Final Project
Reducing CO in Turbine Output

By Group 7:

Mitchell Cappel, Christopher Cebra, Kashyap Ava, Kris Png

For Client: Man Fung Leung

Introduction

Overview

Power plants form an extremely important component of the world's electrical grid, producing vital electricity from materials such as natural gas, coal, and, increasingly, renewables. While these plants have a critical role in electricity production, they also bring several harmful emissions and pollutants, including carbon monoxide (CO). CO, a colorless, odorless gas, is a byproduct of some of the reactions in these turbines.

Data

We received our data from the UCI Machine Learning repository, which contains observations of a Turkish power plant's CO emissions and a series of parameters measured at different locations in the turbine. Our topic of interest is whether any of the parameters in the turbine can predict CO emissions, either higher or lower. Once we have identified significant variables, we can pass on our recommendations to the engineering team, who can modify the turbine to raise or lower the given values. Our analysis contains one dataset, TurbineGroup7.csv. It contains 7,158 observations, each with the following variables represented in Figure 1.

Variable	Abbr.	Unit
Ambient temperature	AT	°C
Ambient pressure	AP	mbar
Ambient humidity	AH	(%)
Air filter difference pressure	AFDP	mbar
Gas turbine exhaust pressure	GTEP	mbar
Turbine inlet temperature	TIT	°C
Turbine after temperature	TAT	°C
Compressor discharge pressure	CDP	mbar
Turbine energy yield	TEY	MWH
Carbon monoxide	CO	mg/m^3
Nitrogen oxides	NO_x	mg/m^3

Figure 1: List of variables in the turbine dataset

The variables in our dataset are divided into three groups. First, carbon monoxide (CO) and nitrogen oxides (NOx) are response variables indicating the pollution produced by the turbine. Since we aim to find methods to reduce CO emissions, our response variable will be CO, while we will discard NOx from our dataset. The second group is the ambient variables. These represent aspects of the environment that might affect the turbine's efficiency but cannot be changed by the engineering team. Three of them are the ambient temperature (AT), pressure

(AP), and humidity (AH), while the fourth, turbine energy yield (TEY), represents electricity demand and thus cannot be changed. We aim to make our turbine more efficient given the same electricity demand. In this case, our five controllable variables are air filter difference pressure (AFDP), gas turbine exhaust pressure (GTEP), turbine inlet temperature (TIT), turbine after temperature (TAT), and compressor discharge pressure (CDP).

Variable	Mean	SD	Min	25%	50%	75%	Max
AT	18.22	6.99	1.14	12.93	18.28	23.68	34.19
AP	1012.40	5.97	994.84	1008.10	1012.00	1016.30	1031.40
AH	82.14	13.00	25.99	73.44	84.66	92.83	100.10
AFDP	3.91	0.72	2.15	3.55	3.86	4.31	6.74
GTEP	25.74	3.68	17.94	23.70	25.19	27.99	36.22
TIT	1078.84	16.57	1002.90	1069.63	1079.20	1092.50	1100.90
TAT	547.64	4.61	511.04	546.97	549.91	550.05	550.60
TEY	132.91	13.60	100.03	123.38	133.42	141.64	168.63
CDP	12.02	0.95	9.85	11.45	11.90	12.64	14.62
CO	2.08	2.01	0.002	1.25	1.66	2.16	34.82
NOX	60.07	9.97	27.18	53.88	59.28	64.03	118.47

Figure 2: Data summaries of all variables in the turbine data. **Red** variables are ambient or environmental, **Blue** variables are controllable variables, and **Purple** variables are pollutants (response variables).

A summary of all the variables and their mean, median, standard deviation, and quartiles is given in Figure 2. As we can see, there are no missing values, meaning we do not have to address any missing data problems. Also notable from Figure 2 is that our response variable, CO, is likely heavily right-skewed–the median is 1.66, the mean is 2.08, and the gap between the 75th percentile value of 2.16 and the maximum of 34.82 is very substantial. This low amount of high values of CO could affect our models.

Finally, the clients were interested in several submodels in addition to the main model. Both submodels are subsets of TEY results, a “medium energy load” model of all data from

130-136 TEY and a “high energy load” model of data greater than 160 TEY. We were asked to produce the best model for the entire dataset and both submodels and issue recommendations in each case.

Method

To analyze the data and address the research questions, we first carried out data cleaning, scaling, and outlier removal to improve model performances. From there, we considered various models, including linear models with and without a transformation. Next, we considered penalized regressions like the LASSO and performed a bootstrap procedure. Finally, we considered nonparametric models, such as decision trees and random forests, with the best overall model performance.

Data Cleaning

First, we performed some data cleaning. In Figure 3, we can look at the correlations between the different variables in our turbine dataset. Any variables that are correlated over 90% should be a potential concern, and in this case, we have several, including the correlation between TEY and GTEP, GTEP and CDP, GTEP and TIT, TIT and TEY, TIT and CDP, and TEY and CDP. These correlations could be problematic if we do a variable selection procedure, as the other highly correlated data could mask the signal from one variable.

