MVP
Jaidaa Alidrisi

Introduction

The goal of this project is to use classification models to predict whether it is going to rain the next day or not based on weather observations from the previous day. Predicting rain is beneficial for various fields. As a software developer in the National Center for Meteorology, I can help weather forecasters and meteorologists in their weather forecasting process and issuing weather alerts to the public or to other governmental sectors. In addition, predicting rain assists farmers managing and improving their produce and it also helps event organizers plan their outdoor events better.

Dataset

The dataset contains 10 years of weather observations (145461 rows) and 23 features. The highlighted features that will help predict rain are:

- MinTemp: The minimum temperature registered in the day
- MaxTemp: The maximum temperature registered in the day
- Humidity9am: Humidity at 9AM
- Humidity3pm: Humidity at 3PM
- Cloud9am: the cloud cover at 9AM
- Cloud3pm: the cloud cover at 3PM
- Pressure9am: Atmospheric pressure at 9AM
- Pressure3pm: Atmospheric pressure at 3pm
- Temp9am : temperature at 9AM
- Temp3pm: temperature at 3PM
- Rainfall: The amount of rain
- RainToday: wether there is rain today or not

And the Target column is:

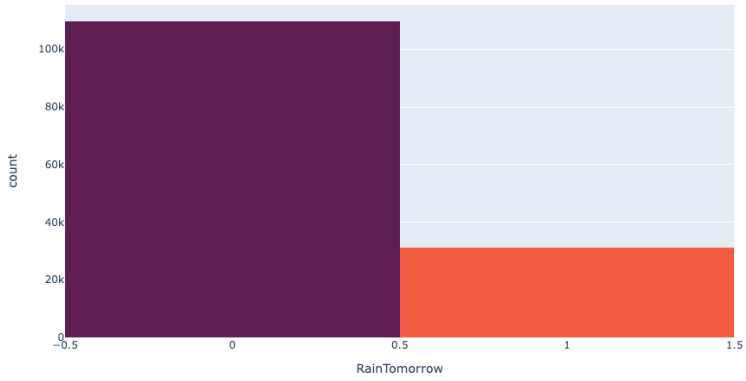
- RainTomorrow: weather there is rain tomorrow or not

MVP

For solving this data science project I started with data cleaning and EDA so I can understand the dataset clearly and I found the following:

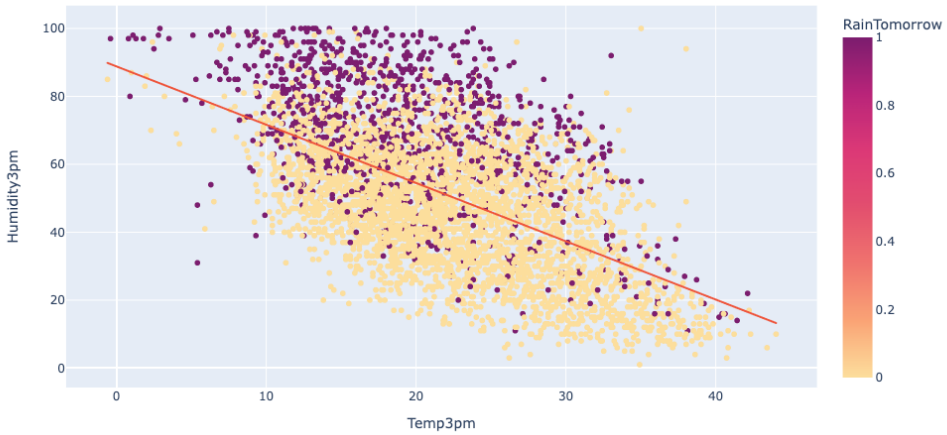
Data Visualization

```
px.histogram(df, x='RainTomorrow',color='RainTomorrow',color_discrete_sequence= ["#5e1e52", "#f15c42"] )
```

