OPEN DATA NATION

FIVAR:

Food Inspection Violation, Anticipating Risk

Nicole Donnelly and Jonathan Boyle
General Assembly DSI-DC-1 Capstone
8 July 2016

Outline

Issue

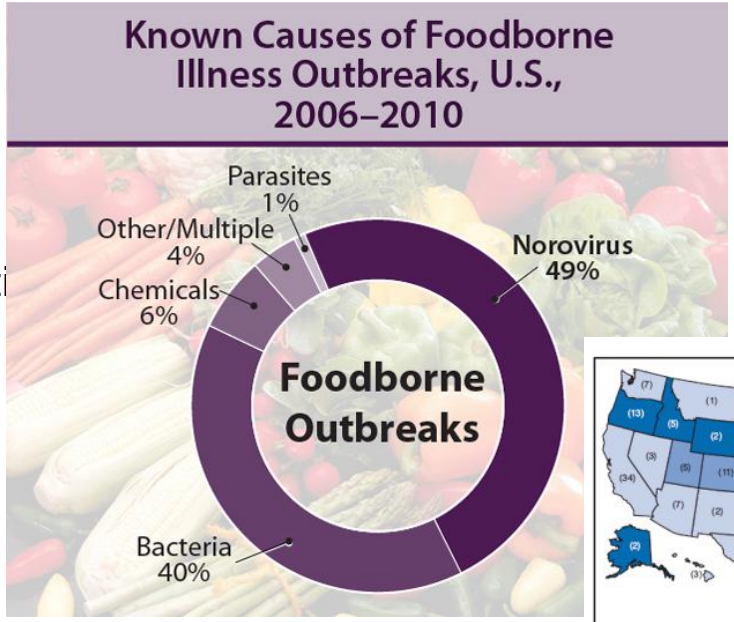
- Food borne illness
  - Better predicti

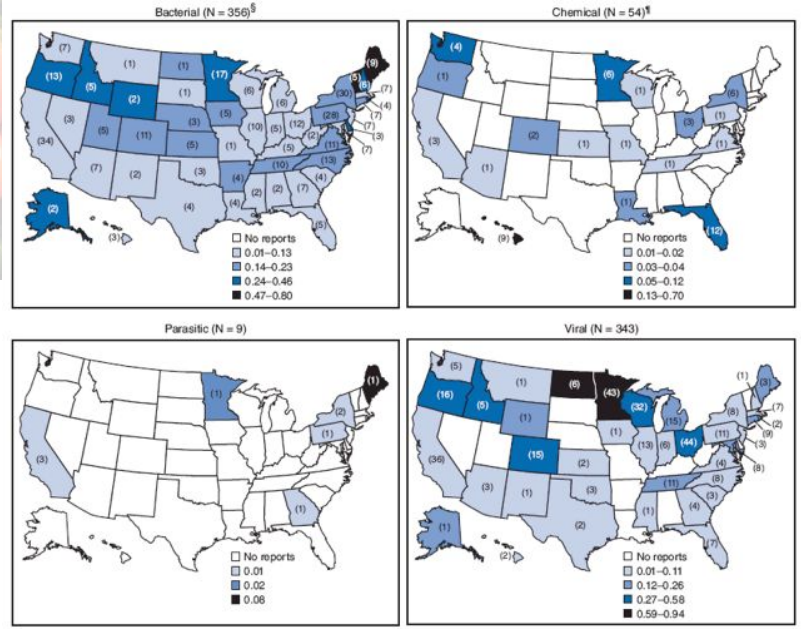
Approach

- Model
  - Selection
  - Optimization
- Results

Plan of Action

- Needed Data
- Next steps
  - Modeling, analysis, and reports
- Product development





http://www.cdc.gov/features/dsnorovirus/figure3.html
http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5822a1.htm

Issue

Food borne illness from restaurants

- Not enough health inspectors
- Annual inspections
- Present approach not optimal for public health protection

