M. ENG PROJECT REPORT

Name: Soham Ray (sr2259), Sudhanshu Khoriya (sdk225)
Title: Data-driven, hyper-local air quality modeling in global metropolitan areas
Professor: Dr. K. Max Zhang (kz33)

INTRODUCTION

Air pollution levels today are higher than ever, and pose a critical threat to well-being. Under the guidance of Professor Zhang, I have gathered data for, designed and constructed a machine learning model to predict air pollution hotspots and possible emission sources, hyper-locally. We focused on 2 locations: West Oakland and London.

DATA

The data had been collected from various sources which has been described in the manual . Here is the description of the features used and their importance for building the final machine learning model.

Features

1) NO_2/PM emissions
   We calculated the NO_2/PM concentrations from the emissions inventory which is present for all the cities in consideration and it contains the annual average of the emission data of the pollutants. For this feature we considered the average concentrations of the emission sites within 500m buffer radius of our sample points. We considered the option of using distance to each emission site as a feature but the emission sites were around 50-60 in number and adding 50-60 more features lead to sparsity in data and most of the features were given very less importance or weights.
2) Bus stops, Restaurants, Subway stations, Gas stations
   For these features we calculated the count of all the above places within a buffer radius of 500m from the sample point to get more accurate prediction of the impact of these features on the air quality at that point as increasing the buffer distance would lead to spatial inconsistency and would have added unnecessary importance and impact of the all the above features.
   For West Oakland we did not consider different types of restaurants as the number of restaurants were very few around 15 and categorizing them was not possible. For London annual average model we also considered a feature as number of restaurants within 500m of our sample of a particular category , as different categories of restaurants amount to different amounts of pollution. There were 8 categories like barbecue, Indian, Chinese, Arabian etc and we calculated the number of each of these categories in a radius of buffer length 500m.

3) Traffic and land features:
   The traffic features which were considered were no of buses and coaches, no of cars and motor vehicles on the roads which are within 500m radius of our sample points to maintain uniformity in the features as we had already established the impact of the pollution causing agents in the vicinity of the sites. The number of vehicles of these roads is an effective measure of the pollution on these roads. The number of vehicles here signifies the traffic density of these roads at the particular time period.

   The road features used were: link length 500m, link length within 200m , primary road within 200m , primary within 500m . Similarly features were found out for secondary roads and tertiary roads. These are important where the actual traffic data is missing but the information about the road use is present and having the data about different categories of roads can be useful for predicting the actual concentration of pollution in those sites as more number of roads implies more pollution . All these data was found on GIS and other traffic sources as listed in the manual.

4) Meteorological features
   The meteorological features that we used in the London model (both PM and NO2) were temperature, dew point, humidity, wind speed and pressure. These features are required for the temporal model as these also affect the air quality and give a good estimate about the concentration of the air pollutants. We took the annual average of each of the mentioned features at each of the measurement sites and used them for model training . But as we had meteorological data only from one site we could not get the spatial difference in these features for annual average model. We can instead use the meteorological data in the daily average model

5) AOD
   Aerosol Optical depth data can be generated from NASA's database and it gives the AOD quantity for a particular region over a period of time. This feature adds to the above features in determining the air quality and is included in the final model.

6) Latitude and Longitude
   These were used as features in order to maintain spatial variance in the West Oakland model. In the London Model using regression we found out that these were not very useful and the results using them as features were not encouraging so we decided to drop them.

Also , we couldn't find any reliable source of data for recent population density in these areas and hence it is not considered as a feature .

We performed an ablation study of the features and found that the airport feature which was distance of all the airports from the sample measurement points do not carry much importance, so

it was omitted from the final model to give good accuracy. Also the buffer size study resulted in 500m being the optimal size of buffer. The results for 1000m were not so good as air quality changes a lot within 1000m radius and it causes inconsistency in the results. For buffer size 200m the data was very sparse with a lot of features being zero and hence it was concluded that 500m is the ideal size of buffer.

