Raleigh Restaurant Inspections

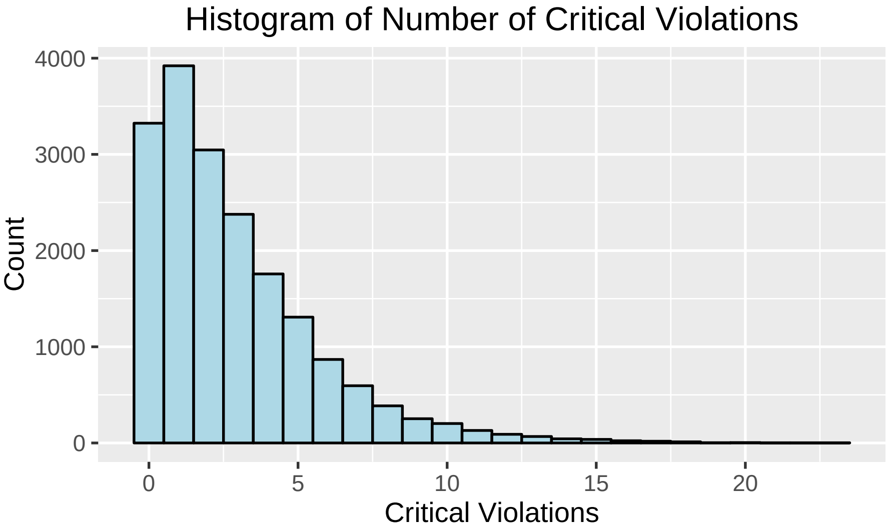
JHUAPL

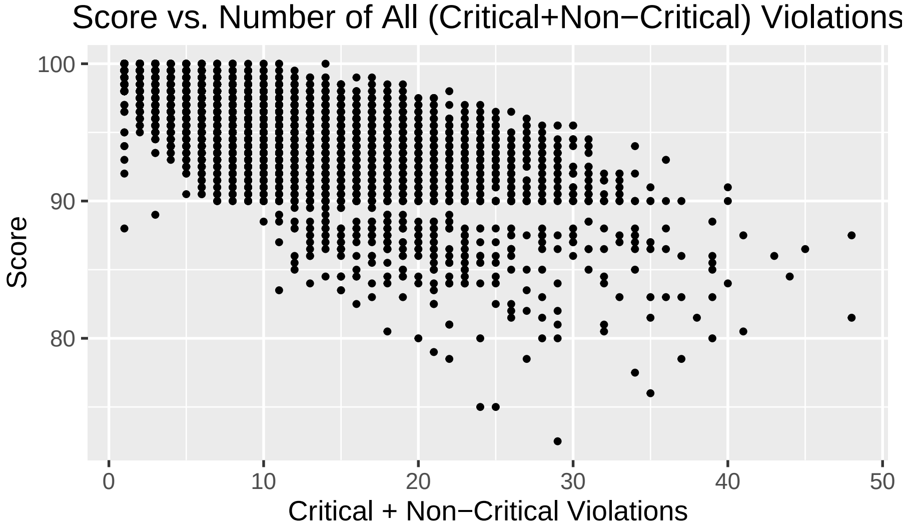
November 17, 2016

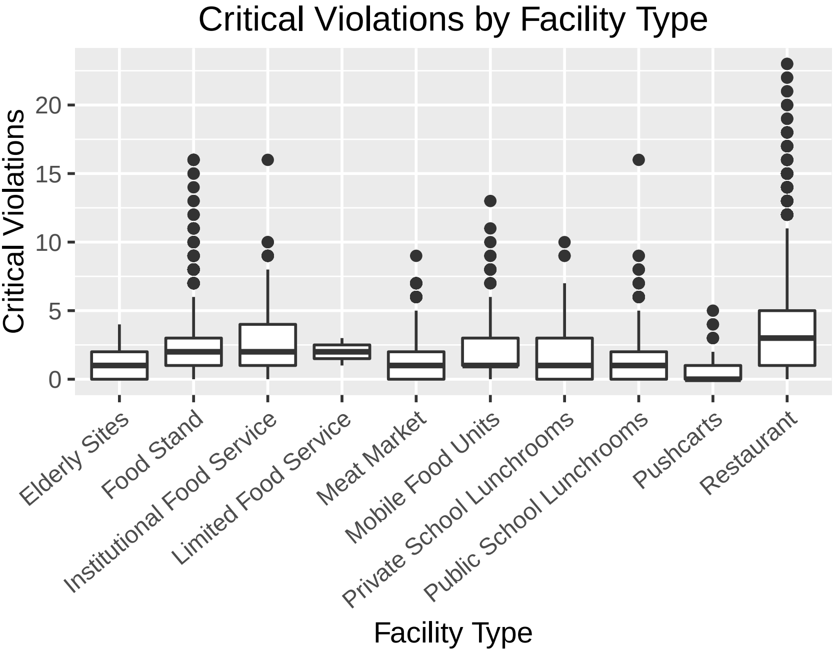
Data Description

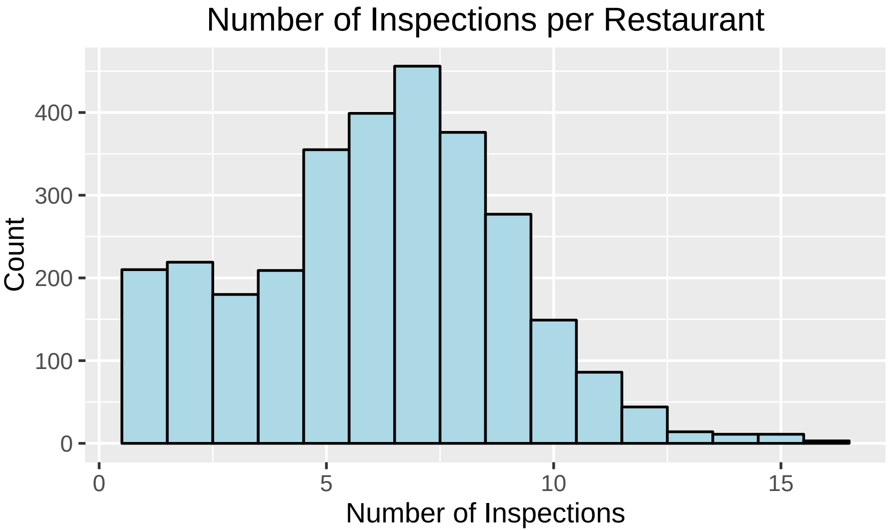
- Data from 9/21/2012 to 11/03/2016.
- 2,809 facilities (1,867 are restaurants)
- 18.469 inspections
- Cities in Wake County
  - top 5: Raleigh, Cary, Wake Forest, Apex, Morrisville

Number of Critical Violations




Score vs. Number of All (Critical+Non−Critical) Violations
