{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

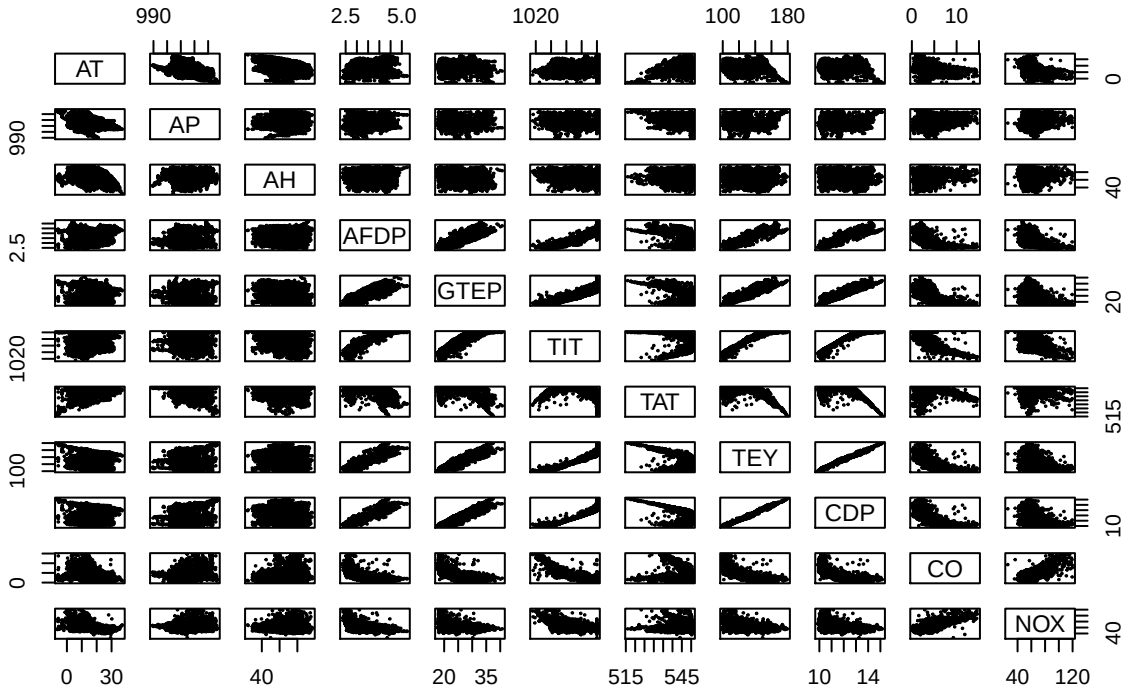


Table 1: Pairwise Correlation Between Variables

	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

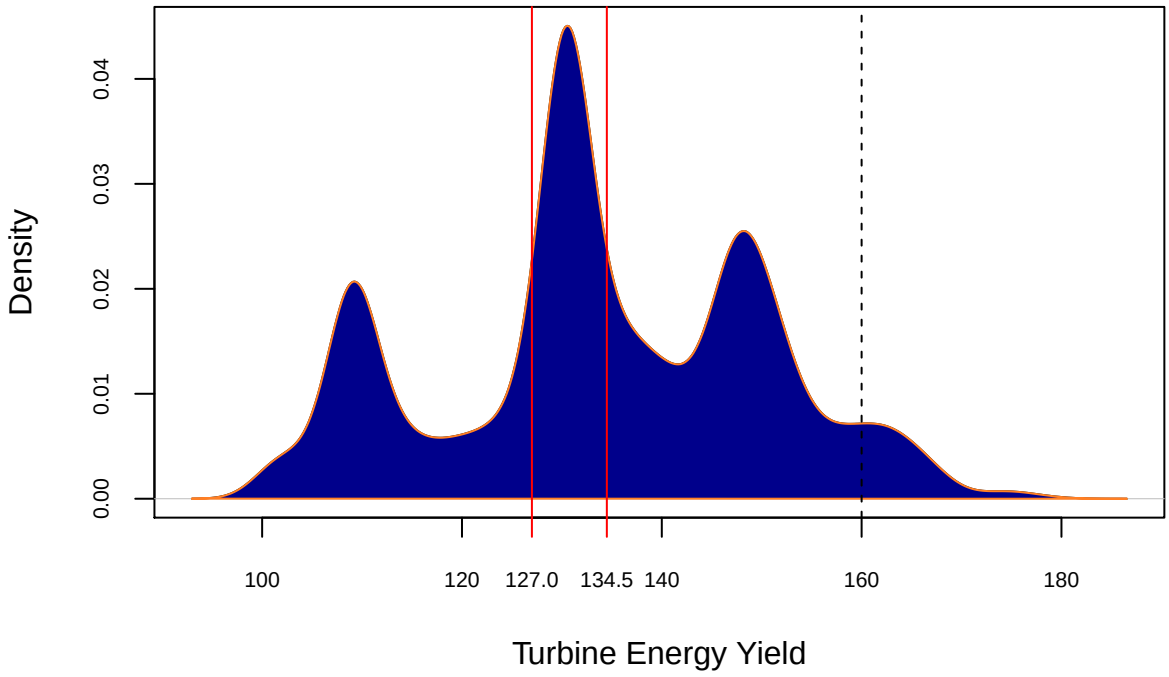
The exploratory analysis shows possible linear relationships between the response variable CO and the feature

