `Rainfall`, `Evaporation`, `WindSpeed9am` and `WindSpeed3pm` columns contain outliers. We will use top-coding approach to cap maximu values and remove outliers from the above variables.

In [189…
```
def max_value(df3, variable, top):
  return np.where(df3[variable]>top, top, df3[variable])

for df3 in [X_train, X_test]:
  df3['Rainfall'] = max_value(df3,'Rainfall', 3.2)
  df3['Evaporation'] = max_value(df3,'Evaporation', 21.8)
  df3['WindSpeed9am'] = max_value(df3,'WindSpeed9am', 55)
  df3['WindSpeed3pm'] = max_value(df3,'WindSpeed3pm', 57)
```

In [190…
```
X_train.Rainfall.max(), X_test.Rainfall.max()
```

Out[190…
```
(3.2, 3.2)
```

In [191…
```
X_train.Evaporation.max(), X_test.Evaporation.max()
```

Out[191…
```
(21.8, 21.8)
```

In [192…
```
X_train.WindSpeed9am.max(), X_test.WindSpeed9am.max()
```

Out[192…
```
(55.0, 55.0)
```

In [193…
```
X_train.WindSpeed3pm.max(), X_test.WindSpeed3pm.max()
```

Out[193…
```
(57.0, 57.0)
```

In [194…
```
X_train[numerical].describe()
```

Out[194...

	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Hu
count	113754.000000	113754.000000	113754.000000	113754.000000	113754.000000	113754.000000	113754.000000	113754.000000	113
mean	12.193497	23.237216	0.675080	5.151606	8.041154	39.884074	13.978155	18.614756	
std	6.388279	7.094149	1.183837	2.823707	2.769480	13.116959	8.806558	8.685862	
min	-8.200000	-4.800000	0.000000	0.000000	0.000000	6.000000	0.000000	0.000000	
25%	7.600000	18.000000	0.000000	4.000000	8.200000	31.000000	7.000000	13.000000	
50%	12.000000	22.600000	0.000000	4.800000	8.500000	39.000000	13.000000	19.000000	
75%	16.800000	28.200000	0.600000	5.400000	8.700000	46.000000	19.000000	24.000000	
max	33.900000	48.100000	3.200000	21.800000	14.500000	135.000000	55.000000	57.000000	

We can now see that the outliers in `Rainfall`, `Evaporation`, `WindSpeed9am`, and `WindSpeed3pm` columns are all capped.

Encode categorical variables

In [195...

```
categorical
```

Out[195...

```
['Location', 'WindGustDir', 'WindDir9am', 'WindDir3pm', 'RainToday']
```

In [196...

```
X_train[categorical].head()
```

Out[196...

	Location	WindGustDir	WindDir9am	WindDir3pm	RainToday
110803	Witchcliffe	S	SSE	S	No
87289	Cairns	ENE	SSE	SE	Yes
134949	AliceSprings	E	NE	N	No
85553	Cairns	ESE	SSE	E	No
16110	Newcastle	W	N	SE	No

In [197...

```
!pip install category_encoders
```

```
Requirement already satisfied: category_encoders in /usr/local/lib/python3.10/dist-packages (2.6.3)
Requirement already satisfied: numpy>=1.14.0 in /usr/local/lib/python3.10/dist-packages (from category_encoders) (1.25.2)
Requirement already satisfied: scikit-learn>=0.20.0 in /usr/local/lib/python3.10/dist-packages (from category_encoders) (1.2.2)
Requirement already satisfied: scipy>=1.0.0 in /usr/local/lib/python3.10/dist-packages (from category_encoders) (1.11.4)
Requirement already satisfied: statsmodels>=0.9.0 in /usr/local/lib/python3.10/dist-packages (from category_encoders) (0.14.2)
Requirement already satisfied: pandas>=1.0.5 in /usr/local/lib/python3.10/dist-packages (from category_encoders) (2.0.3)
Requirement already satisfied: patsy>=0.5.1 in /usr/local/lib/python3.10/dist-packages (from category_encoders) (0.5.6)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.10/dist-packages (from pandas>=1.0.5->category_
encoders) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas>=1.0.5->category_encoders)
(2023.4)
Requirement already satisfied: tzdata>=2022.1 in /usr/local/lib/python3.10/dist-packages (from pandas>=1.0.5->category_encoder
s) (2024.1)
Requirement already satisfied: six in /usr/local/lib/python3.10/dist-packages (from patsy>=0.5.1->category_encoders) (1.16.0)
Requirement already satisfied: joblib>=1.1.1 in /usr/local/lib/python3.10/dist-packages (from scikit-learn>=0.20.0->category_en
coders) (1.4.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.10/dist-packages (from scikit-learn>=0.20.0->cate
gory_encoders) (3.4.0)
Requirement already satisfied: packaging>=21.3 in /usr/local/lib/python3.10/dist-packages (from statsmodels>=0.9.0->category_en
coders) (24.0)
```

