Lab 02 Regression - 190287R

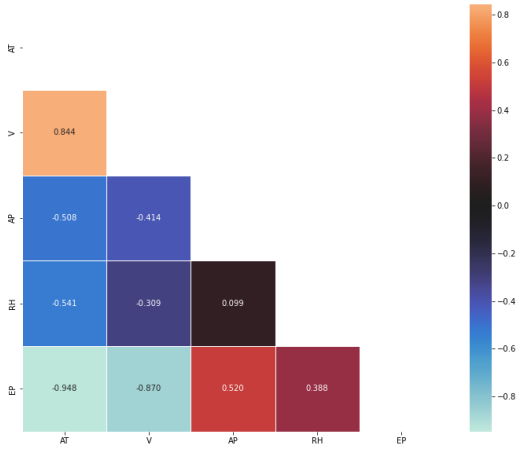
Electrical Energy Output (EP) Predictor

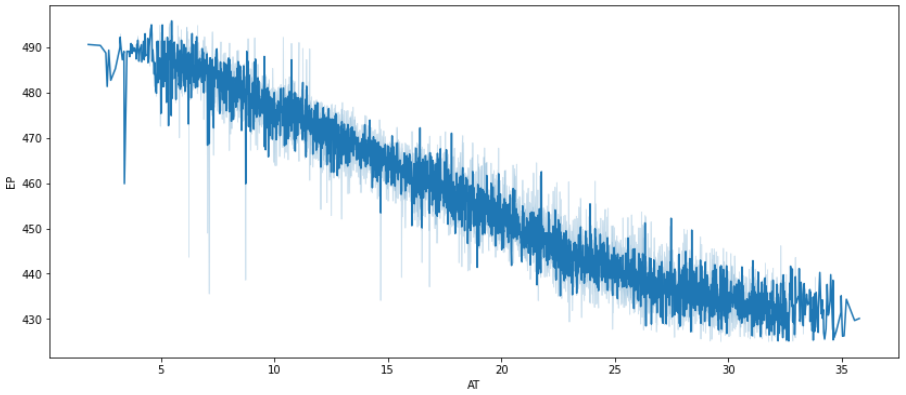
Step 01: Descriptive Analysis

	count	mean	std	min	25%	50%	75%	max
AT	8500.0	19.656261	7.451720	1.81	13.5275	20.360	25.7125	35.77
V	8500.0	54.310672	12.699757	25.36	41.7400	52.080	66.5400	80.25
AP	8500.0	1013.254513	5.936605	992.89	1009.1175	1012.945	1017.2400	1033.29
RH	8500.0	73.354868	14.628062	25.56	63.3800	75.025	84.9200	100.16
EP	8500.0	454.338720	17.068370	420.26	439.7275	451.440	468.4300	495.76

We have training [shape (8500, 6)] and testing [shape (1068, 5)] datasets. In both, there are some features such as Temperature (AT), Ambient Pressure (AP), Relative Humidity (RH) and Exhaust Vacuum (V). Electrical energy output (EP) has been given in the training dataset as a target variable. However, there is no column for EP in the test dataset. We have to predict them. Above picture shows a simple descriptive analysis of the training dataset.

Step 02: Exploratory Analysis on Training Data





Step 03: Separate The Predictors And The Target Variable in The Training Dataset

Step 04: Validation Data Creation (I Used 20% Data from Training Data)

Step 05: Approach 01 Using LinearRegression

I found an almost linear relation between some predictors and the target. Therefore first, I started with the LinearRegression, and it is the simplest also. I trained the LinearRegression model using the training dataset. Then using this model, I predicted the target values for the given training dataset and the created validation dataset. Using the predicted and the actual values of the target variable in both the cases, I found the mean absolute error(MAE) and the r2 score. The value of training MAE, training r2 score, validation MAE and validation r2 score were 3.62, 0.93, 3.6, and 0.93, respectively. The r2 scores are almost 1 in both cases. Therefore points have been fitted OK here. In other words, this is a low bias model. Both the MAE are almost equal. Therefore variance is also low here. As a result, it is a low bias and low variance model with low MAEs.

Step 06: Approach 02 Using RandomForestRegressor

I did min-max scaling to the predictors in the training and validation datasets. I used the scaled predictors and the target from training data to train RandomForestRegressor. Then using this model, I predicted the target values for the scaled predictors in the training dataset and the created validation dataset. Using the predicted and the actual values of the target variable in both cases, I found the mean absolute error(MAE). The value of training MAE and validation MAE were 0.9 and 2.36, respectively.

Step 07: Approach Selection

I used MAE to select the best approach between these two. You can note that the second approach has given a lower MAE. Therefore I chose the second approach.

Step 08: Model Tuning

I did some tuning to the hyperparameters involved in the second approach. As a result, I found we can get better performance than the previous when using 31 for the max_depth with other default values. Here, the value of validation MAE was 2.35.

Step 09: Predict Values of The Target Variable in Test Dataset.