Air Quality Prediction in East Bay Area, CA

Building a Machine Learning Model for Air Quality Predictions

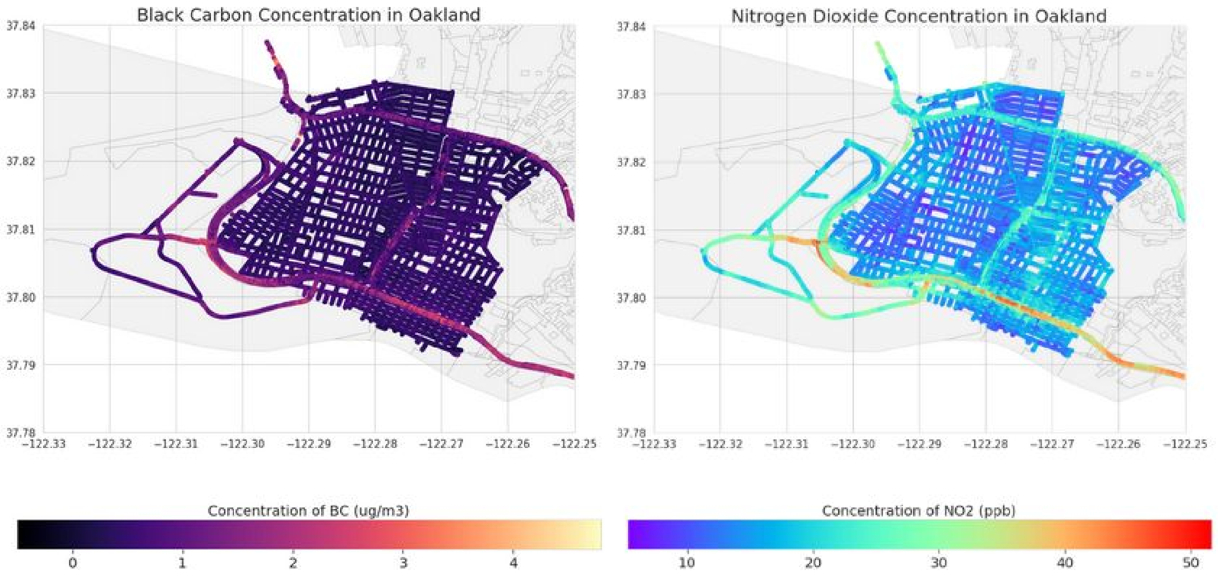

Black Carbon Concentration in Oakland
37.84
37.83
37.82
37.81
37.80
37.79
37.78
−122.33
−122.32
−122.31
−122.30
−122.29
−122.28
−122.27
−122.26
−122.25
Concentration of BC (ug/m3)
0
1
2
3
4
Nitrogen Dioxide Concentration in Oakland
Concentration of NO2 (ppb)
10
20
30
40
50

Table of Contents

Introduction and Problem Statement: 2

Audience: 2

Datasets: 2

Data Wrangling and Cleaning: 3

Exploratory Data Analysis: 12

Statistical Inference and Feature Engineering: 15

Machine Learning for Air Quality Predictions: 22

Introduction and Problem Statement:

In densely packed cities like Oakland or San Francisco, air quality can vary wildly across neighborhoods, due to varying sources of emissions. These include emissions from automobile traffic, industrial sources, local meteorological parameters, and marine vessels and other shipping sources. Current methods to monitor air quality, while certainly very useful, are either too widely spread out (sparsely located monitoring systems set up by local agencies) or too localized (personal air quality monitors that have a small radius of detection).

In this project, we try to answer the following question:

> Can we build a machine learning model to predict air quality per city-block in the City of Oakland and San Leandro, based on previously measured pollutant concentrations, local meteorological conditions, and local sources of emissions such as industries, traffic intersections and automobile traffic on highways, without having to rely on complex physical modeling?

Audience:

Monitoring pollutant concentration within cities is crucial for environmental management and public health experts to understand the impact of air quality and to promote sustainable cities. Apart from being a resource to the general public, this project can be useful to:

- To detect if there is an anomaly in air quality on a given block.
- To detect extreme events in air quality due to anomalies.
- Identify the sources that are contributing to air quality in a given region
- Social fitness apps like Strava, Fitbit, Nike Run who can leverage this data to recommend routes for users to run, bike, hike, etc

Datasets:

The primary dataset is the Oakland Air Pollution Monitoring Data measured by the Environmental Defense Fund (EDF)[1] [2]. Between June 2015 - May 2016, the Environmental Defense Fund (EDF) partnered with Google Earth Outreach and fit Google's cars with mobile sensing equipment to collect air quality measurements. The vehicles drove around the streets of Oakland and San Leandro and collected monitoring data for around 150 days. The dataset was then aggregated over a one year time period, and the estimate of median concentration was generated.

In addition to the above, we will use multiple sources of publicly available data such as:

[1] https://www.edf.org/airqualitymaps/oakland/mapping-pollution

[2] Apte, Joshua S., et al. "High-resolution air pollution mapping with Google street view cars: exploiting big data." *Environmental science & technology* 51.12 (2017): 6999-7008.

- Emissions data from individual sources obtained from the National Emissions Inventory[3]
- Traffic intersection count and distance to closest highway from Open Street Maps using Overpass API[4]
- Local meteorological data on a 1kmx1km grid obtained from Oak Ridge National Lab Daymet dataset[5]

Further details on the datasets are provided in the Project Proposal and Milestone Report.

Data Wrangling and Cleaning:

1) Oakland Air Pollution Monitoring Data

Figure 1 and Figure 2 contain maps of the Oakland Air Pollution Monitoring Data for Black Carbon (BC) and Nitrogen Dioxide (NO_2). The dots on the map represent the locations where the Google Street View cars collected air measurements, and they represent average concentrations measured between June 2015 - May 2016.

Figure 3 shows the histogram for BC and NO_2. The plots show that the data doesn't follow a normal distribution and the dataset is skewed to the right. Hence we need to transform the dataset using Box-cox (log transformation) to convert the data to follow a more normal distribution.

[3] https://www.epa.gov/air-emissions-inventories/2014-national-emissions-inventory-nei-data

[4] https://wiki.openstreetmap.org/wiki/Overpass_API/Overpass_API_by_Example

[5] https://daymet.ornl.gov/

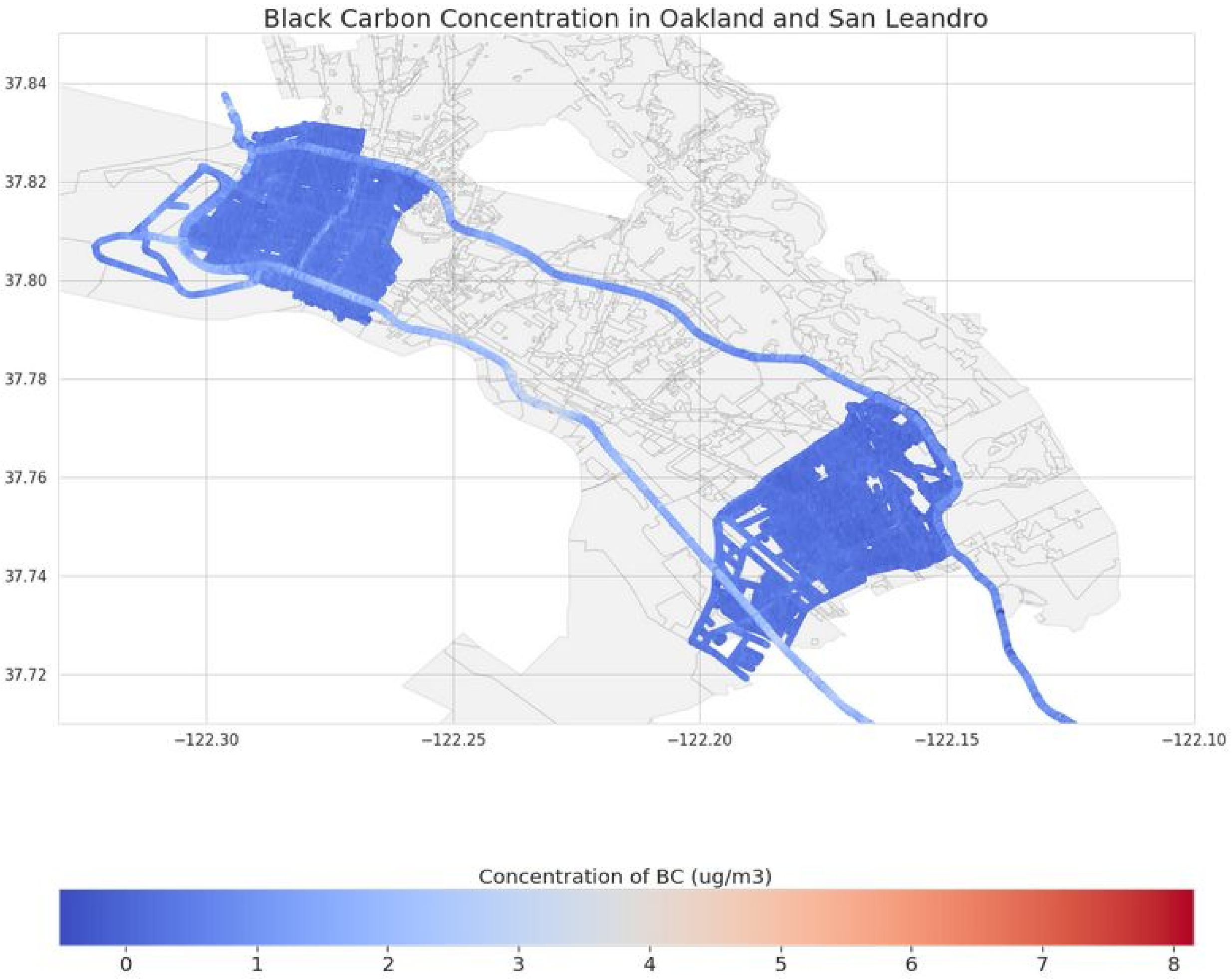
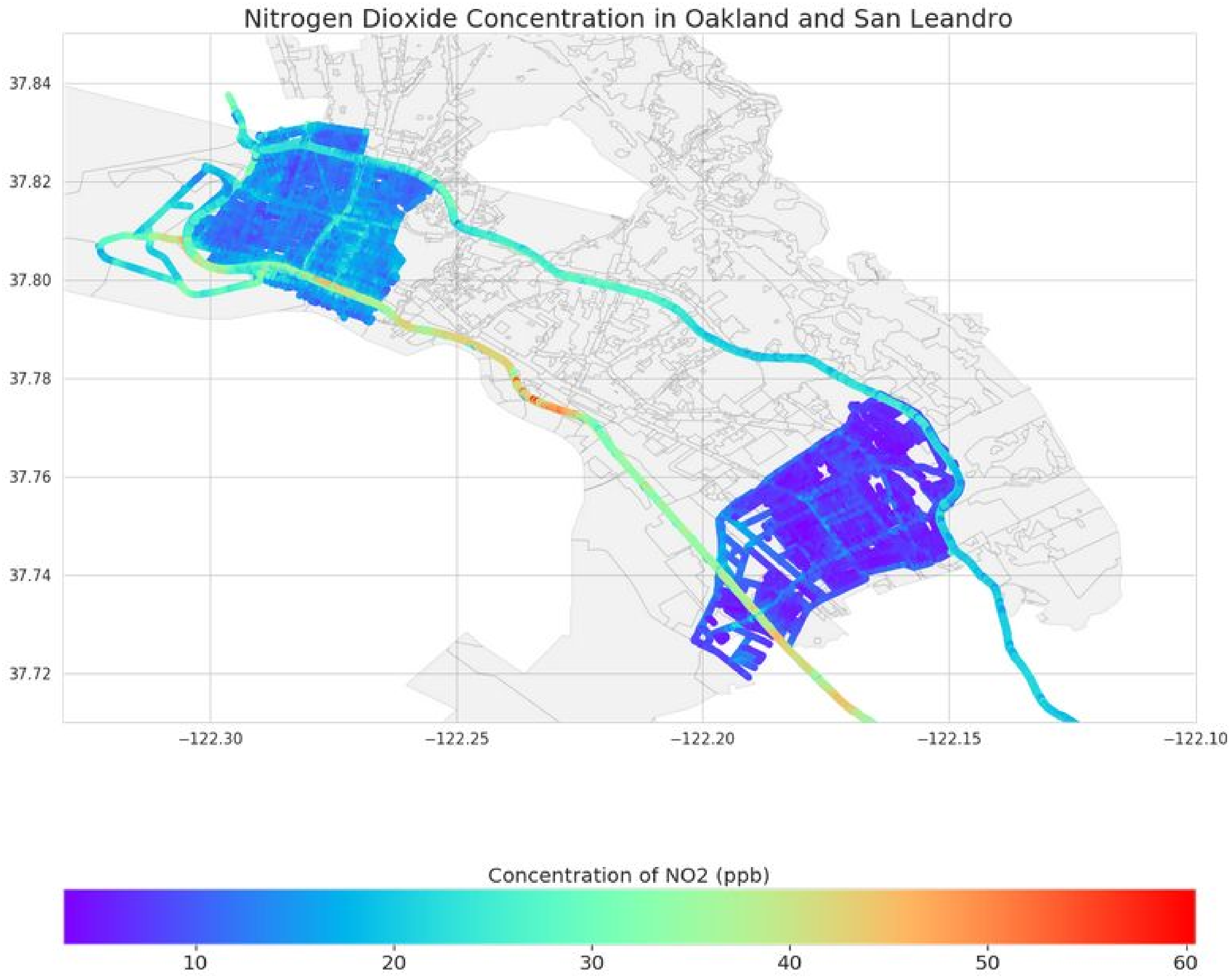

