Ground Level Ozone

Philip Bradfield

February 2017

General Assembly

DC-DSI-3

Introduction

Ground level ozone is an atmospheric pollutant that is formed by chemical reactions between oxides of nitrogen (NOx) and volatile organic compounds (VOC) in the presence of sunlight. Emissions from industrial facilities and electric utilities, motor vehicle exhaust, gasoline vapors, and chemical solvents are some of the major sources of NOx and VOC. Chemically, ozone is a molecule made up of three oxygen atoms (Figure 1) and it is one of the major aspects of smog.

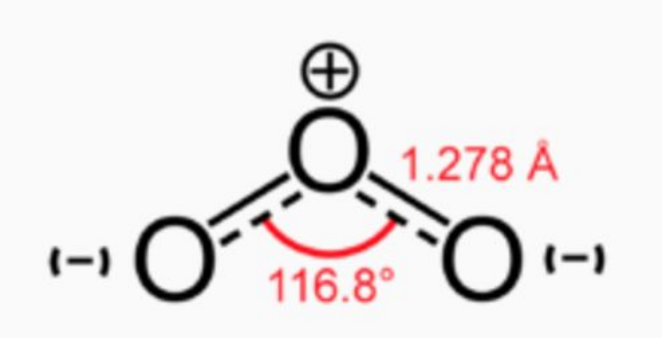

Figure 1. An ozone molecule - O_3

Why ozone is a problem.

Breathing ozone can trigger a variety of health problems, particularly for children, the elderly, and people of all ages who have lung diseases such as asthma. In addition, large long term studies have shown that ozone has significant effect on the mortality of those with diabetes and COPD.

Ozone is also a problem for crops because it interferes with photosynthesis and can causes billions of dollars worth of loss to the world wide agriculture industry. Further, although it has a shorter lifespan, it has heat trapping properties that contribute to regional warming.

The Data

The data for this project came from the EPA public website where, at least for now, anyone can download the data sets which are in a CSV format. The EPA collects pollution and weather related data from thousands of sites across the USA. If you wonder why you pay taxes, this is one of them. These monitors are used to provide information to state, federal and local officials about the health of the air we breath. The six variables needed for this project were ozone, VOC, NOX, temperature, sunlight and wind speed.

Although some the EPA data sets are labelled ‘daily’, it doesn’t mean that they are always collected on a daily basis. For example, in the Los Angeles area the ‘daily’ measurements of VOC are actually done every six days. Since VOC can linger I decided to try and find a better location where the measurements were done more frequently. The Houston area is also known for having high

ozone levels and I found that the data for a site named Channelview TX, a suburb of Houston, had data for 354 days from 2015.

Before presenting all the data from Channelview TX, I found it interesting to plot a histogram of all the ozone data from across the United States for 2015. As seen in Figure 1, that data very closely follows a classic bell-shaped (Gaussian or 'normal') distribution with only a very slight skewing. It is frequently said that many things in nature follow Gaussian distributions if enough data is taken and this is a clear example of that being true.

