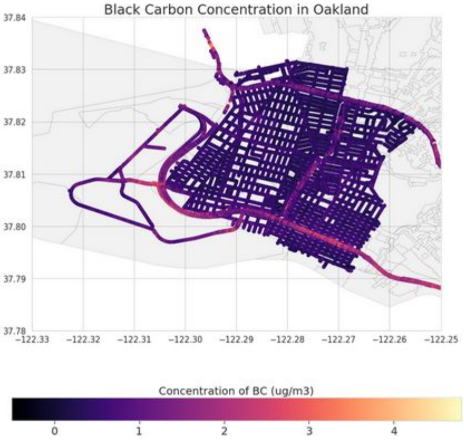
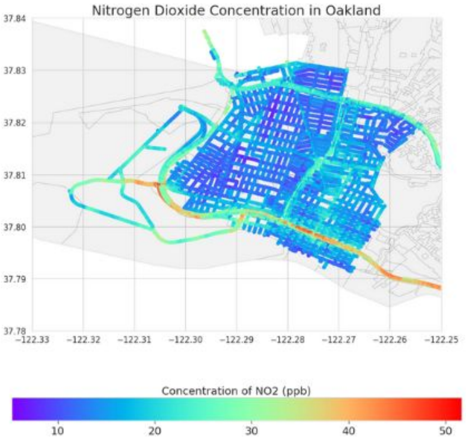
Air Quality Prediction in East Bay Area, CA

Building a Machine Learning Model for Air Quality Predictions

Varsha Gopalakrishnan



Black Carbon Concentration in Oakland



Nitrogen Dioxide Concentration in Oakland

Project Scope

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?

Datasets

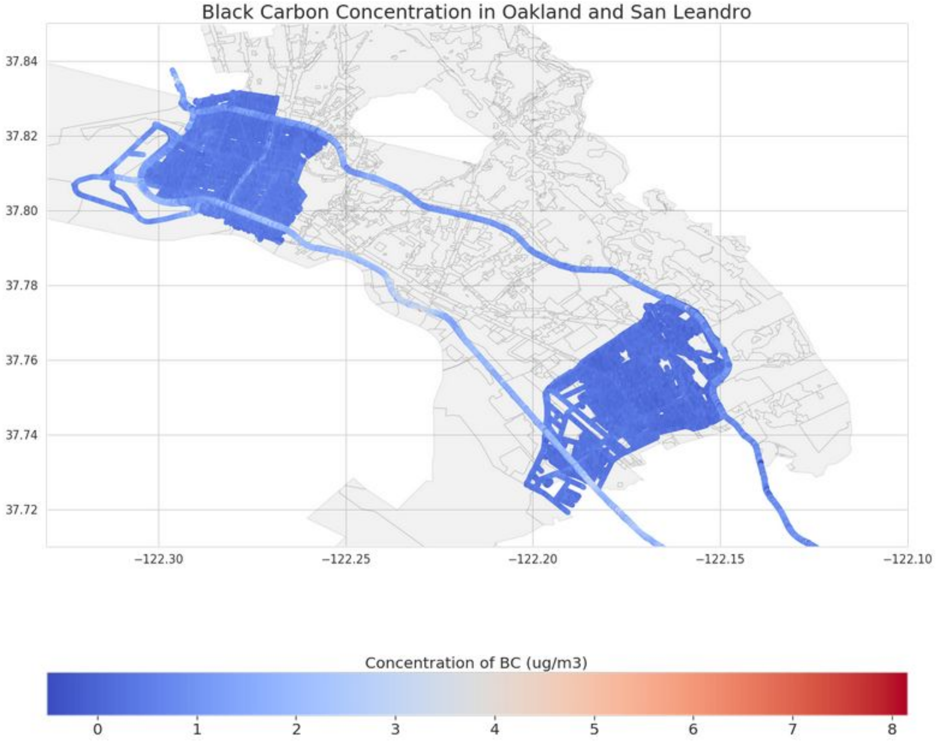
- Oakland Air Pollution Monitoring Data measured by Environmental Defense Fund (EDF)
- Air Emissions data obtained from the National Emissions Inventory for individual sources in Alameda County
- Intersection Count and distance to closest highway from Open Street Maps
- Local meteorological data on a 1kmx1km grid obtained from Oak Ridge National Lab

Oakland Air Pollution Monitoring Data

What does the dataset tell us?

Black Carbon Concentration -

- 21,488 data points
- Each data point - Average BC concentration between 06/15 and 05/16 at a specific lat-long location
- Average concentration across all lat-longs = **0.72 ug/m^3**
  - This is **five times** the 2010 average CA concentration*
- Maximum concentration - **8.15 ug/m^3** (on a highway)


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

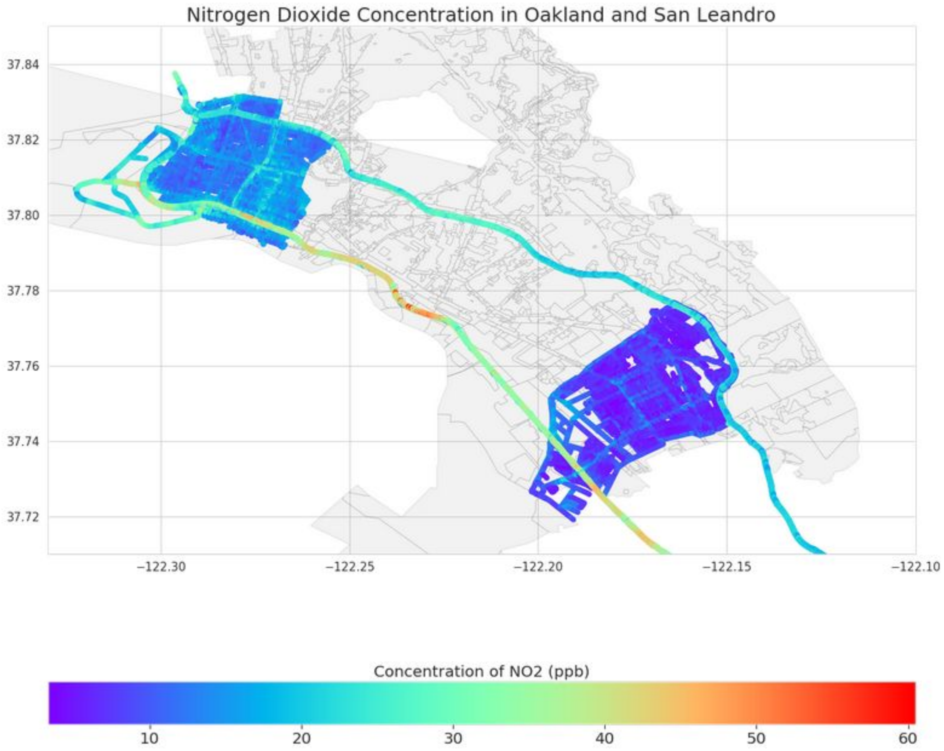
*https://earthjustice.org/sites/default/files/black-carbon/impact-of-ca-air-pollution-laws.pdf

Oakland Air Pollution Monitoring Data

What does the dataset tell us?

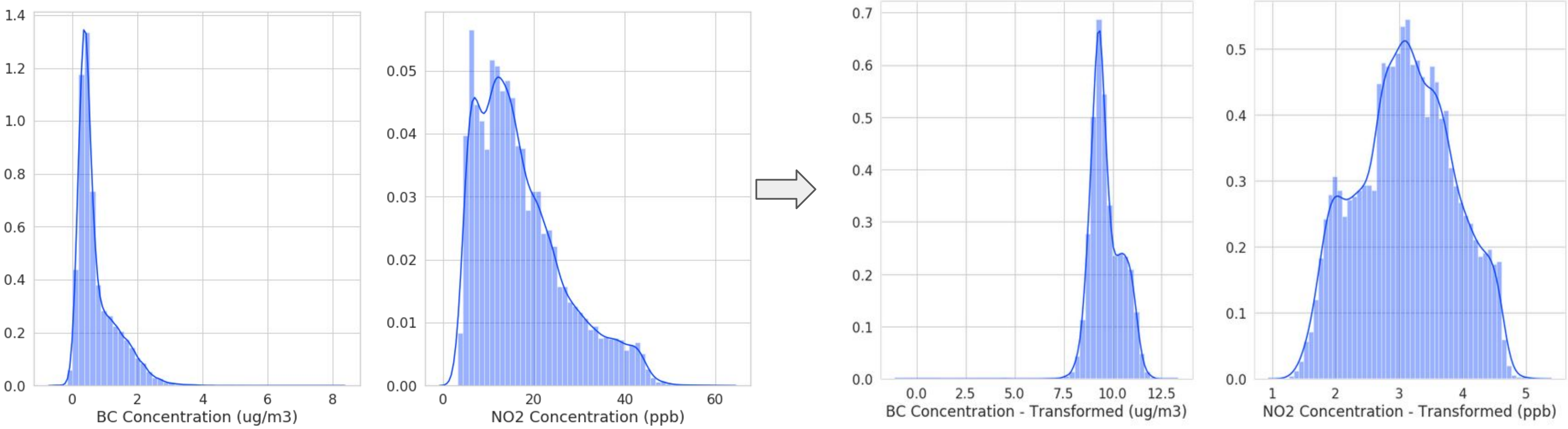
Nitrogen Dioxide Concentration -

- 21,488 data points
- Each data point - Average NO_2 concentration between 06/15 and 05/16 at a specific lat-long location
- Average concentration across all lat-longs - **17.1 ppb**
  - Lower than the National Standard 53 ppb and California Standard of 30 ppb.
- Maximum concentration - **60 ppb** (also close to a highway)


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

Oakland Air Pollution Monitoring Data

What is the distribution of data?



