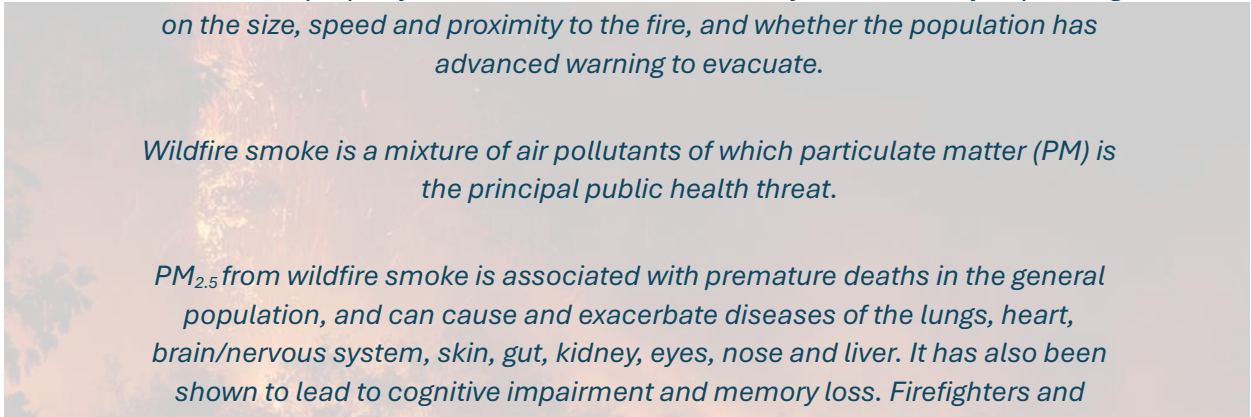


# Final Report: Predicting air quality due to wildfire smoke

## Problem Statement

---



*“Wildfires that burn near populated areas can have significant impact on the environment, property, livestock and human mortality and morbidity depending on the size, speed and proximity to the fire, and whether the population has advanced warning to evacuate.*

*Wildfire smoke is a mixture of air pollutants of which particulate matter (PM) is the principal public health threat.*

*PM<sub>2.5</sub> from wildfire smoke is associated with premature deaths in the general population, and can cause and exacerbate diseases of the lungs, heart, brain/nervous system, skin, gut, kidney, eyes, nose and liver. It has also been shown to lead to cognitive impairment and memory loss. Firefighters and emergency response workers are also greatly impacted by injuries, burns and smoke inhalation, particularly at high concentrations.”<sup>1</sup>*

---

This project was inspired by the orange sky event in NYC on June 6, 2023. The goal of this project is to predict the longevity of PM<sub>2.5</sub>, particulate matter most associated with wildfires (with diameters 2.5 micrometers or smaller), in order to inform the public of how long the air will be unsafe to breathe and when it will be safe to go outside again. Predicting large scale wildfires is not within the scope of this project. Instead, we will look at air quality data and weather data to predict the longevity of PM<sub>2.5</sub>. The problem is conceived as a time series regression analysis, where the explanatory variables come from data from weather stations and air quality monitors and the response variable is air quality measured in PM<sub>2.5</sub>. We will look at the ample data provided for California wildfires and focus on the San Francisco metropolitan area from 2018 to 2019.

---

<sup>1</sup> [https://www.who.int/health-topics/wildfires#tab=tab\\_2](https://www.who.int/health-topics/wildfires#tab=tab_2)

## Datasets

- 'Daily\_Census\_Tract-Level\_PM2.5\_Concentrations\_\_2018\_-\_2019CA.csv':  
[https://data.cdc.gov/browse/select\\_dataset?tags=pm2.5](https://data.cdc.gov/browse/select_dataset?tags=pm2.5)  
The Centers for Disease Control and Prevention is the national public health agency of the United States. The following CDC dataset contains the target feature PM2.5 concentration ('DS\_PM\_pred') which indicates fine particulate matter most associated to wildfires (2.5 micrometers and smaller). According to the WHO and EPA, the recommended short-term (24-hour) PM 2.5 level is 15 µg/m3.
  - 'year', 'date', 'latitude', 'longitude': Self explanatory
  - 'statefips', 'countyfips', 'ctfips': Data naming California
  - 'DS\_PM\_pred': Mean estimated 24-hour average PM2.5 concentration in µg/m3
  - 'DS\_PM\_stdd': Standard error of the estimated PM2.5 concentration
- 'asosCA.csv': <https://mesonet.agron.iastate.edu/ASOS/>  
The Automated Surface Observing System (ASOS) is considered to be the flagship automated observing network. Located at airports, the ASOS stations provide essential observations for the National Weather Service (NWS), the Federal Aviation Administration (FAA), and the Department of Defense (DOD). The primary function of the ASOS stations are to take minute-by-minute observations and generate basic weather reports. The following ASOS dataset located in the IA State Department of Agronomy dataset website is restricted to California stations.
  - 'station': three or four character site identifier
  - 'valid': timestamp of the observation
  - 'lon','lat': longitude and latitude
  - 'tmpf': Air Temperature in Fahrenheit, typically @ 2 meters
  - 'relh': Relative Humidity in %
  - 'drct': Wind Direction in degrees from true north
  - 'sped': Wind Speed in mph

