Prediction of CO and NOx Emission from power plant turbines

OBJECTIVE: Prediction of Gas Turbine CO and NOx Emission from power plant turbines

Description: Predict the Gas Turbine CO and NOx Emission using 11 sensor measures aggregated over one hour from a gas turbine located in Turkey's north western region for the purpose of studying flue gas emissions, namely CO and NOx (NO + NO2)

Motivation: Harmful effect of Flue gas emitted from power plant turbines on environment has always been a substantial concern. In the recent past years many peaceful protest to save environment has been seen. Environmental organization that seeks to protect, analyse or monitor the environment has conducted many events and activities to raise people awareness on environment. This project aims to predict emission of flue gases based on sensor data from gas turbine and various Machine Learning techniques.
The ML model can be used to predict/estimate amount of emission for future operations of Turbine and Turbine of same homologous series. Model output can also be used for validation and backing up of costly continuous emission monitoring systems used in gas-turbine-based power plants. Their implementation relies on the availability of appropriate and ecologically valid data.

Data Source: https://archive.ics.uci.edu/ml/datasets/Gas+Turbine+CO+and+NOx+Emission+Data+Set#

Data Description: The dataset contains 36733 instances of 11 sensor measures aggregated over one hour (by means of average or sum) from a gas turbine located in Turkey's north western region for the purpose of studying flue gas emissions, namely CO and NOx (NO + NO2).

Variable (Abbr.) & Unit
Ambient temperature (AT) in Celsius
Ambient pressure (AP) in mbar
Ambient humidity (AH) (%)
Air filter difference pressure (AFDP) in mbar
Gas turbine exhausts pressure (GTEP) in mbar
Turbine inlet temperature (TIT) in Celsius
Turbine after temperature (TAT) in Celsius
Compressor discharge pressure (CDP) in mbar
Turbine energy yield (TEY) in MWH
Carbon monoxide (CO) in mg/m^3
Nitrogen oxides (NOx) in mg/m^3

Tools Used: Python, Jupyter-lab, Ms- Excel, Tableau

Data Information

Attributes, their count and data type

SL. No	Attribute	Non-Null Count	Data type
0	AT	36674 non-null	float64
1	AP	36674 non-null	float64
2	AH	36674 non-null	float64
3	AFDP	36674 non-null	float64
4	GTEP	36674 non-null	float64
5	TIT	36674 non-null	float64
6	TAT	36674 non-null	float64
7	TEY	36674 non-null	float64
8	CDP	36674 non-null	float64
9	CO	36674 non-null	float64
10	NOx	36674 non-null	float64

Table 1

First Five Rows of data

AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO	NOx
23.06	1019.30	62.78	4.25	30.51	1100.00	542.30	150.94	13.38	1.67	49.31
25.55	1010.50	81.23	4.45	29.85	1099.60	545.38	146.08	13.12	1.06	55.24
18.25	1017.90	81.40	3.91	21.33	1043.60	539.33	113.22	11.02	12.66	71.89
19.74	1016.00	82.36	3.76	24.20	1078.60	549.94	130.07	11.89	2.02	52.26
26.96	1010.20	65.21	5.38	30.73	1099.90	544.02	148.01	13.27	1.10	55.54

Table 2

Attributes are on different scale. Like AT and AH are on the scale of 100 where has AP and TIT are on scale of 1000. So for ML models data set is scaled using standard scalar.

Statistical Analysis

Statistical Summary of data

	AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO	NOx
count	36674	36674	36674	36674	36674	36674	36674	36674	36674	36674	36674
mean	17.72	1013.06	77.87	3.93	25.57	1081.51	546.18	133.54	12.06	2.34	65.26
std	7.45	6.46	14.46	0.77	4.19	17.42	6.83	15.60	1.09	2.12	11.62
min	-6.23	985.85	24.09	2.09	17.70	1000.80	512.45	100.02	9.85	0.00	25.91
25%	11.78	1008.80	68.18	3.36	23.15	1071.90	544.78	124.59	11.44	1.18	57.15
50%	17.82	1012.60	80.47	3.94	25.11	1085.90	549.88	133.73	11.97	1.71	63.83
75%	23.67	1017.00	89.38	4.38	29.07	1097.10	550.04	144.12	12.86	2.83	71.51
max	37.10	1036.60	100.20	7.61	40.72	1100.90	550.61	179.50	15.16	44.10	119.91

