# Inferring Relations between Air Particles and Other Air-based Pollutants

By Mark Collins

# Introduction

Pollution has become an increasing concern among many individuals around the world. A main concern that individuals have regarding pollution often tends to be air quality. In many instances having large amounts of airborne pollutants can cause short- and long-term side effects for people breathing the air containing them. To identify how clean the air in an area is, the amount of particulate matter is used to determined air purity. Having large amounts of particulate matter (abbreviated as PM) is often indicative of poor air quality. However, the most dangerous form of particulate matter is usually the smaller particles that can stay deposited within deeper parts of the lungs. To measure the size of particulate matter, it is often represented in two ways: PM2.5 (particulate matter that is 2.5 microns or less) and PM10 (particulate matter that is less than 10 microns). Understandably, not all air borne particulates will be pollutants. Some particulates may be natural, such as pollen, and cause only mild symptoms for some. Nevertheless, it would be useful to see how measurements such as PM2.5 (and by extension PM10) could relate to other airborne pollutants. It is for this reason that, using data collected over a period of 3 years among different locations, we would like to analyze the possible relationship of PM2.5 with various other pollutants. The pollutants we will be looking at will be carbon monoxide, sulfur dioxide, nitrogen dioxide, and ozone concentrations. Additionally, the cities will also be used as a feature for the project to see if location may affect the level of PM2.5.

# Models

To see how well the pollutants may relate to one another, it is best to see if reliable models can be made to predict these relations. To accomplish this, various machine learning models will be used to try to predict data and see how well these models’ predictions are compared to existing data. To start with, a linear regression model is made based on the existing data in hand. In most cases, real world data is very often linear. Therefore, a linear model is made to be used as a base line. If other models prove to have worse results than the linear model, this result may indicate that the model is either a poor fit for the data given or the relation of the features if very weak. The next chosen models to compare to the base-line linear models will be two tree-based models, and one neural model. The two tree-based models in question are the Bagging and Random Forest tree models. Bagging is a tree model that is often used to reduce variance within a noisy dataset [1]. Bagging takes random samples of data (with replacement) in a training set and generates a final model based on other weaker models. In Random Forest trees, the model is trained with the Bagging method along with feature randomness to create a final model based on many other weaker models [2]. Both work in similar ways, however, Random Forest trees include the randomization of features as well. Lastly, the neural model to be used is Multi-layer Perception (or MLP). Multi-layer perception consists of three layers: the input layer, hidden layer, and output layer. For the purposes of this project, MLP will use the ‘relu’ or (rectified linear unit function) and the ‘adam’ solver.

# Evaluation

To evaluate and process all information, the Python coding language was used to simplify the process. To start with, the data given (within an excel file) was imported into the project code. To do so, the pandas package was used to create a data frame consisting of the information from the excel worksheet using the read\_csv function provided. Next, the data was processed in a way so that it would be accepted by the models’ function when parameterized. The only processing needed to be done was for the “cityname” feature given by the excel sheet. All string data had to be converted over to numeric data corresponding to the values of each city name. Because all x values given to the model functions had to be numerical, “cityname” had to be converted to numerical data to be accepted. Next, the x and y values were assigned. To do this, x was assigned the values of the features we wished to use to predict PM2.5, and y would receive the values for PM2.5. The last thing required for processing was to create test and training data by using the test\_train\_split function given in sklearn’s model selection package using the newly created x and y as parameters.

Text

Description automatically generatedText

Description automatically generatedText

Description automatically generatedNow that data processing is complete, the models could now be generated and compared. To start, functions were created to generate the Accuracy, MAE, MSE, and RMSE of each model as well as generate the permutation feature importance of each feature. The following results (Figure 1) would show that the Random Forest tree method would prove to be the best predictor model compared to the Bagging Tree (Figure 2) and the Multi-Layer Perception neural model (Figure 3). All performed better than the Linear Regression model generated which had an accuracy score of only around .701 and far higher MAE, MSE, and RMSE values (Figure 4).

