NYC Open Data
Case Study

Javier Zapata, PhD(c)

Data Preprocessing

- **NY Data:** weather, SRs requests
- **Complementary data:** Neighborhood Tabulation of NYC
  - Used to cluster observations based on (Lat,Long) into neighborhoods

Python Implementation

- **Visualization:** seaborn, geoplot
- **Machine Learning Library:** h2o
- **Statistics Library:** statsmodels
- **API access library:** sodapy

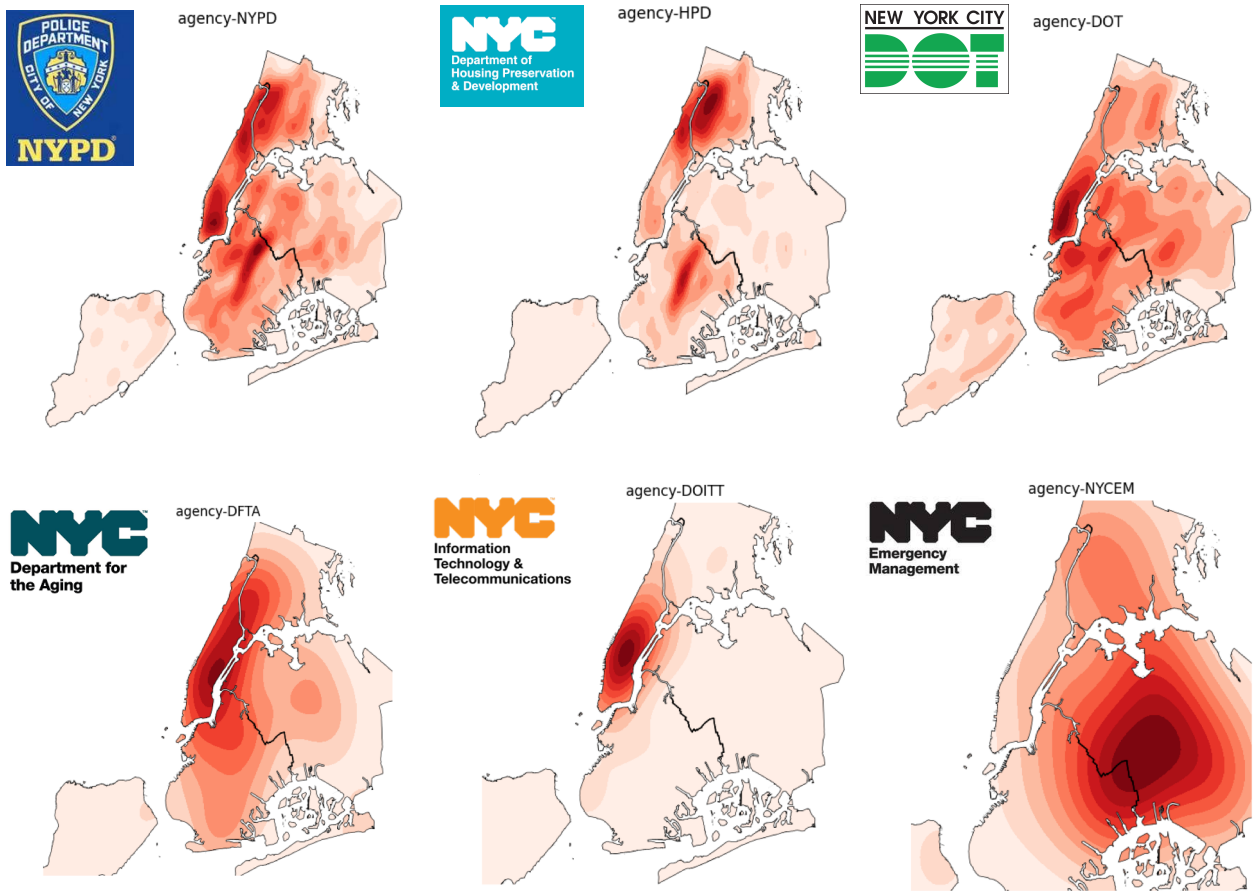
by Agency

Top-5

agency	SRs
NYPD	342965
HPD	235049
DOT	138516
DEP	113928
DSNY	108442

Bottom-5

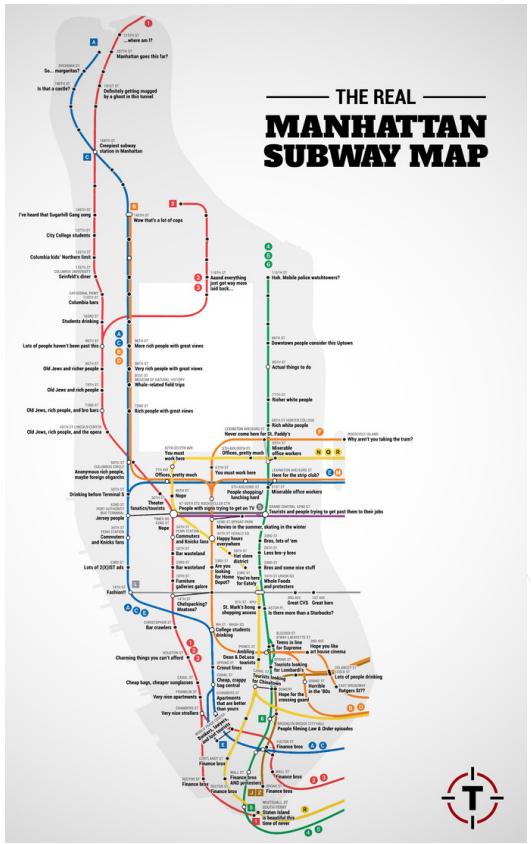
agency	SRs
DOF	571
3-1-1	319
DFTA	309
DOITT	309
NYCEM	30