Black Carbon Concentration in Oakland and San Leandro
37.84
37.82
37.80
37.78
37.76
37.74
37.72
−122.30
−122.25
−122.20
−122.15
−122.10
Concentration of BC (ug/m3)
0
1
2
3
4
5
6
7
8

Figure 1: Concentration of Black Carbon Measured at Different locations



Nitrogen Dioxide Concentration in Oakland and San Leandro
37.84
37.82
37.80
37.78
37.76
37.74
37.72
−122.30
−122.25
−122.20
−122.15
−122.10
Concentration of NO2 (ppb)
10
20
30
40
50
60

Figure 2: Concentration of Nitrogen Dioxide Measured at Different locations

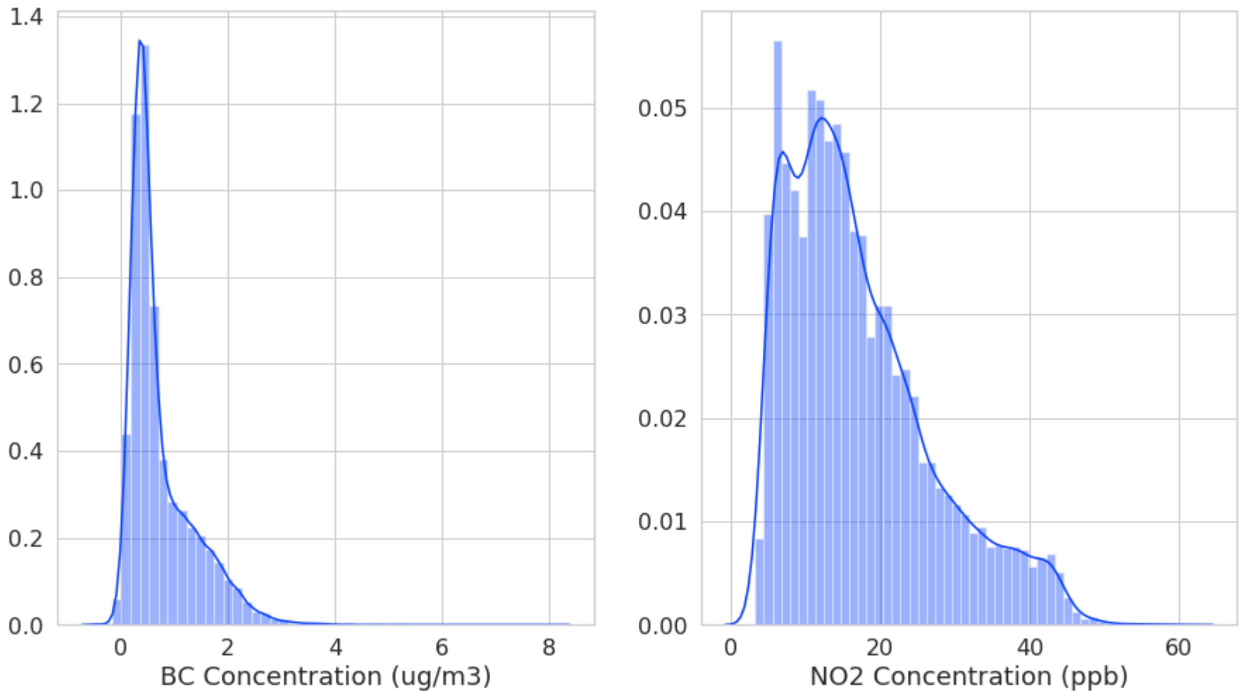


Figure 3: Histogram of BC and NO_2 datasets

The transformed data is shown below in *Figure 4*.

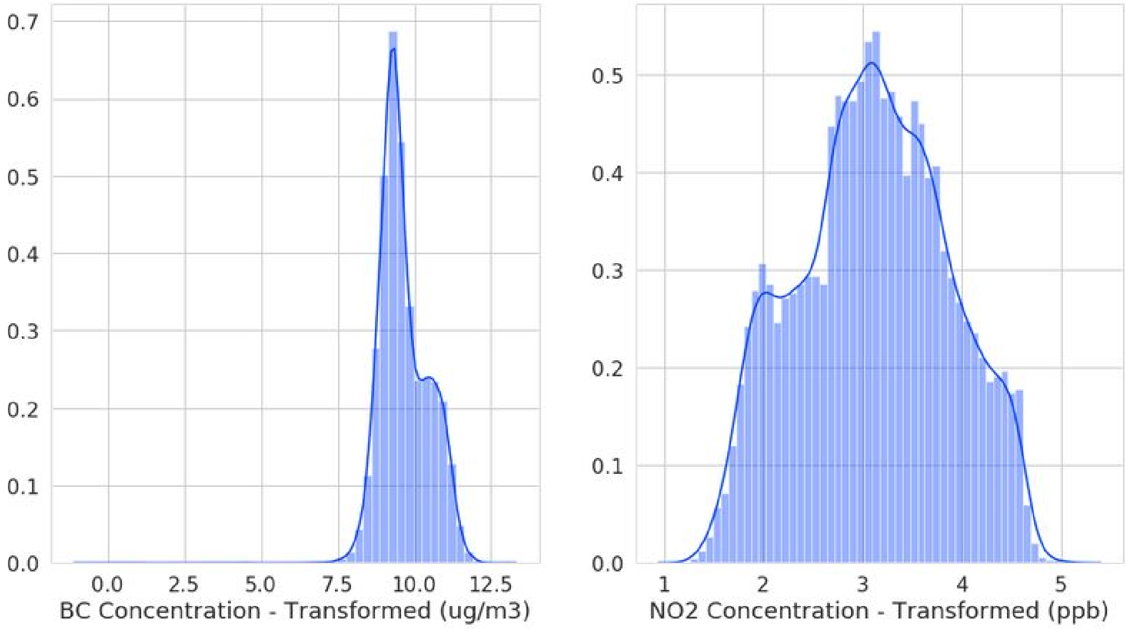


Figure 4: Histogram of BC and NO2 datasets after Box-cox transformation

2) National Emissions Inventory

The National Emissions Inventory contains $PM2_{.5}$ and NO_2 emissions from different stationary sources (or facilities) in the Oakland-San Leandro neighborhoods including Berkeley, Alameda, Oakland and San Leandro. Here, $PM_{2.5}$ was used as a proxy for BC since no data was available. Emissions of BC occur mostly from partial combustion of diesel from engines used in transportation and industries, residential fuel use such as wood and coal and from small boilers.

The dataset contained several facilities that were categorized into several different types but relatively few sources in most of them. In order to minimize the number of categories, sources into some categories were categorized into larger sources such as 'manufacturing', 'transportation' etc. *Figures 5 and 6* contain maps of all the sources that emit $PM_{2.5}$ and NO_2, respectively.

Combining the sources into larger groups resulted in a boxplots as shown in *Figures 7 and 8*. By observing the boxplot, we can see that there are some very large sources of emissions and some relatively smaller sources. The sources were further categorized into 'low','medium' and 'high' depending on the emission quantiles. Low indicates that emissions are lower than first quantile, medium indicates emissions is between first and third quartile, and high indicates emission is above third quartile.

3) Traffic data - number of intersection counts and proximity to highway

Traffic data was obtained from Open Street Maps using the Overpass API. The total number of traffic signals/intersections within a 1,000 ft from each monitoring point was obtained using an API call. The number of traffic intersections varied from 2 intersections to 35 intersections. Proximity or distance to the closest highway from each monitoring point was also obtained using the Overpass API. Since some of points with 'highest' concentrations were observed on highways, this was chosen as one of the features.

