Reducing CO in Turbine Output

Group 7
Mitchell Cappel
Christopher Cebra
Kashyap Ava
Kris Png

Client: Man Fung Leung

Background

- Gas turbines are useful energy sources, but emit harmful emissions
- CO is a major pollutant, being produced at an average rate of 2.081 mg/m^3
- The data includes hourly records of turbine data, ambient variables, and CO emissions from a power plant in Turkey
- Original data retrieved from the UCI Machine Learning Repository

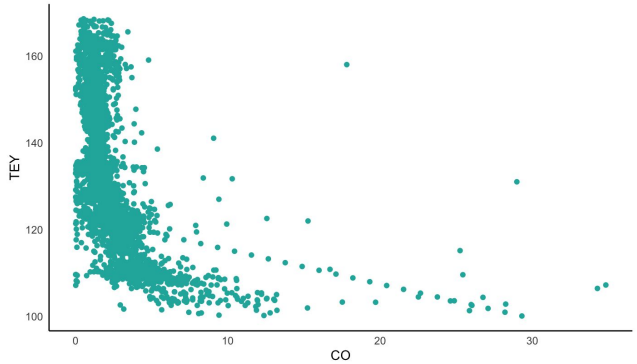
Data

- A dataset with 7,158 hourly turbine measures from a Turkish power plant
- 5 Controllable Variables:
  - AFDP, GTEP, TIT, TAT, CDP
- 4 Environmental:
  - TEY, AT, AP, AH
- Response Variable: CO

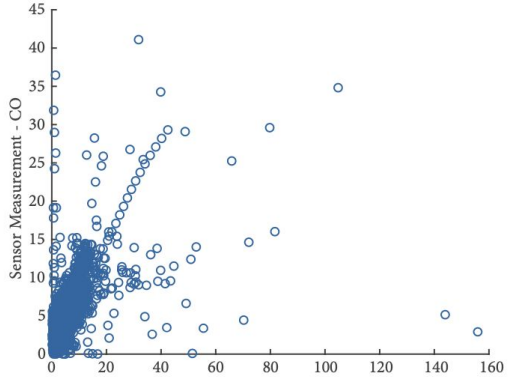
Variable	Abbr.	Unit
Ambient temperature	AT	°C
Ambient pressure	AP	mbar
Ambient humidity	AH	(%)
Air filter difference pressure	AFDP	mbar
Gas turbine exhaust pressure	GTEP	mbar
Turbine inlet temperature	TIT	°C
Turbine after temperature	TAT	°C
Compressor discharge pressure	CDP	mbar
Turbine energy yield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

Data Cleaning

- Initial data contained **25 sequential, evenly-spaced values** for CO we referred to as "**the line**".
- Discussed in mid-project meeting, concluded this was likely due to a problem with the sampling equipment.
- **Removed** "**the line**" from all of our analysis going forward.



CO vs TEY in our dataset



CO vs Predicted CO in the source paper.

Objective and Methodology

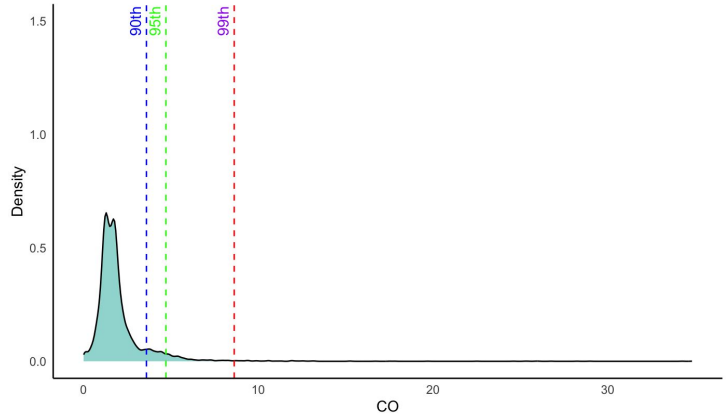
- Find which **controllable variables** the engineers can alter to lead to a **decreased CO output** for their turbine.

- We will test different modelling techniques to gain insights and offer our suggestions to the scientists in Turkey

- Our objective is to find the best model for these three datasets:
  - **Overall**: all TEY values
  - **Medium**: 130-136 MWH
  - **High**: >160 MWH

Data Cleaning

- Outliers can skew the data distribution and negatively impact the performance of predictive models.
- **Removed 72 observations** with CO values greater than the 99th quantile.
- Then the dataset was then **split** into overall, medium, and high data.



Density plot of CO. All observations greater than the red dotted line are removed.

Modeling Techniques

- Linear Regression
  - Baseline model
- LASSO Regression
  - For feature selection and increased interpretability
- Decision Tree
  - Ideal for clear and interpretable recommendations on controllables
- Random Forest
  - Ensemble of decision trees for enhanced prediction and robustness