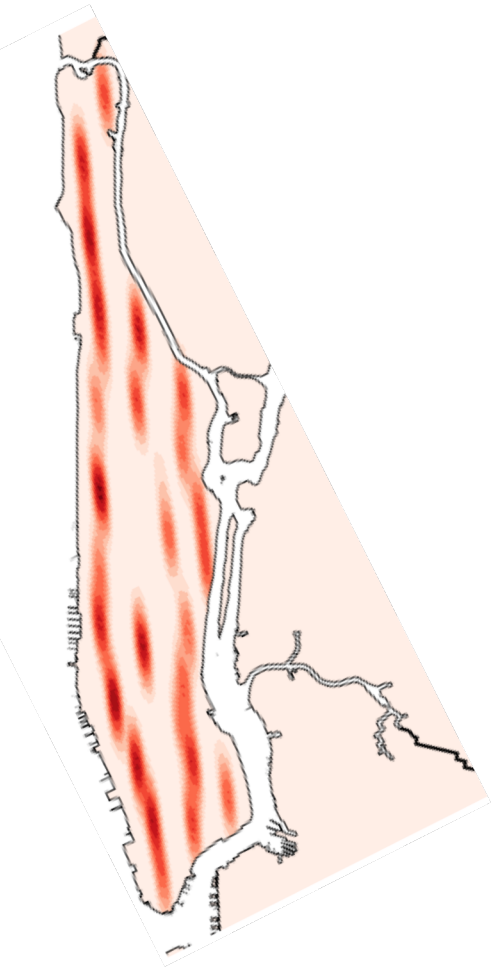


Created by Javier Zapata, PhD(c), UC Santa Barbara,
jzapata@ucsb.edu

by Borough

- Density Plot with SRs coordinates aligns with the main subway lines of Manhattan.
- For other boroughs the results are distributed in all the area.



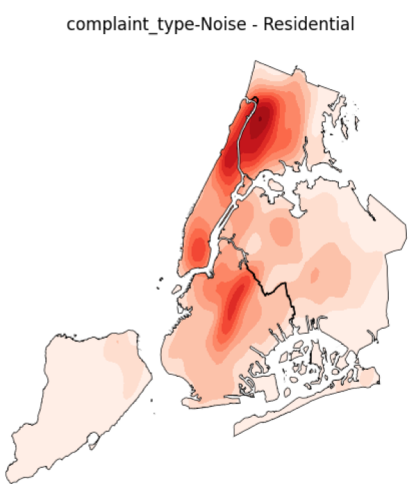


Created by Javier Zapata, PhD(c), UC Santa Barbara,
jzapata@ucsb.edu

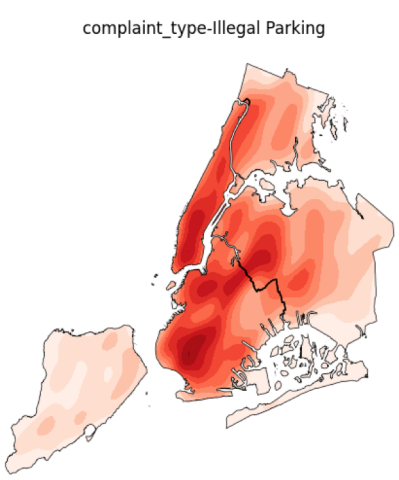
By Complaint Type

Top-5

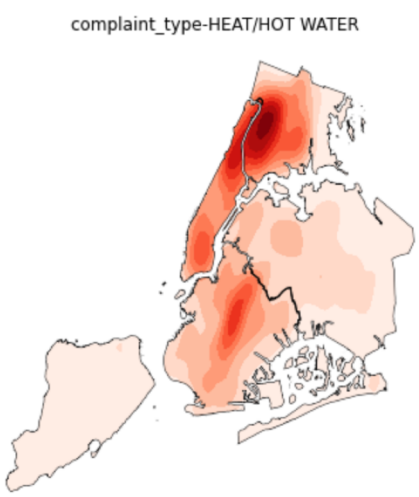
complaint_type	SRs
Noise - Residential	85523
Illegal Parking	69194
HEAT/HOT WATER	59979
Blocked Driveway	57682
Street Condition	47130