Figure 4 -- Linear Regression Data Eval

Figure 2 -- Bagging Data Eval

Figure 1-- Forest Data Eval

Text

Description automatically generated

Figure 3 -- MLP Data Eval

It is possible that MLP could prove to be more accurate than the other models if longer training was allowed. However, lengthening the allowed iterations for MLP could prove to take several hours. Despite this, it does seem that Random Forest and Bagging does show fairly promising results. Both have comparably low average error rates and contain an accuracy score of above .8. While accuracy score may not be that reliable for interpreting regression models, it does give some support that the models do have a fair amount of accuracy when it comes to predicting PM2.5 based on other pollutants and the corresponding city.

# Interpretability

Text

Description automatically generatedText

Description automatically generatedText

Description automatically generatedText

Description automatically generatedRegarding permutation feature importance, the following results were found (see Figures 5-9):

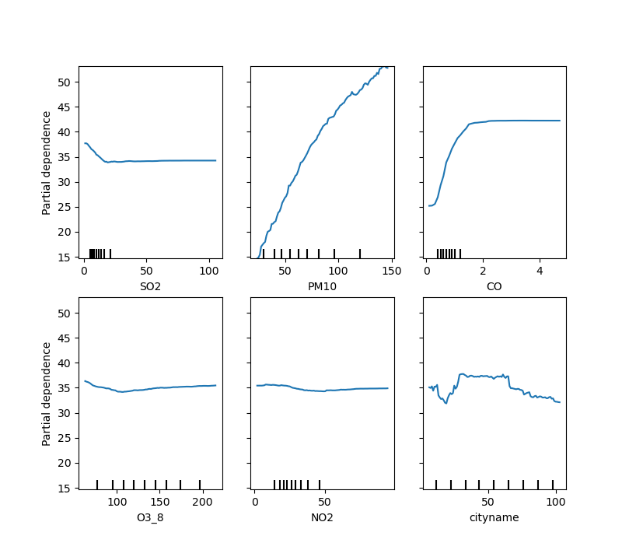
Figure 7 – MLP PFI

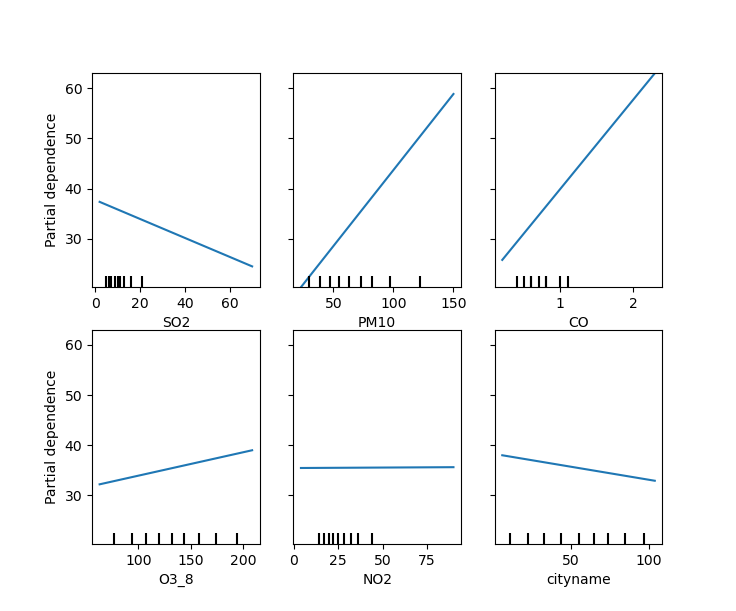
Figure 8 – Linear PFI

Figure 6 – Bagging PFI

Figure 5 – Forest PFI

As should be the case, PM10 proved to have the highest feature importance. If PM10 had not been on the top of the list, all models would have had to been reevaluated seeing as PM10 directly correlates with PM2.5. Furthermore, it can be inferred that the amount of carbon monoxide may have a relation with PM2.5 seeing as it has a fairly high importance value. The other values, however, are up for debate as most of them fluctuate in terms of importance based on what model is being used as a predictor. These trends will continue on to show within the partial dependence plots and Shapely graphs.

Among the partial dependence plots, PM10 and carbon monoxide would prove to be among the higher valued features as their presence in the air increases. PM10 will surely become more important as it increases but finding that carbon monoxide generates a higher dependence score as it increases does further support the findings of the permutation feature importance results. The other pollutants within the plot do not fluctuate, but rather have about the same amount of dependence toward PM2.5 no matter their value. The cities did have a fluctuating dependency depending on the city that was being selected at the time. This, however, is to be expected seeing as no cities will have the same amount of air particles present and the values given to the cities does not represent anything other than the city itself. A higher “cityname” value does not indicate that the city is “worth” more (See Figures 10-13).

