Capstone 1 - In Depth Analysis Report:

Author: Varsha Gopalakrishnan

In the previous report (Milestone Report), we identified some of the features that are important and contribute to concentrations of Black Carbon (BC) and Nitrogen Dioxide (NO2) in the area. In this report, we'll take these results a step further and build machine learning models that give good predictions.

Prior to building machine learning models, the dataset was 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:

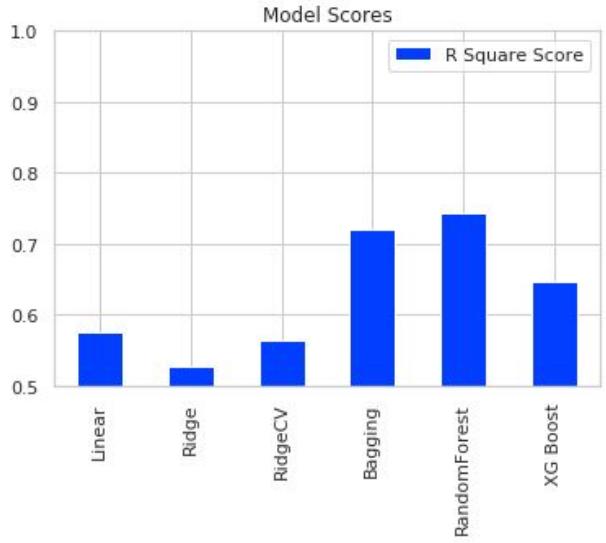
1) **Linear Regression** - Predicts a *dependent variable value (y)* based on a set of given *independent variables (x1, x2.. xn)*. The linear between xs (input) and y(output) is written as $y = \Theta_0 + \Theta_1 x_1 + \Theta_2 x_2 \ldots . \Theta_n X_n$
2) **Ridge Regression** - Ridge regression is 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.

Prior to building and training machine learning models, a Principal Component Analysis (PCA) was performed to reduce the dimensionality of features and to understand whether application of PCA improves the model performance or not.

PCA is a technique to reduce the dimension of the feature space by feature extraction. PCA captures as much variance as possible with a fewer new variables, accounting for a high percentage of variance in the original dataset.

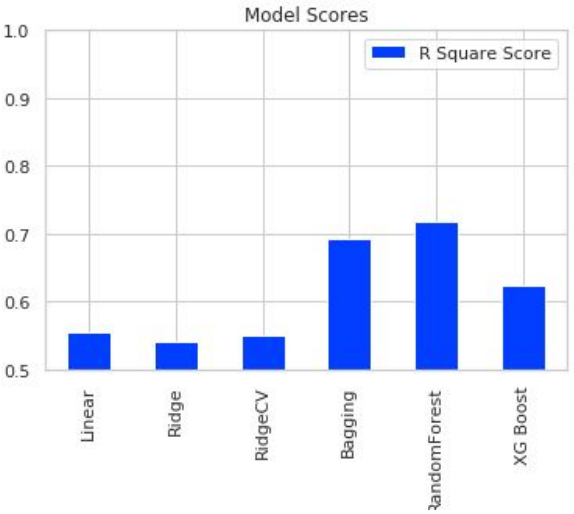
Several different regression models were trained on the training data and tested on the test set, with and without PCA to understand if PCA adds any benefit to the model. The results shown below are for the test set for both BC and NO2.

R2 with PCA for BC



	R Square Score
Linear	0.574383
Ridge	0.527219
RidgeCV	0.563863
Bagging	0.719717
RandomForest	0.742600
XG Boost	0.647342

