Gas Turbine Nitrogen Oxide Emission Reduction Project

Group 3: Hye Rim Ahn, Ron Basak, Ben Burda, Yehjee Kim, Seung Gyun Shin

I. Introduction

Air pollution has been a growing problem with more and more of the world needing energy every day. It is one of the leading causes of death in our population, with around 11.65%[1] of deaths being caused by air pollution alone. With pollution having a huge impact on the environment, it's important to find ways to reduce greenhouse gas emissions despite the increase in demand for energy. The UCI Machine Learning Repository's dataset contains information on Turkish gas turbine carbon monoxide and nitrogen oxide emissions. The dataset contains 36,733 instances with 11 sensor measures, or variables, that were collected over an hour. Of those instances, we focused on select ranges of Turbine Energy Yield values and statistically analyzed 7,158 out of them. The models that we have created aim to identify ways to reduce nitrogen oxide emissions in relation to the various parts of the turbines and environment.

The variables that we considered measure different aspects of the turbine and the environment surrounding it. Ambient temperature (AT), ambient pressure (AP), ambient humidity (AH), and turbine energy yield (TEY) are factors that cannot be controlled, while air filter difference pressure (AFPD), gas turbine exhaust pressure (GTEP), turbine inlet temperature (TIT), turbine after temperature (TAT), and compressor discharge pressure (CDP) are factors that can be controlled. These 9 factors are considered as predictor variables. Nitrogen oxide (NOx) and carbon monoxide (CO) are the two response variables from the observations, and we focused on the nitrogen oxide levels as our response variable for our models.