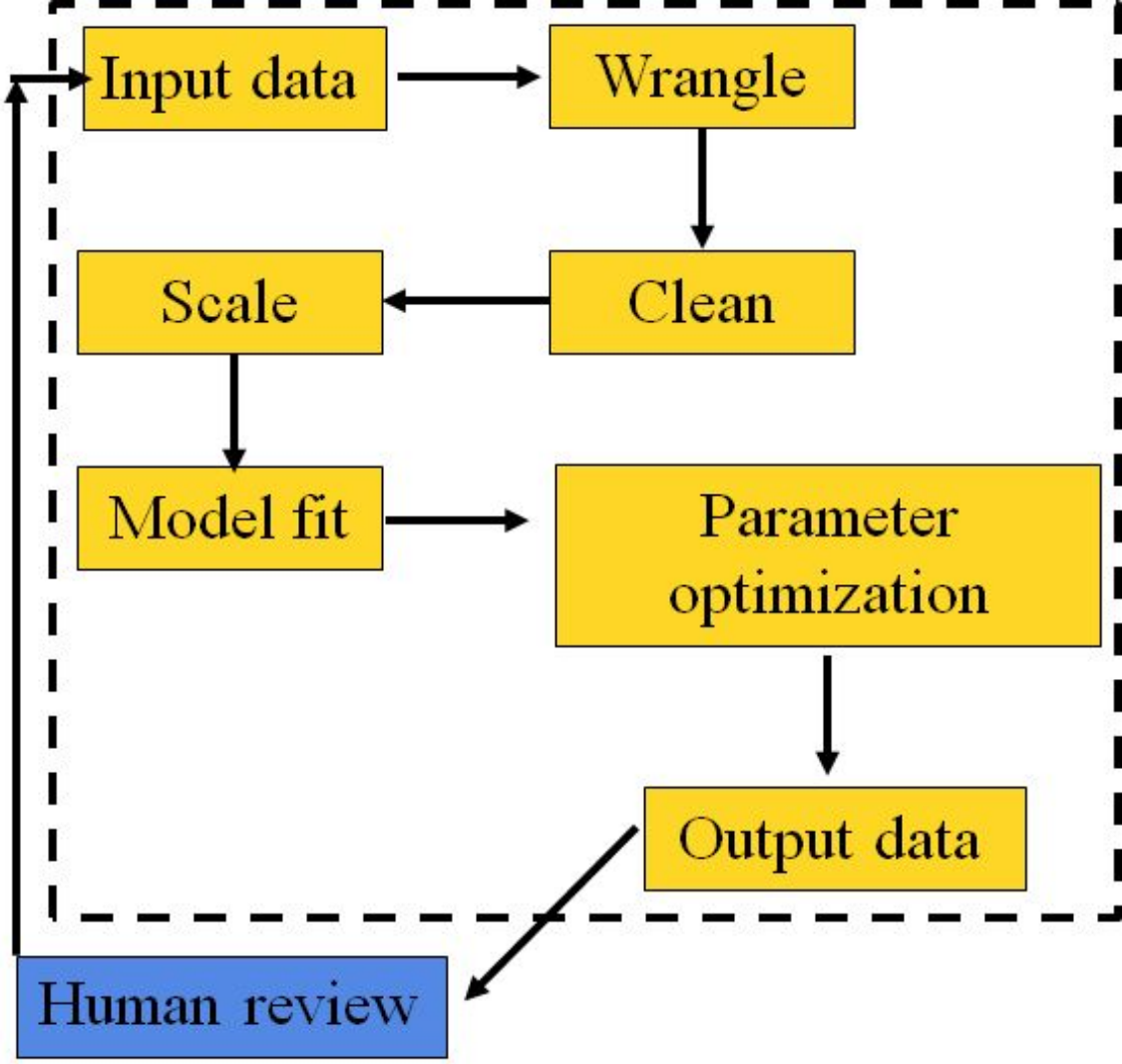
Different approach

- Machine Learning
  - Predict when violations will occur
    - Reduce illness
- Open Data Nation
  - FIVAR model

Model

Data

- Multiple sources compiled
  - DC restaurant health inspection reports
  - 311 complaints
  - Crime records
  - Construction permits
  - Weather data
  - Yelp data
- Additional data reviewed
  - Liquor permits
  - Business licenses



