John Wu

Predicting air quality due to wildfire smoke

The Problem

Wildfire smoke falls under particulate matter (PM).

PM2.5 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. --- WHO

Can we predict how long the air will be dangerous for after wildfires?

Goal

We seek to predict the longevity of PM2.5, particulate matter most associated with wildfires (with diameters 2.5 micrometers or smaller).

We will look at weather station and air quality monitors around the San Francisco metro area from 2018 to 2019, and model on a time series regression problem.

Datasets

Weather Stations

Source: Automated Surface Observing System

https://data.cdc.gov/browse/select_dataset?tags=pm2.5

> Hourly

> Temperature, humidity, wind direction, wind speed

> Key: Longitude and Latitude pair

Air Quality Stations

Source: Center for Disease Control

https://mesonet.agron.iastate.edu/ASOS/

> Daily

> Mean PM2.5 concentration

> Key: Longitude and Latitude pair

Merged Dataset: Jan 2018 – Dec 2019

Combined under closest longitude and latitude of air quality stations, so each weather station has an associated PM2.5

> 4,165,056 rows × 7 columns

> 221 unique locations

> 8 features, 1 label

Features and label

Label: DS_PM_pred': Mean estimated 24-hour average PM2.5 concentration in µg/m3

Features

'DS_PM_stdd': Standard error of the estimated PM2.5 concentration

'valid': timestamp of the observation

'lon','lat': longitude and latitude

'tmpf': Air Temperature in Fahrenheit

'relh': Relative Humidity in

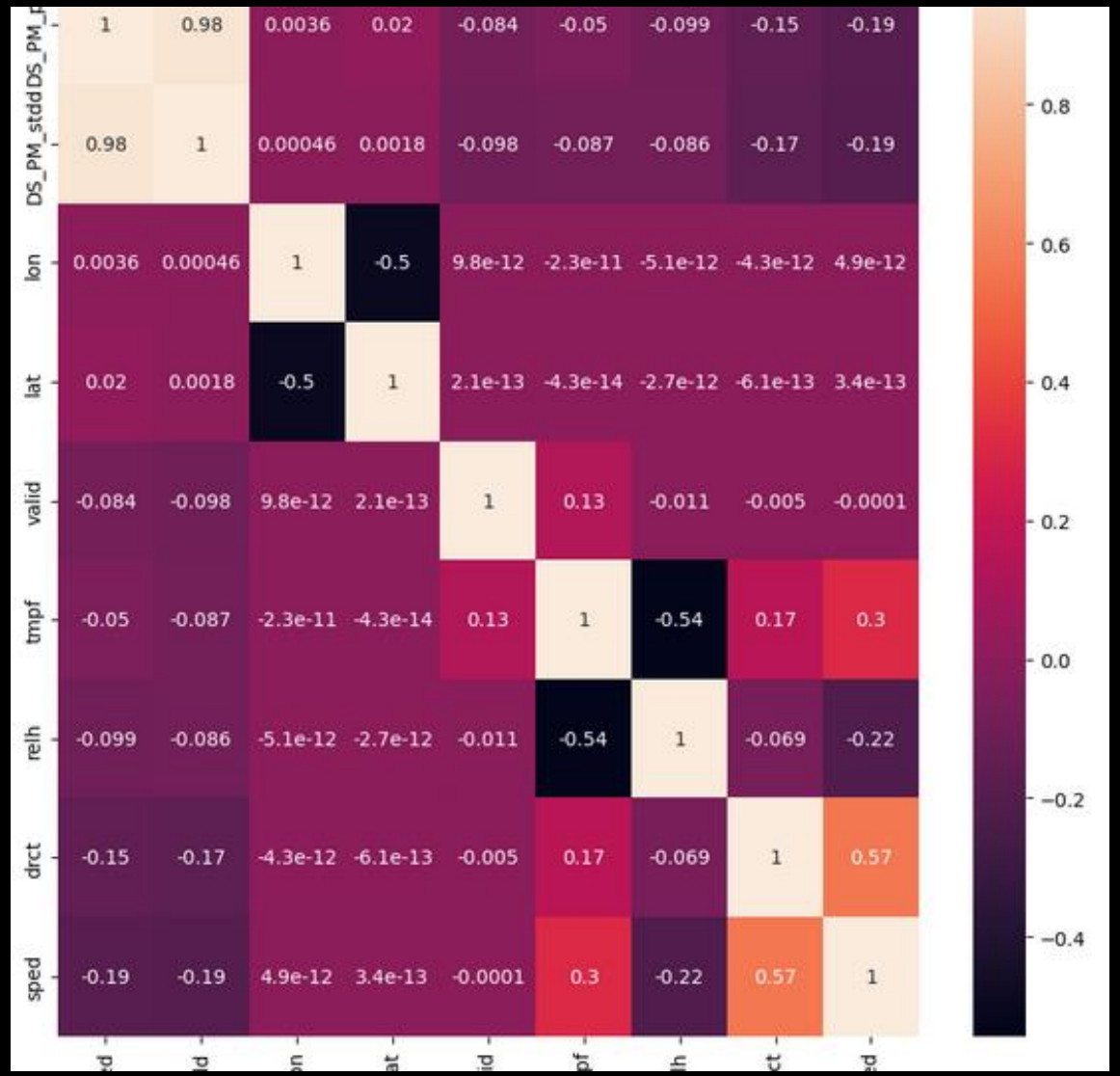
'drct': Wind Direction in degrees from true north