100
90
80
Score
0
10
20
30
40
50
Critical + Non−Critical Violations


Critical Violations by Facility Type
Critical Violations
20
15
10
5
0
Elderly Sites
Food Stand
Institutional Food Service
Limited Food Service
Meat Market
Mobile Food Units
Private School Lunchrooms
Public School Lunchrooms
Pushcarts
Restaurant
Facility Type

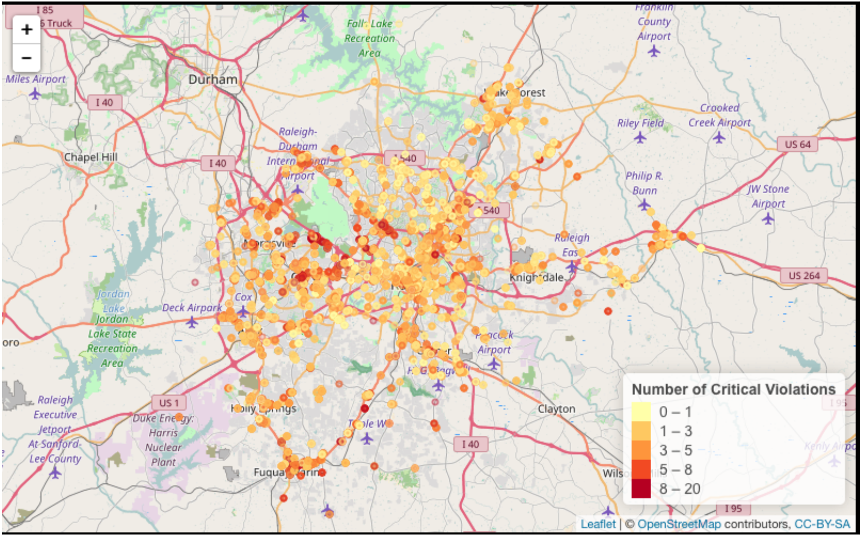

Number of Inspections per Restaurant
Count
400
300
200
100
0
0
5
10
15
Number of Inspections

