Gas Turbine Emission

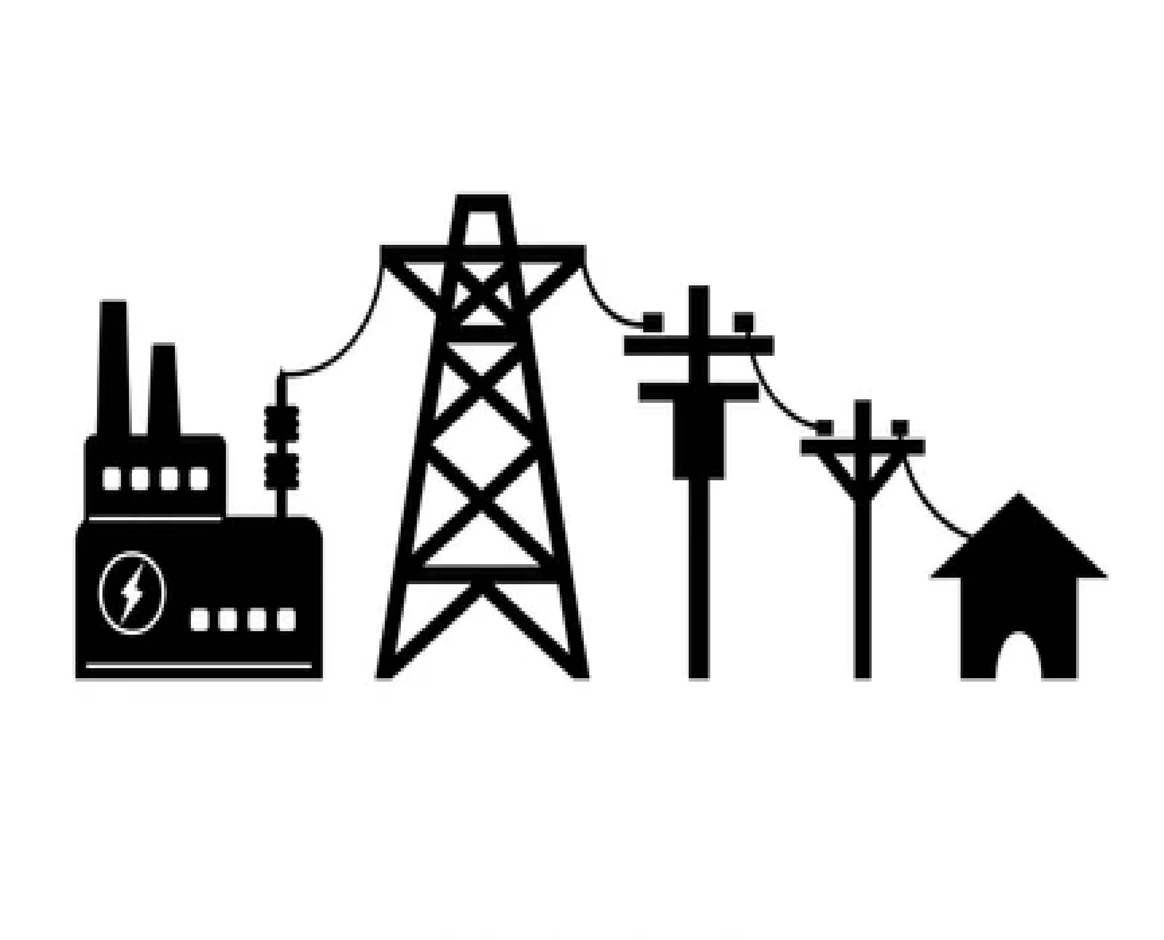
ANALYSIS AND MODELLING OF 5 YEARS' DATA ON CO AND NOX GAS EMISSIONS FROM GAS TURBINES IN A POWER GENERATION PLANT.

Mentor:
Dr. Sayantani Roy Choudhury

Group Members:
Aditya Jaiswal – A22003
Chinmaya Venkataraman – A22014

Research Question

- It is well known that CO and NOx gases are greenhouses gases that are produced in the power generation process and are harmful to the environment.
- Minimisation of emissions is a pressing industry problem and there are regulations with regard to how much can be generated by a power plant.
- We will attempt to predict the most important factors contributing to emissions and provide actionable steps to mitigate them.



