STAT 443 Consulting Project Final Presentation

Presented by Subin, Jackie, Jonas, and Chenfei on May 5, 2021

Agenda

I. Introduction, Methodology & Description

II. Data Modeling in Three Energy Yield Levels

III. Conclusion

IV. Q&A

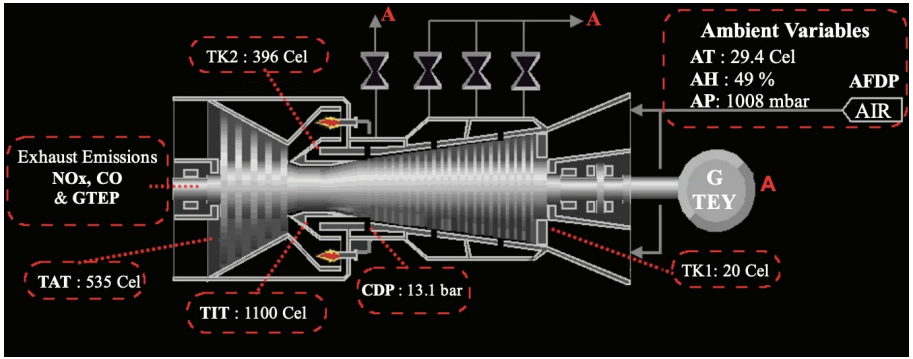
Introduction, Methodology & Description

Dataset Background:

- The dataset contains 36,733 instances of 11 sensor measures aggregated over one hour, from a gas turbine located in Turkey for the purpose of studying flue gas emissions, namely CO and NOx.

Attribute Information:

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



Heysem Kaya, Pinar Tufekci and Erdinc Uzun. '**Predicting CO and NOx emissions from gas turbines: novel data and a benchmark PEMS**', *Turkish Journal of Electrical Engineering & Computer Sciences, vol. 27,* 2019, pp. 4783-4796, [Web link]