R2 without PCA for BC



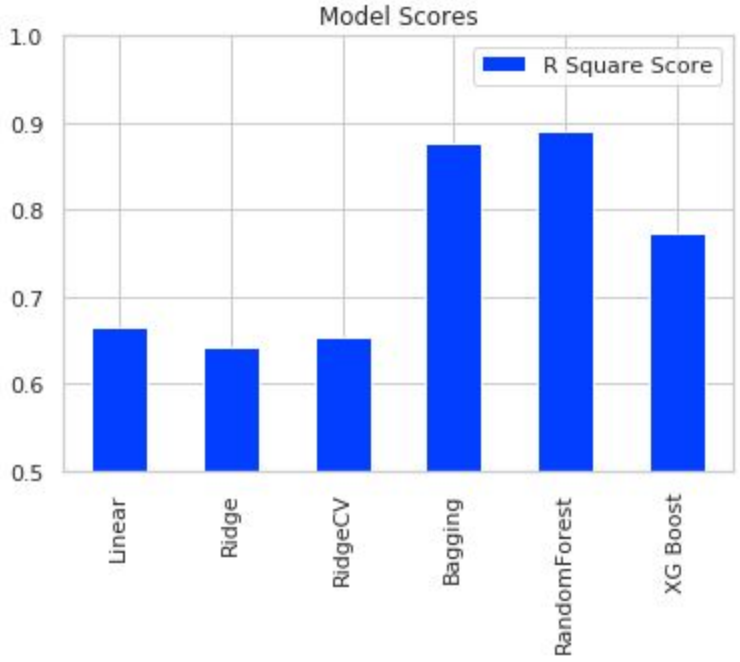
	R Square Score
Linear	0.555311
Ridge	0.541292
RidgeCV	0.550310
Bagging	0.691969
RandomForest	0.717569
XG Boost	0.624146

The above results show that the models for BC dataset perform slightly better with PCA for Linear Regression, Random Forest and XGBoost, but worse for Ridge Regression.

R2 with PCA for NO2

	R Square Score
Linear	0.637097
Ridge	0.598221
RidgeCV	0.627137
Bagging	0.867250
RandomForest	0.876575
XG Boost	0.763445

R2 without PCA for NO2



	R Square Score
Linear	0.663888
Ridge	0.641820
RidgeCV	0.654200
Bagging	0.877379
RandomForest	0.889643
XG Boost	0.773714

For the NO2 dataset, the Linear, Ridge, Random Forest and XGBoost method perform better without PCA.

Gridsearch different models for comparison

Next, 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. Four different models were built and tested including a simple linear regression model, ridge regression, random forest and XGBoost regression.

BC Dataset

Model parameters and R2 score for the BC dataset are shown below. *Figure 1* shows a scatter plot of the predicted vs actual concentrations for each model.

Model Type	**Best Parameter**	**Mean CV score for best estimator**	**Root Mean Squared Error**	**Training R2**	**Test R2**
Linear Regression with PCA		0.588	0.401	0.594	0.577
Ridge regression without PCA	Alpha = 0	0.588	0.401	0.594	0.577
ElasticNet without PCA	Alpha = 0 L1_ratio = 0.1	0.528	0.428	0.530	0.519
Random Forest with PCA	Max depth = 15 Estimators = 300	0.782	0.296	0.921	0.769
XGBRegressor with PCA	Learning rate = 0.1 max depth = 10 estimators = 100	0.798	0.282	0.925	0.791