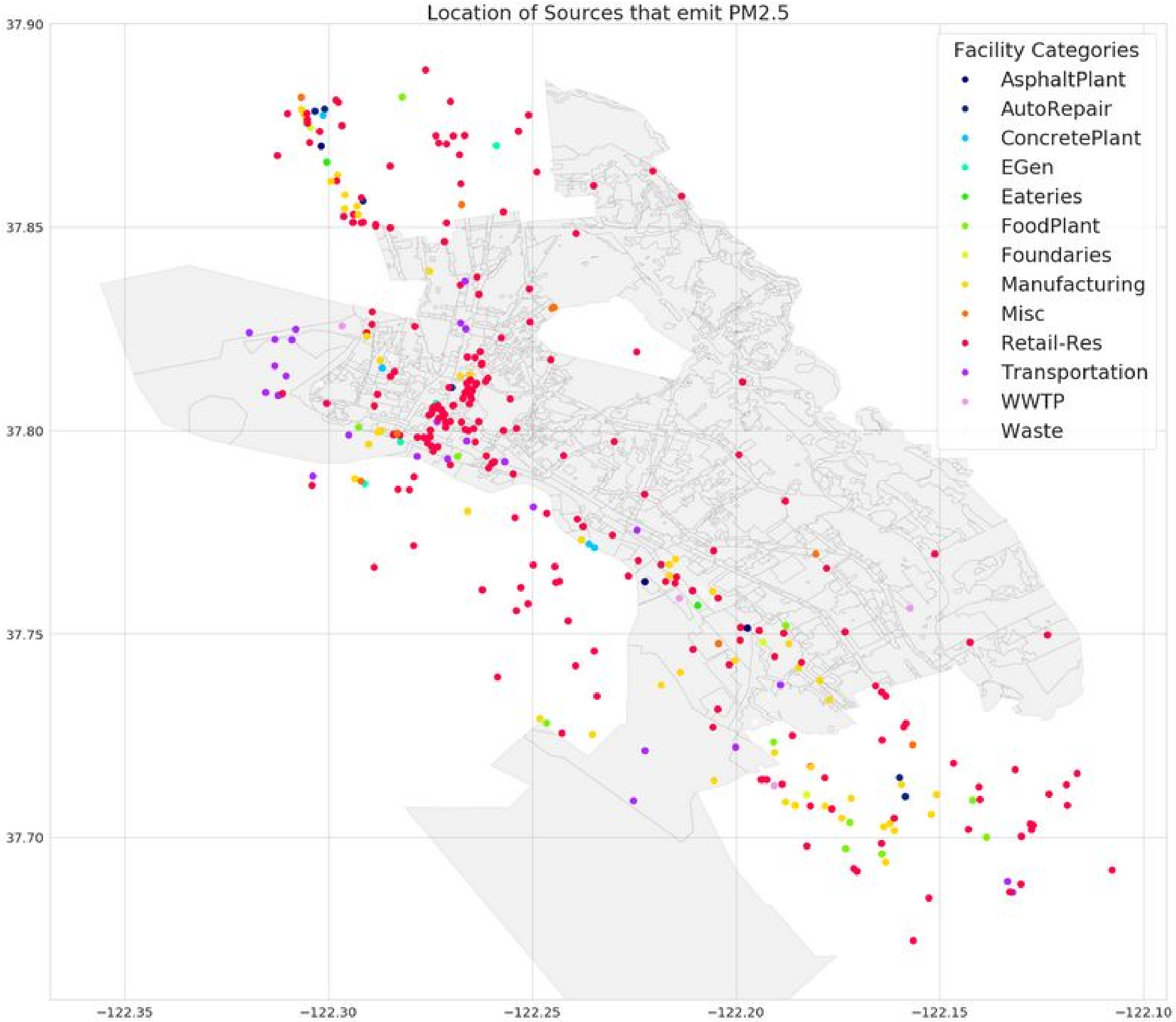


Figure 5: Location of Sources that Emit $PM_{2.5}$

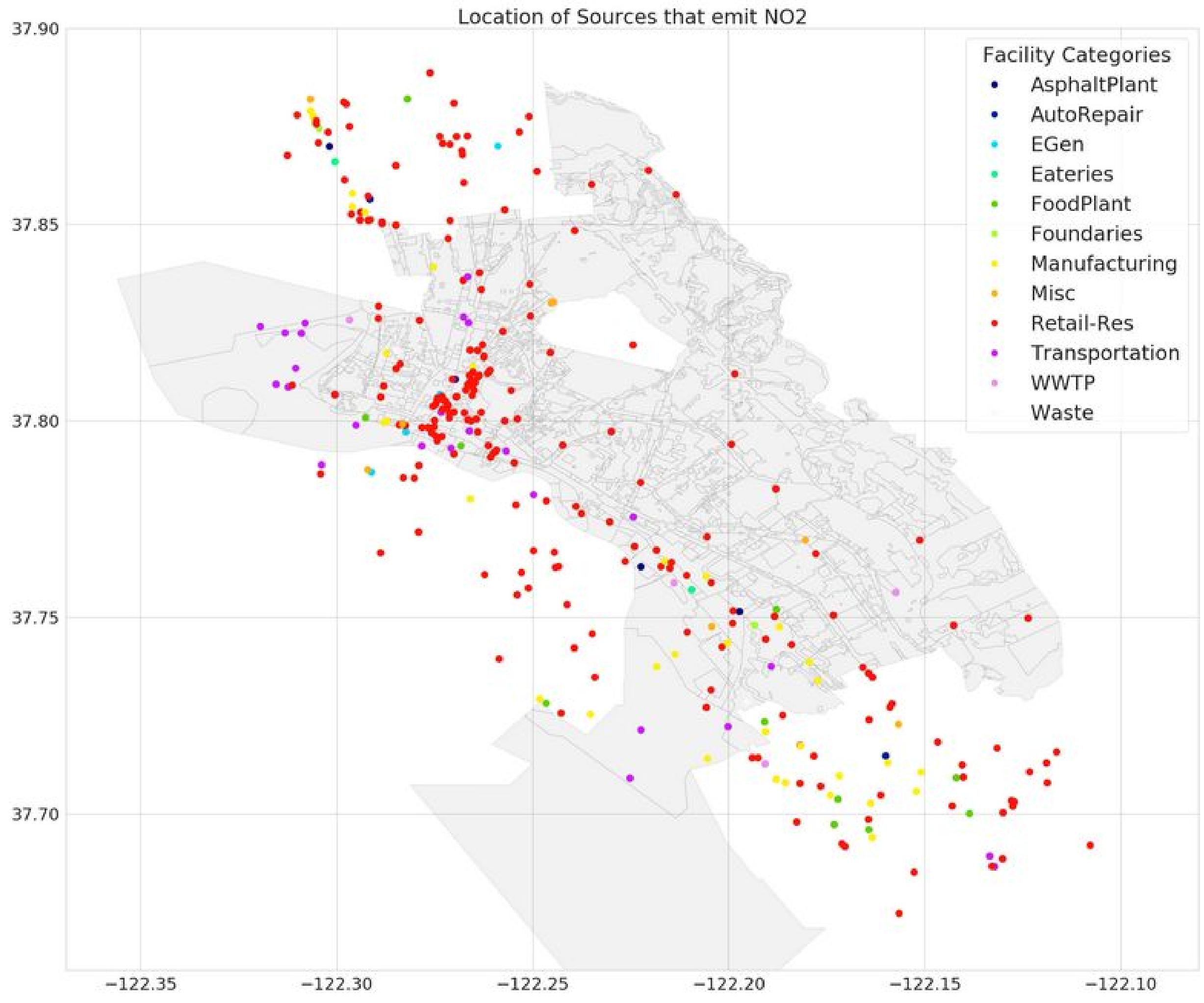


Figure 6: Location of Sources that Emit NO_2

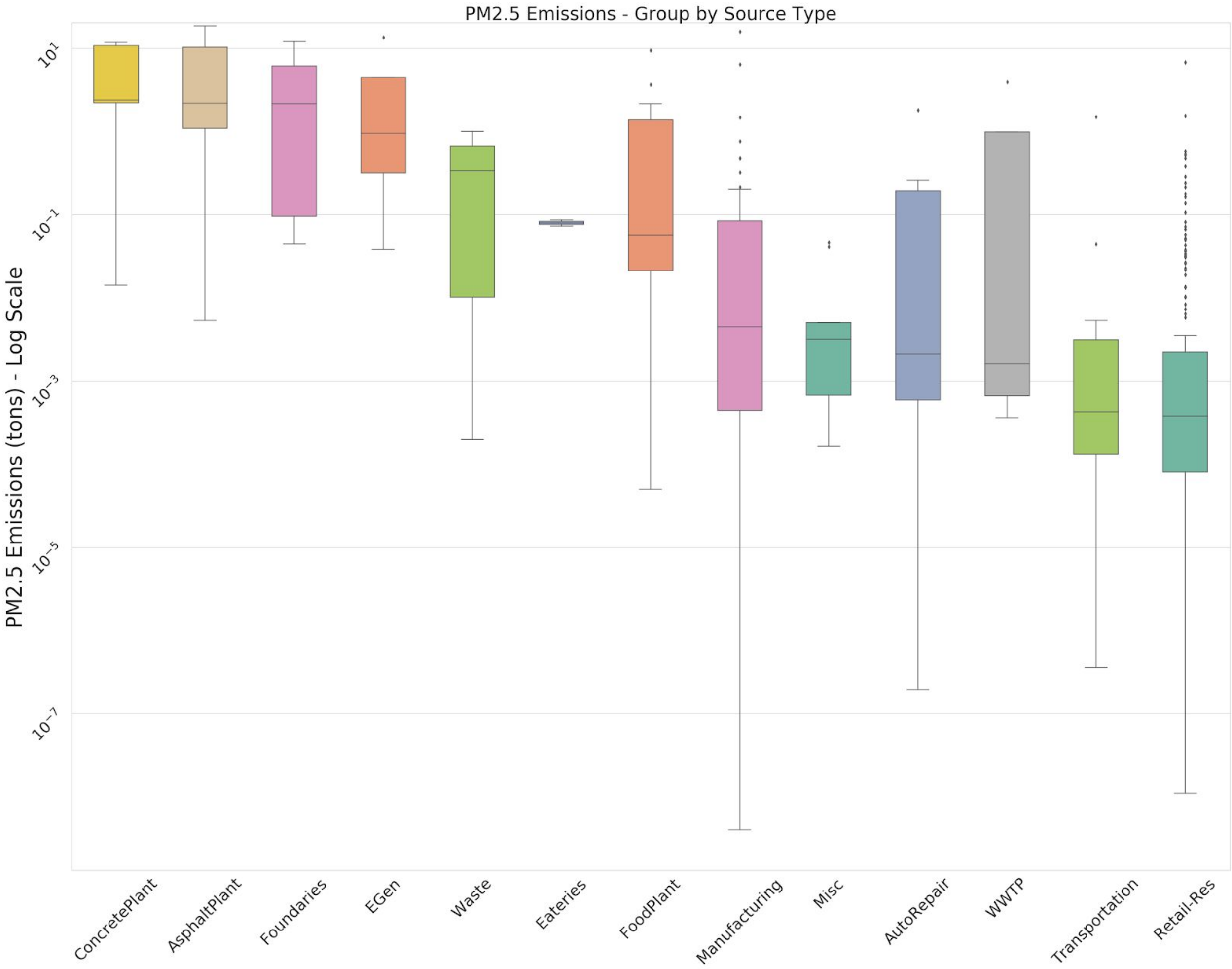


Figure 7: Boxplot of $PM_{2.5}$ emissions grouped by source types - log scale

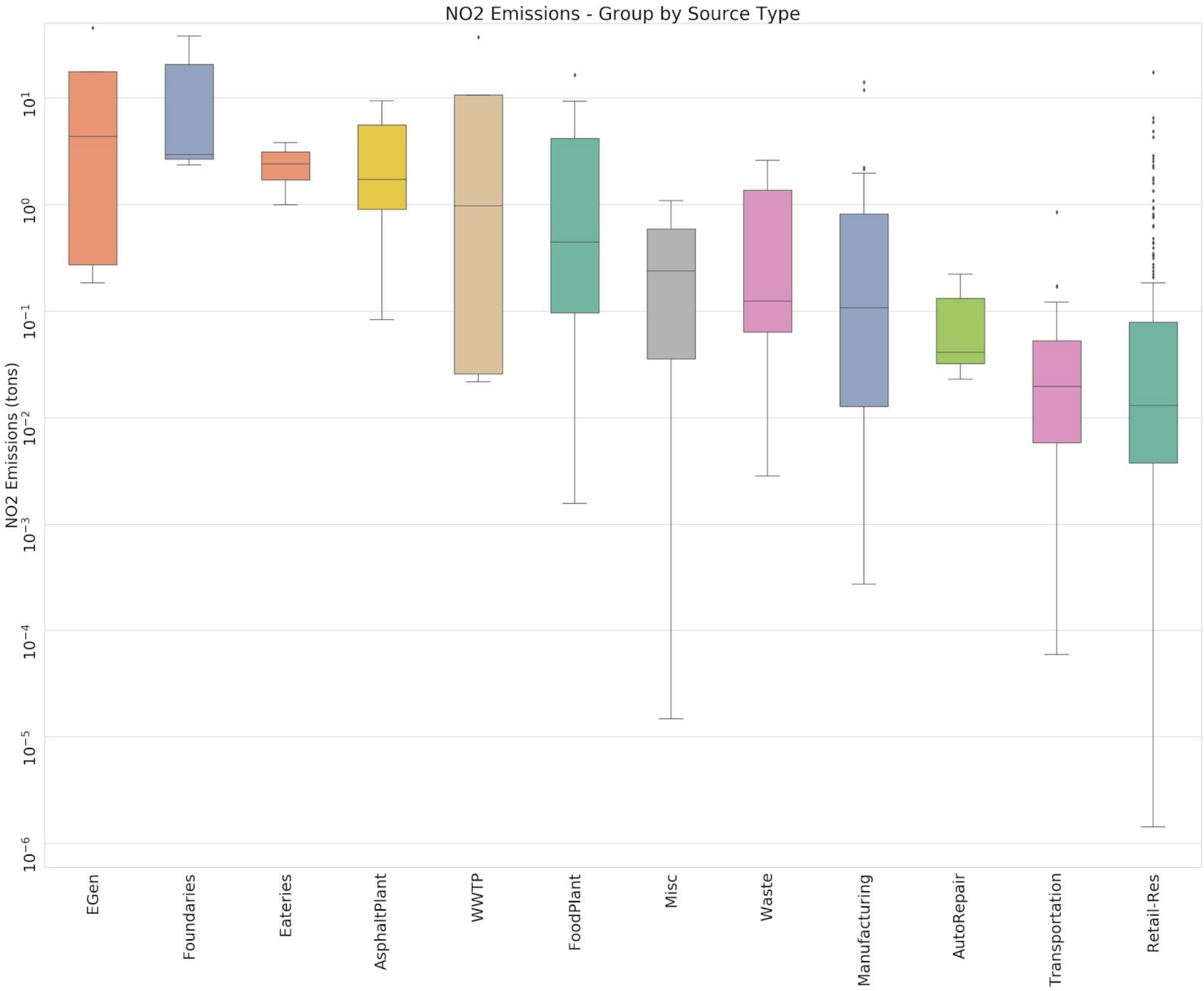


Figure 8: Boxplot of NO_2 emissions grouped by source types - log scale

4) Meteorological Data

Annual average meteorological parameters from June 2015 - May 2016 were obtained from Oak Ridge National Lab's Daymet dataset. The dataset contains daily meteorological data on a 1km by 1km grid basis. The daily measurements were averaged to a time period between June 2015 - May 2016. Parameters include Precipitation, Radiation, Minimum and Maximum Temperature, and Pressure

Exploratory Data Analysis:

At a high level, the question we are trying to answer is “how does air pollution in a given location correlate with sources such as industries, traffic and meteorological parameters?”. Further, “Can we build an air quality prediction model that relies on publicly available datasets to get an accurate prediction of air quality per city block?”.

The hypothesis tree is shown below in *Figure 9*.

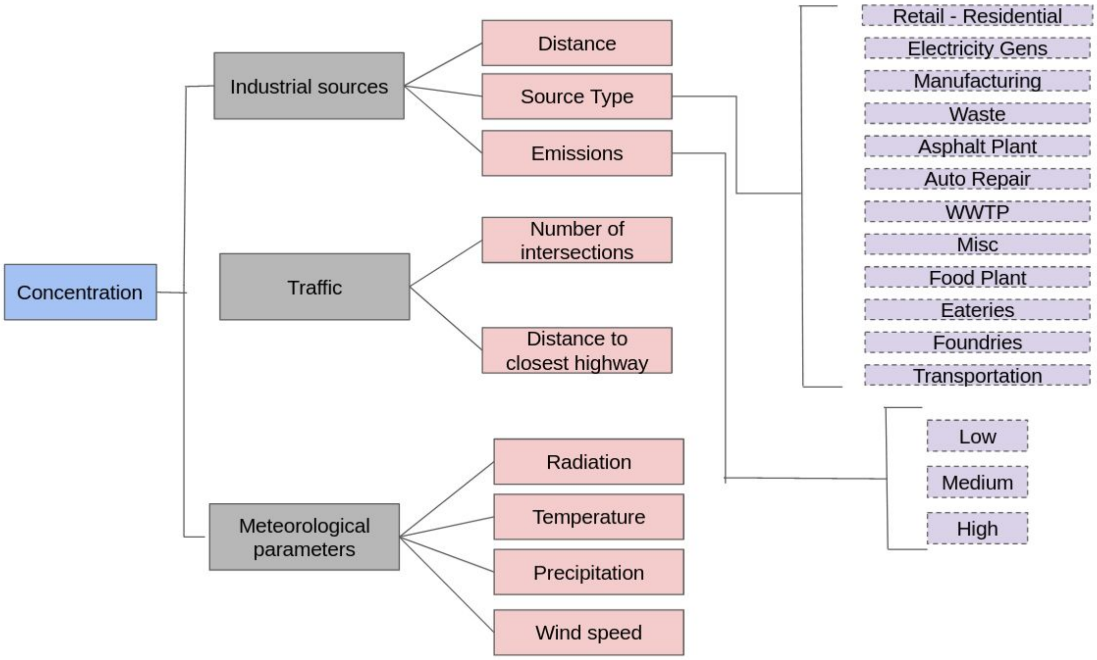


Figure 9: Hypothesis Tree

As a part of the exploratory data analysis, some of the questions that we are trying to answer are listed below:

1. **How does air pollution concentration vary with distance to sources of emissions?**