Research Objective

- Analysing 5 years' data (recorded from 2011 to 2015) on CO and NOx gas emissions from gas turbines in a power generation plant located in Turkey.
- Identifying the most important features in the dataset.
- Performing operations on the data and fitting **Linear Regression models**.
- Training **Machine Learning models** on the data and comparing results to identify the best model using **RMSE score**.

Data Source

- UC Irvine Machine Learning Repository, Gas Turbine CO and NOx Emission Data Set
- Files: gt_2011.csv, gt_2012.csv, gt_2013.csv, gt_2014.csv, gt_2015.csv
- Link: https://archive.ics.uci.edu/ml/machine-learning-databases/00551/pp_gas_emission.zip

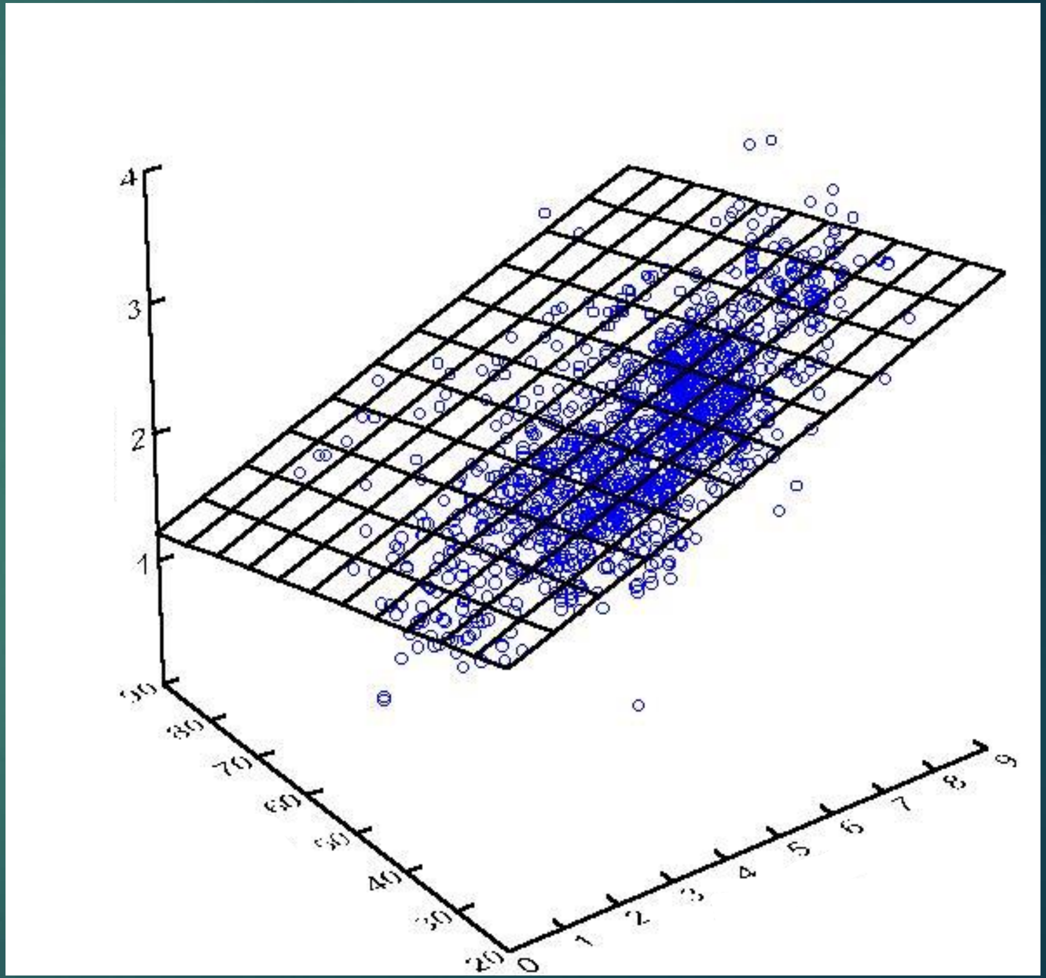
Variables:

Variable	Abbreviation
Ambient temperature	AT
Ambient pressure	AP
Ambient humidity	AH
Air filter difference pressure	AFDP
Gas turbine exhaust pressure	GTEP
Turbine inlet temperature	TIT
Turbine after temperature	TAT
Compressor discharge pressure	CDP
Turbine energy yield	TEY
Carbon monoxide	CO
Nitrogen oxides	NOx

Methodology 1: Linear Regression

Measures Taken

- Outlier Treatment using LOF (Local Outlier Factor)
- Checking for and removing variables that have Variance Inflation factor higher than 5.
- Feature Selection using Lasso Regression.