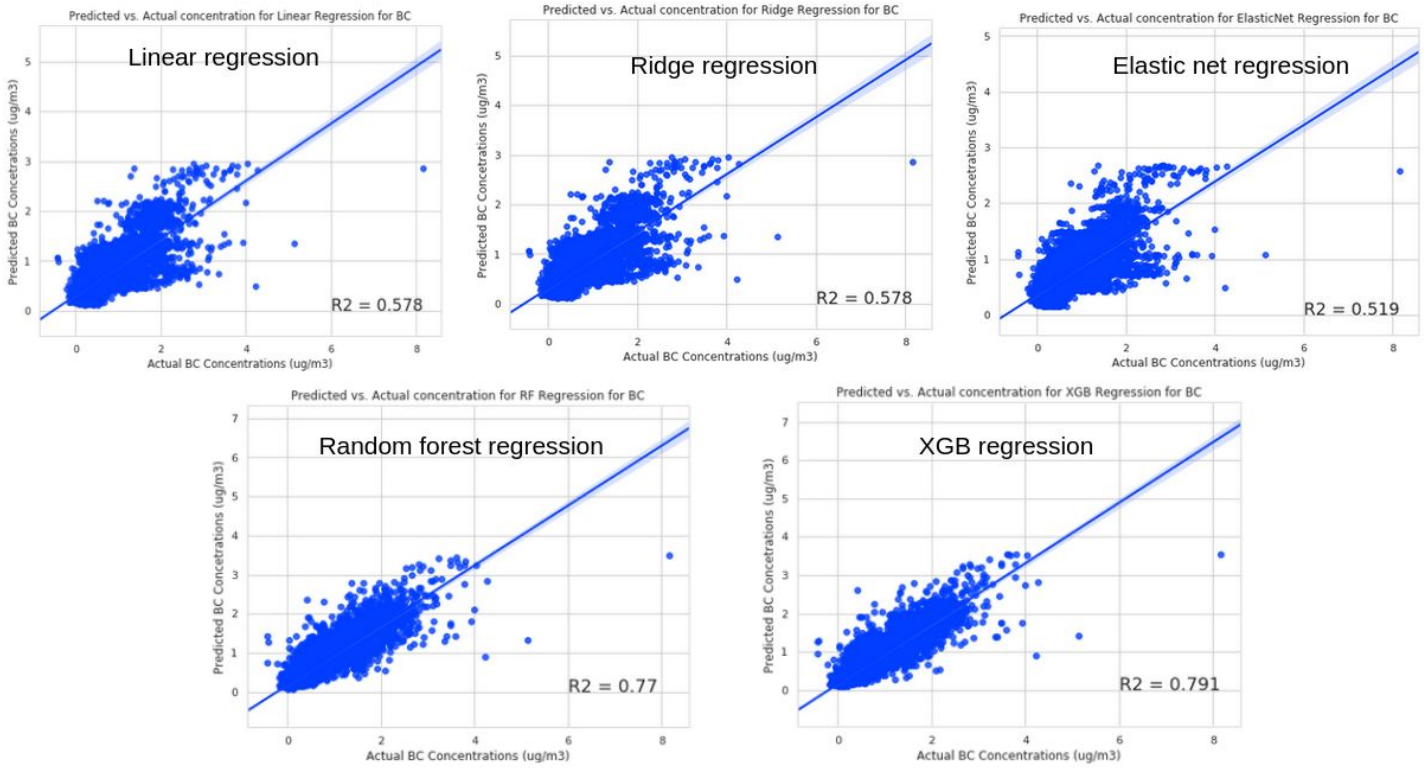


Figure 1: Scatter plot of predicted vs actual BC concentrations for different Machine Learning models

NO2 Dataset

Model parameters and R2 score for the NO2 dataset are shown below. *Figure 2* shows a scatter plot of the predicted vs actual concentrations for each model.

Model Type	**Best Parameter**	**Mean CV score for best estimator**	**Root Mean Squared Error**	**Training R2**	**Test R2**
Linear Regression without PCA		0.649	5.88	0.652	0.640
Ridge regression without PCA	Alpha = 0	0.649	5.88	0.652	0.640
Elastic Net Regression without PCA	Alpha = 0, L1_ratio = 0.1	0.579	6.43	0.580	0.570
Random Forest without PCA	Max depth = 50 Estimators = 400	0.888	3.127	0.985	0.898
XGBRegressor without PCA	Learning rate = 0.1 max depth = 10 estimators = 100	0.899	2.94	0.967	0.91