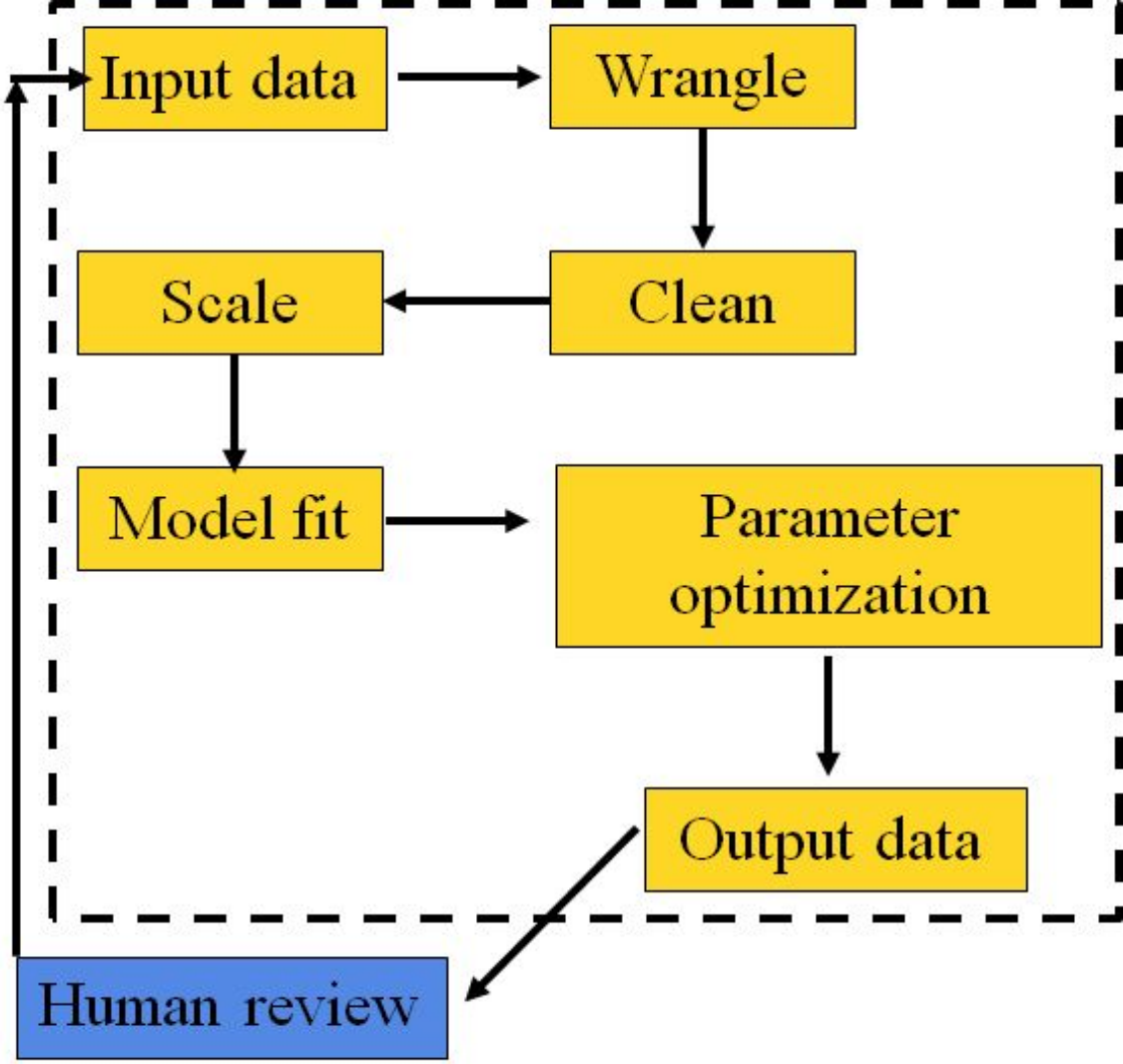
Model

Selection

- Binary classifier
  - RandomForest → Log Reg,KNN
  - Model(s) chosen based on prior work
    - Chicago
    - Montgomery county, MD