Model Evaluation

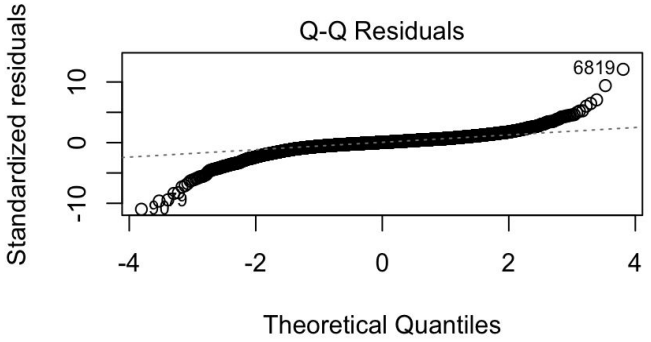
- Train and Test Data:
  - 80:20, randomized split with seed 443
- Metrics used for evaluation:
  - **R-squared:** amount of variation explained by the model
  - **RMSE** (Root Mean Squared Error)
  - **MAE** (Mean Absolute Error)
- Interpretability vs Complexity:
  - Linear methods (LASSO, Linear Regression) are more interpretable, which helps create actionable recommendations
  - Complex methods may fit the data better, but are relatively less interpretable

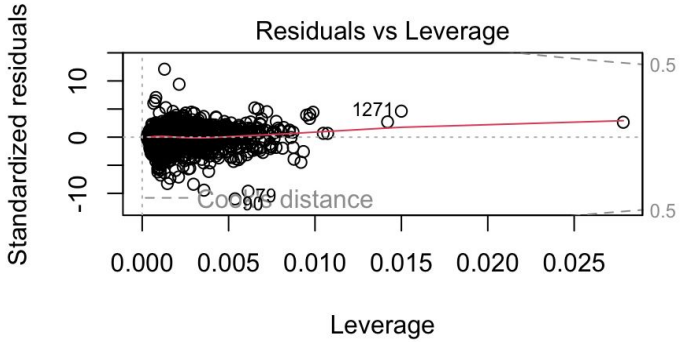
Test Statistics for Overall Data

	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**

Linear Model Assumptions (Overall)

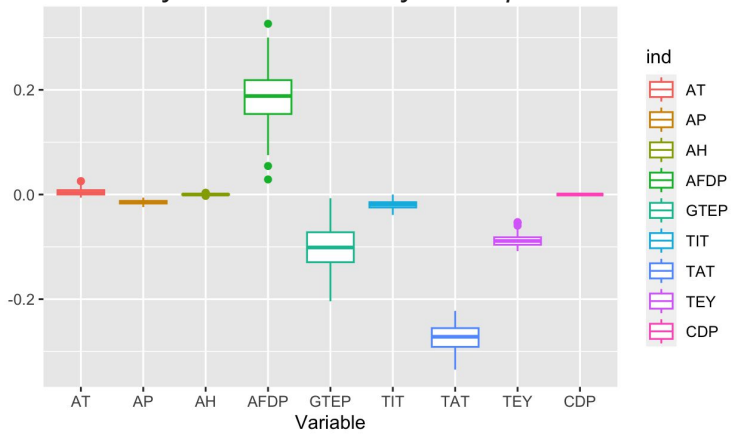
- **Non-normality of errors** and **association among residuals**
- For overall data, these **assumptions are not met**
  - Performed a Box-Cox transformation to attempt linearization





LASSO Model (Overall)

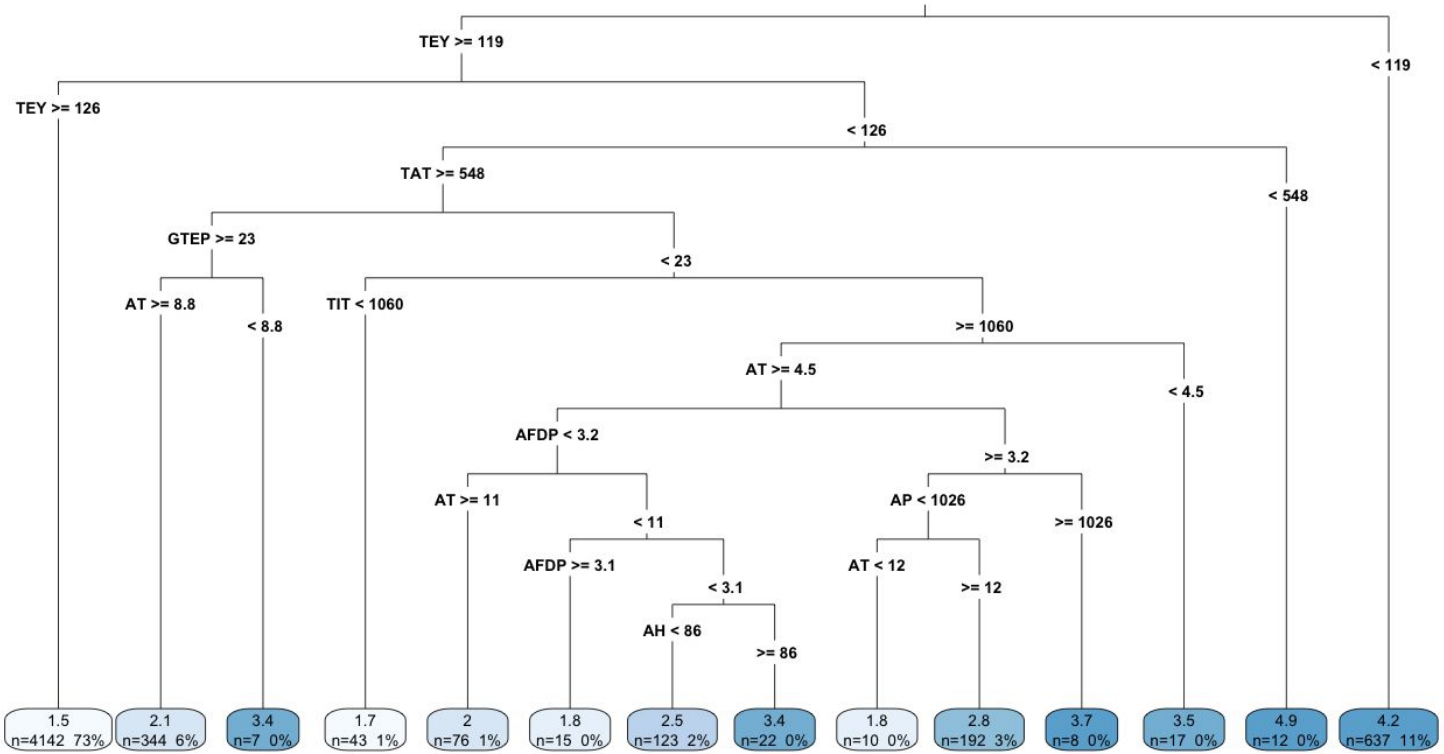
95% Confidence Intervals for the predictors