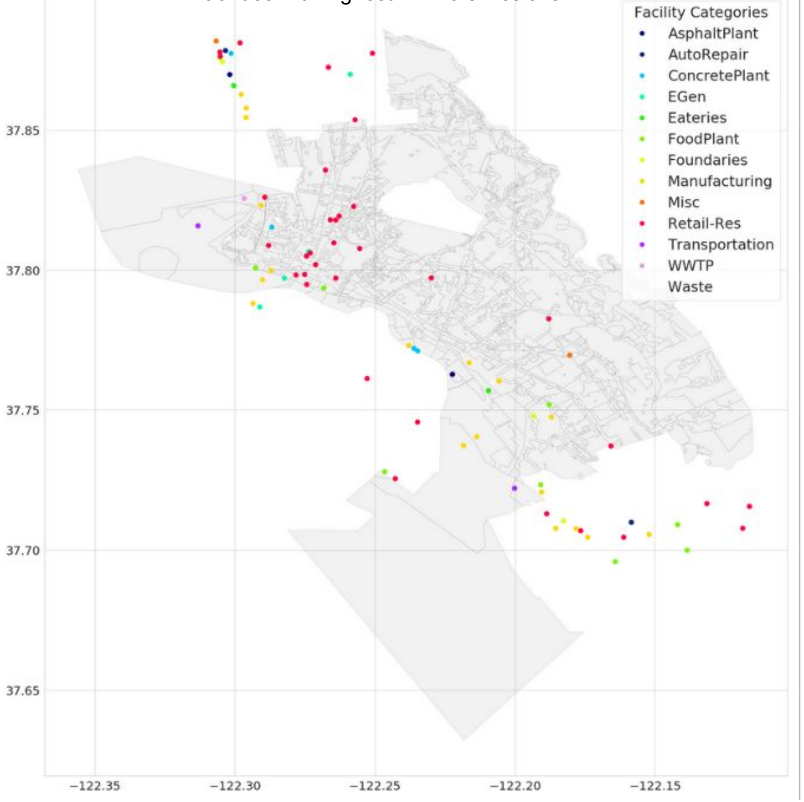
- BC and NO_2 concentrations are “positively-skewed” i.e. skewed towards the right
- Mean concentration is greater than median concentration
- Perform a box-cox transform to transform the data to a normal distribution

National Emissions Inventory

What does the dataset tell us?

$PM_{2.5}$ Emissions Data (used as a proxy for BC)

- **348** stationary sources
- Largest emitters = Asphalt Plants, Concrete batch processing plants, and food manufacturing
- Facilities classified into **'low', 'medium' and 'high'** depending on quantity of emissions
  - *Emissions < first quartile - **'Low'***
  - *Emissions between first and 3rd quartile - **'Medium'***
  - *Emissions > 3rd quartile - **'High'***

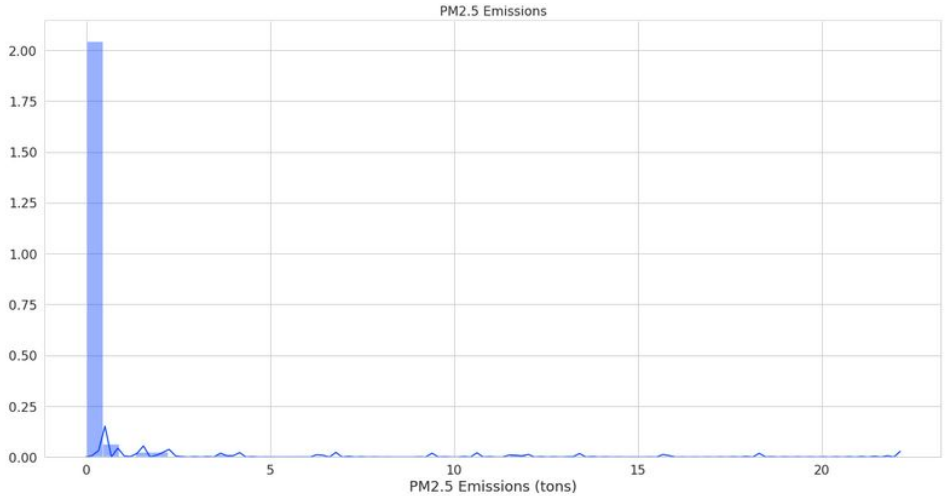


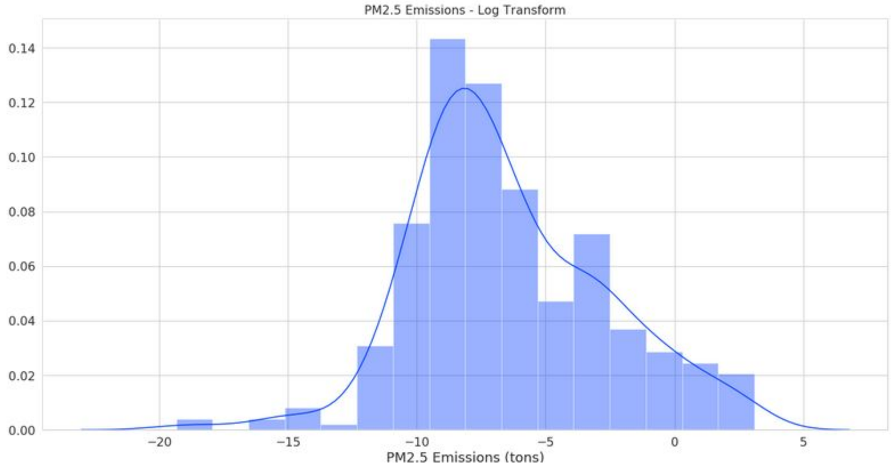
Facilities with highest PM2.5 emissions

National Emissions Inventory

What is the distribution of data?

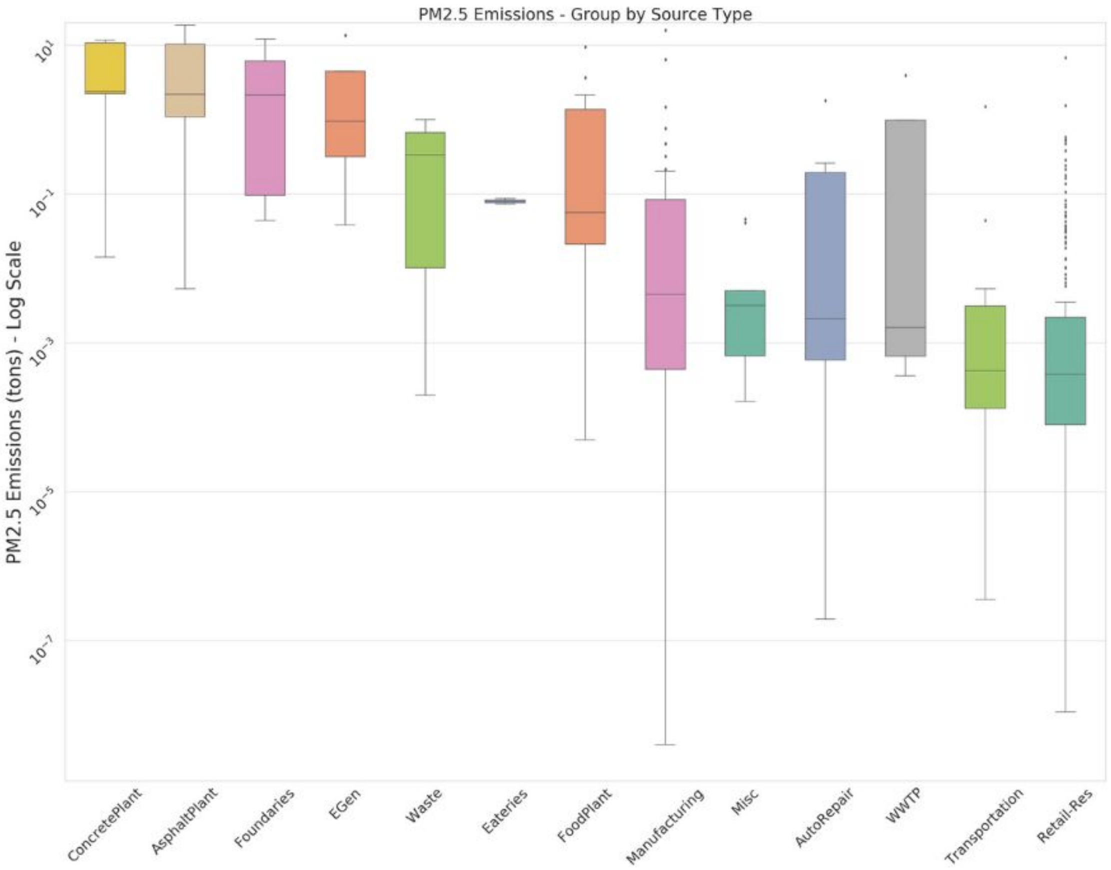
$PM_{2.5}$ Emissions Data





- Histogram indicates the data is heavily skewed right i.e. there are several small sources of emissions and one large emitter.
- Doing a box-cox (log transformation) results in a plot that looks normal