Bronx



Brooklyn, Manhattan and Queens

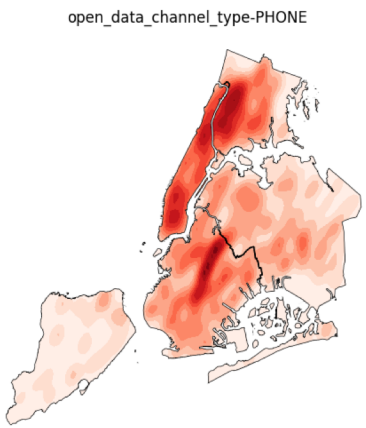


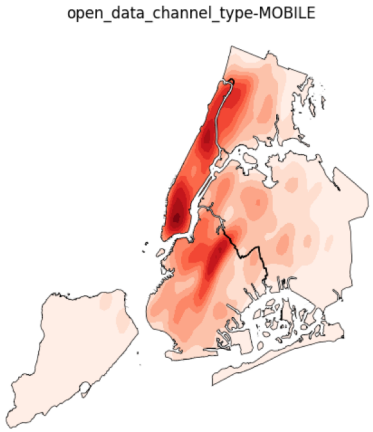
Bronx

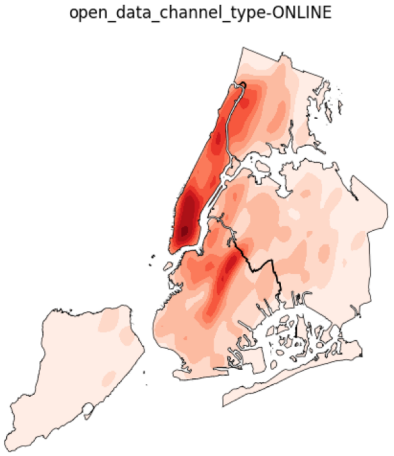
Created by Javier Zapata, PhD(c), UC Santa Barbara,
jzapata@ucsb.edu

