ML Trail Task Report

Problem: To train a model which could compute the Average temperature for a day given some set of features about the day.

Assumptions:

1. The task is a simple regression task
2. Each day has its own independent average temperature which is the average of that specific day's Minimum Temperature and Maximum Temperature

Data preprocessing

1. **Hypothesis 1:**
   a. The data contains 22 features of which only four are concerned with Temperature.
   b. Therefore, I checked the correlation of each feature against the target variable(AvgTemp) and selected features with correlation higher that 0.8

2. **Hypothesis 2:**
   a. The variance in temperature is low on consecutive days hence missing temperature values can be computed as average of the day(day - 1) before and the day after(day + 1)
   b. Therefore I used DataFrame.interpolate() to fill in missing values because it maintains the distribution of the data and fills in the missing values.

Metrics

1. **Mean_Square_error:** It is simply the average of the squared difference between the target value and the value predicted by the regression model.Since the differences are squared, even small errors sum to large values which might lead to overestimation of how bad the model is performing.
2. **R^2 error:** The metric compares each model with a constant baseline and tells how each model performs relative to the mean score.

Experiment

1. Computed target variable as 0.5 x (MinTemp + MaxTemp)
2. Performed feature selection, reducing the dataset to features that have a high correlation with the target variable
3. Performed a 70/30 Test and Train split on the dataset(70% data in train set and 30% data in test set)
4. Create a dictionary of regression models to use on the selected data
5. Fit the model to the data and printed the following:
   a. Model name

b. The regression equation computed by the model
c. Mean_Square_error
d. R^2 error