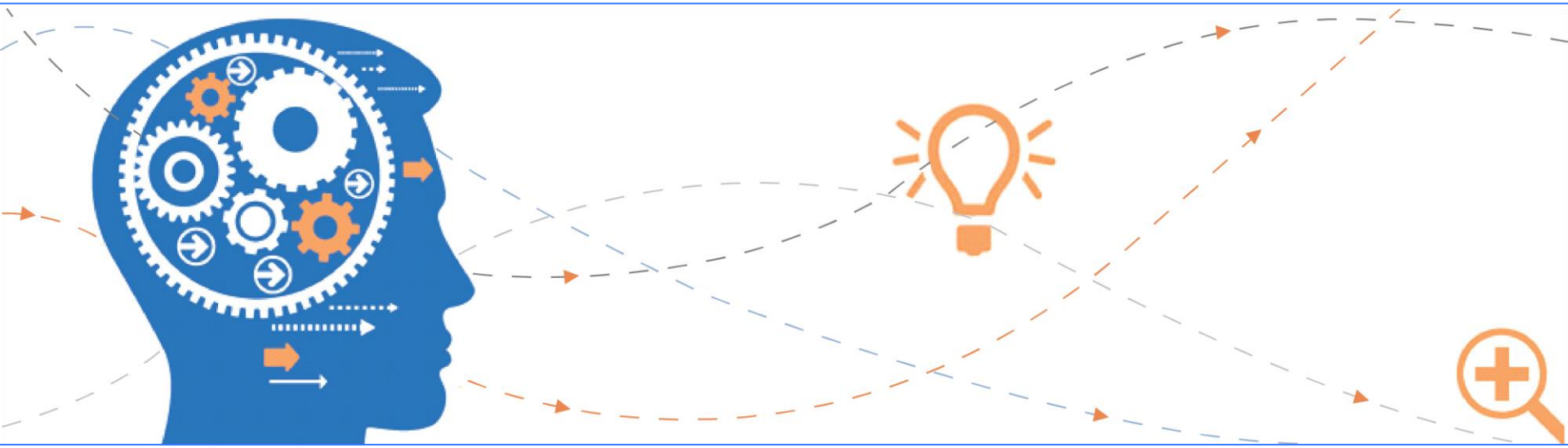
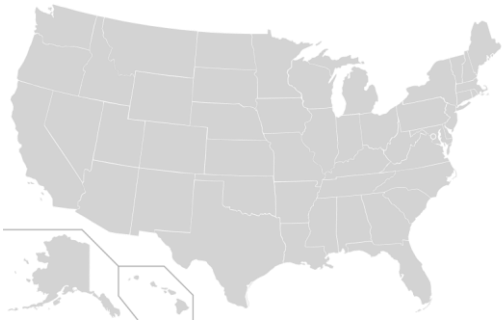
Forecasting fine grained pollutant levels

Nilesh Patil *Jiang Shang*





University of Rochester

Background

What are fine grained pollutants?

- The term fine grained pollutants refers to the particles in atmosphere having size smaller than 2.5µm.
- Natural sources : Oceanic salt sprays , Volcanic activity, storms, forest and grassland fires etc.
- In recent times, human activities like increasing usage of fossil fuels has led to significant jump in anthropogenic aerosols (those made by human activity) and it now accounts for about 10% of total
- PM2.5 is particularly dangerous with a 36% increase in lung cancer per 10 µg/m3 beyond current safety standards

Problem Statement

- Building a predictive model for PM2.5 levels based on historical conditions
  - Input : Temperature, Pressure, RH, Wind speed etc.
  - Output : PM2.5 level in next 24 hours

Potential Significance

- Health alerts
- Supplementing pollution control measures

Methodology Overview

Data

Data source

- EPA(United States Environmental Protection Agency)

Data munging

- From hourly readings to daily averages
- 579 counties spread across US
- Weekly averages for each county
- Monthly average for each county
- Each row represents one county, one day and corresponding values for the previous day
- Replace missing values based on matrix imputation instead of mean for each variable

Parameters used

- Temperature, Pressure, Relative Humidity, Wind speed, Dew point, Ozone, SO_2, CO, NO_2

Model

General modeling approach:

- Use historical readings for physical parameters
- Use month and weeks as categorical inputs