## Data Wrangling

This section details the problems discovered in the datasets above and the steps needed to merge the datasets. The dataset 'Daily\_Census\_Tract-Level\_PM2.5\_Concentrations\_\_2018\_-\_2019CA.csv' becomes 'df1' and 'asosCA.csv' becomes 'df2'.

1. We drop all values not associated to the San Francisco metropolitan area: Before downloading the datasets, we can use the API's to restrict to California, and then to San Francisco. Afterwards, we can drop any extra identifiers associated to San Francisco in 'df1'.
2. Longitude and Latitude are swapped in 'df1': Reassignment is easy.
3. 6% of data in 'df2' is null: Because the data in 'df2' is hourly, we deem it acceptable to drop that.

4. The data between 'df1' and 'df2' must align along the longitude and latitude, but 'df1' are air monitors and 'df2' are weather stations. We must find a way to align each weather station with its closest air quality for that day.
  - a. We convert both data frames into `GeoDataFrames` and take advantage of `geopandas` to deal with geodata.
  - b. We associate each CDC air quality monitor with a unique ASOS weather station. To do this, we convert the degrees of longitude and latitude to meters and find the closest weather station.
  - c. We perform an inner merge along the weather stations and dates.
5. Finally, we need the index to be timestamps and proper datetime format.

## Exploratory Data Analysis

The newly merged dataframe now must be analyzed for important features and experimented on how it can be implemented into a regression analysis framework. Here are some quick facts about this dataframe before proceeding:

- 4,165,056 rows × 7 columns. Before restricting to the dataset to San Francisco, there were over 100 million rows.
- 221 unique locations
- 8 features, 1 label

The first analysis we will look at is a heatmap of the correlation table of features.

- 'DS\_PM\_pred' is our target variable. "date" data is captured in "valid".
- 'DS\_PM\_std': Standard error of the estimated PM2.5 concentration is not a physical metric, rather highly derivative of "DS\_PM\_pred". (Note the  $R^2$  of an OLS would be .9 if 'DS\_PM\_std' were included, which is too dramatic.) Thus, it should be excluded.
- There is higher anticorrelation with wind speed 'sped' and wind direction 'drct' than the time of the year in 'valid' and relative humidity 'relh'. This makes sense as we are not concerned with air quality in general, but the type of particle that is most associated with wildfires which can be blown in. Temperature 'tmpf' does not vary in extremes in the Bay Area where San Francisco is, so it makes sense that Summer (measured in 'valid') more than temperature impacts wildfires from further out in California.
- Latitude is more correlated than longitude by a (very) small order of magnitude. Some visual analysis of isolated stations displayed "DS\_PM\_pred" as generally similar throughout all of the stations around San Francisco. However, latitude may place the station either in a geographical valley or closer to a wildfire. Given the mountains to the East, longitude may play less of a role as the air that blows in is affected.

