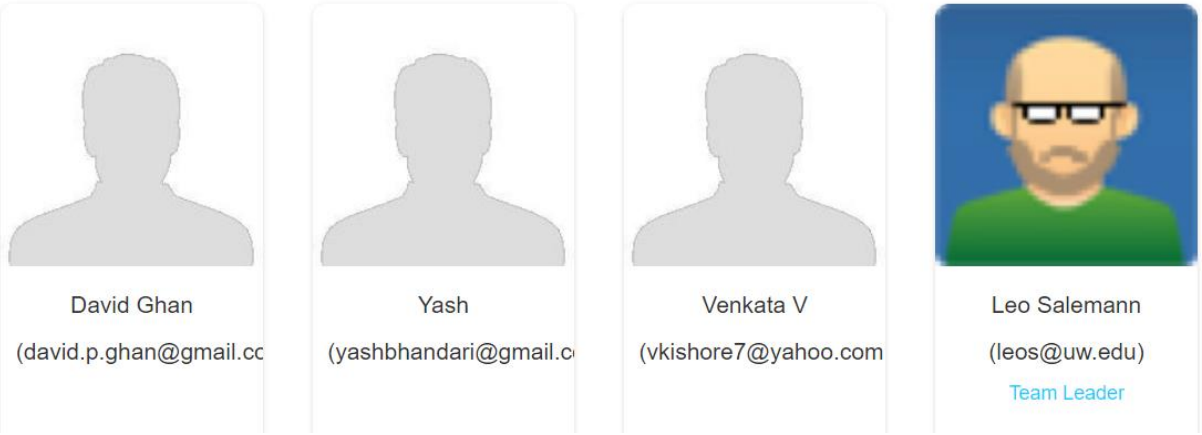
DS420-Cobras
DATASCI 420 BB
Capstone Project, Mid-way Checkpoint
5/8/18

Team Roster:



Leaderboard Ranking: 143 (as of 5/8/18 4:31 pm PDT)

143	↑ 14	DS420-Cobras	2.00	2.00	0.36	0.44	0.54	1.07	5

Current Status

Data Sources

Item	Source	Notes
Beijing Air Quality	https://biendata.com/competition/airquality/bj	
London Air Quality	https://biendata.com/competition/airquality/ld	
Weather Data	Darksky.net via python forecastio library, as described in https://biendata.com/forum/view_post_category/139	Darksky gives us current, historical and forecast data for for specific latitude/longitude.
Beijing Air Quality Station Locations	https://www.dropbox.com/s/5lhxontpbfoyemi/Beijing_AirQuality_Stations_en.xlsx?dl=0	Provides latitude & longitude
London Air Quality Station Locations	https://www.dropbox.com/s/nuy1r6psk46vsi4/London_AirQuality_Stations.csv?dl=0	Provides latitude & longitude
Terrain Map, Backdrop Image	OpenStreetmap, Microsoft PowerBI, MapBox Visual for Power BI	Used for plotting AQP points with map backdrop and manually recording terrain type (flat, mountainous, …) in an handmade spreadsheet.

Satellite Imagery	DigitalGlobe, Microsoft PowerBI, MapBox Visual for Power BI	Used for plotting AQP points with map backdrop and manually recording environment type (park, suburbs, city, forest, …) in an handmade spreadsheet.

Notes

Earlier in the project we were using Beijing and London grid and meo points, but we've dropped these in favor or location-specific weather data from darksky.net.

Feature Engineering

- Perform an inner join between AQI stations and weather data on station_id & time
- Remove duplicate rows
- Drop features with lots of null/empty values.
    - `ozone`
    - `precipIntensity`
    - `precipProbability`
    - `pressure`
    - `uvIndex`
    - `windGust`
    - `cloudCover`
    - `precipType`
    - `visibility`
- Break timestamp into hour, day, month, day of week
- Drop Unneeded pollution features
    - Beijing
        - `CO_Concentration`
        - `O3_Concentration`
        - `SO2_Concentration`
        - `NO2_Concentration`
    - London
        - `NO2_Concentration`
        - `CO_Concentration`
        - `SO2_Concentration`
- Replace negative values with NaN for the following features
- Other work (now deprecated)
    - Developed an aq/met lookup table that mapped each AQI station to its mearnest meo or grid station, though direct observation.

Models

We're running multiple simultaneous models with python/scikit-learn:

- Means Fit(): Return the historical mean for each individual AQI station
- Random Forest Regression
- Linear Regression (currently commented out).

Future Plans

- Investigate Additional data sources available via KDD Cup forums, such as AirVisual data (https://biendata.com/forum/view_post_category/140)
- Refine our feature engineering (replace nulls with means, or something better than outright deletion)
- Gather additional features though map observation (proximity to roads; water, etc.).
- Add more models to the ensemble, such as GBM