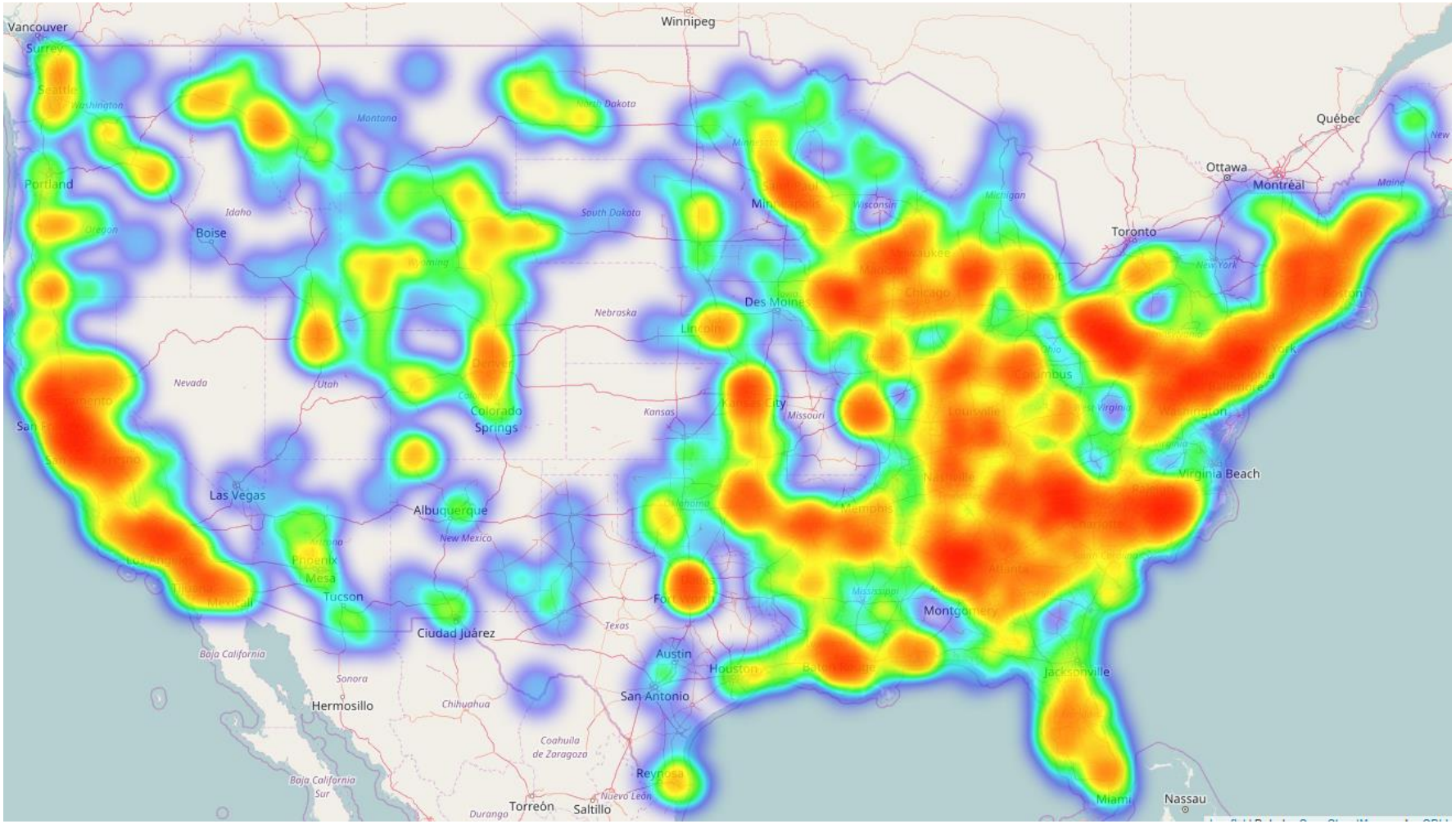
Algorithms of choice

- Linear Regression
- Random Forest
- Xgboost

Validation

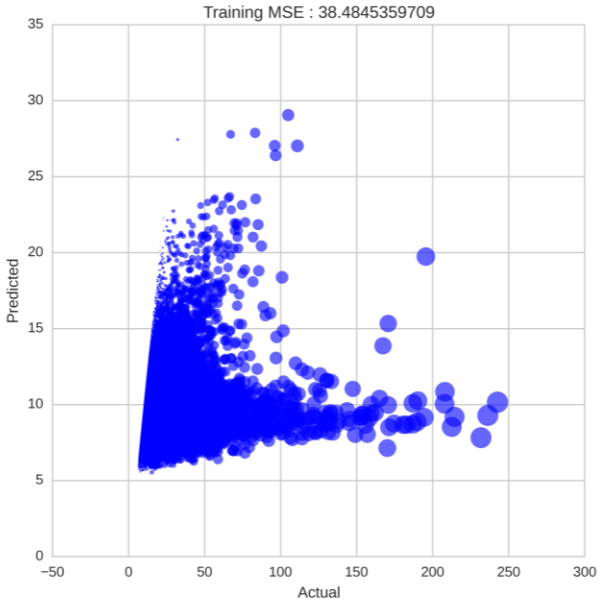
- 70:30 training-test division
- 10 fold cross validation for training dataset (linear regression)