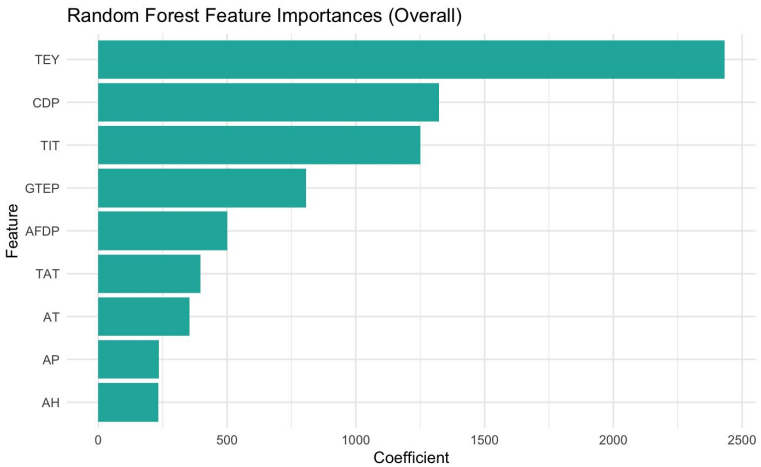
Out of the controllable variables:

- **TAT** is the most significant variable with coefficient **-0.27**.
- **AFDP** is the second most significant variable with coefficient **+0.2**.
- **GTEP** is the third most significant variable with coefficient **-0.1**.

Pruned Decision Tree (Overall)

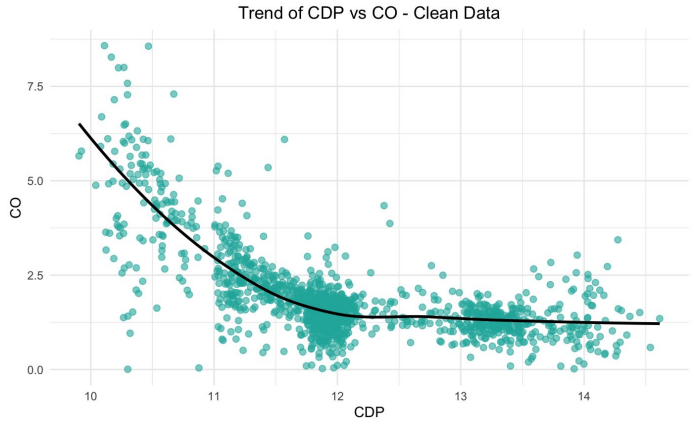


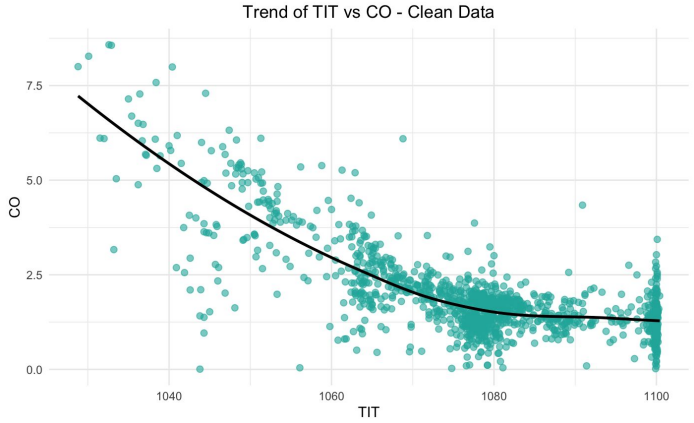
Random Forest – Overall



Out of the controllable variables:

- **CDP** is the most important variable
- **TIT** is the 2nd most important variables
- AFDP, TAT and GTEP are less important





A Closer Look (Overall)

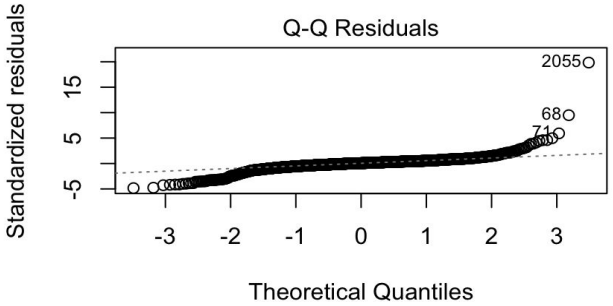
- As CDP increases, predicted CO decreases
- As TIT increases, predicted CO increases