Table 3

Statistical Distribution of data

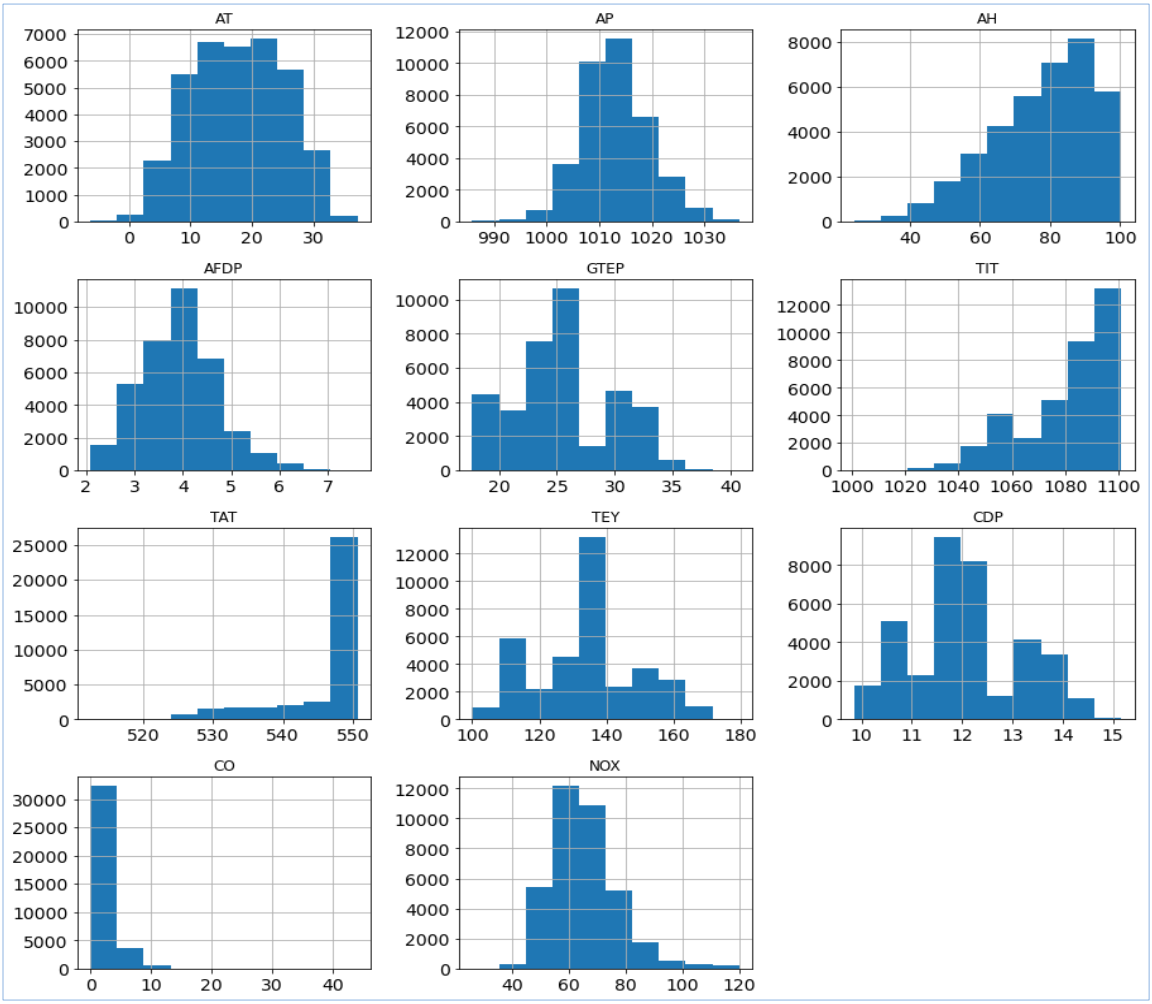

Fig 1

Correlation Analysis

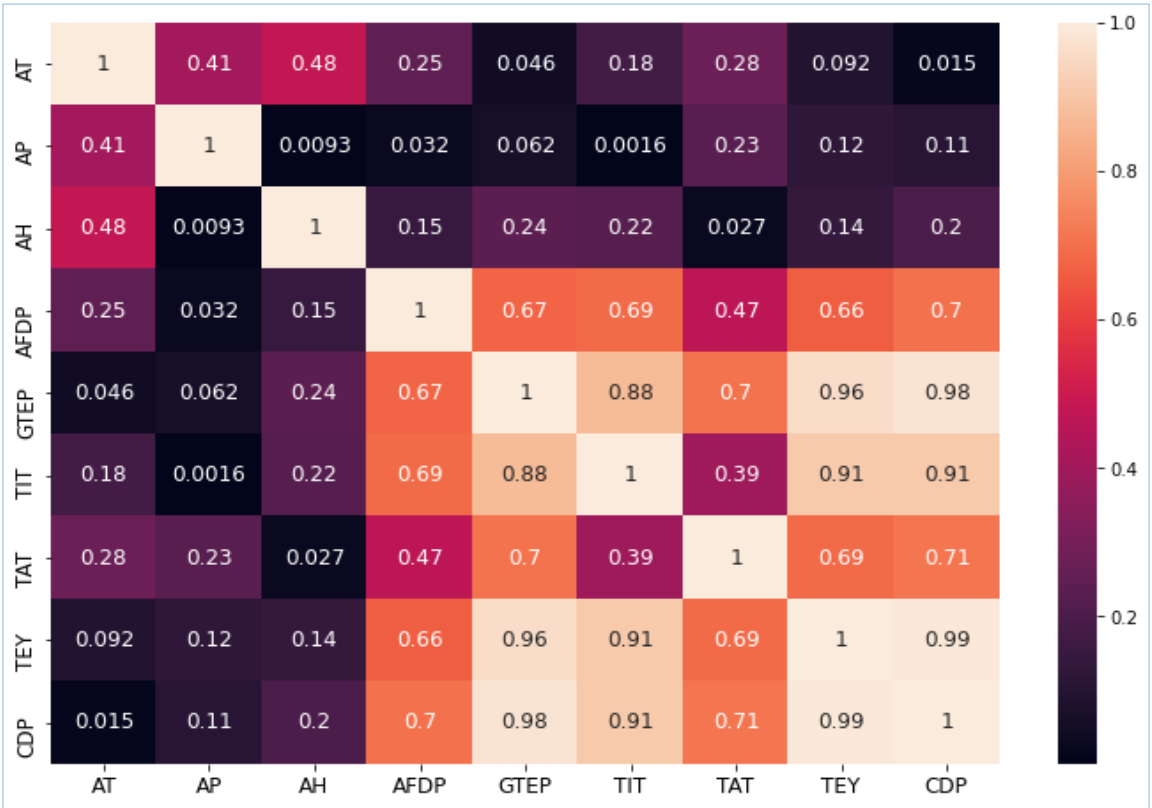

Fig 2

Variation of CO with other features

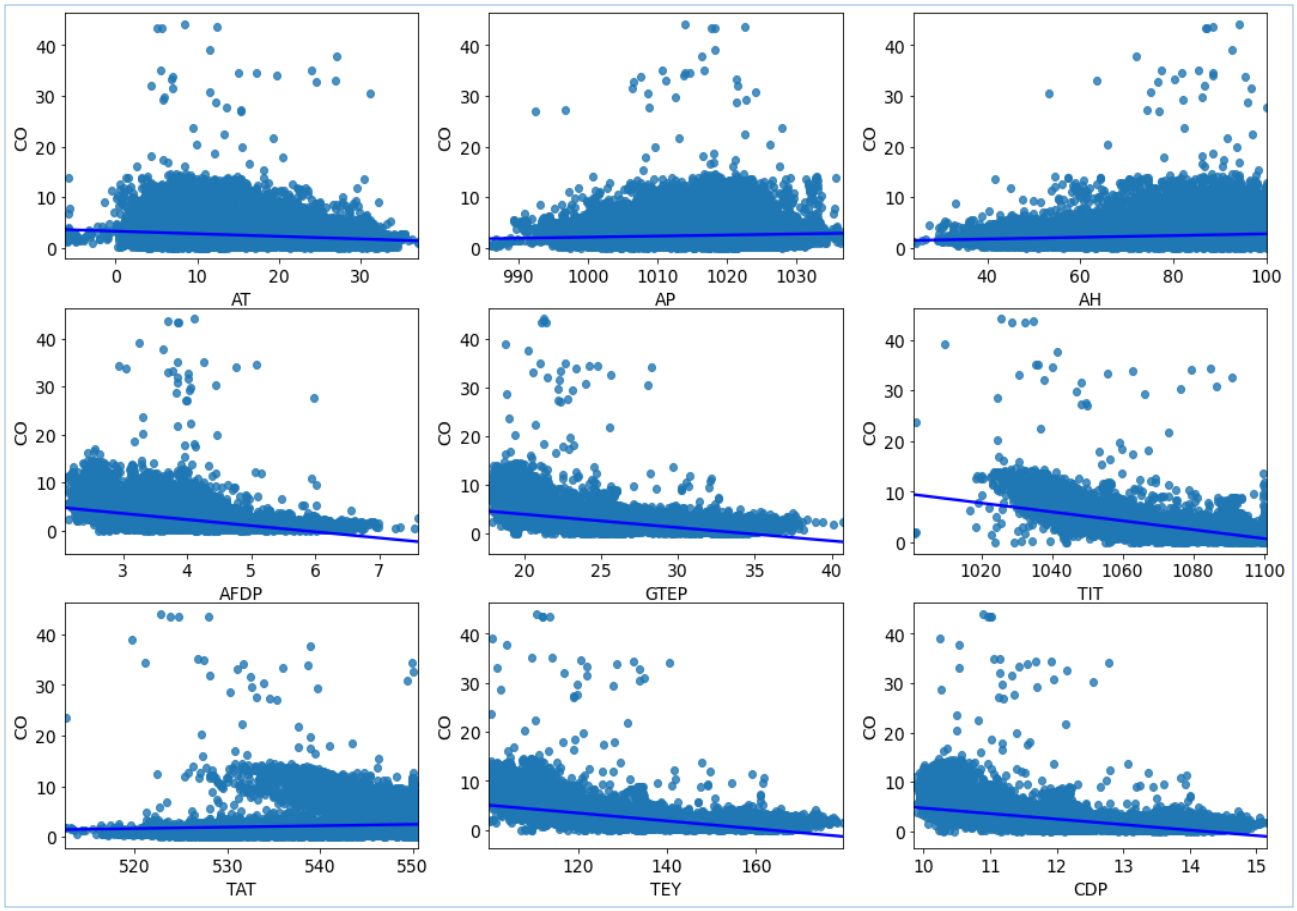

Fig 3