Introduction, Methodology & Description

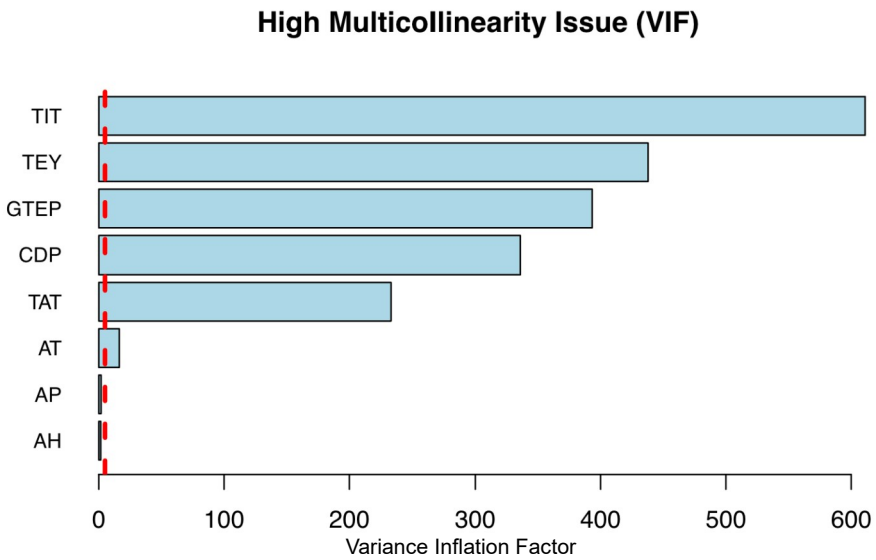
Research Questions

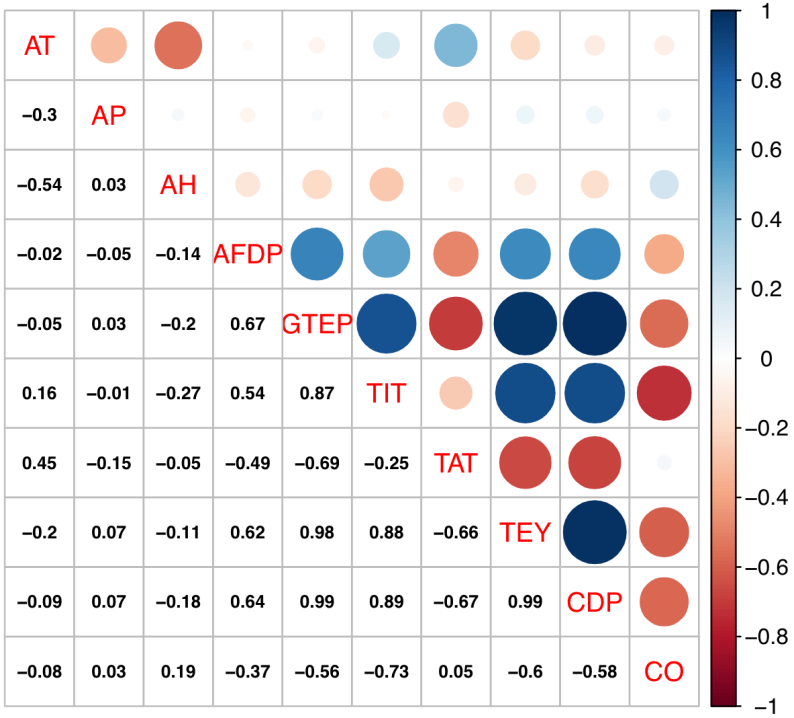
- Is there a relationship between CO emissions and the process variables such that CO emission levels can be reduced through tweaking process variables?
- Can some variables be removed from the model and still explain the variation in CO levels in order to explain the relationship better?
- Identify relationships between CO levels and process variables in different Turbine Energy Yield ranges (i.e. Full Range, Middle Range (130-136 MWH), and High Range (160+ MWH)). Are there any useful models for different ranges?

Data Modeling in Three Energy Yield Levels - All-ranges

All Range Analysis

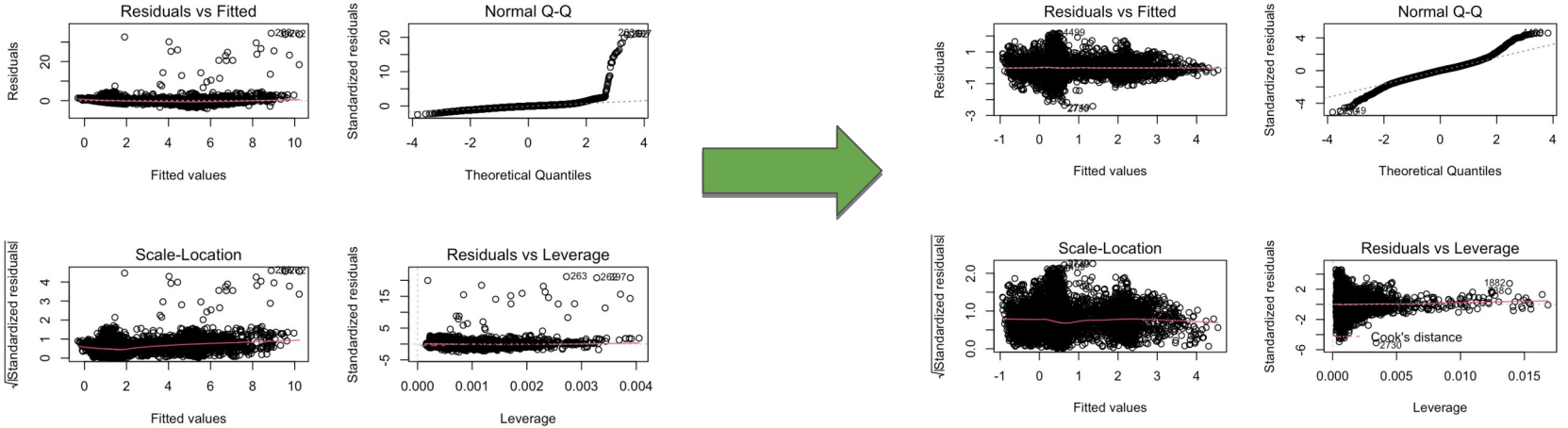
- **Multicollinearity issues** between TIT, TEY, GTEP, CDP, and TAT
  - Removed problematic variables: **CDP** and **TAT**