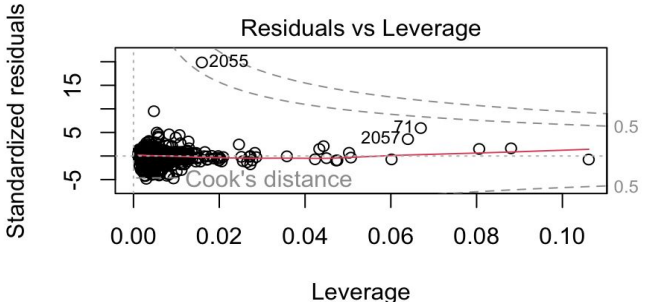
Test Statistics for Medium Yield Data

	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**

Linear Model Assumptions (Medium)

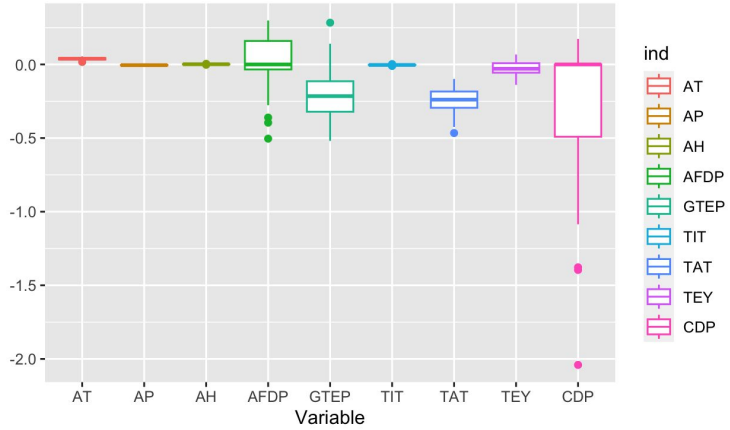
- **Non-normality of errors** and **association among residuals**
- For medium data, these **assumptions are not met**
  - High leverage and high residual points were observed





LASSO Model (Medium)

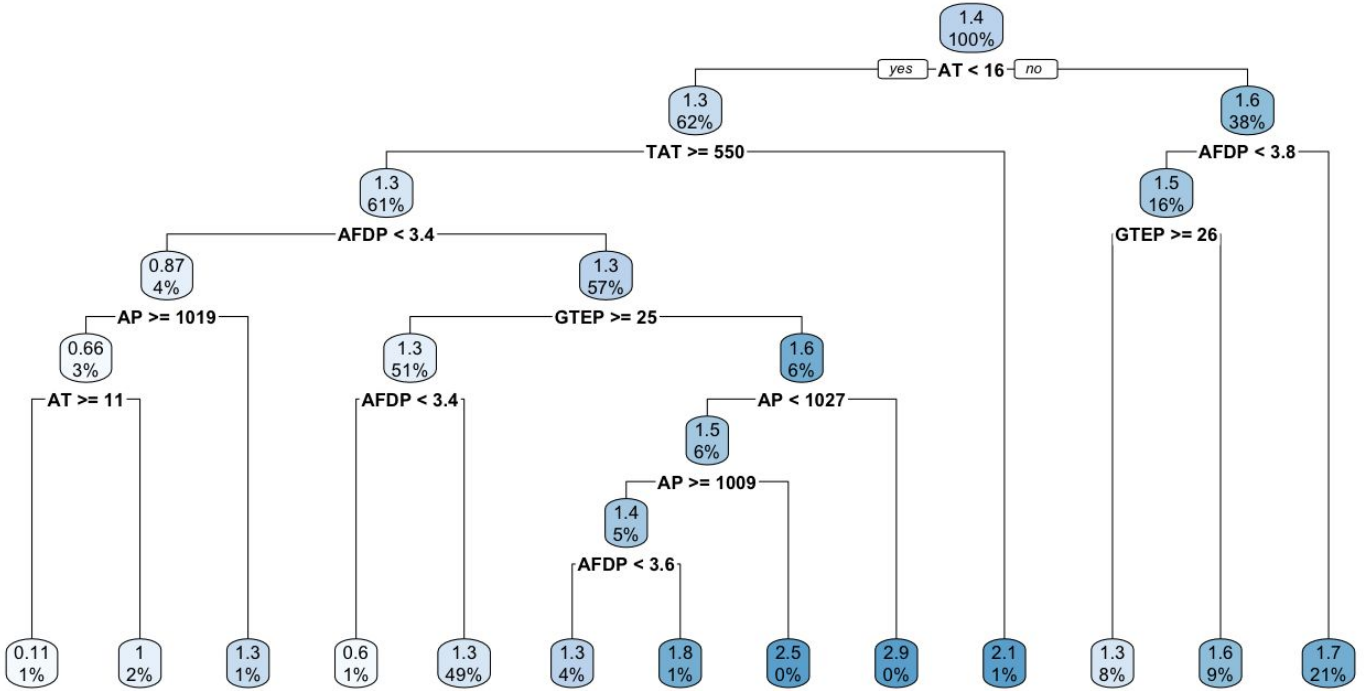
95% Confidence Intervals for the predictors