Results: CO

Lasso Regression coefficients:

```
{'AT': -0.0,
 'AP': 0.0,
 'AH': -0.0,
 'AFDP': -0.0,
 'GTEP': 0.0,
 'TIT': -0.08579434089308881,
 'TAT': -0.018292578288060067,
 'TEY': 0.0,
 'CDP': 0.0}
```

Regression with all variables:

OLS Regression Results

Dep. Variable:	CO	**R-squared:**	0.582
Model:	OLS	**Adj. R-squared:**	0.582
Method:	Least Squares	**F-statistic:**	5536.
Date:	Mon, 30 Jan 2023	**Prob (F-statistic):**	0.00
Time:	02:05:29	**Log-Likelihood:**	-62953.
No. Observations:	35770	**AIC:**	1.259e+05
Df Residuals:	35760	**BIC:**	1.260e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	121.8235	2.119	57.502	0.000	117.671	125.976
AT	-0.0501	0.003	-17.351	0.000	-0.056	-0.044
AP	-0.0020	0.001	-1.389	0.165	-0.005	0.001
AH	-0.0076	0.001	-11.773	0.000	-0.009	-0.006
AFDP	-0.1614	0.015	-10.550	0.000	-0.191	-0.131
GTEP	0.1016	0.010	10.427	0.000	0.082	0.121
TIT	-0.0690	0.003	-26.221	0.000	-0.074	-0.064
TAT	-0.0754	0.003	-21.650	0.000	-0.082	-0.069
TEY	-0.1916	0.008	-25.040	0.000	-0.207	-0.177
CDP	1.9421	0.108	17.977	0.000	1.730	2.154

Omnibus:	45324.595	**Durbin-Watson:**	0.791
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	21766576.561
Skew:	6.593	**Prob(JB):**	0.00
Kurtosis:	123.127	**Cond. No.**	4.52e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.52e+05. This might indicate that there are strong multicollinearity or other numerical problems.

Regression with select variables:

AH
AT

OLS Regression Results

Dep. Variable:	CO	R-squared:	0.033
Model:	OLS	Adj. R-squared:	0.033
Method:	Least Squares	F-statistic:	614.8
Date:	Mon, 30 Jan 2023	Prob (F-statistic):	3.03e-263
Time:	02:05:30	Log-Likelihood:	-77956.
No. Observations:	35770	AIC:	1.559e+05
Df Residuals:	35767	BIC:	1.559e+05
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.7753	0.089	31.036	0.000	2.600	2.951
AT	-0.0476	0.002	-27.406	0.000	-0.051	-0.044
AH	0.0054	0.001	6.069	0.000	0.004	0.007

Omnibus:	33593.422	Durbin-Watson:	0.620
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2933107.711
Skew:	4.298	Prob(JB):	0.00
Kurtosis:	46.521	Cond. No.	640.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

AH
AT
AP
TIT
TAT
TEY

OLS Regression Results

Dep. Variable:	CO	R-squared:	0.553
Model:	OLS	Adj. R-squared:	0.553
Method:	Least Squares	F-statistic:	1.514e+04
Date:	Mon, 30 Jan 2023	Prob (F-statistic):	0.00
Time:	02:05:35	Log-Likelihood:	-67332.
No. Observations:	36733	AIC:	1.347e+05
Df Residuals:	36729	BIC:	1.347e+05
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	164.4816	1.103	149.158	0.000	162.320	166.643
AT	0.0156	0.001	13.315	0.000	0.013	0.018
TIT	-0.1055	0.001	-204.669	0.000	-0.106	-0.104
TAT	-0.0884	0.001	-65.357	0.000	-0.091	-0.086

Omnibus:	47810.815	Durbin-Watson:	0.882
Prob(Omnibus):	0.000	Jarque-Bera (JB):	22010049.939
Skew:	6.977	Prob(JB):	0.00
Kurtosis:	122.104	Cond. No.	1.69e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.69e+05. This might indicate that there are strong multicollinearity or other numerical problems.