Variation of NOx with other features

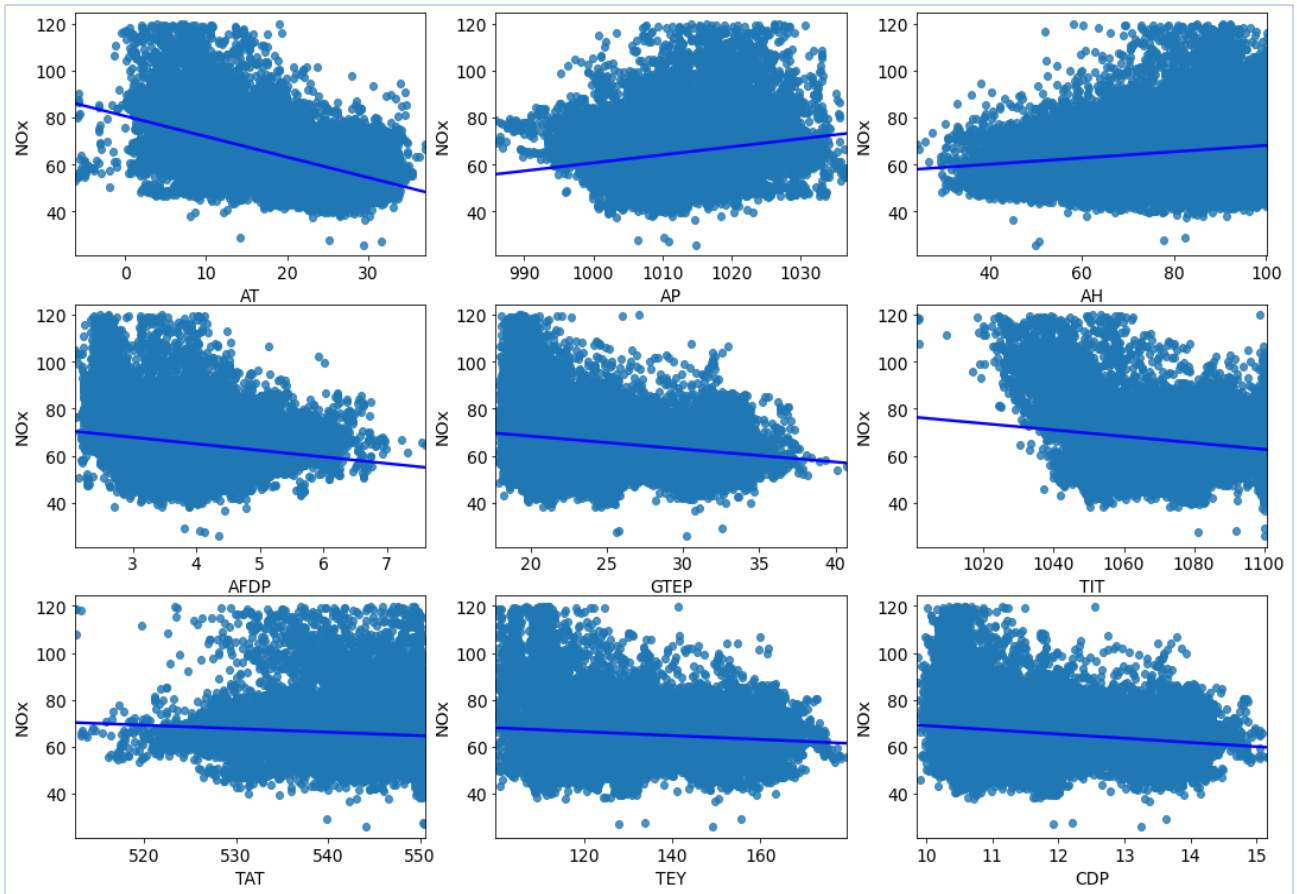

Fig 4

Model Building

First we split the data in train, cross validation and test set in the ratio of 60:20:20.
We built model on train set and tuned it on cross validation. Test data was kept aside for final evaluation.
Separate ML models were built for prediction of CO and NOx emission.
Using Forward addition and Backward Elimination method significant predictor variables were selected for CO and NOx prediction.