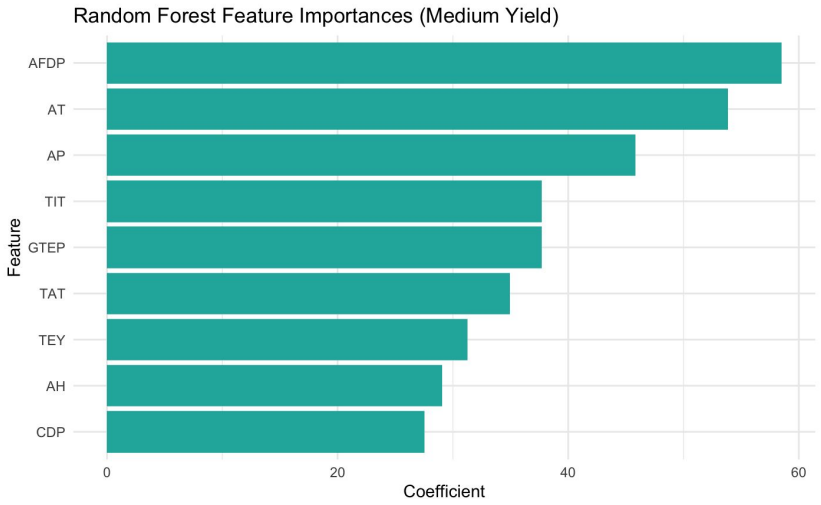
Out of the controllable variables:

- **TAT and GTEP** are the most significant variables with coefficients ≅ **-0.25**.
- **All other variables** have confidence intervals including 0.

Decision tree (Medium Yield)

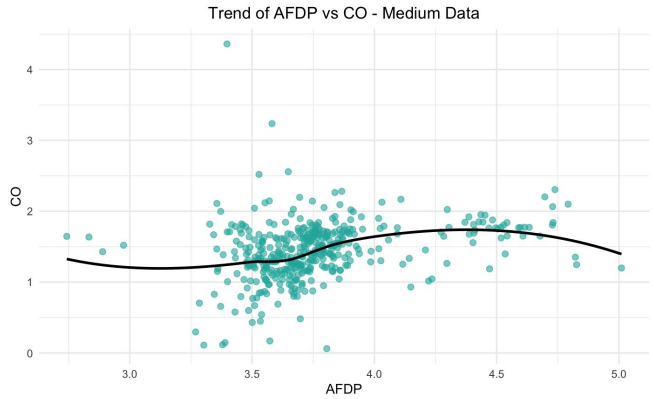


Random Forest – Medium Yield



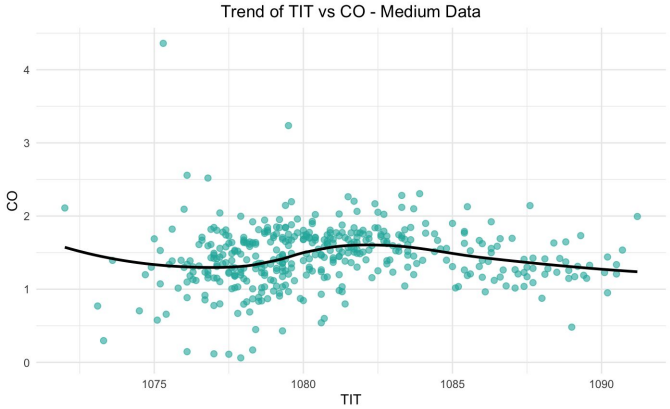
Out of the controllable variables:

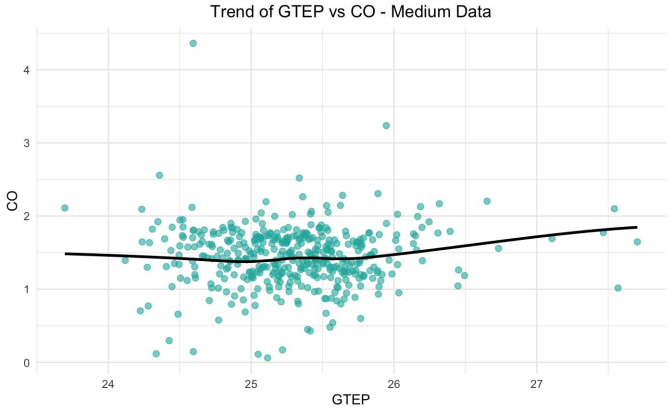
- **AFDP** is the most important variable
- **TIT & GTEP** are 2nd most important variables
- TAT and CDP are less important



A Closer Look (Medium)

- As AFDP increases, predicted CO initially increases then decreases
- As TIT increases, predicted CO initially increases then decreases at the end
- As GTEP increases, predicted CO increases at the end



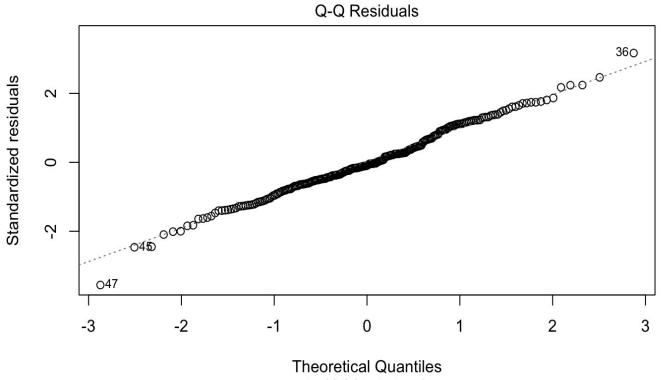


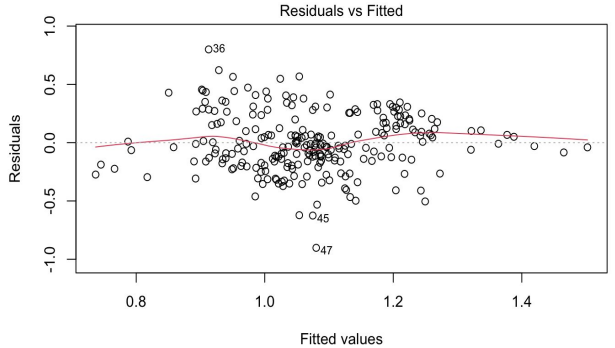
Test Statistics for High Yield Data

	Linear Regression	Lasso	Decision trees	**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**