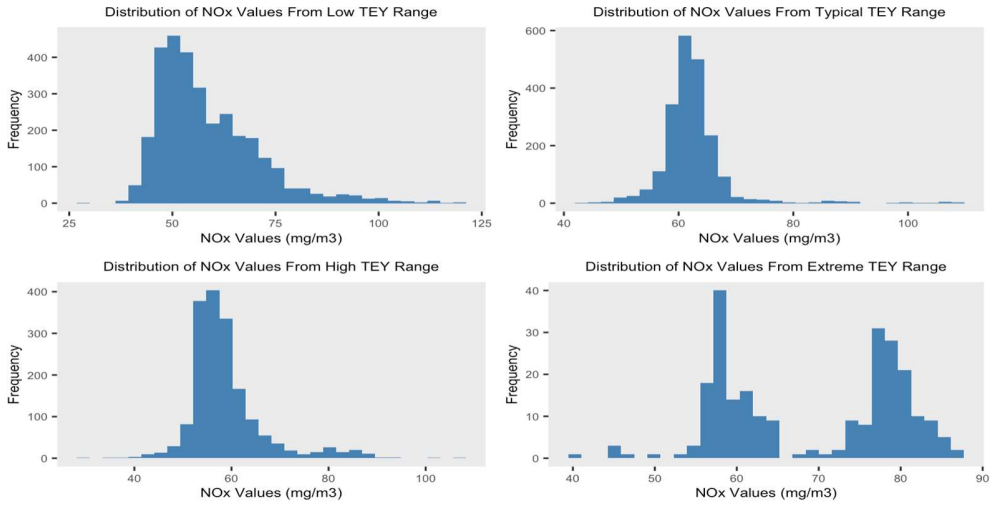
II. Methods - Data Exploration

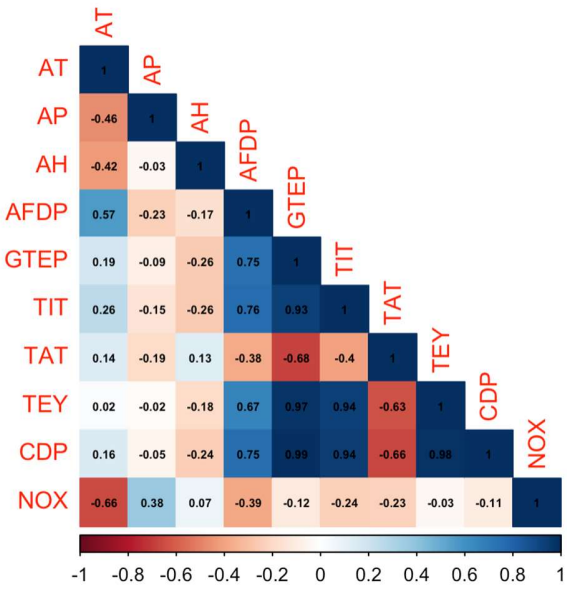
In order to have a well fitted model, the dataset was separated into four different groups based on the turbine energy yield values. Energy yields that were less than 130 megawatts/hour are considered as low energy yields and were grouped together. Energy yields that were between 130 megawatts/hour and 136 megawatts/hour are considered as typical and were grouped together. Energy yields that were between 136 megawatts/hour and 160 megawatts/hour are considered as high and were grouped together. Energy yields that were above 164 megawatts/hour are considered to be extremely high and were grouped together. The groupings allow for the model and its variables to be more accurate and curated given a specific turbine

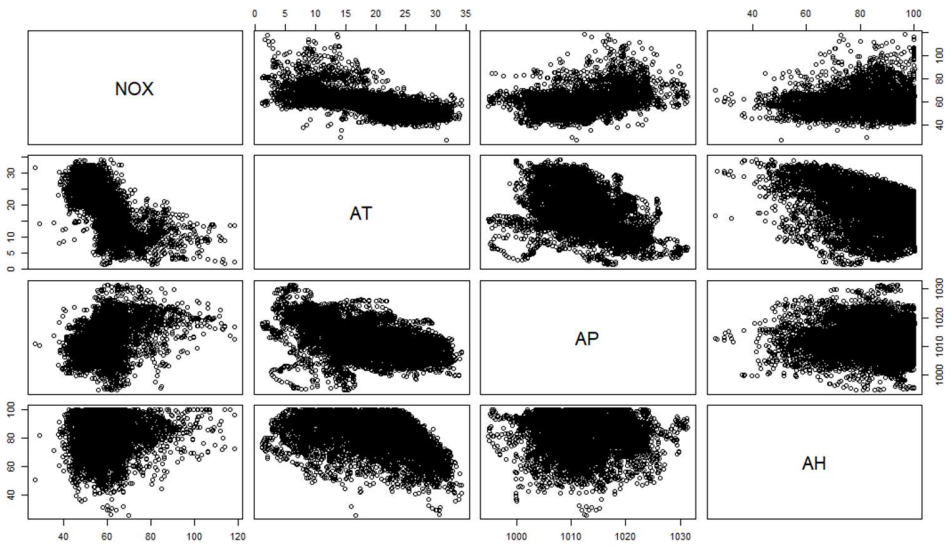
[1]Hannah Ritchie and Max Roser (2017) - "Air Pollution". OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/air-pollution' [Online Resource]

energy yield output. Initially, we created linear models using all four energy yield categories; but, for our final objective and based on our client's needs after the second meeting, we focused on two ranges: the typical range (130 - 136 megawatts/hour) and the extreme range (160+ megawatts/hours).

Our response variable, nitrogen oxide (NOx), overall follows a normal bell curve, meaning that the majority of the values fall within the middle range and few values fall in the extremes. When looking at the data points separated into their respective energy yield groups, we can see that the low, typical, and high turbine energy yield groups also follow a relatively normal bell curve. The group with extremely high turbine energy yields follow a bimodal distribution. The correlation heatmap shows the correlation structure between our variables, with strong negative correlations in red and strong positive correlations in blue.







III. Methods - Models

As part of our methodology to determine which predictors are statistically significant for predicting NOx levels, we have decided to leverage different statistical models. These models are linear regression, ridge regression, k nearest neighbors (KNN), and random forest. Although these models yielded different predictors, we have arrived at a final model that best minimizes nitrogen oxides levels.