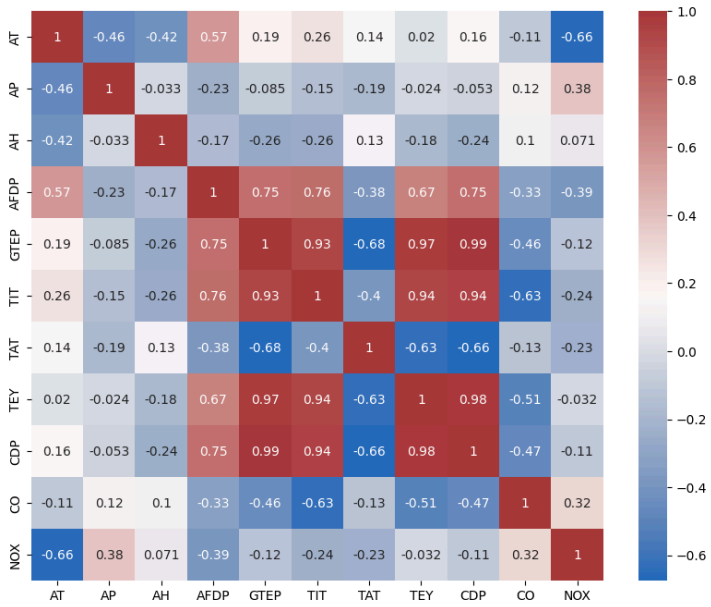
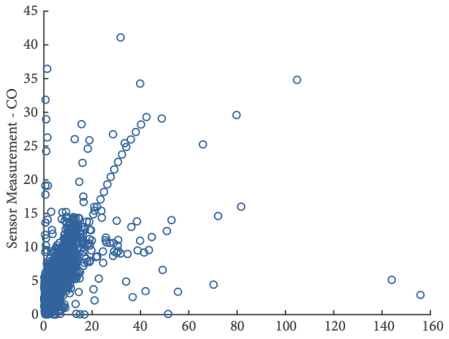


Figure 3: Correlogram of single-variable correlations between all variables in the turbine data.



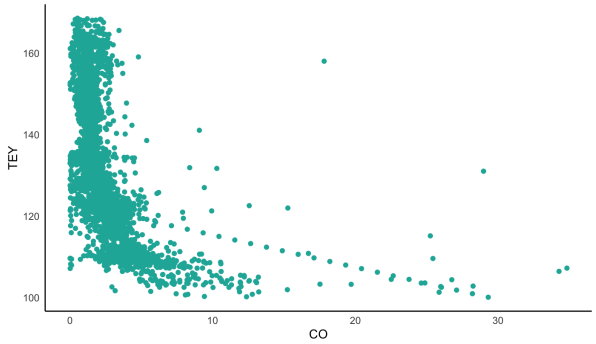


Figure 4: "The line" visible as a source of incorrect/corrupted data. Left: Predicted vs actual CO measurement in the *source paper* for this study (not our constructed data). Right: Scatterplot of TEY (Y-axis) vs CO (X-axis) for our project dataset. "The line" is evident on the bottom left.

The second major data processing step concerned a set of 25 evenly-spaced data points in CO that we referred to as "the line". These points can be seen in the right graph of Figure 4. They are sequential observations in our source CSV file, and the gap between each point in CO (TEY and other variables) is constant. We thought these points likely originated from a problem with the measuring equipment and discussed this in our mid-project check-in meeting with the client, who shared the same view. Therefore, we removed "the line" from our future models.

Finally, we wished to remove other outliers and influential points in CO that might have an outsize effect on our model. Figure A1 shows a density plot of CO by value and marked and dotted 90th, 95th, and 99th quantiles. As we already saw in the variable summary in Figure 2, the overall TEY dataset after the data processing mentioned above had more than 10% of observations with CO values greater than 5. In contrast, the medium and high yield data had no CO values greater than 5. Naturally, any good predictive model would recommend TEY above 130 without emphasizing much of the effects of the other controllable variables. We removed 1% of data with CO values greater than 9 to prevent that and not lose much of the data, hoping to capture more actionable recommendations on the controllable variables.

To answer our research question and find which controllable variables lead to a decreased CO output, we thus aimed to test different modeling techniques to gain insights into the data. These techniques include linear regression, LASSO regression, decision trees, and random forest ensemble methods. The linear regression model provides a simple and interpretable baseline model from which straightforward insights can be gleaned from the data. We use a LASSO regression model for feature selection to understand the data better and combat multicollinearity issues. Moving on to more complex models, we used a decision tree, which can create clear and interpretable recommendations. We also use a random forest ensemble model, an ensemble of multiple decision trees, for enhanced prediction and robustness, especially for non-linear data, at the cost of interpretability. Employing multiple models can bolster each model's gaps with another model's capabilities.

Model Evaluation

In our project, we employed an 80:20 train-test split, where 80% of the data was used for training the model, and 20% was reserved for testing its performance. This approach is essential as it allows us to evaluate how well the model generalizes to unseen data, indicating its real-world performance. By having a separate test set, we can ensure that the model's performance metrics are not overly optimistic due to being evaluated on the same data it was trained on, which helps prevent overfitting. To assess the model's performance, we used several metrics: R-squared (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). R-squared is a statistical measure that indicates the proportion of the variance in the CO that is predictable from the predictor variables. It explains how well the model's predictions approximate the actual data points, with values closer to 1 indicating a better fit. MAE measures the average magnitude of the errors in a set of predictions without considering their direction, offering a straightforward interpretation of prediction accuracy. RMSE, on the other hand, measures the square root of the average squared differences between predicted and actual values. RMSE gives a higher weight to large errors and is thus helpful in understanding the model's performance in scenarios where more significant errors are undesirable. Together, these metrics provide a comprehensive view of the model's accuracy and ability to generalize to new data.

Ultimately, we saw our random forest model perform the best for all TEY ranges in various ways. Our root-mean-squared error and mean-absolute error were lowest for our random forest models, and our r-squared values were highest. These values for all TEY ranges can be found in the appendix tables.

Linear Regression

We began by fitting standard linear regression models for all TEY ranges, but we first wanted to check that they met the assumptions of linearity and normality. We found that the linear regression model may be inadequate at modeling the data for the overall and medium yield cases due to the non-normality of errors and association between residuals, as shown through the full model and medium TEY's Q-Q Residuals and Residuals vs. Leverage plots in Appendix Figures A2 through A3. In these cases, the points deviate from the reference line in the Q-Q residual plots, especially at the tails, indicating the non-normality of errors. In their Residuals vs Leverage plots, we see a non-horizontal line, which may indicate the presence of an association between residuals. However, this is not the case for our linear model for high TEY values, which we have deemed to fulfill the assumptions necessary for employing a linear regression model. The plot of high TEY shows that the linear model assumptions are met in Appendix Figure A4.