2. ***2.*** **Is concentration directly correlation with distance, or to (distance)^2 or to emissions/distance?**
3. **Is there a correlation between concentration and number of traffic intersections?**
4. **Is there a correlation between concentration and distance to closest highway?**
5. **Is there a correlation between concentration and meteorological parameters?**

For the BC dataset, *Figure 10* shows the correlation between concentration and distance to closest highway, when all the points are combined together. However, on further observation, we notice that there are two distinct regions or clusters that arise. Separating the correlations into two clusters, below and above 6 km, we observe a stronger positive correlation for the two regions as shown in *Figure 11*.

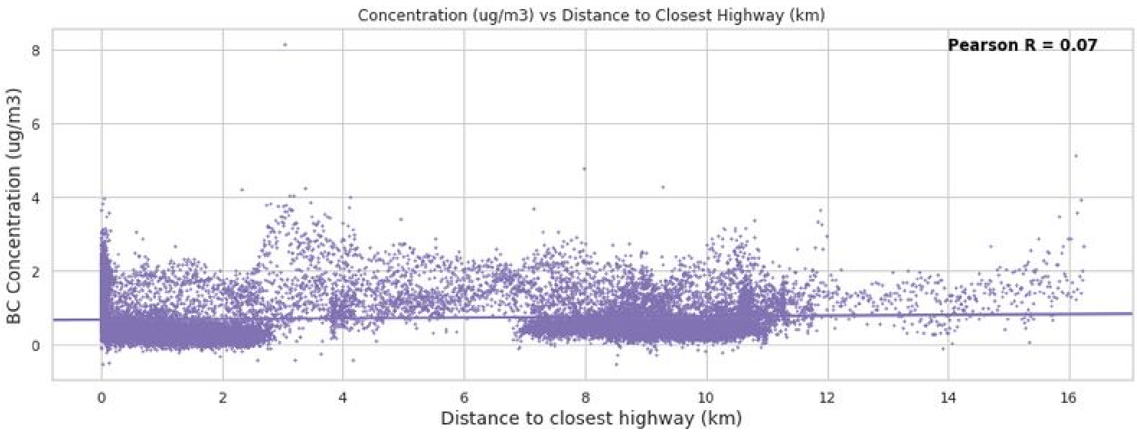


Figure 10: Scatter plot of BC concentration vs distance to closest highway when all points are combined

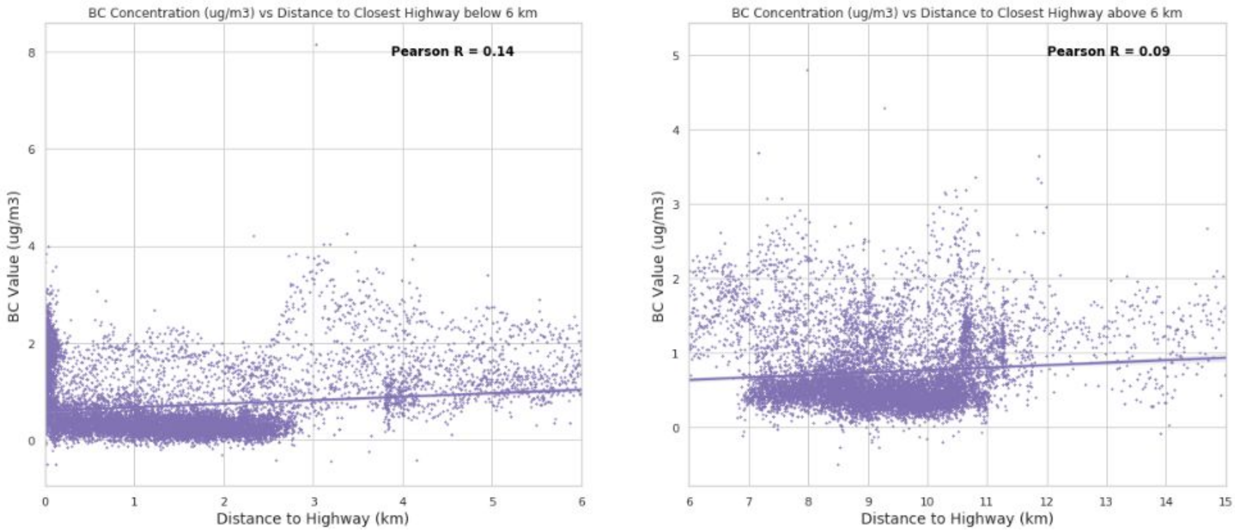


Figure 11: Scatter plot of BC concentration vs distance to closest highway split into two clusters

Here, points with distance = 0 are points on the highway and the concentrations for these points are high, indicating that some of the highest BC concentrations are measured on the highway. Points with distance < 6km have a slightly larger positive correlation with distance to highway. These could be points on on and off ramps, and points very close to highways. The lower correlation with increasing distance indicates that BC concentration tends to reduce with distance from highway.