Each dataset was split up into test and training groups, meaning that a portion of the dataset will be used for creating the model while the other portion will be used to test and validate that model to avoid overfitting or comparing from previous data points.

A. Linear Regression Model

We created a multiple linear regression model as our first model. We believed this was a reasonable starting point since the model is used to estimate the relationship between two or more independent variables and one dependent variable. There are a few assumptions that need to be met in order for us to successfully use a multiple linear regression model. These are little to no multicollinearity, normality, and homoscedasticity. Thus, as part of the prerequisites to using a linear regression model, we checked for multicollinearity issues; these were illustrated in the

data exploration stage. To account for this, we used the variance inflation factor (VIF) with the cut of value of 10. In other words, if predictors have a value of 10 or higher, we dropped the predictor from the model since it posed a multicollinearity issue. From this, we utilized backward elimination to determine the significant predictors that would be used for our model.

For the low energy range, variables AFDP, GTEP, TAT, AT, AP, and AH were chosen for the model. For the typical energy range, variables AFDP, CDP, TIT, AT, AH, AP, and TEY were chosen for the model. For the high energy range, variables AFDP, TAT, AT, AP, and AH were chosen for the model. For the extreme energy range, variables AFDP, TIT, TAT, AT, AP, and TEY were chosen for the model.

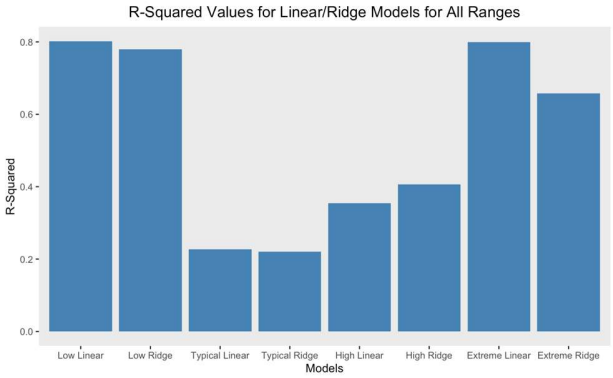
The summaries of our models give information regarding our models, such as the R^2 value and the Mean Squared Prediction Error (MSPE). The R^2 value is a representation of the goodness-of-fit of a statistical model that represents what proportion of variation in a response variable is explained by the independent variables. The Mean Squared Prediction Error is the mean of the squared sum of the differences between the predicted value from the model and the actual value. This can measure the accuracy of our model.The model has a R^2 values of 0.226 and 0.799 for the typical TEY and the extreme TEY energy ranges respectively. The MSPE for the typical energy range and extreme range are 91.02 and 113.29 respectively. Though the model does well for extreme energy values, the model does not perform well for the typical energy range (130 - 136 megawatts/hour).