Based on these results, linear regression models will not be adequate to derive our overall and medium yield data results. Therefore, we must pivot to non-linear models to give a clear result to our client. As for the high-yield data, while the normality assumption is met, we still observe trends in the residuals versus the leverages, so we use non-linear models to ensure we end with the best possible model.

LASSO Regression

Before moving on to non-linear models, we wanted to improve our standard linear models using LASSO penalized residuals. Through LASSO, we can fix the multicollinearity issues noted earlier in Figure 3 by shrinking the coefficients of some correlated variables to zero. Through this process, we saw a slight increase in our R-squared values compared to our standard linear model for all TEY ranges.

To analyze the coefficients, we performed a bootstrap procedure on the LASSO regression output, which creates an analog of a confidence interval for LASSO results, which are fitted as point estimates only.

However, our model assumptions still conclude that linear models do not fit our data well. While we have improved our R-squared values slightly, we expect a more significant improvement once we move away from linear-based modeling.

Decision Tree

The decision tree model works by recursively splitting the data into nodes based on conditions that maximize the homogeneity (or minimize the impurity) of the data within each node. At each step, the model evaluates possible splits on variables (such as TEY, TIT, CDP, or AFDP) and selects the condition that best separates the observations into groups that are more similar in terms of the target variable. The model minimizes the prediction error by grouping observations with similar characteristics into the same node. For example, the decision tree trained for the high-yield data shown in Figure A7 splits on AH >= 66; it identifies a threshold that separates observations into two groups where predictions are more accurate within each group. Subsequent split (GTEP >= 33) refined this grouping to reduce variability and improve prediction accuracy. Each node's prediction represents the average or majority class of the observations within that node. The percentages at each node indicate how much of the total data is represented, ensuring the tree effectively balances model complexity and predictive power. This hierarchical splitting approach reduces error by focusing on creating smaller, more predictable subsets of the data.

A further advantage of the decision trees is that they allow for predictions more efficiently based on ambient variables. The optimal recommendation for some of the controllable variables differs based on the impact of one of the ambient variables.

Although the decision tree performed well compared to the previously discussed regression techniques, there is still scope for improvement as we can reduce the error and increase R-squared even more using multiple trees.

Random Forest

To improve the predictive performance and address the limitations of a single decision tree, we employed a Random Forest model with 400 trees. Random Forest combines the

predictions of multiple decision trees, each trained on different subsets of the training data, to create a more robust and accurate model. By averaging the outputs of these trees, the Random Forest reduces the variance typically associated with individual trees, leading to more stable and generalizable predictions. Including 400 trees ensures that the model captures a diverse set of patterns and minimizes the impact of overfitting to any specific dataset split. Using Random Forest significantly improved the model's performance metrics. The Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) were reduced, indicating that the predictions were closer to the actual values on average, and the error distribution was narrower. Additionally, the test R-squared value increased, demonstrating a more substantial explanatory power of the model and its ability to account for a larger proportion of the variance in the target variable. These improvements highlight the effectiveness of Random Forest in leveraging multiple decision trees to create a more accurate and reliable model. The variable importance was used to identify the most important controllable variables in the trained random forest model.

Results

Linear Regression

Overall Data: CO ~ AT + AP + AH + AFDP + GTEP + TIT + TEY + TAT + CDP			
Coefficient	**Estimate**	**Standard Error**	**P-value**
Intercept	122.64	3.76	< 2e-16
AT	-0.01	0.004	0.003
AP	-0.003	0.002	0.001
AH	-0.002	0.001	0.07
AFDP	0.31	0.03	< 2 e-16
GTEP	-0.04	0.02	0.10
TIT	-0.004	0.005	0.42
TEY	-0.09	0.01	< 2e-16
TAT	-0.18	0.01	< 2e-16
CDP	-0.13	0.16	0.41

Table 1: Linear Regression output summary for the overall model. Variables highlighted in **Green** indicate significant predictors.

Table 1 shows the output of the linear regression model fitted for the overall data. We observed that out of the controllable variables, AFDP, TET, and TAT are significant concerning the p-values. Still, the coefficient of AFDP was the highest, indicating that it is the most impactful variable. The results for the medium and high-yield datasets are shown in Tables A1 and A2. We observed that AFDP, GTEP, TEY, and TAT are significant for the medium yield data. Out of these, GTEP was the most impactful. However, the controllable variables were insignificant for the high-yield data.

LASSO Regression

Figure A5 shows the output from this bootstrap procedure on our three TEY models. Some coefficients are zeroed, and if the confidence interval does not include zero, we can conclude that these coefficients may be part of our model. For our overall data, we concluded that TAT, AFDP, and GTEP have the most non-zero coefficients and should be focused on in our model. TAT and GTEP are the most significant factors in our medium data. Finally, for our high-yield TEY data, we found that GTEP and TAT are the most significant predictors of CO when reducing multicollinearity issues through LASSO. Several ambient variables have nonzero coefficients as well. Also, as the client noted in the presentation, these “confidence intervals” do not guarantee that a coefficient is or is not significant, and individual LASSO passes could select different variables.

Decision Tree

Figure 5 displays our pruned decision tree for our overall dataset. The trees from our medium and high-yield sets can be found in Figures A6 and A7. Using this information, the engineers can analyze changes needed for controllable variables based on certain ambient conditions. This data can provide specialized insights but may not help when one wants a clear overall conclusion.

For this specific case, the overall model, we can see that TEY is the most significant variable in our decision tree. Recommendations based on variables other than TEY are only performed when TEY is between 119 and 126. Therefore, 84% of the data has a prediction based on TEY only. From there, however, our tree makes predictions based on four controllable variables–TAT, GTEP, TIT, and AFDP. The predicted fitted value from the tree is equal to the top of the three numbers at the base of the tree, while the number of observations in this branch and that percentage of the overall observations are also listed.