Emissions Inventory grouped by source type - $PM_{2.5}$



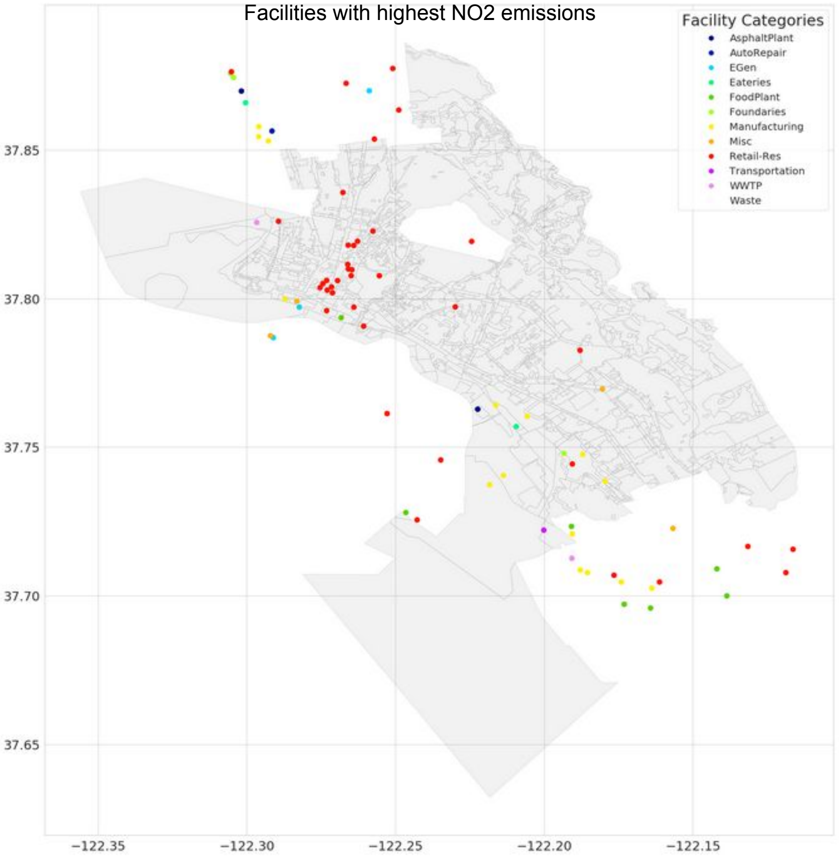
- Sources of similar types are further grouped into larger categories to reduce dimensionality of input features
  - Pharmaceuticals, metals, pipe plants and solvents - “Manufacturing”
  - Port of Oakland, Airport, Parking and US Coast Guard - “Transportation”

National Emissions Inventory

What does the dataset tell us?

NO_2 Emissions Data

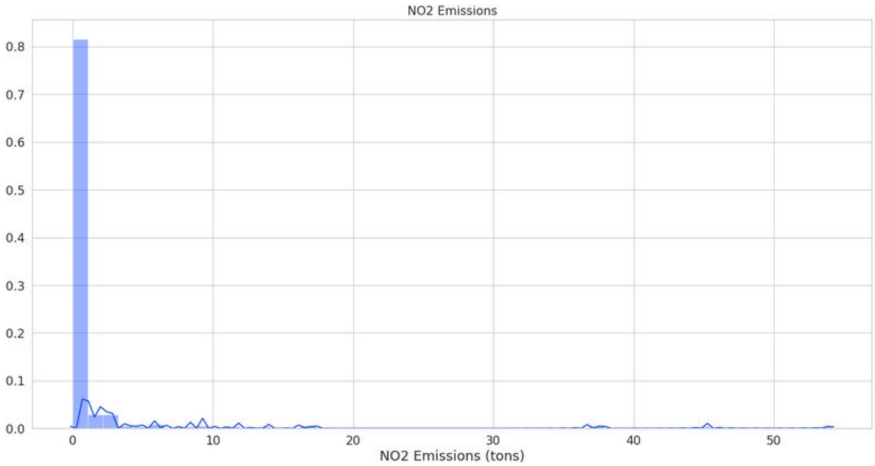
- **317** stationary sources
- Largest emitters = Food Plants, Foundries, Electricity Generating Units and Pharmaceutical Plants.
- Facilities classified into **'low', 'medium' and 'high'** depending on quantity of emissions
  - *Emissions < first quartile - **'Low'***
  - *Emissions between first and 3rd quartile - **'Medium'***
  - *Emissions > 3rd quartile - **'High'***

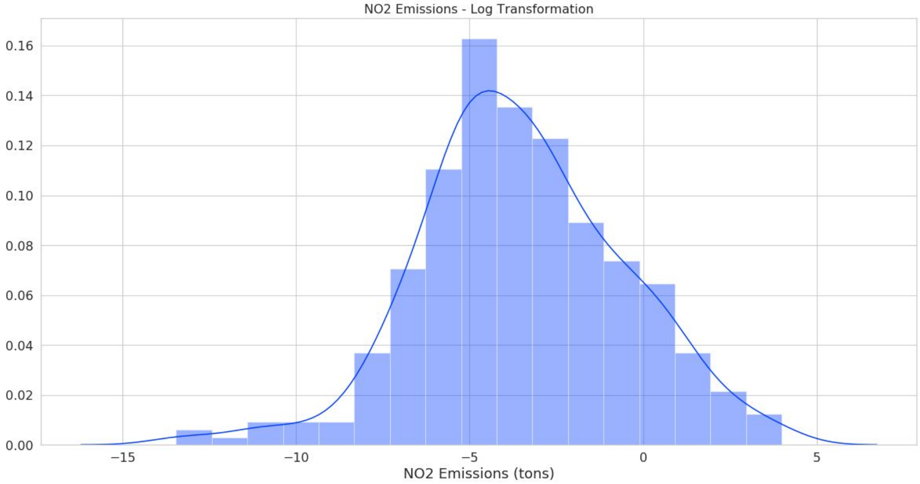


National Emissions Inventory

What is the distribution of data?

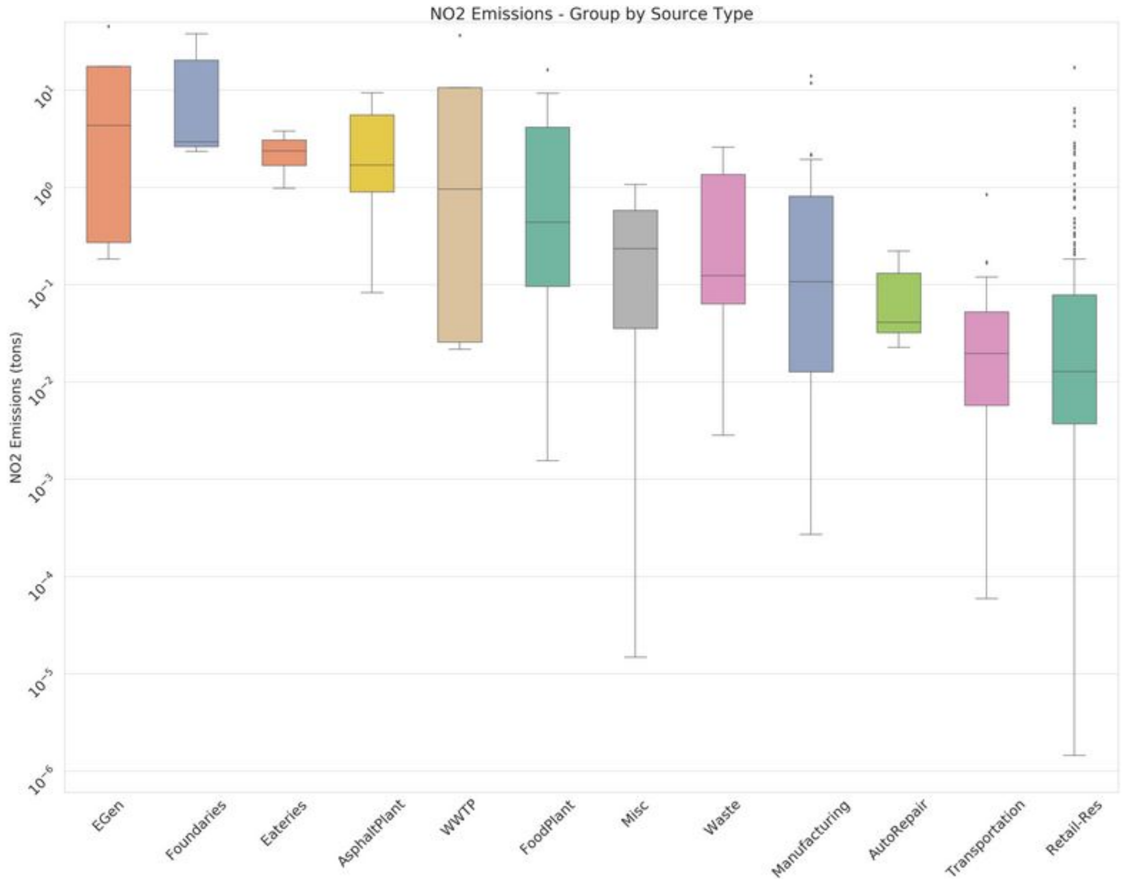
NO_2 Emissions Data





- Histogram indicates the data is heavily skewed right i.e. there are several small sources of emissions and one large emitter.
- Doing a box-cox (log transformation) results in a plot that looks normal

Emissions Inventory grouped by source type - NO_2



- Sources of similar types are further grouped into larger categories to reduce dimensionality
  - Pharmaceuticals, metals, pipe plants and solvents - “Manufacturing”
  - Port of Oakland, Airport, Parking and US Coast Guard - “Transportation”

Traffic Data

What does the dataset tell us?