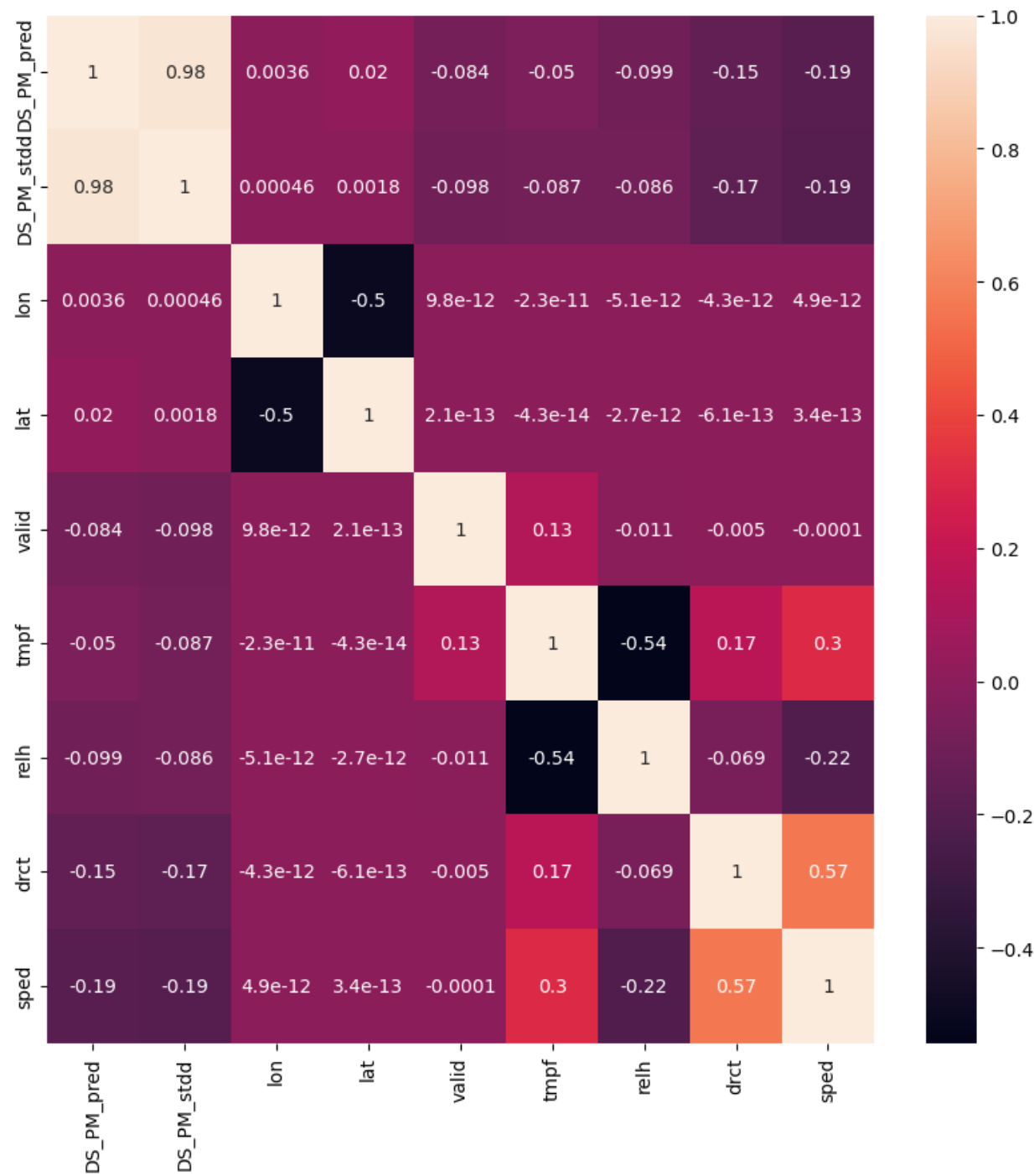


Figure 1: Heat Map of Correlation Table

The next analysis comes from visualizing the cross-validation of a time-series. We can only split the data and not shuffle because we do not want future data to leak into past validation data. However, we shall note a large peak in PM2.5 that is not captured in every cross-validation split (the colors below represent each split and fewer colors last for the entire span).

