Hyperlocal Air Quality Prediction in Houston

Kaveri Chhikara, Vignesh Venkatachalam, Carolyn Vilter, Vishal Vincent Joseph

Introduction

When the project began, our general goal was to investigate the intersection between climate science and machine learning, particularly in the Global South. A foray into the literature supported the idea that machine learning predictions were a promising climate science approach: though complex in their own right, compared to dynamic or physical climate models they require less developmental time, take minimal inputs, and are relatively simple.

Literature and data review suggested that meteorological variables like rainfall and temperature, as well as data on air quality, were readily available at a relatively granular level in most countries. Initially we intended to focus on data from India, motivated by personal interest and the fact that climate issues in the Global South are particularly pressing yet understudied. Unfortunately, for the same reason, we found it difficult to source adequate data for India. So our focus shifted to the United States: specifically Houston, Texas.

Between 2017 and 2018, The Environmental Defense Fund (EDF) partnered with Google Earth Outreach to map air pollution at the hyperlocal level. In various cities around the world, mobile air quality sensors were used to gather air pollution data at street level, showing how air pollution varies over very short distances. One of those cities was Houston, where low-cost sensors were outfitted on city fleet vehicles, Google Street View cars, and stationary locations.

Given this unique data source, we were motivated to focus our project on air quality prediction in Houston. For simplicity, and on the basis of data availability, we decided to focus on data for just one pollutant: NO_2. Unlike ozone and particle pollution, which can be of concern over large regions, NO_2 levels are appreciably higher in close proximity to pollution sources (e.g., vehicles on major freeways, factories)[ref] – hence we felt it was an important pollutant to analyze.

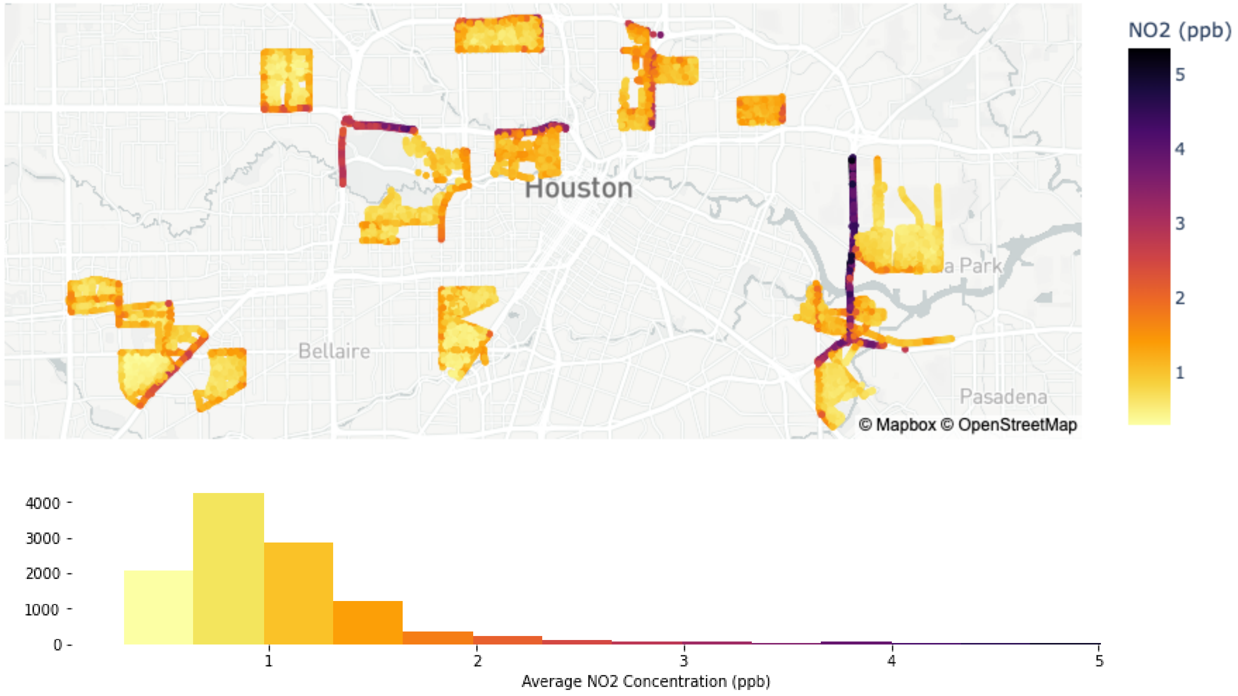
In particular, we undertook to develop machine learning models that could learn to associate EDF's air quality data with hyperlocal data on meteorological conditions, local emissions, and traffic. From there, our goal was to predict pollution levels in nearby Houston neighborhoods that have weather, emissions, and traffic data available but did not take part in EDF's air quality monitoring.

It's important to note that our project is inspired and informed by Varsha Gopalakrishnan's "Hyperlocal Air Quality Prediction using Machine Learning" project.[ref] Gopalakrishnan performed her analysis using Oakland data, and our intention was to adapt most aspects of her approach to data from Houston.

Datasets

Our target variable, hyperlocal NO_2 pollution levels, comes from the EDF Houston air quality dataset available for download through the OpenAQ portal. NO_2 levels were collected over nine months in certain designated areas in Houston. The result is a dataset with average NO_2 readings in ppm, latitude, and longitude for 11,534 precise locations around the city.