Optimization

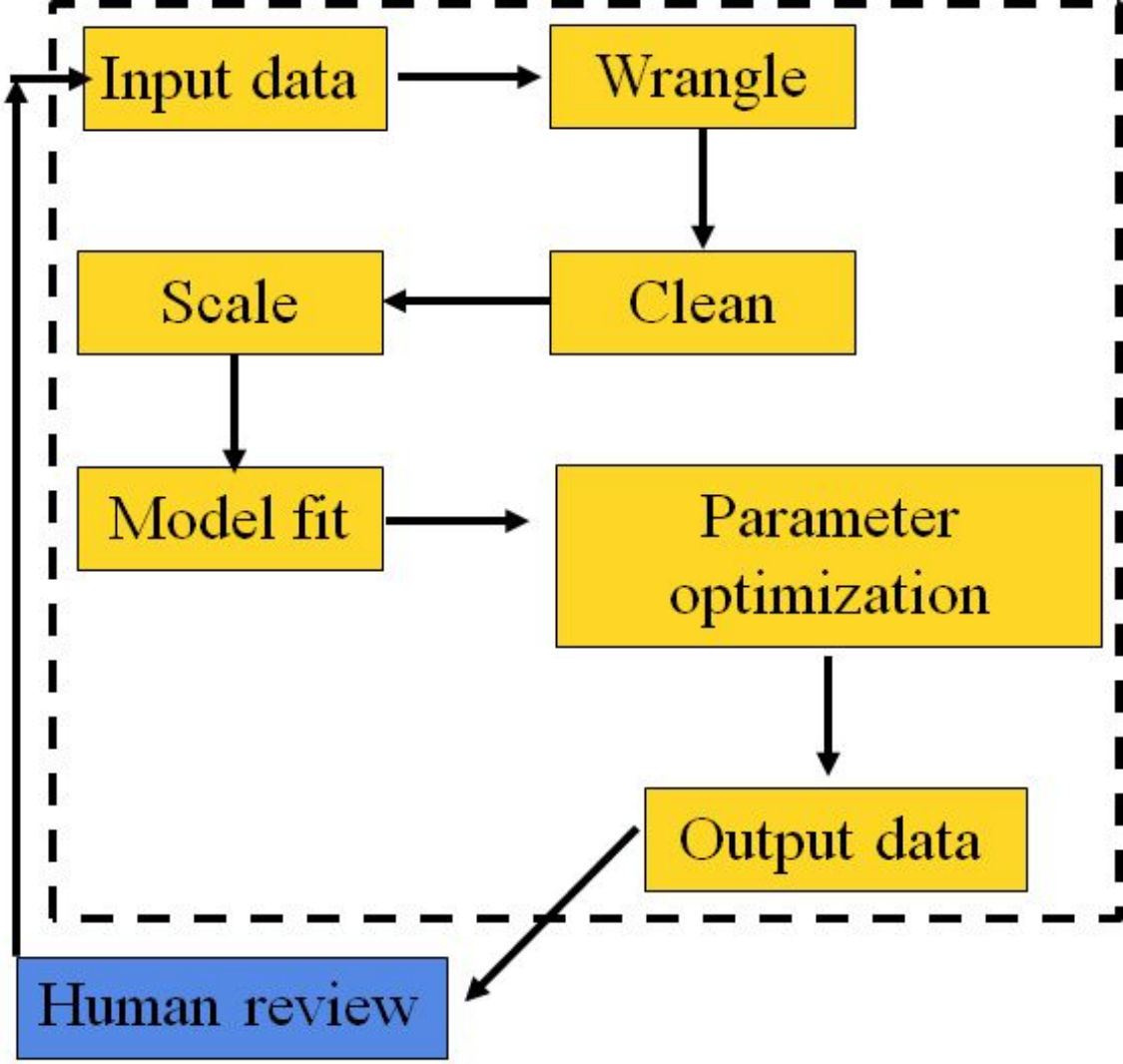
- Mix
  - Prototype code
  - Automated functions and scripts
- Feature selection
- Parameter optimization