In [198...

```
#encode RainToday variable

import category_encoders as ce

encoder = ce.BinaryEncoder(cols=['RainToday'])

X_train = encoder.fit_transform(X_train)

X_test = encoder.transform(X_test)
```

In [199...

```
X_train.head()
```

Out[199…

	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	...
110803	Witchcliffe	13.9	22.6	0.2	4.8	8.5	S	41.0	SSE	S	...
87289	Cairns	22.4	29.4	2.0	6.0	6.3	ENE	33.0	SSE	SE	...
134949	AliceSprings	9.7	36.2	0.0	11.4	12.3	E	31.0	NE	N	...
85553	Cairns	20.5	30.1	0.0	8.8	11.1	ESE	37.0	SSE	E	...
16110	Newcastle	16.8	29.2	0.0	4.8	8.5	W	39.0	N	SE	...

5 rows × 25 columns

We can see that two additional variables `RainToday_0` and `RainToday_1` are created from `RainToday` variable

Now, We will create the `X_train` training set.

In [200…

```
X_train = pd.concat([X_train[numerical], X_train[['RainToday_0', 'RainToday_1']],
                     pd.get_dummies(X_train.Location),
                     pd.get_dummies(X_train.WindGustDir),
                     pd.get_dummies(X_train.WindDir9am),
                     pd.get_dummies(X_train.WindDir3pm)], axis = 1)
```

In [201…

```
X_train.head()
```

Out[201…

	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity
110803	13.9	22.6	0.2	4.8	8.5	41.0	20.0	28.0	65.0	
87289	22.4	29.4	2.0	6.0	6.3	33.0	7.0	19.0	71.0	
134949	9.7	36.2	0.0	11.4	12.3	31.0	15.0	11.0	6.0	
85553	20.5	30.1	0.0	8.8	11.1	37.0	22.0	19.0	59.0	
16110	16.8	29.2	0.0	4.8	8.5	39.0	0.0	7.0	72.0	

5 rows × 118 columns

Similarly, We will create the `X_test` testing set.

In [202…

```
X_test = pd.concat([X_test[numerical], X_test[['RainToday_0', 'RainToday_1']],
                     pd.get_dummies(X_test.Location),
                     pd.get_dummies(X_test.WindGustDir),
                     pd.get_dummies(X_test.WindDir9am),
                     pd.get_dummies(X_test.WindDir3pm)], axis = 1)
```

In [203…

```
X_test.head()
```

Out[203…

	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity
86232	17.4	29.0	0.0	3.6	11.1	33.0	11.0	19.0	63.0	
57576	6.8	14.4	0.8	0.8	8.5	46.0	17.0	22.0	80.0	
124071	10.1	15.4	3.2	4.8	8.5	31.0	13.0	9.0	70.0	
117955	14.4	33.4	0.0	8.0	11.6	41.0	9.0	17.0	40.0	
133468	6.8	14.3	3.2	0.2	7.3	28.0	15.0	13.0	92.0	

5 rows × 118 columns

We now have training and testing set ready for model building. Before that, we should map all the feature variables onto the same scale. It is called `feature scaling`. We will do it as follows.

11. Feature Scaling

In [205…

```
X_train.describe()
```

Out[205...

	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Hu
count	113754.000000	113754.000000	113754.000000	113754.000000	113754.000000	113754.000000	113754.000000	113754.000000	113
mean	12.193497	23.237216	0.675080	5.151606	8.041154	39.884074	13.978155	18.614756	
std	6.388279	7.094149	1.183837	2.823707	2.769480	13.116959	8.806558	8.685862	
min	-8.200000	-4.800000	0.000000	0.000000	0.000000	6.000000	0.000000	0.000000	
25%	7.600000	18.000000	0.000000	4.000000	8.200000	31.000000	7.000000	13.000000	
50%	12.000000	22.600000	0.000000	4.800000	8.500000	39.000000	13.000000	19.000000	
75%	16.800000	28.200000	0.600000	5.400000	8.700000	46.000000	19.000000	24.000000	
max	33.900000	48.100000	3.200000	21.800000	14.500000	135.000000	55.000000	57.000000	

8 rows × 21 columns

In [206...

```
cols = X_train.columns
```

In [207...

```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()

X_train =scaler.fit_transform(X_train)

X_test = scaler.transform(X_test)
```

In [208...

```
X_train = pd.DataFrame(X_train, columns = [cols])
```

In [209...

```
X_test = pd.DataFrame(X_test, columns = [cols])
```

In [210...

```
X_train.describe()
```

Out[210...

	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Hu
count	113754.000000	113754.000000	113754.000000	113754.000000	113754.000000	113754.000000	113754.000000	113754.000000	113
mean	0.484406	0.530004	0.210962	0.236312	0.554562	0.262667	0.254148	0.326575	
std	0.151741	0.134105	0.369949	0.129528	0.190999	0.101682	0.160119	0.152384	
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.375297	0.431002	0.000000	0.183486	0.565517	0.193798	0.127273	0.228070	
50%	0.479810	0.517958	0.000000	0.220183	0.586207	0.255814	0.236364	0.333333	
75%	0.593824	0.623819	0.187500	0.247706	0.600000	0.310078	0.345455	0.421053	
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	

8 rows × 118 columns

We now have `X_train` dataset ready to be fed into the Logistic Regression classifier. We will do it as follows.