Number of Traffic Intersections within 1,000 ft

- Location of traffic intersections in Oakland and San Leandro obtained from Open Street Maps
- Total number of traffic intersections within a 1,000 ft radius from each monitoring location
- Minimum number of intersections = 2
- Maximum number of intersections = 35

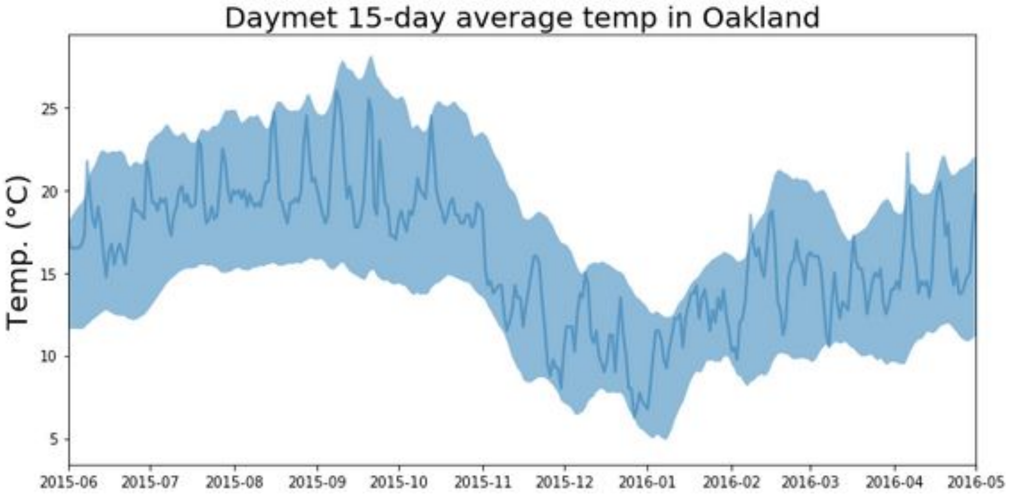
Distance to closest highway

- Distance to closest highway from monitoring location obtained from Open Street Maps

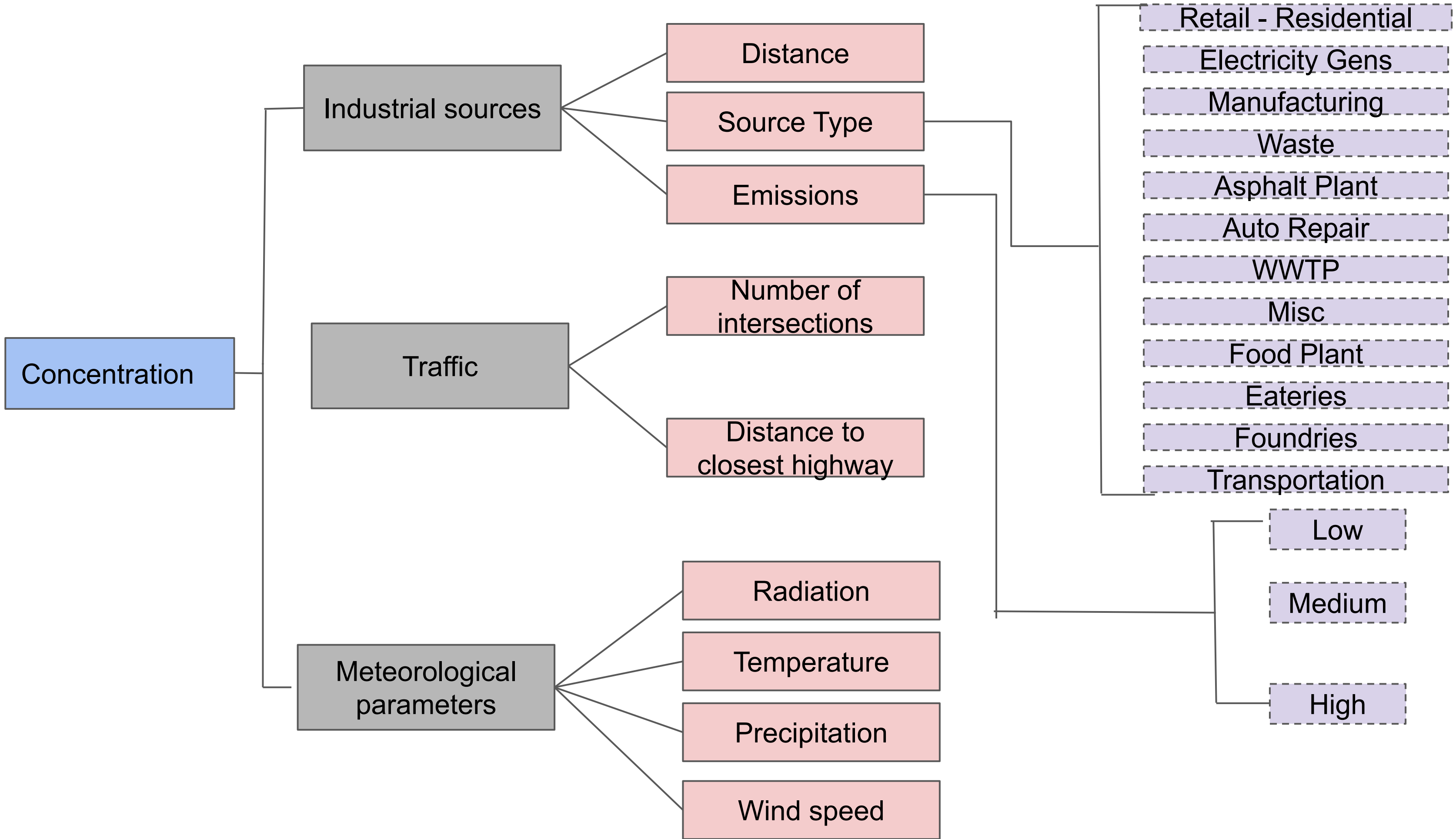
Meteorological Data

What does the dataset tell us?

- Obtained from Oak Ridge National Lab **Daymet** dataset.
- Dataset contains daily meteorological parameters on a 1km by 1km grid basis.
- Average daily measurements estimated between June 2015 - May 2016.
- Parameters include Precipitation, Radiation, Minimum and Maximum Temperature, and Pressure

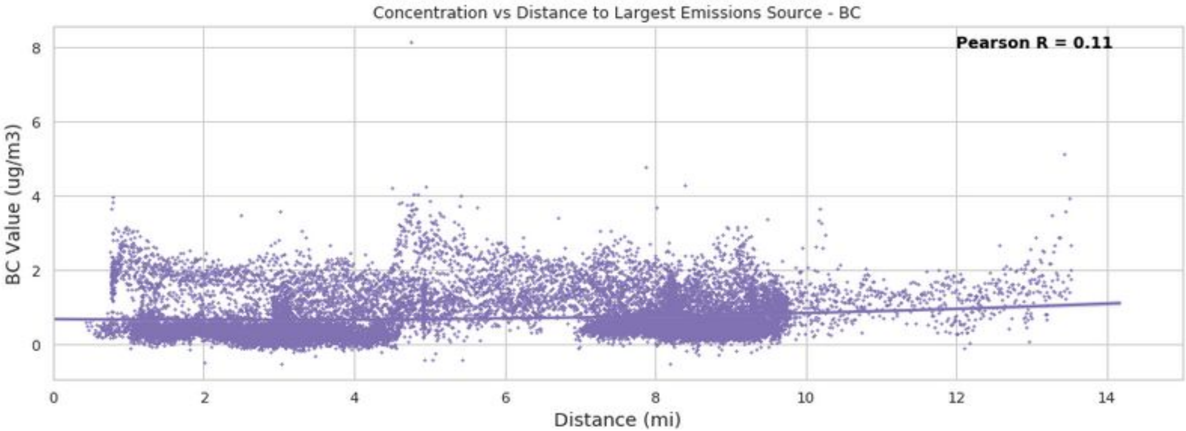


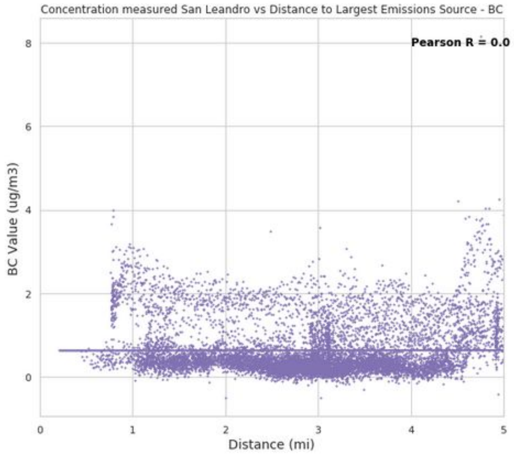
Hypothesis Tree

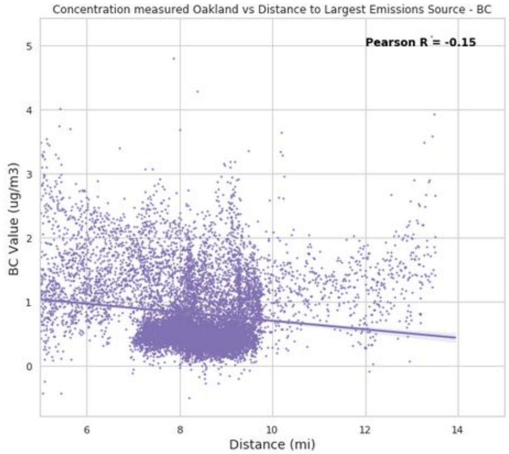


Exploratory Data Analysis

How does concentration vary with distance for largest emitter for BC?





