Gas Turbine CO Emission Analysis

Aayushi Gupta, Kyle Kaminski, Rosa Lin, Ruben Martinez

5/5/2021

Contents

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 and 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). However, our client did tell us to not consider nitrogen oxide, so we will only be focusing on carbon monoxide in this report. 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 (more details in Appendices 1):

- 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
- CO - Carbon Monoxide
- NOX - Nitrogen Oxide (Removed from data)

Here's a quick peek at the data set:

—— **ADD TABLE HERE** ——

Exploratory Data Analysis

Pairwise Correlations

—— ADD TABLE 1 HERE ——

—— ADD TABLE 2 HERE ——

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

Carbon Monoxide Distribution

The client provided us a set of production ranges to analyze. An overall production range that analyzes all of the data points from the carbon monoxide emission output, a typical production range which looks at data points from 130 to 136, and a high production range that looks at data points higher than 160.

—— ADD FIGURE 1 HERE ——

The typical production range the client provided did not fully capture the typical production range that we observed in our data sample (see Figure 1 above). This could be a result of the data values from the 2015 data set having lower values compared to other data sets. Therefore, we decided to shift the typical production range to 127 to 133 given that it is a better representation of the typical production range of the carbon monoxide emission output.

Data Preparation

The first step to preparing the data was to remove the response variable nitrogen oxide, since our analysis solely focuses on carbon monoxide emissions.

Since we were able to anticipate variables that could cause some problems in our linear based analyses due to collinearity, we decided to remove the following variables from our linear based models:

- TIT
- CDP
- TEY

For all of our models, we split our data into training and testing dataset to avoid overfitting the models. By doing so, we minimized the effects of data discrepancies and effectively evaluated our models.

