CO2 Emissions Prediction Model

CO2 emissions are steadily rising in the World and are starting to drive climate change. We are starting to feel these effects through Fires, Hurricanes, and other climate impacts that are costing people their homes and making life more dangerous on planet earth. In order to gain a better understanding of the energy industry and how the different types of plants are contributing to climate change, I wanted to build a prediction model to see if I could predict how much CO2 and Methane were emitted given an amount of power created for coal and Natural Gas plants.

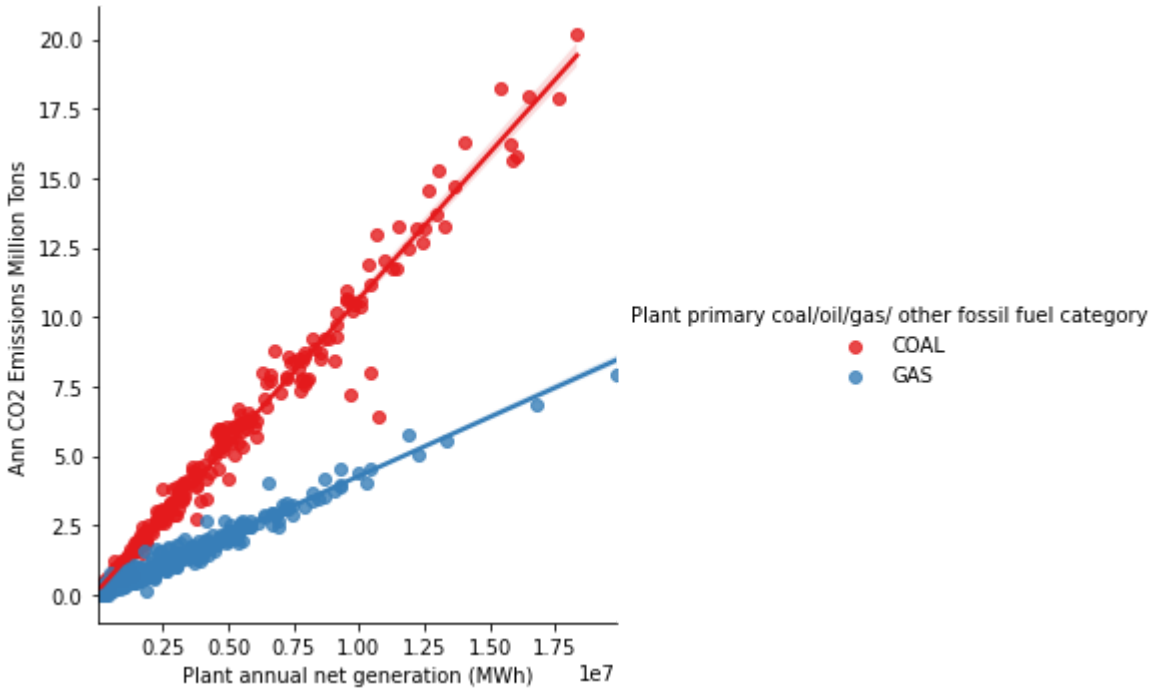
Approach

I decided that the best way to predict this was to use the data reported to the EPA in 2018 for all of the Energy Plants in the United States (https://www.epa.gov/ghgreporting/ghgrp-reported-data (https://www.epa.gov/ghgreporting/ghgrp-reported-data)). I decided to start with a linear model, then try to predict with a multiple regression model, and finally to use supervised learning models to see which of these models made the best prediction. I created three types of models: **Model 1:** The first model was simply to predict CO2 emissions given Megawatt Hour output of energy and type of plant. **Model 2:** This model was a multiple regression model that looked at emissions data from coal Plants and used the Type of Coal as a dummy variable to see if this could enhance the prediction. **Model 3:** This model took emissions data from both gas and coal Plants. Using this data, I built a model to see if I could predict whether or not the plant was a gas or coal plant. **Model 4:** Finally, built a number of supervised learning models to predict whether or not for coal plants if we could predict the type of coal by looking at the CO2 emissions.