Model Building Process of CO

Feature Engineering

Principle Component Analysis

In statistical analysis using heat map we saw there was high multi – collinearity among the predictors.
So we applied PCA for removing multi – collinearity. Below figure represent how predictors were combined after PCA. VO, V1, V2, V3 are principle components we found after PCA

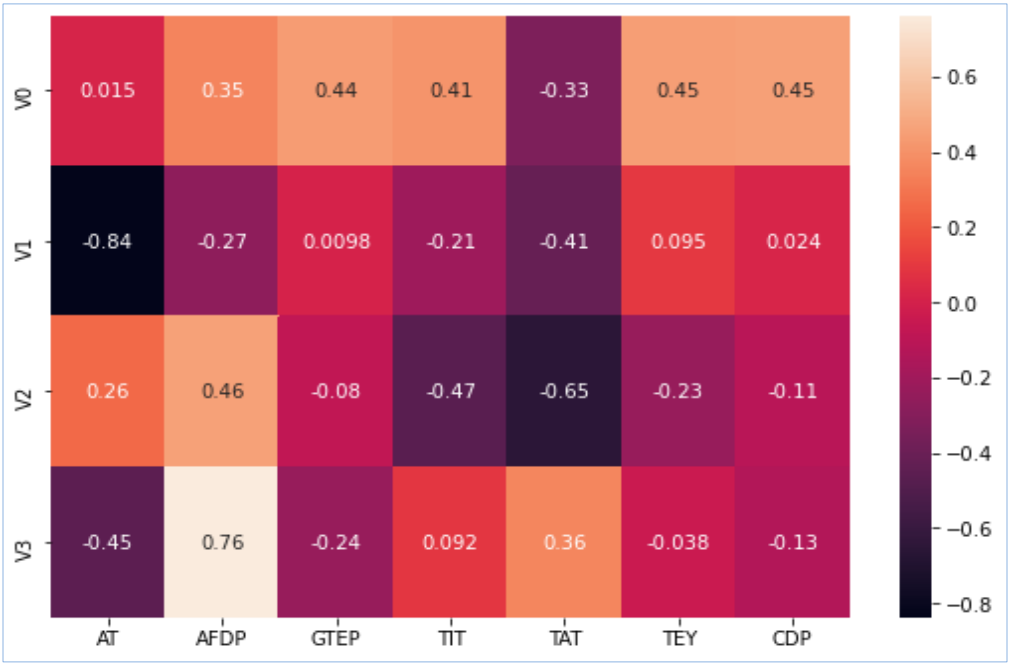

Fig 5

Linear Regressing Model built on above described PCA showed Adjusted_R^2 = 0.57 with MSE = 1.78 on cross validation set.

Polynomial Feature Transformation

Next we applied *Polynomial Feature transformation to get new features as a combination of old*
Below figure represent how features were combined after polynomial feature transformation

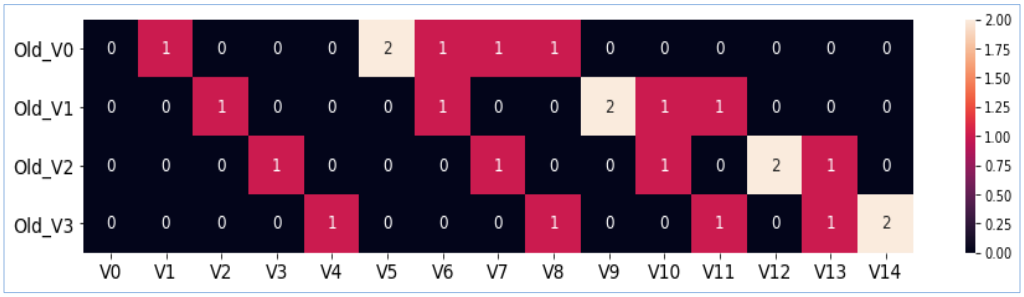

Fig 6

Old_Vi's are principle components and Vi's are new set of features after feature transformation.
After Feature transformation we again Linear Regression model and found Adjusted_R^2 = 0.669 with MSE = 1.38 on cross validation set.

Next we applied ridge regression, Decision Tree, Random Forest and Support Vector Machine on train data and tuned it on cross validation set.

Performance of Decision Tree on train and cross validation for different depth

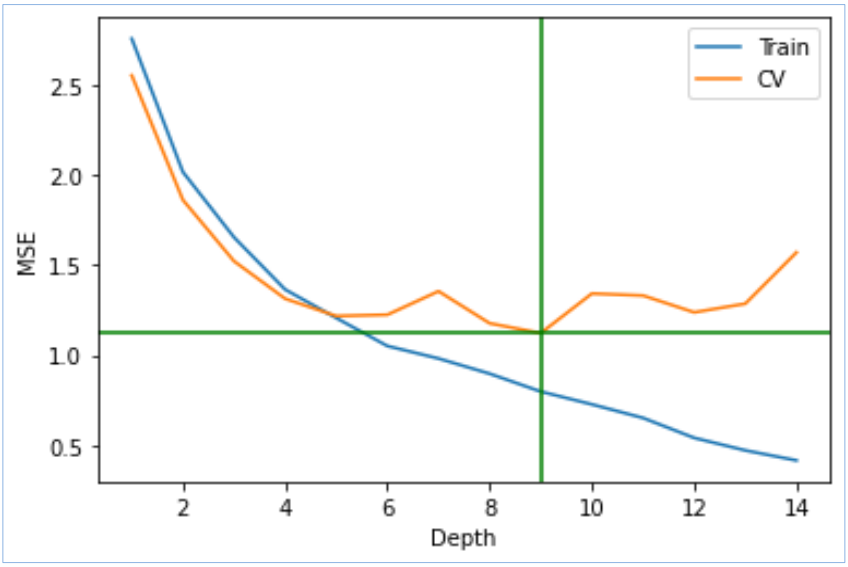

Fig 7

Performance of Random Forest on train & cross validation for different no. of features & max depth 9

