Justin Carter

3/24/22

Midterm Project: Executive Summary

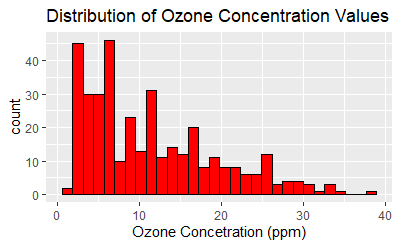
**Executive Summary**

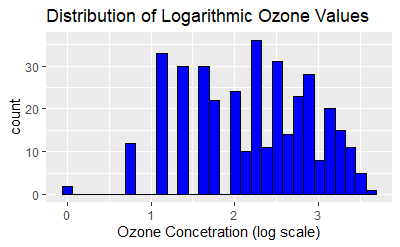
Currently, there is much work being done in order to better understand the world around us. Millions of sensors have been applied in a myriad of venues in order to collect meaningful and impactful data about our surroundings. One such venue for this collection is Earth's atmosphere. As phenomena such as climate change continue to occur, it becomes increasingly important to understand which factors affect important characteristics of the atmosphere, like its concentration of ozone. In furtherance of this pursuit, atmospheric data were analyzed to determine which variables were most important in predicting ozone concentration.

**Data Exploration and Transformation**

Variables in the data set were checked for multicolinearity after missing data from the variables had been imputed via the median value of each column. Variables that had correlations of an absolute value of .7 or higher with other predictor variables were removed from consideration in terms of creating models. Then, the distributions of the variables were observed.

Upon further analysis of the distributions of the quantitative variables, it was determined that several of them had irregular distributions. Since the Month and day\_of\_month variables had low correlations with hour\_average\_max (the hourly average maximum value of ozone concentration in parts per million), and had highly irregular distributions, they were dropped from the data set. In addition, the day\_of\_week categorical variable was dropped from the data set due to causing errors in the cross-validation process. hour\_average\_max and visibility were observed to have irregular distributions, and were transformed. visibility (visibility in miles) was replaced with the square root of visibility, while hour\_average\_max was replaced with the logarithmic transformation of that variable. The original values for those variables were dropped from the data set. Below, the results of the logarithmic transformation of hour\_average\_max (our response variable) can be seen:





**Fitting the Models, 10 Fold Cross-Validation, and "Double" Cross-Validation**

A linear regression model, as well as a random forest model were fit on the transformed data. Then, a 10 fold cross-validation process was conducted on both models in order to determine their approximate error in predicting new data. During the 10 fold cross-validation process, the random forest model was tuned with the number of variables randomly sampled as candidates at each split. These values were set to a vector of 2 through 7. Through the cross-validation process, it was determined that the best model between the two categories in predicting values of ozone concentration was the random forest model, with a split tuning parameter of 3. This model was determined to be the best model considered due to having the lowest CV(10) value.

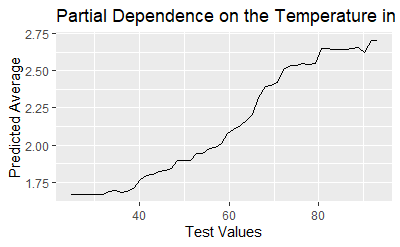
**Key Variables and Model Interpretation**

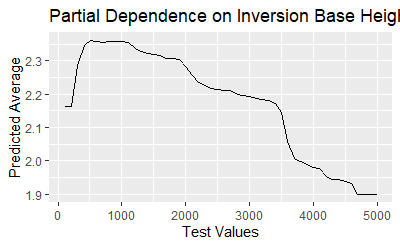
After the random forest predictive model was finalized, two variables emerged as the most important: the Temperature in Sandburg and the inversion base height (in feet). While details of these variables are sparse in the source of this data, there are still some generalizations that one can make regarding these findings. Primarily, the higher the value of the temperature in Sandburg, the higher the predicted value of maximum 1-hour average of ozone (measured in parts per million). This could be the case due to the phenomenon of particles in the air rising when exposed to increased temperatures. Although Sandburg was not the location in which these measurements were taken, it is within 100 miles of where the measurements were taken (Upland, CA). This means that it is likely that temperatures for these areas would be highly-correlated.

Inversion base height was the second-most important variable in the random forest model. This is the height in which warmed air from below begins to become cold. The relationship between inversion base height and ozone becomes somewhat less intuitive here. Initially, as inversion base height values increase, so do the predicted values of ozone concentration. However, this trend reverses quickly and becomes a highly negative relationship. This may be because the cold air is acting in such a way as to trap warm air (and thus areas of concentrated ozone) below it at a certain threshold.

The trend between inversion base height and ozone concentration initially surprised me, but upon further thought, it makes sense that there would be, after an initial uptick, a negative relationship between the height in which air becomes cold and concentration of ozone (for reasons mentioned previously). Plots of these observations can be seen below:

Double cross-validation was conducted in order to assess how much of the variability of ozone concentration is explained by this model-fitting process. The result of this process was a value of about 0.940. This is to say, about 94% of the variability of ozone concentration values is explained by this model-fitting process. The RMSE of the model selection process was determined to be about 0.399. Based on this low RMSE, this model is more than sufficiently accurate, and could be used for predicting new data points.





**Conclusion**

There is little doubt that humans have much more to learn about the world around them. Part of this learning process will involve an increased understanding of how Earth's atmosphere behaves. In furthering this process of learning, an analysis was conducted in order to find a model that has high accuracy in predicting values of atmospheric concentrations of ozone. This process yielded a model that has a low error rate in predicting these values, with much of the predictive power coming from variables concerning inversion base height, and the temperature of Sandburg, California.