Figure 1: Average NO_2 Concentration in Houston: heatmap and histogram



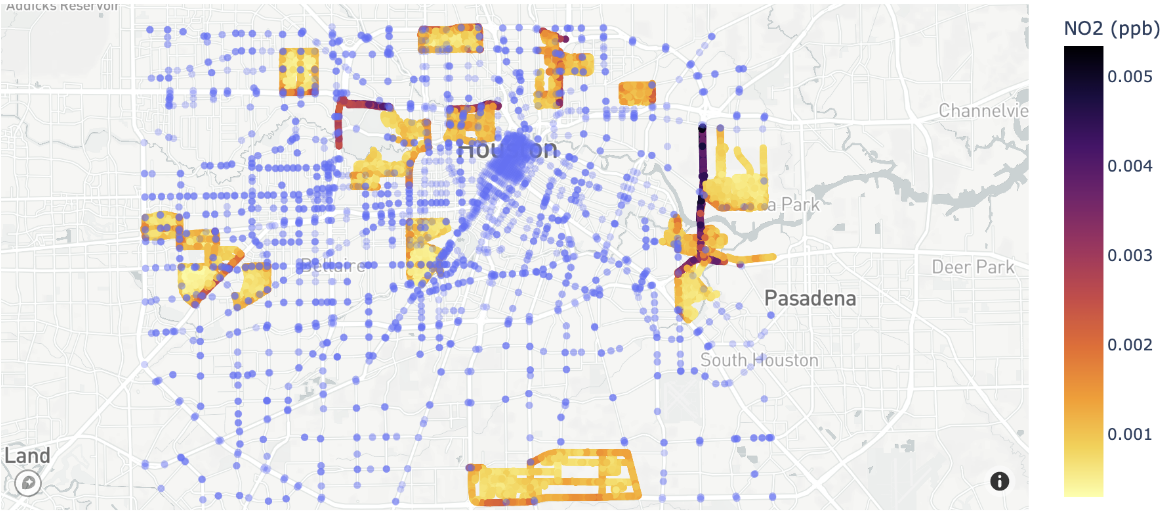
Interestingly, despite Houston's sizable chemical, manufacturing, and recycling industries and history of relatively poor air quality,[ref] we found that NO_2 concentrations in the dataset ranged from 1 to 5 ppb, with the overwhelming balance of measurements on the lower side of the range (figure 1). These values are well within acceptable NO_2 pollution levels; for reference, the EPA's annual average NO_2 standard is 53 ppb.[ref]

Still, the dataset contained substantial variation within that range. The level of pollution at any one location depends on a number of factors, including traffic on nearby streets; emissions from facilities like railroads, ports, and factories; and meteorological phenomena. Accordingly, to predict EDF air quality, we gathered feature data across these three buckets.

Traffic Data

To define bounds for our traffic data gathering, we first referenced the geographical range for which EDF collected the air quality data in Houston, specifically the minimum and maximum latitude and longitude values. We chose this area as the bounding box and undertook to determine the location of all traffic signals and intersections within that boundary (figure 2).

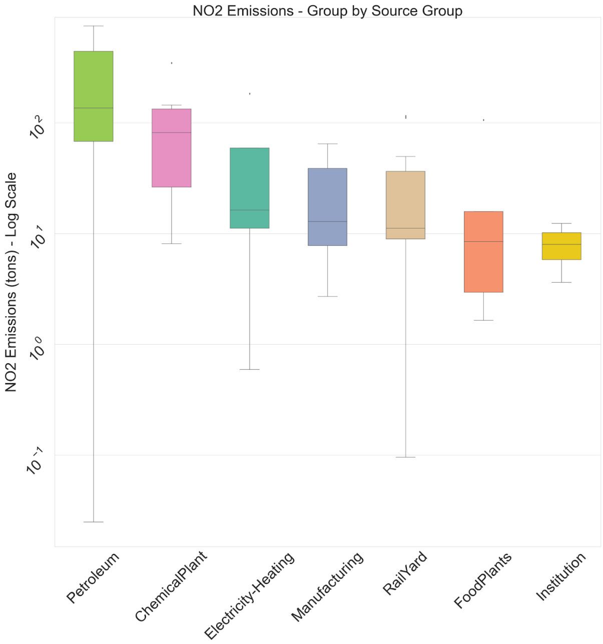
Figure 2: Intersections across Houston



The Overpass API from Open Street Maps is used to determine the location of all traffic signals (intersections) within a given bounding box. The Overpy library is used to send the request to the API and this call returns the latitude and longitude of all traffic signals. Next, the distance between each traffic intersection and each point in the monitoring data is measured. For a given monitoring point, the traffic score is calculated as the number of traffic intersections within 1,000 ft of each point in the monitoring data.

Emissions Data

Figure 3: Emission sources by type



We use NEI's data for point sources – which provides facility specific pollutant information from across the US. Similar to how we bound based on a boundary box region, we consider only those facilities that fall within the (lat, long) range for which EDF data was collected.

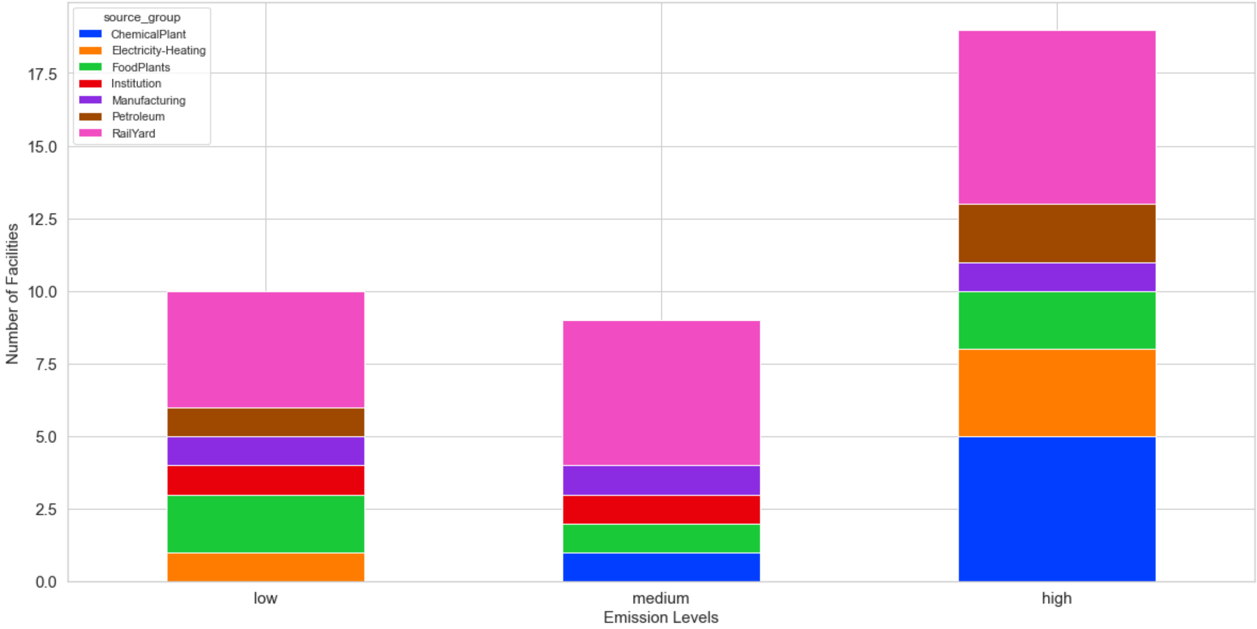
After grouping some of the sources based on their types (figure 3) and range of emission values, we calculate the distance between the location of every pollutant data read (from the EDF data) to each facility. We also calculate the emissions per distance measure for each such row, with the idea being that both the quantity of NO_2 emission from the facility, as well as the distance from the facility will affect NO_2

concentration at a certain point. Keeping explainability in mind, we dropped data from a few sources marked “unknown”, as they do not help us understand the type of polluting facility, even if they help us predict NO_2.