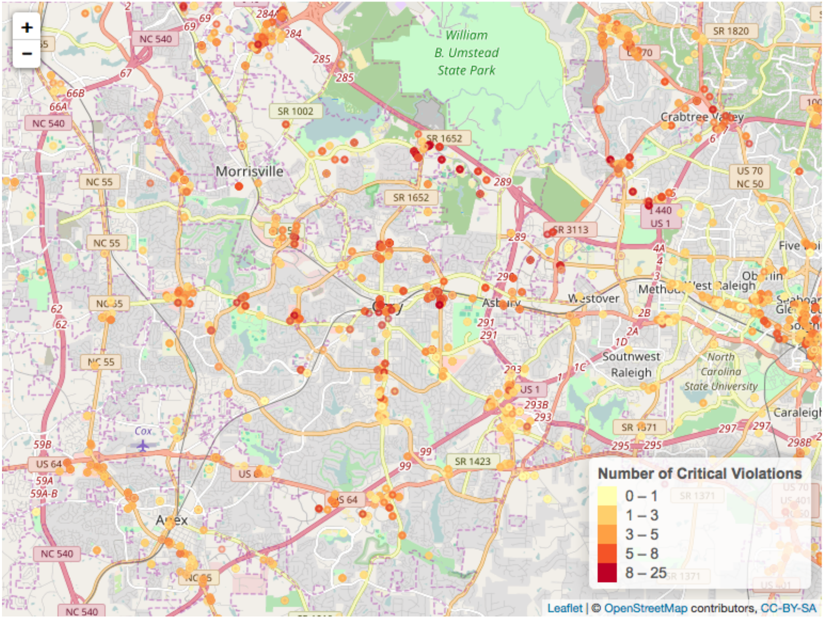
Other Variables in Model / Other Data Sources

Besides number of previous critical violations, days from last inspection, and days since opening (extracted from inspections), other variables include:

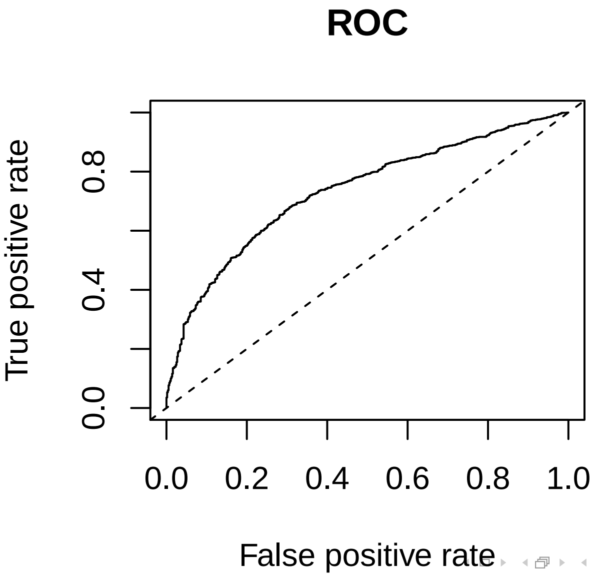
- Geographical Information
  - Average Number of Critical Violations for all Prior Inspections of Nearest 5 Neighbors
- Census data: extracted income information by ZIP
  - Median Household Income
  - Percent below Poverty Line
- Yelp data: extracted information on restaurants
  - Rating (out of 5 stars)
  - Price ($-$$$$)
  - Restaurant Category (top 20, e.g. Mexican, Sushi, Chinese, Coffee)