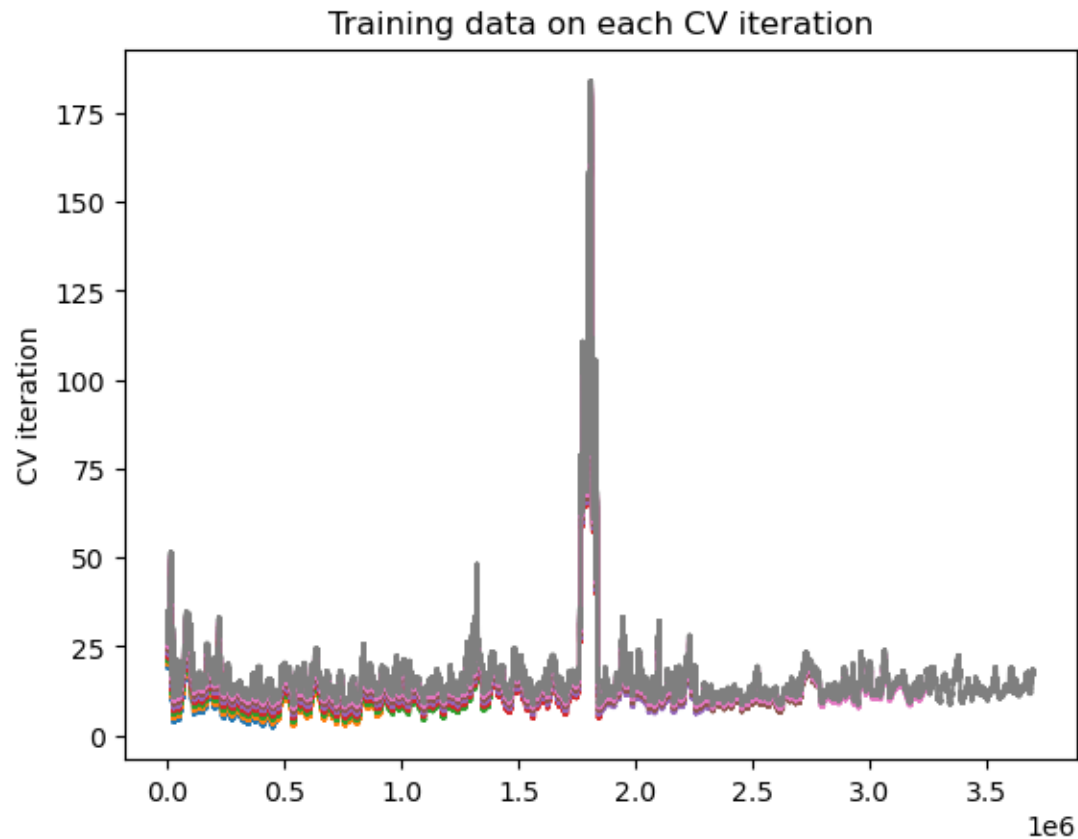


Figure 2: Cross-validation splits. PM2.5 vs Time

The quote below explains the large peak that appears in PM2.5.

*“The deadliest and most destructive wildfire in California’s history*

*The Camp Fire started on Thursday, November 8, 2018, in Northern California’s Butte County. Ignited by a faulty electric transmission line, the fire originated above several communities and an east wind drove the fire downhill through developed areas.... The fire caused at least 85 civilian fatalities, and injured 12 civilians and five firefighters. It covered an area of 153,336 acres, and destroyed more than 18,000 structures, with most of the destruction occurring within the first four hours. The towns of Paradise and Concow were almost completely destroyed, each losing about 95% of their structures.”<sup>2</sup>*

