Aayushi Gupta, Kyle Kaminski, Rosa Lin, Ruben Martinez

## Section 1: 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.

Our plan is to analyze a dataset that contains 7384 instances of 11 sensor measures that have been aggregated over one hour (by means of average or sum) from a gas turbine located in Turkey for the purpose of studying flue gas emissions, namely CO and NOx (NO + NO2). The data comes from the same power plant as the dataset (Source) used for predicting hourly net energy yield. By contrast, this data is collected in another data range (01.01.2011 - 31.12.2015), includes gas turbine parameters (such as Turbine Inlet Temperature and Compressor Discharge pressure) in addition to the ambient variables. Note that the dates are not given in the instances but the data are sorted in chronological order. See the attribute information and relevant paper for details. Kindly follow the protocol mentioned in the paper (using the first three years' data for training/ cross-validation and the last two for testing) for reproducibility and comparability of works. The dataset can be well used for predicting turbine energy yield (TEY) using ambient variables as features.

### Goal

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.

### 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 |

**Section 2: Exploratory Descriptive Analysis**

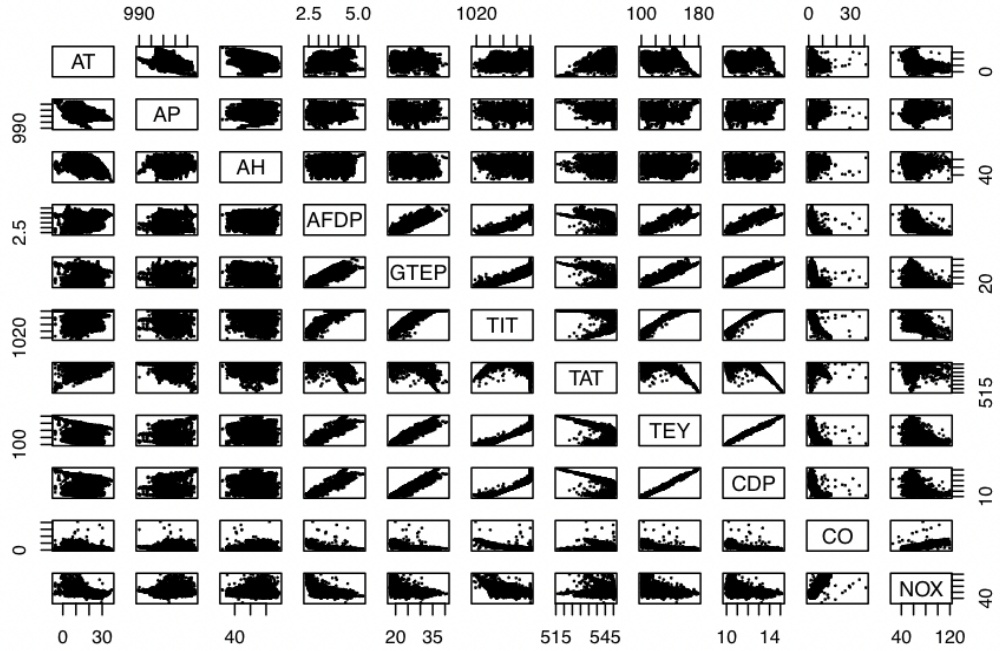


Table 2: Pairwise Correlation Between Variables

| | 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 |

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.

**Section 3: 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 $\beta$ 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. (insert source, STAT 420 textbook)

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.

**Lasso**

The Lasso model is similar in structure to the linear model, but it differs in how the variable selection process is treated. Lasso models often perform better than a simple/multiple linear regression because the Lasso model can penalize unimportant variables by shrinking their corresponding coefficients, which decreases the influence those variables have on the model. This is preferable over the linear regression model because the variance can be decreased without largely impacting the model's bias.

$$\underset{\theta}{argmin}\ SSE + \lambda \sum_{i=1}^{K} |\theta_i|$$

**Tree based models**

Decision trees are nonparametric models and work by taking in all of the characteristics of the observations, and then splitting the data into separate groups based on the optimal splitting characteristics. These models are called decision tree models because each split can be thought of as a branch in a tree. The leaves are thus called terminal nodes in this model because that is where the model outputs the prediction based on all the splitting criteria up until

that point. A decision tree can be used to predict both categorical outcomes and quantitative outcomes. In this analysis, we are looking for a numeric outcome so a regression tree is used.

$$\text{Gini}(K) = \sum_{i \in N} P_{i,K}(1 - P_{i,K}) = 1 - \sum_{i \in N} P_{i,K}^2$$

**Citation:**

Pınar Tüfekci, Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods, International Journal of Electrical Power & Energy Systems, Volume 60, September 2014, Pages 126- 140, ISSN 0142-0615

**Current Progress:**

We have written the introduction, objectives and methodology, cleaned and imported the data, provided a few visuals, started our methods section, and will begin to fit a multiple linear regression model, lasso model, and decision tree model. Current results are shown above. After fitting those models, we will determine which results are the most suitable for our client and write our conclusion.

**Unresolved issues:**

Multicollinearity in scatter plot (Resolve using VIF, remove variables)

**Questions:**

Focus solely on carbon monoxide or also include nitrogen?
Any specific methods the client wants us to do?
Are they okay with our current visuals?
Will you be presenting these results to someone else?
Do you have any questions for us?