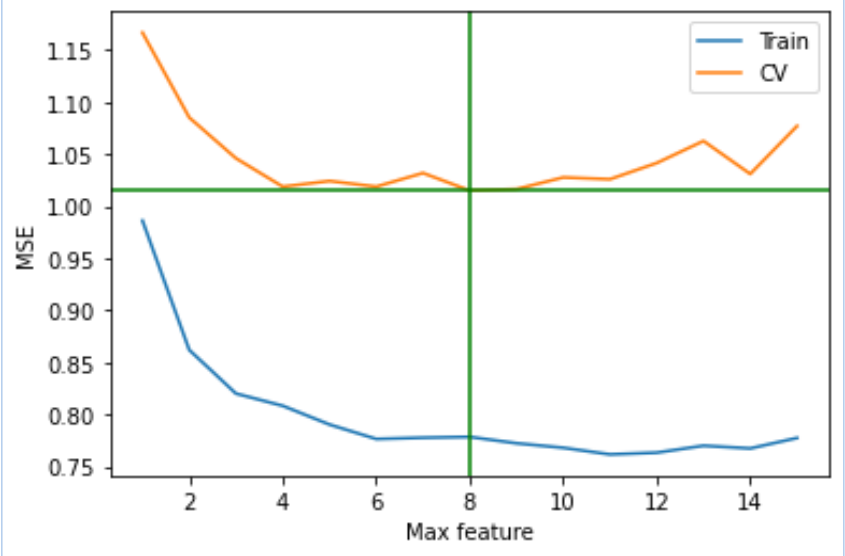

Fig 8

Below table summarizes the performance of these models on cross validation set.

ML MODEL	MSE on CV
Linear Regression	1.38
Ridge Regression	1.38
Decision Tree	1.12
Random Forest	1.01
SVM Regressor	1.23
Mean Line Prediction	30682.35

Table 4

We found random forest performs best on cross validation set with MSE of 1.01.

HYPERPARAMETER OPTIMIZATION USING BAYESIAN OPTIMIZATION

We applied Bayesian Optimization and 5 fold cross validation to find best parameter for Random Forest.
In below table shows the range over which hyper parameters of Random Forest were varied for selection of best parameter.

Hyper parameter	Description	Value
max_depth	maximum depth of trees in RF	10 - 25
n_estimators	no. of tree in RF	80 - 200
max_features	maximum no. of features to be considered for splitting	3 - 15
criterion	criteria for split	mse, friedman_mse

Table 5

Best Set of Parameters for Random Forest

Hyper parameter	Description	Best Value
max_depth	maximum depth of trees in RF	16
n_estimators	no. of tree in RF	161
max_features	maximum no. of features to be considered for splitting	4
criterion	criteria for split	friedman_mse

Table 6

Finally we built Random Forest model on train and cross validation combined with best set of hyper parameter and tested it on test data.
We found train MSE =0.34 and test MSE = 1.16.
Next we built the Random forest with the same best set of forest on all data and deployed it locally using gradio library.

Feature Importance

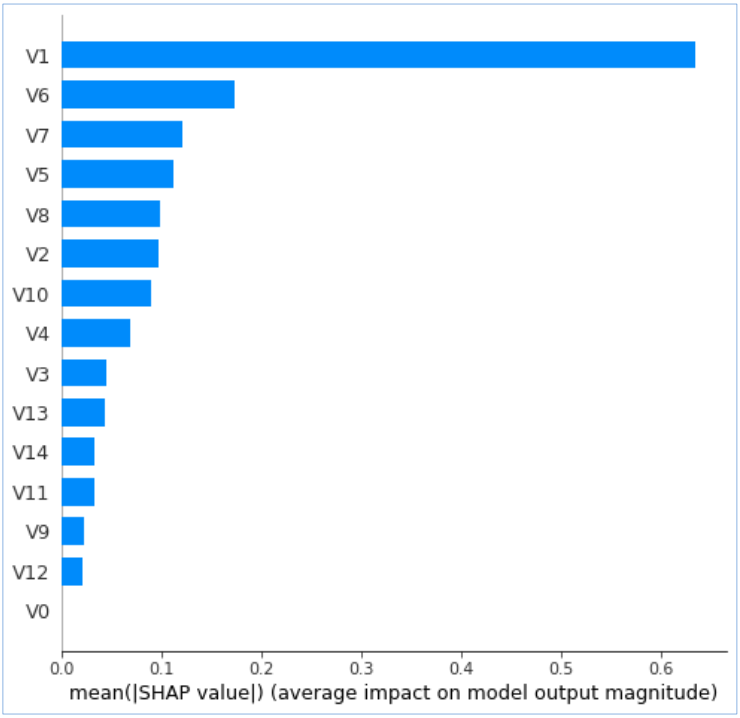

Fig 8

Top 5 most important features we found were V1 followed by V6, V7, V5 and V8.
Where V1 = Old_V0 (PC V0) = 0.015*AT + 0.35*AFDP + 0.44*GTEP + 0.41*TIT -0.33*TAT + 0.45*TEY +0.45*CDP

Model Building Process of NOx

Similar steps were followed to build model for NOx emission prediction.