12. Model training

In [211...

```
#train a logistic regression model on the training set
from sklearn.linear_model import LogisticRegression

#instatiate the model
logreg = LogisticRegression(solver='liblinear', random_state=0)

#fit the model
logreg.fit(X_train, y_train)
```

Out[211...

```
LogisticRegression
LogisticRegression(random_state=0, solver='liblinear')
```

13. Predict Results

In [212…

```
y_pred_test = logreg.predict(X_test)

y_pred_test
```

Out[212…

```
array(['No', 'No', 'No', ..., 'No', 'No', 'Yes'], dtype=object)
```

predict_proba method

predict_proba method gives the probabilities for the target variable(0 and 1) in this case, in array form.

`0 is for probability for no rain` and `1 is for probability for rain.`

In [213…

```
#probability of getting output as 0 - no rain

logreg.predict_proba(X_test)[:,0]
```

Out[213…

```
array([0.91385464, 0.8357268 , 0.82036795, ..., 0.97674976, 0.79859481,
       0.30737669])
```

In [214…

```
#probability of getting output as 1 -  rain

logreg.predict_proba(X_test)[:,1]
```

Out[214…

```
array([0.08614536, 0.1642732 , 0.17963205, ..., 0.02325024, 0.20140519,
       0.69262331])
```

14. Check accuracy score

In [215…

```
from sklearn.metrics import accuracy_score

print('Model accuracy score: {0:0.4f}'.format(accuracy_score(y_test,y_pred_test)))
```

```
Model accuracy score: 0.8502
```

Here, **y_test** are the true class labels and **y_pred_test** are the predicted class labels in the test-set.

Compare the train-set and test-set accuracy

Now, we will compare the train-set and test-set accuracy to check for overfitting.

In [216…

```
y_pred_train = logreg.predict(X_train)

y_pred_train
```

Out[216…

```
array(['No', 'No', 'No', ..., 'No', 'No', 'No'], dtype=object)
```

In [218…

```
print('Training-set accuracy score: {0:0.4f}'.format(accuracy_score(y_train, y_pred_train)))
```

```
Training-set accuracy score: 0.8477
```

check for overfitting and underfitting

In [219…

```
# print the scores on  training and test set

print('Training set score: {:.4f}'.format(logreg.score(X_train,y_train)))

print('Test set score: {:.4f}'.format(logreg.score(X_test, y_test)))
```

```
Training set score: 0.8477
Test set score: 0.8502
```

The training-set accuracy score is 0.8477 while the test-set accuracy to be 0.8502. This two values are quite comparable. So, there is no question for overfitting.

In Logistic Regression, we use default value of C = 1. It provides good performance with approximately 85% accuracy on both the training and the test set. But the model performance on both the training and test set are very comparable. It is likely the case of underfitting.

We will increase C and fit a more flexible model.

In [220…

```
# fit the Logistic Regression model with C=100

#instantiate the model
```

```
logreg100 = LogisticRegression(C=100, solver = 'liblinear', random_state=0)

#fit the model
logreg100.fit(X_train, y_train)
```

Out[220…

```
LogisticRegression
LogisticRegression(C=100, random_state=0, solver='liblinear')
```

In [221…

```
# print the scores on  training and test set

print('Training set score: {:.4f}'.format(logreg100.score(X_train,y_train)))

print('Test set score: {:.4f}'.format(logreg100.score(X_test, y_test)))
```

```
Training set score: 0.8478
Test set score: 0.8505
```

We can see that, C= 100 results in higher test set accuracy and also a slightly increased training set accuracy. So, we can conclude that a more complex model should perform better.

Now, We will invesstigate, what happens if we use more regularized model than the default value of C= 1, by setting C= 0.01

In [222…

```
# fit the Logistic Regression model with C=0.01

#instantiate the model

logreg001 = LogisticRegression(C=0.01, solver = 'liblinear', random_state=0)

#fit the model
logreg001.fit(X_train, y_train)
```

Out[222…

```
LogisticRegression
LogisticRegression(C=0.01, random_state=0, solver='liblinear')
```

In [223…

```
# print the scores on  training and test set

print('Training set score: {:.4f}'.format(logreg001.score(X_train,y_train)))

print('Test set score: {:.4f}'.format(logreg001.score(X_test, y_test)))
```

```
Training set score: 0.8409
Test set score: 0.8448
```

So, if we use more regularized model by setting C=0.01, then both training and test set accuracy decrease relative to the default parameters.

Compare model accuracy with null accuracy

So, the model accuracy of 0.8501. But, we cannot say that our model is very good based on the above accuracy. We must compare it with the **null accuracy**. Null accuracy is the accuracy that could be achieved by always predictiong the most frequesnt class.