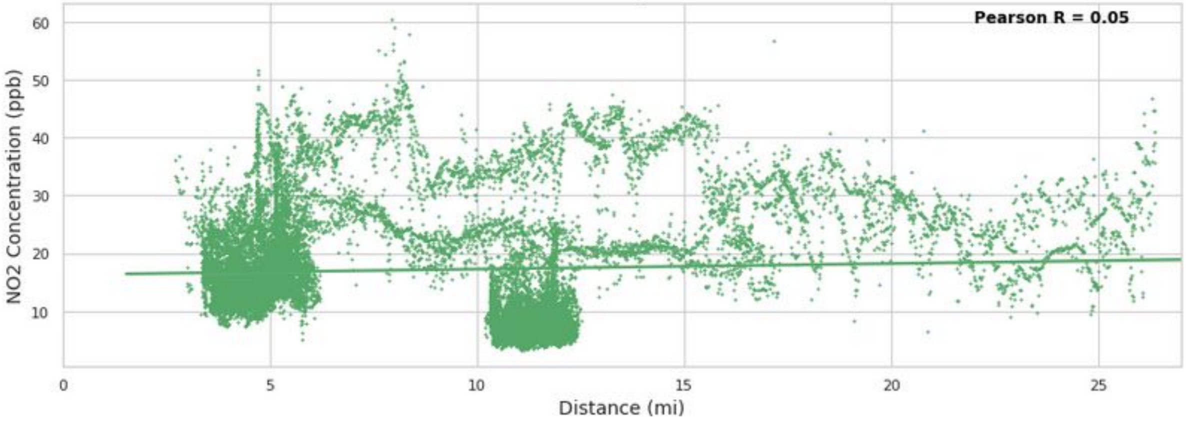


- Overall low correlation
- Two separate clustered points arise because of the location of the source
- Source located in San Leandro;
  - Points < 5 mi = measurements in San Leandro region
  - Points > 5 mi = measurements in Oakland region
- Zero correlation for points in San Leandro
  - Indicates facility is not an important feature for predicting BC concentration.
- Negative correlation for points in Oakland
  - Interesting and unusual; could indicate distances farther away have a negative impact on concentrations

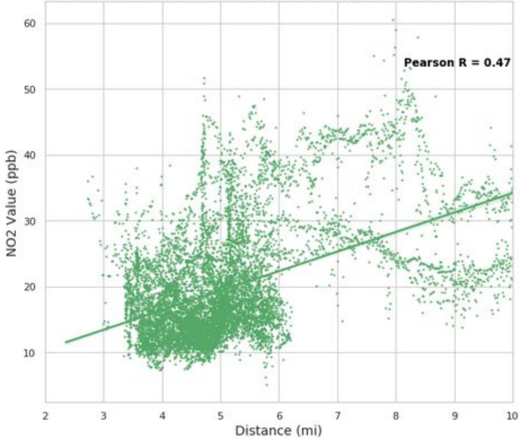
Exploratory Data Analysis -

How does concentration vary with distance for largest source for NO_2?

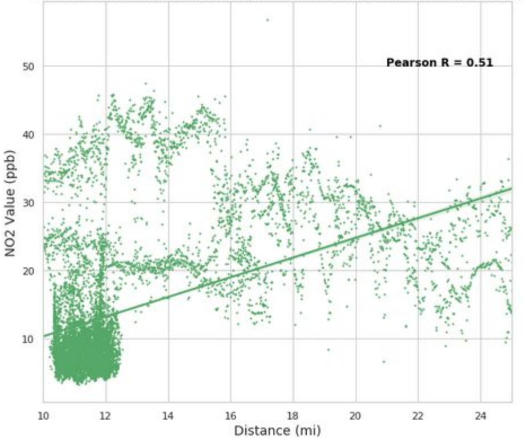
Concentration vs Distance to Largest Emissions Source - NO2



Concentration measured in Oakland vs Distance to Largest Emissions Source - NO2



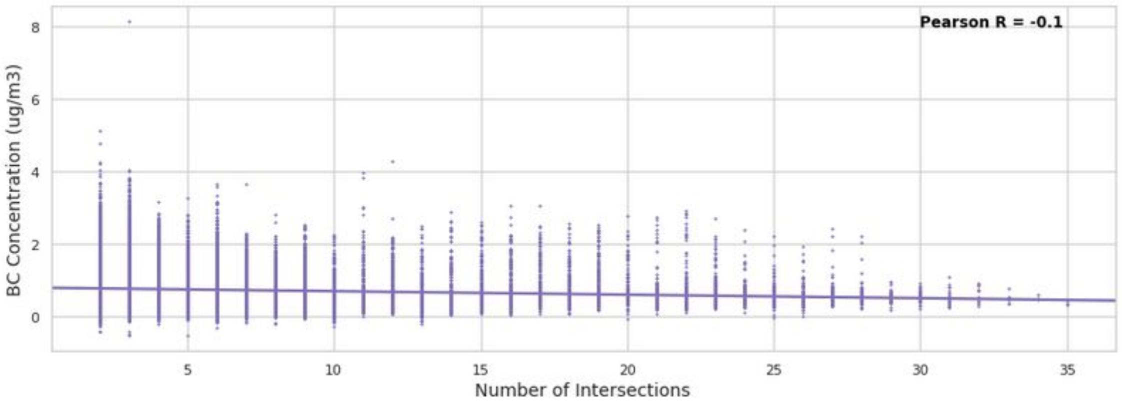
Concentration measured in San Leandro vs Distance to Largest Emissions Source - NO2

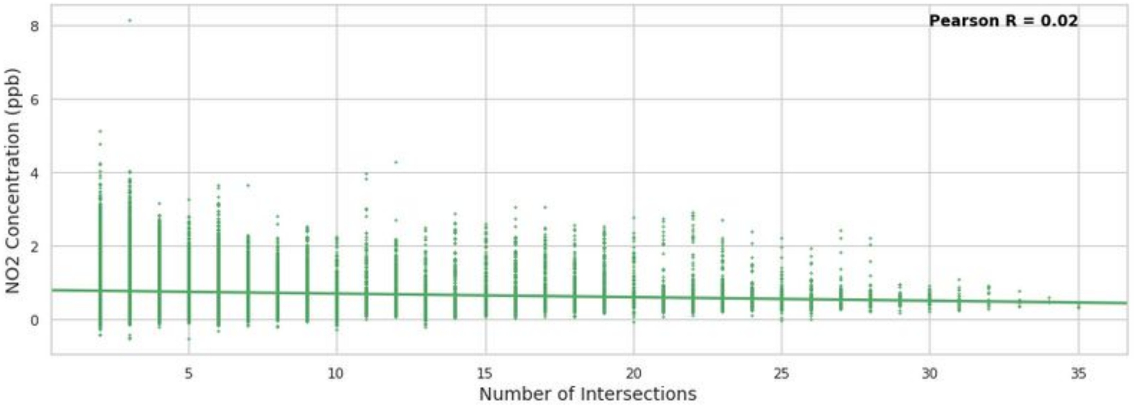


- Overall low correlation
- Two separate clustered points arise because of the location of the source
- Source located in Berkeley
  - Points < 10 mi = measurements in Oakland
  - Points >10 mi = measurements in San Leandro.
- When points are split into Oakland vs San Leandro - a strong positive correlation for points in Oakland and San Leandro.
- This could indicate that this facility could be an important feature in predicting NO_2 concentrations.

Exploratory Data Analysis -

How does concentration vary with number of traffic intersections?

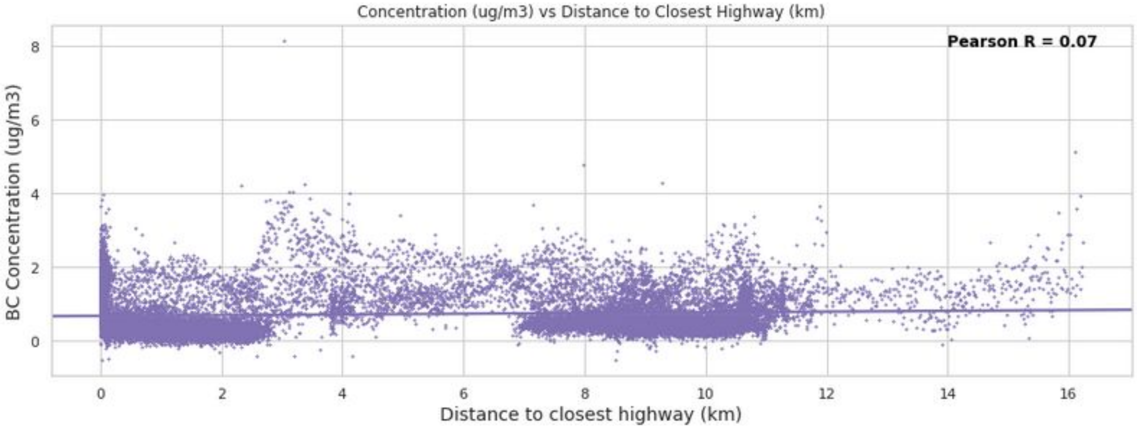


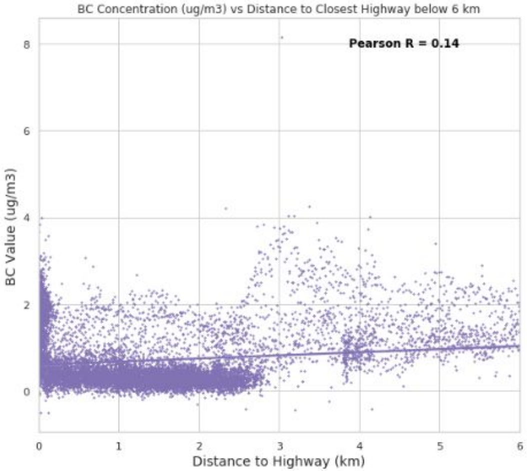


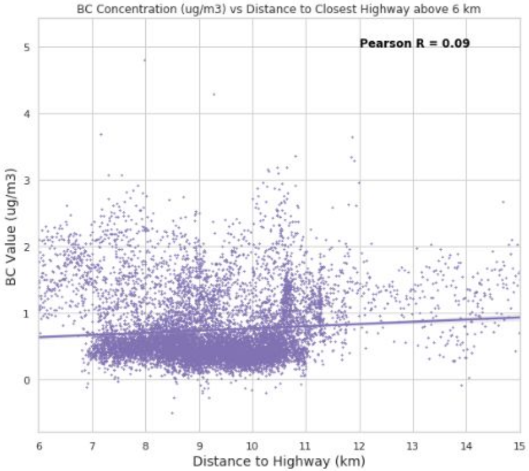
- Low correlation between concentration and number of intersections
- Indicates number of traffic intersections may not be an important feature in predicting concentrations.
- Negative correlation between number of traffic intersections and BC concentration is also interesting.

Exploratory Data Analysis -

How does BC concentration vary with distance to closest highway?