Data Modeling in Three Energy Yield Levels - All-ranges

- **Outlier Inspection**
  - Removed 125 influential points (1.6% of the data)
- **Transformation**
  - Quadratic transformation on predictors: AP, TIT, and TEY
  - A Box Cox transformation on the dependent variable CO

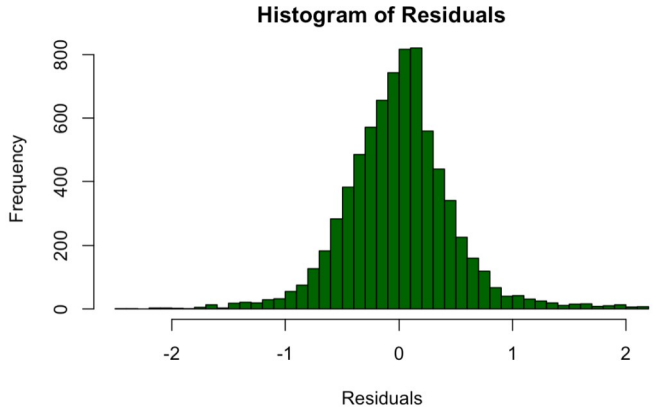


Data Modeling in Three Energy Yield Levels - All-ranges

Final selected model:

- The most effective predictors appear to be: **AT, AP, AH, TIT,** and **TEY**
- Relationships between CO and predictors in All-ranges models:
  - $(\mathbf{CO}^{0.465} - 1) / 0.465 = -0.4 + (0.031 * \mathbf{AT}) + (4.71 * \mathbf{AP} - 4.55 * \mathbf{AP}^2) + (0.0077 * \mathbf{AH}) + (-108.2 * \mathbf{TIT} + 7.29 * \mathbf{TIT}^2) + (22.27 * \mathbf{TEY} - 14.54 * \mathbf{TEY}^2) + \varepsilon$
  - Adjusted $R^2 = 0.789$
  - Root Mean Square Error (RMSE): 0.476

- To decrease the CO emission:
  - **AT** ↓ **AH** ↓ **TIT** ↓
  - **AP** ↑ **TEY** ↑



Conclusion - All-range Analysis

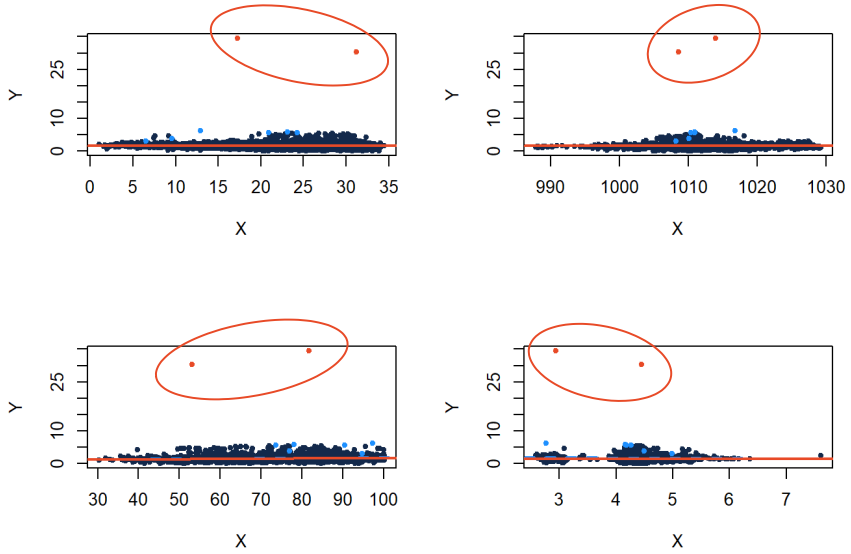
Summary

- AFDP was statistically insignificant, so it was removed first
- There was severe multicollinearity problem between predictors, such as TIT, TEY, GTEP, CDP, and TAT
  - Their VIF (Variance inflation factor) values were over 200
  - By comparing each variable, CDP, TAT, and GTEP were removed
- There were 125 influential points that had a large effect on the model. Since they were only 1.6 % of the full data, we decided to remove the outliers
- Box Cox and quadratic transformations were used to improve the model regression
- With the final chosen model, we can say that about 79% of the data fit the regression model

