Ground Level Ozone

Philip Bradfield
GA - DC-DSI-3
10 February 2017

What is ozone?

Agenda

What is ozone?

Why it matters.

Data

Modeling Approach

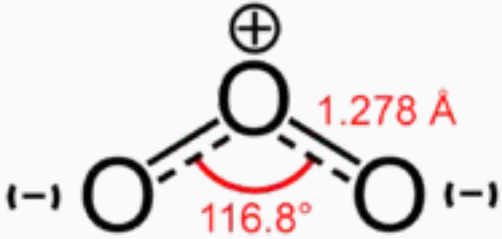
Modeling Results

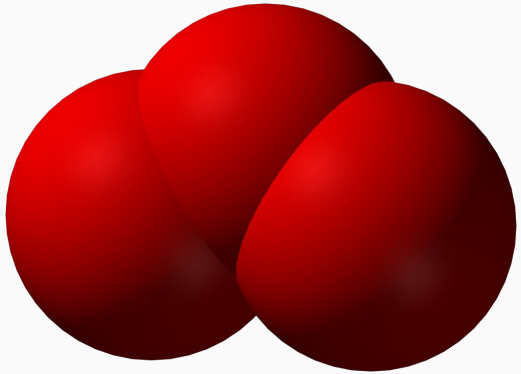
Discussion

Future

Concluding Remarks

A molecule made up of three oxygen atoms - O_3





What is ozone?

Agenda

What is ozone?

Why it matters.

Data

Modeling Approach

Modeling Results

Discussion

Future

Concluding Remarks

Ground level ozone is an atmospheric pollutant that is formed by 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.

What is ozone?

Agenda

What is ozone?

Why it matters.

Data

Modeling Approach

Modeling Results

Discussion

Future

Concluding Remarks

Ozone is one the of the major aspects of smog.

Why it matters.

Agenda

What is ozone?

Why it matters.

Data

Modeling Approach

Modeling Results

Discussion

Future

Concluding Remarks

Human Health

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. Large long term studies have shown that ozone has significant effect on the mortality of those with diabetes and COPD Am J Respir Crit Care Med. “Long-Term Ozone Exposure and Mortality in a Large Prospective Study.” MC Turner, et al.

Why it matters.

Agenda

What is ozone?

Why it matters.

Data

Modeling Approach

Modeling Results

Discussion

Future

Concluding Remarks

Agriculture

Ground level ozone can also has harmful effects on sensitive vegetation and ecosystems. Numerous studies indicate that the monetary value of crop loss due to ozone will be many billions of dollars in the coming years. [J. of Exp. Botany, 1 Jan. 2012, “How is ozone reducing our food supply?”, S. Willinson et al.]

Greenhouse Gas

Ozone effects are regionalized due to its shorter life span, but it can have a heat trapping effect greater than carbon dioxide.

Wikipedia - ‘Ozone’

Data Source: EPA

Agenda

What is ozone?

Why it matters.