variables: CDP, TEY, TIT, GTEP and AFDP. The analysis 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.

Turbine Energy Yield Distribution

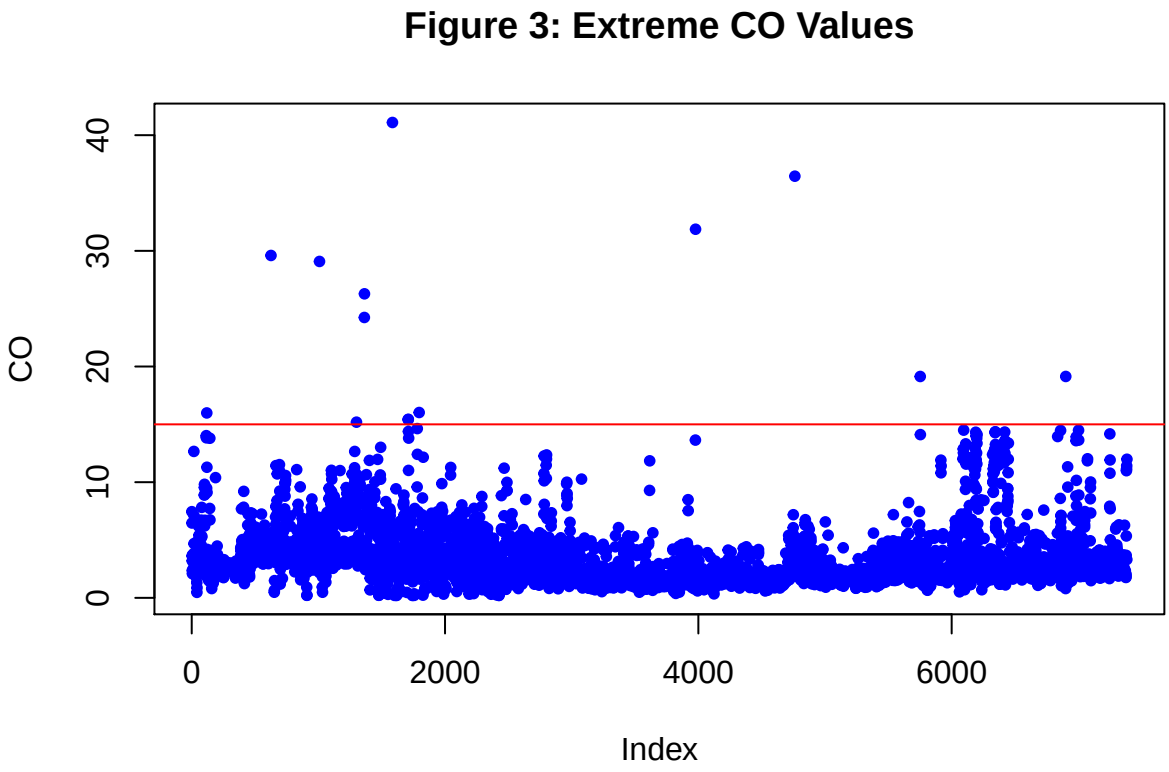
The client provided us a set of turbine energy production ranges to analyze. An overall production range that analyzes all data, a typical production range which looks at data connected to energy yields from 130 to 136, and a high production range that looks at data connected to energy yields higher than 160.