Spatially Similar

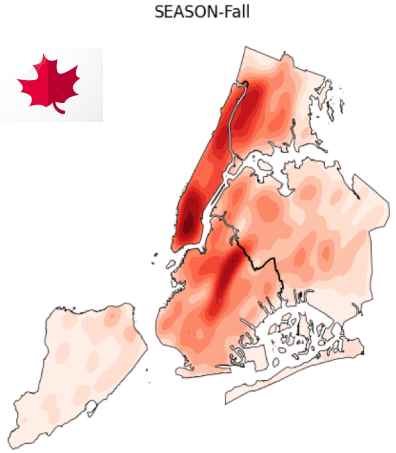
by Data Channel

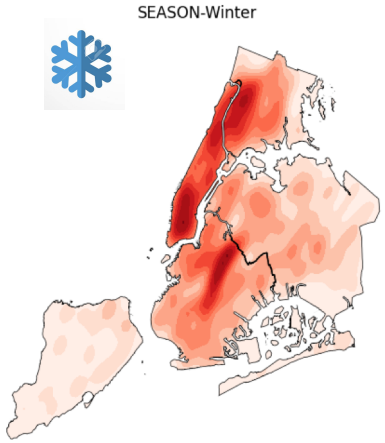


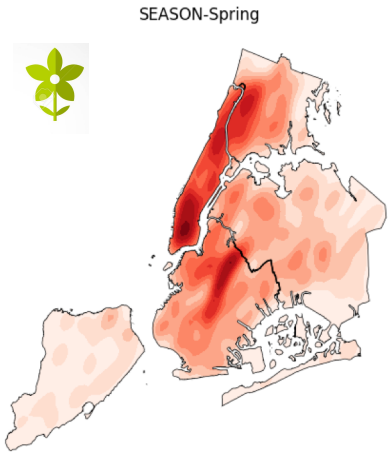


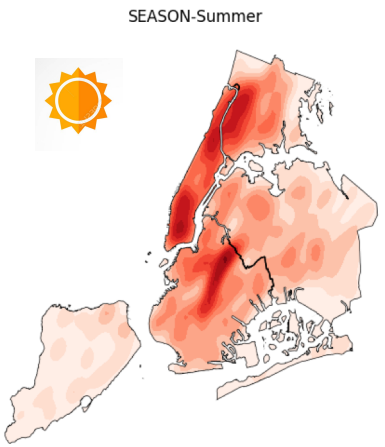


by Seasons







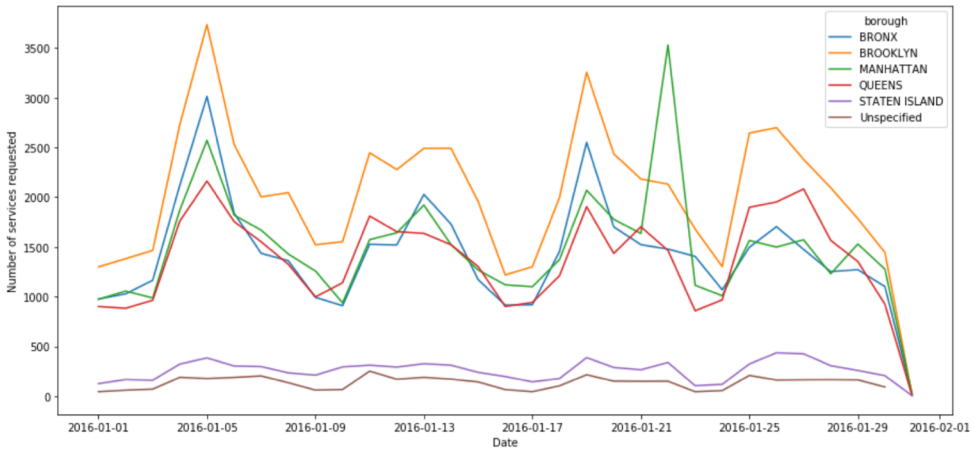


Created by Javier Zapata, PhD(c), UC Santa Barbara,
jzapata@ucsb.edu

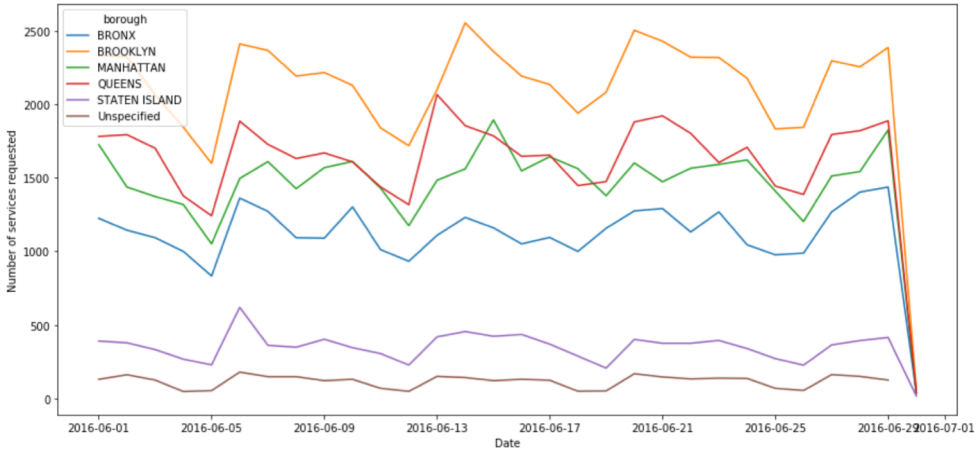
Weekly Seasonality

- Counts by Borough: Day of Week effect

Jan-2016



Jun-2016



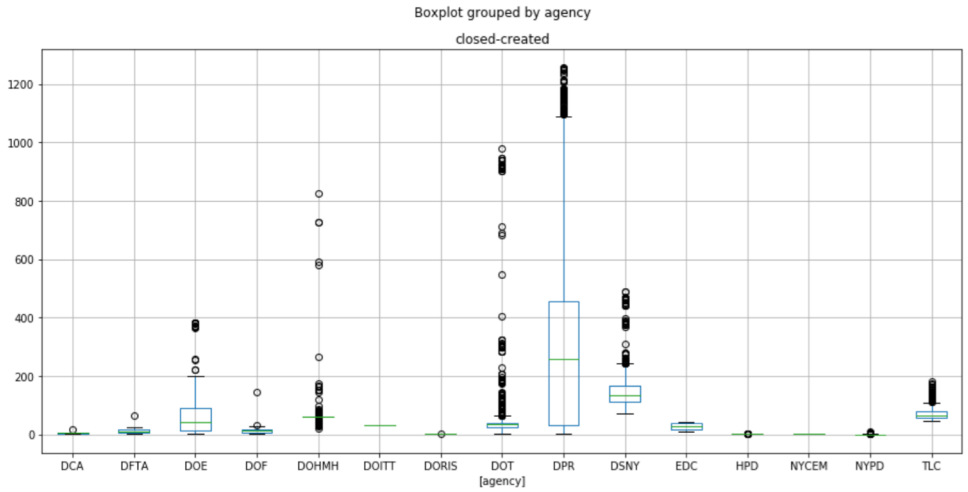
Created by Javier Zapata, PhD(c), UC Santa Barbara,
jzapata@ucsb.edu

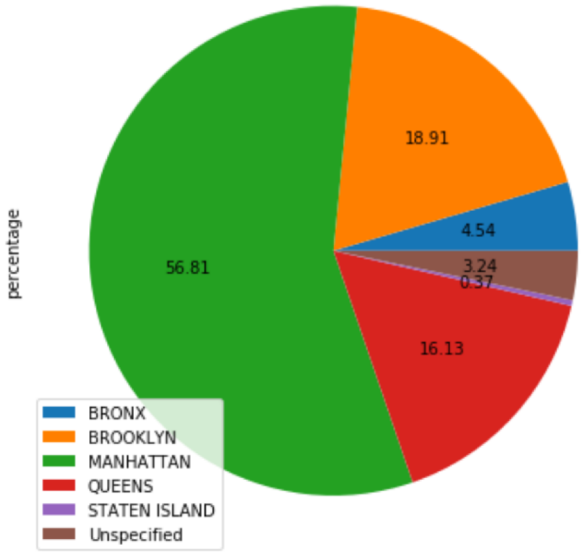
Past Due SRs

- SRs with 'closed-date' > 'due-date'
- Agencies
  - Department of Transportation (DOT)
  - Department of Parks and Recreation (DPR)
  - Taxi & Limousine Commision (TLC)
  - Department of Health and Mental Hygiene (DOHMH)