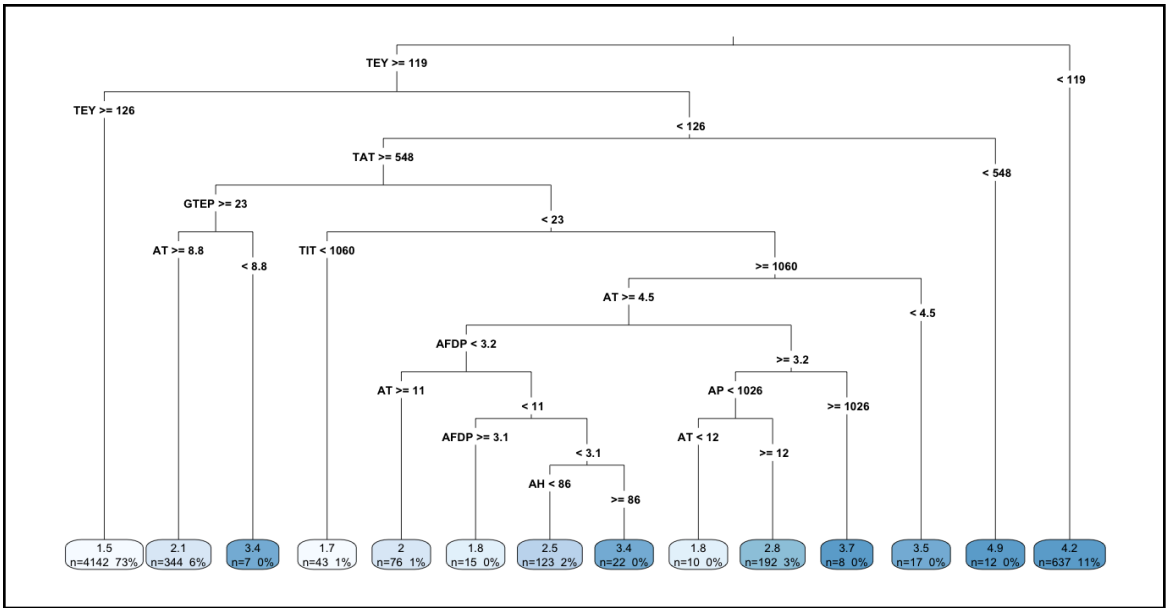


Figure 5: Pruned decision tree for overall data

Random Forest

From Figure 6, we can see that TEY was the most important feature, and as for controllable variables, we see that CDP and TIT are the two most important features following TEY.

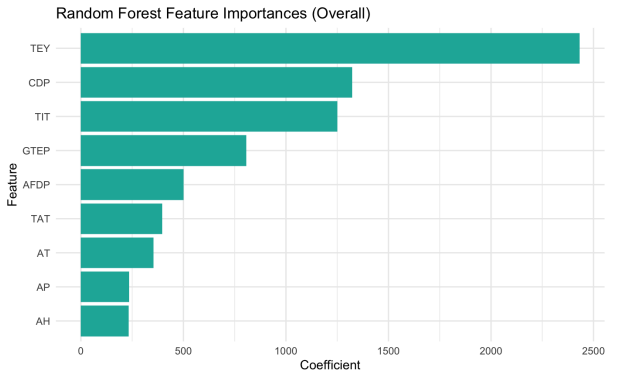


Figure 6: Barplot of the Values of Random Forest Feature Importances for Overall Yield

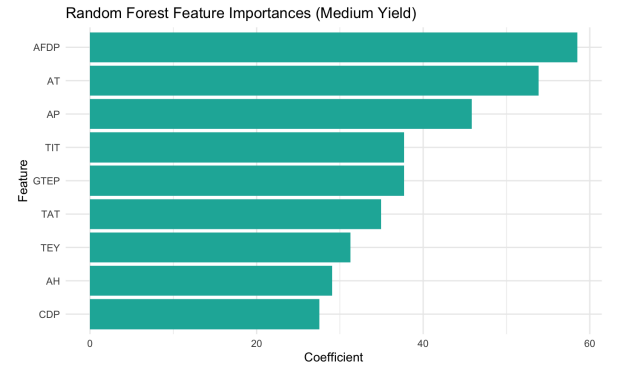


Figure 7: Barplot of the Values of Random Forest Feature Importances for Medium Yield

From Figure 7, we see that AFDP is the most important feature overall, and out of the controllable variables, we also see TIT and GTEP being deemed as important features.

Based on our random forest model's variable importances in Figure 8, we see that AFDP and GTEP are considered the most important features besides AH, which cannot be controlled.

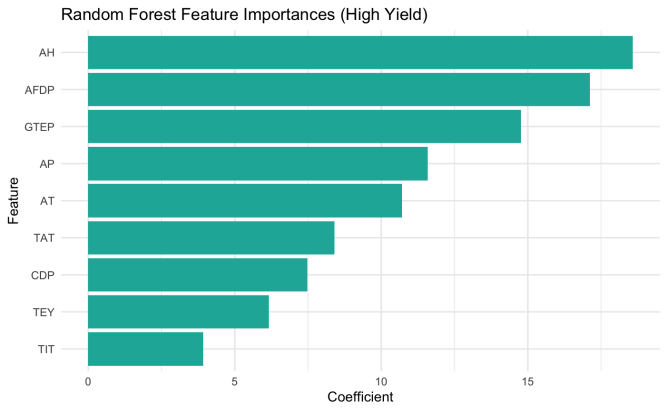


Figure 8: Barplot of the Values of Random Forest Feature Importances for High Yield

Based on the output of our various models, we can thus extract actionable suggestions for the engineers on decreasing CO output. The test results for the four models used on the overall, medium, and high-yield datasets are in Tables 2, 3, and 4 below. The random forest performed the best in all three cases with the lowest MAE and RMSE and highest R-squared.

