Guanyu Huo, Duru Kahyaoglu, Andrea Lopez, Daniel Pak, Naishadh Parmar
Applied Machine Learning Group 24

Climate change has been a grave concern in modern times. One major contributor towards climate change is air pollution. One of the indicators of air pollution is the amount of ozone in the atmosphere. While ozone concentration in the upper levels of the atmosphere is beneficial, increased ozone concentration in lower levels, due to the emission of various chemicals and increased heat, can severely damage our respiratory system. If we can forecast ozone levels, we can guide public policy well and curb air pollution levels that would in turn address one of the causes of climate change. In this project, we will examine the factors that increase ozone concentration and develop a model to predict the amount of ozone in the atmosphere given the relevant environmental conditions.

Our dataset comes from the U.S. Environmental Protection Agency (EPA), which has the purpose of analyzing and enacting environmental regulations. We will be using the Clean Air Status and Trends Network (CASTNET) dataset. CASTNET provides trends in pollutant concentrations, acidic deposition, and air pollution impacts to ecosystems..

From CASTNET, we will use 3 datasets: site information where the measurements are taken (e.g. latitude, longitude, and elevation), meteorological conditions (e.g. temperature, humidity or windspeed), and the hourly gas concentrations (e.g. ammonia or ozone). These datasets can be joined with one another to create one big table to create our feature matrix and label vector. In total, we will have 55 features, and 1 label (ozone concentration). CASTNET provides csv files of these datasets going back to 1987, however we will be looking at data from the last decade (2013 to 2022). Using this dataset, we will frame finding the concentration of ozone as a regression problem.

To solve this problem, we will be using the supervised learning framework that we learned in this course. We will first perform exploratory data analysis to see the space and distribution of our features and label. Then we will preprocess the data by encoding it. Next we will split the data into training, validation and testing sets based on the year the measurements were taken. Specifically, our training set will have data from 2013 to 2018, our validation set will have data from 2019 to 2020, and our testing set will have data from 2021 to 2022. The data will then be scaled based on our training set. Our initial approach involves constructing a multiple linear regression model utilizing carefully chosen independent variables from our exploratory data analysis. If the linear regression model proves to be an inadequate fit for the data, we will explore regularization techniques such as ridge regression and lasso regression for more generalizability.

We will also look into different tree based models like regression trees, Random Forest, and XGBoost to try to improve performance. Finally, we will explore using deep learning methods like neural networks. Throughout all model development, we will also tune the available hyperparameters and use cross validation to ensure our model will not overfit on the training data.