Model

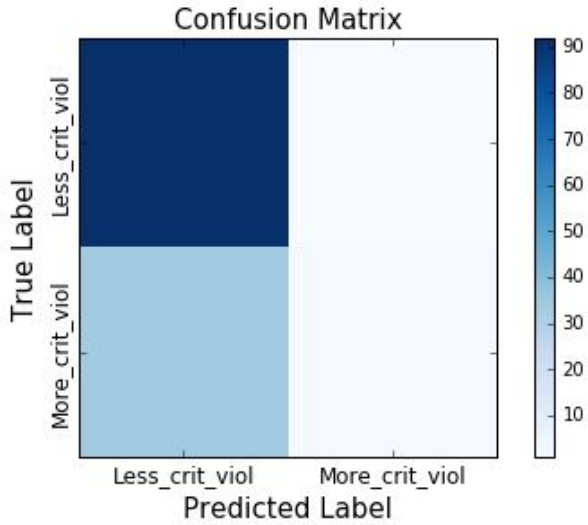
Testing

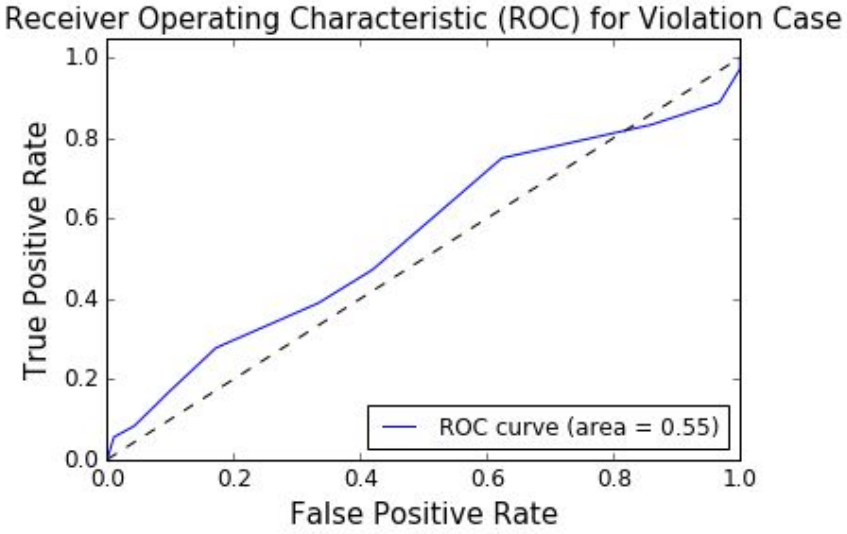
- Test-train split
  - 2013-2015 data
- Out-of-sample
  - 2016 data



Results

Metric	75/25	
	Log Reg	KNN
Accuracy	0.69	0.73
F1	0.00	0.10
Roc_Auc	0.60	0.55





Testing

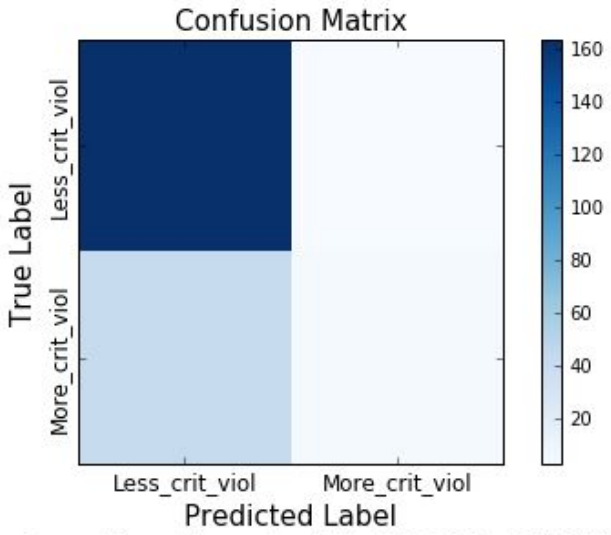
- Test-train split
  - 2013-2015 data
  - Effects from unbalanced data
    - ~6% (77 of 1284) - 0 violations
  - More work/different data needed