DATA PREPROCESSING

Before building the model, we first convert the data into a format that can be understood by it. The various features are co-related by latitude/longitude (spatially) or data/time (temporally). Land use features, bus stops, gas stations, traffic and road features as well as AOD (meteorological) data are merged with NO_2, $PM_{2.5}$ concentration based on latitude and longitude. For example, if we have land use features for a particular latitude longitude, and we have the NO_2 concentration for the same latitude longitude, we merge the features. Once this is done, the data is normalized to improve prediction accuracy. In addition, outliers and NaN values are removed as well. Finally, this data is fed into the model to generate predictions.

MODEL BUILDING AND EVALUATION

West Oakland

For West Oakland, we were able to gather multiple predictors over a large amount of data, as described above. Since we had a large data set, we decided to go with a deep learning approach. In this case, we used feed forward neural networks. Ideally, with temporal data, we could use a recurrent neural network but unfortunately, we don't have time-series data. So we tried multiple approaches to feed forward neural networks.

We try to classify areas in West Oakland by their pollutant concentration, so we can create a heatmap in the future. Before training, we normalized and cleaned all data. Next, pollutant concentration was divided into five categories: Very low, low, medium, high and very high. Once normalized, all values are between 0 and 1. For both pollutants, we considered pollutant concentration levels above 0.8 as very high, above 0.6 as high, above 0.4 as medium, above 0.2 as low and below 0.2 as very low. Locations with very high pollutant concentration are considered to be pollution hotspots.

Initially, we started with a very small model of 2 hidden layers and 16 hidden units each. However, this seemed to overfit the data since despite getting high training accuracy, validation accuracy was low. With further tweaking, results improved. A lower learning rate with a larger network seemed to solve the overfitting problem. Finally, we applied Bayesian optimization, where we provided the hyperparameters as features and loss as output. Since Bayesian Optimization looks

through the features trying to minimize output, we finally get the ideal hyperparameter set, as highlighted in the results table. Since our dataset was adequately large, we chose a 90:10 train test split to create a train dataset and a test dataset.

Features used (same features for both pollutants):

- Land Use: Distance and count of primary, secondary and tertiary roads for buffer distances 200m, 500m and 1km. Distance and count of residential roads in 500m. Distance from all 8 airports in the city.
- Temporal: Date and time
- Spatial: Latitude, Longitude
- Other: Car speed. Number of restaurants, bus stops, gas stations and subway stations in 500m
- Emission Inventory: NO_2 emissions are used for NO_2 prediction and PM emissions are used for PM prediction

Results for NO_2

No. of layers	Hidden Units/layer	Optimization Function	Learning Rate	Training accuracy (NO_2) in %	Test accuracy (NO_2) in %
2	16	Adam	0.01	43	20
2	32	Adam	0.01	44	35
2	**64**	**Adam**	**0.001**	**59**	**56**
2	64	SGD	0.001	52	51

Results for BC

No. of layers	Hidden Units/layer	Optimization Function	Learning Rate	Training accuracy (BC) in %	Test accuracy (BC) in %
2	16	Adam	0.01	44	20
2	32	Adam	0.01	44	32
2	**64**	**Adam**	**0.001**	**60**	**54**
2	64	SGD	0.001	53	49

<u>London</u>

Here, we first use the Breathe London 2019 dataset which gives us pollutant concentration and location. However, for annual average data, we only have around 100 points for NO_2 and around 80 points for $PM_{2.5}$. This greatly restricts performance.

With such little data, neural networks are not a possibility. Neural networks typically need a large dataset for training. With only 100 datapoints, deep neural networks are bound to underfit and provide inaccurate results. Also, other classifiers are shown to provide better results than shallow neural networks. Thereby, we use different regression algorithms from the scikit-learn library. Results are given in the table below. Initially, we only predict pollution hotspots (i.e, locations with very high pollutant concentration). To do this, we compared our normalized validation data with our predicted data such that if both state very high pollutant values [top 15% of pollutant concentration], we consider it as correct. This gave us an accuracy of **59%**. We also evaluate the models based on root mean squared error between predicted and true pollutant concentration. This time, due to data sparsity, we use a 80:20 train test split to create the training dataset and test dataset. The results are provided for test dataset.