Results: NOx

Lasso Regression coefficients:

```
{'AT': -1.292347600752513,
 'AP': -0.07887427101161554,
 'AH': -0.14400977007209276,
 'AFDP': 0.0,
 'GTEP': -0.0,
 'TIT': 0.6276786731438204,
 'TAT': -0.6384334423096925,
 'TEY': -0.9796940258821495,
 'CDP': -0.0}
```

Regression with all variables:

OLS Regression Results

Dep. Variable:	NOX	R-squared:	0.518
Model:	OLS	Adj. R-squared:	0.518
Method:	Least Squares	F-statistic:	4268.
Date:	Mon, 30 Jan 2023	Prob (F-statistic):	0.00
Time:	06:52:20	Log-Likelihood:	-1.2546e+05
No. Observations:	35770	AIC:	2.509e+05
Df Residuals:	35760	BIC:	2.510e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-61.5671	12.162	-5.062	0.000	-85.406	-37.728
AT	-1.7649	0.017	-106.504	0.000	-1.797	-1.732
AP	-0.2360	0.008	-28.839	0.000	-0.252	-0.220
AH	-0.2224	0.004	-60.016	0.000	-0.230	-0.215
AFDP	0.7286	0.088	8.295	0.000	0.556	0.901
GTEP	-0.1015	0.056	-1.815	0.070	-0.211	0.008
TIT	1.4144	0.015	93.655	0.000	1.385	1.444
TAT	-1.5245	0.020	-76.244	0.000	-1.564	-1.485
TEY	-1.9461	0.044	-44.304	0.000	-2.032	-1.860
CDP	-1.9019	0.620	-3.067	0.002	-3.117	-0.686

Omnibus:	7249.078	Durbin-Watson:	0.369
Prob(Omnibus):	0.000	Jarque-Bera (JB):	26041.843
Skew:	0.996	Prob(JB):	0.00
Kurtosis:	6.675	Cond. No.	4.52e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.52e+05. This might indicate that there are strong multicollinearity or other numerical problems.

Regression with select variables:

AH
AT

OLS Regression Results

Dep. Variable:	NOX	R-squared:	0.324
Model:	OLS	Adj. R-squared:	0.324
Method:	Least Squares	F-statistic:	8574.
Date:	Mon, 30 Jan 2023	Prob (F-statistic):	0.00
Time:	06:56:40	Log-Likelihood:	-1.3151e+05
No. Observations:	35770	AIC:	2.630e+05
Df Residuals:	35767	BIC:	2.630e+05
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	90.5104	0.400	226.492	0.000	89.727	91.294
AT	-0.9703	0.008	-125.090	0.000	-0.985	-0.955
AH	-0.1038	0.004	-25.944	0.000	-0.112	-0.096

Omnibus:	5163.068	Durbin-Watson:	0.301
Prob(Omnibus):	0.000	Jarque-Bera (JB):	13428.258
Skew:	0.807	Prob(JB):	0.00
Kurtosis:	5.531	Cond. No.	640.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

AH
AT
AP
TIT
TAT
TEY

OLS Regression Results

Dep. Variable:	NOX	R-squared:	0.517
Model:	OLS	Adj. R-squared:	0.517
Method:	Least Squares	F-statistic:	6546.
Date:	Mon, 30 Jan 2023	Prob (F-statistic):	0.00
Time:	07:01:56	Log-Likelihood:	-1.2904e+05
No. Observations:	36733	AIC:	2.581e+05
Df Residuals:	36726	BIC:	2.582e+05
Df Model:	6		
Covariance Type:	nonrobust		

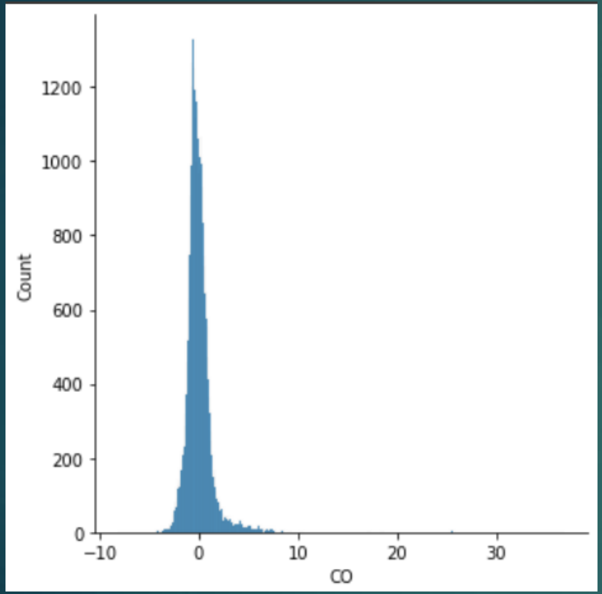
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-83.1243	10.420	-7.978	0.000	-103.547	-62.701
AT	-1.7945	0.010	-174.485	0.000	-1.815	-1.774
AP	-0.2419	0.008	-32.171	0.000	-0.257	-0.227
AH	-0.2174	0.004	-60.970	0.000	-0.224	-0.210
TIT	1.4467	0.014	105.278	0.000	1.420	1.474
TAT	-1.5370	0.016	-96.666	0.000	-1.568	-1.506
TEY	-2.1188	0.019	-113.363	0.000	-2.155	-2.082