'sped': Wind Speed in mph

Exploratory Data Analysis

We look at a heat map of the correlation table of features

- ‘DS_PM_pred’ is our target variable.
- 'DS_PM_stdd': Standard error of the estimated PM2.5 concentration is not a physical metric; 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’.
- Latitude is more correlated than longitude by a (very) small order of magnitude.

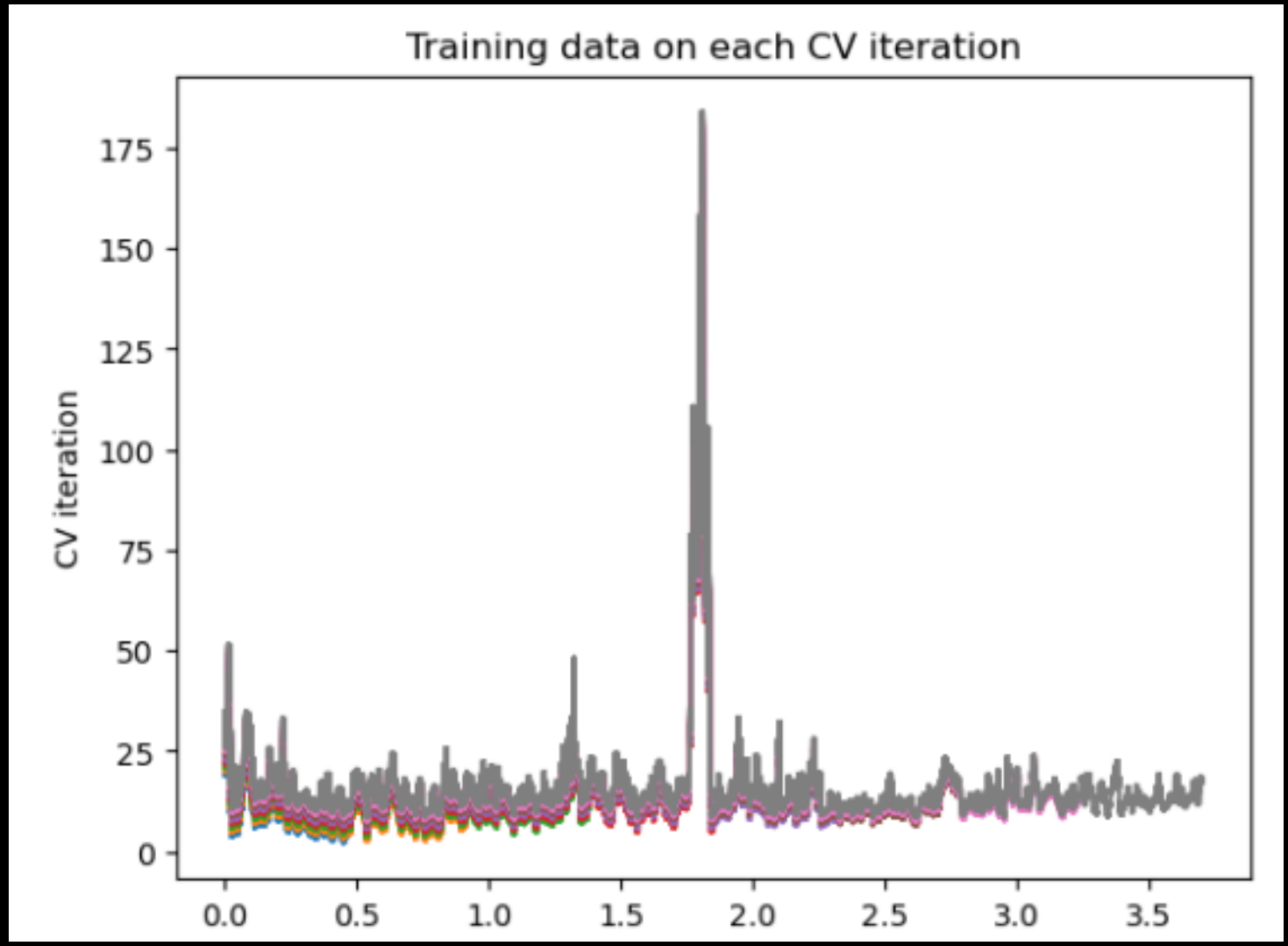


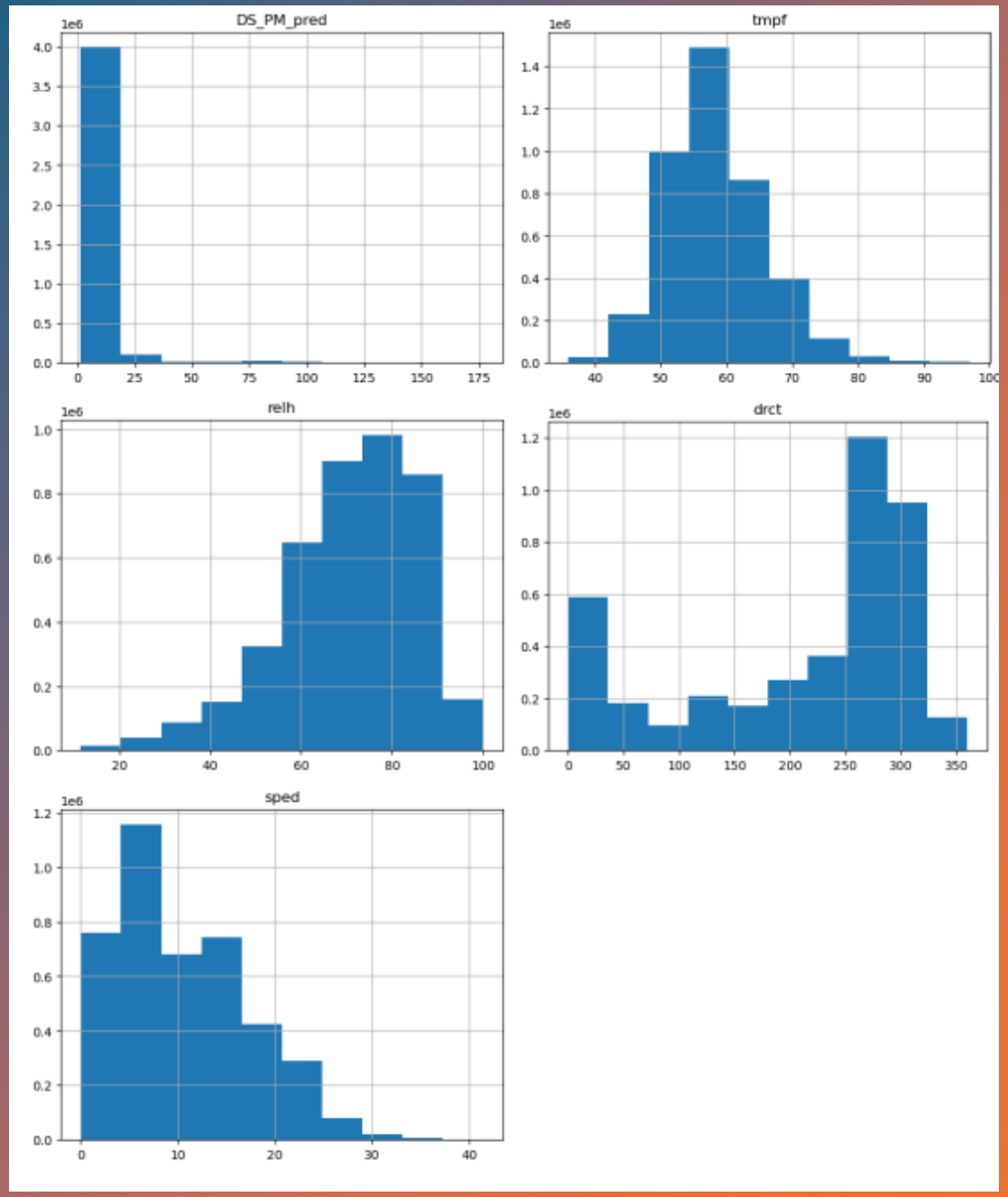