The box plot shows the groups we bucketed the facilities into based on their types and emission distributions. We see that Petroleum and Chemical plants are amongst the larger pollution sources.

Additionally, to get a sense of the number of facilities distributed across Houston, we bucketed these facilities into Low, Medium and High polluting groups and calculated their numbers in each bucket (figure 4).

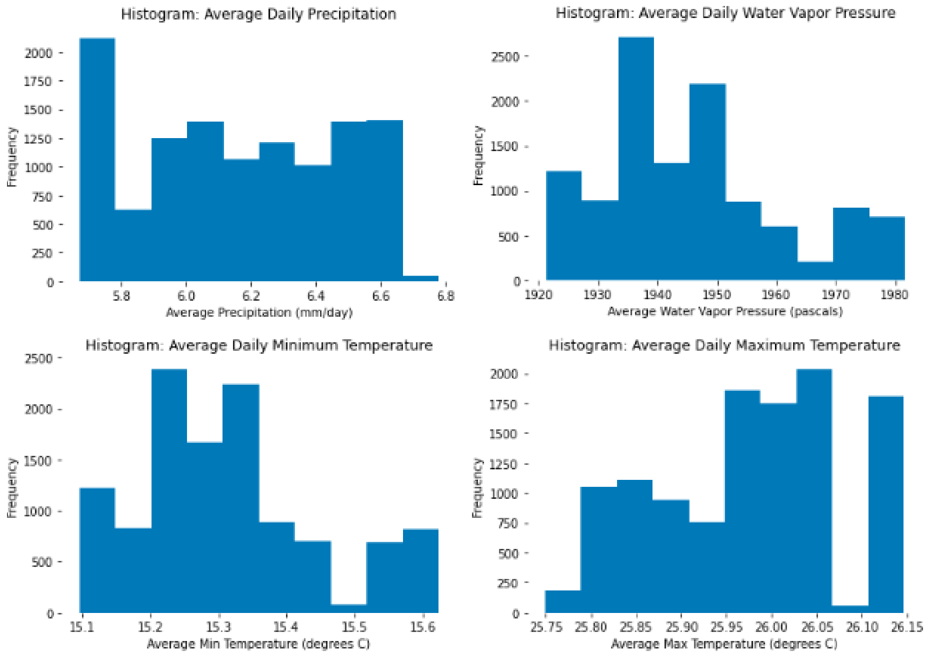
Figure 4: Number of point source facilities by group and emission levels



Meteorological Data

Following Gopalakrishnan's approach, we use the Daymet API[ref] to collect our meteorological data. Daymet's "daily surface weather and climatological summaries" provide historical weather data at a daily, 1km grid granularity, extrapolated from less-granular meteorological observations. Through approximately 200 calls to the Daymet API, we were able to gather the available daily data (maximum temperature, minimum temperature, shortwave radiation, vapor pressure, and precipitation) at latitudes and longitudes corresponding to the locations of our air quality measurements and produce averages for each metric across the 9 months when EDF was measuring air quality in Houston.

Figure 5: Histograms of relevant meteorological metrics



Experiment Setup

The first step in our analysis was to combine all four data sources – air quality, traffic, emissions, and weather – into one master dataset, which we did by wrangling the three feature datasets to correspond with the latitude and longitude observations in the air quality data. The final dataset has 11,534 observations for each of the monitoring points in the EDF dataset and 82 features from traffic, facilities and meteorological datasets.

The next step before running the baseline model was to normalize the features. We used `StandardScaler()` to remove the mean and scale by unit variance. Normalizing the features allows us to effectively gauge their relative importance.

Our objective was to use this combined dataset to develop a model that can accurately predict NO_2 pollution levels in Houston neighborhoods where EDF air quality data was not collected. In particular, using several different approaches, we train models to predict NO_2 emissions on the same locations as the EDF data and then use the trained model and additional traffic, features, and emissions data to predict concentrations elsewhere in Houston.

Our baseline model uses a simple OLS linear regression for feature selection. This is a relatively simple model, which we could use to assess if our data munging process had been successful and if the analytical dataset had been created correctly. Additionally, the model gives a quick idea of what could potentially be important features in the prediction, and helps us check if certain features we anticipate to be crucial predictors are recognized as important by the model. To begin with, we used a threshold of +/- 90% to indicate very high correlation between features (figure A1), dropping such features to avoid multicollinearity. We then use the remaining features to run a simple linear regression to gauge the significance of each feature for predicting NO_2 emissions. These basic preliminary results indicated that minimum temperature, precipitation, and number of intersections were the most important of our observable features for predicting air quality.

Beyond the baseline, we chose to implement 6 more complex models for our prediction, namely Linear Regression with and without PCA, Lasso Regression, Ridge Regression, Random Forest, Extreme Gradient Boosting (XGBoost) and Neural Networks. For each model, we calculate negative mean squared error as a measure of performance and to compare our models' relative effectiveness.

The reason for selecting the above set was to account for models with different underlying approaches to prediction. Between these 6 models, we encompass simple linear regression, regularization, tree-based models and neural networks – thereby spanning a wide range of machine learning approaches.

The models are discussed in greater detail in the next section.

Methods

A summary of our feature selection, cross validation approaches, and the hyperparameters we tuned can be found in figure 6.