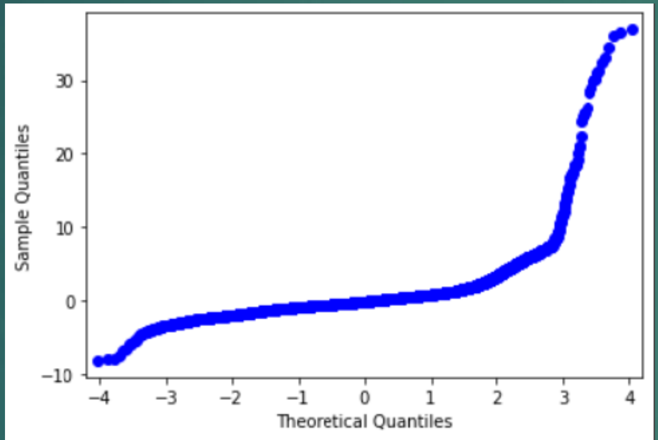
Omnibus:	7100.246	Durbin-Watson:	0.374
Prob(Omnibus):	0.000	Jarque-Bera (JB):	24481.310
Skew:	0.963	Prob(JB):	0.00
Kurtosis:	6.505	Cond. No.	3.90e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.9e+05. This might indicate that there are strong multicollinearity or other numerical problems.

Distribution of errors:

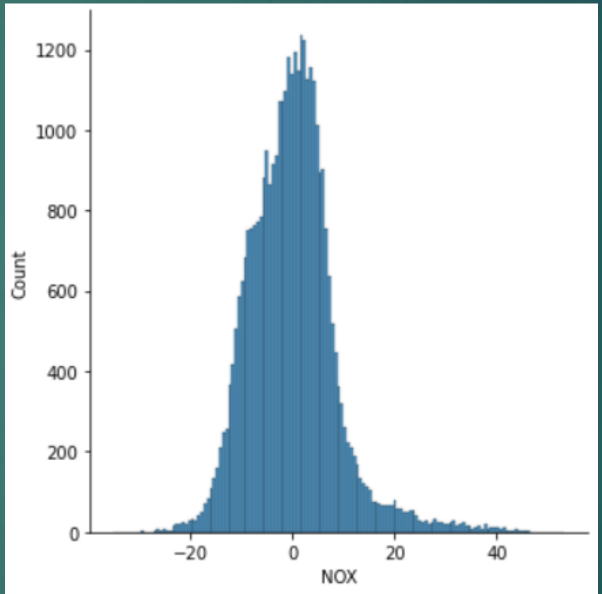
CO

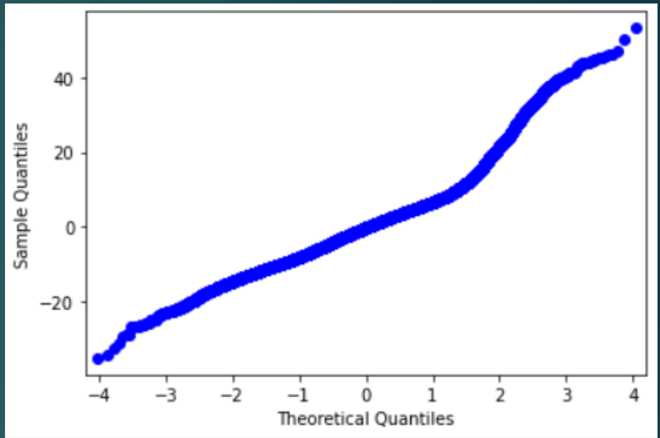




Lasso reg score - 0.5375

NOx





Lasso reg score - 0.4634

Analysis and results

- Durbin-Watson statistic indicates that there is autocorrelation in residuals of both regressions.
- Prediction of NOx has higher error in prediction compared to CO and the errors are normally distributed, indicating that there is information that is not being captured.
- Variance inflation factor is very high for all variables except **Ambient Temperature** and **Ambient Humidity**.
- There is high multicollinearity in the data even among the significant variables.