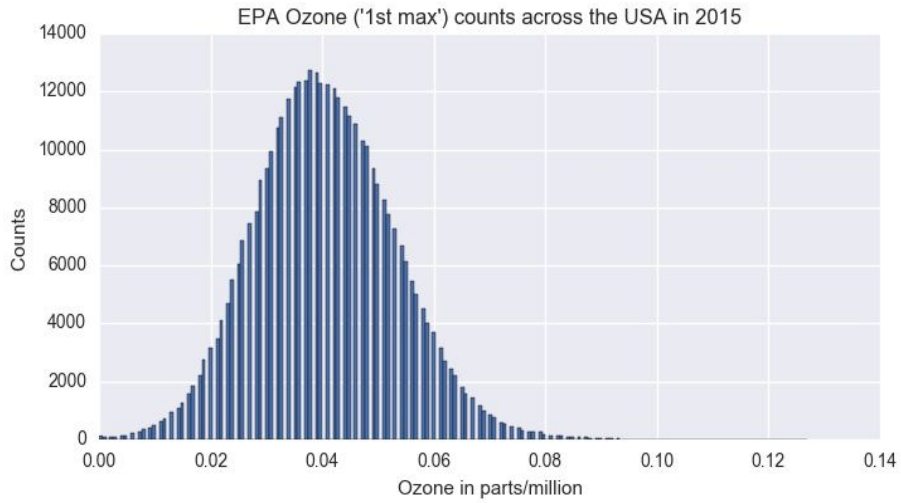

Figure 1. Ozone counts from all of the EPA sites across the USA for 2105.

Data Figure 2 is a histogram of ozone from Channelview TX and Figure 3 is a plot of the daily measurements of ozone for Channelview TX for 2015.

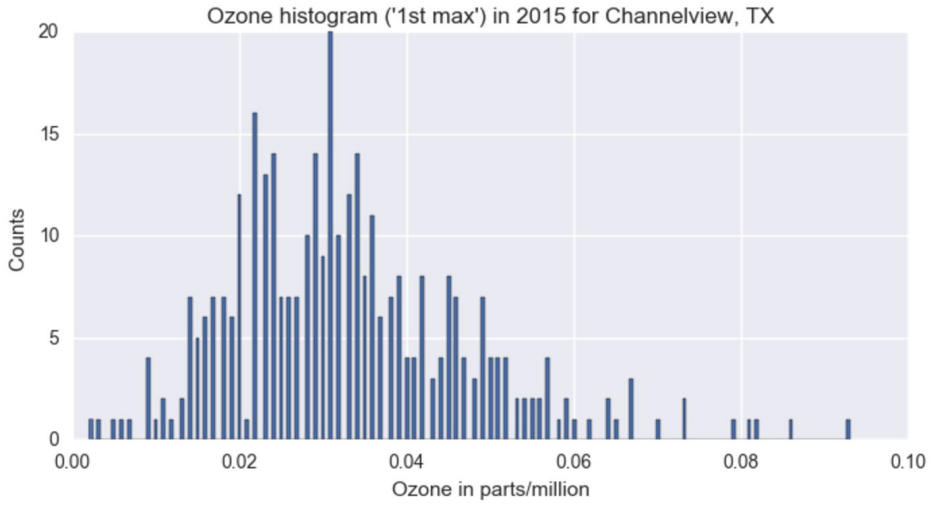

Figure 2 Histogram of levels of ozone for Channelview TX for 2015

While not as distinct, the outline of a bell-shaped curve can also be seen in Figure 2 so it is understandable that if you gathered data from hundreds of sites, a nice Gaussian would occur.