Data

Modeling Approach

Modeling Results

Discussion

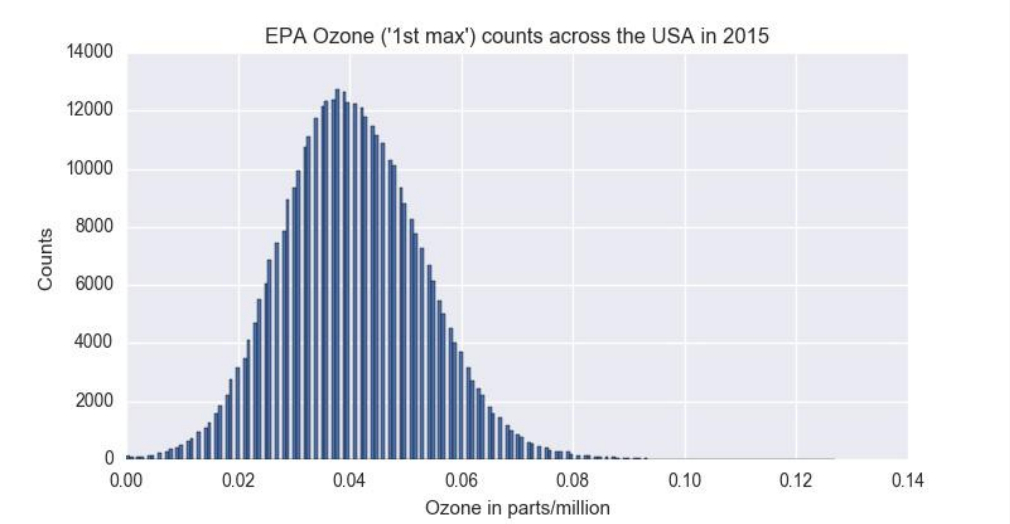
Future

Concluding Remarks

Open source data at

www.epa.gov (for now??)

Ozone, VOC, NOX,
Sunlight levels ,
Temperature, Wind speed.



Data - Ozone Channelview TX

Agenda

What is ozone?

Why it matters.

Data

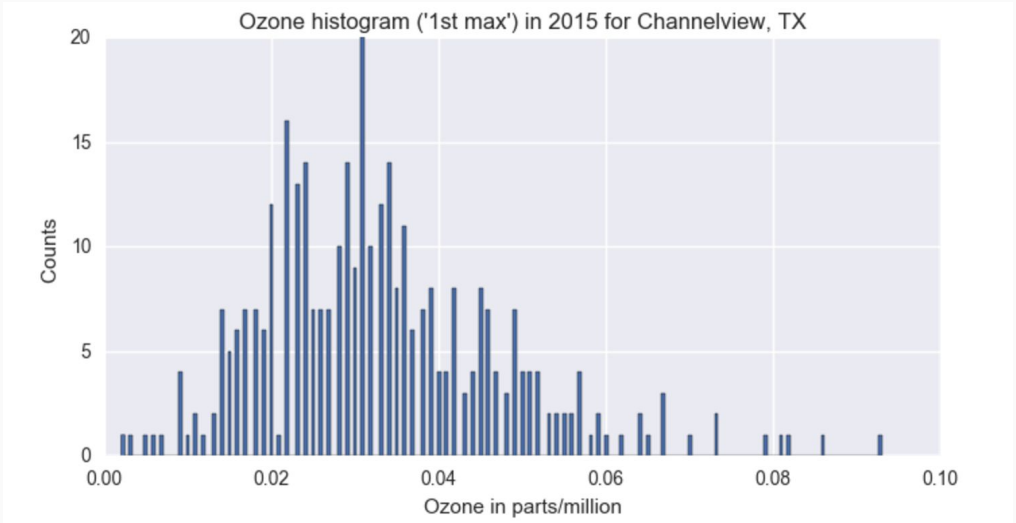
Modeling Approach

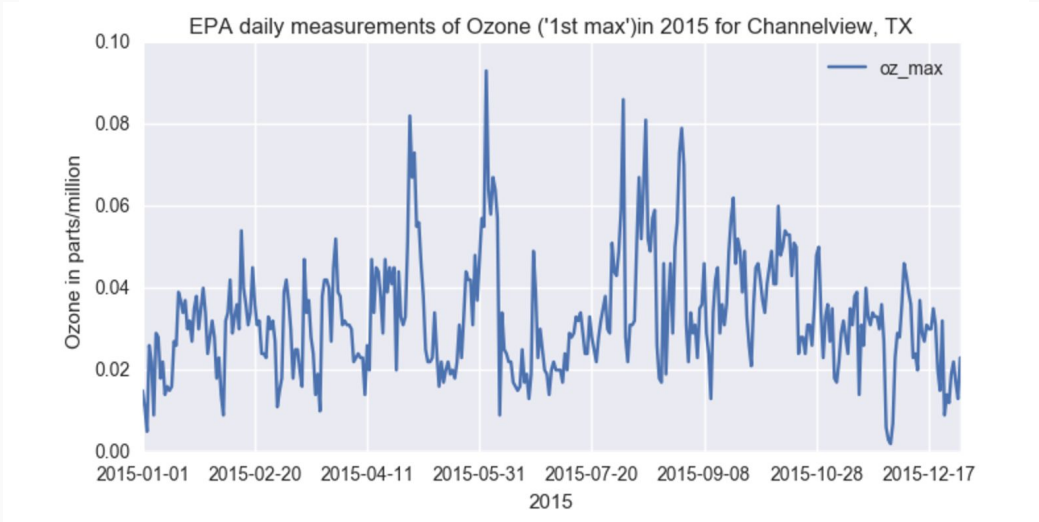
Modeling Results

Discussion

Future

Concluding Remarks





Data
Wind - Solar

Agenda

What is ozone?

