{Gas Turbine CO Emission Analysis}

Aayushi Gupta, Kyle Kaminski, Rosa Lin, Ruben Martinez

Client: Darren Glosemeyer

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 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, turbine, and 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:

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

Methods

Exploratory Data Analysis

Pairwise Correlations

Figure 1: Pairwise Correlation Plot

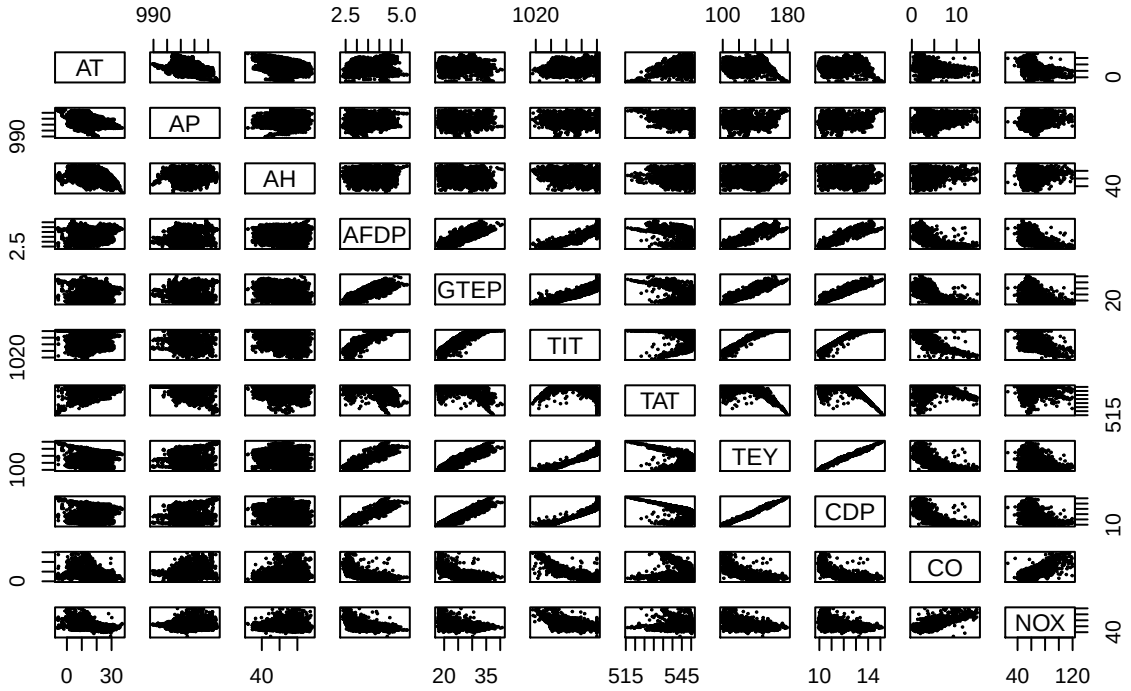


Table 2: Pairwise Correlation Between Variables

	AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO	NOX
AT	1.00	-0.49	-0.47	0.47	0.19	0.33	0.20	0.11	0.20	-0.43	-0.59
AP	-0.49	1.00	0.08	-0.09	-0.04	-0.08	-0.29	0.05	0.03	0.23	0.22
AH	-0.47	0.08	1.00	-0.25	-0.30	-0.26	0.02	-0.18	-0.22	0.20	0.07
AFDP	0.47	-0.09	-0.25	1.00	0.84	0.92	-0.53	0.88	0.92	-0.71	-0.58
GTEP	0.19	-0.04	-0.30	0.84	1.00	0.89	-0.63	0.93	0.94	-0.62	-0.36
TIT	0.33	-0.08	-0.26	0.92	0.89	1.00	-0.41	0.95	0.95	-0.80	-0.51
TAT	0.20	-0.29	0.02	-0.53	-0.63	-0.41	1.00	-0.65	-0.67	0.06	0.07
TEY	0.11	0.05	-0.18	0.88	0.93	0.95	-0.65	1.00	0.99	-0.68	-0.40
CDP	0.20	0.03	-0.22	0.92	0.94	0.95	-0.67	0.99	1.00	-0.67	-0.44
CO	-0.43	0.23	0.20	-0.71	-0.62	-0.80	0.06	-0.68	-0.67	1.00	0.72
NOX	-0.59	0.22	0.07	-0.58	-0.36	-0.51	0.07	-0.40	-0.44	0.72	1.00