Overall	Linear Regression	LASSO	Decision Tree	**Random Forest**
RMSE	0.721	0.719	0.603	**0.504**
R-Squared	0.618	0.619	0.732	**0.813**
MAE	0.493	0.491	0.374	**0.313**

Table 2: Model Performances for Overall Data

Medium	Linear Regression	LASSO	Decision Tree	**Random Forest**
RMSE	0.393	0.391	0.398	**0.357**
R-Squared	0.099	0.113	0.074	**0.252**
MAE	0.263	0.263	0.275	**0.241**

Table 3: Model Performances for Medium-Yield Data

High	Linear Regression	LASSO	Decision Tree	**Random Forest**
RMSE	0.648	0.639	0.711	**0.549**
R-Squared	0.091	0.113	-0.094	**0.346**
MAE	0.476	0.478	0.567	**0.433**

Table 4: Model Performances for High-Yield Data

Conclusion and Discussion

Overall

For our overall model, where our random forest model performed the best, plotting the trends of CDP vs. CO and TIT vs. CO in Figure 9, we can see the negative relationship between CDP and CO and TIT and CO. For CDP and CO, there is a steep negative trend initially before gradually plateauing at around 12 mbar. There is also a negative trend between TIT and CO, which gradually plateaus at around 1085C. With these insights, we thus suggest engineers keep CDP above 12 mbar and TIT above 1085 C to decrease CO output in a general case.

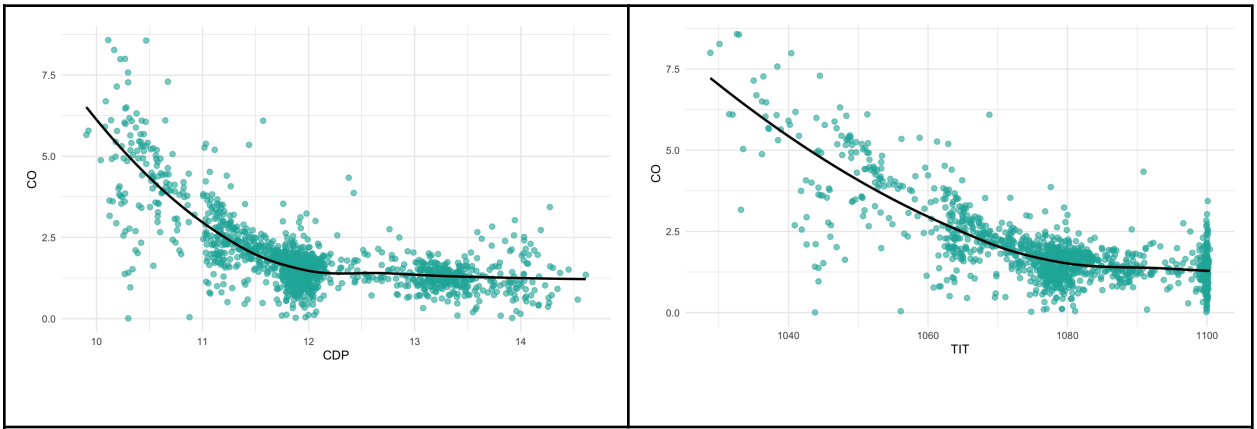


Figure 9: Plot of the Trend of CO vs CDP and TIT for Overall Yield

Medium-Yield

For our medium yield model, we saw that the random forest model performs better than the other models, with a test r-squared of 0.252, greater than the test r-squared values of linear regression (0.099), LASSO (0.113), and decision trees (0.074). Hence, we looked at the results of

our random forest model for insights. The plots of these variables against CO see scattered, clustered data, as shown in Figures 10a and 10b. For the plot of AFDP vs CO, there is an initial downward trend before increasing and decreasing again. It is similar to the plot of TIT vs CO. For GTEP vs CO, there is a reasonably flat trend before increasing toward the end. As such, we suggest the following recommended values of the aforementioned controllable variables to decrease CO output in the case of a TEY of 130-136 MWH: 3.6 mbar and lower for AFDP, 1088 C and higher for TIT, and around 25 mbar for GTEP.