So, we should first check the class distribution in the test set.

In [224…

```
# check class distribution in test set

y_test.value_counts()
```

Out[224…

```
RainTomorrow
No     22067
Yes     6372
Name: count, dtype: int64
```

We can see that the occurences of most frequent class is 22067. So, we can calculate null accuracy by dividing 22067 by total number of occurences.

In [225…

```
#check null accuracy score

null_accuracy = (22067/(22067+6372))

print('Null accuracy score: {0:0.4f}'.format(null_accuracy))
```

```
Null accuracy score: 0.7759
```

We can see that our model accuracy score is 0.8502 but the null accuracy is 0.7759. So, we conclude that our Logistic Regression model is doing a very good job in predicting the class labels.

Now, based on the above analysis we can conclude that our classification model accuract is very good. Our model is doing a very good job in terms in predicting the class labels.

But, it does not gie the underlying distribution of values. Also, it does not tell anything about the type of errors our classifier is making.

We have another tool called `Confusion matrix` that comes to our rescue.

15. Confusion Matrix

A confusion matrix is a tool for summarizing the performance of a classification algorithm. A confusion matrix will give us a clear picture of classfication model performance and the types of errors produced by the model. It gives us a summary of correct and incorrect predictions broken down by each category. The summary is represented in tabular form.

Four types of outcomes are possible while evaluating a classfication model performance. These four outcomes are describe below: -

True Positives (TP) - True Positives occur when we predict an observation belongs to a certain class and the observation actually belongs to that class.

True Negatives (TN) - True Negatives occur when we predict an observation does not belongs to a certain class and the observation actually does not belong to that class.

False Positives (FP) - False Positives occurs when we predict an observation does not belong to a certain class but the observation actually does not belong to that class. This type of error is called **Type I error**.

False Negatives (FN) - False Positives occurs when we predict an observation does not belong to a certain class but the observation actually belong to that class. This is a very serious type of error and it is called **Type II error**.

These four outcomes are summarized in a confusion matrix given below.

In [226…

```
# Print the confusion matrix and slice it into four pieces

from sklearn.metrics import confusion_matrix

cm = confusion_matrix(y_test, y_pred_test)

print('Confusion Matrix \n\n', cm)

print('\nTrue Positives(TP) = ', cm[0,0])

print('\nTrue Negatives(TN) = ', cm[1,1])

print('\nFalse Positives(FP) = ', cm[0,1])

print('\nFalse Negatives(FN) = ', cm[1,0])
```

```
Confusion Matrix

 [[20892  1175]
 [ 3086  3286]]

True Positives(TP) =  20892

True Negatives(TN) =  3286

False Positives(FP) =  1175

False Negatives(FN) =  3086
```

The confusion matrix shows `20892 +3285 = 24177 correct predictions` and `3087 + 1175 = 4262 incorrect predictions.`

In this case, we have

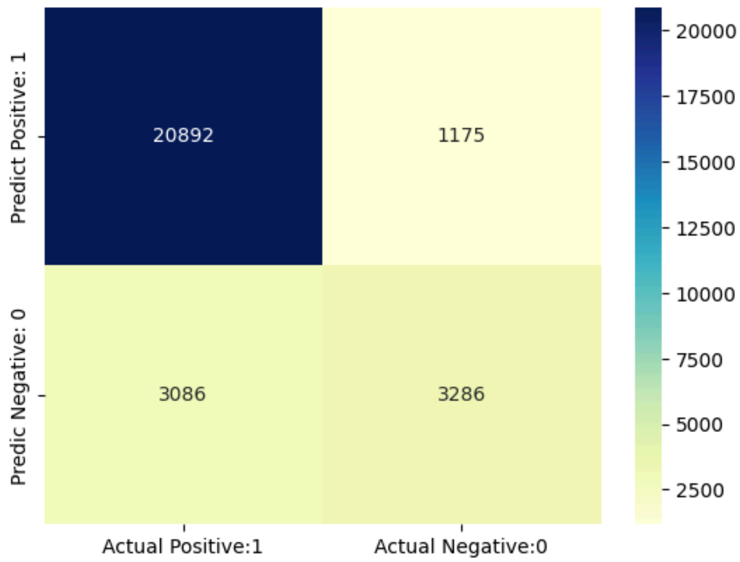
- `True Positives` (Actual Positive: 1 and Predict Positive: 1) - 20892
- `True Negatives` (Actual Negative: 0 and Predict Negative: 0) - 3285
- `False Positives` (Actual Negative: 0 but Predict Positive: 1) - 1175 `(Type I error)`
- `False Negatives` (Actual Positive: 1 but Predict Negative: 0) - 3087 `(Type II error)`

In [228…

```
# visualize confusion matrix with seaborn heatmap

cm_matrix = pd.DataFrame(data = cm, columns = ['Actual Positive:1','Actual Negative:0'],
                         index = ['Predict Positive: 1', 'Predic Negative: 0'])
```

```
sns.heatmap(cm_matrix, annot = True, fmt = 'd', cmap='YlGnBu')
```

Out[228… <Axes: >



16. Classification Report

Classification Report

Classification report is another way to evaluate the classification model performance. It displays the **precision, recall, f1** and **support** score for the model. We have described thse terms in later.