Diagram, engineering drawing

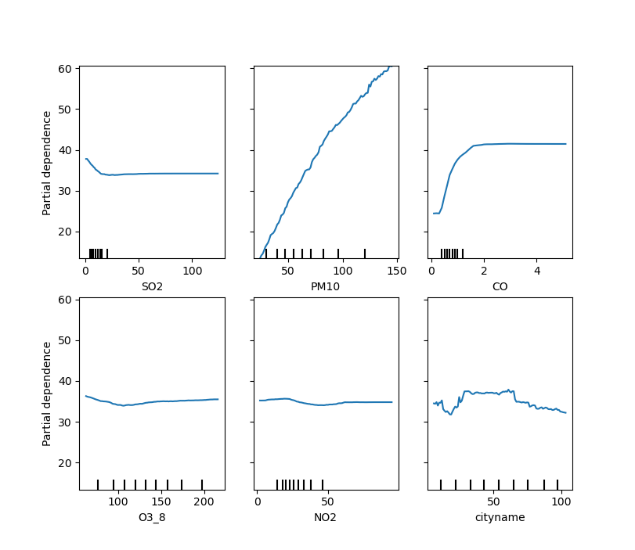
Description automatically generated

Figure 13 – Linear PDP

Figure 12 – MLP PDP

Figure 10 – Random Forest PDP

Figure 11 – Bagging PDP

In the Shapely graphs for PM10 and carbon monoxide, the higher the feature values the more impact the values would have on the model. Again, “cityname” wildly fluctuates in comparison to its values. In the Shapely graphs, ozone and nitrogen oxide would show to have very congregated shapely values regardless of its feature value indicating that their value has little affect to the amount of PM2.5. Sulfur dioxide’s high feature values do seem to trend toward a negatively shapely value. However, the impact is so slight that nothing can be said toward the pollutant with a high amount of certainty (See Figures 14 – 16). Regarding Bagging, no Shapely graph could be generated as the SHAP package does not support Bagging.