For the NO2 dataset, *Figure 12* shows that there is only a low correlation between concentration and distance to closest highway for NO2, when all the points are combined together. Separating the correlations into two clusters, below and above 6 km, we observe that correlation coefficient reduces slightly for the points < 6km, and drops to almost zero as the distance increases as shown in *Figure 13*.

Once again, points with distance = 0 are the points on the highway and the concentrations for these points are high, indicating that some of the highest NO2 concentrations are measured on the highway. Points with distance < 6km have a slightly larger positive correlation with distance to highway. These could be points on on and off ramps, and points close to highways. The close to zero correlation with increasing distance indicates that NO2 concentration tends to reduce with distance from highway.

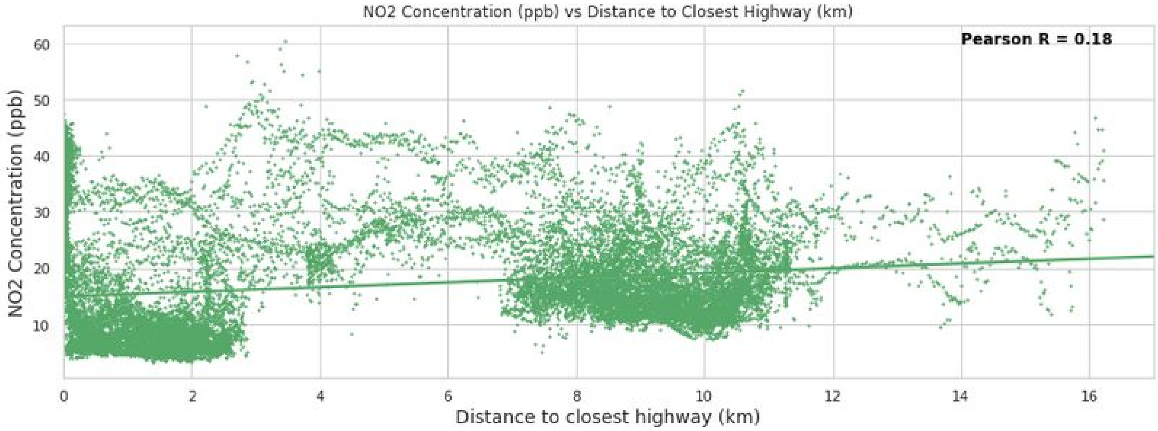


Figure 12: Scatter plot of NO_2 concentration vs distance to closest highway when all points are combined

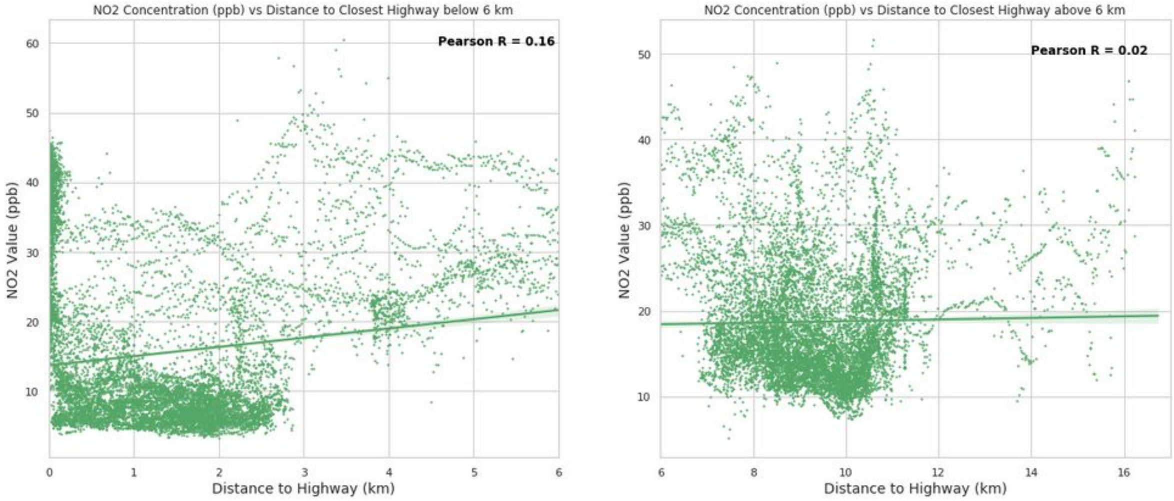


Figure 13: Scatter plot of NO_2 concentration vs distance to closest highway when all points are combined

Exploring the correlation between some of the meteorological variables and concentrations, we notice that there is a low negative correlation between precipitation and concentrations as shown in *Figure 14*. This could indicate that precipitation may be an important feature while building the machine learning model for prediction.

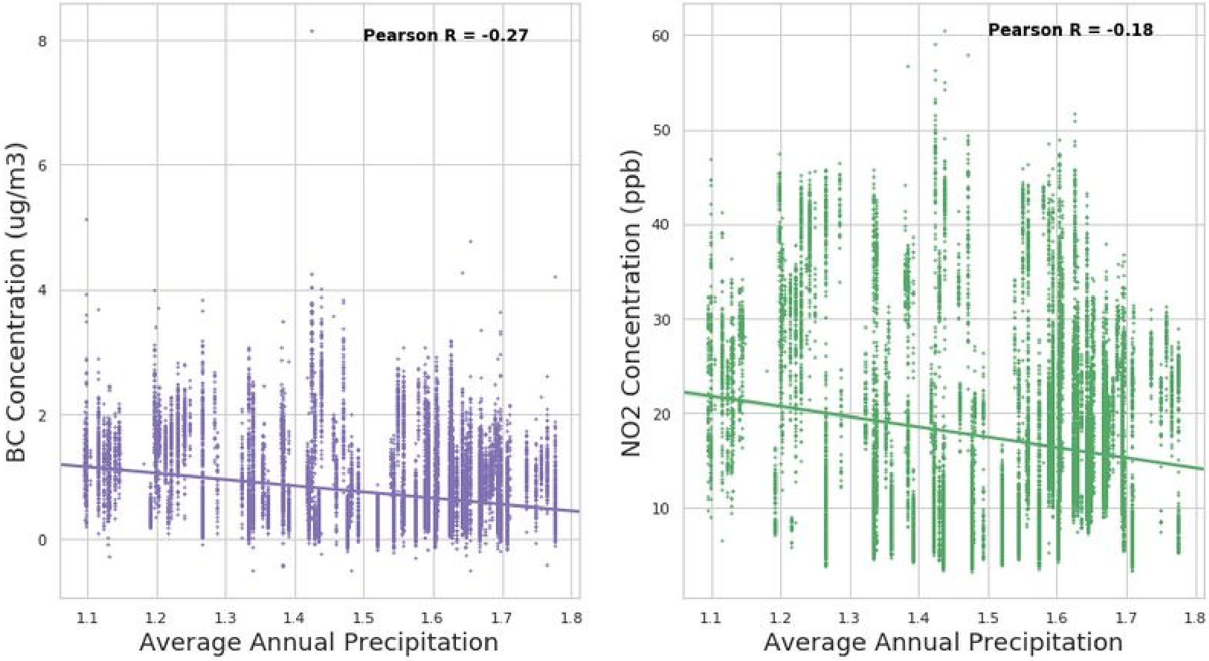


Figure 14: Scatter plot of concentrations vs annual precipitation

Further details on the exploratory data analysis can be found in the data story report.

Statistical Inference and Feature Engineering:

Based on the exploratory data analysis, the input features to the machine learning model includes the following:

1) Distance to industrial sources from all monitoring locations. Facilities are classified as low, medium and high, along with their source category classification.
2) Number of traffic intersections within 1,000 ft of each monitoring location
3) Distance to the closest highway from each monitoring location
4) Meteorological parameters at each monitoring location including precipitation, radiation, minimum and maximum temperature, and pressure

In total, there are 355 features in the BC dataset and 324 features in the NO2 dataset.

Resolving Multicollinearity

A correlation matrix of features in the BC and NO2 datasets revealed that multicollinearity is a major issue in this dataset. Several features had a correlation coefficient as high as 0.99. It is important to eliminate features that are multicollinear because multicollinearity can undermine the statistical significance of an independent variable. While multicollinearity does not necessarily affect a model's predictive accuracy, it affects the variance associated with the prediction, as well as, reduces the quality of interpretation of independent variables i.e. effect of the data on the model isn't trustworthy

To resolve multicollinearity and understand feature importance, several methods were explored including:

1) Identifying and dropping features that had a correlation coefficient above 0.9
2) Step forward feature selection with Variance Inflation Factor (VIF) and R2 estimation
3) Lasso regularization with step forward feature selection

Further details on the first approach is available in the milestone report.

The algorithm shown in *Figure 15* was used to eliminate features that are multicollinear and understand feature importance.