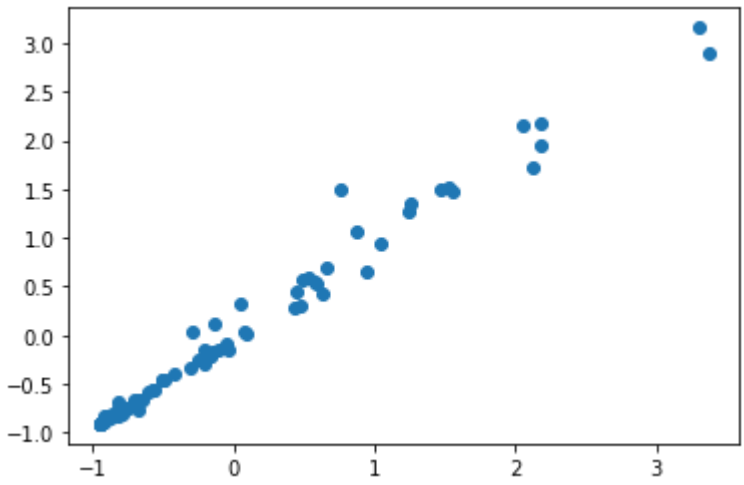
Findings

Simple Linear Regression

My first finding was that there is a direct, linear relationship between energy output in Mega-Watt Hours (MWh) for both Coal and Gas plants. As you can see in the chart below, there is significantly more CO2 output per MWh for Coal than there is for Gas. You can also see that the relationship between power output and pollution doesn't decrease with more or less output. This simple model gave an R-squared value of .982, which is very strong.



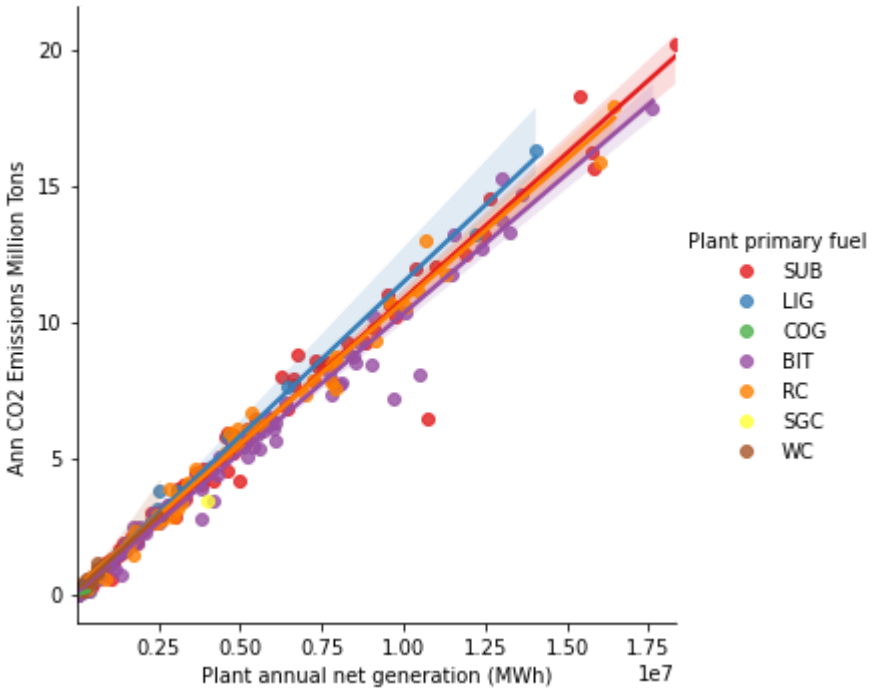
Looking at the first linear coal model between energy output and CO2 output, you can see that the prediction vs the test values are all closely correlated. A simple model comparing only these values gives the following scatter plot:



Multiple Regression with Consideration of Coal Fuel Source (Model 2)

The next iteration of the model that I did was to include the different fuel sources for coal and to assign them as dummy variables in the new multiple regression model to see if it could be improved. Additionally, I added a variable called Plant Annual Heat Input to see if I could get additional predictive power.