We can print a classification report as follows:-

In [229…

```
from sklearn.metrics import classification_report

print(classification_report(y_test, y_pred_test))
```

```
              precision    recall  f1-score   support

          No       0.87      0.95      0.91     22067
         Yes       0.74      0.52      0.61      6372

    accuracy                           0.85     28439
   macro avg       0.80      0.73      0.76     28439
weighted avg       0.84      0.85      0.84     28439
```

Classification accuracy

In [231…

```
TP = cm[0,0]
TN = cm[1,1]
FP = cm[0,1]
FN = cm[1,0]
```

In [232…

```
# print classification accuracy

classification_accuracy = (TP + TN)/float(TP + TN + FP + FN)

print('Classsification accuracy : {0:0.4f}'.format(classification_accuracy))
```

```
Classsification accuracy : 0.8502
```

Classification error

In [233…

```
# print classification error

classification_error = (FP + FN)/float(TP + TN + FP + FN)

print('Classsification error : {0:0.4f}'.format(classification_error))
```

```
Classsification error : 0.1498
```

Precision

Precision can be defined as the percentage of correctly predicted outcomes out of all the predicted positive outcomes. It can be given as the ratio of true positives (TP) to the sum of the true and false positives (TP + FP)

So, **Precision** identifies the proportion of correctly predicted positive outcome. It is more concerned with the positive class than the negative class

Mathematically, precision can be defined as the ratio of `TP to (TP+FP)`.

In [234…

```
# print precision score

precision = TP/float(TP +FP)

print('Precision : {0:0.4f}'.format(precision))
```

```
Precision : 0.9468
```

Recall

Recal can be defined as the percentage of correctly predicted positive outcomes out of all the actual positive outome. It can be given as the ratio of true positives (TP) to the sum of true positives and false negative (TP + FN). **Recall** is also called **Sensitivity**

Recall identifies the proportion of correctly predicted actual positives.

Mathematically, Recall can be given as ratio of `TP to (TP+FN)`.

In [235…

```
recall =TP/ float(TP+FN)

print('Recall or Sensitivity: {0:0.4f}'.format(recall))
```

```
Recall or Sensitivity: 0.8713
```

True Positive Rate

True Positive Rate is synonymous with **Recall**.

In [236…

```
true_positive_rate = TP/ float(TP + FN)

print('True Positive Rate: {0:0.4f}'.format(true_positive_rate))
```

```
True Positive Rate: 0.8713
```

False Positive Rate

In [237…

```
false_positive_rate = FP/ float(FP + TN)

print('False Positive Rate: {0:0.4f}'.format(false_positive_rate))
```

```
False Positive Rate: 0.2634
```

Specificity

In [238…

```
specificity = TN/ (TN + FP)

print('Specificity: {0:0.4f}'.format(specificity))
```

```
Specificity: 0.7366
```

f1-score

f1-score is weighted harmonic mean of precision and recall. The best possible **f1-score** would be 1.0 and the worst would be 0.0. **f1-score** is the harmonic mean of precision and recall. So, **f1-score** is always lower than accuracy measures as they embed precision and recall into their computation. The weighted average of `f1-score` should be used to compare classifier models, not global accuracy.

Support

Support is the actual number of occurence of the class in our dataset.

17. Adjusting the treshold level

In [239…

```
#print the first 10 predicted probabilities of two classes - 0 and 1

y_pred_prob = logreg.predict_proba(X_test)[0:10]

y_pred_prob
```

Out[239…

```
array([[0.91385464, 0.08614536],
       [0.8357268 , 0.1642732 ],
       [0.82036795, 0.17963205],
       [0.99025342, 0.00974658],
       [0.957263  , 0.042737  ],
       [0.97992825, 0.02007175],
       [0.1782761 , 0.8217239 ],
       [0.23474596, 0.76525404],
       [0.90052756, 0.09947244],
       [0.85490676, 0.14509324]])
```

Observations

- In each row, the numbers sum to 1.
- There are 2 columns which correspond to 2 classes- 0 and 1.
  - Class 0 - predicted probability that there is no rain tomorrow.
  - Class 1 - predicted probability that there is rain tomorrow.
- Importance of predicted probabilities
  - We can rank the observations by probability of rain or no rain.
- predict_proba process
  - Predicts the probabilities
  - Choose the class with the highest probability
- Classification threshold level
  - There is a classification threshold level of 0.5.
  - Class 1 - probability of rain is predicted if probability > 0.5.
  - Class 0 - probability of no rain is predicted if probability < 0.5.