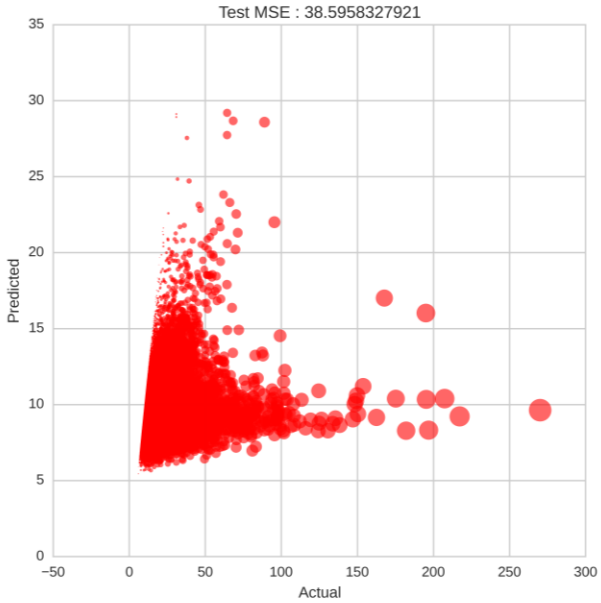
Relative average PM values (yearly)

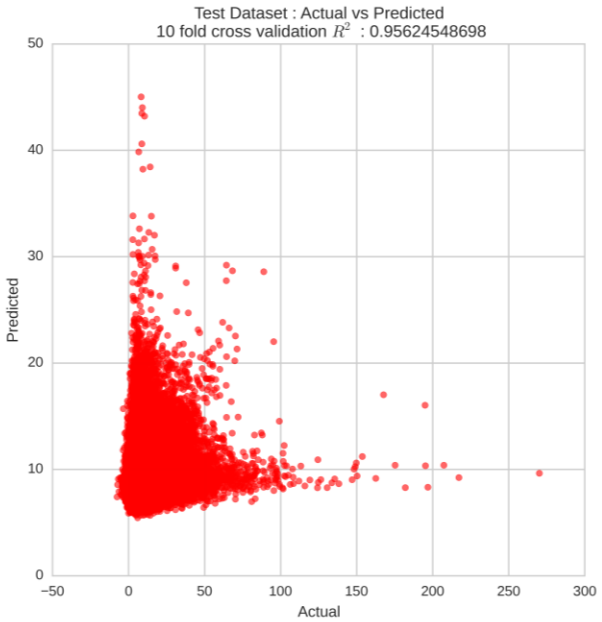


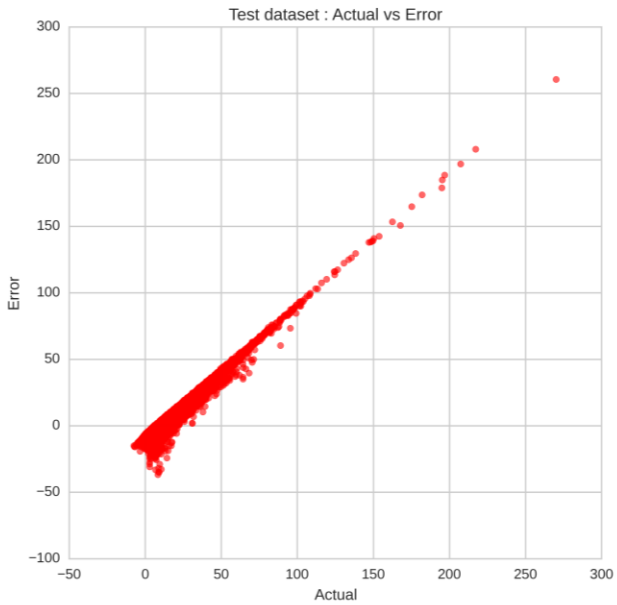
PM2.5 concentration across USA : correlates highly with human activity and sea face