Features used (same features for both pollutants):

- Land Use: Distance from primary, secondary and tertiary roads for buffer distances 200m, 500m and 1km. Distance from all 8 airports in the city.
- Traffic: Number of two wheeled motor vehicles, cars and taxis, buses and coaches, heavy vehicles with 2, 3, 4 and 6 axles
- Other: Number of restaurants, bus stops, gas stations, restaurants and subway stations in 500m
- Emission Inventory: NO_2 emissions are used for NO_2 prediction and PM emissions are used for PM prediction

Results for NO_2

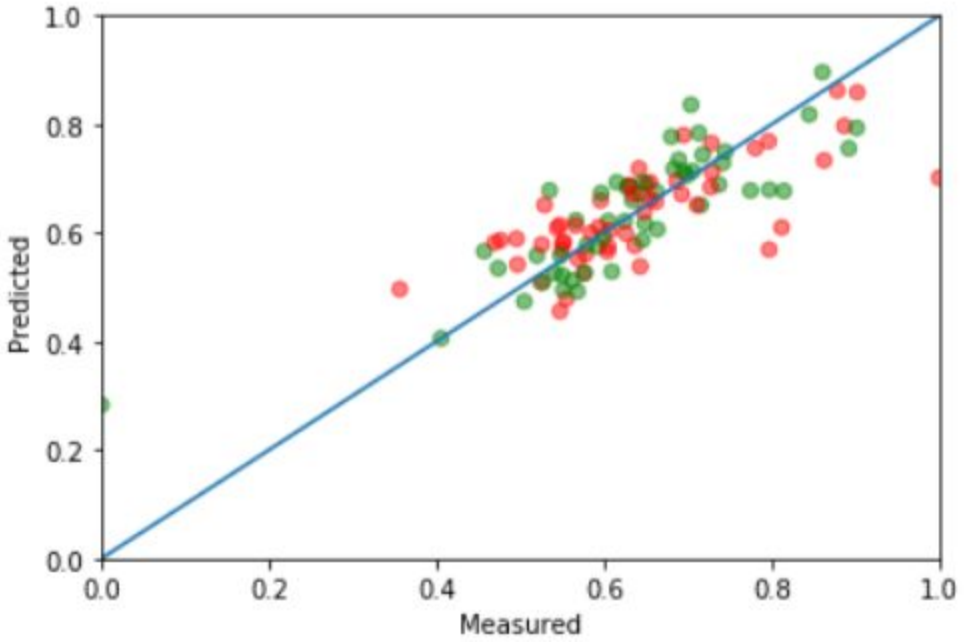
Model	Hotspot Prediction accuracy for NO_2	Root Mean Squared Error for NO_2	R2 score for $PM_{2.5}$
Random Forest	41%	0.06	0.77
Linear Regression	**58%**	**0.08**	**0.69**
Multilayer Perceptron	0%	Very large	-611

Results for $PM_{2.5}$

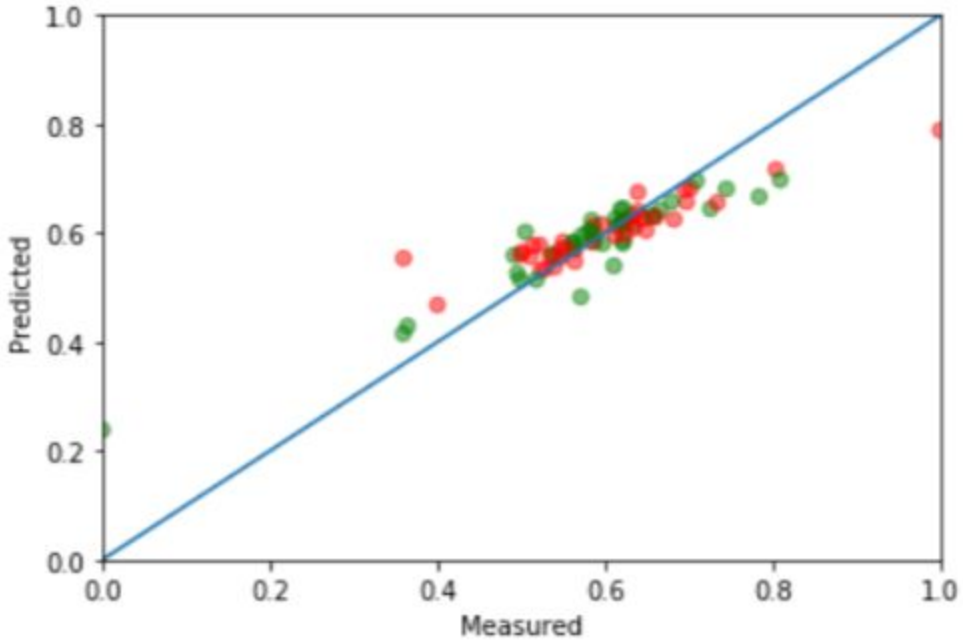
Model	Hotspot Prediction accuracy for $PM_{2.5}$	Root Mean Squared Error for $PM_{2.5}$	R2 score for $PM_{2.5}$
Random Forest	**25%**	**0.062**	**0.72**