In [240…

```
#store the probabilities in dataframe

y_pred_prob_df = pd.DataFrame(data = y_pred_prob, columns = ['Prob of - No rain tomorrow (0)', 'Prob of - Rain tomorrow (1)'])

y_pred_prob_df
```

Out[240…

	Prob of - No rain tomorrow (0)	Prob of - Rain tomorrow (1)
0	0.913855	0.086145
1	0.835727	0.164273
2	0.820368	0.179632
3	0.990253	0.009747
4	0.957263	0.042737
5	0.979928	0.020072
6	0.178276	0.821724
7	0.234746	0.765254
8	0.900528	0.099472
9	0.854907	0.145093

In [241…

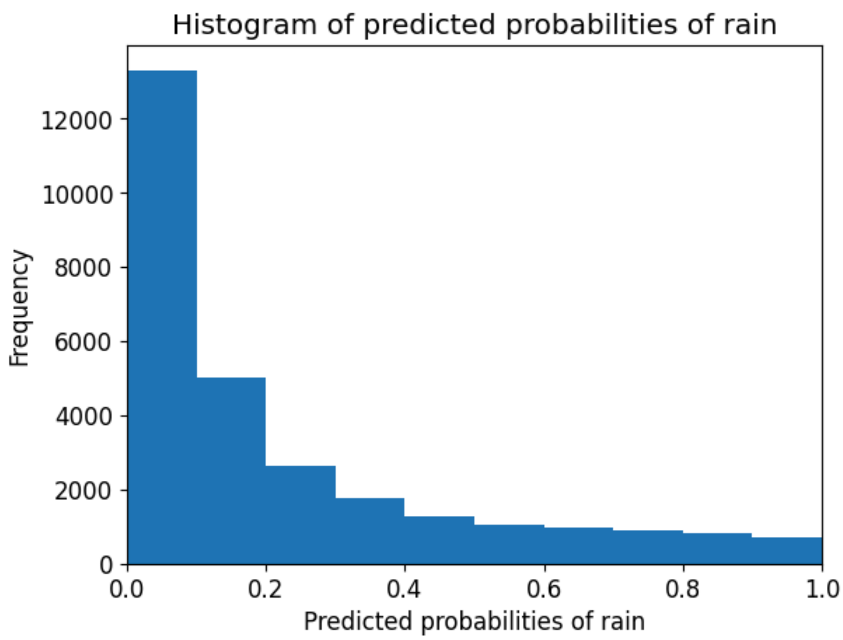
```
# print the first 10 predicted probabilities for class 1 - Probability of rain

logreg.predict_proba(X_test)[0:10, 1]
```

Out[241… array([0.08614536, 0.1642732 , 0.17963205, 0.00974658, 0.042737 ,
       0.02007175, 0.8217239 , 0.76525404, 0.09947244, 0.14509324])

In [242…
```
# store the predited probabilities for class 1 - Probability of rain

y_pred1 = logreg.predict_proba(X_test)[:,1]
```

In [243…
```
# plot histogram of predicted probabilities

#adjust the font size
plt.rcParams['font.size'] = 12

#plot histrogram with 10 bins
plt.hist(y_pred1, bins = 10)

#set the title of predicted probabilities
plt.title('Histogram of predicted probabilities of rain')

#set the x-axis limit
plt.xlim(0,1)

#set the title
plt.xlabel('Predicted probabilities of rain')
plt.ylabel('Frequency')
```

Out[243… Text(0, 0.5, 'Frequency')



Observations

- We can see that the above histogram is highly positive skewed.
- The first column tell us that there are approximatelty 15000 observations with probability between 0.0 and 0.1
- There are small number of observations with probability > 0.5.
- So, these small number of observations predict that there will be rain tomorrow.
- Majority of observation predict that there will be no rain tomorrow.