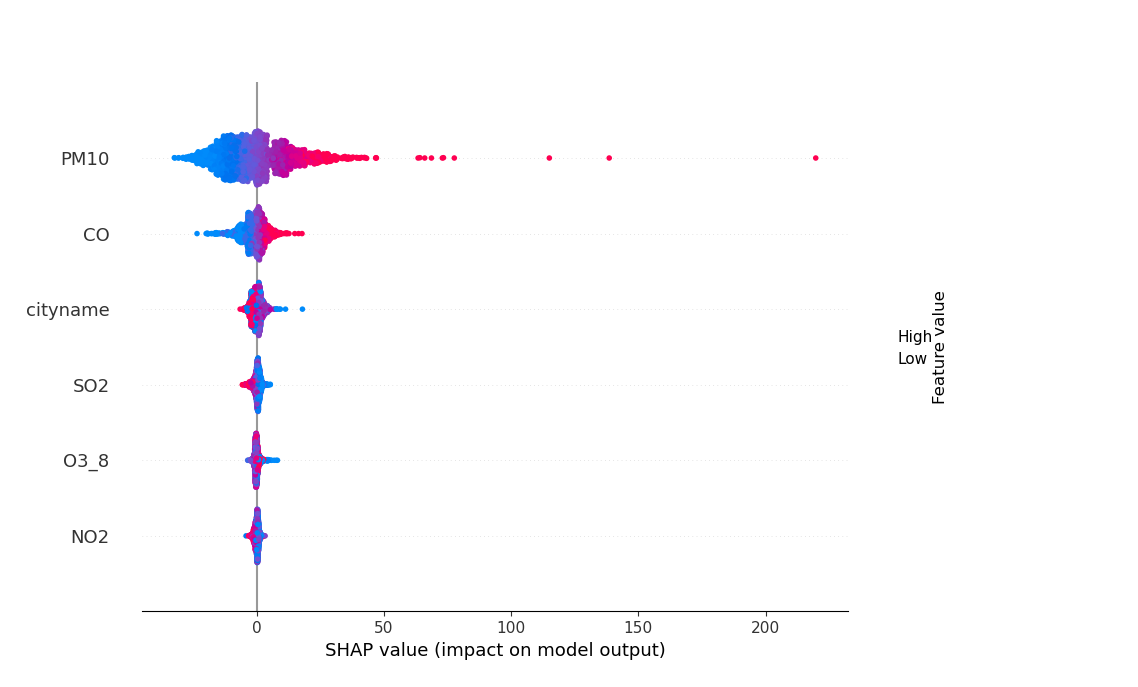


Figure 14 – Random Forest SHAP value

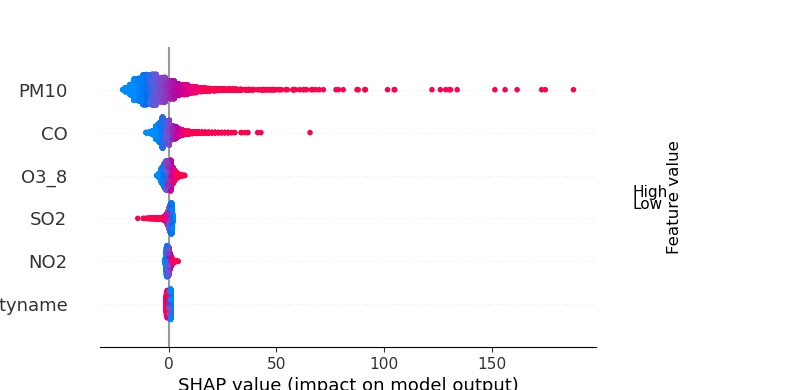


Figure 15 – MLP SHAP value

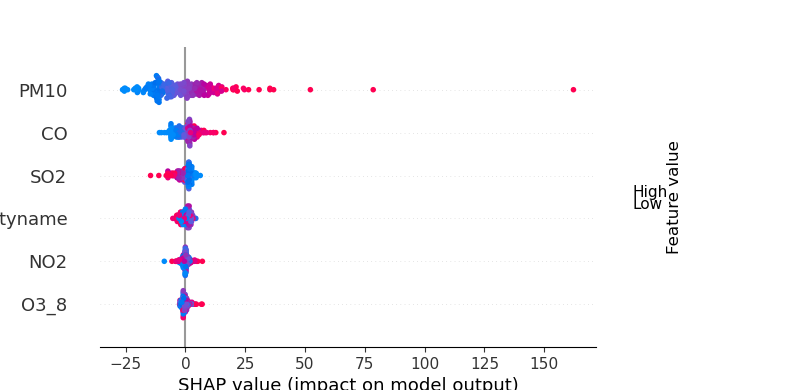


Figure 16 – Linear SHAP value

# Conclusion

Based on the models generated from the data given, it does seem that carbon monoxide does have some positive correlation with the amount of PM2.5 present in the air. Additionally, the location does seem to also have an effect on the amount of PM2.5. Although the methods used in the interpretability section did not support that the city has a substantial effect on PM2.5, this would most likely be because the value of the city is not inherently worth more other than an identifier. The reasoning behind this is because when locational attributes such as “cityname” or station are added to the models, the models in question do tend to perform better by around .05 in terms of accuracy score. Additionally, MAE, MSE, and RMSE do have lower values when these features are added. Other features such as humidity, temperature, and other features not included within this project may have some relation to the amount of PM2.5 present in the air. However, within the amount of allotted time given, it was not possible to generate models with all different feature combinations. Nevertheless, all models generated did show a fair amount of correlation between the features provided and the amount of PM2.5 as a whole. The linear model would prove to generate results that would be acceptable enough to suggest the relation on its own, and the other models would support that claim. However, the weight of the pollutants, excluding PM10 and carbon monoxide, is still in question and further research could be conducted to see exactly how much these pollutants affect the presence of PM2.5 if any. Given the amount of data within the excel format provided, with further research, a more accurate claim may be formed.

# References

[1] IBM Cloud Education, *What is Bagging?*, IBM, Dec. 2020. Accessed on: Apr. 2, 2022

[2] IBM Cloud Education, *What is Random Forest?*, IBM, Dec. 2020. Accessed on: Apr. 2, 2022