Linear Regression









Easier to analyze the model & straight forward to explain using just linear coefficients

Predictions :

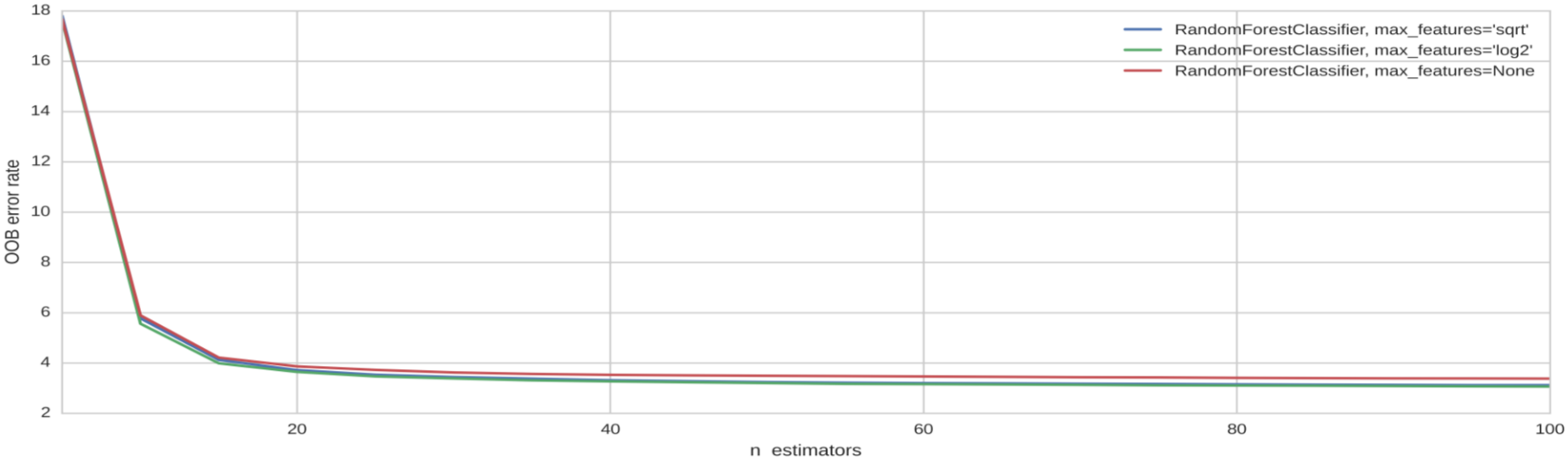
First 2 plots show results on training and testing datasets with size set to the error