Data Modeling in Three Energy Yield Levels – Mid-range

Middle Range Analysis (130-136 MWH)

- Multicollinearity issues between AT, GTEP, TIT, and CDP
  - Removed problematic variables.
- Outlier Inspection
  - Investigated the outliers, but kept them.
- Transformation
  - Quadratic transformation on predictors helped improve the model performances.
- Two models were chosen:
  - **Model 1** features AT, AP, AH, AFDP, TAT, TIT, CDP (Pro: More control over variable values, Con: Less reliable model)
  - **Model 2** features AT, AP, AH, AFDP, TAT, TIT, NOX (Pro: More reliable, Con: Less control)



Data Modeling in Three Energy Yield Levels – Mid-range

Relationships between CO and predictors in Mid-Range models:

- $\sqrt{CO} = a_0 + a_1\mathbf{AT} + a_2\mathbf{AP} + a_3\mathbf{AH} + -a_4\mathbf{AFDP} + -a_5\mathbf{AFDP^2} + -a_6\mathbf{TIT} + -a_7\mathbf{TIT^2} + -a_8\mathbf{TAT} + a_9\mathbf{TAT^2} + a_{10}\mathbf{CDP} + a_{11}\mathbf{CDP^2} + \varepsilon$
  - Adj. $R^2 = 0.1607$
- $\sqrt{CO} = b_0 + b_1\mathbf{AT} + b_2\mathbf{AT^2} + b_3\mathbf{AP} + -b_4\mathbf{AP^2} + b_5\mathbf{AH} + -b_6\mathbf{AH^2} + -b_7\mathbf{AFDP} + -b_8\mathbf{AFDP^2} + -b_9\mathbf{TAT} + b_{10}\mathbf{TAT^2} + -b_{11}\mathbf{TIT} + -b_{12}\mathbf{TIT^2} + b_{13}\mathbf{NOX} + -b_{14}\mathbf{NOX^2} + \varepsilon$
  - Adj. $R^2 = 0.2814$

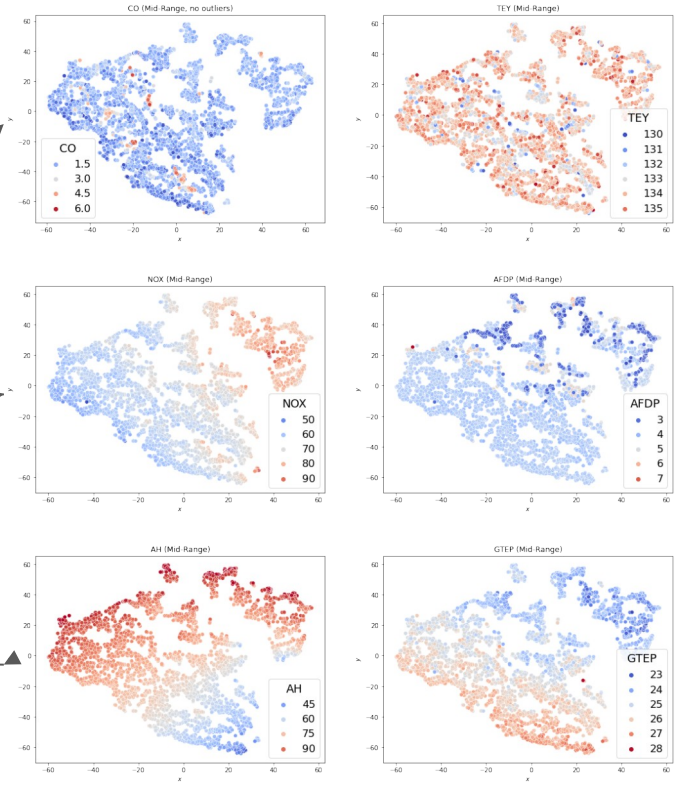
- **First model** (desired effect CO ↓)
  - Ambient ↓ AFDP ↑ TAT ↓ TIT ↑↓ CDP (stay the same)
  - Least Effective: AP, CDP*
  - Somewhat Effective: AT, AH, AFDP, TAT
  - Most Effective: TIT

- **Second model**
  - Ambient ↓ AFDP ↑ TAT ↓ TIT ↑ NOX ↓
  - Least Effective: AP, TAT*
  - Somewhat Effective: TIT, AFDP
  - Most Effective: AT, NOX, AH