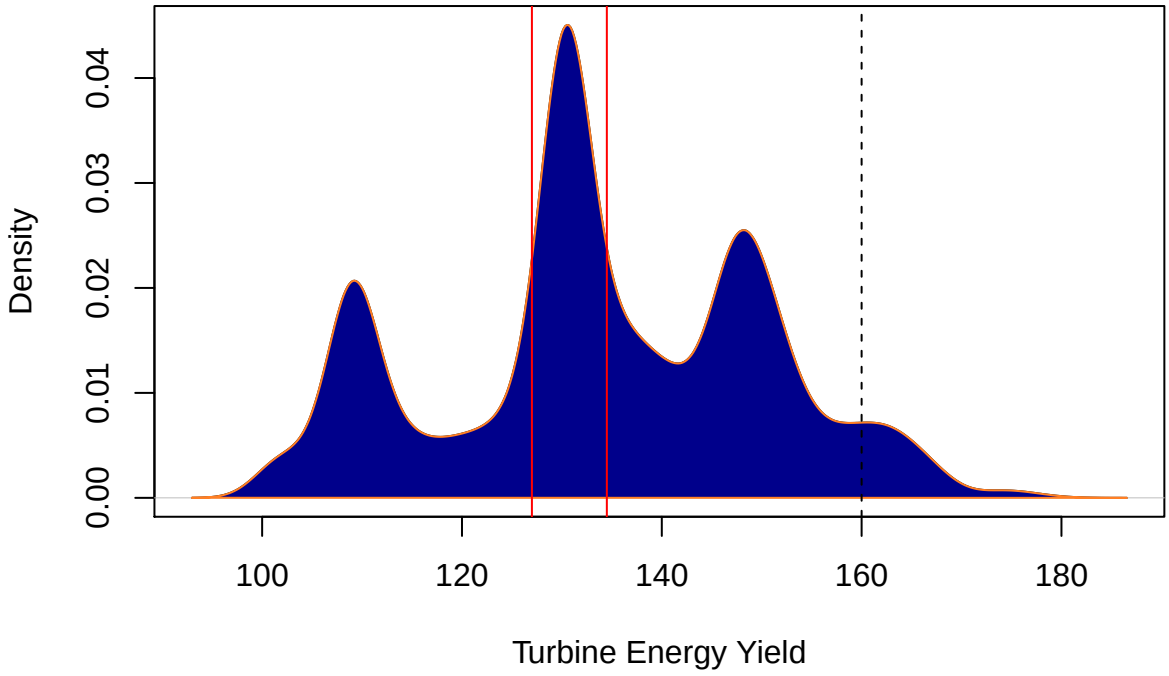
The exploratory analysis shows possible linear relationships between the response variable CO and the feature variables CDP, TEY, TIT, GTEP and AFDP. The analysis also indicates possible collinearity between some

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.

Figure 2: Turbine Energy Yield Distribution



The typical production range the client provided did not fully capture the typical production range that we observed in our data sample (see Figure 2 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 134.5 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

Model Selection

To accurately identify the process variables that impact carbon monoxide emissions, we decided to examine three different models to make sure that the model we selected was the most useful and effective way of analyzing the data set. The three models we used were **Multiple Linear Regression**, **Lasso**, and **Decision Trees**.