Figure 2: Turbine Energy Yield Distribution



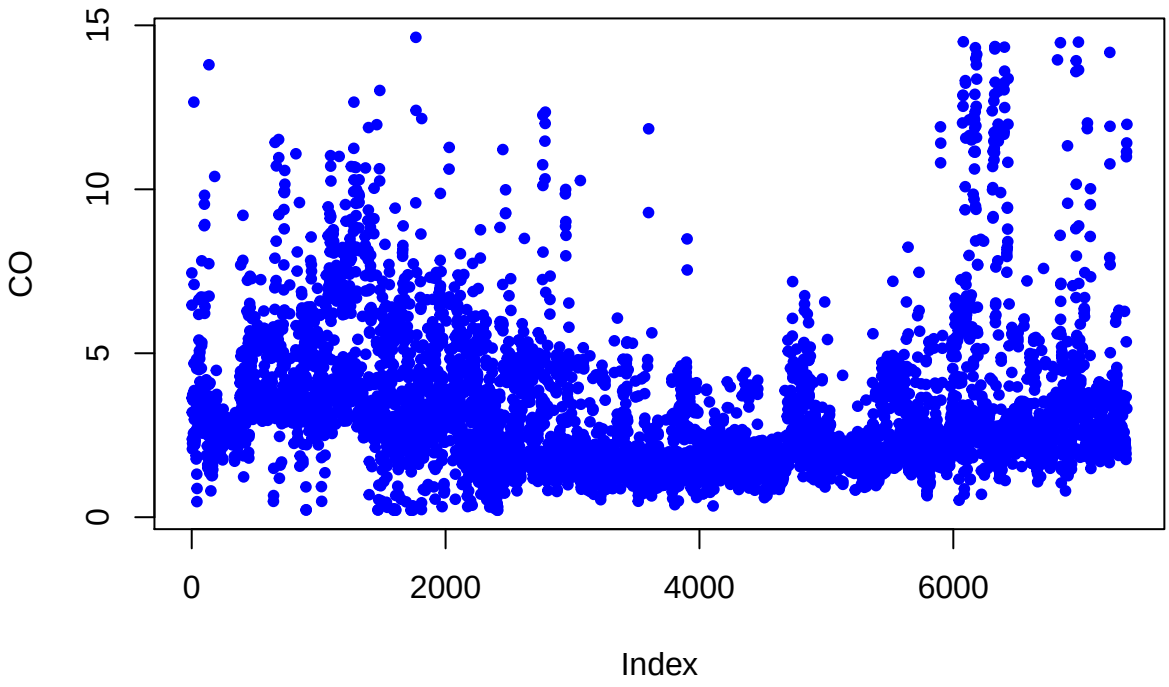
The typical production range the client provided did not fully capture the typical production range that we observed in our data sample (see Figure 2 above). This could be a result of the data values from the 2015 data set having lower values compared to other data sets. Therefore, we decided to shift the typical production range to 127 to 134.5 given that it is a better representation of the typical production range of the turbine energy yield.

Carbon Monoxide Extreme Values



We quickly encountered some issues with the data collected from the power plant. As you can see from Figure 3 above, there are several CO values which deviate significantly from the rest of the values. These outliers look to be somewhat random in Figure 3, but after some investigating we found that the observations with extreme CO values typically fell into one of two categories (see Appendix 8). After appropriately dealing with the observations that fell into each category, we found the CO values to be much more reasonable as is evident in Figure 4.

Figure 4: CO Values



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 multicollinearity, 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. We calculate the RMSE 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.