Results

Overall Decision Tree Model

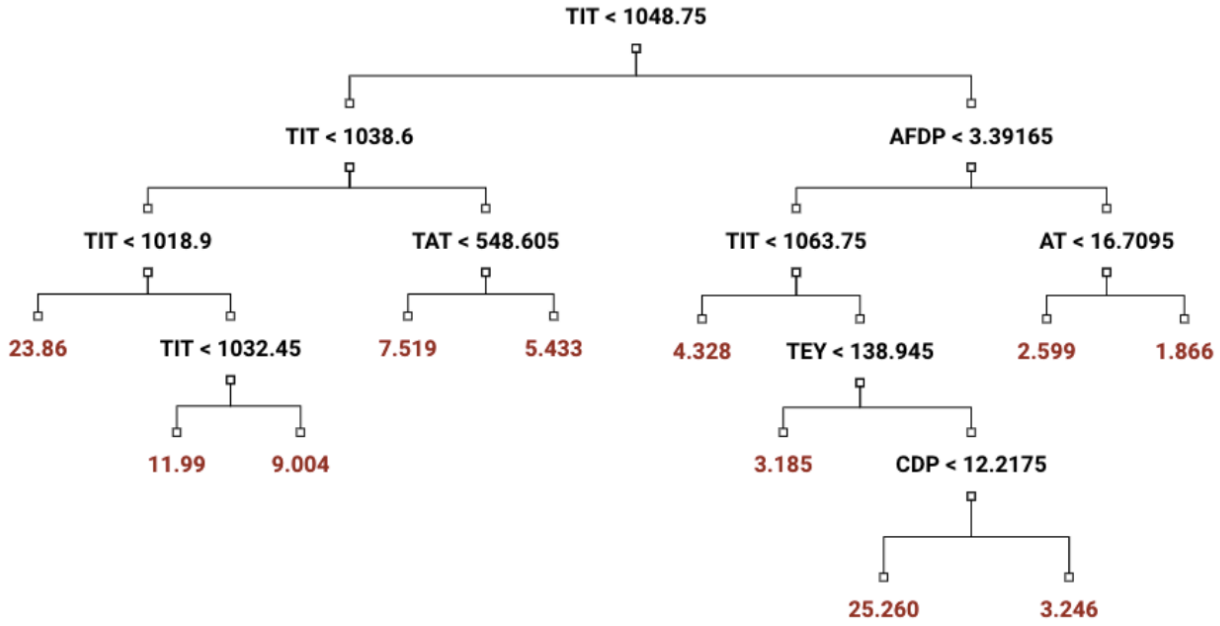

Figure 1: Overall Production Range Decision Tree

The decision tree above represents our final tree model that was trained on the entire data set supplied to us. The first split the tree made was on the turbine inlet temperatures, separating observations where the TIT was less than 1048.75 to the left and the remaining observations to the right. If we observe all of the terminal nodes on each side of the tree after this first split, it is clear that the higher TIT values resulted in lower CO values with the exception of 1 observation where the CO output was very high at 25.26. This value is an anomaly and we do believe it to be a result of incorrect data entry (maybe it should have been 2.526), or some equipment malfunction (see section 5). Just like the TIT values, It is also observed that higher TAT, AT, and AFDP values also result in lower CO output as well.

Typical Decision Tree Model

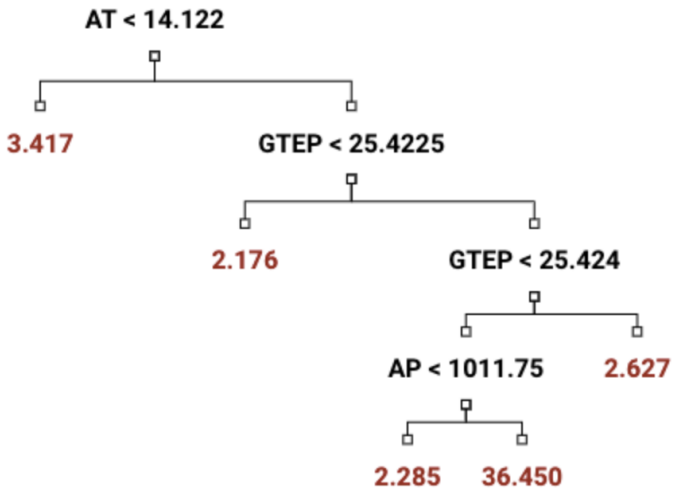

Figure 2: Typical Production Range Decision Tree

This decision tree represents our final tree model that was trained on the typical energy production range with TEY values between 127 and 133. This tree first split on AT, and actually terminates when the AT is less than 14.122 with an output, arguing that AT is likely the most important variable in this energy production range with the higher AT values resulting in lower CO, agreeing with our tree in the previous slide. Again, the single anomaly is a CO output with a very high value at 36.45, again believed to be an error in the data. We also observe that the lower GTEP values look to result in lower CO output, same being with AP, however because of that anomaly AP is not a strong argument.

High Decision Tree Model

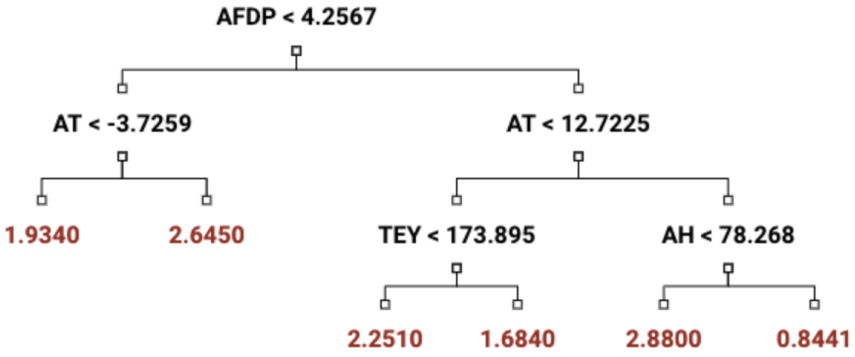

Figure 3: High Production Range Decision Tree

And finally, this decision tree represents our final tree model that was built on the high production range data with TEY values over 160. This tree argues that higher AFDP values on average result in lower CO output because the average value of the nodes on the right side is lower than those on the left. Unlike our previous models, AT does not show a very strong relationship with the CO output values. Higher TEY and AH values look to have lower CO outputs.

Appendix

Variance Inflation Factor (VIF)

Variance Inflation Factor detects multicollinearity in regression analysis. Multicollinearity is when the correlation between predictors affects regression results. We only used VIF in our linear based models.

VIF for overall production range

—— **ADD VIF 1 HERE** ——

VIF for typical production range

—— **ADD VIF 2 HERE** ——

VIF for high production range

—— **ADD VIF 3 HERE** ——