Data Modeling in Three Energy Yield Levels – Mid-range

Clustering Analysis

- While there is an underlying clustering structure for the full dataset, there isn't a strong one within the Mid-Range dataset (~2-3 clusters).
- CO levels tend to be higher in the same cluster with TEY levels below the Mid-Range. Otherwise, there are no patterns in other clusters.
- Unable to detect strong/useful patterns for CO, TAT, TEY, and AP.
- Detected strong/useful patterns for NOX, AFDP, GTEP, TIT, CDP, AT, and AH. Note: GTEP, TIT, CDP, and AT share very similar patterns (i.e. multicollinearity).
- It is likely that CO might not have a strong relationship with these predictors within this range.



Conclusion - Mid-Range Analysis

Suggestions/Recommendations

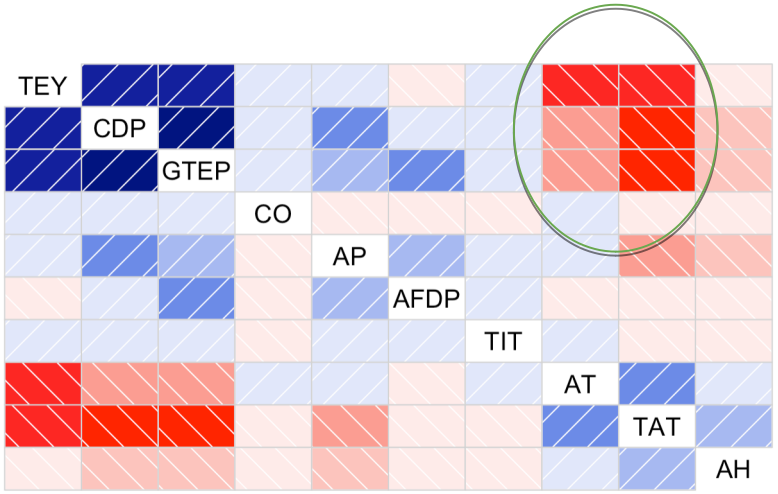
- Model Building and Diagnostics didn't yield a very trustworthy model for the Mid-Range Dataset.
  - TEY range is too small to avoid dealing with more noise than meaningful relationships.
  - The model for the Full Dataset or for a larger TEY range would be more useful.
- NOX seems to have a stronger relationship than CO with these predictors.
  - The client could have very good models for NOX if interested. Otherwise, there might be unknown variables that would be better suited for a CO model.
  - An NOX model could help decrease NOX emissions, which could help decrease CO emissions.

- If the client wants to use a Mid-Range model…
  - The most effective predictors appear to be AT, AH, AFDP, TIT, and NOX.
    - Decrease AT, AH, and NOX (Less control)
    - Increase AFDP and TIT (More control)
  - Results from these models should be met with some skepticism.

Data Modeling in Three Energy Yield Levels – High-range

High Model Analysis (TEY > 160 MWH)

- Transformation
  - Model 1: Boxcox: sqrt CO <- diagnostics plots look better(normality)
- Multicollinearity issues between TEY, CDP, GTEP, TAT
  - **Model1**:
    - Perform Farrar - Glauber Test to check for overall and individual multicollinearity diagnostic
    - Use t-test of partial correlation of coefficients to check the pattern of multicollinearity
    - Remove AT, TIT, & CDP & perform ridge regression _
- Outlier Inspection
  - Model1: removed 1 outlier: 585 observations
  - Model2: removed 4 outlier: 582 observations
- Two models
  - **Model1:** (without NOx) features AP, AH, AFDP, GTEP, TAT, TEY (Pro: More control over variable values, Con: Less reliable model)
  - **Model 2:** (with NOX) features AT, TAT, GTEP, TEY, NOX



Data Modeling in Three Energy Yield Levels – High-range

Relationships between CO and predictors in High-Range models:

- $\sqrt{CO} = 47.187 - 0.005 * \mathbf{AP} - 0.002 * \mathbf{AH} - 0.025 * \mathbf{AFDP} - 0.112 * \mathbf{GTEP} - 0.062 * \mathbf{TAT} - 0.029 * \mathbf{TEY} + \varepsilon$
  - Average R^2(train) = 0.2033
  - Average RMSE(train) = 0.2882
  - Average R^2(test) = 0.1713
  - Average RMSE(test) = 0.2942
- $CO = 98.842 + 0.0286 * \mathbf{AT} - 0.155 * \mathbf{TAT} - 0.373 * \mathbf{GTEP} - 0.0265 * \mathbf{TEY} + 4.093 * \mathbf{NOX} - 0.718 * \mathbf{NOX^2} + \varepsilon$
  - Adj. R^2 = 0.3949

- **First model (TO decrease the CO emission)**
  - AP↑ AH↑ AFDP↑ GTEP ↑ TAT↑ TEY↑ **(Ambient variables have less impact on the CO emission compared to the process variables)**
  - Least Effective: AP, AH
  - Somewhat Effective: AFDP, TAT, TEY
  - Most Effective: GTEP

- **Second model with NOX (TO decrease the CO emission)**
  - AT ↓ TAT↑ GTEP ↑ TEY↑ NOX↑
  - Somewhat Effective: AT, TEY
  - Most Effective: AT, GTEP, NOX

Table for Model 1

	variables	values
1	rmse_train_mean	0.288227266
2	rmse_test_mean	0.294173494
3	r2_train_mean	0.203316833
4	r2_test_mean	0.171290259
5	intercep_mean	47.186992653
6	ap_mean	-0.005201394
7	ah_mean	-0.001808442
8	afdp_mean	-0.025073188
9	gtep_mean	-0.111746843
10	tat_mean	-0.061647308
11	tey_mean	-0.029287514

Table for Model 2

Coefficients:

	Estimate
(Intercept)	98.842006
AT	0.028563
TAT	-0.154917
GTEP	-0.373090
TEY	-0.026454
poly(NOX, 2)1	4.093483
poly(NOX, 2)2	-0.717589

Conclusion - High-Range Analysis

Summary & Suggestions:

- Multicollinearity issue is a big concern for modeling high range data and is not fully addressed regardless of the improvements performed
- For High-Range data, though the average R^2 for the model without NOx is relatively low, the coefficient interpretation matches common sense that ambient variables have less impact on CO emission compared to the process variables

- The model with NOX provides a good adjusted R-squared (highly reflected the CO emission) and with comparatively small VIF value (low multicollinearity issue risk). Hence, based on that model with NOX, our recommendation is that to decrease the Ambient Temperature and increase the Gas turbine exhaust pressure (GTEP), Total energy yield (TEY), and Nitrogen oxides (NOX)

Conclusion - All Models

Summary

- All range: The multicollinearity issue was the main problem, but removing influential observations and a power transformation solved the issue and helped improve the model.
- Mid range: The main issue about the mid range data is about the small adjusted R-square value
- High range:
  - Model1: multicollinearity issue is a big concern
    - Box Cox transformation helps with diagnostics plots, Farrar - Glauber Test helps identify and remove variables causing multicollinearity issues, ridge regression better address multicollinearity to improve model fit
  - Model2: Including NOX value does solve the multicollinearity issue to some degree. Thus, we provide two models for high range for clients to select the most practical one from their perspective (i.e from the car turbine engineering facet).

Conclusion: Takeaways

Real-World Takeaways

- The model for the full dataset would be best one to use
  - Mid-Range model isn't very trustworthy because the data is too "narrow".
  - High-Range model also has limited sample size

Final Model Selection: Full-Range Model

- $(\mathbf{CO}^{0.465} - 1) / 0.465 = - 0.4 + (0.031 * \mathbf{AT}) + (4.71 * \mathbf{AP} - 4.55 * \mathbf{AP}^2) + (0.0077 * \mathbf{AH}) + (-108.2 * \mathbf{TIT} + 7.29 * \mathbf{TIT}^2) + (22.27 * \mathbf{TEY} - 14.54 * \mathbf{TEY}^2) + \varepsilon$
- The most effective predictors are **AT**, **AP**, **AH**, **TIT**, and **TEY**
- With this model, to reduce CO emissions…
  - Ambient temperature, ambient humidity, and turbine inlet temperature need to be decreased
  - Ambient pressure and turbine energy yield need to be increased

Q&A

Thanks for listening!
Any questions?