K-Fold Cross-Validation

When evaluating our models we used 5-Fold Cross-Validation. This minimizes the effect that a single train-test split has on our model. Instead of being highly dependent on a single train-test split of our data we average the results of five different train-test splits when evaluating our model. This results in a better generalized model that does not overfit our data.

	Overall Production Range	Typical Production Range (127-134.5)	High Production Range (160+)
Linear Regression	1.0975	0.6379	0.3986
Lasso	1.1056	0.6564	0.4339
Decision Tree	0.8708	0.6169	0.4188

Results

Decision Tree Model Selection

In the table below, the decision tree outperforms linear regression and lasso in the overall production range and typical production range. Even though the linear regression model performs better in the high production range, the decision tree model comes in as a close second. Therefore, we decided to use the Decision Tree Model to examine the biggest impact on emission levels in order to decrease emissions overall.

Overall Decision Tree Model

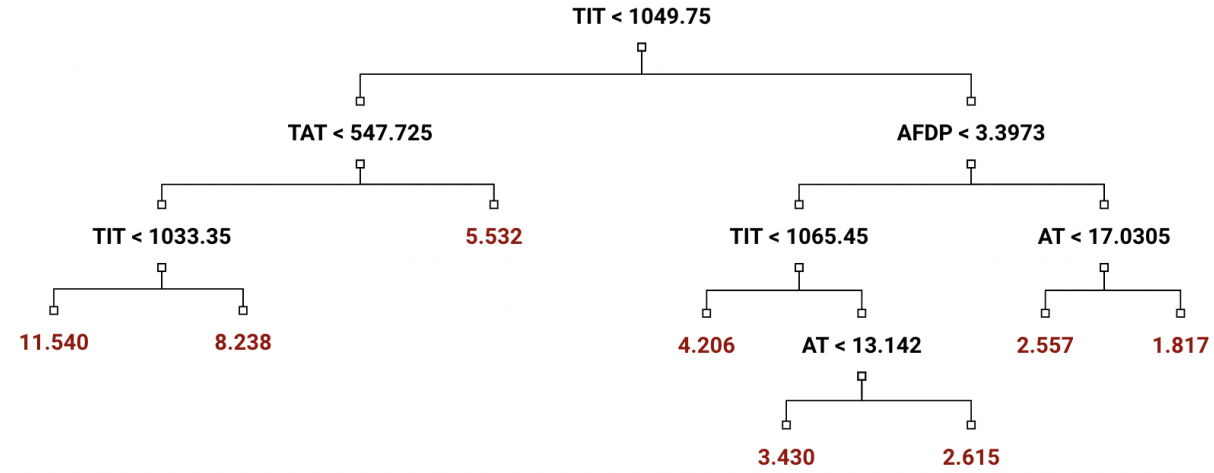


Figure 1: Overall Production Range Decision Tree

The decision tree above represents the 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 1049.75 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. 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

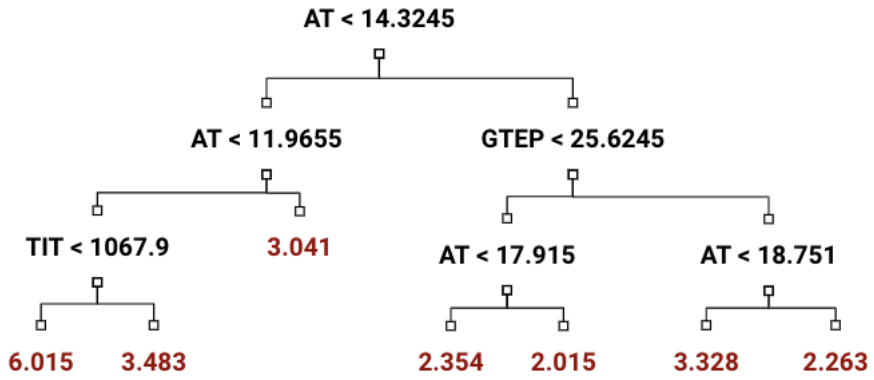


Figure 2: Typical Production Range Decision Tree

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 greater than 11.9655. We can infer AT is likely the most important variable in this energy production range with the higher AT values resulting in lower CO, agreeing with our overall tree model. We can conclude **lower GTEP and higher AT** values appear to result in lowering CO output.