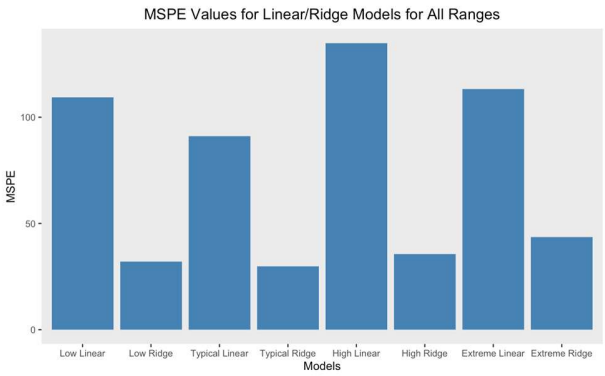
B. Ridge Regression Model

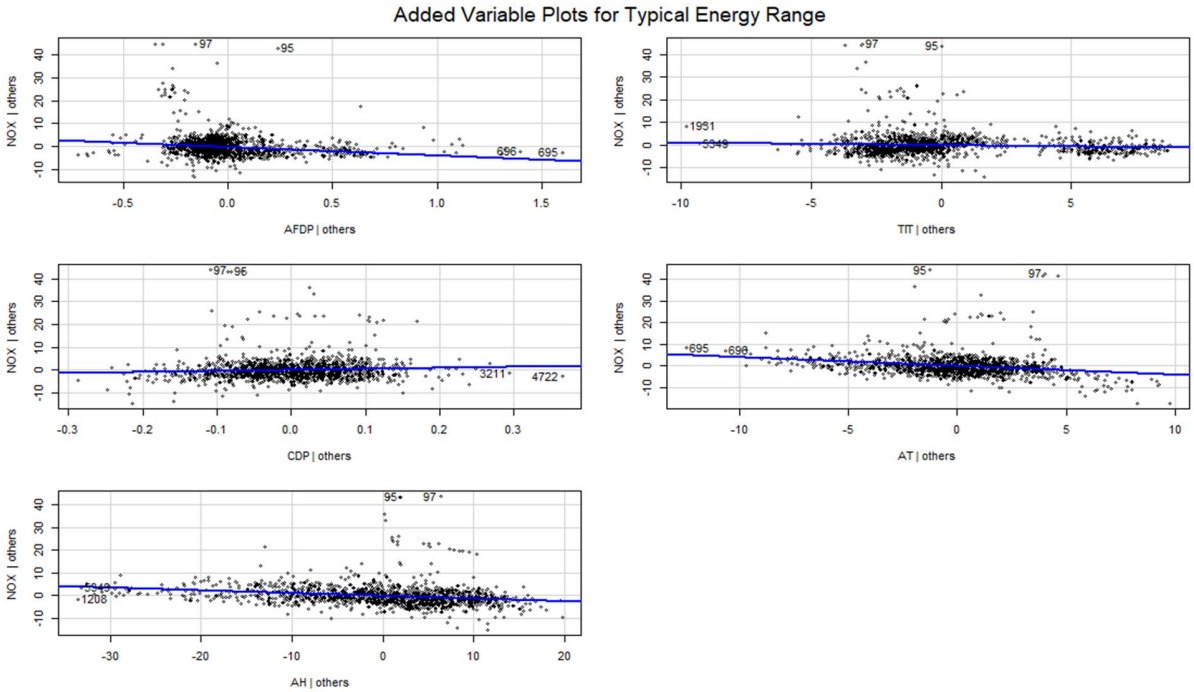
Due to the levels of multicollinearity, ridge regression was a model we considered as well. We chose to user ridge regression as our next model since there are high levels of multicollinearity within the variables and there are steps to standardize the variables to combat this issue. However, note that the lasso regression model is also a viable option but we chose to use ridge regression over lasso regression. This was because we wanted to preserve all the variables in the model for further statistical analysis and ridge regression is regarded as a better option over lasso regression when there is extremely high multicollinearity between features since it reduces variance in exchange for bias.

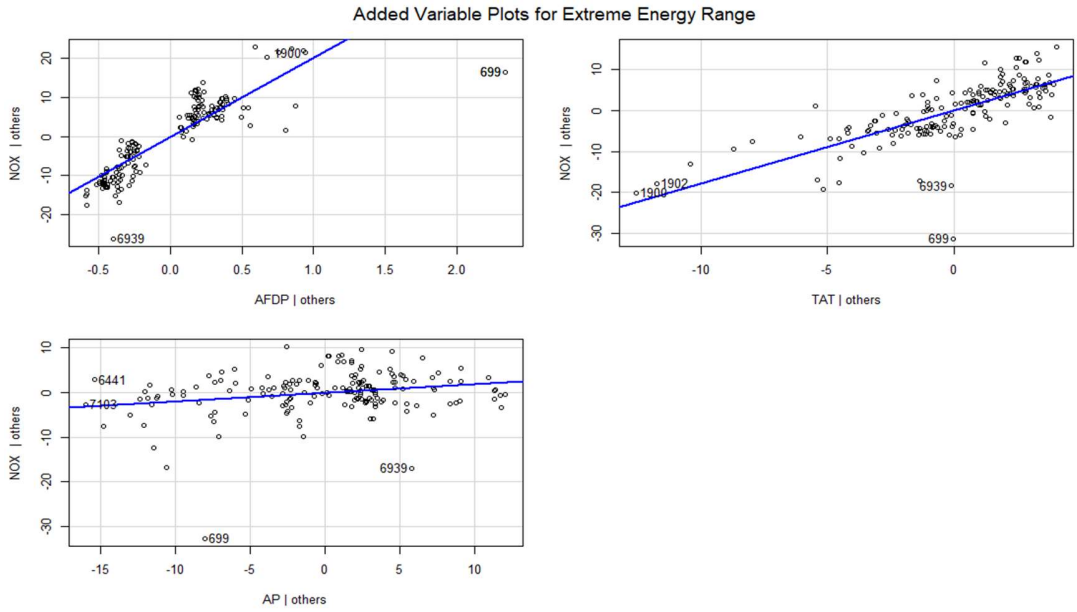
For the dataset containing low TEY values, variables AFDP, GTEP, TAT, AT, and AH were selected from the model. For the dataset containing typical TEY values, variables AFDP,