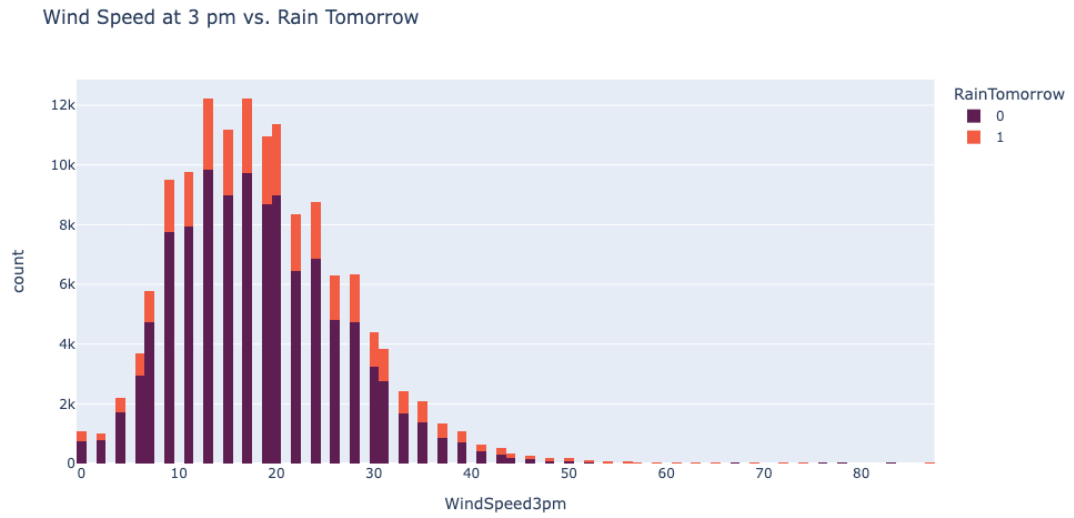


The first thing I noticed is that the data is imbalanced and I will handle that by using compute_class_weight from sklearn.utils.class_weight

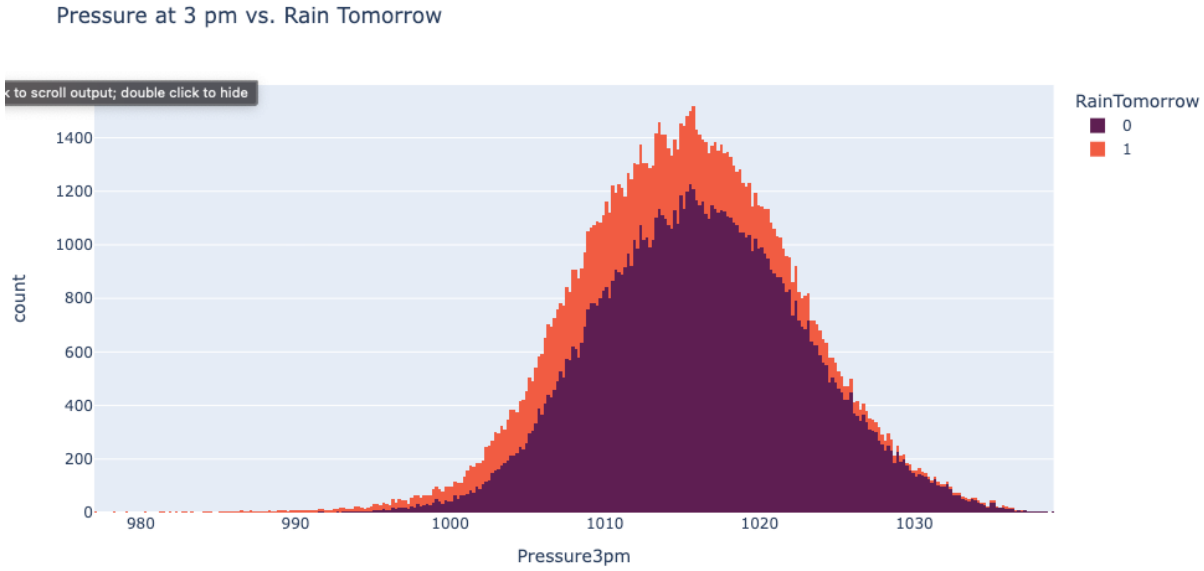
What are the features that affect Rain tomorrow ?



