Gas Turbine CO and NOx Emission Analysis

Aayushi Gupta, Kyle Kaminski, Rosa Lin, Ruben Martinez

Introduction

The combined cycle power plant, also known as combined cycle gas turbine plant, is an assembly of heat engines that combine to generate electricity (Tüfekci). A combined-cycle power plant (CCPP) is made up of gas turbines, steam turbines, and heat recovery steam generators. The electricity is generated and combined in one cycle by gas and steam turbines and then transferred from one turbine to another.

We are interested in identifying the process variables that impact carbon monoxide emissions. By determining the process variables that impact carbon monoxide emissions we will be able to find opportunities to reduce carbon monoxide emissions.

Gas Turbine CO and NOx Emission Data Set

The data comes from a gas turbine located in Turkey that studies the flue gas emissions of specifically carbon monoxide (CO) and nitrogen oxide (NOx) gases. The data set provides hourly statistics of 11 sensors. Data points were collected from a gas turbine from Jan 01 2011 to Dec 13 2015.

Description

The data file `gt_2015.csv` has 7384 observations and 11 variables from the UCI Gas Turbine CO and NOx Emission Data Set. We are going to explore and analyze the following variables:

- `AT` - Ambient Temperature
- `AP` - Ambient Pressure
- `AH` - Ambient Humidity
- `AFDP` - Air filter difference pressure
- `GTEP` - Gas turbine exhaust pressure
- `TIT` - Turbine inlet temperature
- `TAT` - Turbine after temperature
- `TEY` - Turbine energy yield
- `CDP` - Compressor discharge pressure

Here's a quick peek at the data set:

AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO	NOX
1.95320	1020.1	84.985	2.5304	20.116	1048.7	544.92	116.27	10.799	7.4491	113.250
1.21910	1020.1	87.523	2.3937	18.584	1045.5	548.50	109.18	10.347	6.4684	112.020
0.94915	1022.2	78.335	2.7789	22.264	1068.8	549.95	125.88	11.256	3.6335	88.147
1.00750	1021.7	76.942	2.8170	23.358	1075.2	549.63	132.21	11.702	3.1972	87.078
1.28580	1021.6	76.732	2.8377	23.483	1076.2	549.68	133.58	11.737	2.3833	82.515
1.83190	1021.7	76.411	2.8410	23.495	1076.4	549.92	133.58	11.829	2.0812	81.193

Here's some descriptive statistics of the data set:

```
##        AT               AP               AH              AFDP
```

```
##  Min.   :-6.235   Min.   : 989.4   Min.   :24.09   Min.   :2.369
##  1st Qu.:11.073   1st Qu.:1009.7   1st Qu.:59.45   1st Qu.:3.117
##  Median :17.456   Median :1014.0   Median :70.95   Median :3.538
##  Mean   :17.225   Mean   :1014.5   Mean   :68.65   Mean   :3.599
##  3rd Qu.:23.685   3rd Qu.:1018.3   3rd Qu.:79.65   3rd Qu.:4.195
##  Max.   :37.103   Max.   :1036.6   Max.   :96.67   Max.   :5.239
##       GTEP            TIT            TAT             TEY
##  Min.   :17.70   Min.   :1016   Min.   :516.0   Min.   :100.0
##  1st Qu.:23.15   1st Qu.:1070   1st Qu.:544.7   1st Qu.:126.3
##  Median :25.33   Median :1080   Median :549.7   Median :131.6
##  Mean   :26.13   Mean   :1079   Mean   :546.6   Mean   :134.0
##  3rd Qu.:30.02   3rd Qu.:1100   3rd Qu.:550.0   3rd Qu.:147.2
##  Max.   :40.72   Max.   :1100   Max.   :550.6   Max.   :179.5
##       CDP              CO               NOX
##  Min.   : 9.871   Min.   : 0.2128   Min.   : 25.91
##  1st Qu.:11.466   1st Qu.: 1.8082   1st Qu.: 52.40
##  Median :11.933   Median : 2.5334   Median : 56.84
##  Mean   :12.097   Mean   : 3.1300   Mean   : 59.89
##  3rd Qu.:13.148   3rd Qu.: 3.7026   3rd Qu.: 65.09
##  Max.   :15.159   Max.   :41.0970   Max.   :119.68
```

Goals

The goal for this project is to utilize this data set for the purpose of studying flue gas emissions, specifically carbon monoxide(CO) and nitrogen oxides (NOx). Our focus will be to find statistically significant relationships between the ambient and turbine variables and the emissions variables. We will limit the size of our model to more clearly demonstrate these relationships. Ultimately we will suggest which variables make the biggest impact on emission levels in order to decrease emissions overall.