RMSE

In order to determine which model was the most effective, we compared the RMSE of multiple linear regression, lasso, and decision tree models. Root Mean Squared Errors are the standard deviation of residuals. The point of calculating the RMSE is to measure how spread out these variables are. The rule of thumb is, the lower the RMSE, the better.

Training and Testing Data

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

Results

Decision Tree Model Selection

In the table below, the decision tree outperforms linear regression and lasso in the overall production range, typical production range, and high production range. Therefore, we decided to use the Decision Tree Model to examine the biggest impact on emission levels in order to decrease emissions overall.

	Overall Production Range	Typical Production Range (127-134.5)	High Production Range (160+)
Linear Regression	1.1088	0.6284	0.4442
Lasso	1.7272	0.7946	0.4780
Decision Tree	0.8843	0.6442	0.3462

Overall Decision Tree Model

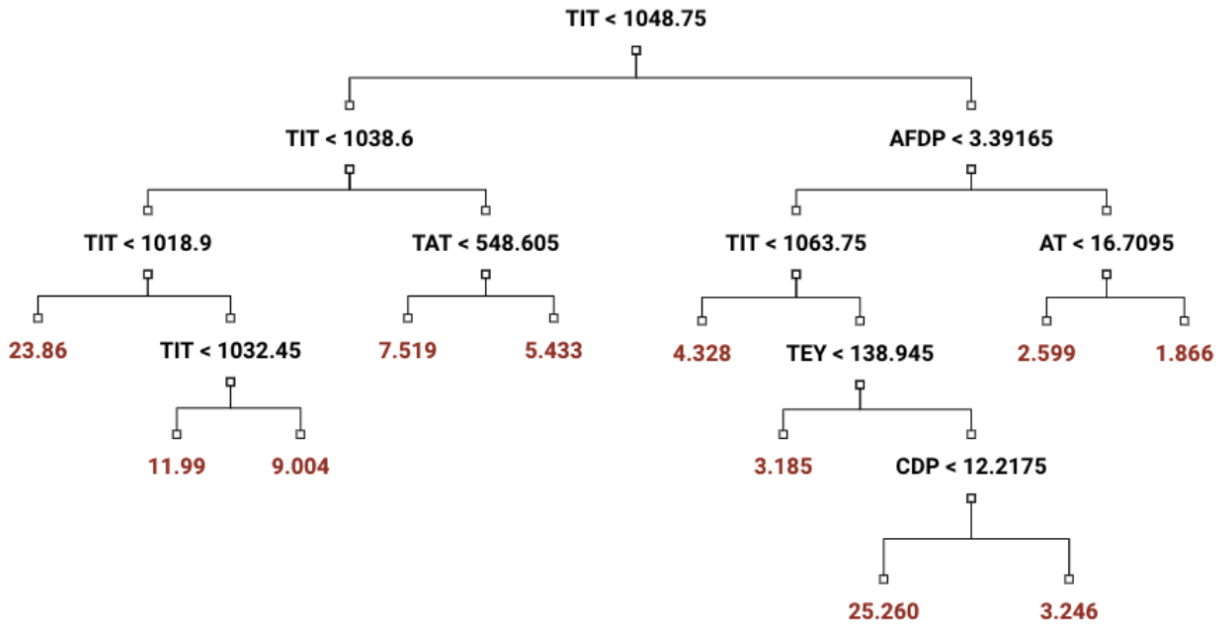


Figure 1: Overall Production Range Decision Tree

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). Similar to the TIT values, it is also observed that higher TAT, AT, and AFDP values also resulted in lower CO output as well.