TIT, CDP, AT, and AH were selected in the model. For the dataset containing high TEY values, variables AFDP, TAT, AT, AP, and AH were selected in the model. Finally, for the dataset containing extreme TEY values, variables AFDP, TAT, and AP were selected in the model. Below are the R^2 and Prediction Error values taken from the results of our modeling processes.









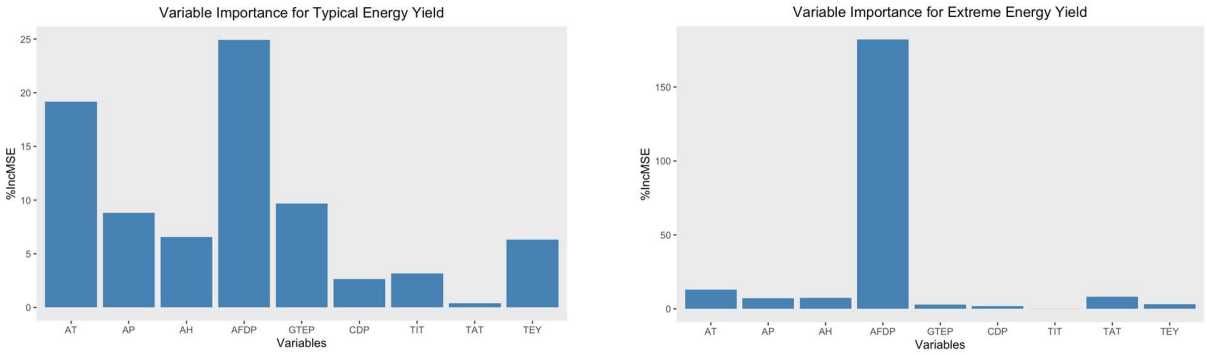
C. KNN Model

Given the results of the initial modeling methods, we determined it was appropriate to consider a k-nearest-neighbor (KNN) model with the data. K-Nearest-Neighbor is an algorithm that makes predictions depending on the values of k-number of the nearest data points, which can be used in both classification and regression modelings. This case as our response variable is numeric, we used KNN regression. In k-nearest-neighbor modeling, it is important to find the best k-value to make the best prediction. We chose the best k-value for each energy level by parameter tuning depending on MSPE from repeated cross validation. For both energy ranges, this k was equal to four. After finding fitting models with appropriate k-values, we have calculated prediction errors to compare performances of models with other models. We could get significant results on the extreme energy range. However, we still got a poor prediction error on the typical range. In addition, as the k-nearest-neighbors regressor is a non-parametric method, regardless of how good our predictions are, we could not introduce meaningful insights and suggestions to reduce emission rate of NOx. With this, we decided to try other regression models.