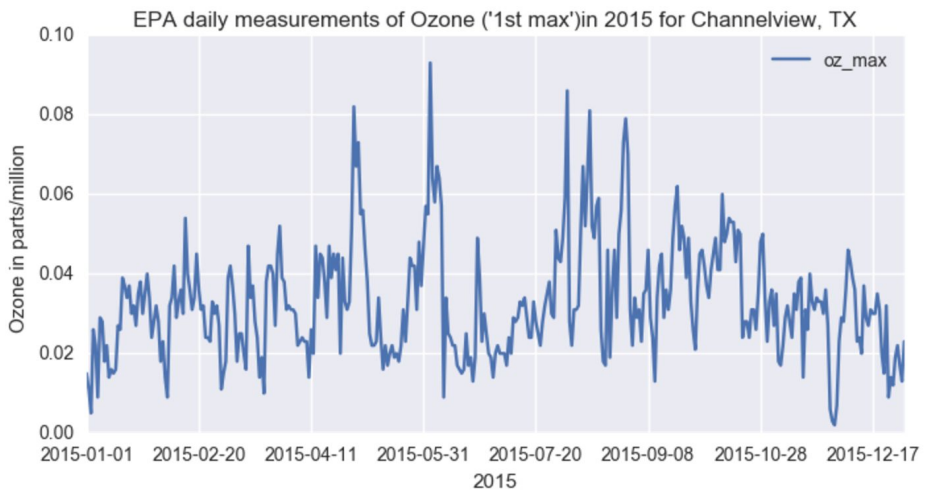

Figure 3. Daily ozone counts for Channelview TX for 2015

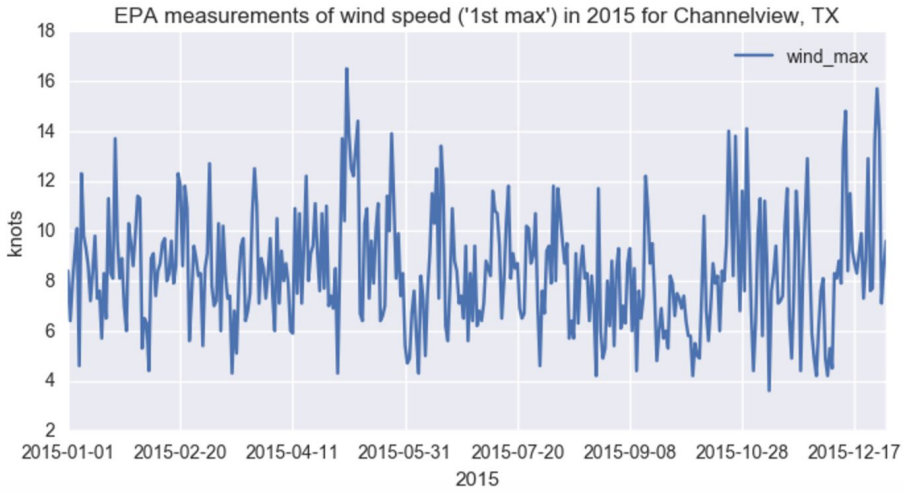

Figure 4. The wind speed for Channelview TX for 2105

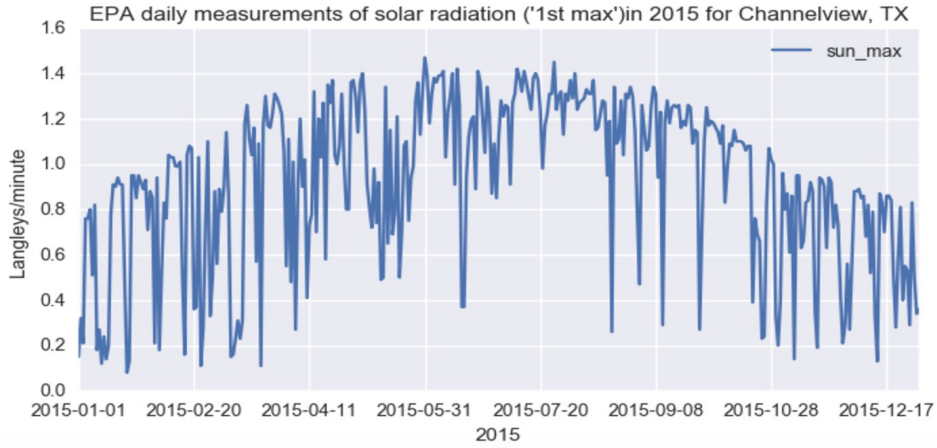

Figure 5. Daily measurements of sunlight for Channelview TX for 2015

Figure 4 shows the wind speed for Channelview TX for 2015 and Figure 5 shows the levels of sunlight for 2015.

