***Introduction (5 points)***

*This should be similar in structure to the midterm report. In your introduction, make sure to motivate your problem, describe your dataset, and the previous work. If someone solved the exact same ML problem as you, describe the previous work and the model’s predictive power. Your model’s predictive power will likely be similar to previous work. If it is not, something is likely incorrect in your code.*

The Breathe Providence project has developed a network of 25 low-cost air sensors distributed throughout Providence, RI. The project’s goal is to characterize neighborhood-scale pollution dynamics and disentangle the impacts of local versus transported pollution sources. Each sensor measures a suite of common air pollutants, including PM2.5, CO2, CO, O3, NO, and NO2.

NO2, or nitrogen dioxide, is a common air pollutant known to vary at sub-kilometer spatial scales. It is produced by a variety of natural and anthropogenic processes, including fossil fuel combustion, lightening, and photochemical reactions with precursor pollutants. NO2 has been linked to the development and exacerbation of respiratory diseases such as asthma, making of particular concern in Rhode Island: as of 2022, the state had the highest statewide asthma prevalence in the country. Historically, the RI Department of Health has measured NO2 at two places in the state: at a suburban site in East Providence and at a near-road site in Cranston. These “reference” monitoring locations represent the lower and upper bounds of NO2 concentrations likely to be observed in Providence, but do not offer insight into local-scale pollution variability.

Low-cost sensors’ advantages in cost, size, and ease of operation make it possible to deploy them in high-resolution networks; however, with these advantages come significant tradeoffs in data quality. Electrochemical gas sensors are known to be affected by meteorological conditions (e.g. temperature and humidity), cross-sensitivities to pollutants other than the target gas, and drift over time. These sensor-unique biases can be mitigated through calibration schemes, which often rely on the relationship of one or more co-located low-cost sensors to reference-grade measurements. Previous work has investigated the use of machine learning methods—including linear regression, random forest, and support vector machine methods—for calibrating co-located sensors. Though more complex ML models can achieve XYZ, physically interpretable models are preferable…. This project seeks to develop a physically interpretable machine learning calibration for the Breathe Providence NO2 measurements using a co-located sensor at the East Providence reference monitoring site. Though this work addresses only the co-located calibration, this method will eventually be used for field calibration, which will provide insight into intra-urban NO2 dynamics.

For this problem, the target variable is the reference NO2 measurement made hourly at the East Providence site, which is a Federal Reference Method using a chemiluminescence analyzer, operated by the Rhode Island Department of Health. Reference NO2 data were accessed via the Air Quality System (AQS) API, the US Environmental Protection Agency’s repository of ambient air quality data. The features for this problem are measurements taken by the co-located Breathe Providence low-cost sensor; all features are continuous. Concentrations of NO2, NO, and O3 (the latter two being known cross-sensitivities) are measured by an Alphasense XYZ and have units of voltage. Meteorological variables (temperature and relative humidity) are measured by a BME and have units of degrees C and percentage, respectively. Low-cost sensor measurements were accessed via the Berkeley Environmental Air-quality & CO2 Network (BEACO2N) API; BEACO2N, a group at the University of California, Berkeley, supplied the sensors used by Breathe Providence. All feature variables are measured every few seconds and aggregated to hourly averages. The dataset covers the entire year of 2023. Though it is time series data, this is not a time series problem; thus, the dataset is IID.