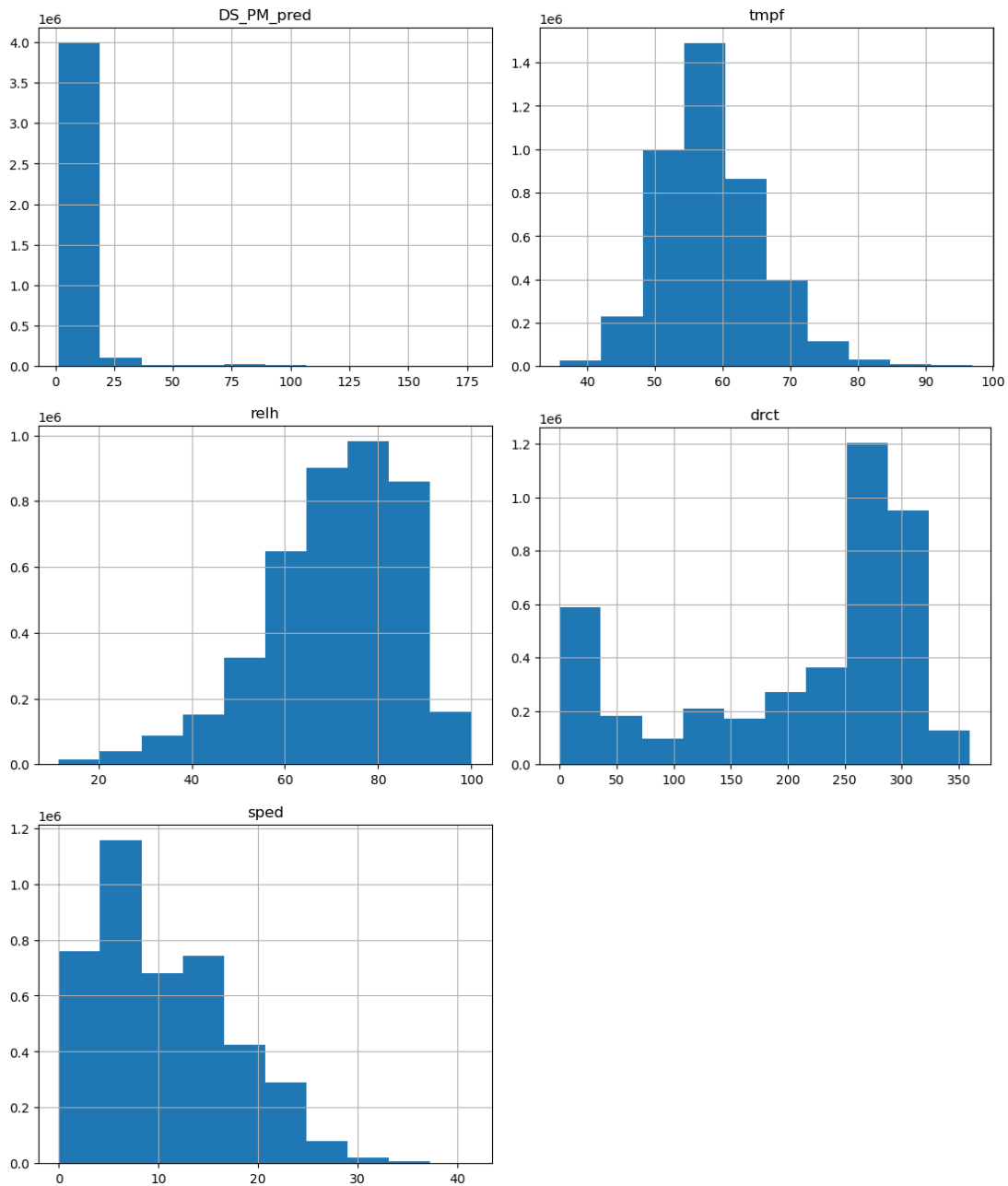
<sup>2</sup> <https://www.fire.ca.gov/our-impact/remembering-the-camp-fire>

From here, there are many ways to proceed onto modeling:

1. Remove the data associated with the peak in PM2.5.

This goes against the goal of the project as we want to know how long after a peak it should be safe to breathe the air. Different models can consider the peaks outliers but not for the purposes of this project.

2. For many machine learning algorithms, we need the data we input to be appropriately scaled. A standard normal distribution without outliers would be ideal. Below are the frequency distributions of the numerical features.

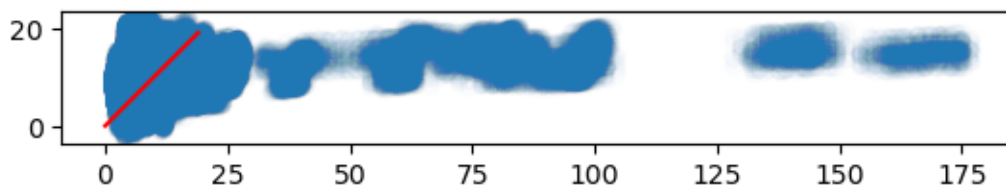


- 'DS\_PM\_pred': We expect for the most part to be right skewed because of wildfire, while the majority of the data reflects safe air quality levels.
- 'tmpf': This is almost normal and may just need scaling.
- 'relh': There is a slight left skew.
- 'drct': The wind direction may be default 0 when there is no wind speed. Otherwise, as a coastal city, the wind direction remains one way most of the time.
- 'sped': There is a slight right skew.

A comparison was made between 'StandardScaler' and 'RobustScaler' to see if accounting for outliers would make a difference. In a simple OLS model, the  $R^2$  did not change between either. Recall that the  $R^2$  score shows how close the regression line is to fitting the data. In the end, a 'MinMaxScaler' was used as it was more appropriate for the LSTM model later used.