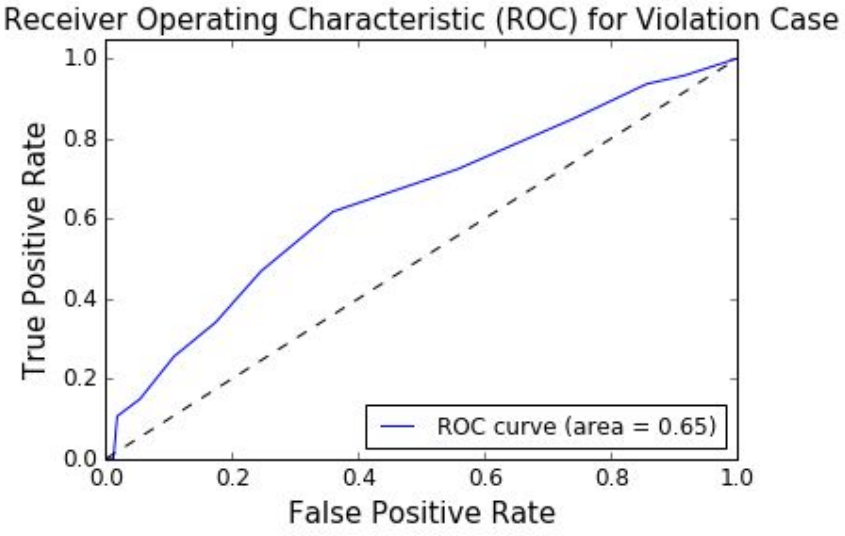
Results

Metric	75/25	
	Log Reg	KNN
Accuracy	0.69	0.73
F1	0.00	0.10
Roc_Auc	0.60	0.55
2016_accuracy	0.78	0.79
2016_F1	0.00	0.18
2016_Roc_Auc	0.56	0.65

Other models in progress:

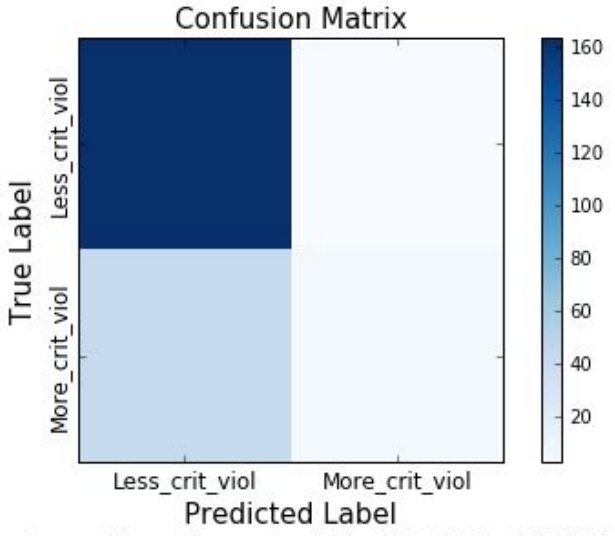
- Naive Bayes
- RandomForest (full process)
- Regression models for split-off analysis

