Objective:

To predict the air pollution index for the dates given using historical weather data.

Feature Engineering:

* There are 2 categorical variables and rest are numeric variables. Categorical variables are transformed using one hot encoding technique.
* Weather squall is not available in test data. So ignored for prediction.
* Date time feature has been engineered to yield additional columns such as year, month, day, day of week and hour.
* Have done univariate analysis and found the following observations

1. Both Visibility in miles and Dew points are having same values. So one of the column is ignored
2. Features are in different units and need proper scaling of data to avoid disruption in prediction
3. There are no missing values
4. There are few duplicates in test data. However those can’t be ignored as dimensions violates while submission

Feature Scaling:

Since features are distributed in different units. I have done both normalization and Standardization to achieve scaling of data

This is to check and evaluate the metrics based on data available after both types of scaling

Model Building:

To run and evaluate different linear models, have created the function runmodel with model and data as parameters

I have chosen the following models:

1. Linear Regression
2. Lasso Regression
3. Ridge Regression
4. Elasticnet Regression
5. XGBoost

Hyperparameter Tuning:

Have done hyper parameter tuning on both linear based and tree based models. However used randomized search for tree based models to get optimal time

Final Result:

Have used weighted average of combinational models to arrive the prediction. As Lasso and XGB gave the better result. Have used 80% of Lasso and 20% of XGB.

Model Score:

Have got 91.51522 as highest score using (0.8\* Lasso)+(0.2\*XGB)