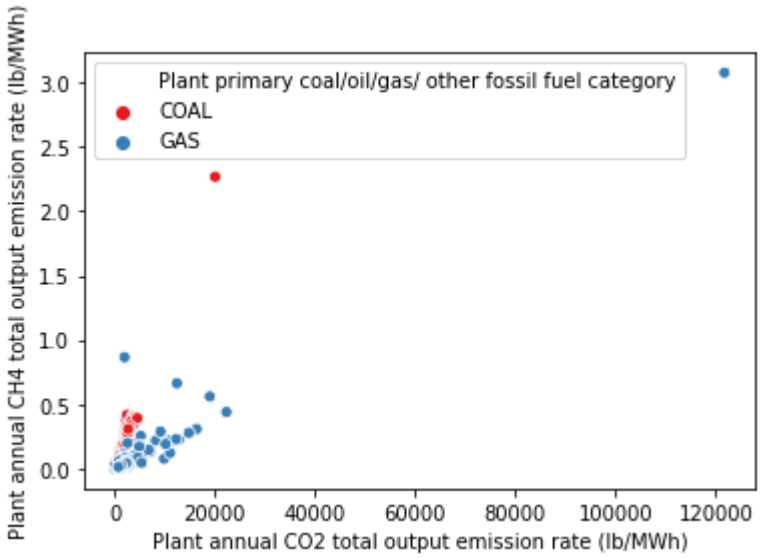
There were differences in the CO2 output based on the type of CO2 burne. You can see in the chart below a layout of the comparisons of CO2 output based on type of COAL Fuel source:



This model was stronger than the linear regression model, giving me a value with an R-squared of .997.

Model 3:

Model 3 is where I did the majority of my work on supervised learning models. The prediction that I made with these models was very basic. Can we accurately predict a gas vs coal plant by looking only at the methane and CO2 output per MWh? From the graph below, it looks like we should be able to predict this.



The graph below shows the types of models used and the accuracy scores of these models to solve this problem.

Model Type	Accuracy Score

Model Type	Accuracy Score
Logistic Regression Model	0.983
Random Forest	0.996
Decision Tree - Entropy Model - No Max Depth	0.984
Decision Tree - Max Depth 3	0.987
Gradient Boosting - Learning Rate .25	0.983
KNN Neighbors - 6 Neighbors	0.920

Model 4:

Model 4 was intended to see if we could predict the type of fuel source used in a Coal Plant based on the CO2 outputs. I used a random forest model to try to make this prediction, and I tried out different max depths of 1, 2 and 3 for the model.

Max_depths: 1 Accuracy score (training): 0.460 Accuracy score (validation): 0.391

Max_depths: 2 Accuracy score (training): 0.528 Accuracy score (validation): 0.422

Max_depths: 3 Accuracy score (training): 0.532 Accuracy score (validation): 0.406

Also using a confusion matrix and classification model, I was not able to yield strong results for this model after I saw these results, I abondoned this model.

Ideas for Further Research and Recommendations

Recommendations for Use of Analysis:

1. I believe that with my multiple regression model, we have a way to predict very precisely how much CO2 will be emitted into the environment given Plant Annual Heat Input, enenrgy output, and type of fuel used. This kind of prediction would be useful to determine whether or not differnt types of plants would be compliant with restraints on pollution in environmental regulations.
2. This can go the other way as well. If we could figure out a more accurate model that could be built to predict what kind of coal fuel was being used based on methane and carbon dioxide output, we would have a way to know whether or not the plant operators were being truthful about what kind of coal they were burning. There might be an incentive to use cheaper, dirtier coal for example for them to save money, but a good model could ensure that they don't do that.
3. Finally, this type of analysis could be used to compare the outputs of CO2 for gas vs coal and to determine which would be more profitable given future limitations on emissions. Since we can so accurately predict CO2 emissions for both gas and coal, we should be able to project these emissions out into the future and take them into account for the investment cost of a plant.

Ideas for Further Use:

There is a lot of potential use of these types of models if we start taxing carbon and methane to judge the profitability of these plants given assumptions on methane and carbon dioxide output per MWh of energy produced. As soon as the market starts to account for these costs to society, we will need to have accurate ways to incorporate the effect of these emissions ultimately on profitability.