They have a dedicated column in the dataset

- They concentrate in Manhattan





Created by Javier Zapata, PhD(c), UC Santa Barbara,
jzapata@ucsb.edu

Generalize Additive Model

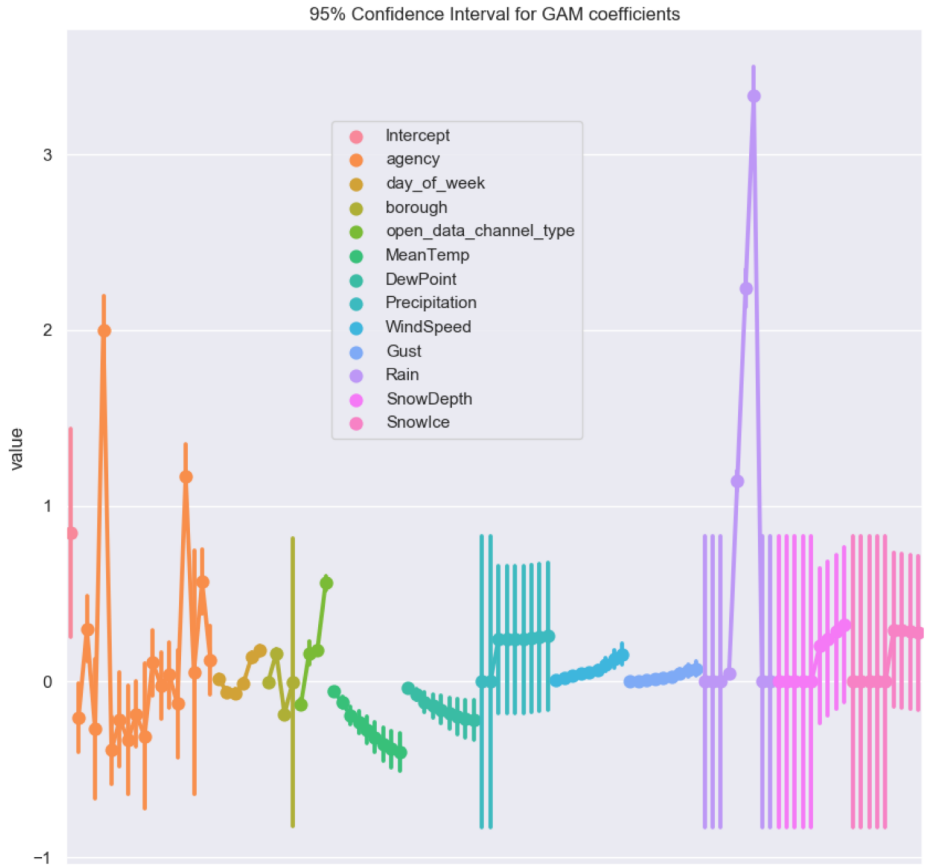
We regress SRs counts with a Generalize Additive Model. Predictors are categorical variables as fixed effects and weather variables with splines

Discarded weather variables:

- SnowDepth
- Snowice
- Precipitation

Weather variables with explanatory power:

- Temperature
- DewPoint
- WindSpeed
- Gust
- Rain???



All categorical variables seem to have explanatory power

Created by Javier Zapata, PhD(c), UC Santa Barbara,
jzapata@ucsb.edu

Predictive Model Construction

- Model SRs counts as Poisson Distribution

$$Y_{SRs} | \ (NYC\ data + Weather\ data) \sim Poisson(\lambda)$$

- We build models to predict the mean count

- **Model 1** : Ensemble: Gradient Boosting Machine + Random Forest

$$g(E[Y_{SRs}]) = \omega_0 + \omega_{\mathrm{GBM}} f_{\mathrm{GBM}}(X) + \omega_{\mathrm{RF}} f_{\mathrm{RF}}(X)$$

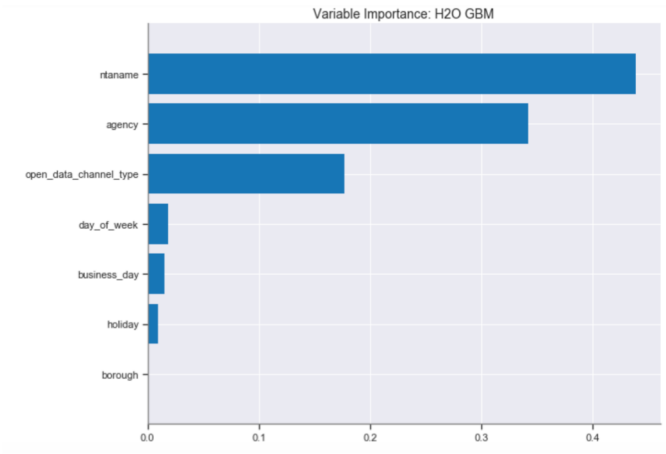
- **Model 2**: GBM only
- **Model 3**: 1 Layer ReLU Deep Net

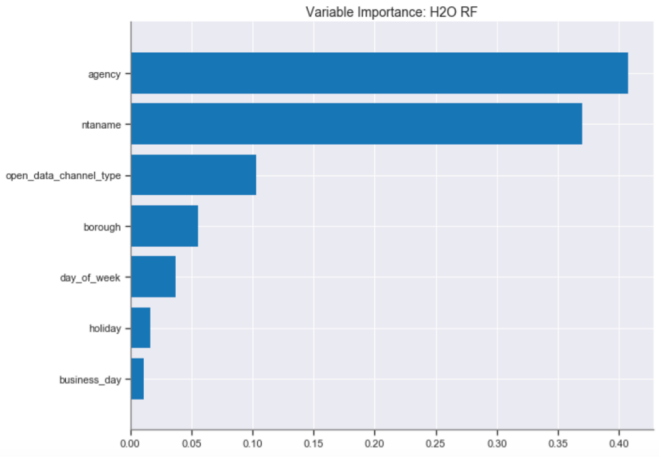
Created by Javier Zapata, PhD(c), UC Santa Barbara,
jzapata@ucsb.edu