- From this scatter plot we can see that there is a negative trend between Temperature and Humidity. It is most likely to rain when humidity is high and temperature is low.



- From this histogram graph we can see the frequency of raining tomorrow is higher when the wind speed is median



- From this graph we can see the frequency of raining tomorrow is higher when the pressure is lower

Data Cleaning

After splitting the data to test and train the following preprocessing has been performed:

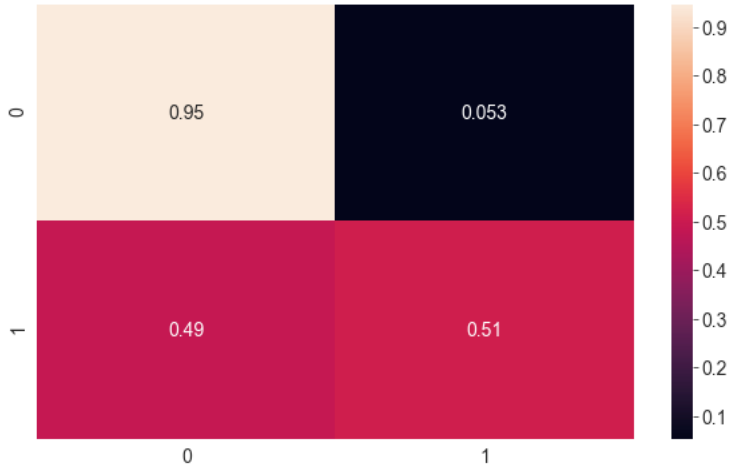
- Define input and target columns for test and train dataset
- Impute missing data for categorical and numerical features

Modeling

For predicting rain tomorrow I started with logistic regression and I got the following results:

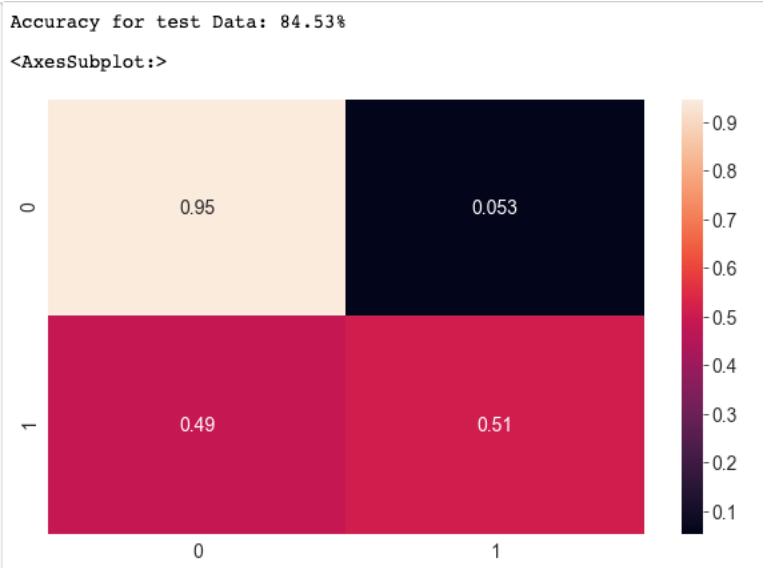
- Confusion matrix and accuracy for training data

```
Accuracy for Train Data: 85.07%
: <AxesSubplot:>
```



- Confusion matrix and accuracy for test data

```
Accuracy for test Data: 84.53%
<AxesSubplot:>
```



- Classification Report

	precision	recall	f1-score	support
0	0.94	0.87	0.90	23844
1	0.50	0.72	0.59	4314
accuracy			0.85	28158
macro avg	0.72	0.79	0.75	28158
weighted avg	0.88	0.85	0.86	28158

We can conclude from these scores that predicting true negative is more accurate than predicting True positive so predicting that it will not rain tomorrow is more accurate than if it will rain and this result could be due to class imbalance.

Final Project Plan

- Apply Class Weight and see if it will enhance the score
- Scale Numeric Features
- Apply another Classification models and compare results