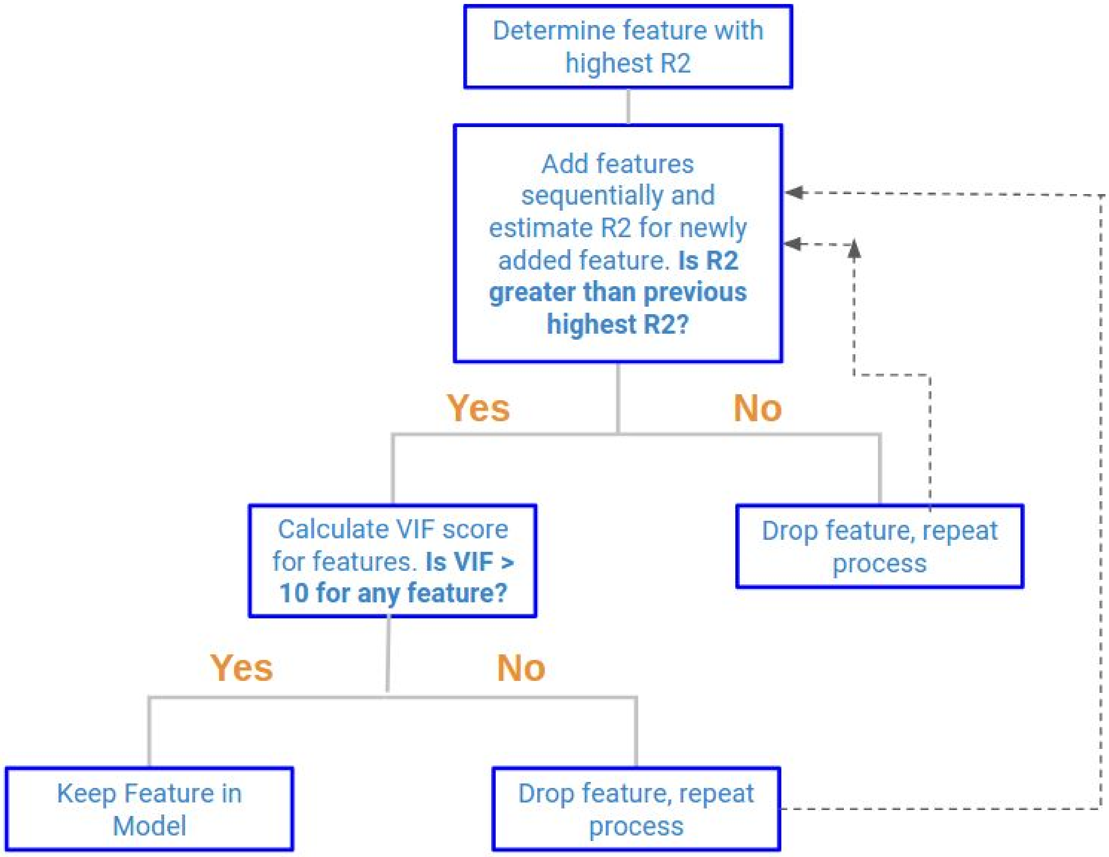

Determine feature with highest R2
Add features sequentially and estimate R2 for newly added feature. **Is R2 greater than previous highest R2?**
Yes
No
Calculate VIF score for features. **Is VIF > 10 for any feature?**
Drop feature, repeat process
Yes
No
Keep Feature in Model
Drop feature, repeat process

Figure 15: Algorithm for step forward feature selection with VIF and R2 estimation

For the BC and NO2 datasets, only ‘Radiation’ was selected as a feature using the step forward approach, with an R2 of 0.584 for BC and 0.757 for NO2.

Next, a lasso regularization method was applied for feature selection where the coefficient of certain features that did not have an impact on the target variable were penalized to zero. To eliminate multicollinearity, the step forward selection approach discussed above was applied to the features selected by the Lasso model. Lasso regularization selected 171 features out of the 355 features for the BC dataset and 137 features out of 324 features for the NO2 dataset. After applying the step-forward VIF approach, only ‘Minimum Temperature’ was selected as a feature.

Random Forest for Feature Importance

Since the previous methods did not give us too much insight into the important features, a Random forest with cross validation method was used for feature selection. Random forests are commonly used for feature selection because the tree-based models naturally rank features by how well they improve the purity of the node. The BC and NO2 datasets were split into test/train data and a 4-fold cross validation approach was applied on the training dataset to select the features. For each fold, feature importance was calculated and an average of all feature scores was used to determine feature importance in order to rank features. *Figure 16* and *Figure 17* show the feature importance for the BC and NO2 datasets, respectively.

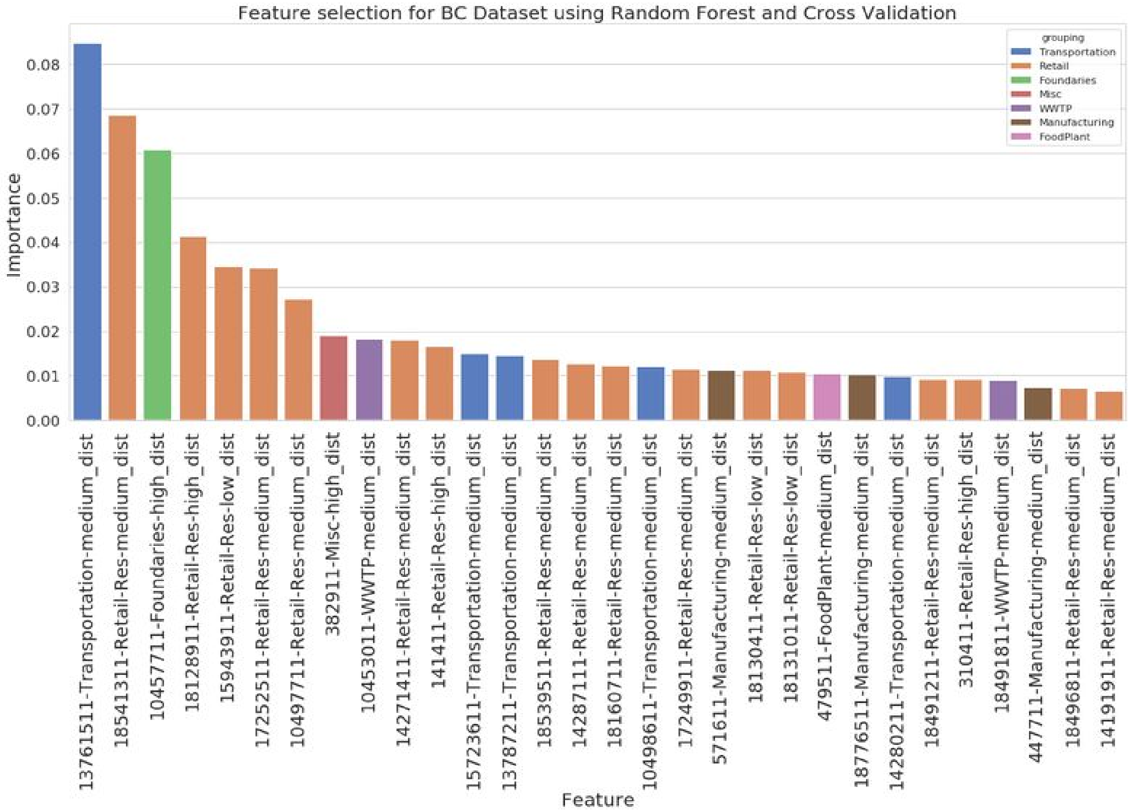


Figure 16: Feature importance for BC dataset

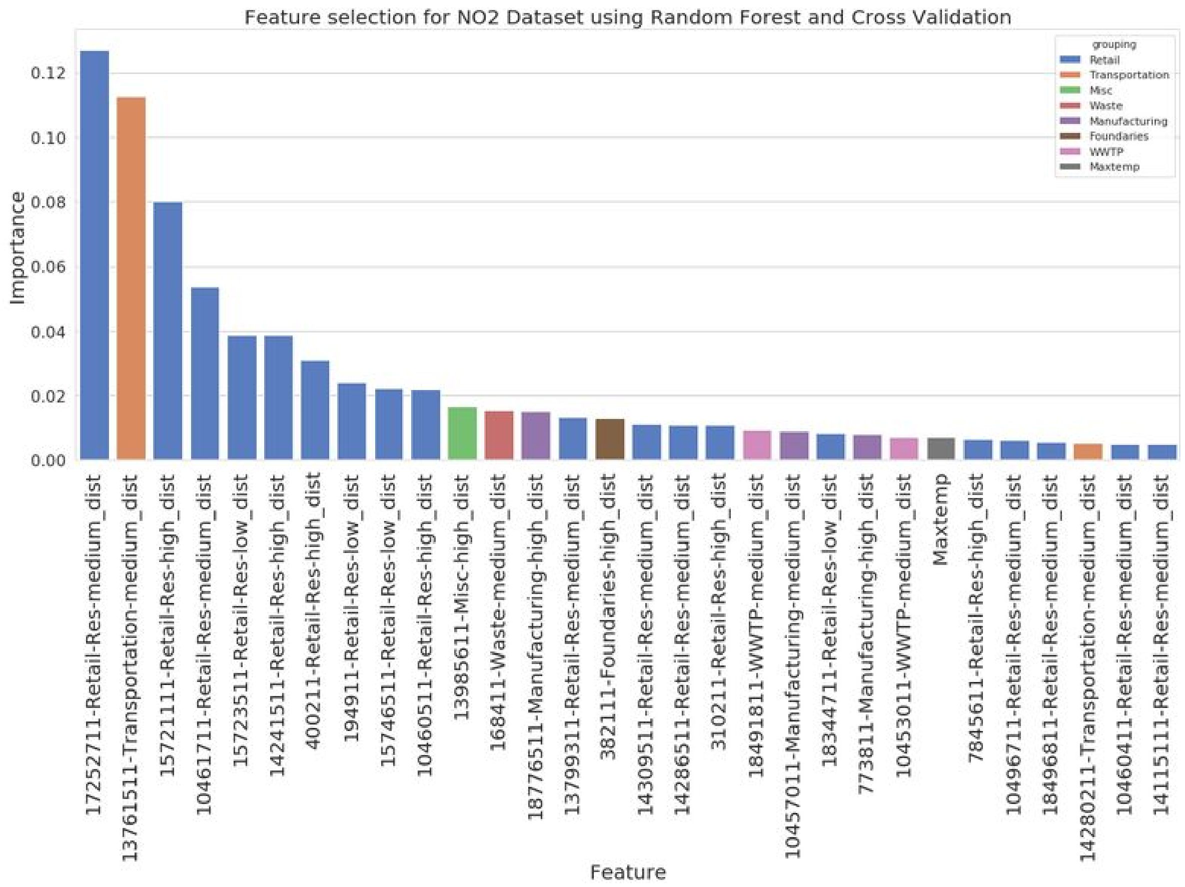


Figure 17: Feature importance for NO_2 dataset

Figure 18 and Figure 19 show the location of the top 100 features from the Random Forest model for the BC and NO2 dataset, respectively. The features shown in these figures only include the location of the important features, in this case location of facilities that contribute most to air quality in the region. The size of the dots indicate the feature importance, with larger dots indicating higher importance. The color of the points show the sector to which each facility belongs.

For the BC dataset, the top five features that contribute to concentration in the region includes the Oakland International Airport, a medical/outpatient unit, the Ridge Foundry in San Leandro, a retail unit in Berkeley and a large residential complex in Oakland.