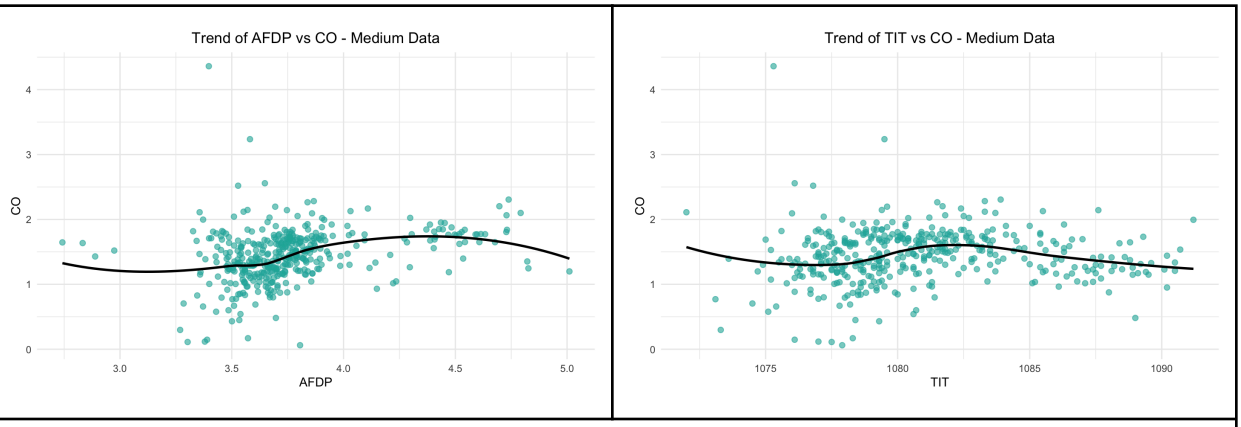


Figure 10a: Plot of the Trend of CO vs AFDP and TIT for Medium Yield

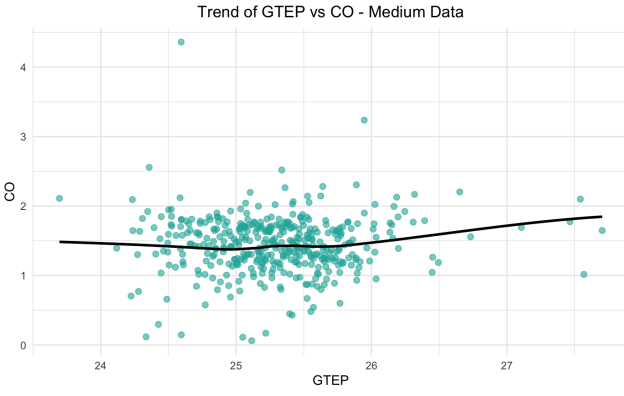


Figure 10b: Plot of the Trend of GTEP vs CO for Overall Yield

High-Yield

Lastly, the random forest model performed the best among our chosen models for our high-yield model, with a test r-squared value of 0.346. However, this subset of data did see a very small number of points, which may have contributed to the model's relatively poor performance. We see the effect of the small amount of data in their corresponding plots in Figure

11. However, we may still attempt to glean some insights based on these plots. For AFDP against CO, there is an upward trend, which turns downward before slowly trending upward again. For GTEP against CO, there is a sharp downward trend, which then turns upward at around 33.5 mbar. Thus, we recommend the values of around 4.4 mbar for AFDP and 33.5 mbar for GTEP for reducing CO output in the case of a TEY of 160 MWH and above.

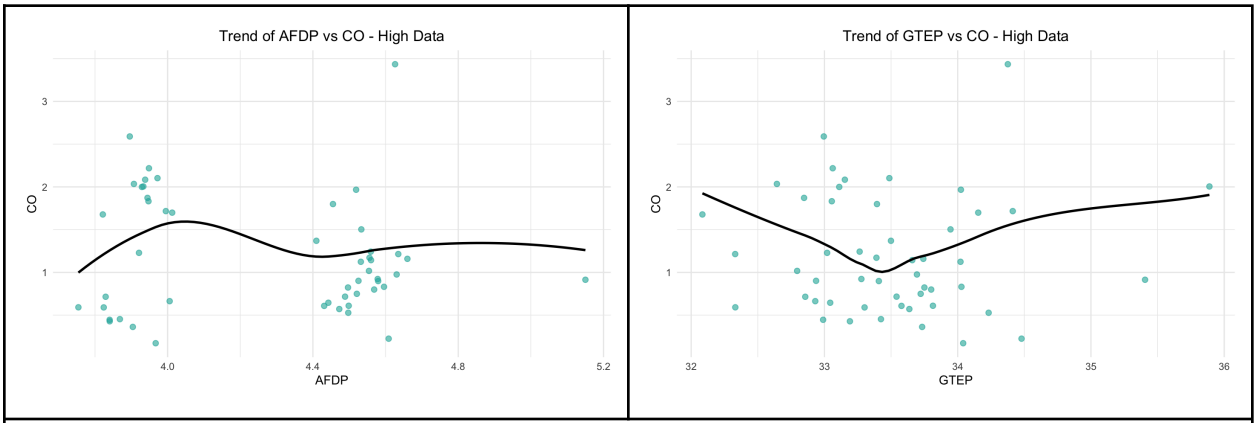


Figure 11: Plot of the Trend of CO vs AFDP and GTEP for Medium Yield

A summary of our final suggestions is presented below:

Overall	Keep: - CDP above 12 mbar - TIT above 1085 C
Medium	Keep: - AFDP to 3.6 mbar and lower - TIT above 1088 C - GTEP around 25 mbar
High	Keep: - AFDP around 4.4 mbar - GTEP around 33.5 mbar

A few general conclusions about our models and parameters are as follows: Firstly, the models with the highest explanatory power of CO emissions are the random forests, followed by the decision trees, with the linear models having the lowest performance. The nonlinearity of the data can explain this. Secondly, the models on the full data had the highest predictive power towards changes in CO emissions compared to either submodel, including our R-squared and RMSE/MAE results. Our conclusions, based on the random forests, are listed above.

We also have suggestions for future actions and analysis based on our research. We mainly want to focus on improving our medium and high-yield models. Increasing the amount of data in each of these subgroups would help lead to stronger models and suggest collecting more data over a longer period. With this expanded data, we could run all of our modeling and tests again and hopefully see an improved result. It is also possible that the turbine is just running reasonably efficiently in the submodels, and most variables do not significantly impact its performance. Most cases where the CO emissions were very high were in the full model rather than each submodel.

Secondly, one difficulty we may have, especially with the full model, is the unfound impact of TEY. From our correlation plot, it is clear that TEY has a strong correlation with CO (again, most of the emissions arise from turbine inefficiencies, which are more common at lower TEYs). The random forest feature importances indicate this–the strongest variable in our full model is TEY, followed by two variables that are highly correlated with TEY. It can also be seen in our full model decision tree, where the most significant predictor of most of our tree splits are inequalities involving TEY. The submodels help in this capacity–since they remove the broader TEY vs CO trend by focusing on smaller subsets of this. We can see this by the absence of TEY from the top of the feature importances for the submodel. It would allow the client to focus only on the impact of the controllable variables.

Appendices

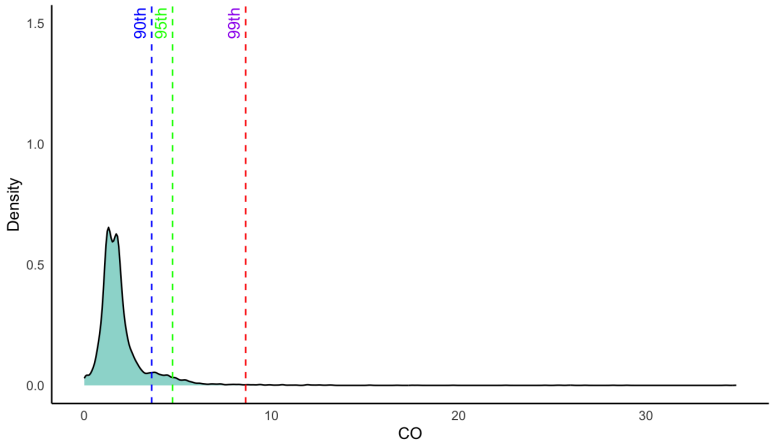
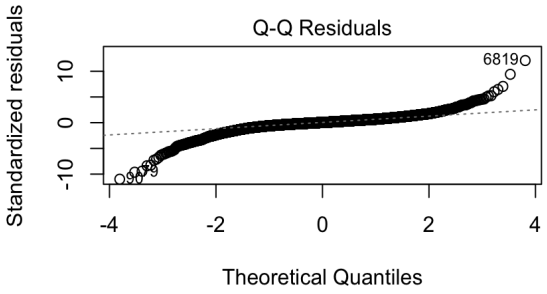