Typical Decision Tree Model

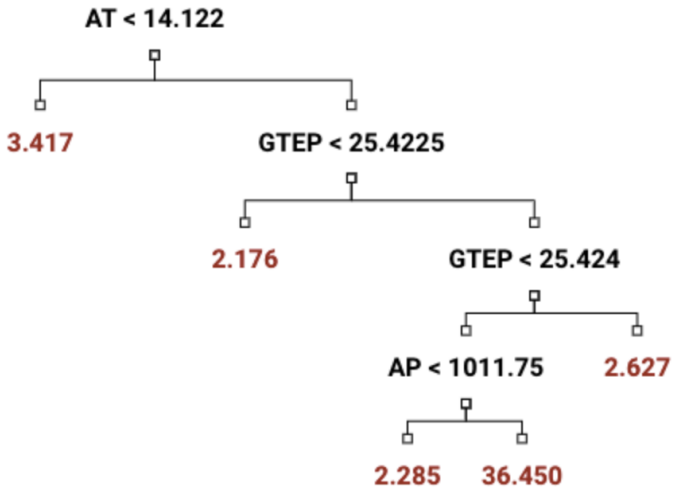


Figure 2: Typical Production Range Decision Tree

This decision tree represents our final tree model that was trained on the typical energy production range with TEY values between 127 and 134.5. 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

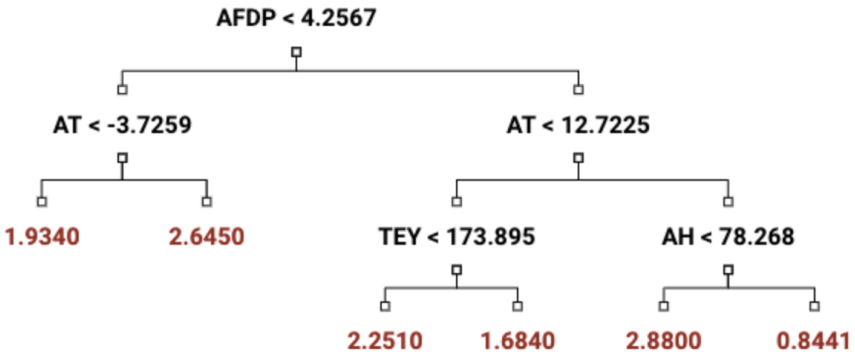


Figure 3: High Production Range Decision Tree

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.