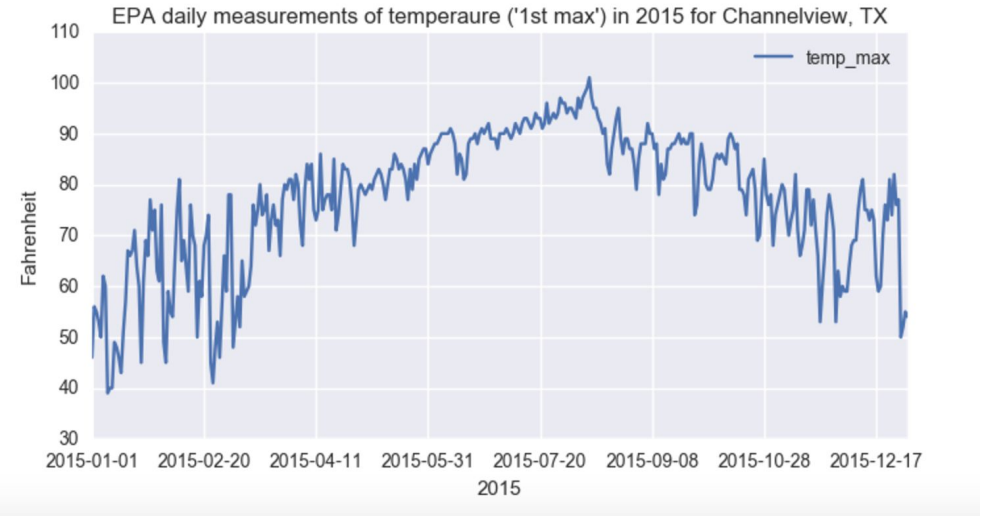

Figure 6 Daily temperature measurements for Channelview TX for 2015

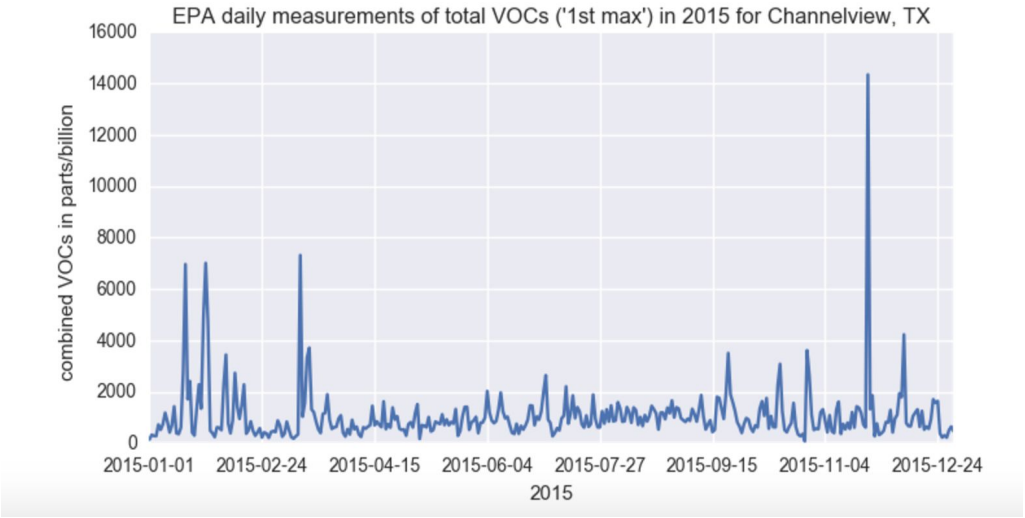

Figure 7 Daily VOC measurements for Channelview TX for 2015.

Each variable was part of a separate EPA data set. Once they were all pulled together into a single data set and the nulls were eliminated (giving 354 days out of 365) twelve different modelling approaches were used with ozone for each day being the target (typically referred to as “y”) and the other variables (for the same day) being used to build a model (“X”).

The approaches included -K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Logistic Regression, Decision Trees, Random Forest, ‘Extra Trees’. “Bagging” (bootstrap aggregating) was done on KNN, Decision Trees, Random Forest and Extra Trees and “AdaBoost” was done on

Decision Trees and Extra Trees. (See appendix for explanation of Bagging and AdaBoost.)

Results

One method of scoring how well a classification model is working is with a 'confusion matrix.' Simply put, a confusion matrix is a table that tells you how often the model was correct and how often it was wrong. Figure 8 (below) show a confusion matrix result from the Logistic Regression model where the total number of correct predictions was 102 (97 +5) out of 107 trials. This resulted in an accuracy score of 0.953 with 1.0 being a perfect score.

	predicted Below	predicted Above
Below Threshold	97	1
Above Threshold	4	5

Figure 8. Confusion matrix for Logistic Regression model.