Linear Regression	25%	0.07	0.57
Multilayer Perceptron	0%	Very large	-339

Breathe London - ScatterPlot of normalized values of Y_true vs Y_pred for RandomForest Regressor (NO_2)



Breathe London - ScatterPlot of normalized values of Y_true vs Y_pred for RandomForest Regressor ($PM_{2.5}$)



We also use another dataset for London, London Air dataset. The annual average model for this has even lesser datapoints (approximately 30 for $PM_{2.5}$ and 30 for NO_2). Again, this greatly restricts performance and we get similar results over the same predictors.

Features used (same features for both pollutants):

- Spatial: Site name, latitude and longitude
- Traffic: Number of two wheeled motor vehicles, cars and taxis, buses and coaches, heavy vehicles with 2, 3, 4 and 6 axles
- Other: Number of restaurants, bus stops, gas stations and subway stations in 500m
- Emission Inventory: NO_2 emissions are used for NO_2 prediction and PM emissions are used for PM prediction

Results for NO_2

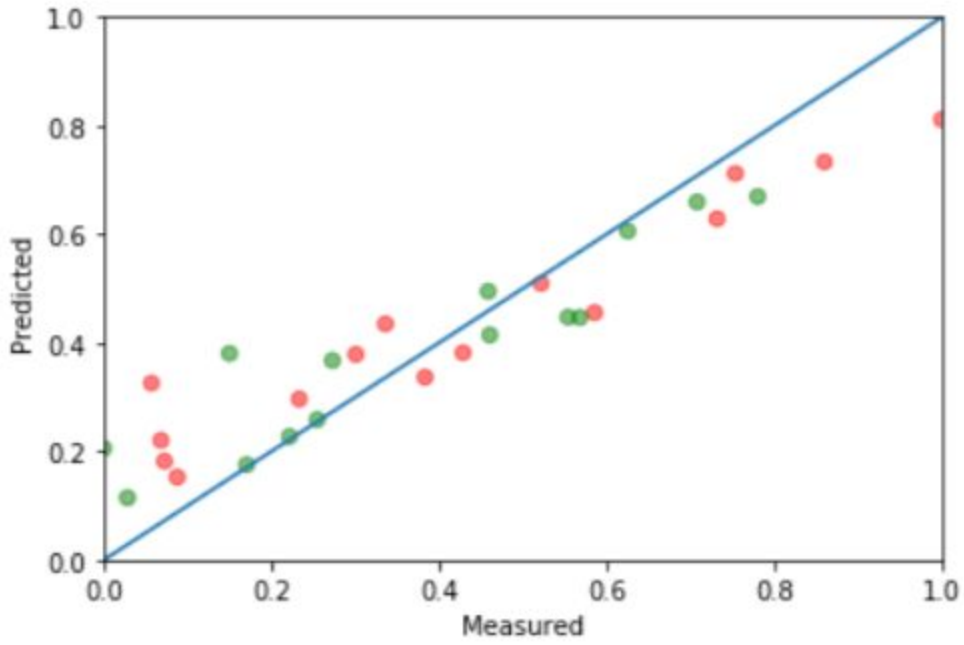
Model	Hotspot Prediction accuracy for NO_2	Root Mean Squared Error for NO_2	R2 score for NO_2
Linear Regression	57%	0.17	0.61
Random Forest	**83%**	**0.11**	**0.82**
Multilayer Perceptron	0%	2.75	-100