Why it matters.

Data

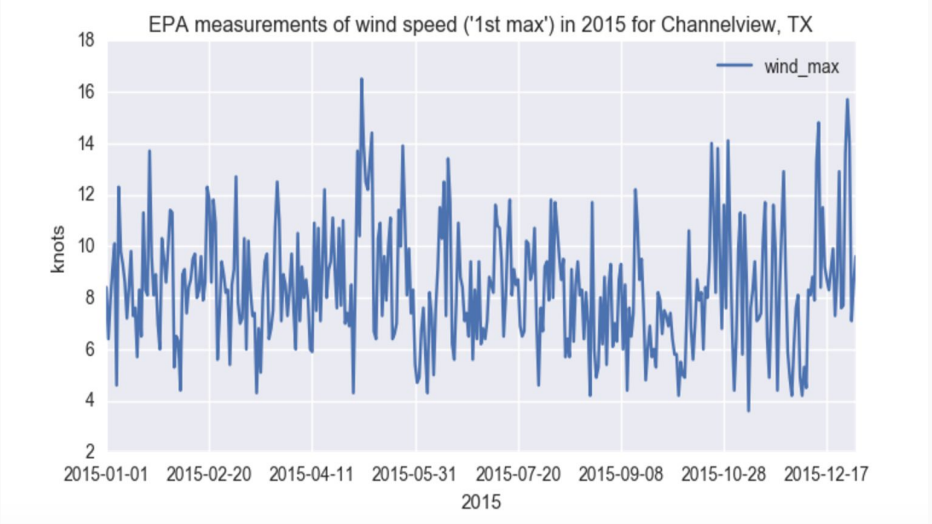
Modeling Approach

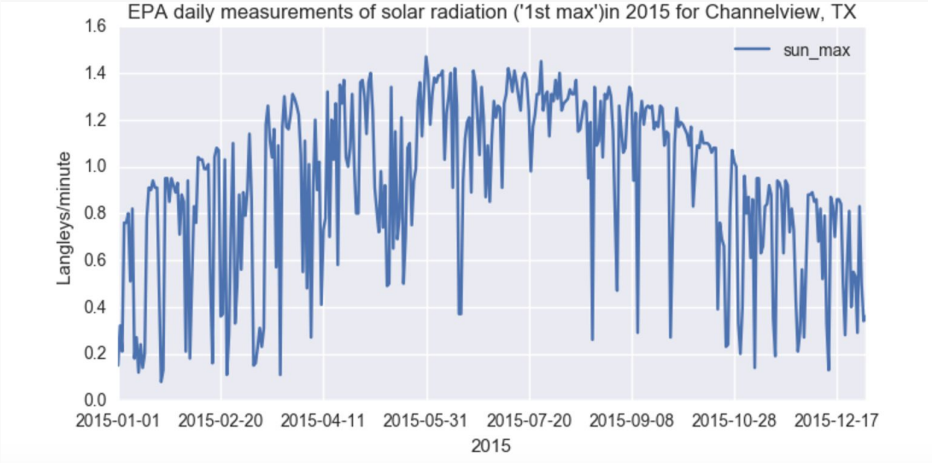
Modeling Results

Discussion

Future

Concluding Remarks





Data
Temp. & VOC

Agenda

What is ozone?

Why it matters.