High Decision Tree Model

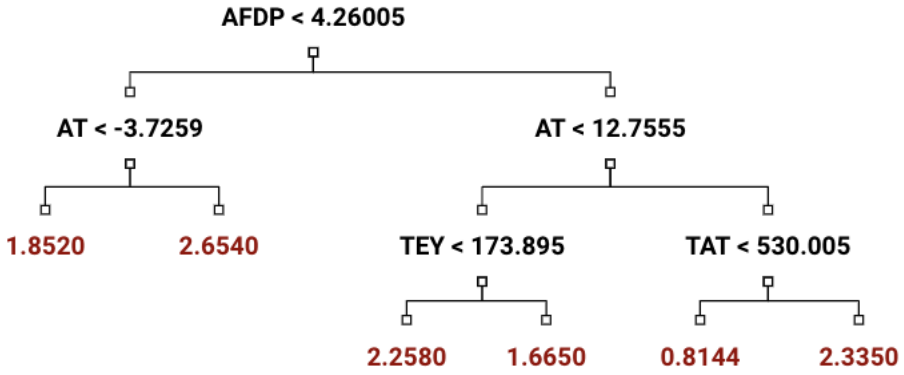


Figure 3: High Production Range Decision Tree

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 lower TAT** values appear to have lower CO outputs.

Conclusion

Most Sensitive Process Variables

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

- AFDP
- AP
- TAT
- TIT

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

- AT
- GETP
- TIT

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

- AFDP
- AT
- TAT
- TEY

Process Variables Impact on CO

1. Overall Production Range: Higher TAT, AT, and AFDP values will lower CO outputs.
2. Typical Production Range (127-134.5): Higher AT and lower GTEP will lower CO outputs.
3. High Production Range (160+): Higher TEY and lower TAT will lower CO outputs.

Recommendations

1. Explore possible ways to increase AFDP.
2. Ensure the typical production range that was supplied to us is the correct range.
3. Client should make sure all data is entered correctly and note any events which require the turbines to shut down.

Appendix

1. Extended description of variables

The table below shows the full description of variables not included in our description.

Variable (Abbr.)	Unit	Min	Max	Mean
Ambient Temperature (AT)	C	6.23	37.10	17.71
Ambient Pressure (AP)	mbar	985.85	1036.56	1013.07
Ambient Humidity (AH)	%	24.08	100.20	77.87
Air Filter Difference Pressure (AFDP)	mbar	2.09	7.61	3.93
Gas Turbine Exhaust Pressure (GTEP)	mbar	17.70	40.72	25.56
Turbine Inlet Temperature (TIT)	C	1000.85	1100.89	1081.43
Turbine After Temperature (TAT)	C	511.04	550.61	546.16
Compressor Discharge Pressure (CDP)	mbar	9.85	15.16	12.06
Turbine Energy Yield (TEY)	MWH	100.02	179.50	133.51
Carbon Monoxide (CO)	mg/ m^3	0.00	44.10	2.37
Nitrogen Oxides (NOx)	mg/ m^3	25.90	119.91	65.29

Figure 4: Variable Description

2. Multiple Linear Regression

We created 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 found 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.

3. 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.

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

[**Source**]

4. 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**]

5. Model Training and Testing Procedure

All of our models were built following the same procedure. We first split the provided data set into training and testing data sets, with 80% of the data randomly assigned to the training data and the remaining 20% assigned to the testing data. We built every comparable model on the same training data as well as tested every comparable model on the same testing data. We did this so we would not risk a model outperforming another from chance due to being applied to a different subset of the original data. All models were also trained using Cross Validation. Cross Validation lowers a models dependency on the data it is trained on by splitting the data into k different subsets (in our case, 5) and training the model k times, each time using k-1 subsets to train the model and the final subset to test the models performance. The k different models built are then averaged together to a single model which does not overfit our training data. The figure below visually illustrates the process that we followed.