Figure 6: Feature selection, cross validation, and hyperparameters

Class	Model	Feature Selection	Cross Validation	Hyperparameters
Linear	OLS (Baseline)	Univariate Screening (Pearson's Correlation)		
	Linear Regression	w/ & w/o PCA		
	Lasso Regression	Embedded	GridSearchCV	`alpha`: [$1e^{-5}$, $1e^{-4}$, $1e^{-3}$, $1e^{-2}$, $1e^{-1}$, 1]
	Ridge Regression	Embedded	GridSearchCV	`alpha`: [$1e^{-5}$, $1e^{-4}$, $1e^{-3}$, $1e^{-2}$, $1e^{-1}$, 1] `solver`: ['auto', 'svd', 'cholesky', 'sag']
Tree Based	Random Forest	Embedded	RandomizedSearchCV	`n_estimators`: [100, 300, 500] `max_features`: ['auto', 'sqrt', 'log2'] `max_depth`: range(10, 50, 10)
	XGBoost	Embedded	RandomizedSearchCV	`n_estimators`: range(50, 500, 50) `max_depth`: range(5, 50, 5)
NN	82 -(30-30-20) -1 with tanh activation	-	-	`Learning_rate, batch size, num_epochs, optimizer`

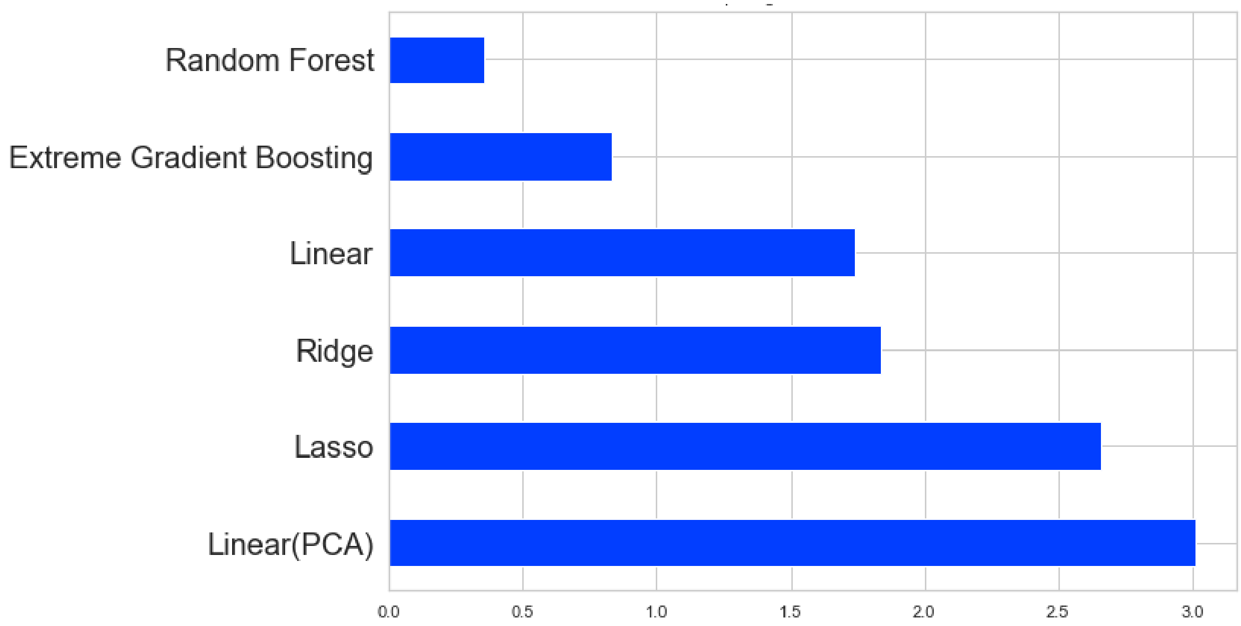
To ensure robust parameter tuning during cross validation, for each of these models we use Scikit Learn's GridSearchCV and RandomizedSearchCV approaches that exhaustively consider all/most parameter combinations. The number of folds considered for cross validation is k = 10.

For the tree based methods, feature selection was done after carrying out hyperparameter tuning using all features. Using the shortlisted hyperparameters,various models were created for different subsets of features using the feature importance threshold as a hyperparameter (incremented in steps of 0.002). The subset that resulted in the lowest mean squared error was chosen as the final set of features. These features were then used to fit the final model along with the best hyperparameters chosen in the first step.

For the linear models, PCA was carried out as the feature selection technique by considering all principal components that together explained 90% of the total variance. In the case of lasso and ridge, no extra steps were required for feature selection and the coefficients were assigned values based on the type of regularization.

For the neural network model, we first split the data to use 70% for our training set and 30% for our testing set. We used a model with 8 hidden layers. We compared the model performance for different activation functions – ReLU and Tanh – and found that the model performed better with a non-linear activation function. We used Stochastic Gradient Descent (SGD) as an optimizer. We did not use standardized features for this model because we felt that a neural network would form combinations in dimensions that were the most optimal.

Figure 7: Model evaluation: mean squared error



We found that the random forest model performed best, with the lowest mean squared error value (figure 7).

Results

Finally, we undertook to predict NO_2 levels in new neighborhoods in Houston using the random forest model given its low MSE. In particular, we chose two regions in Houston that did not have NO_2 data from Google's data collection process. We selected these neighborhoods on the basis of our model features as well as spatial demographic and health data compiled by EDF. Out of a desire to induce variation in predicted NO_2 levels, we attempted to choose regions with disparate characteristics.

Inspecting our traffic data and emissions data alongside EDF data[ref] on the prevalence of adult asthma, prevalence of COPD, and life expectancy across Houston, we observed that neighborhoods in western Houston exhibited more favorable factors across the board: lower asthma and COPD incidence, higher life expectancy, fewer emissions point sources, and in the case of our chosen neighborhood, fewer street intersections. In comparison, many communities in southern and eastern Houston generally have higher incidence of asthma and COPD, lower life expectancy, and a higher density of emissions point sources. In accordance with this trend, we chose two neighborhoods – specifically two grids of points bounded by latitude and longitude – in western and southern Houston.