Tree Model Explanations

Carbon Monoxide Correlations

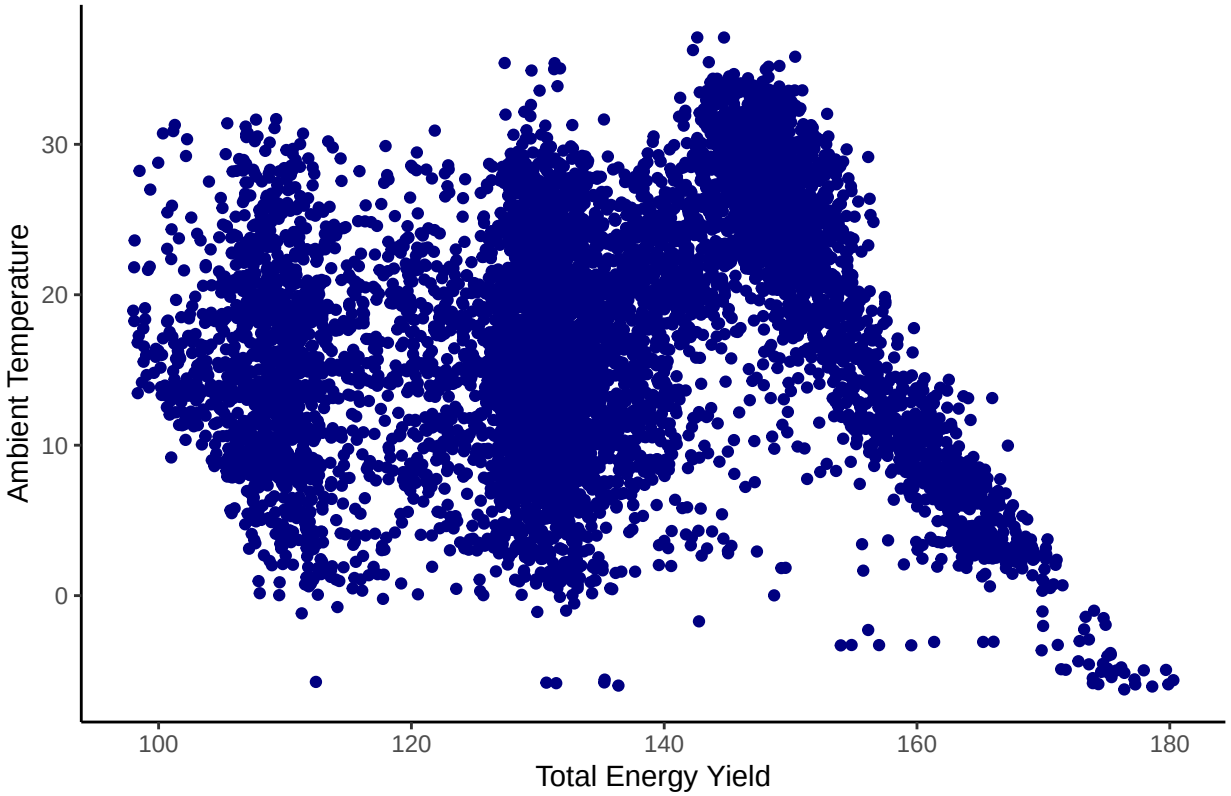
—— ADD TABLE 6 HERE ——

The table above shows us the single correlations between CO and the explanatory variables that were used in the tree model splits. We found it surprising that high TIT, TAT, and AT resulted in lower CO outputs as we would think that cooler temperatures would result in more efficient energy production. The correlation table does show us that TIT and AT both have negative correlations with CO, meaning as one increases the other decreases. This supports the arguments made in our tree models. TAT has an almost 0 correlation with CO according to the data, which does not agree or disagree with our models since the TAT splits were always after other splits.

Ambient Temperature Plots

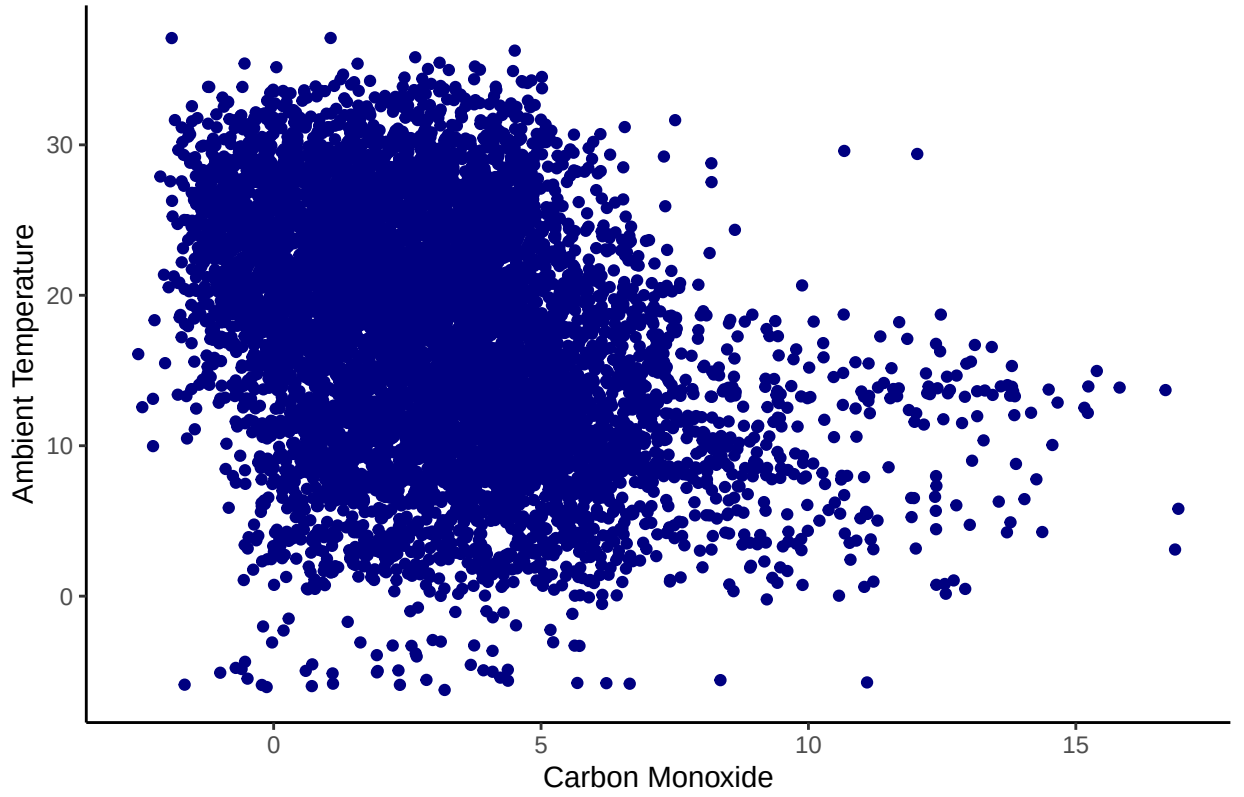
```
model = lm(gt_2015$AT ~ gt_2015$TEY)
ggplot(model, aes(x = gt_2015$TEY, y = gt_2015$AT), xlim = c(100,180)) + geom_jitter(width = 3, color =
```