Lower the threshold

In [247…
```
from sklearn.preprocessing import binarize

for i in range(1,5):

  cm1=0

  y_pred1 = logreg.predict_proba(X_test)[:,1]
  y_pred1 = y_pred1.reshape(-1,1)
```

```
y_pred2 = binarize(y_pred1, threshold = i/10)
y_pred2 = np.where(y_pred2 == 1, 'Yes', 'No')

cm1 = confusion_matrix(y_test, y_pred2)

print('With', i/10, 'threshold the Confusion Matrix is','\n\n',cm1,'\n\n',
        'with',cm1[0,0]+cm1[1,1],'correct predictions, ','\n\n',
      cm1[0,1],'Type I errors ( False Positives ), ','\n\n',
      cm1[1,0],'Type II errors ( False Negatives ), ','\n\n',
      'Accuracy Score: ', (accuracy_score(y_test, y_pred2)), '\n\n',
      'Sensitivity: ',cm1[1,1]/(float(cm1[1,1]+cm1[1,0])), '\n\n',
      'Specificity: ',cm1[0,0]/(float(cm1[0,0]+cm1[0,1])), '\n\n',
      '====================================================','\n\n')
```

```
With 0.1 threshold the Confusion Matrix is

 [[12727  9340]
 [  547  5825]]

 with 18552 correct predictions,

 9340 Type I errors ( False Positives ),

 547 Type II errors ( False Negatives ),

 Accuracy Score:  0.6523436126446077

 Sensitivity:  0.9141556811048337

 Specificity:  0.5767435537227534

 ===========================================================

With 0.2 threshold the Confusion Matrix is

 [[17066  5001]
 [ 1233  5139]]

 with 22205 correct predictions,

 5001 Type I errors ( False Positives ),

 1233 Type II errors ( False Negatives ),

 Accuracy Score:  0.7807939800977531

 Sensitivity:  0.806497175141243

 Specificity:  0.7733720034440568

 ===========================================================

With 0.3 threshold the Confusion Matrix is

 [[19079  2988]
 [ 1873  4499]]

 with 23578 correct predictions,

 2988 Type I errors ( False Positives ),

 1873 Type II errors ( False Negatives ),

 Accuracy Score:  0.829072752206477

 Sensitivity:  0.7060577526679221

 Specificity:  0.8645941904200843

 ===========================================================

With 0.4 threshold the Confusion Matrix is

 [[20192  1875]
 [ 2517  3855]]

 with 24047 correct predictions,

 1875 Type I errors ( False Positives ),

 2517 Type II errors ( False Negatives ),

 Accuracy Score:  0.8455641900207461

 Sensitivity:  0.6049905838041432

 Specificity:  0.9150314949925228

 ===========================================================
```

Comments

- In binary problems, the threshold of 0.5 is used by default to convert predicted probabilities into class predictions.
- Threshold can be adjusted to increase sensitivity or specificity.
- Sensitivity and specificity have an inverse relationship. Increase one would always decrease the other and vice versa.
- We can see that increasing the threshold level results in increased accuracy.
- Adjusting the threshold level should be on of the last step you do in the model-building process.

18. ROC - AUC

ROC Curve

Another tool to measure the classification model performance visually is **ROC Curve**. ROC Curve stands for **Receiver Operating Characteristic Curve**. An **ROC Curve** is a plot which shows the performance of a classification mdoel at various classification threshold levels.

The **ROC Curve** plots the **True Positive Rate (TPR)** against the **False Positive Rate** at varius threshold levels.

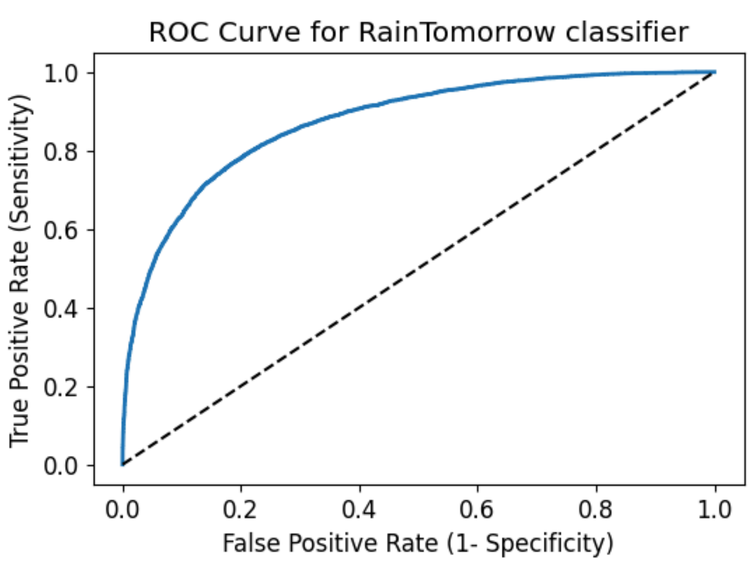
True Positive Rate (TPR) is also called **Recall**. It is defined as the ratio of `TP to (TP+FN).`

False Positive Rate is defined as ratio of `FP to (FP +TN).`

In the ROC Curve, we will focus on the TPR(True Positive Rate) and FPR(False Positive Rate) of a single point. This will give us the general performance of the ROC Curve which consists of the TPR and FPR at various threshold levels. So, an ROC Curve plots TPR vs FPR at different classification of threshold levels. If we lower the threshold levels, it may result in mroe items being classified as positive. It will increase both True Positive (TP) and False Positive(FP).

In [248…

```
# plot ROC Curve

from sklearn.metrics import roc_curve

fpr,tpr, thresholds = roc_curve(y_test,y_pred1, pos_label = 'Yes')

plt.figure(figsize=(6,4))

plt.plot(fpr,tpr, linewidth =2)

plt.plot([0,1],[0,1],'k--'  )

plt.rcParams['font.size'] = 12

plt.title('ROC Curve for RainTomorrow classifier')
plt.xlabel('False Positive Rate (1- Specificity)')
plt.ylabel('True Positive Rate (Sensitivity)')

plt.show()
```



ROC curve help use to choose a threshold level that balances sensitivity and specificity for a particular context.