Figure A1: Density plot of CO with 90th, 95th, and 99th quantiles

These show the diagnostics for the residuals for the linear models, Q-Q plot, and residuals vs leverage:



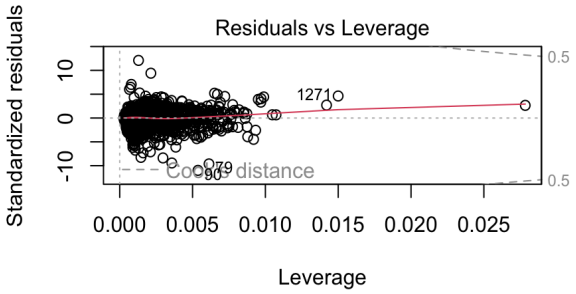
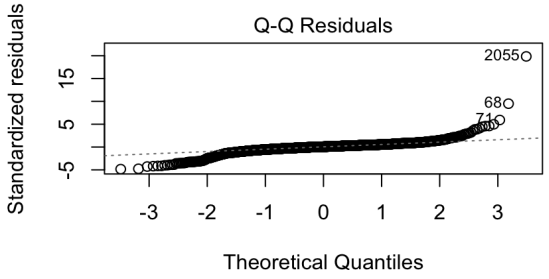


Figure A2: Q-Q Residuals Plot and Residuals vs Leverage Plot for Overall Yield



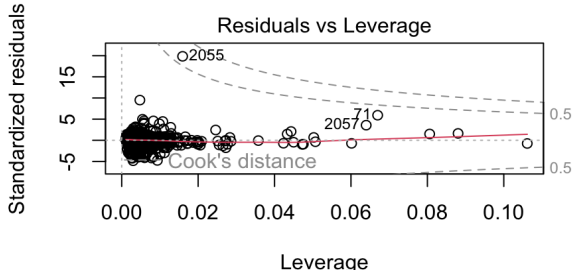
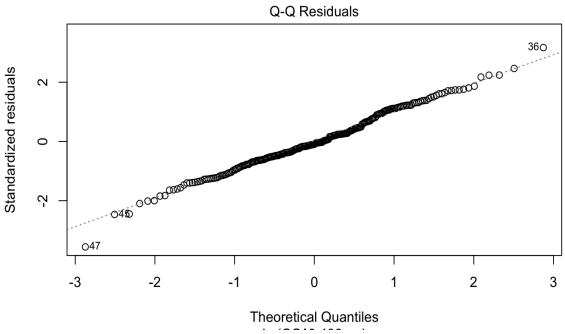


Figure A3: Q-Q Residuals Plot and Residuals vs Leverage Plot for Medium Yield



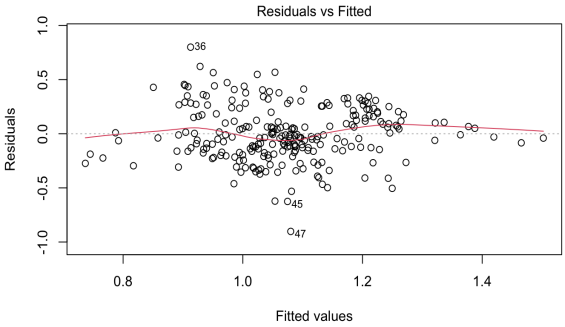


Figure A4: Q-Q Residuals Plot and Residuals vs Leverage Plot for High Yield

Medium Data: CO ~ AT + AP + AH + AFDP + GTEP + TIT + TEY + TAT + CDP			
Coefficient	**Estimate**	**Standard Error**	**P-value**
Intercept	152.22	13.02	< 2e-16
AT	0.04	0.005	2.8 e-13
AP	-0.001	0.002	1.9 e-05
AH	-0.001	0.001	0.5
AFDP	0.26	0.04	3.1 e-08
GTEP	-0.37	0.04	< 2e-16

TIT	-0.004	0.003	0.19
TEY	0.03	0.01	0.352
TAT	-0.24	0.02	< 2e-16
CDP	0.05	0.18	0.79

Table A1: Linear Regression output summary for the medium model. Variables highlighted in Green indicate significant predictors.

High Data: CO ~ AT + AP + AH + AFDP + GTEP + TIT + TEY + TAT + CDP			
Coefficient	**Estimate**	**Standard Error**	**P-value**
Intercept	245.47	245.75	0.32
AT	-0.12	0.04	0.003
AP	-0.02	0.009	0.01
AH	-0.01	0.006	0.003
AFDP	0.20	0.16	0.21
GTEP	0.16	0.09	0.06
TIT	-0.19	0.23	0.41
TEY	-0.13	0.08	0.09
TAT	0.02	0.05	0.73
CDP	-0.23	1.34	0.86

Table A2: Linear Regression output summary for the high model. Variables highlighted in Green indicate significant predictors.