Linear Model Assumptions (High)

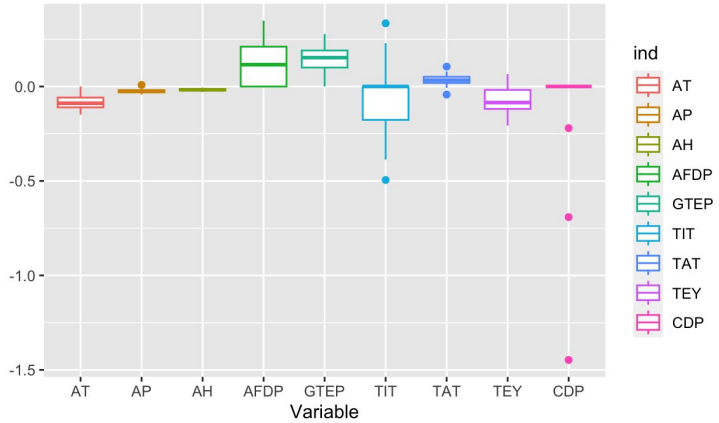
- **Normality of errors** but **association among residuals**
- For high data, the **second assumption is not met**
  - Non-linear trends in residuals were observed.





LASSO Model (High)

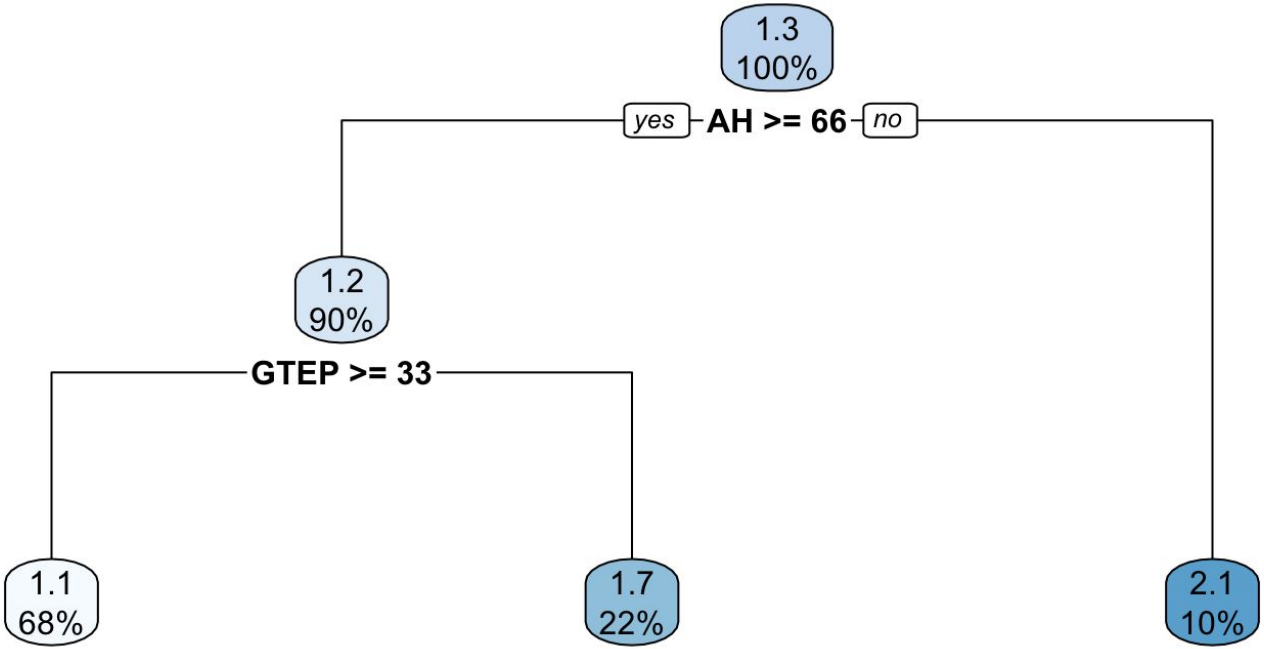
95% Confidence Intervals for the predictors



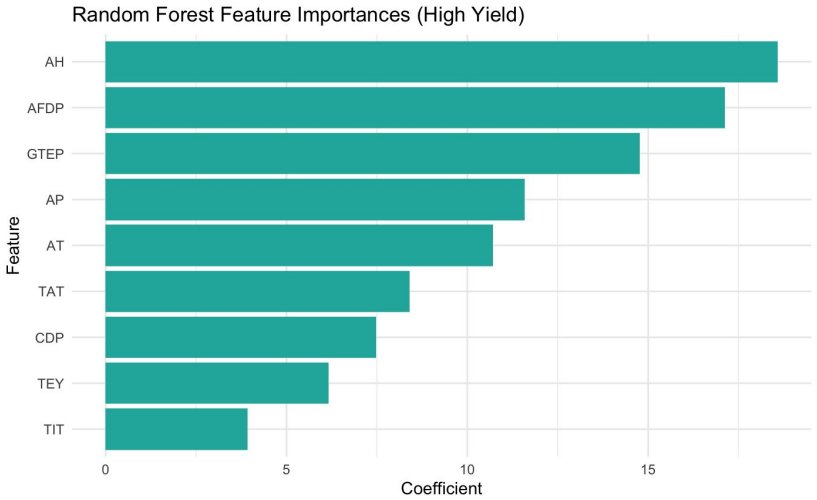
Out of the controllable variables:

- **GTEP** is the most significant variable with coefficients ≅ **0.2**.
- **TAT** is also significant, with coefficient < **0.05.**
- **All other variables** have confidence intervals including 0.