D. Random Forest Model

Though our KNN model did not work as expected, we found the random forest model to

be our best model. We chose to explore the dataset using random forest for three main reasons. Firstly, it is a versatile, relatively easy model to construct since there is no implicit formal distribution assumption that needs to be checked compared to other models such as the linear regression. Moreover, since random forests are non-parametric, they can handle skewed and multi-modal data well, which were evident during our data exploration phase. Secondly, random forest is a robust model since it leverages bootstrapping and consists of several independent decision trees. By adding randomness and searching the best feature among a random subset of features, the model prevents underfitting. Lastly, it is easy to measure the relative importance of each feature on the prediction. This would enable us to offer the ideal recommendations to mitigate nitrogen oxide levels to the client.

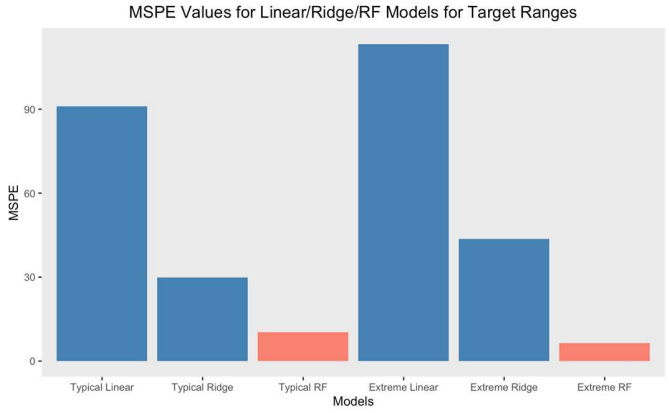


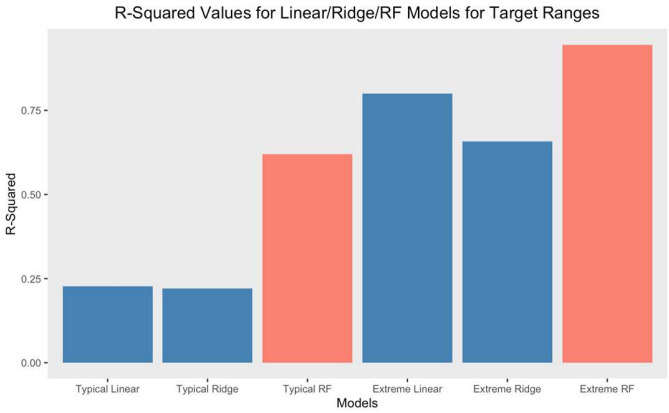
Using this model in the typical range, we determined the order of predictors from most significant to least significant as follows: AFDP, AT, GTEP, AP, AH, TEY, TIT, CDP, and TAT, which can be seen in the figure above. The best parameters to tune that are within the client's control to lower nitrogen oxide are AFDP and GTEP. Overall, the model achieved a R^2 value of 0.62 and MSPE of 10.36.

However, in the extreme energy range, we determined the order of predictors from most significant to least significant as AFDP, AT, TAT, AH, AP, TEY, GTEP, CDP, TIT. From the variable importance bar graph above, we can notice that AFDP is an extremely influential predictor in the model. The overall performance of the model in this range is must better at predicting than in the typical energy range since it has an R^2 value of 0.94 and a MSPE of 6.36.

IV. Results / Best Model

Our best model for both the typical and extreme energy ranges ended up being the random forest models. The models produced the highest R^2 values for each range and gave us information on what variables are most important to reduce nitrogen oxide. In the typical energy range AFDP and GTEP were the most important predictors that are possible to manipulate, with the three ambient variables also being very important. In the extreme energy yield range AFDP was by far the most important variable. The relationship between AFDP and NOx in the extreme range showed two distinct clusters, one where AFDP was less than 4.25 and one where AFDP was greater than 4.25. The first cluster with the lesser AFDP showed lower nitrogen oxide emissions as well.