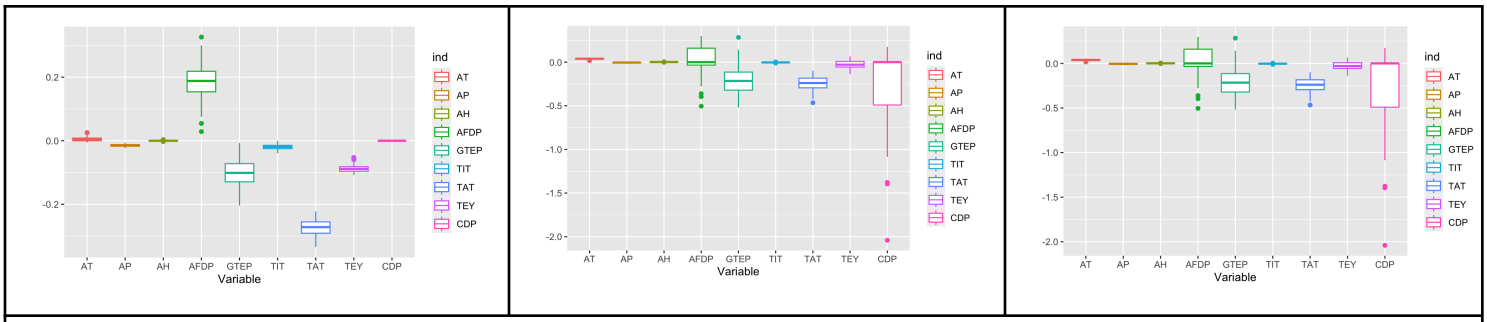


Figure A5: Bootstrap LASSO outputs for the three cases–Overall (left), Medium TEY (middle), High TEY (right)

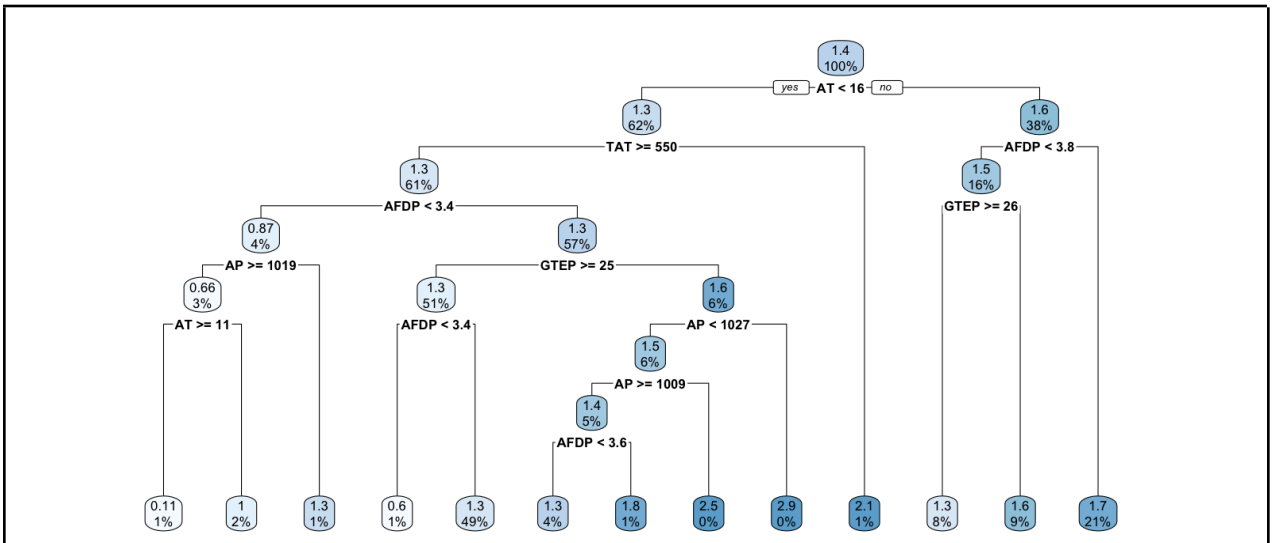


Figure A6: Decision tree for medium-yield data

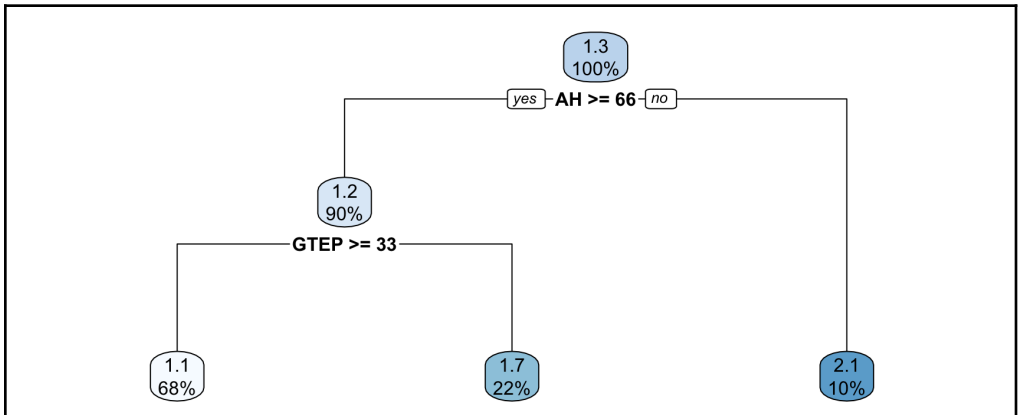


Figure A7: Decision tree for high-yield data

Table A3 briefs about the R/ Python code written in the code files submitted.

R File Name	Contents
OutlierSearching.Rmd	Linear model assumptions and outlier/high leverage discovery. Shows difference in model performance between standard models and outlier-free models.
Stat443BootstrapLasso.Rmd	Performed LASSO and bootstrap procedure on each model. Creates graphics of confidence interval for variable (other than Intercept) coefficients.
stat443_consulting_project_introductory_analysis.py	Introductory data analysis, correlation investigations, 'the line', starting linear regression models, model selection processes, Random Forest analysis and predicted performance plots
STAT443_Project.Rmd	Data cleaning, train-test split, linear regression, LASSO, decision trees, and random forests

Table A3: R Code Submissions with listed contents

Contributions

The group members contributed equally to all models and reports in the project.