Another method of scoring is with the area under the receiver operating characteristic (ROC) curve which is made up of ratios related to the number of correct predictions and the number of wrong predictions. This scoring approach is typically referred to as a ROC - AUC score (AUC standing for 'area under curve') As with accuracy score, the closer the ROC-AUC score is to 1.0, the better. Figure 9 shows the ROC-AUC for the Logistic Regression model with a score of 0.94.

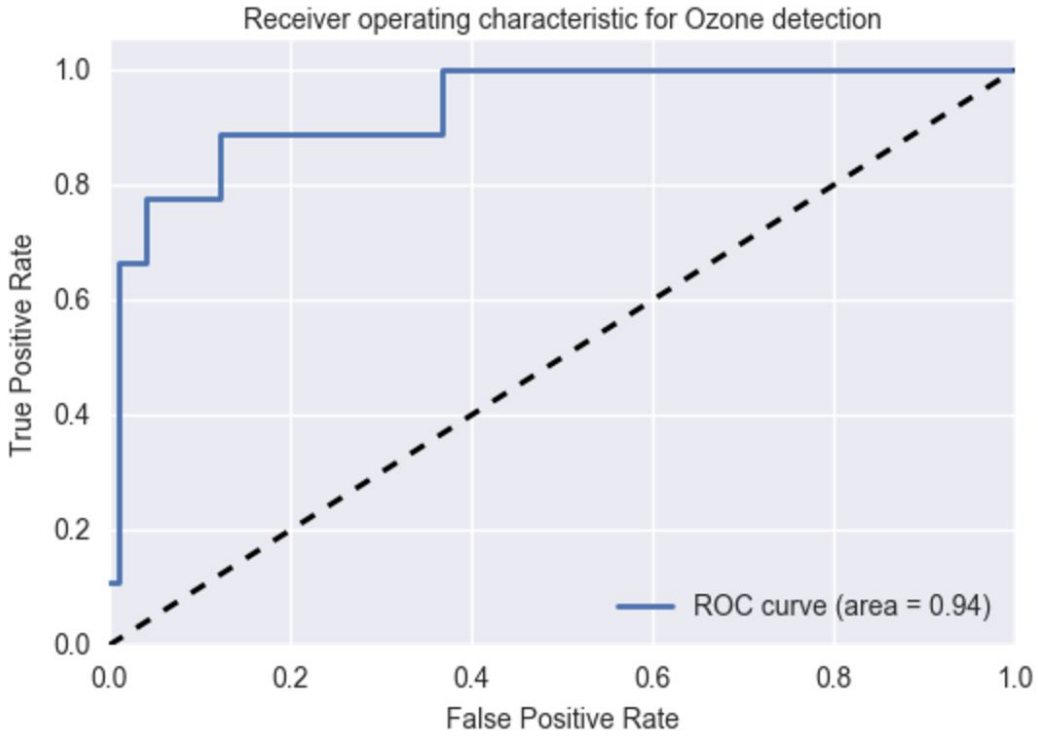

Figure 9. ROC-AUC plot for the Logistic Regression model.

The accuracy scores and area under the ROC curves for all the models are plotted in Figure 8 and Figure 9 respectively.