Visualizing the Time Series

Visualizing the cross validation split of the time series shows a large spike (time vs PM2.5, both standardized).

This has a historical explanation as the Camp Fire, which started on Thursday, November 8, 2018, in Northern California's Butte County.

This is 160 miles away from the air quality stations in San Francisco and was the deadliest wildfire in California history.





Frequency Distributions

For purposes of scaling, we look at the frequency distributions of the numerical features. A standard normal distribution is ideal.

- 'DS_PM_pred': We expect for the most part to be right skewed because of wildfires
- '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.
- 'sped': There is a slight right skew

Modeling

The following models will be compared with 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

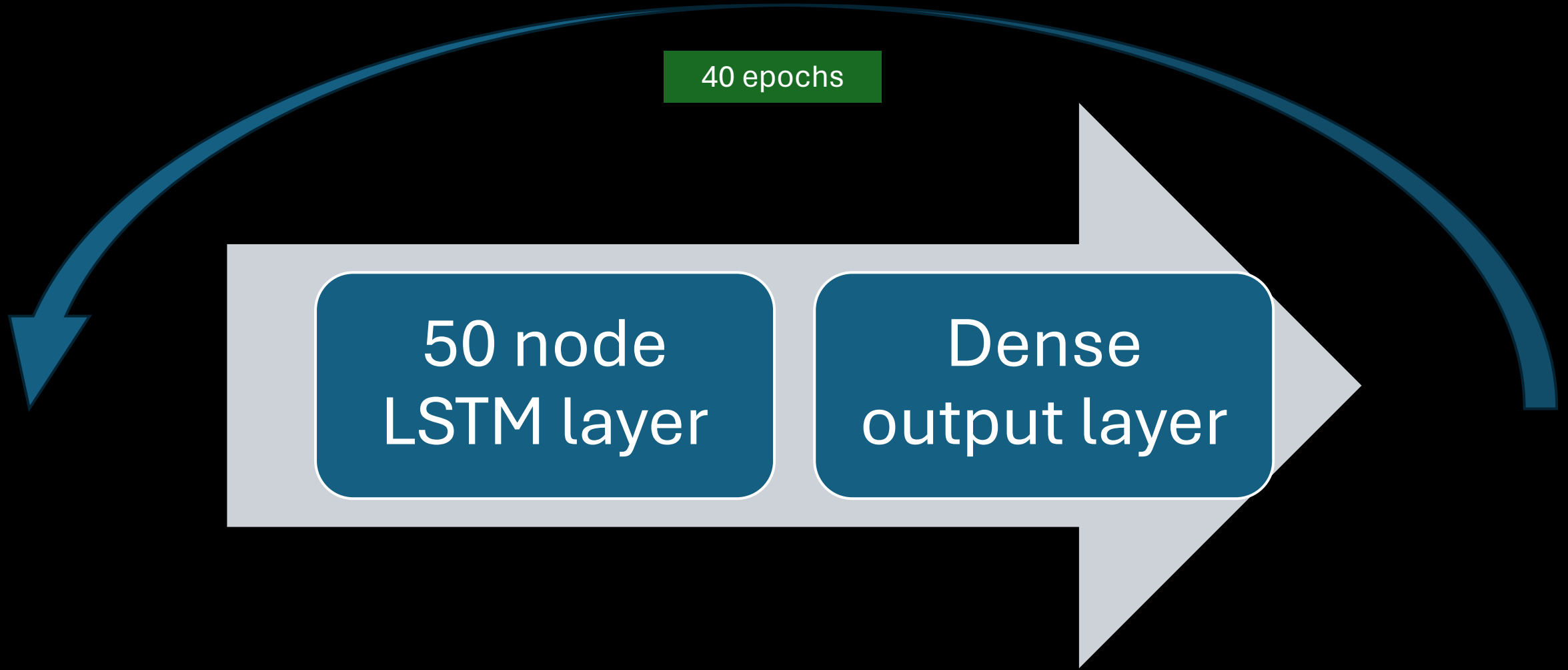
    Root mean squared error as comparison.