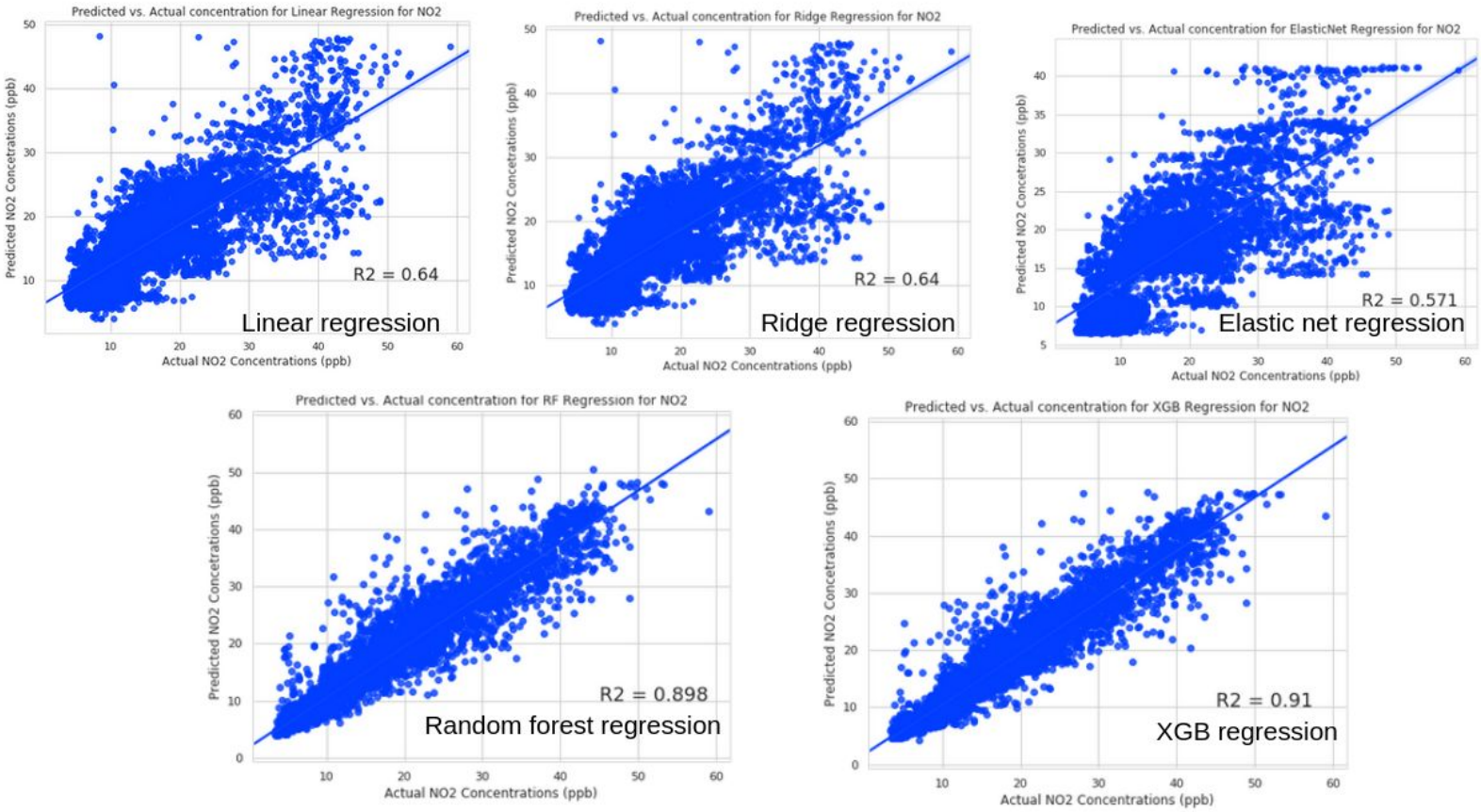


Figure 2: Scatter plot of predicted vs actual NO2 concentrations for different Machine Learning models

Summary of findings:

Results from the analysis show that tree based approaches such as Random Forest and XGBoost method performed far better than Linear or Ridge Regression models for both BC and NO2. The main reason for this is because tree based methods make no assumptions on the relationship between features. If two features are highly correlated, then little or no information is gained from splitting the second feature after splitting on the first feature. As a result, the second feature is dropped in favor of the next one.

The Random Forest method results in an R2 value of 0.77 for the BC dataset (RMSE = 0.296) and 0.898 for the NO2 dataset (RMSE = 3.127). The XGBoost method results in a score of 0.791 for the BC dataset (RMSE = 0.282) and a prediction score of 0.91 for the NO2 dataset (RMSE = 2.94). Since multicollinearity is a big issue in this dataset, tree based methods outperform linear regression methods.

We previously noted that the top five features that contribute to BC concentration in the region includes a soft drink manufacturing company (SVC manufacturing Inc, division of Pepsi Co), electricity generation units located in the Berkeley campus, a waste treatment facility, an office building located on Clay Street in Oakland, and the Port of Oakland.

The Port of Oakland, the manufacturing facility, electricity generating units and waste treatment facilities are all large sources of PM emissions in the area. Here PM data is used as a proxy for BC.

For the NO2 dataset, the top five features that contribute to concentration in the region includes a wholesale facility, a residential complex with generators, a commercial printing facility (Consolidated Printers Inc.), annual precipitation, and a coffee roasting plant (Peets Coffee roasters). Generators, natural gas burners used in coffee roasting plants, and printing facilities are major sources of NO2 in the area.

While the Random forest approach gave us some insight into the features that are important, this still does not give us the major sources of emissions (or features) that contribute to air quality in a local area. 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. Emissions from vehicles and trucks traveling on highways are actually one of the highest contributors to BC and NO2 emissions on a local-level, and lack of emissions data from traffic is one of the limitations of this work. Based on my analysis so far, I observed that it is an aspirational goal to try and predict the major sources of pollution at a local level, even though the model was able to identify some features that are important.