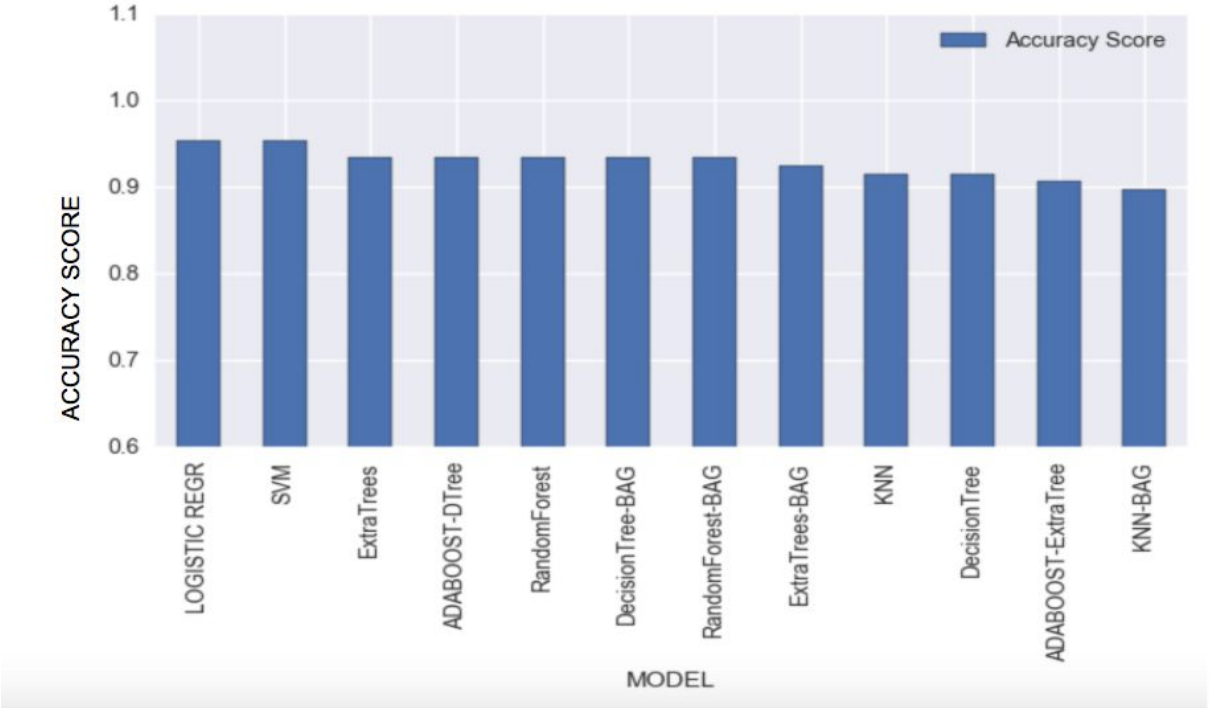

Figure 8. Accuracy Scores for the different modelling approaches.