Figure 8: NO_2 predictions for the selected regions: heatmap and histogram

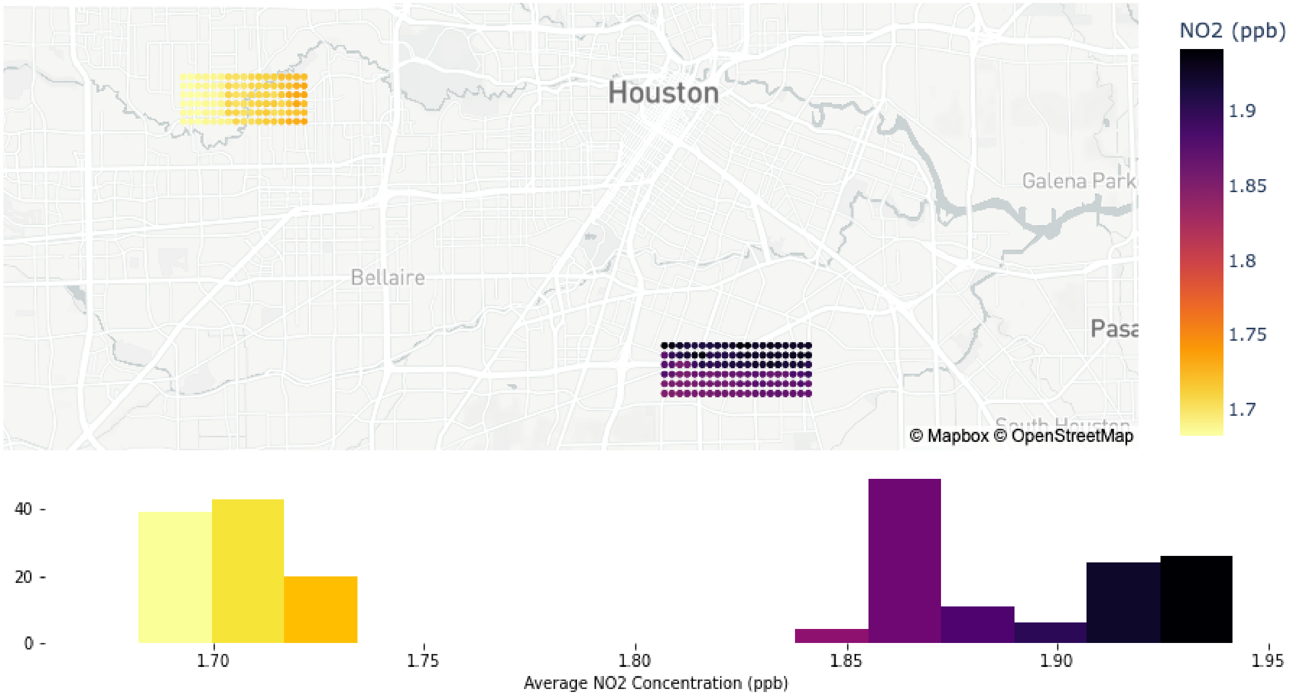
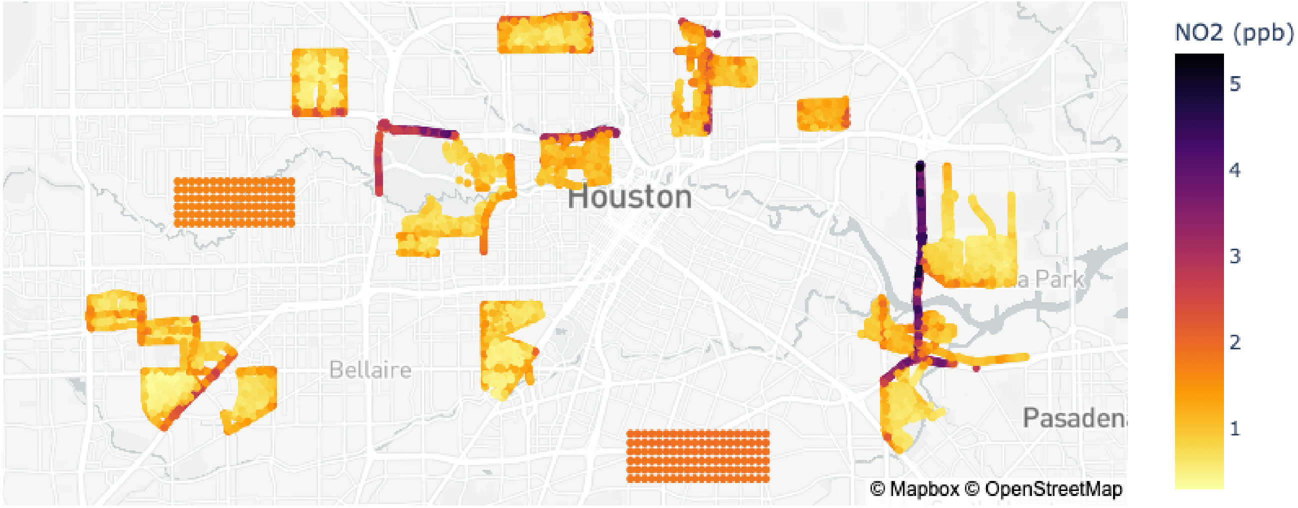


Figure 9: NO_2 predictions in the context of the existing Houston NO_2 data



After collecting additional meteorological, intersection distance, and point source distance data for our chosen grids, we were able to produce predicted NO_2 values for the regions using the random forest model. Our results are pictured in figure 8. We found that our NO_2 predictions occupied a range from about 1.65 ppb to 1.95 ppb. In the context of the original EDF NO_2 data, these predicted values occupy a relatively dense section of the histogram. Our predictions are presented in the context of the training dataset in figure 9.

One interesting aspect of our predicted results can be observed in figure 8. The map and histogram show a clear split between the predicted values for our western and southern neighborhoods. Predicted NO_2 data points for our western neighborhood are represented exclusively in the left portion of the histogram and data points for our southern neighborhood are represented exclusively in the right portion, with a distinct gap in between. This disparity is consistent with our expectations in choosing these regions and suggests that further analysis might find an association between Houston's NO_2 pollution levels and policy-relevant factors like illness and life expectancy.

Conclusion and Key Lessons

Summary

Hyperlocal air monitoring is the need of the hour in order to keep air pollution in check. Motivated by this, we leveraged a rich hyperlocal air quality dataset compiled by EDF and Google in Houston to predict the air quality in other at-risk neighborhoods. Using multiple data sources that span facility emissions, traffic hubs and key meteorological factors, we attempted to estimate the effectiveness of various machine learning models in predicting NO_2 concentration levels. All models underwent robust hyperparameter tuning and feature selection to eliminate any potential bias. The random forest model seemed to perform the best on the available training data and was used to predict the NO_2 concentration levels at two neighborhoods chosen on the basis of spatial, demographic and health data.

Our results indicate that the southern neighborhood experiences poorer air quality as compared to the western neighborhood, which is in line with the inherent characteristics of these neighborhoods across asthma and COPD incidence, higher life expectancy and number of emissions point sources. This suggests that further analysis might find an association between Houston's NO_2 pollution levels and policy-relevant factors like illness and life expectancy.

