Predicting Low-Level Ozone

Robert Dusel

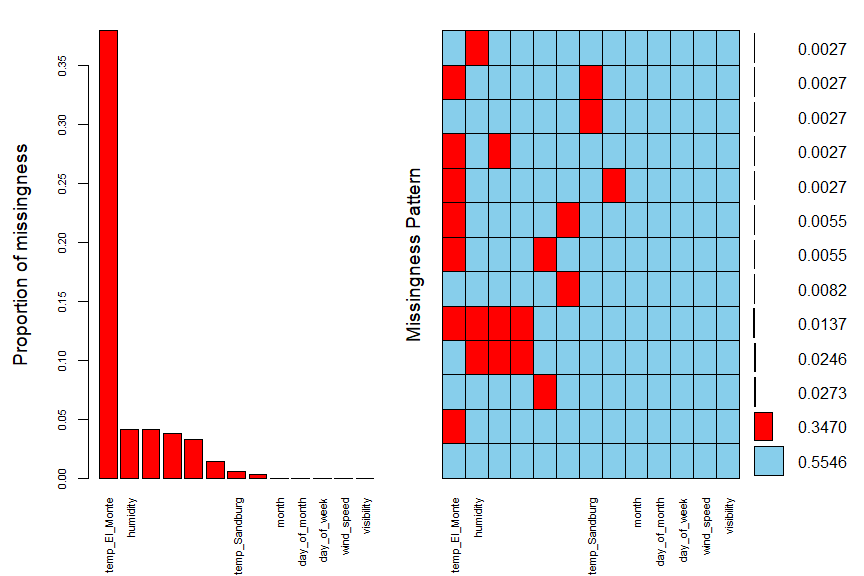
11/02/2021

## Introduction

Is the air I breath safe? This is a question that many people never ask themselves. Air quality is often taken for granted and its neglect can lead to serious health concerns. This analysis will be looking at the gas ozone. Most of the ozone in Earth’s atmosphere is located way up high in the stratosphere, where it serves an important role in blocking much of the Sun’s harmful UV radiation. Ozone has a darker side. It can also form in the lower parts of the atmosphere where people live. When people breath in this low-level ozone, it could cause health problems. Long-term exposure to low-level ozone is associated with an increased risk of mortality from respiratory causes [3]. Being able to predict low-level ozone levels and understanding the variables that lead to its formation can result in effective policy that controls or reduces it. This analysis will focus on predicting low-level ozone levels from a provided data set. It begins with cleaning the data, then follows up with model building. It ends with interpreting the model and assessing the full process.

## The Data and Cleaning

The data comes from a 1976 study by Breiman and Friedman (JASA, 1985, p. 580). It contains weather data from the Los Angeles area in California. The data set has 13 variables and 366 observations. The response variable, hour\_average\_max, is the maximum one hour average of ozone level measured in ppm. This analysis begins with cleaning. There are 203 missing values. 68% of these missing values fall into one variable, temp\_El\_Monte, which is a temperature recording of the city of El Monte. These graphs breakdown the missing values:



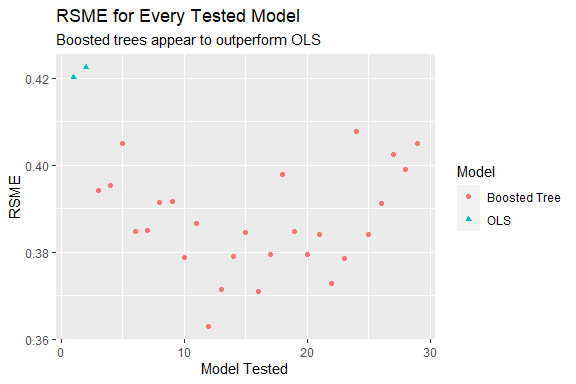
The graph on the left shows that almost 40% of the temp\_El\_Monte variable is missing! It turns out that this variable is highly correlated with the other temperature variables. It can be safely dropped from the data set. There are a few more observations that need to be removed too. There are five observations missing a response value. These will be dropped. With the more problematic missing values taken care of, the rest of the missing values will be filled in using KNN imputation. With the missing values handled, feature engineering is up next. There are three variables related to time/day. Two will be dropped. The variable that will remain will be what month the observation happened in. There is another variable that could be a problem, inversion\_base\_temp. It has a VIF of over 17 and will be dropped from the data set. With the predictors handled, it is time to look at the response. The response, hour\_average\_max, is highly skewed to the right. A log transformation will be applied to it to make it more normal. This concludes cleaning!

## Model Building and Assesment

This analysis will focus on predicting the response using ordinary least squares regression and boosted trees. The model forms that will be tested are

The first model, with four predictors, was chosen through the use of the regsubsets function. This function finds the best model, in respects to linear regression, with ‘m’ predictors for values of ‘m’ from 1 to the total number of predictors. In other words, that is the best 4 predictor model, and its adjusted R^2 is close to the full model.