Model 1:
Ensemble : GBM + RF

Only Categorical Predictors

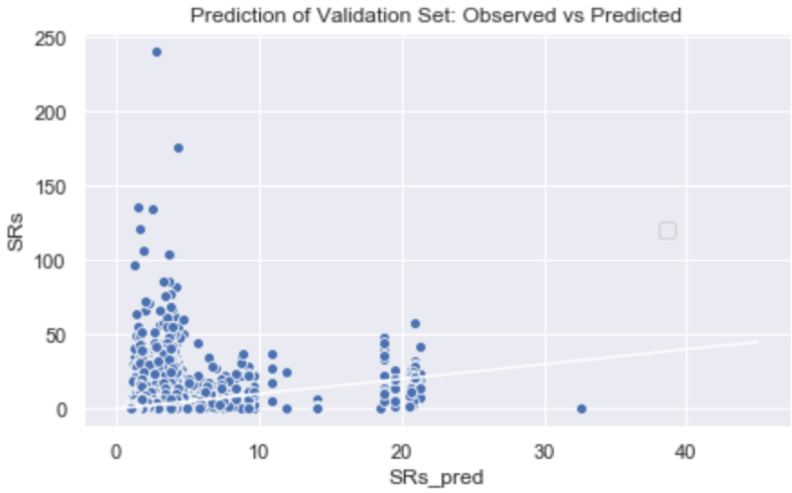
Most explanatory variables:
- Neighborhood/Borough
- Agency
- Open data channel





Validation:
- Mean Squared Error = 4.686
- Median Absolute Error = 0.579
- Mean Squared Log Error = 0.146

Outliers!