Regression Models

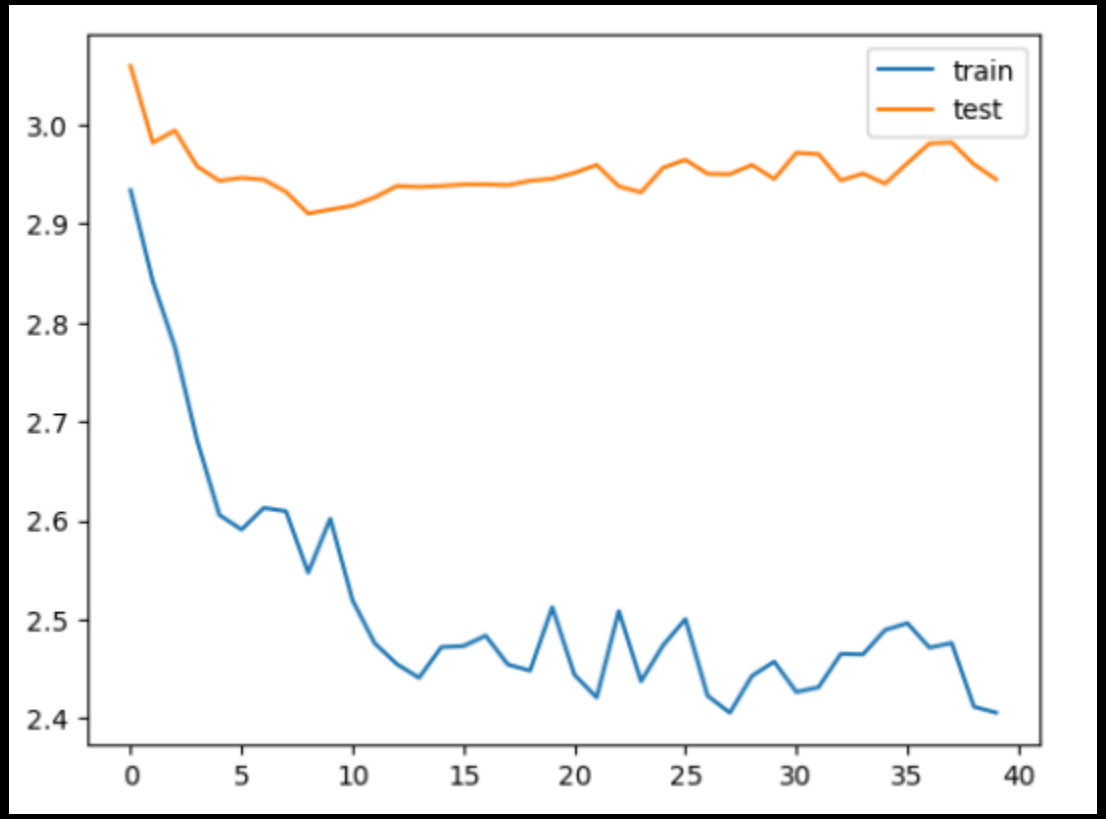
Each regression model (linear, ridge, lasso) implemented a similar pipeline