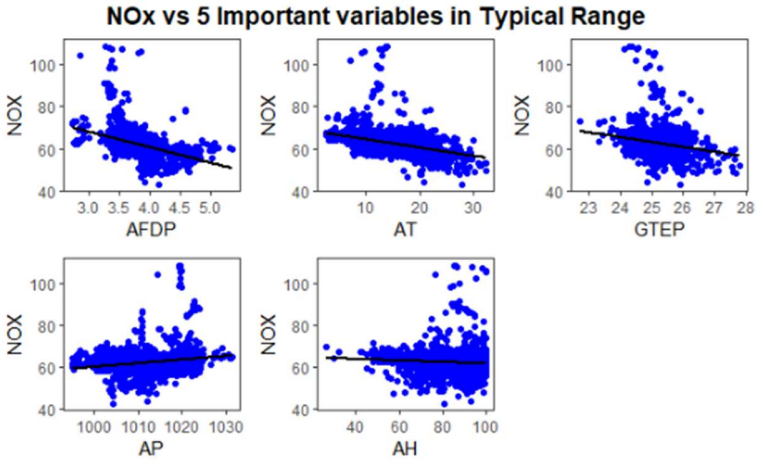


	AFDP	AT	GTEP	AP	AH	TEY	TIT	CDP	TAT
%IncMSE	24.92	19.81	9.69	8.82	6.56	6.32	3.16	2.65	0.38
Node Purity	12634.81	8597.78	3765.17	7155.25	4811.96	4482.14	3107.24	2277.10	1675.36

	AFDP	AT	TAT	AH	AP	TEY	GTEP	CDP	TIT
%IncMSE	182.16	12.98	8.28	7.81	7.04	3.27	2.86	1.94	0.06
Node Purity	13236.81	1580.02	1883.13	1346.26	638.38	269.36	497.39	292.61	55.06

IV. Conclusion and Suggestions

Based on our final model, random forest, we recommend the following to decrease NOx levels for the typical and extreme energy ranges:



For the typical energy range, we recommend increasing AFDP and GTEP. From a scientific perspective, we speculate that by increasing these variables, particularly AFDP, this would create a higher pressure drop, making the filter more restrictive to the air flow. In turn, this could lead to a decrease in NOx levels as a result (usaairfilter). In the case of the ambient predictors, our analysis shows that NOx is lower when the temperature is low, the humidity is low, and the pressure is high.

Regarding the extreme energy range, we could find two distinct clusters in the AFDP vs NOx. Even though we could observe slight tendencies where NOx emission decreases when AFDP increases, we recommend maintaining AFDP below 4.25 as the NOx emission drastically increases when AFDP increases over 4.25. From our random forest model, AFDP was by far the most important predictor, and its influence dwarfed those of the other variables. From our scientific research, we found that high energy leads to high temperature since the particles would have a higher kinetic energy.

As a result of this increase in temperature, it accelerates the rate of combustion leading to high flue gas temperature. This increase in flue gas temperature drives the chemical reaction towards the formation of nitrogen rather than nitrogen monoxide. As a result, this leads to a lower NOx level (Li et al., 2012).

V. Peer Review / Appendix and Code

* The appendix and corresponding code can be found at the end of the document s well as in the additional RMD file.

*Everyone contributed greatly in our project group. We all discussed and reached a consensus on what models and methods to try. We also worked as a group to code our models. design our graphics, and create our presentation. Everyone did great, and no one slacked off.

VI. References

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science

Hannah Ritchie and Max Roser (2017) - "Air Pollution". Published online at OurWorldInData.org. Retrieved from:'https://ourworldindata.org/air-pollution' [Online Resource]

Heysem Kaya, PÄ±nar TÃ¼fekci and ErdinÃ§ 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

Li, Zhengqi, and Yong Liu. “ACS Publications: Chemistry Journals, Books, and References Published ...” *Effect of the Air Temperature on Combustion Characteristics and NOx Emissions from a 0.5 MW Pulverized Coal-Fired Furnace with Deep AirStaging*, ACS Publications, 23 Mar. 2012, https://pubs.acs.org/doi/abs/10.1021/ef300233k.