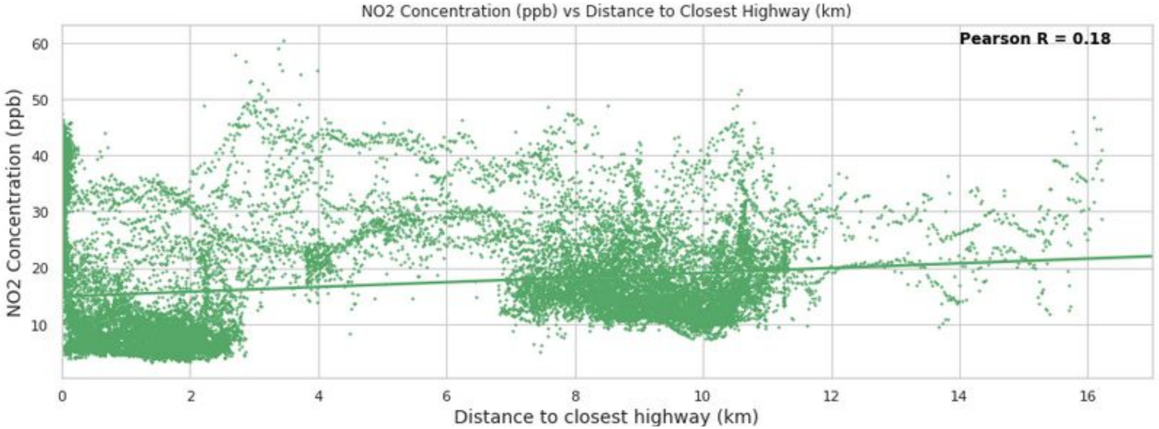




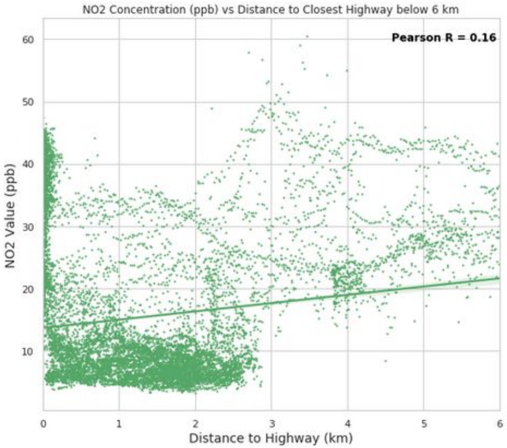
- Overall low correlation
- When points are split into two separate regions, below and above 6 km, we observe a stronger positive correlation for the two regions.
- Points with distance = 0 are points on highways with high concentrations
- Points with distance < 6km have a slightly larger positive correlation. These could be points on ramps, and points in close proximity to highways.
- The lower correlation with increasing distance indicates that BC concentration tends to reduce with distance from highway.

Exploratory Data Analysis -

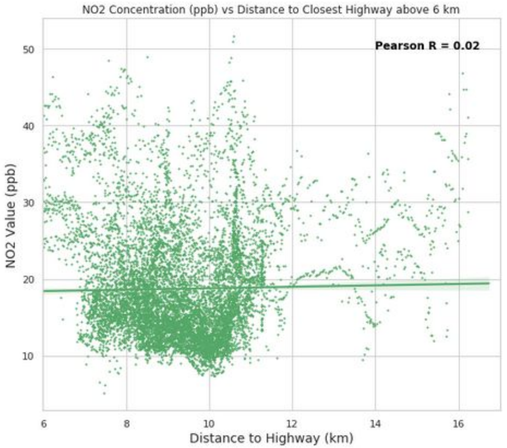
How does NO2 concentration vary with distance to closest highway?



NO2 Concentration (ppb) vs Distance to Closest Highway (km)



NO2 Concentration (ppb) vs Distance to Closest Highway below 6 km

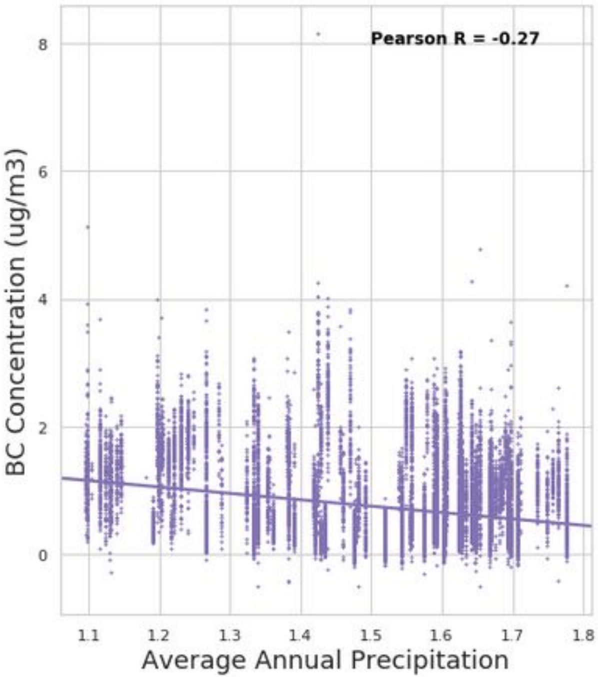


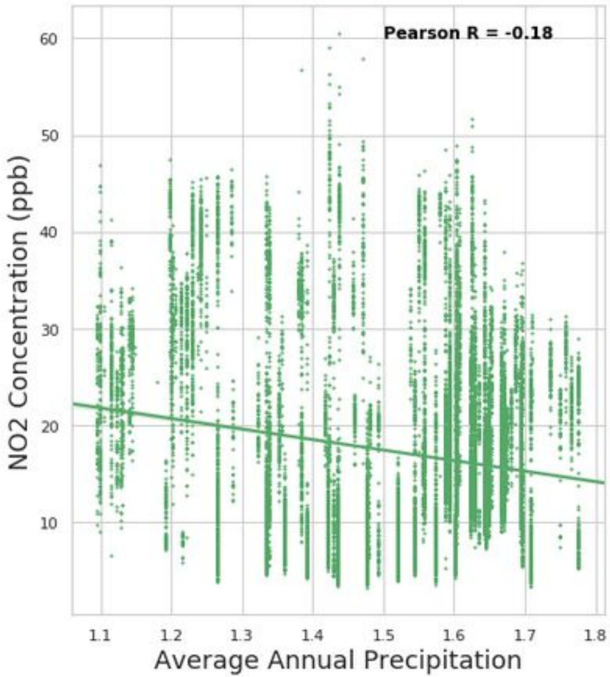
NO2 Concentration (ppb) vs Distance to Closest Highway above 6 km

- Overall low positive correlation
- When points are split into two separate regions, below and above 6 km, correlation coefficient reduces slightly for the points < 6km, and drops to almost zero with distance
- Points with distance = 0 are points on highways with high concentrations
- Points with distance < 6km have a slightly larger positive correlation. These could be points on ramps, and points in close proximity to highways
- Close to zero correlation with increasing distance indicates that NO_2 concentration tends to reduce with distance from highway.

Exploratory Data Analysis -

How does concentration vary with some meteorological parameters





- Low negative correlation between concentration and annual precipitation.
- Correlations indicate that this could be an important feature in the prediction model.

Feature Selection

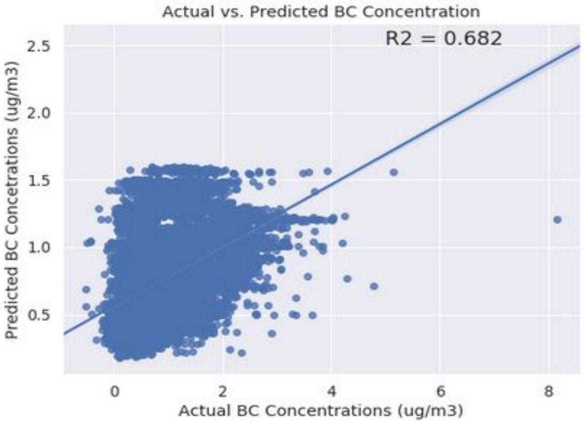
Resolving multicollinearity:

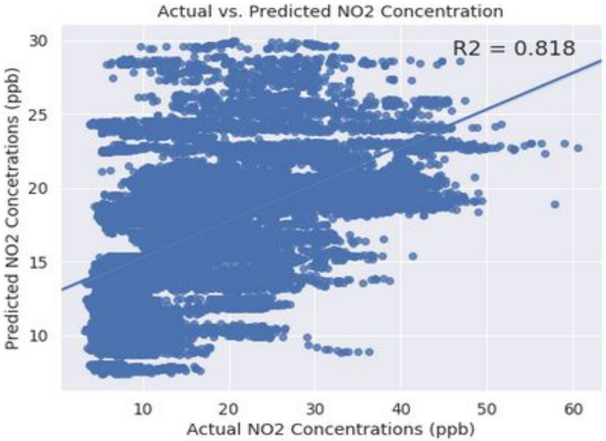
- Correlation matrix of features in the BC and NO_2 datasets revealed that multicollinearity is a major issue in this dataset.
- It is important to eliminate features that are multicollinear because multicollinearity can undermine the statistical significance of an independent variable.
- While multicollinearity 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.

Feature selection

Resolving multicollinearity by dropping features with a high correlation coefficient

- In this approach, we identify all features that have a correlation coefficient greater than 0.9 or lesser than -0.9, and drop those features.
- A simple OLS fit on the resulting features indicated that the model has an R^2 value of 0.682 for BC and 0.818 for NO_2.
- However, the Variance Inflation scores for the features are still pretty high (>10) indicating the multicollinearity is not entirely removed and **this method of feature selection is not ideal**.