ROC - AUC

ROC-AUC stands for **Receiver Operating Characteristic - Area Under Curve**. It is a technique to compare classifier performance. In this technique, we meausre the `area under the curve(AUC)`. A perfect classfier will have a ROC - AUC, whereas a purelt random classifier will have a ROC AUC equal to 0.5.

So, **ROC AUC** is the percentage of the ROC plto that is underneath the curve.

In [251...
```
#compute ROC AUC

from sklearn.metrics import  roc_auc_score

ROC_AUC = roc_auc_score(y_test, y_pred1)

print('ROC AUC: {:.4f}'.format(ROC_AUC))
```

```
ROC AUC: 0.8729
```

Comments

- ROC AUC is single number summary of classifier performance. The higher the value, the better the classifier.
- ROC AUC of our model approaches towards 1. So, we can conclude that our classifier does a good job in predicting whether it will rain tomorrow or not.

In [252...
```
# calculate cross-validated ROC AUC

from sklearn.model_selection import cross_val_score

Cross_validated_ROC_AUC = cross_val_score(logreg, X_train, y_train, cv = 5, scoring='roc_auc').mean()

print('Cross validated ROC AUC: {:.4f}'.format(Cross_validated_ROC_AUC))
```

```
Cross validated ROC AUC: 0.8695
```

19.k-Fold Cross Validation

In [253...
```
# Applying  5-fold Cross Validation

from sklearn.model_selection import cross_val_score

scores = cross_val_score(logreg, X_train, y_train, cv = 5, scoring='accuracy')

print('Cross-validation scores:{}'.format(scores))
```

```
Cross-validation scores:[0.84686387 0.84624852 0.84633642 0.84963298 0.84773626]
```

We can summarize the cross validation accuracy by calculating its mean.

In [254...
```
# compute Average cross-validation score

print('Average cross-validation score: {:.4f}'.format(scores.mean()))
```

```
Average cross-validation score: 0.8474
```

Our original model score found to be 0.8476. The average cross-validation score is 0.8474. So, we can conclude that cross validation does not result in performance improvement.

20. Hyperparameter Optimization using GridSearch CV

In [259...
```
from sklearn.model_selection import GridSearchCV

parameters = [{'penalty':['l1','l2']},
              {'C':[1, 10, 100, 1000]}]

grid_search = GridSearchCV(estimator = logreg,
                           param_grid = parameters,
                           scoring = 'accuracy',
                           cv = 5,
                           verbose = 0)

grid_search.fit(X_train, y_train)
```

Out[259…

```
▸ GridSearchCV
▸ estimator: LogisticRegression
  ▸ LogisticRegression
```

In [260…

```
#examin the best model

#best score achieved during the GridSearchCV
print('GridSeach CV best score : {:.4f}\n\n'.format(grid_search.best_score_))

#print parameters that give the best results
print('Parameters that give the best results :','\n\n',(grid_search.best_params_))

#print estimator that was chosen by GridSearc
print('\n\nEstimate that was chosen by the search :','\n\n',(grid_search.best_estimator_))
```

```
GridSeach CV best score : 0.8474


Parameters that give the best results :

 {'C': 1}


Estimate that was chosen by the search :

 LogisticRegression(C=1, random_state=0, solver='liblinear')
```

In [261…

```
#calculate GridSearch CV score on test set

print('GridSeach CV score on test set: {0:0.4f}'.format(grid_search.score(X_test, y_test)))
```

```
GridSeach CV score on test set: 0.8502
```

Comments

- Our original model test accuracy is 0.8501 while GridSearch CV accuracy was 0.8502
- We can see that GridSearch CV improve the performance for this particular model.

21. Results and conclusion

1. The logistic regression model accuracy score is 0.8501.So, the model does a very good job in predicting whether or not it will rain tomorrow in Australia.
2. Small number of observations predict that there will be rain tomorrow. Majority of observations predict that there will be no rain tomorrow.
3. The model shows no sign of overfitting
4. Increasing the value of C results in higher test set accuracy and also a slightly increased training set accuracy. So, we can conclude that a more complex model should perform better.
5. Increasing the treshold level results in increased accuracy
6. ROC AUC of our model approaches towards 1. So, we can conclude that our classifier does a good job in predicting whether it will rain tomorrow or not.
7. Our Original model accuracy score is 0.8501 whereas accuracy score after RFECV is 0.8500. So, we can obtain approximately similar accuract but with reduced set of features.
8. In the original model, we have FP = 1175 whereas FP1 = 1174. So, we get approximately same number of false positives. Also, FN =3087 wheareas FN1 = 3091.So, we get slightly higher false negatives.
9. Our, original model score is found to be 0.8476. The average cross validation score is 0.8474. So we can conclude that cross-validation does not result in performance improvement.
10. Our original model test accuracy is 0.8501 while GridSearch CV accuracy is 0.8502. We can see that GridSearch Cv improve the performance for this particular model.

In [ ]:

In [ ]:
```

```