Air Quality Prediction:
U.S. Counties 2017-2019

•••

DS5110 Semester Project

Daniel Heffley (dh3by)
Camille Leonard (cvl7qu)
Shah Shahrokhabadi (ss3qs)
Stephanie Verbout (sv8jy)

Executive Summary

- Air Quality Index (AQI)
  - A daily measure of how polluted the air is in a certain jurisdiction
- Research Questions
  - Can we use daily readings of gas, pollutant, particulate matter and meteorological data with socio-economic data to predict daily AQI?
  - Can we use gas, pollutant, particulate matter and meteorological data to predict median income?
- Random Forest (RF) Regression and Gradient-Boosted Tree (GBT) Classifier
  - Predicted AQI extremely well
  - No model predicted median income adequately

Air Quality Index (AQI) Values	Levels of Health Concern	Colors
0 to 50	Good	Green
51 to 100	Moderate	Yellow
101 to 150	Unhealthy for Sensitive Groups	Orange
151 to 200	Unhealthy	Red
201 to 300	Very Unhealthy	Purple
301 to 500	Hazardous	Maroon

Data Summary

- Environmental Protection Agency (EPA)
  - Air Quality System (AQS) database
  - Supplemented by socio-economic data from OpenIntro
- Response Variables
  - AQI and Median Income
- Predictors
  - Daily readings from air quality monitors
  - Pollutant, gas, particulate matter and meteorological data
  - Population density and median income
- Sampling
  - Taken from 2017-2019 EPA data, 2010 Census

Transformations/Preprocessing

- Merge daily criteria gas, pollutant, meteorological, and AQI readings
  - Supplement with county-level socio-economic information
    - Median income
    - Population density
  - Dropped categorical variables and dates
- Impute missing data for predictor columns with median and drop missing AQI values
- ‘Healthy’/’Unhealthy’ AQI binary indicator
  - Binary Logistic Regression and Gradient-Boosted Tree Classification
- Standard Scaling
  - Scaled features for all models except Random Forest
- Feature Indexer
  - For Random Forest Model

Model Development

- 10 Models total
- Split 80% Training Data; 20% Holdout Data
- Pipelines, Param Grid tuning, 5-fold Cross-Validation
- Linear Regression, RF, Binary Logistic Regression, and GBT Classification
  - Baseline model included only criteria gas features (CO, SO2, NO2, O3)
  - Four models were fit with sets of criteria gas, pollutant, meteorological predictors and ***without*** socio-economic factors as predictors
  - Three models were fit with sets of criteria gas, pollutant, meteorological and socio-economic factors as predictors
  - Two models predicted median income with criteria gas, pollutant, meteorological, and population density predictors

Model Performance

Linear Regression and Random Forest Models

	Model 0	***Model 1***	***Model 2***	***Model 4***	***Model 5***	***Model 6***	***Model 7***	***Model 8***
Type	Linear Regression	Linear Regression	Random Forest Regression	Linear Regression	Linear Regression	Random Forest Regression	Linear Regression	Linear Regression
Features/ Predictors	criteria gas	criteria gas, pollutants, particulates, met. data	criteria gas, pollutants, particulates, met. data	criteria gas, med. income, pop. density	criteria gas, pollutants, particulates, met. data, med. income, pop. density	criteria gas, pollutants, particulates, met. data, med. income, pop. density	criteria gas, AQI, pop. density	criteria gas, pollutants, particulates, met. data, AQI, pop. density
Response Variable	AQI	AQI	AQI	AQI	AQI	AQI	med. income	med. income
Best Tuning Params	reg = 0.01 maxIter = 1 elasticNet = 0	reg = 0.01 maxIter = 1 elasticNet = 0	#Trees = 50 maxDepth= 10	reg = 0.01 maxIter = 1 elasticNet = 0	reg = 0.01 maxIter = 1 elasticNet = 0	#Trees = 20 maxDepth= 10	reg = 0.1 maxIter = 10 elasticNet = 0.5	reg = 0.1 maxIter = 1 elasticNet = 0
RMSE	26.4	21.8	3.26	26.26	21.8	3.03	1787.26	1653.43
adjusted-R^2	0.511	0.666	0.993	0.515	0.666	0.994	0.122	0.25

Model Performance

Binary Response Models

	Model 3	***Model 9***
Type	Binary Logistic Regression	Gradient-Boosted Tree Classifier
Features/ Predictors	criteria gas, pollutants, particulates, met. data	criteria gas, pollutants, particulates, met. data
Response Variable	General AQI *("Healthy"/"Unhealthy")*	General AQI *("Healthy"/"Unhealthy")*
Best Tuning Params	reg = 0.01 maxIter = 10 elasticNet = 0	maxDepth = 10 maxBins = 40 maxIter = 10
Accuracy	84.7%	99.9%
Precision	78.8%	99.9%
Recall	63.6%	99.8%
AUROC	78.4%	99.9%
AU PR Curve	69.7%	99.9%

Conclusions

- GBT Classifier and RF Regression
  - Very high performance in predicting AQI
  - Models ***did not*** predict median income accurately
- Feature importance from RF model
  - ★ ***particulate matter, ozone concentration, and temperature***
  - Temperature is not considered in AQI calculations (AQI-technical manual, 2016)
  - Further point of research: temperature
- Future use of model
  - Useful in cases where not all pollutant or gas data is available for certain sites
- Future work
  - Test other methods: xgBoost, ensemble methods, other socio-economic factors