Data

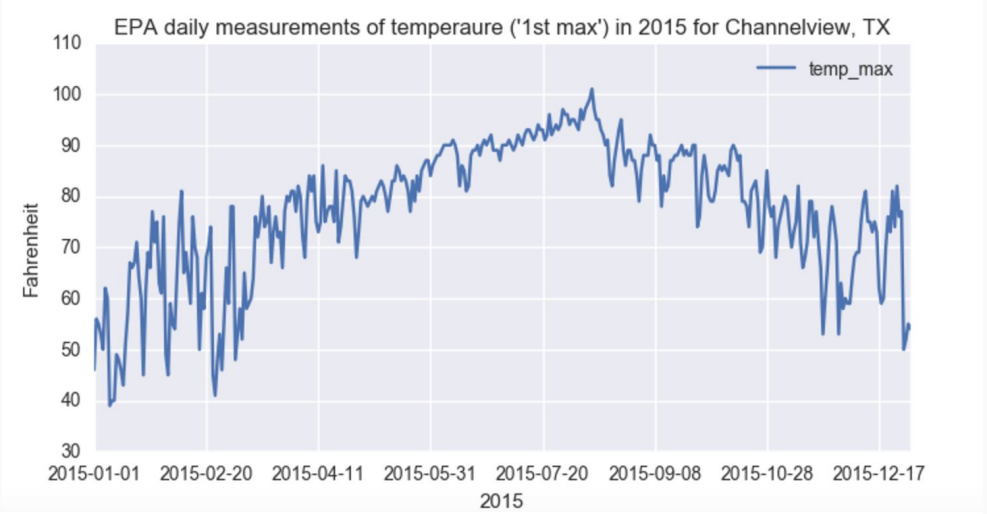
Modeling Approach

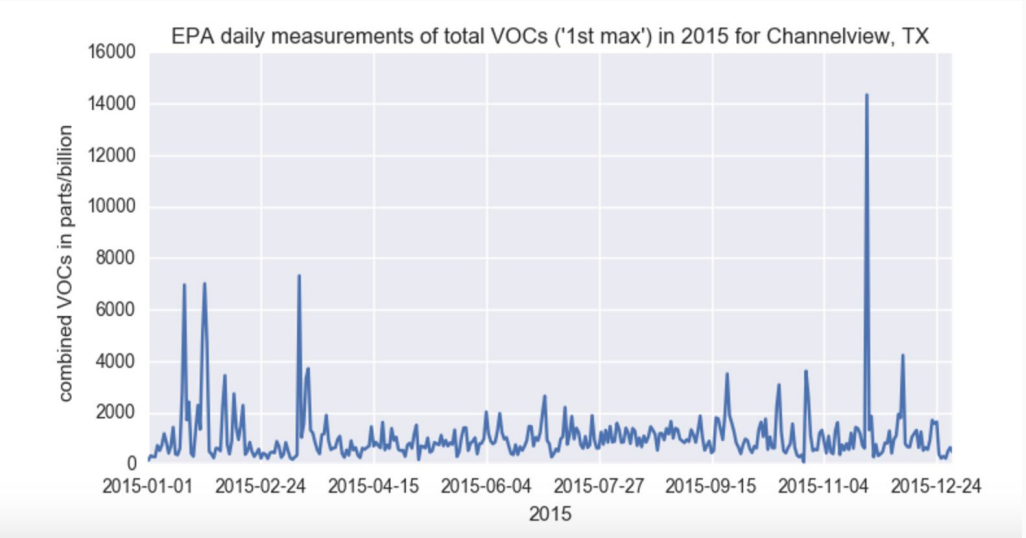
Modeling Results

Discussion

Future

Concluding Remarks





Modeling Approach

Investigated regressive and classification modeling.

A decision tree type classification model gave the best 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.

	predicted Below	**predicted Above**
Below Threshold	96	2
Above Threshold	5	4

Modeling Results

Agenda

What is ozone?

Why in matters.