Results for $PM_{2.5}$

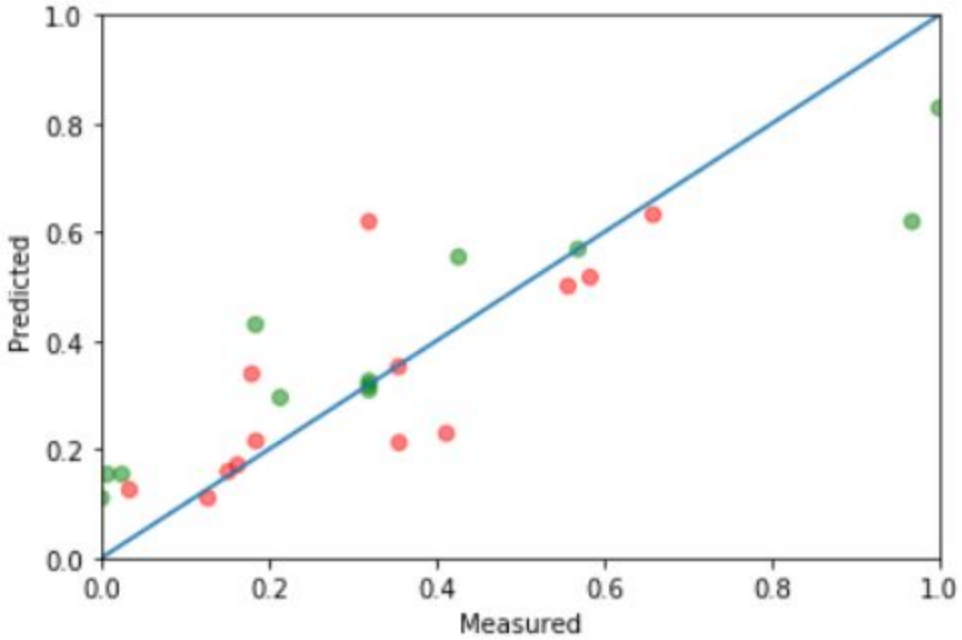
Model	Hotspot Prediction accuracy for $PM_{2.5}$	Root Mean Squared Error for $PM_{2.5}$	R2 score for $PM_{2.5}$
Linear Regression	33%	0.18	0.49
Random Forest	**33%**	**0.13**	**0.72**
Multilayer Perceptron	0%	0.47	-2.33

The high accuracy can be misleading since our dataset is very small.

London Air - ScatterPlot of normalized values of Y_true vs Y_pred for RandomForest Regressor (NO_2)



London Air - ScatterPlot of normalized values of Y_true vs Y_pred for RandomForest Regressor ($PM_{2.5}$)



IMPLEMENTATION STEPS

1. Import libraries
    a. Numpy
    b. Scikit
    c. TensorFlow
2. Data Preprocessing
    a. Replace NA values
    b. Correlate different datasets based on spatial features
        i. Scripts: xxx
    c. Split data into train and test or use k-fold cross validation
3. Model building and evaluation
    a. Tried with multiple classification/regression techniques
        i. Random Forest Regression
        ii. Logistic Regression
        iii. Decision Trees
        iv. Linear Regression
        v. Multi-Layer Perceptron
        vi. Neural Network using Tensorflow
    b. Used classification accuracy/RMSE for evaluation
4. HyperParameter Tuning
    a. Import scikit optimize
    b. Use gp_opt bayesian optimization to select best hyperparameters
    c. Record best model and best accuracy

CODE

- Annual Average data model: xxx
- Breathe London data model: xxx
- West Oakland Neural Network Model: xxx

CONCLUSION

With minimal data and features, we are able to predict some form of air pollution. In the future, with more data, we should be able to more accurately predict air pollution and also label its sources. Additionally, we can focus on automating this process so it can run for any location.