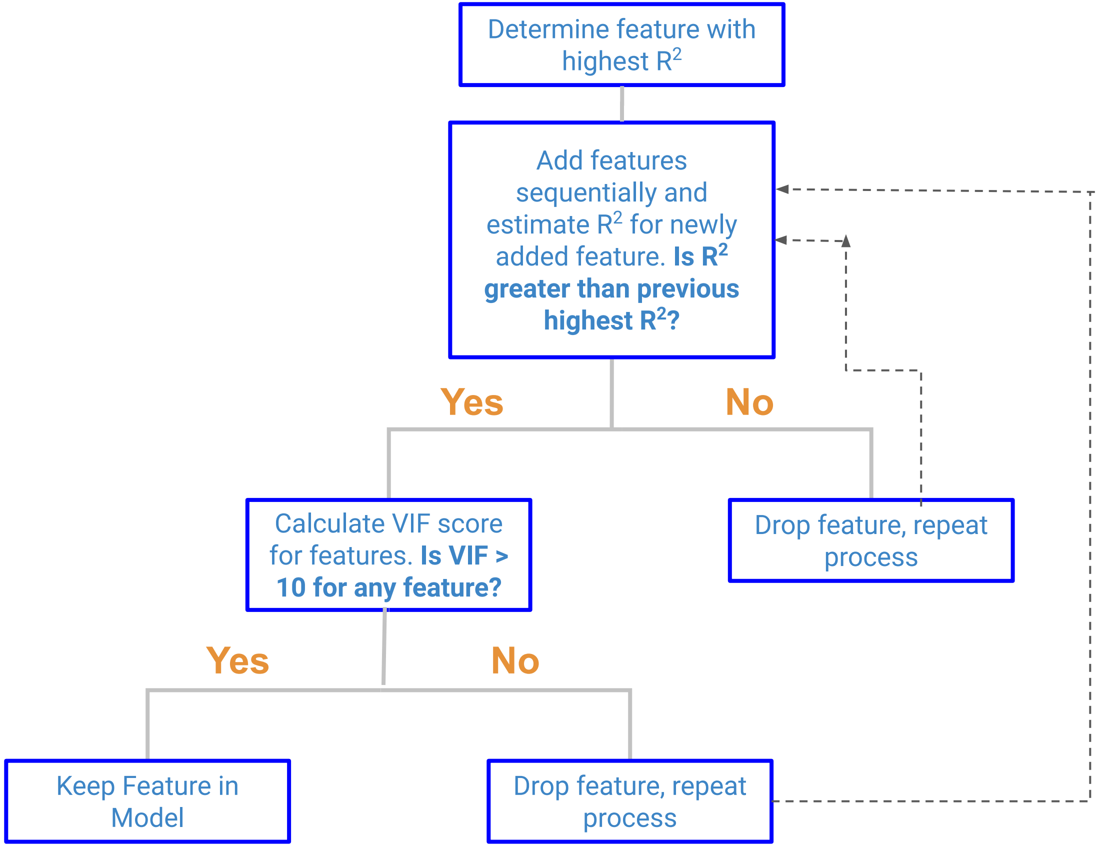




Feature Selection

Resolving multicollinearity by a step forward feature selection with VIF and R^2 estimation

- We use the algorithm listed here to drop features that are multicollinear.
- For the BC and NO_2 datasets, only '**Radiation**' was selected as a feature with R^2 of 0.584 for BC and 0.757 for NO_2.



Determine feature with highest R^2

Add features sequentially and estimate R^2 for newly added feature. **Is R^2 greater than previous highest R^2?**

Calculate VIF score for features. **Is VIF > 10 for any feature?**

Drop feature, repeat process

Keep Feature in Model

Drop feature, repeat process

Feature Selection

Resolving multicollinearity by Lasso Regularization and a step forward feature selection with VIF and R^2 estimation

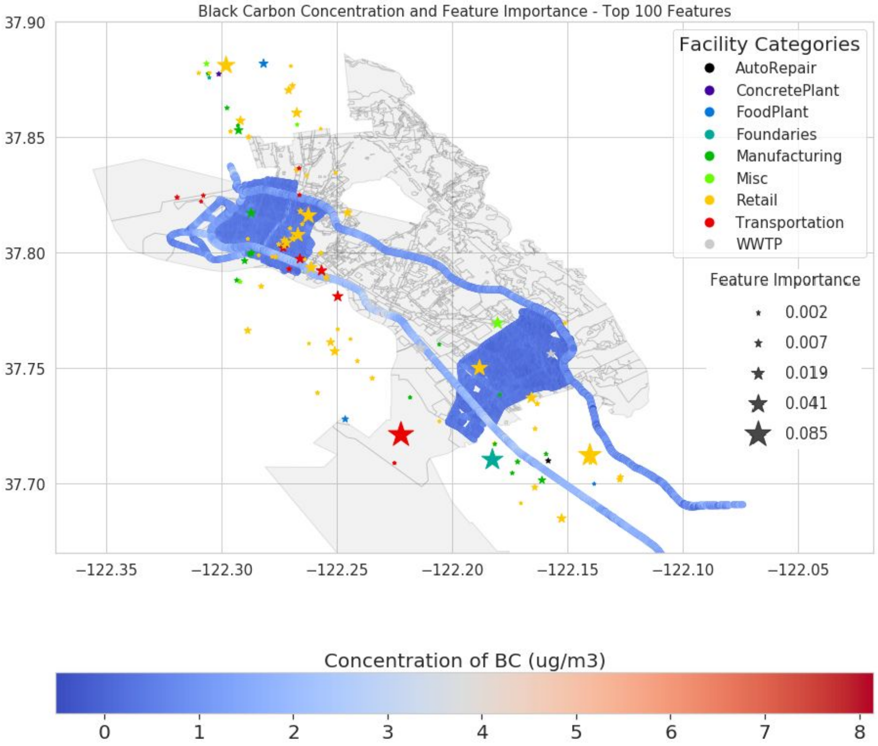
- The same step forward procedure was used but this time, a lasso regularization was applied to eliminate features that don't have an effect on the target variable.
- Lasso regularization selected 49 features out of the 68 features for the BC dataset and 42 features out of 50 features for the NO2 dataset.
- After applying the step-forward VIF approach, only **'Minimum Temperature'** was selected as a feature

Feature Selection

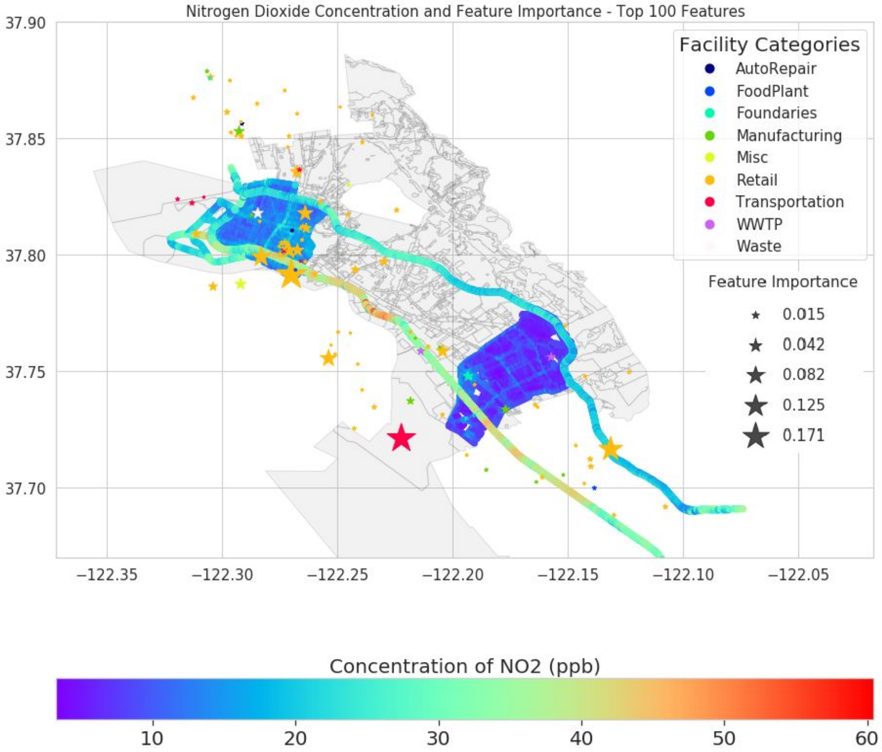
Feature Selection using Random forest with cross validation

- 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
- BC and NO_2 datasets were split into test/train data, 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.
- The next slide shows location of the top 100 features BC and NO_2. These only include location of facilities that contribute 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.

Feature Selection



Top 5 Features: Oakland International Airport, medical/outpatient unit, the Ridge foundry in San Leandro, a retail unit in Berkeley and a large residential complex in Oakland.

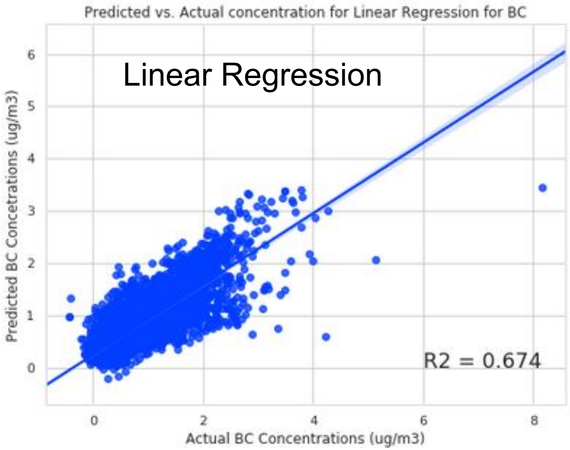


Top 5 Features: commercial complex, the Oakland International Airport, the Kindred hospital in San Leandro, Digital Realty data center in Oakland and the California Supreme Court of Alameda.