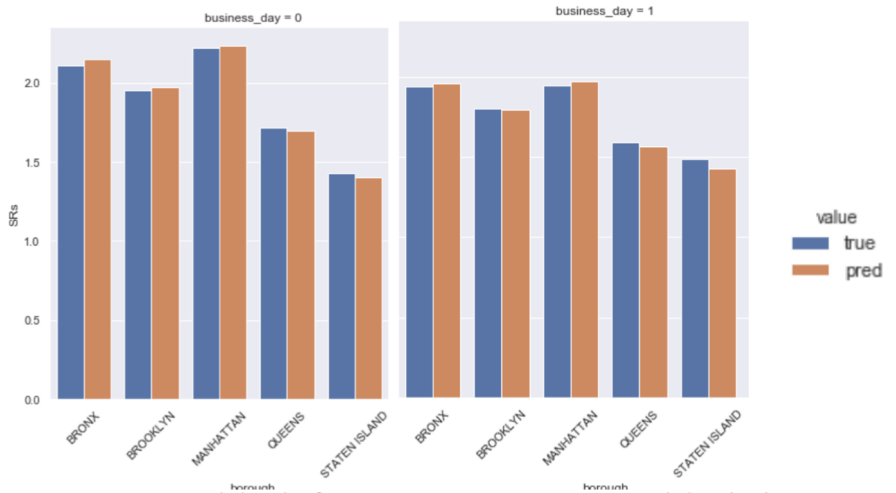


Created by Javier Zapata, PhD(c), UC Santa Barbara,
jzapata@ucsb.edu

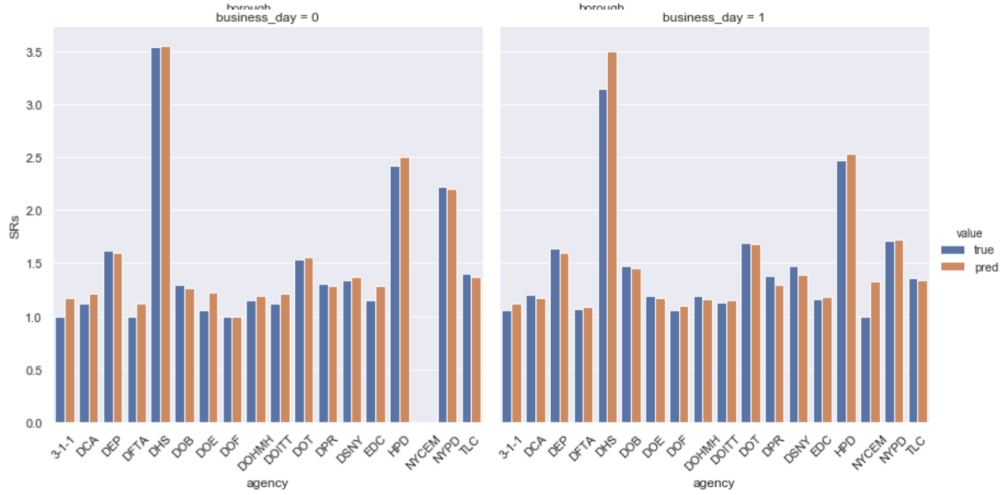
Model 1:
Ensemble : GBM + RF

- Comparison of aggregated mean estimates

By Borough



By Agency



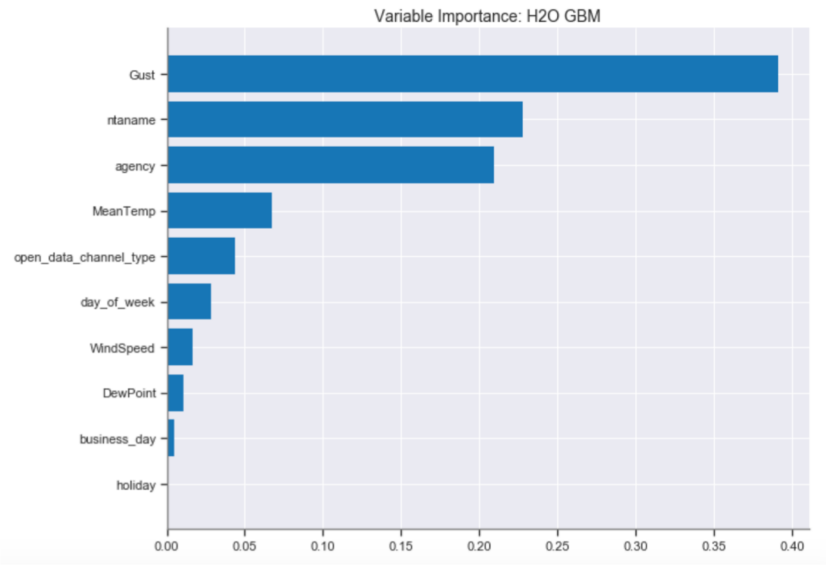
Created by Javier Zapata, PhD(c), UC Santa Barbara,
jzapata@ucsb.edu

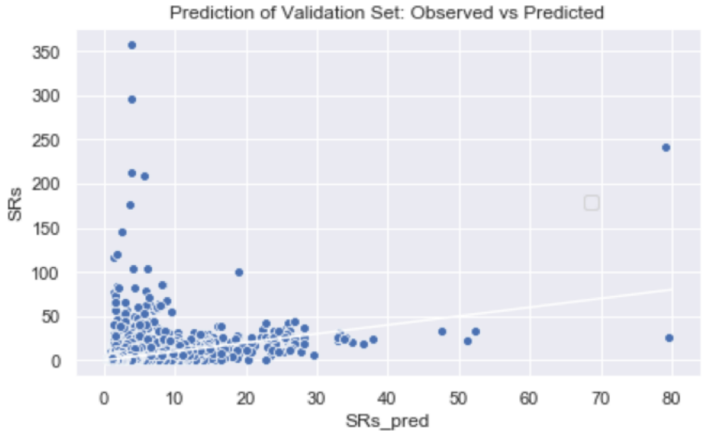
Model 2:
GBM

Categorical + Weather Predictors

Most explanatory variables:
- Gust
- Mean Temp
- Neighborhood/Borough
- Agency
- Open data channel

Validation:
- Mean Squared Error = 5.509
- Median Absolute Error = 0.567
- Mean Squared Log Error = 0.138



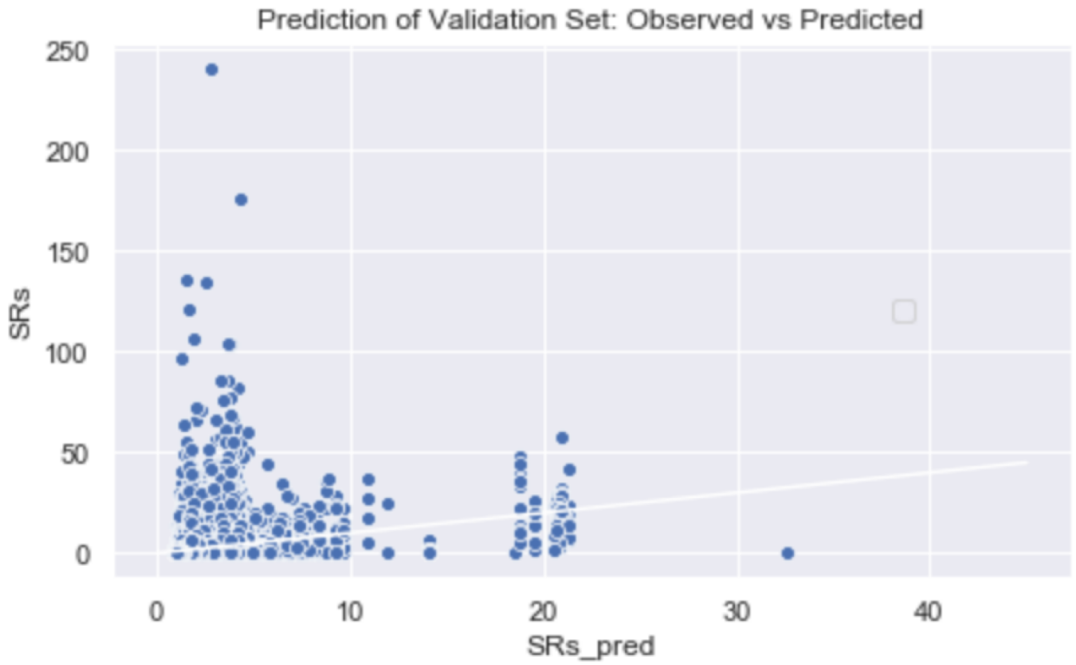


Created by Javier Zapata, PhD(c), UC Santa Barbara,
jzapata@ucsb.edu

Model 3:
1 Layer ReLU Deep Net

Only Categorical Predictors

Worse performance than previous model



Validation:

- Mean Squared Error = 4.909
- Media Absolute Error = 0.626
- Mean Squared Log Error = 0.155

Created by Javier Zapata, PhD(c), UC Santa Barbara,
jzapata@ucsb.edu

Improving the Results

- More time and computational power
  - Tree based model somputational power is always better
- Use all the data available, not just 2016.
- Data Exploration:
  - Implement exponential family PCA to:
    - observe directions of maximum variation with categorical data
    - fill missing data
- Feature creation:
  - Clustering observations by coordinates, time of day and the other predictors
    - K-Means, Spectral Clustering
  - Use Anomaly Detection on predictors to create weight for training.