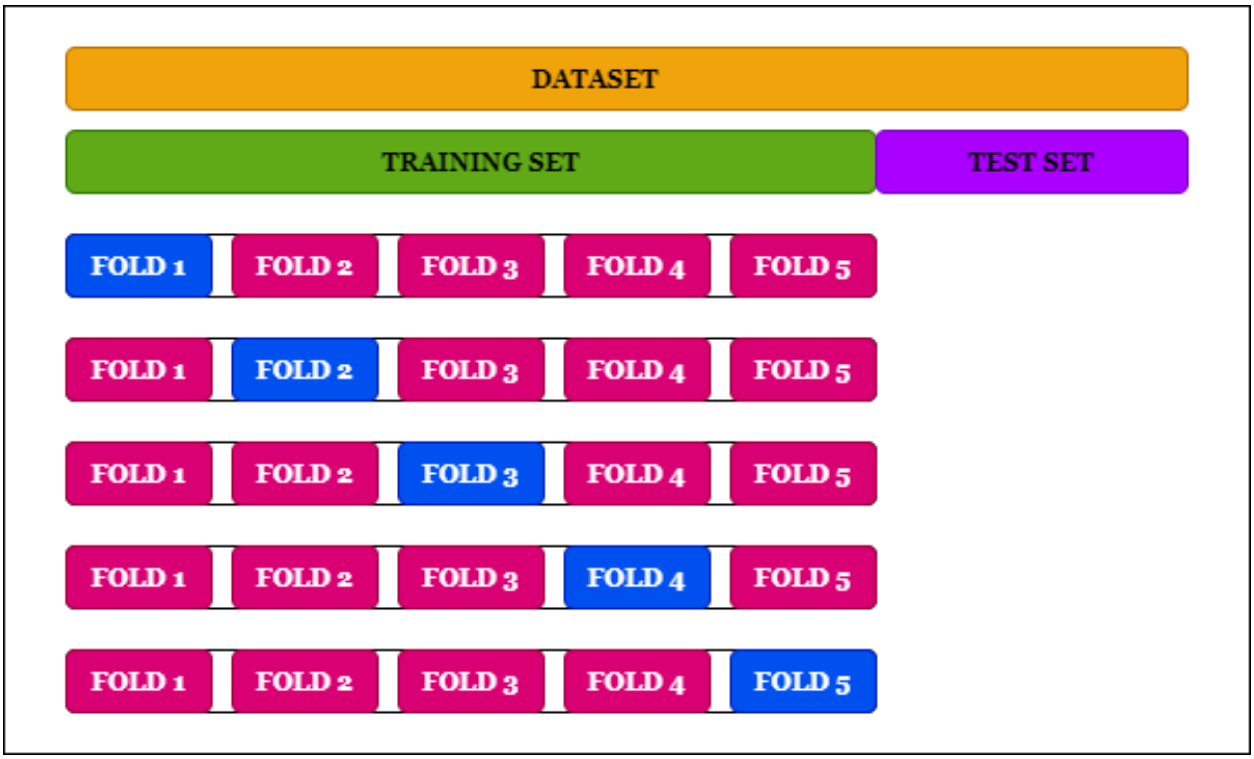


Figure 5: 5-Fold Cross Validation

[**Source**]