Feature Engineering

Principle Component Analysis

Below figure represent how predictors were combined after PCA. VO, V1, V2, V3, V4, V5 are principle components we found after PCA

Fig 9

Linear Regressing Model built on above described PCA showed Adjusted_R^2 = 0.38 with MSE = 81.13 on cross validation set. Linear regression model has very poor performance.

Polynomial Feature Transformation

Below figure represent how features were combined after polynomial feature transformation.

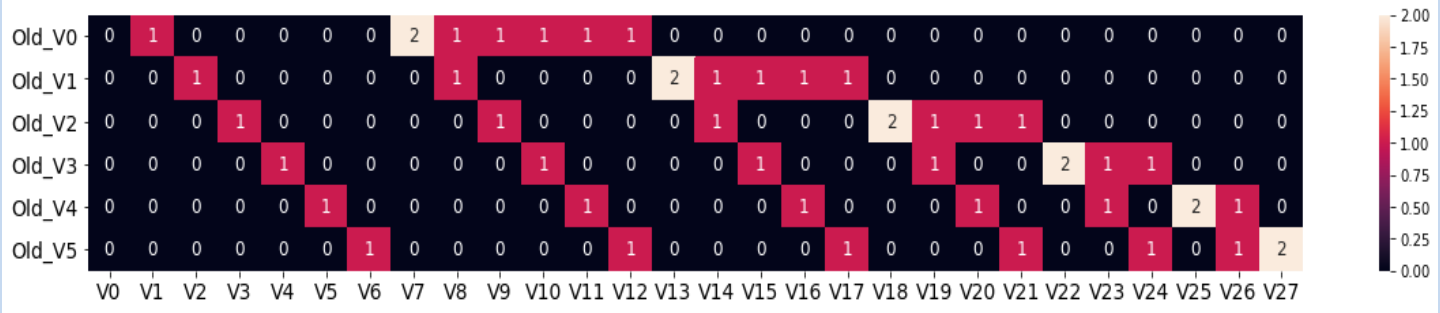

Fig 10

Old_Vi's are principle components and Vi's are new set of features after feature transformation.
After Feature transformation we again Linear Regression model and found Adjusted_R^2 = 0.596 with MSE = 52.81 on cross validation set.

Next we applied ridge regression, Decision Tree and Random on train data and tuned it on cross validation set.

Performance of Decision Tree on train and cross validation for different depth

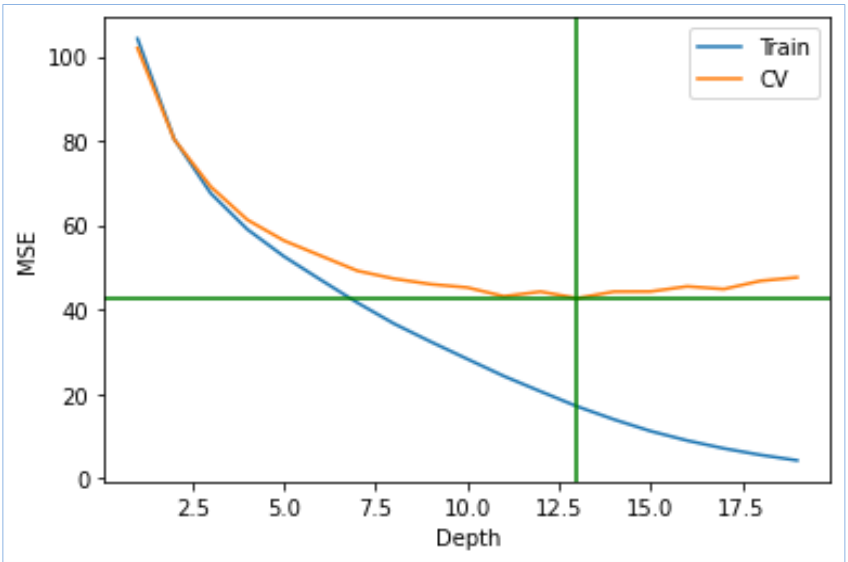

Fig 11

Performance of Random Forest on train & cross validation for different no. of features & max depth 13