Decision tree (High Yield)

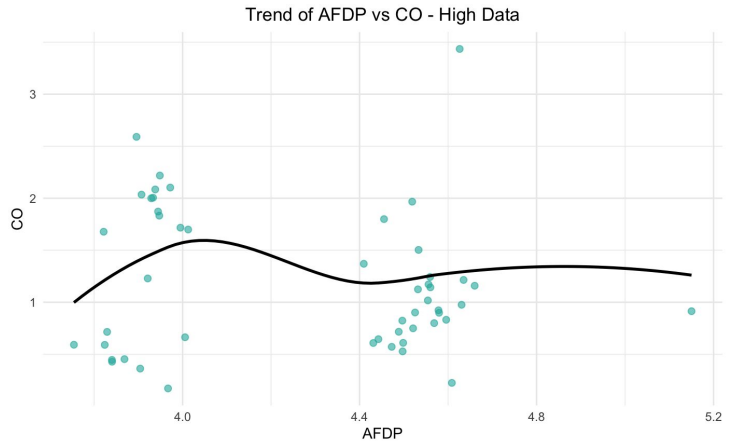


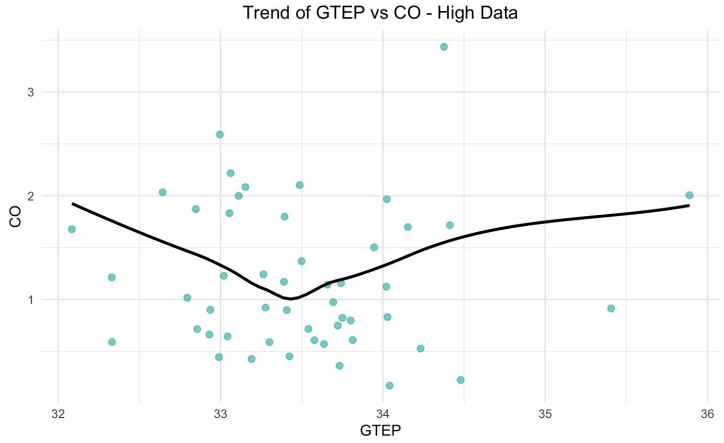
Random Forest – High Yield



Out of the controllable variables:

- **AFDP** is the most important variable
- **GTEP** is the 2nd most important variable
- TAT, CDP and TIT are less important





A Closer Look (High)

- As AFDP increases, CO initially increases and then decreases
- As GTEP increases, CO initially decreases then increases

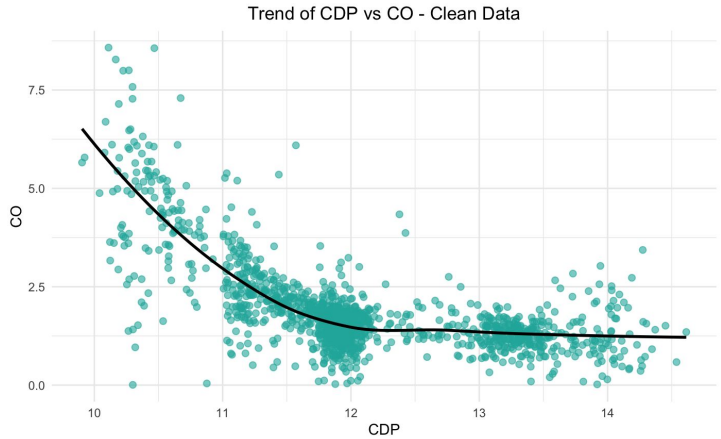
Summary: Modeling

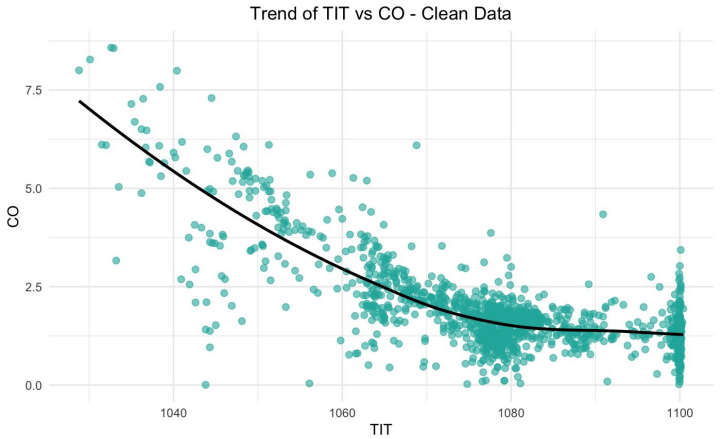
- Random Forest performed the best at overall, medium and high yields
- Model results can be used to generate actionable suggestions to lower CO
- Model performance is low for medium and high yield
    - Low number of data points for model training
    - Limited range of CO values for these data sets