Machine Learning

- Prior to building and training machine learning models, a **Principal Component Analysis (PCA)** was performed to reduce dimensionality of features and to understand whether application of PCA improves the model performance or not.
- Results of the PCA are not shown here as the benefits of PCA are relatively low; increase in R^2 value is not significant.
- Since PCA is a computationally intensive process, especially while using Random Forest or XGBoost, PCA transformation is not considered.
- Several different machine learning models were built and tested for its predictive performance.
  - Linear Regression
  - Ridge Regression
  - Elastic Net Regerssion
  - Random Forest
  - XGBoost

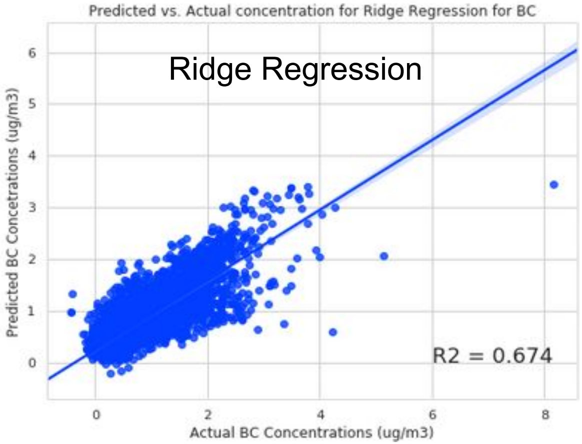
Machine Learning - BC dataset



Predicted vs. Actual concentration for Linear Regression for BC

Linear Regression

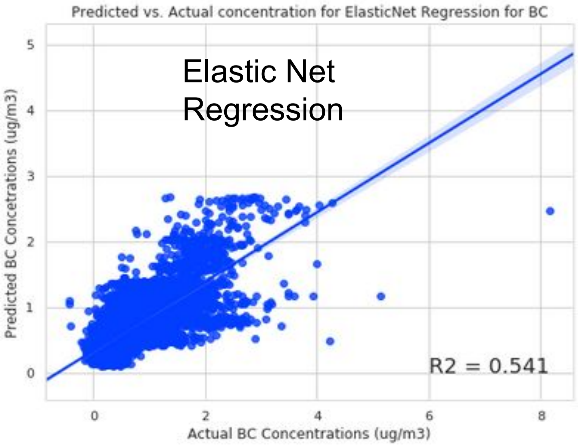
R2 = 0.674



Predicted vs. Actual concentration for Ridge Regression for BC

Ridge Regression

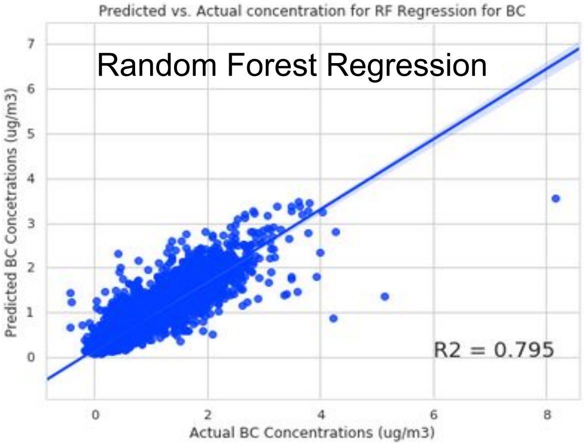
R2 = 0.674



Predicted vs. Actual concentration for ElasticNet Regression for BC

Elastic Net Regression

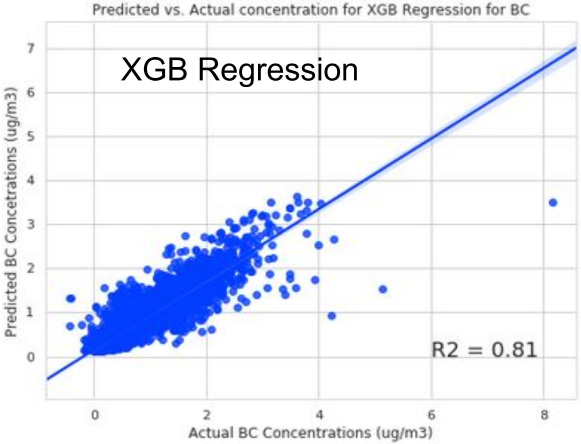
R2 = 0.541



Predicted vs. Actual concentration for RF Regression for BC

Random Forest Regression

R2 = 0.795

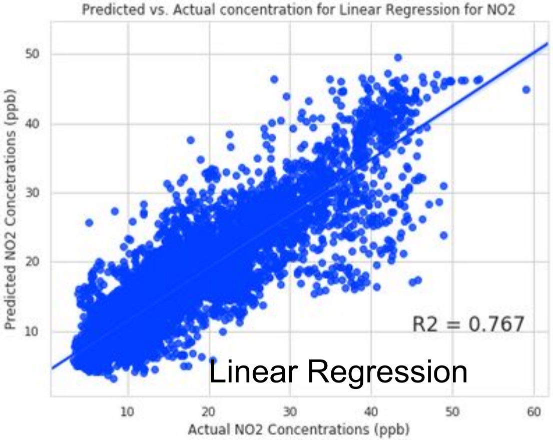


Predicted vs. Actual concentration for XGB Regression for BC

XGB Regression

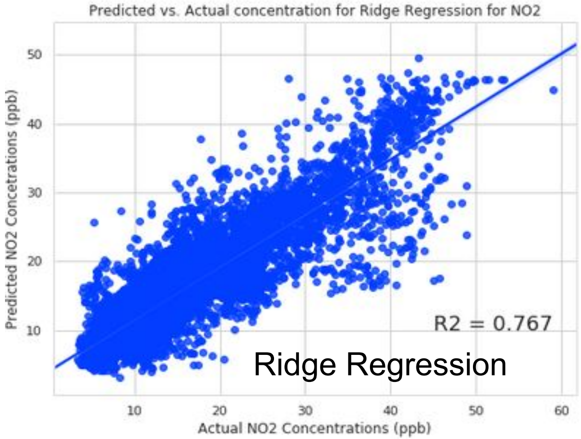
R2 = 0.81

Machine Learning - NO$_2$ dataset


Predicted vs. Actual concentration for Linear Regression for NO2

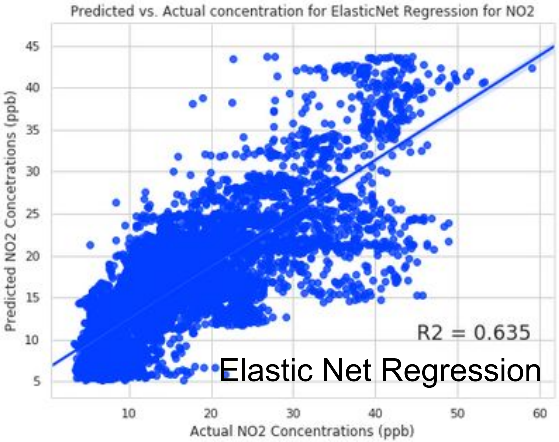
R2 = 0.767

Linear Regression


Predicted vs. Actual concentration for Ridge Regression for NO2

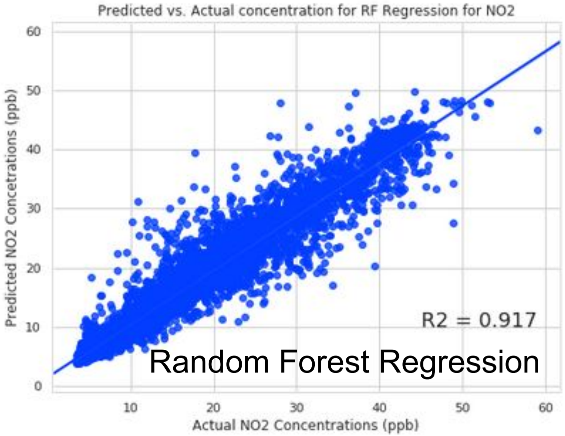
R2 = 0.767

Ridge Regression


Predicted vs. Actual concentration for ElasticNet Regression for NO2

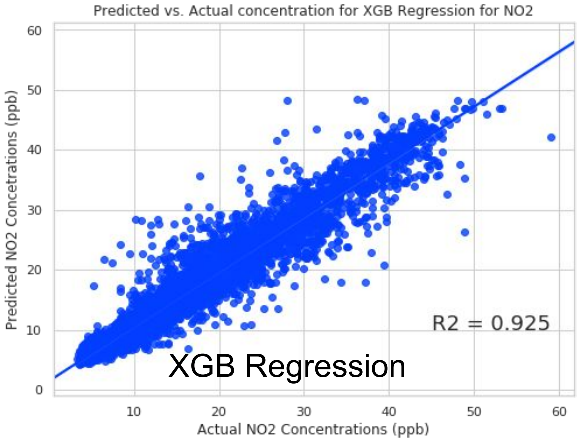
R2 = 0.635

Elastic Net Regression


Predicted vs. Actual concentration for RF Regression for NO2

R2 = 0.917

Random Forest Regression


Predicted vs. Actual concentration for XGB Regression for NO2

R2 = 0.925

XGB Regression

Summary

- Objective was to build a machine learning model to predict air pollution concentration using data on major sources of emission, number of traffic intersections, proximity to closest highway and local meteorological data.
- Several input features (industrial sources) were correlated with each other - severe multicollinearity issues, evident from the high VIF scores.
- Random Forest approach was used for feature selection due to multicollinearity issue
- Features with largest effect on BC concentration - Oakland International Airport, a medical/outpatient unit, a foundry, a retail unit and a large residential complex.
- NO_2 dataset - commercial complex, Oakland International Airport, a hospital, a data center and the supreme court of California building.
- Tree based Machine Learning models such as Random Forest and XGBoost methods performed far better than Linear or Ridge Regression models for BC and NO_2.

Summary

- **Applications and Future Work**
  - Expand spatial extent of prediction - generate a heatmap for grid of points across East Bay Area, and make predictions based on models developed in this work.
  - Develop an app where user can enter address to get air quality in a location.
  - Helpful for city planners, public health experts to identify ‘hot-spots’ or locations with unusually high concentrations
  - Public can identify locations/neighborhoods with good air quality in case they are interested in purchasing, renting or selling houses
- **Limitations:**
  - Insight into only the main sources that contribute to concentration in the entire region, not on a hyper-local level even though air quality varies hyper-locally.
  - Cannot identify major sources of emissions in a particular neighborhood in Oakland vs. San Leandro; only sources in the entire region can be identified
  - Emissions from vehicles and trucks which are highest contributors to BC and NO_2 emissions on a local-level are not considered due to lack of data.