Report

Nico Ceresa

1. **A description of your dataset and where you found it.**

This dataset contains data on pollution levels and different metrics surrounding weather such as temperature, dew point, pressure, wind speed and direction, and precipitation. I found this data on Kaggle while looking at time series data.

2. **A description of the task you are trying to accomplish with the data via a machine learning technique.**

My goal with this project is to create a model that can predict pollution levels given data about the atmosphere.

3. **A description of any preprocessing you did to the dataset.**

Luckily for me, the dataset did not require a lot of preprocessing. I only had to encode one label: the wind direction. All of the other features of the data were quantitative metrics, so the only preprocessing that was necessary was to scale the data.

4. **A description of your ML Technique and why you chose it.**

The machine learning technique that I chose was a Recurrent Neural Network. Specifically, a Long Short-Term Memory model. I chose this model for this data as this data is a time series model; I could have used an ARIMA model, but I wanted to practice with some sort of neural network. This model uses information gates consisting of sigmoid and tanh functions.

5. **A description of the summary statistics.**

For the summary statistic, I chose to evaluate my model using the Root Mean Squared Error. I chose to use it as my accuracy metric as RMSE gives a relatively high weight to large errors. This means the RMSE is most useful when large errors are particularly undesirable. On the training data, I had an RMSE of 28.9. On the validation data, an RMSE of 26.64. On the testing data, an RMSE of 27.3.

6. **A contextualization of whether the results were good.**

I think that considering how non-complex my model is, the results were great. Sure, if I was to go deeper into this and really make a super complex model, I could achieve much better results with lower RMSEs, but that is not necessary for what I am predicting. Predicting pollution levels, though requires general accuracy, does not require extreme precision as the difference between 110 and 120 is not very noticeable. So, having RMSEs around 27 is not horrible.