Data

Modeling Approach

Modeling Results

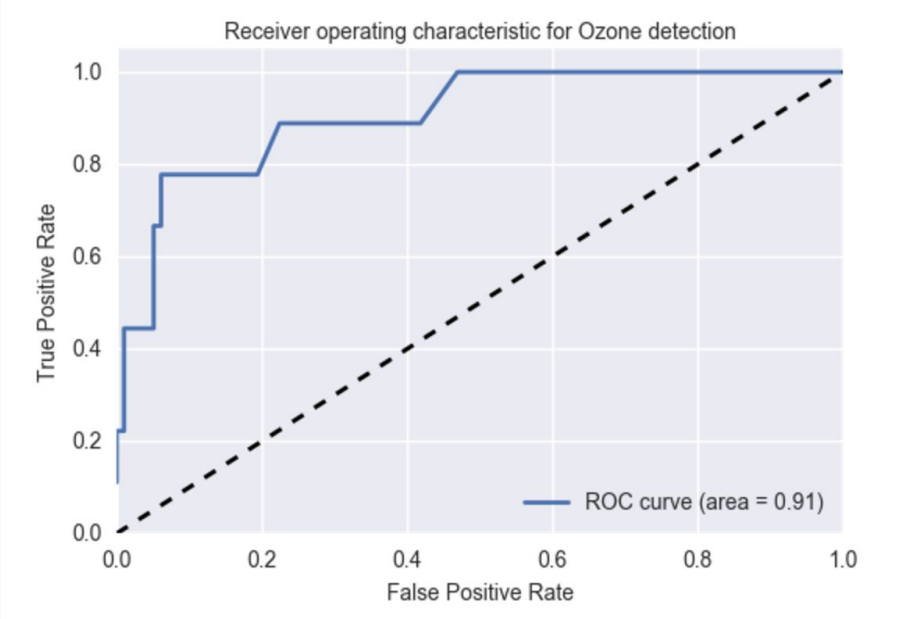
Discussion

Future

Concluding Remarks

ROC - AUC curve scoring of 0.91

(Receiver Operating Characteristic - Area Under Curve)



Modeling Results

A natural question for this type of modeling is which variable was the most important. Not surprisingly, the model indicated that sunlight was the most important or influential factor. That is not surprising since the it is a critical part of the formation of ozone.

According to the model, the second most influential input variable was the temperature.

Future

1. When complete data is available - model other locations.
2. Add other variables - humidity, barometric pressure?
3. Turn the program such into a teaching tool for high school students - environment, computers and (saving the best for last)...data science.
4. Weather modeling?

Concluding Remarks

Modeling something where weather is an important aspect of how it works has limitations.

People spend many years and millions of dollars building sophisticated computer models to predict weather patterns.

Additional research shows that the interaction that occurs to form ozone from VOC and NOX is not a simple one. A more sophisticated (non-weather modeling) model could be used - neural network?

ROC-AUC explanation

Agenda

What is Ozone?

Why it matters.

Data

Modeling Approach

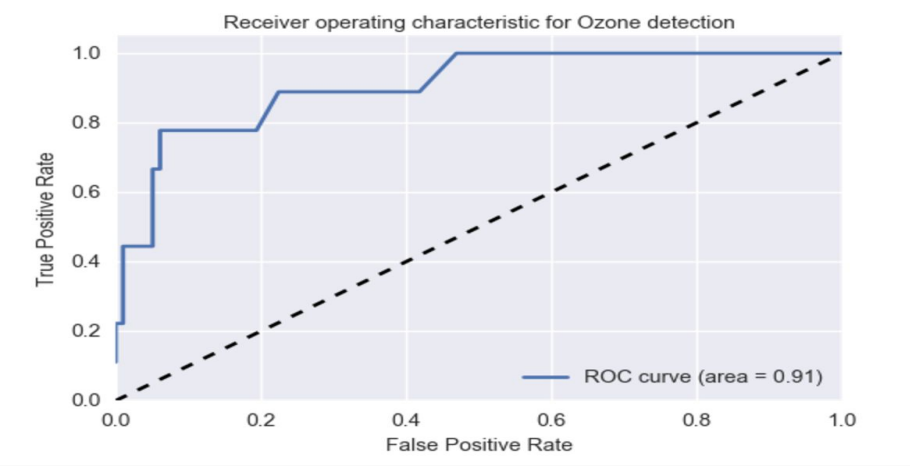
Modeling Results

Discussion

Future

Concluding Remarks

ROC - AUC curve explanation



$$\mathrm{TPR} = \frac{\mathrm{TP}}{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{N} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

“Extra Trees” explanation

Agenda

What is Ozone?

Why it matters.

Data

Modeling Approach

Modeling Results

Discussion

Future

Concluding Remarks

“Extra Trees - “Extreme Random Forest” - a type of decision tree classifier

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 trees with a random subset of the features - reduce the variance aspect with a small increase in bias. Improves performance though you have some loss of interpretability

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

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

‘Extreme’ - 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.