What we learned

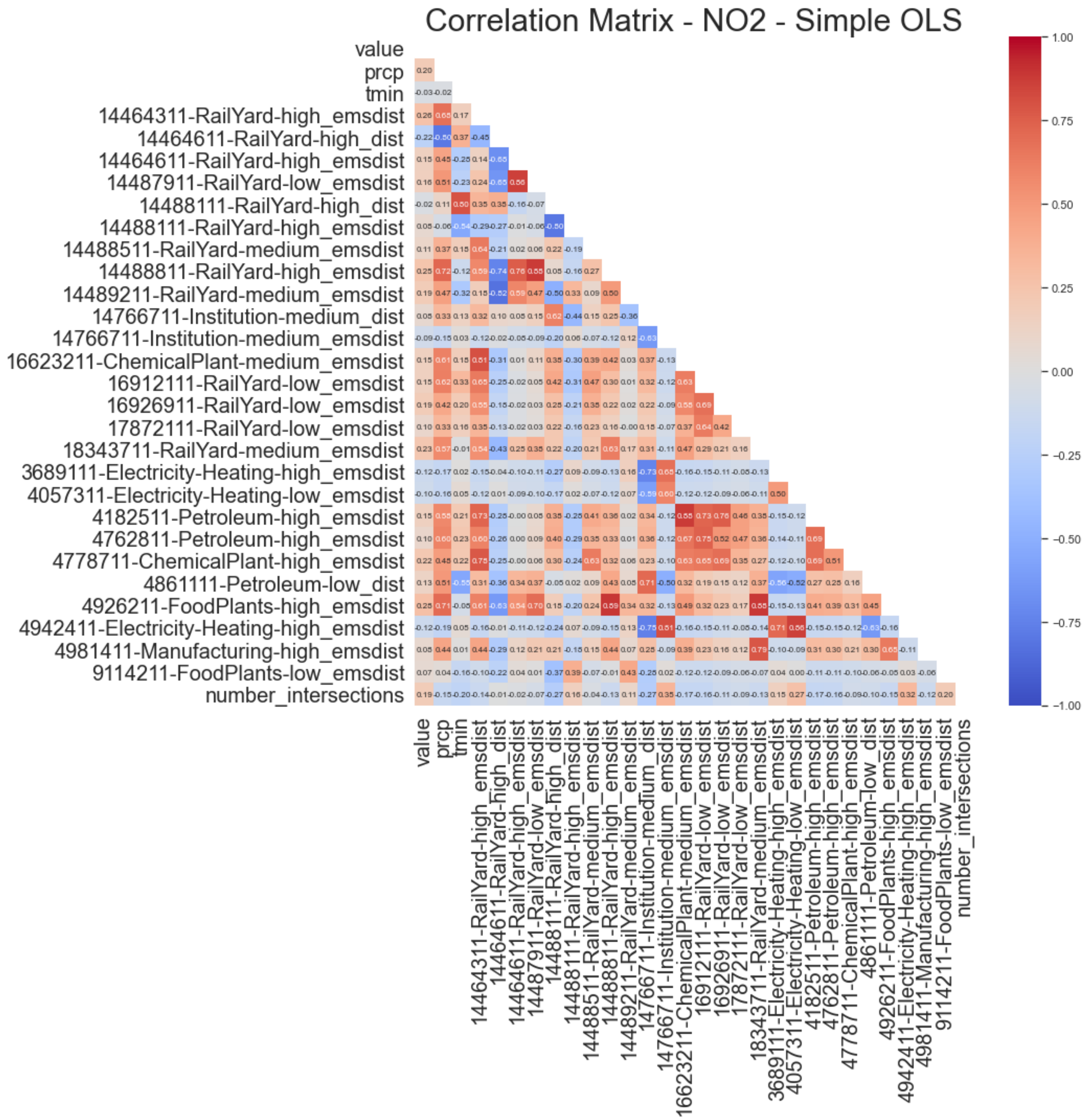
Our greatest learning outcome was gaining exposure to structuring a machine learning project and building the pipeline from scratch. We realized over time that hyperparameter tuning is more an art than a science, and requires a combination of research along with trial and error to achieve reasonable results. Most importantly, we experienced first-hand the necessity for data quality, since a lack of variability in our target variable gave us predictions that occupy a very narrow range (1.65 ppb - 1.95 ppb).

Our next steps to build on the progress we have made so far would be to:

- Address missing data issues in the facility emissions dataset
- Further explore the association between air quality and key policy-relevant factors like life expectancy, asthma prevalence, and income
- Optimize existing models by improving upon our hyperparameter tuning strategy and exploring more advanced architectures for the neural regression model

Appendix

Figure A1: Basic correlation matrix from Simple OLS regression



Figures A2-A5 show the results of our feature selection processes for each model.