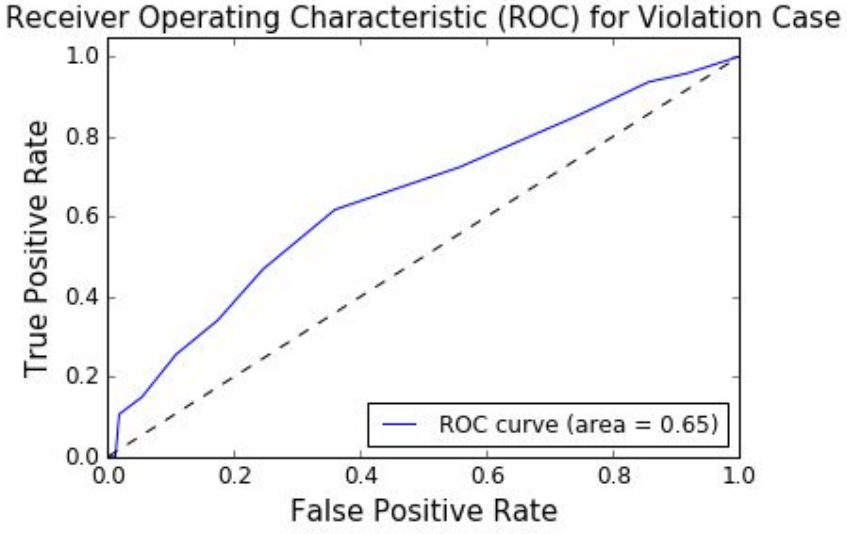




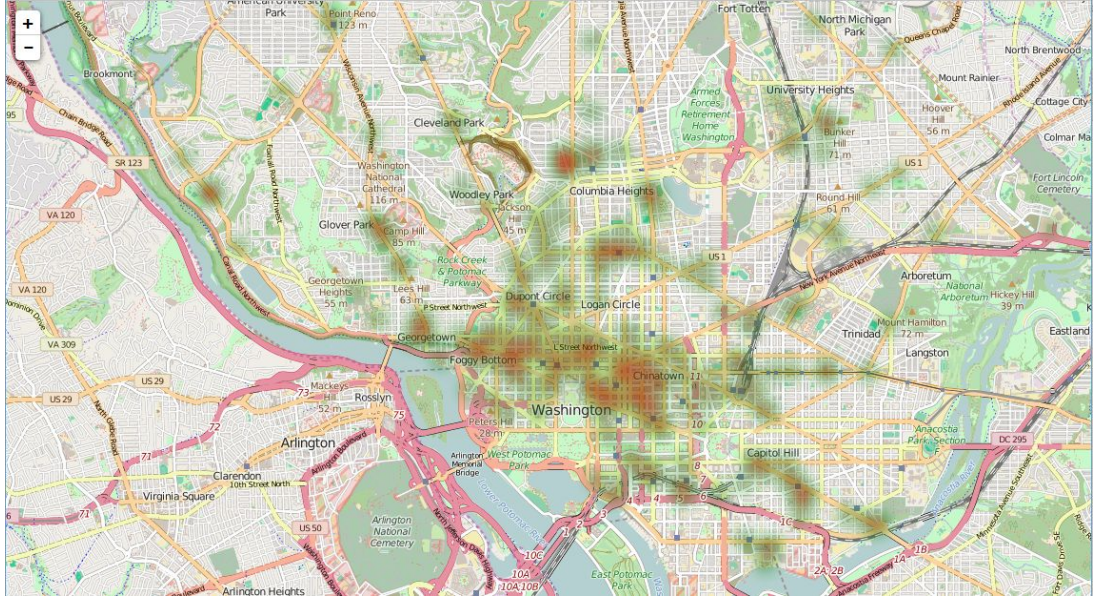
Results

Rank	Important Features
1	Inspector badge #
2	Time since last inspection
3	3-day average high temp
4	# yelp reviews
5	Local crime count
6	Local construction permit count
7	Yelp rating
8	Sandwich shop





Next steps



Immediate Actions

- Perform other model methods
- Expand dataset with additional Yelp matches
- Statistical tests for feature selection
- Develop metrics for comparison to other cities/prior work

Immediate Actions

- Analyze data concerning specific inspectors
- Coordinate further with Open Data Nation about specific metrics and direction of efforts
    - Discussions with DC government about “openness/accessibility” of data