6. Variance Inflation Factor (VIF)

Variance Inflation Factor detects multicollinearity in regression analysis. Multicollinearity is when two or more predictors are linearly associated. Linear associations between predictors can cause issues in linear regression results because we assume that independent variables are independent of one another. Linearly related predictors move very similarly, this means we can not really infer that an change in one predictor will impact the outcome by a certain amount because the correlated predictor will also move. Fortunately, we can figure out which predictors are most responsible for multicollinearity issues by checking their VIF values (high VIF values indicate issues with that predictor). We set the VIF cutoff at a value of 10, removing the highest VIF predictor variables one by one until all had VIF values less than 10. It is important to emphasize that only linearly based models are effected by multicollinearity, so we only removed predictor variables that caused issues in 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
```

AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO
11.5830	1026.0	44.346	3.2080	24.339	1075.6	549.93	130.28	11.681	2.5116
11.9170	1025.5	43.388	3.2533	24.415	1076.2	550.06	130.86	11.711	2.2051
11.9360	1025.2	44.874	3.2461	24.335	1076.6	550.09	131.10	11.717	3.7674
11.7650	1024.8	47.251	3.2628	24.592	1077.8	550.11	132.08	11.798	3.7105
11.5740	**1024.4**	**48.852**	**3.0075**	**22.900**	**1040.6**	**533.85**	**117.17**	**11.132**	**29.0840**
11.1090	1024.4	45.556	3.1811	23.939	1078.5	550.06	132.91	11.748	3.4775
10.3270	1024.6	49.863	3.1489	23.520	1075.9	550.06	130.97	11.684	3.5792
9.5762	1024.7	54.655	3.1173	23.738	1076.7	550.04	131.95	11.765	3.5510
8.0657	1024.3	57.378	3.0706	23.518	1075.0	550.08	131.43	11.663	3.8083

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
```

We preprocessed our data by using Variance Inflation Factor (VIF) to tune our model and remove any redundant predictor variables. This selection prefers smaller models which aligns with our goal of limiting the size of our final model. 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 for the overall production range: TIT, CDP, and TEY. For our typical production range we removed TIT and AT while for our high production range we removed TEY and CDP.

7. Correlations

Correlation is a statistical measure that measures how strongly 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**]

8. Carbon Monoxide Extreme Values

The first relationship that we found between observations with extreme CO values was a shock to their TIT, TAT, and TEY values. In the figure above, the row with an extreme CO value is highlighted in yellow. Comparing that observation with the neighbor observations above and below, it can be seen that the observation had a drop in TIT, TAT, and TEY values while all other values (except for CO) stayed relatively constant. We believe that this was due to some sort of malfunction or temporary shutdown of the turbines, which would explain why the turbine related temperatures and the energy yield dropped significantly for a short time period. The high CO could be explained by the turbine starting back up after the shut down, which would likely require significantly more energy than a running machine and produce high amounts of CO. These observations are not relevant to analyzing which process variables are directly related to CO output as their CO values are likely heavily influenced by some outside factor not accounted for in our data. Therefore, we chose to remove the observations which fit this explanation. 16 observations were removed.

AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO
29.053	1010.4	56.689	4.3808	30.297	1100.1	544.22	148.53	13.296	1.16390
30.021	1010.5	51.452	4.3562	30.013	1100.0	545.15	148.11	13.198	0.89122
30.571	1010.6	43.295	4.4434	30.438	1099.9	543.82	149.06	13.253	0.86881
30.479	**1010.6**	**41.500**	**4.3378**	**29.650**	**1099.8**	**546.17**	**147.87**	**13.066**	**13.63900**
30.217	**1010.8**	**39.650**	**4.4695**	**30.633**	**1099.9**	**542.97**	**149.90**	**13.341**	**31.86900**
29.865	1011.0	36.924	4.4806	30.725	1099.8	542.61	150.31	13.281	3.30280
29.275	1011.2	41.059	4.4016	30.363	1099.9	543.84	149.52	13.289	2.67400
28.358	1011.6	47.163	4.3168	30.038	1100.0	544.89	148.87	13.157	1.44270

The observations that had high CO values but did not show any evidence of a shock to the process variables all indicated a different relationship. The figure above shows an example of two observations, again highlighted in yellow, which we found a possible explanation for their extreme CO values. It is obvious that the CO values belonging to the two observations in question are vastly different from their neighbor observations despite all the process variables being very similar. If you divide their CO values by 10 (move the decimal once to the left), the CO values of the two observations look more in line with their neighboring values. We believe that the observations like these in the data could have been incorrectly entered. One explanation for this is if the data is manually entered, the employee entering the data could have accidentally made this error. We decided to change the observations which fit this explanation by dividing them by 10. There were 9 observations that were impacted by this.

9. Individual Contributions

Aayushi

-

Kyle

-

Rosa

-

Ruben

-