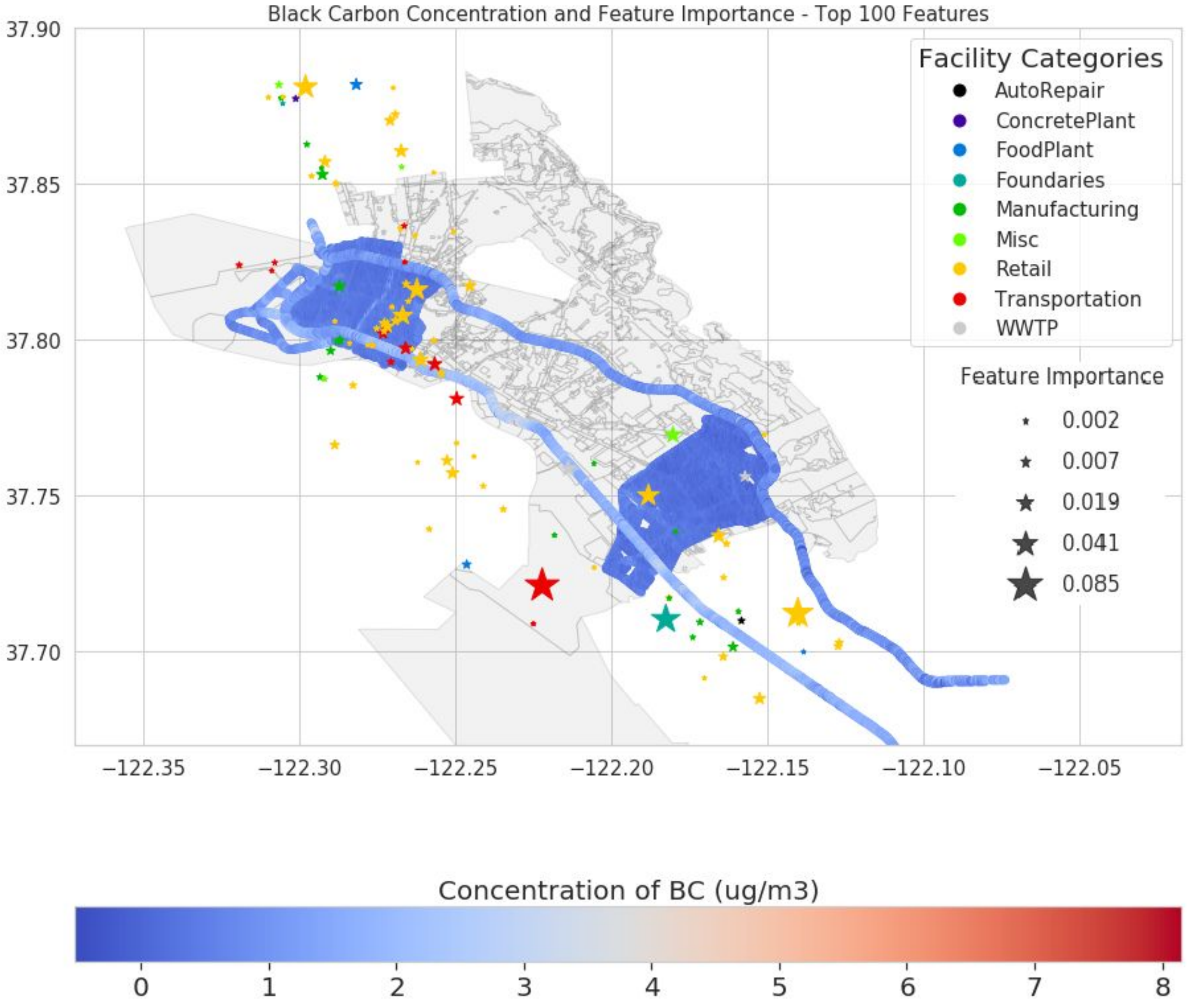


Figure 18: Location of top 100 features for predicting BC concentration

For the NO2 dataset, the top five features that contribute to concentration in the region include a commercial complex (KTUV Fox 2 office), the Oakland International Airport, the Kindred hospital in San Leandro which may have generators and boilers, Digital Realty Data Center in Oakland which has a lot of generators and the California Supreme Court of Alameda, which may also have a lot of generators.

While it's hard to view this on the map, the BC concentration at some points close to the facilities that are important are high indicating that these sources contribute to the concentration in that area. Similarly, for NO2, concentrations close to the ‘Retail’ facilities are high indicating that these facilities contribute to the concentration in the area.

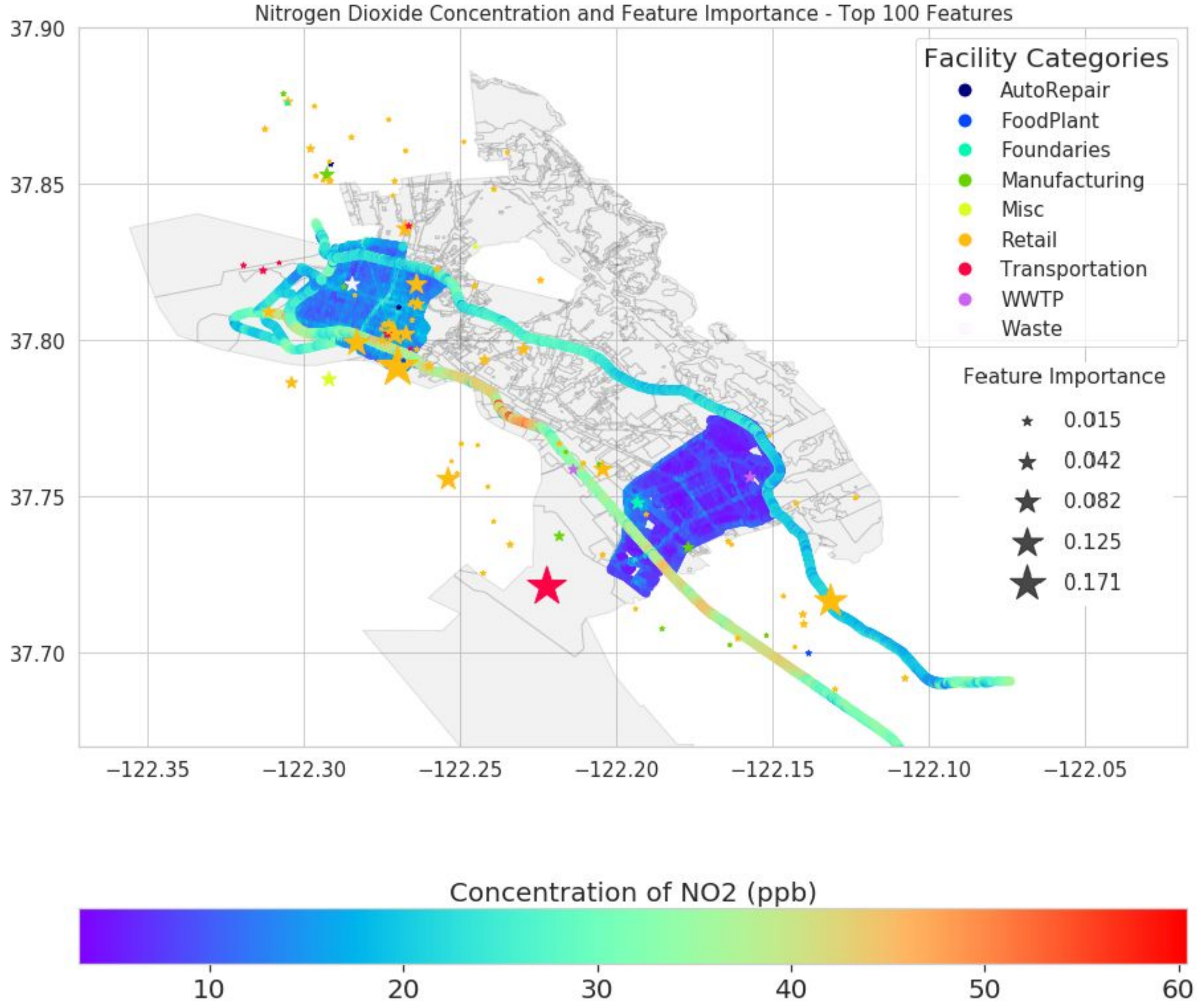


Figure 19: Location of top 100 features for predicting NO2 concentration

The features listed above are only the top five features in terms of feature importance. However, there are more features shown in the graphs in Figure 16 and 17 that actually contribute to concentration of BC/NO2 near the areas where they are high. It is also important to note that most of the facilities that have a high feature importance are closely clustered around the I-880 highway, which has some of the highest concentration of BC and NO2.

Even though feature importance gives an insight into the features that contribute to concentration in the region, one of the challenges of this particular dataset has been the lack of data on emissions from traffic in the region. In other words, the lack of inclusion of emissions from traffic as a feature in the dataset has resulted in some challenges. Since traffic contributes highest to black carbon concentrations (which comes from partial combustion of fuel) and nitrogen dioxide concentrations, this analysis still does not give us a full picture on the major sources that contribute to air pollution in an area. Besides, the concentration in each location is so hyper-localized that it's hard to identify one single source that contributes to concentration in the entire region.

Machine Learning for Air Quality Predictions:

Prior to building machine learning models, the BC and NO2 datasets were first split into training and test data. The training dataset was used to train and evaluate the model, while the testing dataset was used as the final evaluation.

Several different machine learning models were built and tested for its predictive performance. Some of the most important models that were built and tested include:

1) **Linear Regression** - Predicts a dependent variable value (y) based on a set of given independent variables (x1, x2.. xn). The linear between xs (input) and y(output) is written as $y = \Theta 0 + \Theta 1x1 + \Theta 2x2.....\Theta nXn$
2) **Ridge Regression** - Ridge regression is very similar to linear regression (sum of squares) except that a small amount of bias is introduced. In return, we get a significant drop in variance. In other words, by starting with a slightly worse fit, Ridge Regression can provide better long term predictions. A ridge regression with cross-validation selects the model parameters by performing cross validation and selecting the best parameters for the fit.
3) **Bagging** - Bootstrap Aggregation (or Bagging for short), is a simple and very powerful ensemble method. Bootstrap Aggregation is a general procedure which involves the application of Bootstrapping to reduce the variance for the algorithms that have high variance such as decision trees.
4) **Random Forest** - Random Forests are an improvement over bagged decision trees. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.
5) **XGBoost** - XGBoost is an ensemble learning method. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance, or in other words it combines bagging and boosting methods.

Principal Component Analysis

Prior to building and training machine learning models, a Principal Component Analysis (PCA) is usually performed to reduce the dimensionality of features and to understand whether application of PCA improves the model performance or not. Upon further investigation, the benefits of PCA was found to be relatively low and the increase in R2 value with PCA was not significant. Since PCA is a computationally intensive process, especially while using Random Forest or XGBoost, PCA transformation is not applied in this work.

GridSearching Different Models for Comparison

A gridsearch and cross validation was performed on the different models based to identify the best parameters (such as optimal tree depth, number of estimators etc.) for each model as well as to achieve the best accuracy score. Five different models were built and tested including a simple linear regression model, ridge regression, elastic net, random forest and XGBoost regression.