Figure 6: Relationship between Total Energy Yield and Ambient Temperatur



```
model2 = lm(gt_2015$AT ~ gt_2015$CO)
ggplot(model, aes(x = gt_2015$CO, y = gt_2015$AT), xlim = c(100,180)) + geom_jitter(width = 3, color =
```

Figure 7: Relationship between Carbon Monoxide and Ambient Temperature



The plots above help further explain the why higher AT result in lower CO outputs. The plot on top shows no clear trend with TEY values less than around 145, but a clear negative trend when the TEY values are above 145. This argues that the extreme TEY values are impacted by AT, likely because at very high temperatures the machines can not work as hard or they will overheat. The plot on the bottom shows AT against CO, where we see 9 CO values greater than 17 whereas the other 7300 observations are less.

Conclusion

Most Sensitive Process Variables

Based on our results, the following variables are the most sensitive process variables for the overall production range:

- NA
- NA

Based on our results, the following variables are the most sensitive process variables for the typical production range (127 - 134.5):

- NA
- NA

Based on our results, the following variables are the most sensitive process variables for the high production range (160+):

- NA
- NA

As you can see...

Suggestions

Appendix

Multiple Linear Regression

We will create a multiple linear regression model using the feature variables remaining after preparing our data – AT, AP, AH, AFDP, GTEP, and TAT. The implementation and parameters of this model can be obtained by the following equation where we will find estimates for the parameters β using:

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

[**Source**]

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.

We will then use model selection using VIF 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.

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

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

```
##       AT       AP       AH     AFDP     GTEP      TAT
## 3.849314 1.608266 1.723129 7.382685 5.916211 2.357876
```

VIF for typical production range

```
##       AP       AH     AFDP     GTEP      TAT      TEY      CDP
## 1.179915 1.406863 3.351823 1.496232 1.038742 1.836942 4.823918
```

VIF for high production range

```
##       AT       AP       AH     AFDP     GTEP      TIT      TAT
## 3.084771 1.948954 1.321587 1.996187 1.897431 1.027967 4.191605
```

[**Source**]

Decision Tree

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.

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

[**Source**]

Correlations

Correlation is a statistical measure that measures which two variables are linearly related. It is commonly used to describe simple relationships without discussing the cause and effect.

$$r = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2 \sum(y - m_y)^2}}$$

[**Source**]

Individual Contributions:

Aayushi:

Kyle:

Rosa:

Ruben