Values for CO

- RMSE Value, CO - 1.945
- R-squared, Adjusted R-squared – 0.582
- Lasso regression score, CO - 0.5375

Values for NOx

- RMSE Value - 64.184
- R-squared, Adjusted R-squared – 0.518
- Lasso regression score, NOx - 0.4634

Interpretation

- Lasso regression indicates that air filter difference pressure, gas turbine exhaust pressure and compressor discharge pressure are not important for prediction of emissions.
- **Ambient Temperature** and **Ambient Humidity** are not good linear predictors for CO emissions but give good results for prediction of NOx.
- Some of the variables that improve prediction of emissions have very high values for variance inflation factor, such as turbine inlet temperature and turbine after temperature. However, they still are observed to be significant and give better R-squared, AIC and BIC values compared to models that drop these variables.

Methodology 2: Machine Learning

We applied the following algorithms to the data and documented the results and RMSE values after optimising the parameters using the python library, hyperopt:

- Decision Tree
- Random Forest
- Gradient Boosting Machines
- XGBoost (Extreme Gradient Boosting)
- Support Vector Machine

Parameter Search Space and Best Parameters

```
param_dt={
    'max_depth': scope.int(hp.quniform('max_depth',2,20,1)),
    'ccp_alpha': hp.uniform('ccp_alpha',0.001,0.1)
}

param_rf={
    'n_estimators':scope.int(hp.quniform('n_estimators',50,500,1)),
    'max_features':hp.choice('max_features',list(range(2,7)))
}

param_gbm = {
    'max_depth':scope.int(hp.quniform('max_depth',1,6,1)),
    'n_estimators':scope.int(hp.quniform('n_estimators',50,500,1)),
    'learning_rate':hp.uniform('learning_rate',0.001,0.1)
}

param_xgb = {
    'max_depth':scope.int(hp.quniform('max_depth',1,6,1)),
    'n_estimators':scope.int(hp.quniform('n_estimators',50,500,1)),
    'learning_rate':hp.uniform('learning_rate',0.001,0.1),
    'colsample_bytree':hp.uniform('colsample_bytree',0.2,0.8)
}
```

- Decision Tree:
  - CO: ccp_alpha: 0.024, max_depth: 17.0
  - NOx: ccp_alpha: 0.024, 'max_depth': 18.0
- Random Forest:
  - CO: 'max_features': 2, 'n_estimators': 229
  - NOx: 'max_features': 3, 'n_estimators': 317
- Gradient Boosting:
  - NOx: learning_rate: 0.0664, max_depth: 6 n_estimators: 497
  - CO: learning_rate: 0.0663, max_depth: 5, n_estimators: 254
- XgBoost:
  - CO: colsample_bytree: 0.797, learning_rate: 0.098, max_depth: 5, n_estimators: 380
  - NOx: colsample_bytree: 0.751, learning_rate: 0.082, max_depth: 6, n_estimators: 420

Analysis and Results:

RMSE Values for CO:-

- Random Forest:1.048
- XGBoost:1.210
- Gradient Boosting: 1.324
- SVM: 1.478 (linear)
- Decision Tree: 1.725
- Linear Regression: 1.945

RMSE Values for NOx:-

- Random Forest:16.901
- Gradient Boosting: 18.438
- XGBoost:18.797
- Decision Tree: 31.0495
- Linear Regression: 64.184
- SVM: 65.546 (linear)

Interpretation:

- Random forest performs best for prediction of emission of CO and NOx, giving much better results than all other models
- CO prediction shows some degree of success with linear models while the best models are still the non-linear ones
- NOx prediction shows the best results for non-linear models and the worst with linear models.

CONCLUSION

- There are strong non-linear relationships between emissions of CO and NOx and the predictor variables, which were best predicted using a random forest model.
- Ambient Temperature and Ambient Humidity are seen to be negatively correlated to emissions.
- Air filter difference pressure was seen to be negatively correlated to CO but positively correlated to NOx.
- Ambient Pressure is insignificant even at 90% confidence level for CO prediction and gas turbine exhaust pressure is insignificant at 95% confidence level but significant at 90% confidence level.
- CO and NOx have been seen to steadily increase from 2011 to 2015, therefore measures need to be taken to mitigate this.

THANK YOU!