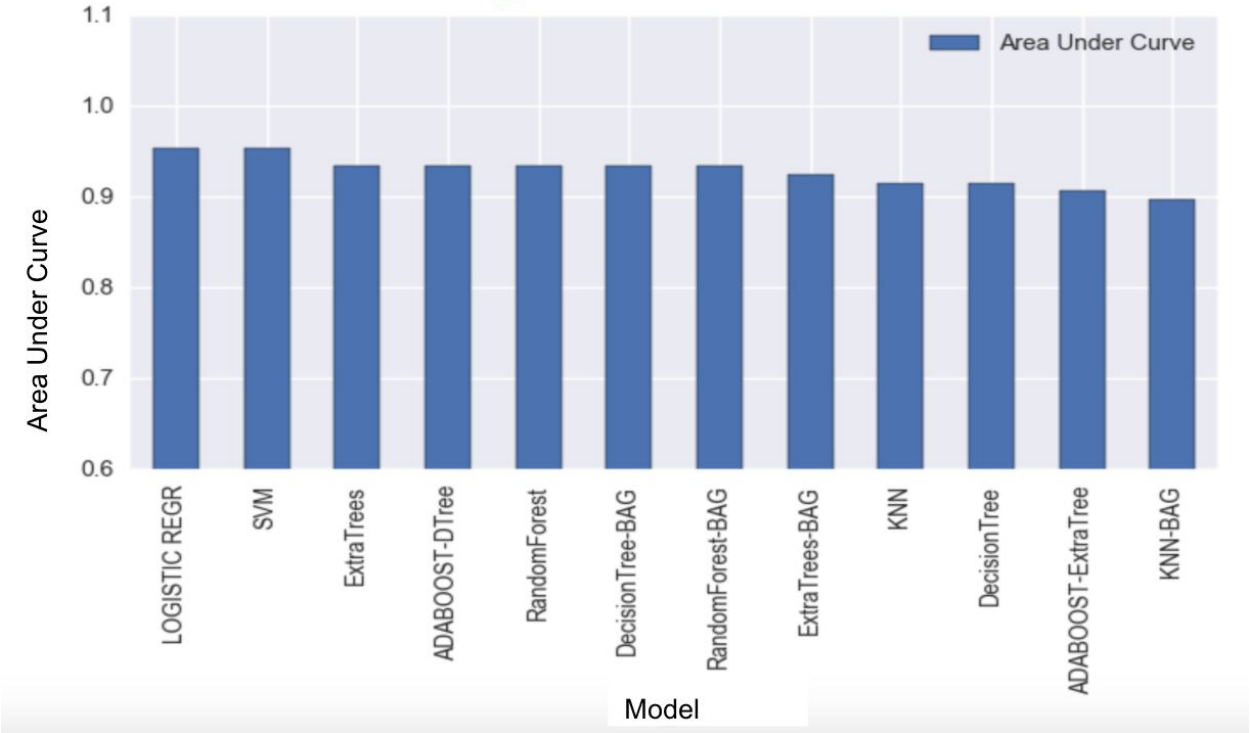

Figure 9. Area under the Curve for different modelling approaches.

Since sunlight is crucial for the formation of ozone from VOC and NOX it isn't surprising that the models indicate that it is most important input variable.

Future

One of the difficult aspect of this project was finding a location where all of the input variables were available for a large part of the year. I look at over a dozen cities and Channelview was the only one with more than 300 days of data for the year that had ozone levels that could be a serious health problem. The most disappointing city I looked at was Bakersfield CA because it has a serious ozone problem. The problem with Bakersfield was that half-way through the year, the sunlight data for it made an unnatural shift that indicated that they had been performing those measurements incorrectly. Welcome to data science! With more time I could fInd other cities with fairly complete data sets and seeing if the model can do as well or better.

Combining two or more different models and weighting them appropriately could also improve the scores. Adding in some other weather data like humidity might also improve the model.

With some modifications, this program could be set up to be used in a high school situation to bring in young students who have an interest in the environment and hopefully, data science. There is already a program in North Carolina where high school students make actual measurements on air pollution so that might be a good starting place for this idea.

Until other locations are tested, I remain a little skeptical of my results. The more research I read on this topic, the more I learned that people spend enormous amounts of money and time trying to build sophisticated weather modeling programs so they can better understand when, why and where pollution such as ozone will be a problem. Both the weather and the chemical reactions between VOC and NOX are not linear systems so my scientific background makes me a little skeptical my simple model can work fairly well. On the other hand, Alexander the Great's approach to the Gordian Knot was simple and effective. Maybe that is true here as well.

Appendix:

Types of decision tree classifiers:

Decision Tree classification - a machine learning approach that is not affected by scaling and various other transformations of feature values, is robust to inclusion of irrelevant features, and easily understood. Minus side - they can 'overfit' and can be susceptible to high variance

Random Forest - average over many decision trees with a random subset of the features - reduce the variance aspect with a small increase in bias. Improves performance though you have loss of interpretability

Extra Trees - at each candidate split in the learning process, a random subset of the features is used (Random Forest) AND for each feature under consideration, a random value is selected for the split.

'Bagging' - ('bootstrap aggregating') repeatedly (*B* times) selecting a random sample with replacement of the training set and fitting trees to these samples and aggregating the results.

The bootstrapping procedure leads to better model performance because it decreases the variance of the model, without increasing the bias.

AdaBoost -"Adaptive Boosting" AdaBoost is adaptive in the sense that the incorrectly classified instances from the initial fitting are focused with more weight on subsequent fittings.