Conclusion: Overall

Suggestions for Reducing CO – Overall

- **CDP** & **TIT** are the most important features according to RF
  - Recommended values of CDP: **12.5 mbar** and higher
  - Recommended values of TIT: **1085 C** and higher

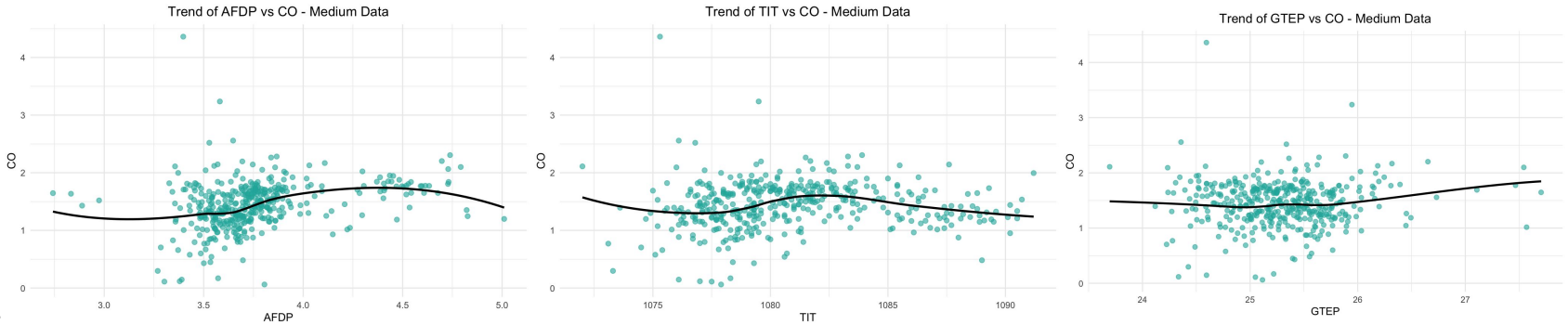




Conclusion: Medium Yield

Suggestions for Reducing CO – Medium Yield

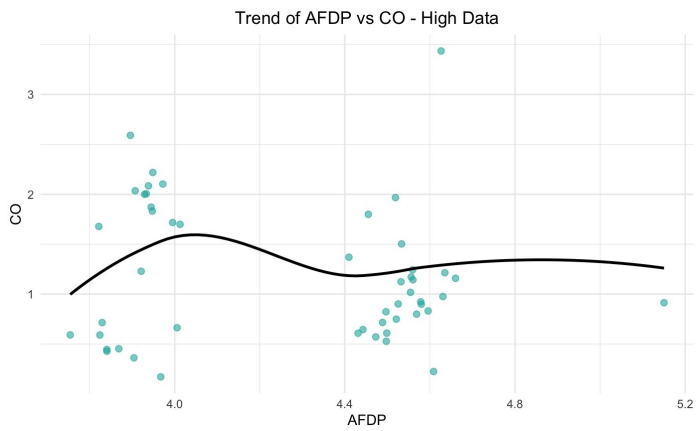
- **AFDP, TIT** & **GTEP** are the most important features according to RF
  - Recommended values of AFDP: **3.6 mbar** and lower
  - Recommended values of TIT: **1088 C** and higher
  - Recommended values of GTEP: around **25 mbar**

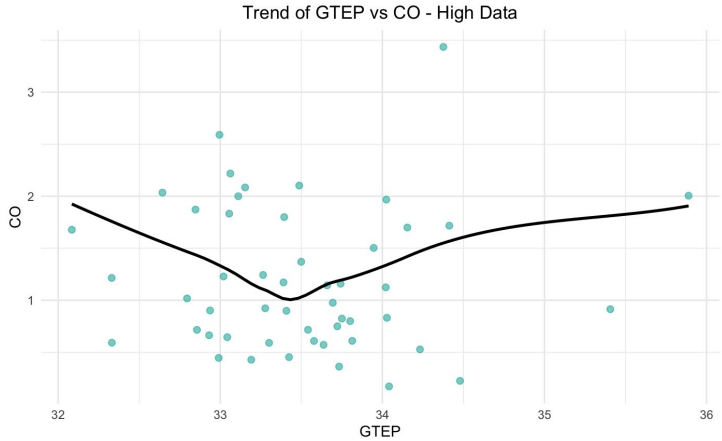


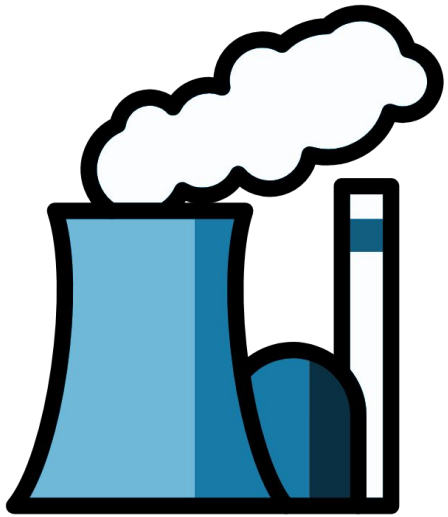
Conclusion: High Yield

Suggestions for Reducing CO – High Yield

- **AFDP** & **GTEP** are the most important features according to our RF
  - Recommended values of AFDP: **4.4 - 4.5 mbar**
  - Recommended values of GTEP: **33.5 mbar**







Thank You

Questions?