BC Dataset

Model parameters, R2 score and Root Mean Squared Error (RMSE) for the BC dataset are shown below. *Figure 20* shows a scatter plot of the predicted vs actual concentrations for each model. The best parameter of alpha = 0 for the elastic net and ridge regression is a strange occurrence here - this indicates that the ridge regression converges to an OLS. This is likely due to the test and training data being very similar. The Random Forest model also results in a max depth of 15 which is likely an overfitting issue.

Model Type	Best Parameter	Mean CV score for best estimator	Root Mean Squared Error	Training R2	Test R2
Linear Regression without PCA		0.671	0.352	0.692	0.674
Ridge regression without PCA	Alpha = 0	0.671	0.352	0.692	0.674
ElasticNet without PCA	Alpha = 0 L1_ratio = 0.1	0.549	0.418	0.553	0.541
Random Forest without PCA	Max depth = 15 Estimators = 800	0.809	0.279	0.926	0.795
XGBRegressor without PCA	Learning rate = 0.1 max depth = 10	0.818	0.269	0.930	0.81

	estimators = 100				

Table 3: Model parameters, R2 score and RMSE for different models for BC dataset

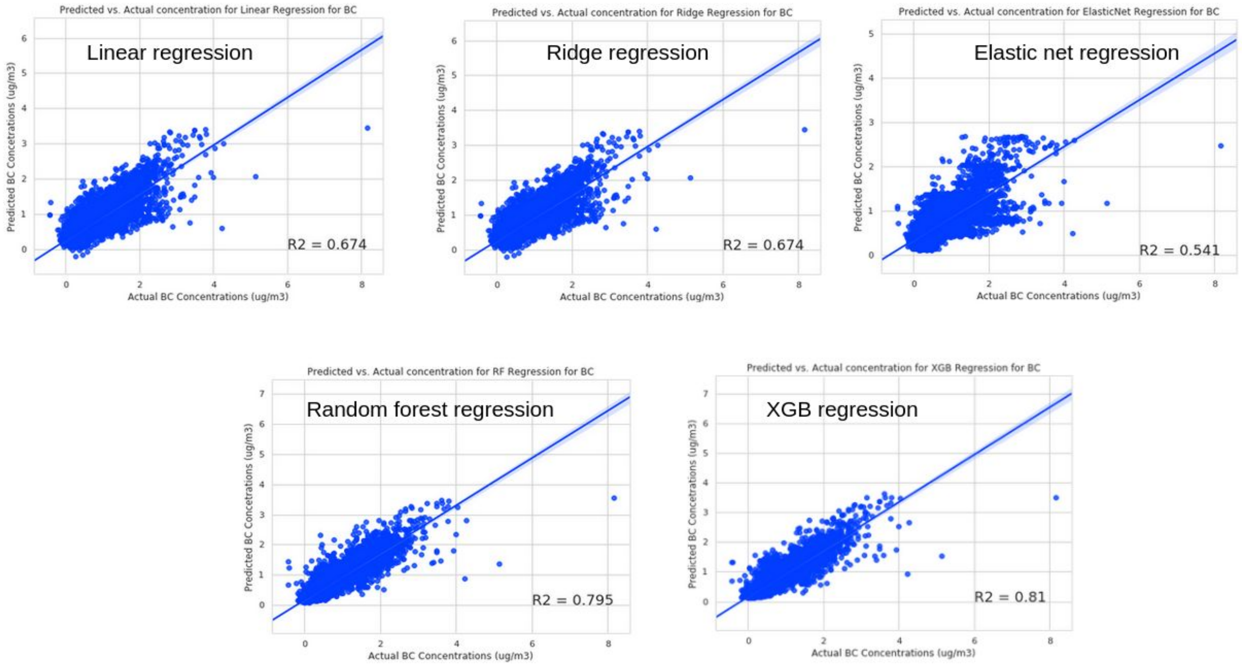


Figure 20: Scatter plot of predicted vs actual BC concentrations for different Machine Learning models

NO2 Dataset

Model parameters, R2 score and Root Mean Squared Error (RMSE) for the NO2 dataset are shown below. *Figure 21* shows a scatter plot of the predicted vs actual concentrations for each model. Similar to the BC dataset, the alpha = 0 for ridge and elastic net indicates that they converge to an OLS.

Model Type	**Best Parameter**	**Mean CV score for best estimator**	**Root Mean Squared Error**	**Training R2**	**Test R2**
Linear Regression without PCA		0.760	4.733	0.771	0.767

Ridge regression without PCA	Alpha = 0	0.759	1.733	0.771	0.767
Elastic Net Regression without PCA	Alpha = 0, L1_ratio = 0.1	0.638	5.92	0.640	0.635
Random Forest without PCA	Max depth = 50 Estimators = 800	0.910	2.828	0.988	0.917
XGBRegressor without PCA	Learning rate = 0.1 max depth = 10 estimators = 100	0.920	2.69	0.973	0.925

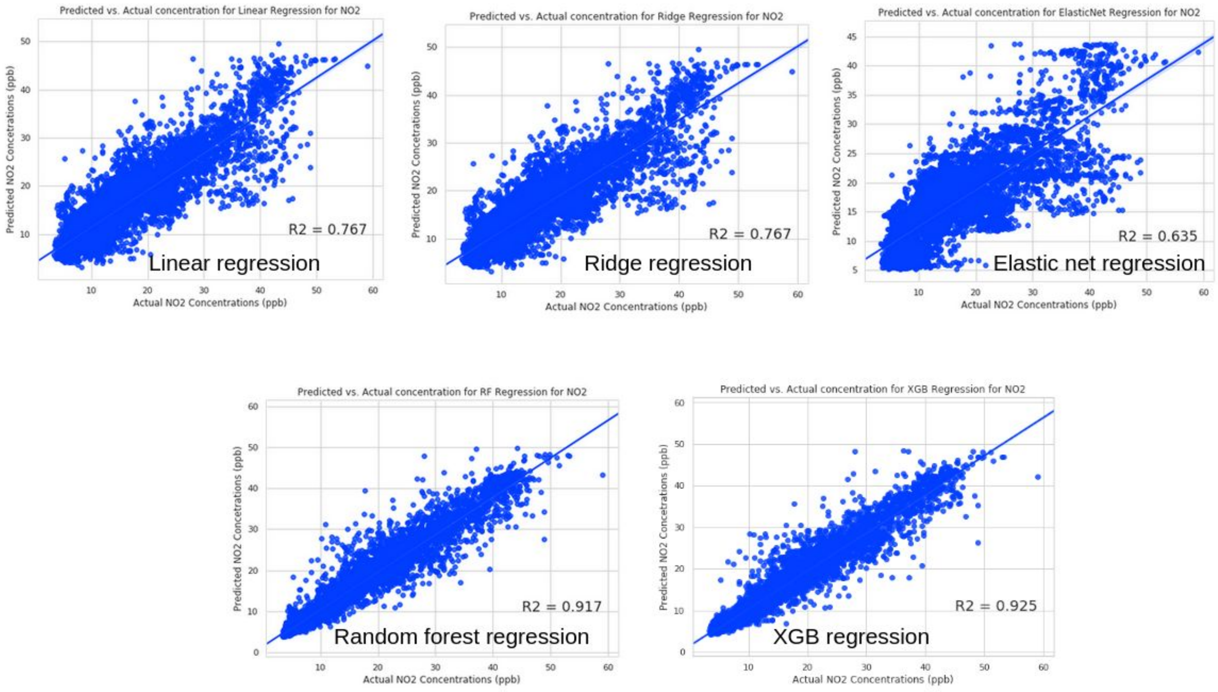


Figure 21: Scatter plot of predicted vs actual NO2 concentrations for different Machine Learning models

Summary

The objective of my work was to build a machine learning model to predict air pollution concentration on a block-by-block basis using publicly available data on major sources of emission, number of traffic intersections, proximity to closest highway and local meteorological data. This work leverages the air pollution monitoring dataset collected by the Environmental Defense Fund in partnership with Google Street View to collect air pollution concentrations in Oakland and San Leandro.

The input features used in the model includes distance to industrial sources of air emissions from a given X, Y location (in this case monitoring data), number of traffic intersections within 1,000ft, distance to closest highway and local meteorological data.

To understand the importance of different features in air quality prediction, a Random Forest approach was used due to severe multicollinearity issues with the dataset. The Random Forest approach resulted in the best method for feature selection because tree-based models naturally rank features by how well they improve the purity of the node, and it ignores features that are very similar. We found the following:

1) For the BC datasets, the top 5 features that had the largest effect on the concentration included the Oakland International Airport, a medical/outpatient unit with emergency generators, a foundry, a retail unit and a large residential complex with emergency generators.
2) For the NO_2 dataset, the facilities that resulted in a large contribution to NO_2 concentration in the area included a commercial complex, Oakland International Airport, a hospital, a data center and a supreme court.

This is an interesting finding, and it helps us pinpoint the major sources of emissions in the area that result in a poor air quality. Typically, Airports are major sources of Black Carbon, Particulate Matter and Nitrogen Dioxide because of the fuel burnt in aircrafts and smaller ground support equipment. Data centers with large amounts of diesel generators and other smaller generators and natural gas boilers are other major sources in the area.

It is important to note that the current extent of this work is to perform feature engineering and build a machine learning model based on the training and test data available. The next steps in this work would be to expand the spatial extent of prediction to generate a heatmap for a grid of points across all the cities in the East Bay Area, based on the same feature engineering procedure. This would involve identifying all the major sources of emissions, the distance from the point of interest to each source, number of traffic intersections in the area, proximity to a highway and the local meteorological parameters. This could be done by having the user enter an address, and use the location to generate features and make air quality predictions using the Random Forest and XGBoost models that were trained and tested in this work.

This next step would be particularly helpful for city planners, environmental management teams and public health experts to predict air quality for any location in East Bay Area, CA, and identify ‘hot-spots’ or locations where concentrations are unusually high. This would also help trace down the sources of emissions that contribute to high-concentrations in the area. Another application of this work would be for the public to identify locations/neighborhoods with good air quality in case they are interested in purchasing, renting or selling their houses. Typically houses located in areas with poor air quality are priced lower since they may be located close to industrial sources.

Finally, while we were able to build machine learning models to predict air pollution concentration, the features we identified here only give us an insight into the main sources that contribute to concentration in the entire region, and not on a hyper-local level. Since air pollution varies at a hyper-local level, predicting the major sources of pollution at a fine resolution is a hard problem to solve. Particularly, this work does not help identify the major sources of emissions in a particular neighborhood in Oakland vs. a neighbourhood in San Leandro as there would be several other factors such as traffic that contribute to differences in air quality. This work only identifies all the major sources in the entire East Bay Area that contribute to air pollution.

In addition, emissions from vehicles and trucks traveling on highways are actually one of the highest contributors to BC and NO_2 emissions on a local-level, and lack of emissions data from traffic is another limitation of this work.