### 3. Is linear regression appropriate for the data?

Below is a graph of a simple OLS with x-axis as actual data and y-axis as predicted, and an "ideal" regression line in red of  $R^2 = 1$ .



The  $R^2$  for this model is 0.070, so it does not have much predictive power. Still, using an OLS as a baseline still gives us something to compare other models to.

### 4. Should we consider location data to be categorical?

One hot encoding can be used to turn the unique location data into a categorical variable. A simple OLS with and without the one hot encoded location data were compared and the RSME's were compared. Recall that RSME is the root mean squared error and that the closer to 0, the better the model is at producing an accurate prediction. Without gave 9.3 and with gave 13.3. Additionally, processing time was a lot less without one hot encoding.

## Modeling

Although the previous section indicates that a simple linear regression will not have great predictive power, other regression models will still be my baseline in comparing models. Aside from regression models, a simple Long Term Short Memory Neural Network was chosen to model this time series. The following models will be compared along similar set-ups:

- Training-testing split via appropriate unshuffled split.
- Min-max scaling
- 4 models

1. LSTM
  2. Linear regression
  3. Ridge regression
  4. Lasso regression
- Root mean squared error as comparison.

For the regression models, a pipeline was used with:

- 'MinMaxScaler',
- alpha's [0.001, 0.01, 0.1, 0, 1, 10] where appropriate,
- 'TimeSeriesSplit' with 8 splits (1 split for each season).

A simple LSTM model was used with:

- 80-20 train test split,
- 'MinMaxScaler',
- A 50 node LSTM layer
- 1 Dense output layer
- 40 epochs

The loss function (MAE) was graphed for the 40 epochs of the training and testing sets of the LSTM.

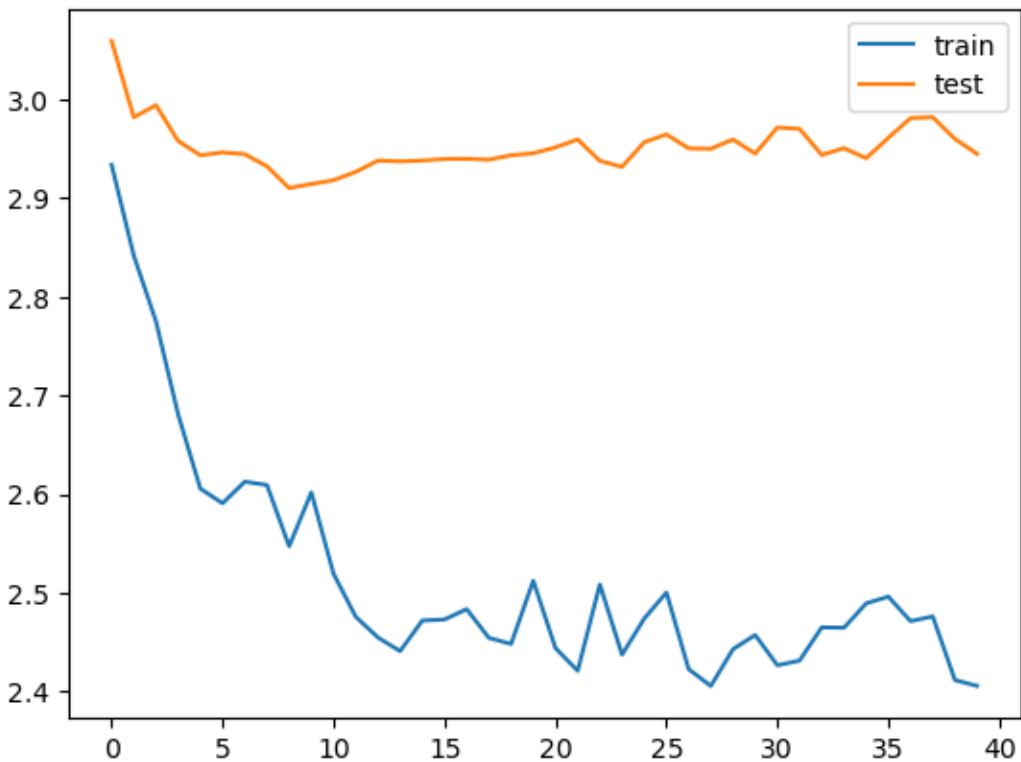


Figure 3: Loss vs Epochs



Although the training loss showed a steady decline, the testing loss stayed consistent. Adding different layers to make a deeper network is one way of making the loss match between the training and testing sets. An experiment was tried with a more complicated LSTM consisting of a convolution layer and drop out layers, but the training loss became worse while holding the testing loss at similar levels. This points to the lack of deeper features that the model can learn from and a preference for the initial simpler model.