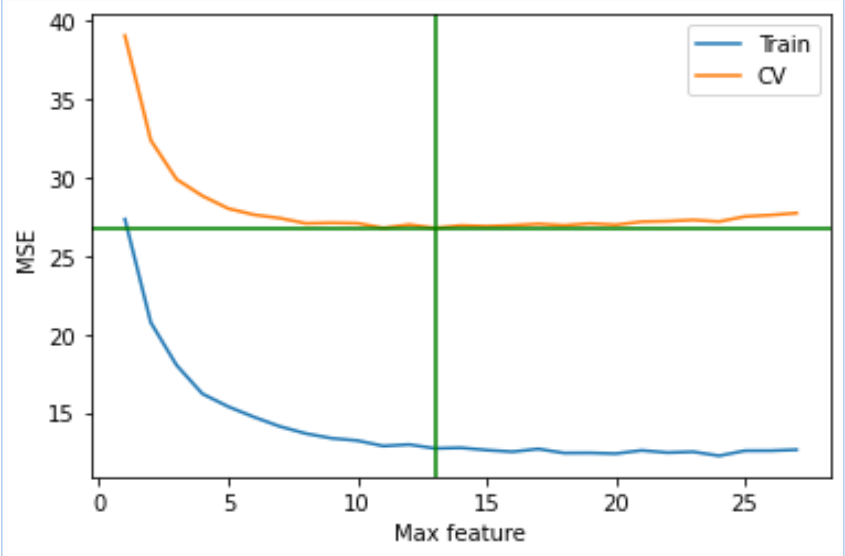

Fig 12

Below table summarizes the performance of these models on cross validation set.

ML MODEL	MSE on CV
Linear Regression	52.81
Ridge Regression	52.81
Decision Tree	46.68
Random Forest	26.78
Mean Line Prediction	958520.56

Table 7

We found random forest performs best on cross validation set with MSE of 26.78.

HYPERPARAMETER OPTIMIZATION USING BAYESIAN OPTIMIZATION

We applied Bayesian Optimization and 5 fold cross validation to find best parameter for Random Forest.
In below table shows the range over which hyper parameters of Random Forest were varied for selection of best parameter.

Hyper parameter	Description	Value
max_depth	maximum depth of trees in RF	10 - 35
n_estimators	no. of tree in RF	80 - 200
max_features	maximum no. of features to be considered for splitting	3 - 28
criterion	criteria for split	mse, friedman_mse

Table 8

Best Set of Parameters for Random Forest

Hyper parameter	Description	Best Value
max_depth	maximum depth of trees in RF	23
n_estimators	no. of tree in RF	187
max_features	maximum no. of features to be considered for splitting	11
criterion	criteria for split	mse

Table 9

Finally we built Random Forest model on train and cross validation combined with best set of hyper parameter and tested it on test data.
We found train MSE = 3.23 and test MSE = 22.43
Next we built the Random forest with the same best set of forest on all data and deployed it locally using gradio library.

Feature Importance

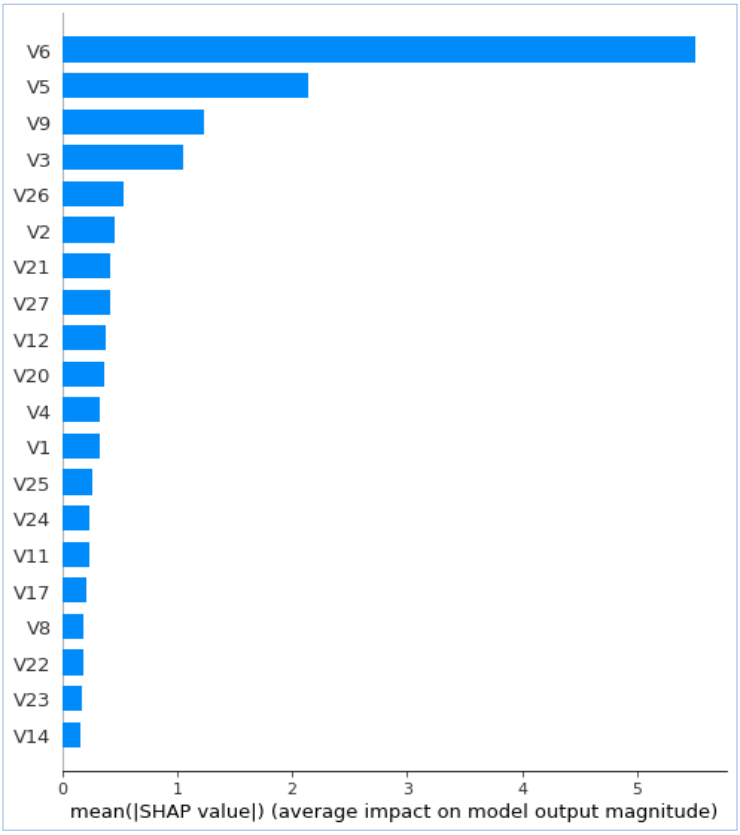

Fig 13

Top 5 most important features we found were V6 followed by V5, V9, V3 and V26.
Where V5 = Old_V4 (PC V4) = 0.28*AT + 0.15*AP + 0.12*AH + 0.65*AFDP - 0.14*GTEP - 0.38*TIT -0.47*TAT - 0.24*TEY - 0.13*CDP

References & Citation

Heysem Kaya, PÄ±nar TÃ¼fekci and ErdinÃ§ Uzun. 'Predicting CO and NOx emissions from gas turbines: novel data and a benchmark PEMS', Turkish Journal of Electrical Engineering & Computer Sciences, vol. 27, 2019, pp. 4783-4796