Other than the number of predictors, there are still more tuning parameters to consider for the boosted tree training method. This analysis focuses on adjusting the bag fraction, learning rate, and maximum depth of the trees. With the use of the caret function, optimal tuning parameters can be chosen. Here are the results from using the caret function to run a 10-fold cross-validation on the data set.

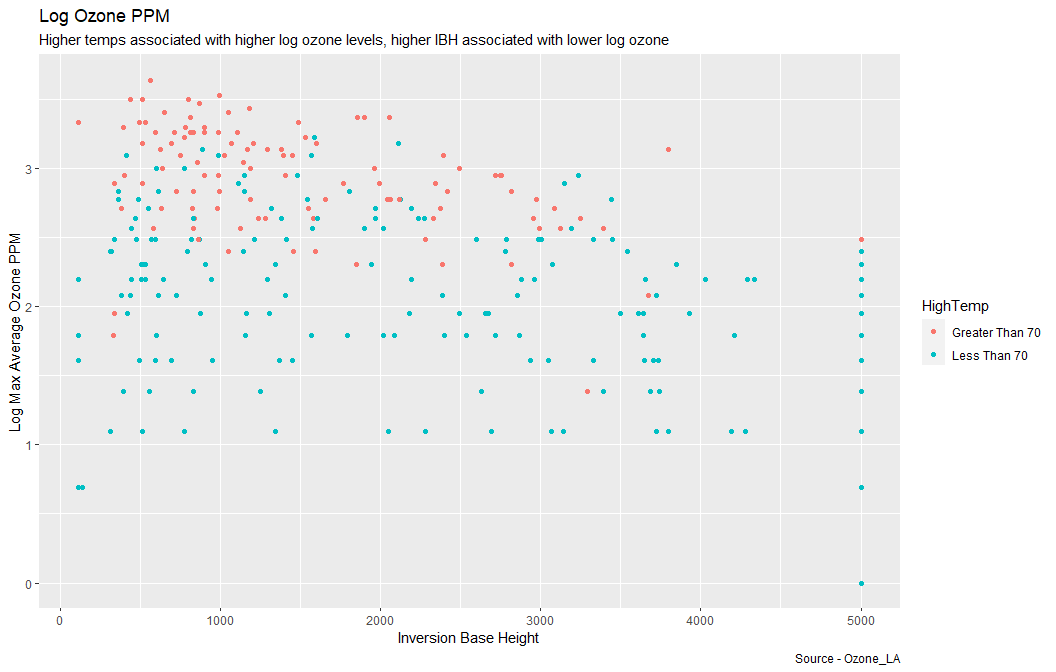


It is clear from the graph that the boosted tree training method is outperforming OLS in terms of lower RSME. The best boosted tree model had an RSME of 0.3629, R2 of 0.7792, max\_depth of 2, bag fraction of 0.5, and a learning rate of 0.1.

Before locking in our best model choice, assessment of the entire modeling process has to happen. This done with an outer five fold cross-validation loop. This process reveals how much variability in the response can be accounted for through this modeling process. Boosted trees performed the best in terms of RMSE on every fold, although the parameters for the best model changed each time. The outer validation process resulted in an R2 of 0.7420. About 74% of the variability seen in log Ozone is accounted for in this modeling process. With the modeling process accounted for, the best model can be chosen. The best model is a boosted tree with a max\_depth of 2, bag fraction of 0.5, and a learning rate of 0.1.

## Model Interpretation

According to the model, the two most important predictors in this data set for log ozone are temp\_Sandurb and inversion\_base\_height. This makes a lot of sense. Let’s look at a graph of the relationships.



The graph reveals the interesting associations between the two most important predictors and the response. Higher temperatures are associated with higher ozone levels and a higher base inversion height is associated with lower ozone levels. Ozone forms when pollutants emitted by cars, power plants, and other sources react to sunlight [1]. With our data set, the variable temp\_Sandburg may function as a proxy for the amount of sunlight. Hot sunny days probably have more overall sunlight. So what is ‘inversion base height’ (IBH)? So usually, temperature decreases as you ascend in height. This norm can invert and the temperature can start to increase with height [4]. IBH tells you as what height this inversion starts to happen. Hot air can trap ozone underneath it. This explains why when the IBH is lower, its associated with higher levels of ozone.

## Conclusion

The analysis of this data set, through the use of a boosted tree, found that the strongest predictors for ozone, are temperature and inversion base height. This modeling process explains about %74 of the variability seen in log ozone levels. There is room for improvement in the model and an inclusion of other predictors (like air pollution concentrations) can probably increase accuracy.

Sources

[1] <https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics>

[2] <https://www.sciencedirect.com/science/article/abs/pii/S1364815206002076>

[3] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4105969/>

[4] <https://w1.weather.gov/glossary/index.php?word=inversion>