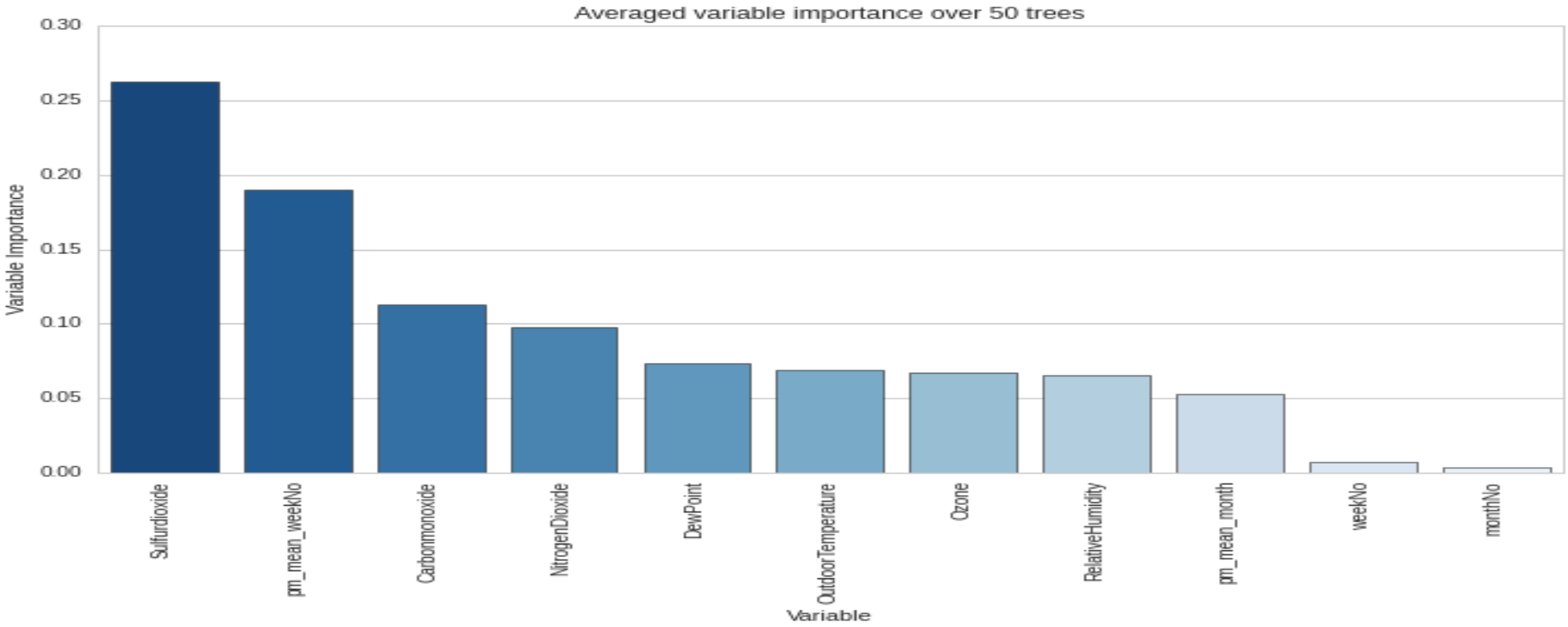
As shown, the linear regression model is not very accurate even though MSE and R2 are low, especially R2

The error increases linearly w.r.t. The actual PM2.5 value

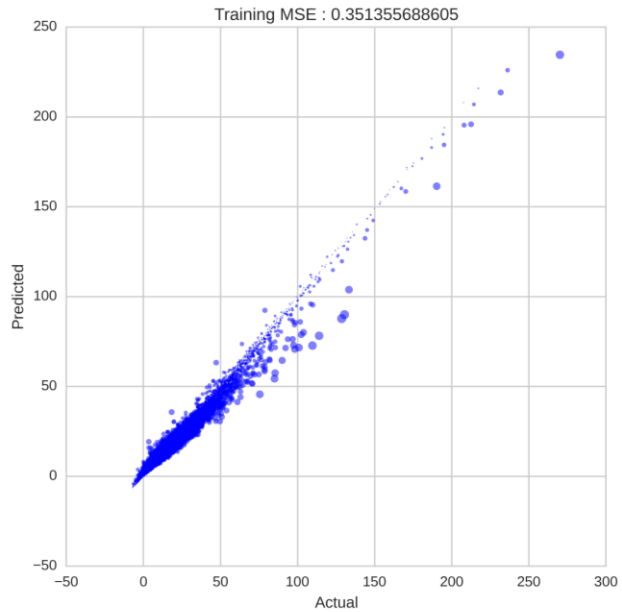
Even at lower PM values, the model isn't predictive at all

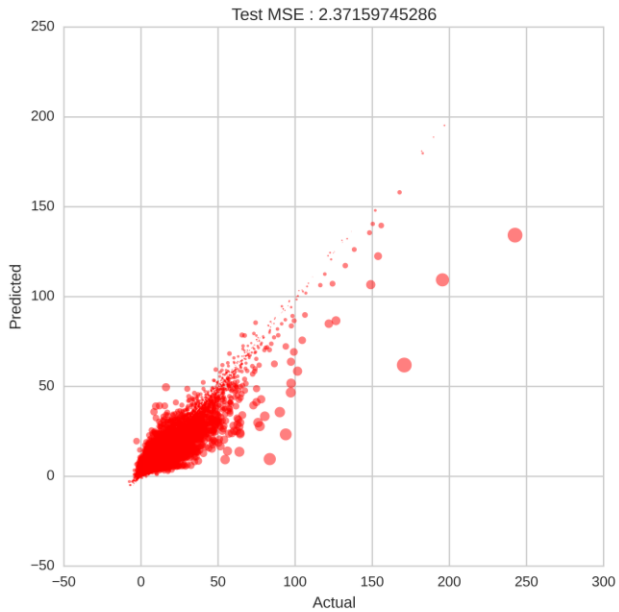
Random Forest - preliminary parameter search