LSTM Model

The simple LSTM model used an 80-20 train test split and a MinMaxScaler



MAE Loss function for LSTM

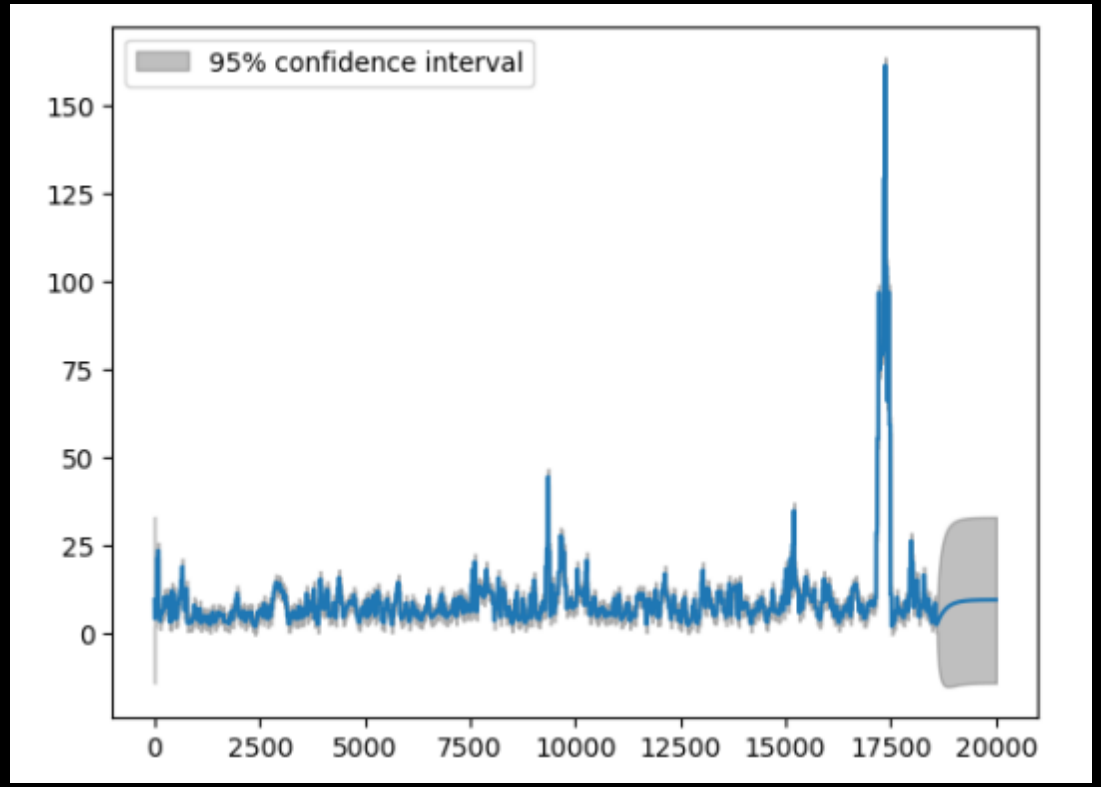


Comparison

Model	RMSE
LSTM	0.707
Lasso regression	8.583
Ridge regression	9.340
Linear regression	9.340

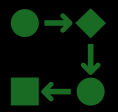
Recommendations on use

- Given the same feature measurements, 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.
- 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.

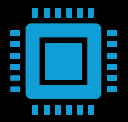


Next Steps

Although the relationship between the features and the label are non-linear, more can be done to predict for non-outliers.



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.



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