Created by Javier Zapata, PhD(c), UC Santa Barbara,
jzapata@ucsb.edu

Improving the Results

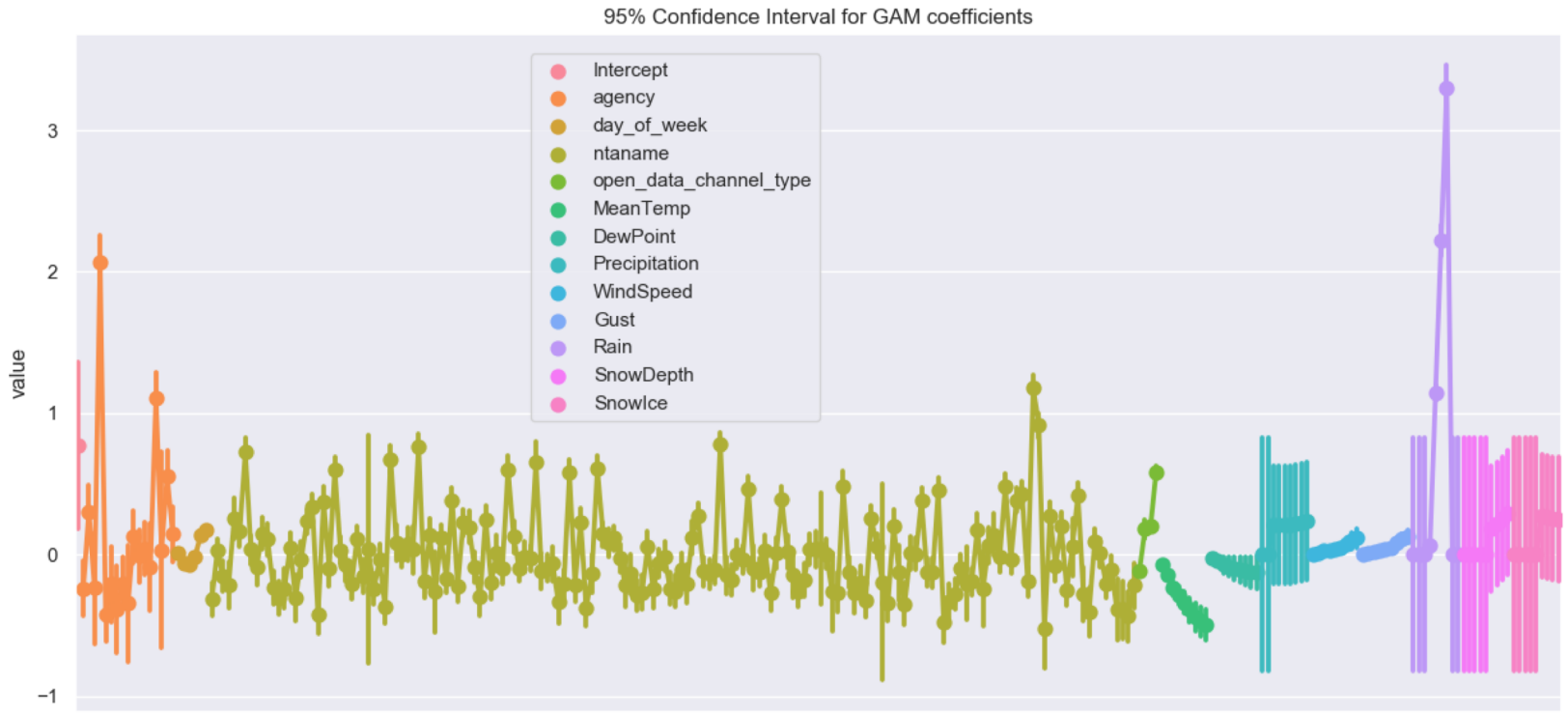
- Feature Selection in GAM:
  - Bootstrap Confidence Intervals for GAM coefficients
  - Grid search for the splines penalty
- More Training:
  - Analyze class imbalance in the predictors
  - Enhance robustness to hyperparameters with Kernel Mean Matching formulations of the algorithms.
- Outliers control with weigths:
  - Define a weight on observations as a new hyper parameter
  - Fit the model using Bayesian Optimization or Random Grid Search

Improving the Results

- Predictive Model:
  - Include XGBoost which is a Gradient Boosting Machine with more hyperparameters to tune.
  - Incorporate Bayesian and other latent variable models (very time consuming for training)
  - Grid search to tune the hyperparameters of the model

Appendix: Generalize Additive Model

Using NTAs instead of Boroughs, same conclusion.



Created by Javier Zapata, PhD(c), UC Santa Barbara,
jzapata@ucsb.edu