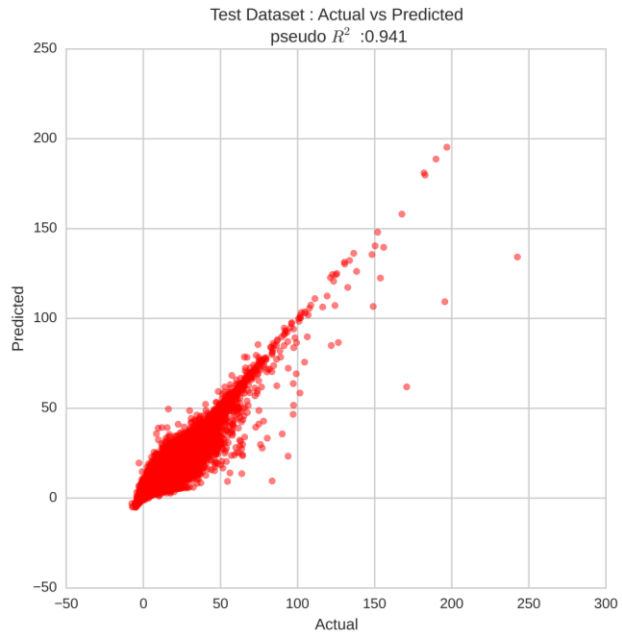


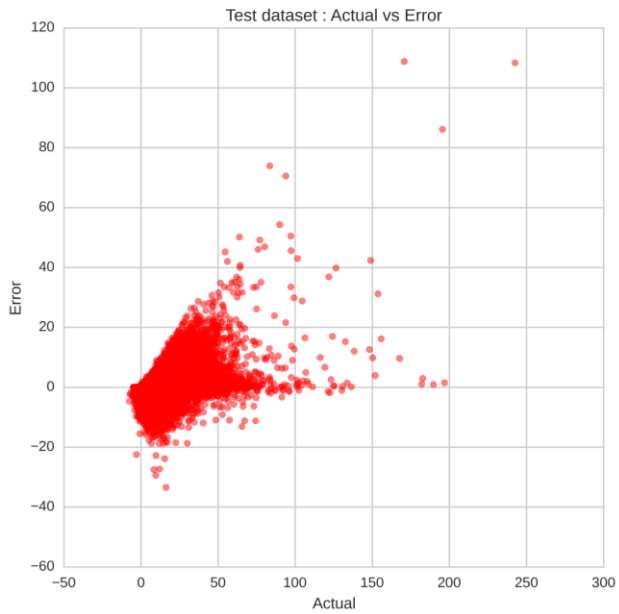


Random Forest – current model results









Model:

Using grid search over a limited space, we used 35 trees and nvars at each split = log2(subset_vars)

Predictions :

- First 2 plots show results on training and testing datasets with size set to the error
- The third plot lays out predicted vs actual only and as we can see, our random forest model is doing a much better job than linear regression model, yet there are a few values being predicted incorrectly by huge margins as shown in the second plot
- The fourth plot shows error distribution w.r.t. actual values in the test dataset

Problems & further work

Missing values :

- Using matrix imputation, we filled in the missing values
- We used :
  - multiple matrix imputation averaged for each value
  - Multivariate Imputation by Chained Equations
- The results aren't representative of the original dataset, most values are close to zero and we still haven't figured out how to solve this issue.
- One way is to use convex optimization based & KNN-based matrix imputation but the current implementations require 4.3 Tb memory

Variables :

- We have used historical PM aggregates (weekly, monthly)
- We still haven't tried out historical values for other parameters
- We would welcome suggestions for additional data that can be included on the city level.

Reference

I. *Yu Zheng, Furui Liu, and Hsun-Ping Hsieh. 2013. U-Air. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13 (August 2013). DOI:http://dx.doi.org/10.1145/2487575.2488188*

II. *Anon. Air pollution and lung cancer incidence in 17 European cohorts: prospective analyses from the European Study of Cohorts for Air Pollution Effects (ESCAPE). Retrieved July 10, 2013 from https://www.ncbi.nlm.nih.gov/pubmed/23849838*

III. *https://aqsdr1.epa.gov/aqsweb/aqstmp/airdata/download_files.html*