Exploratory Data Analysis

Relationships between feature variables

Figure 1: Scatterplot Matrices to decide which feature variables have a linear relationship

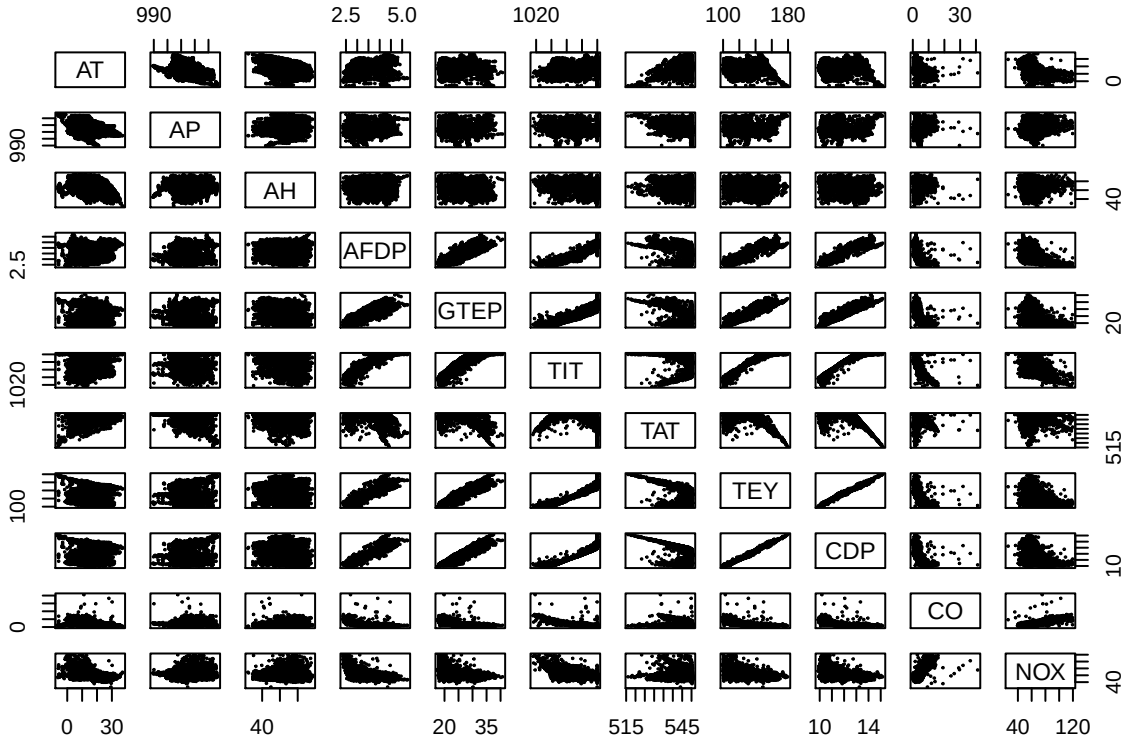

Figure 2:

Table 2: Pairwise Correlation Between Variables

	AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO	NOX
AT	1.00	-0.49	-0.47	0.47	0.19	0.33	0.21	0.11	0.20	-0.39	-0.59
AP	-0.49	1.00	0.08	-0.09	-0.04	-0.08	-0.29	0.05	0.03	0.20	0.21
AH	-0.47	0.08	1.00	-0.25	-0.30	-0.26	0.03	-0.18	-0.22	0.16	0.07
AFDP	0.47	-0.09	-0.25	1.00	0.84	0.92	-0.52	0.88	0.92	-0.64	-0.58
GTEP	0.19	-0.04	-0.30	0.84	1.00	0.89	-0.62	0.93	0.94	-0.56	-0.37
TIT	0.33	-0.08	-0.26	0.92	0.89	1.00	-0.40	0.95	0.95	-0.74	-0.52
TAT	0.21	-0.29	0.03	-0.52	-0.62	-0.40	1.00	-0.63	-0.66	0.03	0.05
TEY	0.11	0.05	-0.18	0.88	0.93	0.95	-0.63	1.00	0.99	-0.62	-0.40
CDP	0.20	0.03	-0.22	0.92	0.94	0.95	-0.66	0.99	1.00	-0.61	-0.44
CO	-0.39	0.20	0.16	-0.64	-0.56	-0.74	0.03	-0.62	-0.61	1.00	0.68
NOX	-0.59	0.21	0.07	-0.58	-0.37	-0.52	0.05	-0.40	-0.44	0.68	1.00

Remove variables that are highly correlated.

```
##       AT       AP       AH     AFDP     GTEP      TAT
## 3.866424 1.600597 1.718769 7.412520 5.909197 2.301015
```

Exploratory analysis shows possible linear relationships between the response variable CO and the feature variables CDP, TEY, TIT, GTEP and AFDP. Collinearity between some of the feature variables (TIT, CDP, and TEY) could cause some problems in our analysis and will likely lead to the removal of the redundant variables.

Methods

Linear Regression

We will create a multiple linear regression model using all feature variables mentioned in the description of Section 1. The implementation and parameters of this model can be obtained by the following equation where we will find estimates for the parameters β using:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Key assumptions are stated as:

- **L**inearity: can be written as a linear combination of the predictors.
- **I**ndependence: the errors are independent of each other (not highly correlated).
- **N**ormality: the distribution of the errors follow a normal distribution.
- **E**qual Variance: the error variance is the same.[1]

We will then use model selection using backward BIC to tune our model and remove any insignificant predictor variables. This selection prefers smaller models which aligns with our goal of limiting the size of our final model.

[1]Dalpiaz David, Applied Statistics in R, https://daviddalpiaz.github.io/appliedstats/model-diagnostics.html