Recall that the closer to 0 that RSME is, the better. The resulting RSME's for each model are:

LSTM	0.707
Linear regression	9.340
Ridge regression	9.340
Lasso regression	8.583

## Recommendations on use

- Given the same feature measurements (temperature, wind speed, prior PM2.5 measurements, etc.), one could make a prediction for PM2.5 within a small timeframe using the LSTM model.
- Even with a 1-dim ARIMA model, we can still make decent predictions with 95% confidence, as shown in the graph below. Negative values, of course, must be ignored in this prediction.

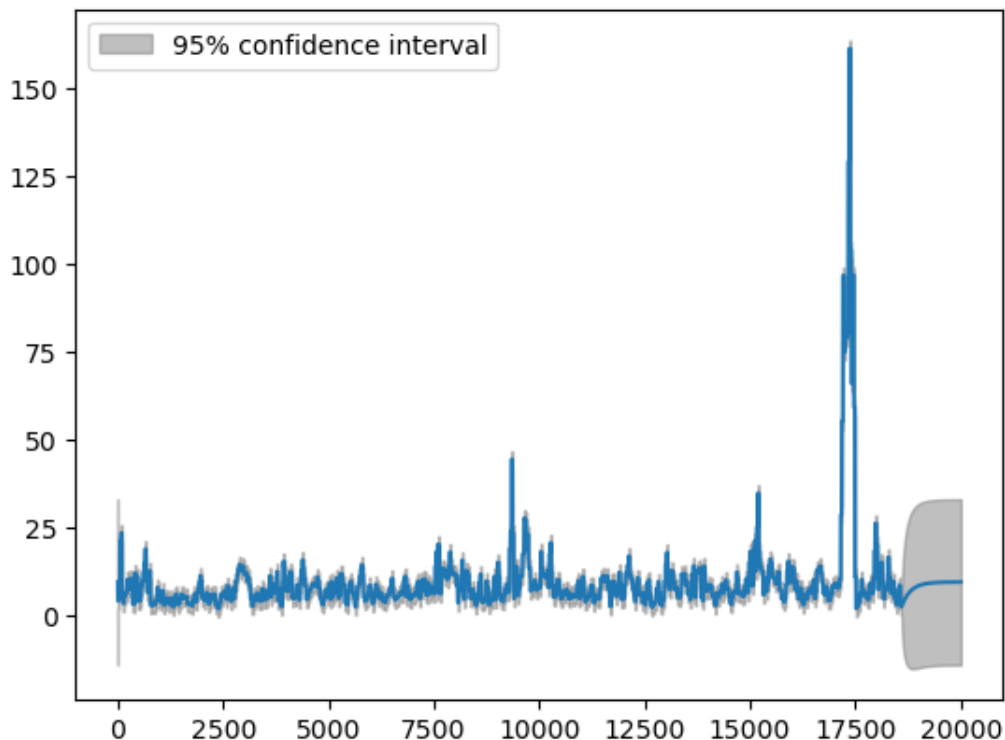


Figure 4: AR(1) model

- The data wrangling steps can be replicated on other state data from ASOS, so a model for other states can be easily derived from this one.

## Next Steps

- Although the relationship between the features and the label are non-linear, more can be done to predict for non-outliers. The spike indicated in previous can be taken out and more time series analysis can be applied to find more seasonality in air quality. Of course, the goal of the project is to predict safe (and unsafe) levels of air due to wildfires, so adding more datasets that use wildfires can also be done. Many things in nature have seasonality, so implementing another model that predicts the seasonality of wildfires can also be done but is surely outside the scope of this current project.
- Further steps like adding more layers to the LSTM, more parameters to the regression models, and other preprocessing methods can be applied. The ideal would be to get the RSME down to at most 0.5. However, without large changes like more added data (including the same data but for different years), the current changes may not lower the RSME enough.
- A larger scope can also be implemented, such as adding data from the rest of California. To really test the model's predictive power, we can look at data from other states as originally intended.
- Part of doing this project was learning how to use AWS to do some of the computing. More can be done in the optimizing of this code as well as learning more about AWS.