Supplementary Slides

Data Sources

DOH Inspections
Crime
ABRA
DCRA
Construction

Rating
Number of Reviews
Category

Places

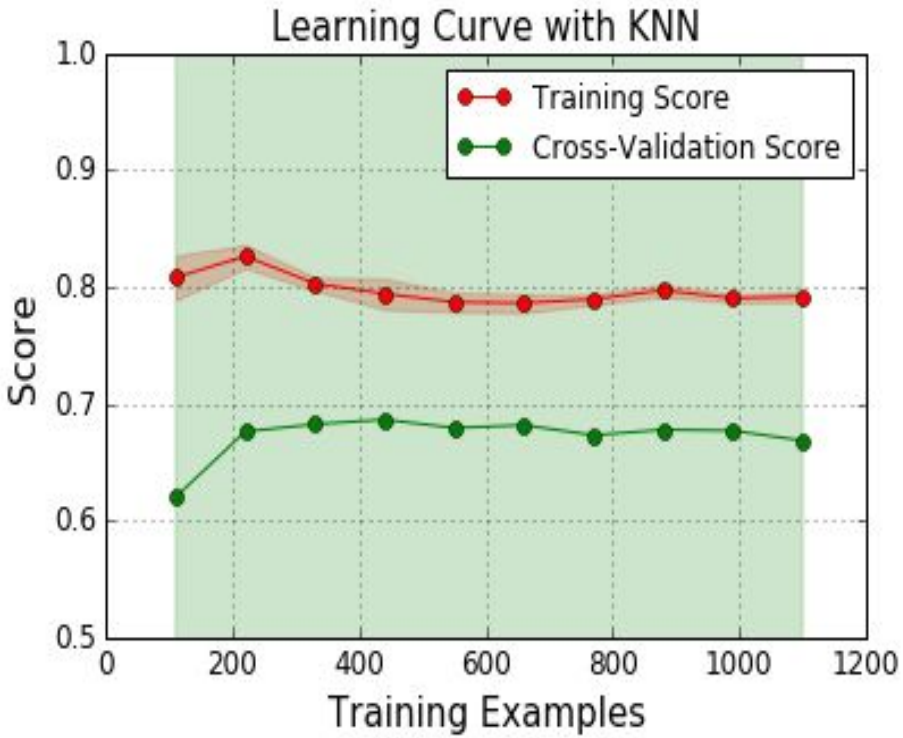
Weather

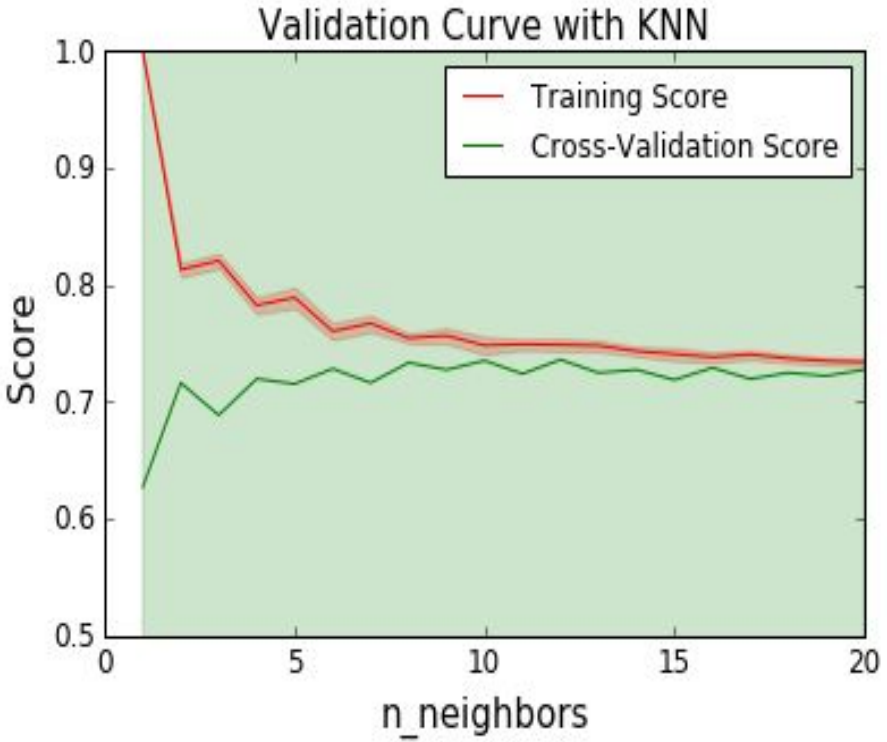
Non-emergency City Issues

Model

Metric	"As is"		50/50		75/25	
	Log Reg	KNN	Log Reg	KNN	Log Reg	KNN
Accuracy	0.93	0.95	0.56	0.56	0.69	0.73
F1	0.96	0.97	0.61	0.58	0.00	0.10
Roc_Auc	0.52	0.52	0.60	0.57	0.60	0.55
2016_accuracy	0.93	0.93	0.57	0.59	0.78	0.79
2016_F1	0.96	0.96	0.46	0.45	0.00	0.18
2016_Roc_Auc	0.56	0.59	0.59	0.57	0.56	0.65

Model


Learning Curve with KNN
Training Score
Cross-Validation Score
Score
Training Examples
1.0
0.9
0.8
0.7
0.6
0.5
0
200
400
600
800
1000
1200


Validation Curve with KNN
Training Score
Cross-Validation Score
Score
n_neighbors
1.0
0.9
0.8
0.7
0.6
0.5
0
5
10
15
20