Figure A2: Feature selection using Lasso Regression

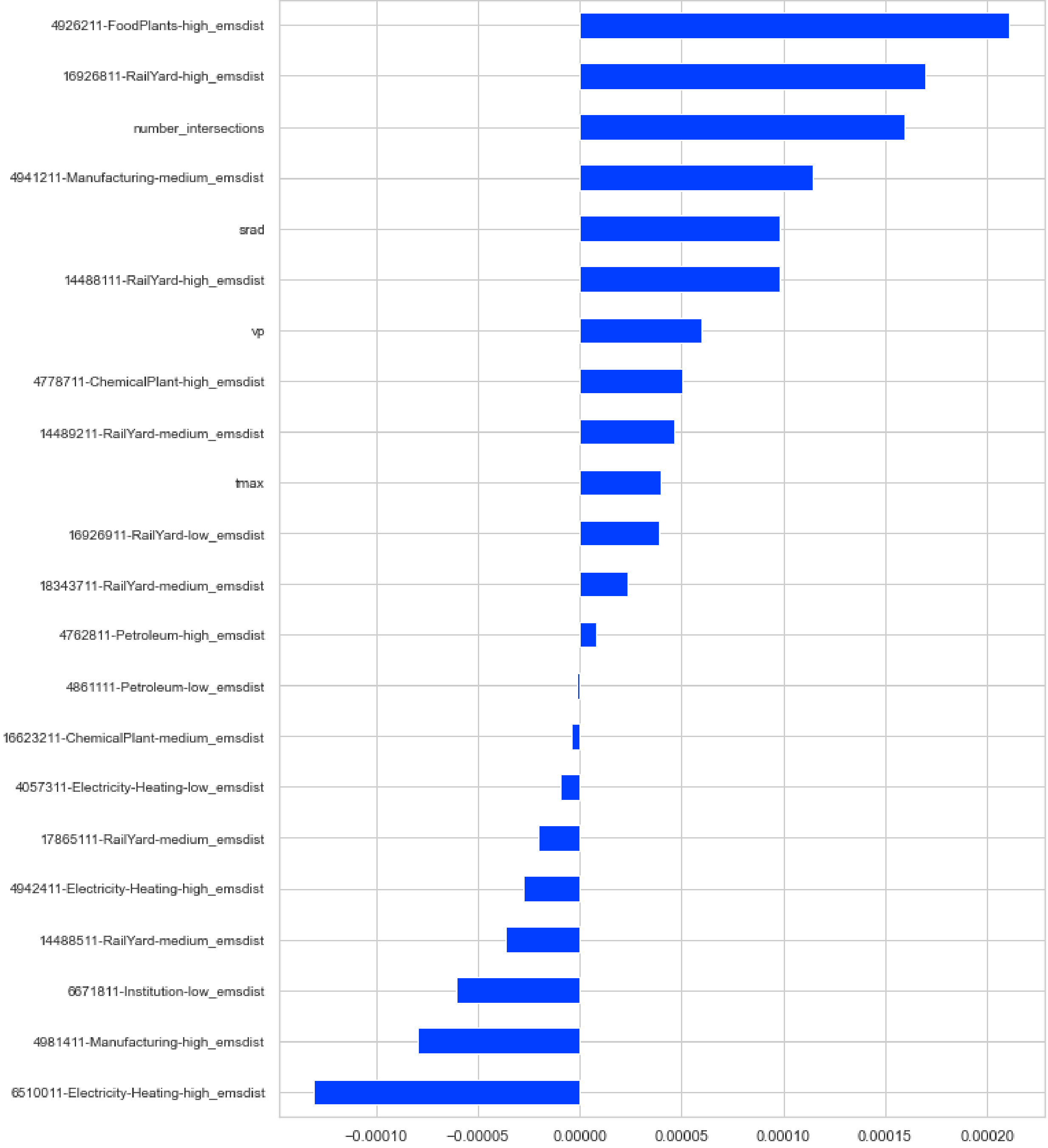


Figure A3: Feature selection using Ridge Regression

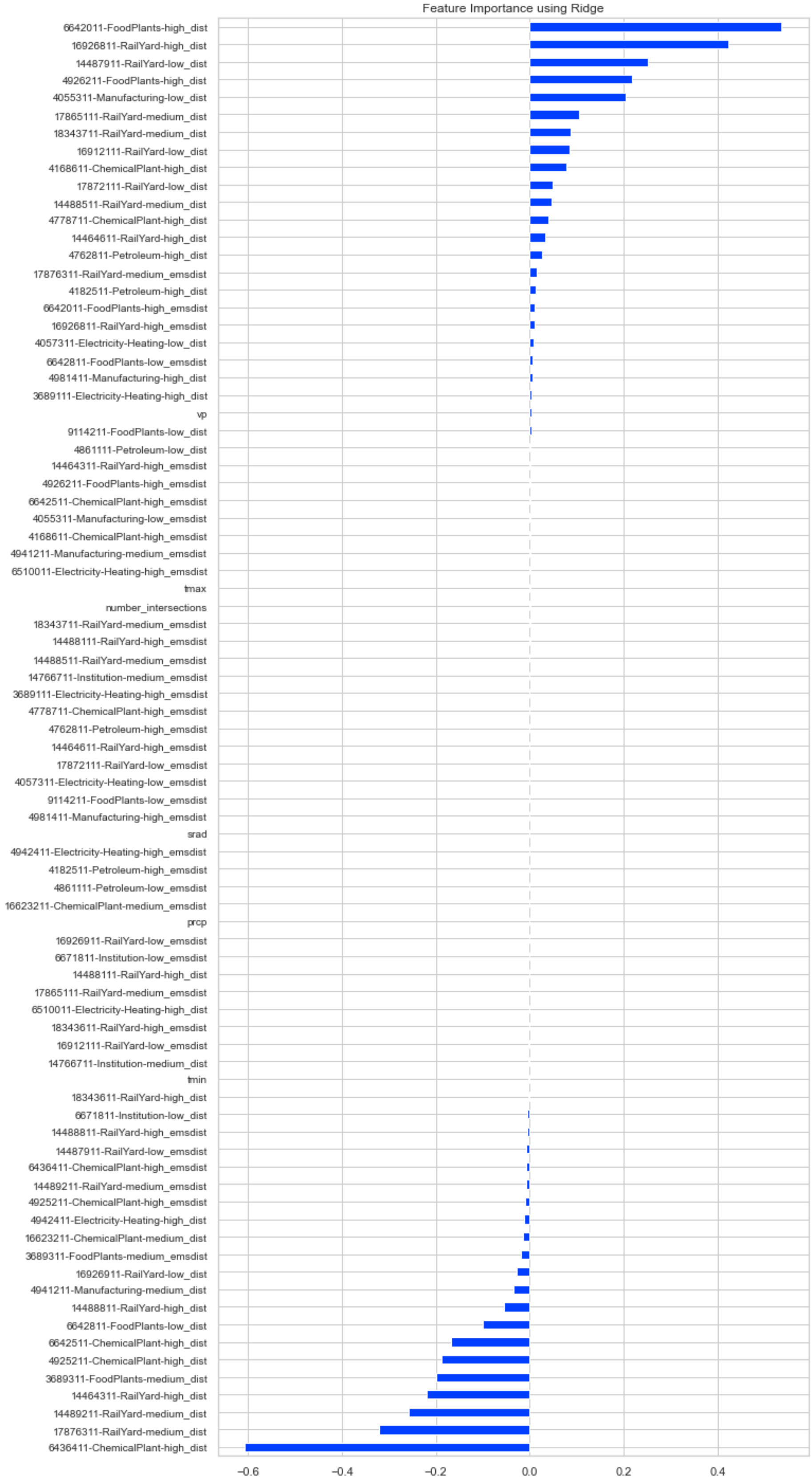


Figure A4: Feature selection using Random Forest

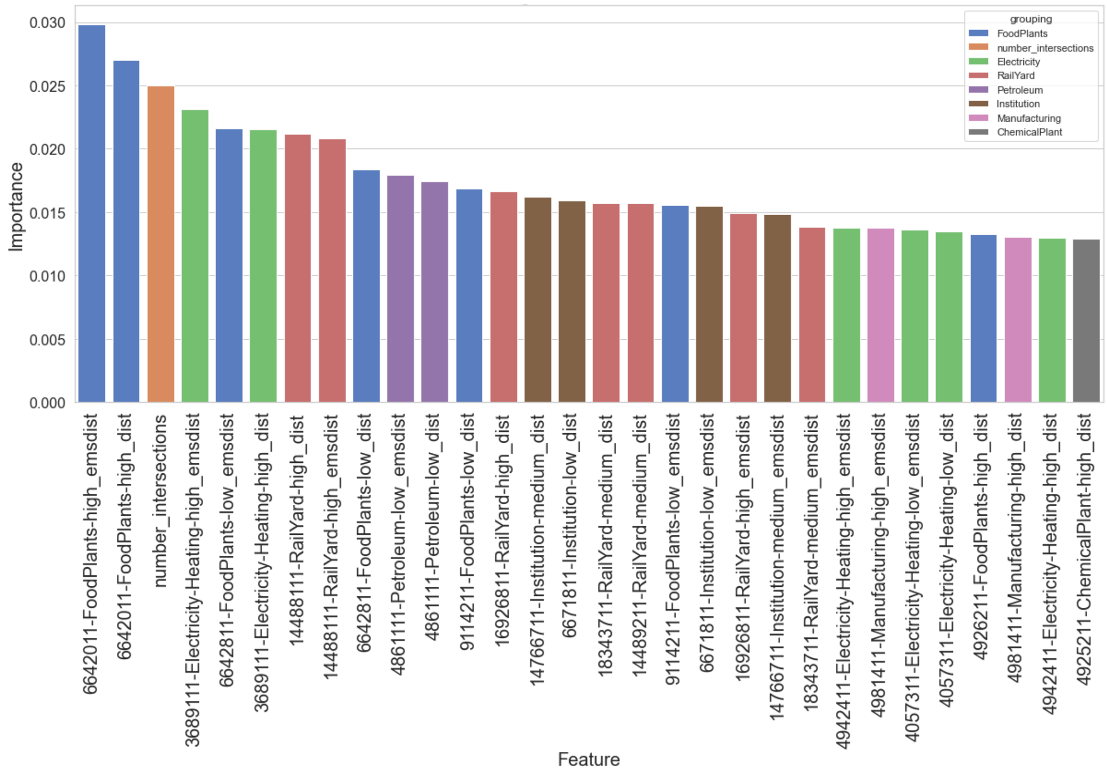
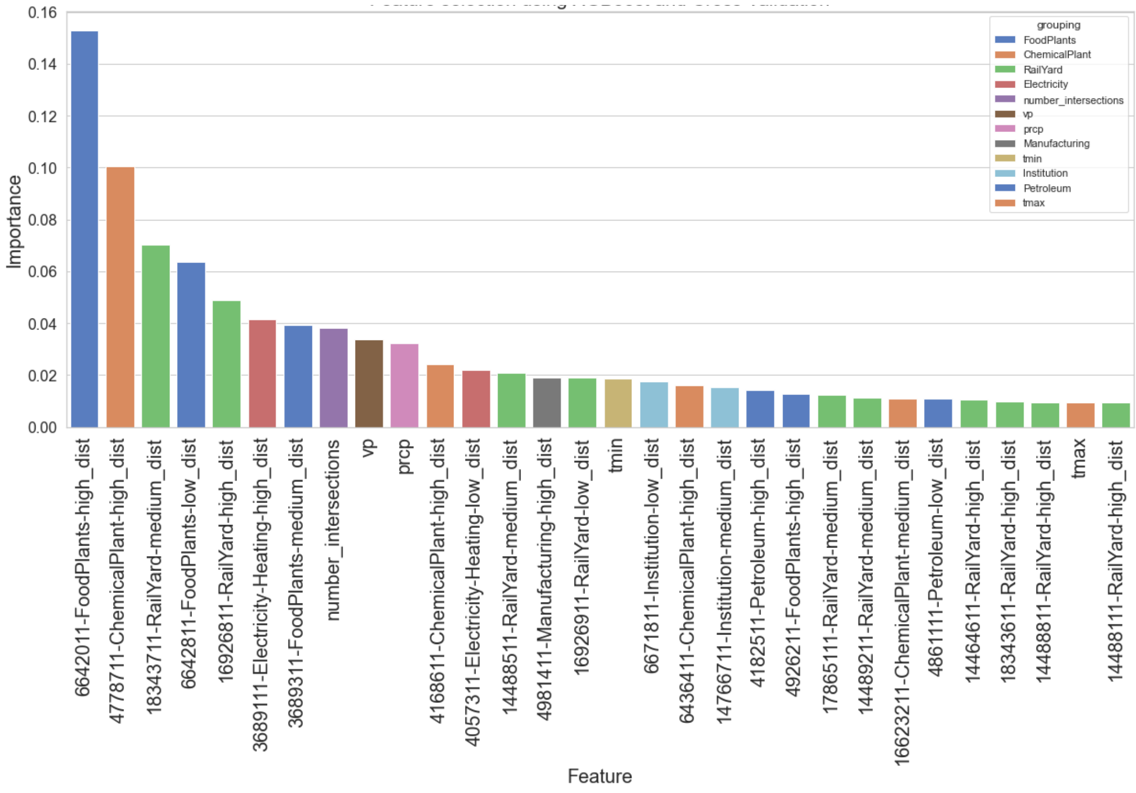


Figure A5: Feature selection using XGBoost


grouping
FoodPlants
ChemicalPlant
RailYard
Electricity
number_intersections
vp
prcp
Manufacturing
tmin
Institution
Petroleum
tmax
Importance
6642011-FoodPlants-high_dist
4778711-ChemicalPlant-high_dist
18343711-RailYard-medium_dist
6642811-FoodPlants-low_dist
16926811-RailYard-high_dist
3689111-Electricity-Heating-high_dist
3689311-FoodPlants-medium_dist
number_intersections
vp
prcp
4168611-ChemicalPlant-high_dist
4057311-Electricity-Heating-low_dist
14488511-RailYard-medium_dist
4981411-Manufacturing-high_dist
16926911-RailYard-low_dist
tmin
6671811-Institution-low_dist
6436411-ChemicalPlant-high_dist
14766711-Institution-medium_dist
4182511-Petroleum-high_dist
4926211-FoodPlants-high_dist
17865111-RailYard-medium_dist
14489211-RailYard-medium_dist
16623211-ChemicalPlant-medium_dist
4861111-Petroleum-low_dist
14464611-RailYard-high_dist
18343611-RailYard-high_dist
14488811-RailYard-high_dist
tmax
14488111-RailYard-high_dist
Feature