Capstone Project Proposal

Goal

This project proposal is based on the dataset for a Kaggle competition from 2015. The goal is to establish a model that can predict the dangerous levels of air pollutants on an hourly basis.

Client

The EPA's Air Quality Index is used daily by people suffering from asthma and other respiratory diseases to avoid dangerous levels of outdoor air pollutants, which can trigger attacks. According to the World Health Organisation there are now estimated to be 235 million people suffering from asthma. Globally, it is now the most common chronic disease among children, with incidence in the US doubling since 1980. The model we build could be used as the basis for an early warning system that is capable of accurately predicting dangerous levels of air pollutants on an hourly basis.

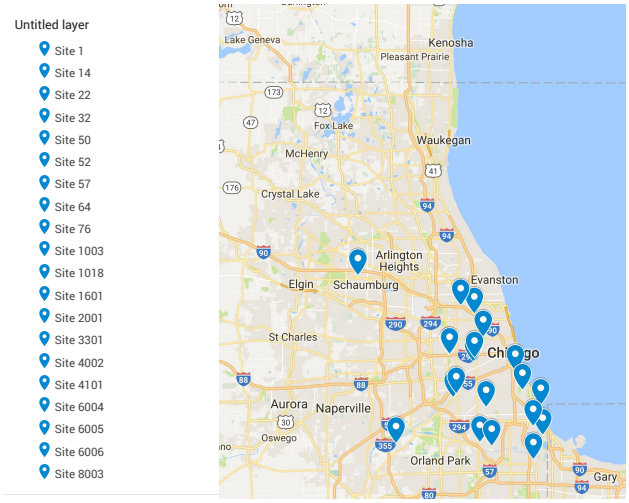
Data

Hourly data on various targets were collected for 8 days from various sites around Chicago was provided as training data. The aim of the challenge is to predict various time points within the next 3 days after the training period (1, 2 ,3, 4, 5, 10, 17, 24, 48, and 72 hours after the end of the 8-day training data).

Target Code:

measured_quantity	PARAMETER_DESC
target_8	Carbon monoxide
target_4	Sulfur dioxide
target_3	SO2 max 5-min avg
target_10	Nitric oxide (NO)
target_14	Nitrogen dioxide (NO2)
target_9	Oxides of nitrogen (NOx)
target_11	Ozone
target_5	PM10 Total 0-10um STP
target_15	OC CSN Unadjusted PM2.5
target_2	Total Nitrate PM2.5 LC
target_1	EC CSN PM2.5 LC TOT
target_7	Total Carbon PM2.5 LC TC
target_8	Sulfate PM2.5 LC

Mapping of the Sites in EMC Data Set.



Summary of the data on Targets and Sites:

		Sites																		
		1	14	22	32	50	52	57	64	76	1003	1018	1601	2001	3301	4002	4101	6004	6005	6006	8003
Targets	target_1							x													
	target_2							x													
	target_3	x				x		x					x			x				x	
	target_4	x				x		x				x	x	x		x	x			x	x
	target_5																			x	
	target_7							x													
	target_8							x								x		x			x
	target_9															x					x
	target_10															x					x
	target_11	x			x	x			x		x		x			x					x
	target_14															x					x
	target_15							x													

Project Outline:

1. Data Wrangling
2. Test Models
3. Optimize Models

Deliverables:

1. Models and Prediction Values for each “target” variable.
2. Evaluation of prediction model by mean absolute error score.
3. Paper / Slide deck