What is NOT in model: Inspector information, ZIP code


Number of Critical Violations
0 – 1
1 – 3
3 – 5
5 – 8
8 – 20
Durham
Chapel Hill
Knightdale
Clayton
Raleigh-Durham International Airport
Falls Lake Recreation Area
Jordan Lake State Recreation Area
Leaflet | © OpenStreetMap contributors, CC-BY-SA


William B. Umstead State Park
Morrisville
Crabtree Valley
Westover
Southwest Raleigh
North Carolina State University
Caraleigh
NC 540
NC 55
SR 1002
SR 1652
SR 3113
SR 1820
SR 1423
US 64
US 1
US 70
NC 50
Number of Critical Violations
0 – 1
1 – 3
3 – 5
5 – 8
8 – 25
Leaflet | © OpenStreetMap contributors, CC-BY-SA

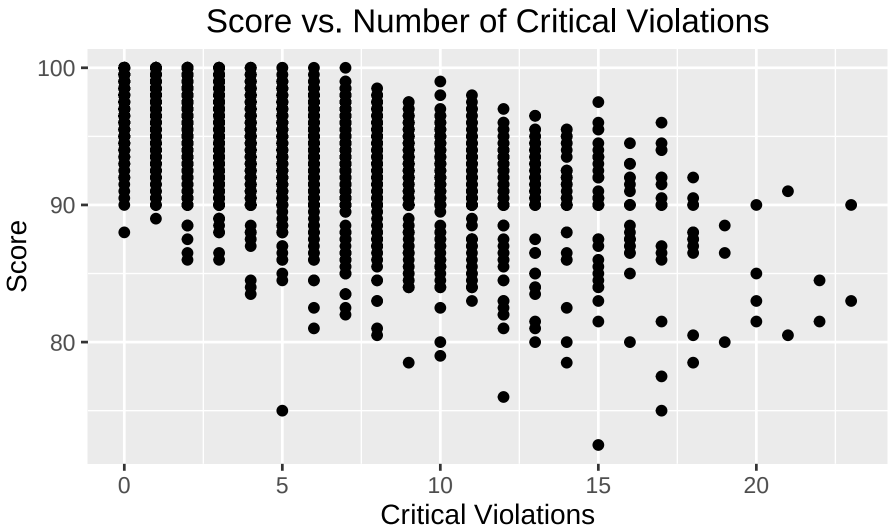
Logistic regression model trained on 2012-2015, tested on 2016 (Jan-Nov), approx. 70/30 split. The AUC on test-set is 0.734.



Next Steps

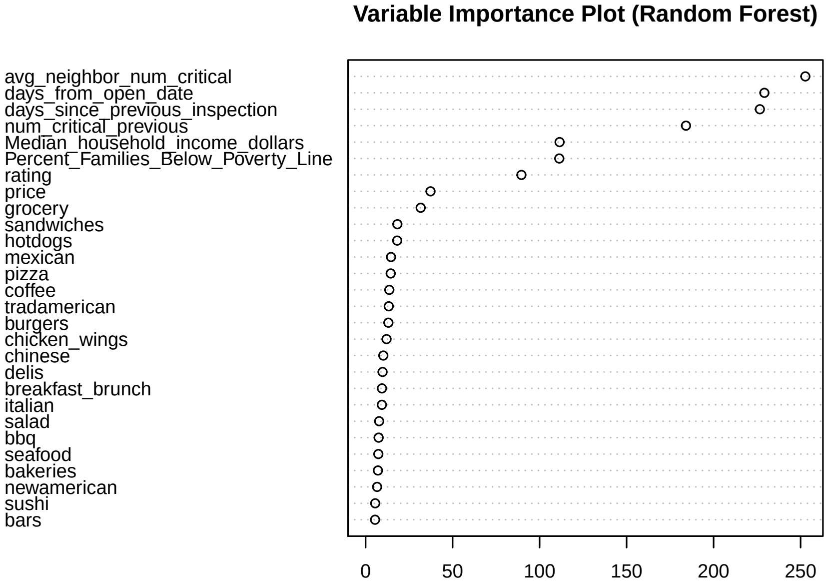
- For binary response, what threshold would be most useful?
  - Results for cutoffs of 2, 3, 4, or 5 are similar (AUC 0.73-0.77) - beyond that gets rare
- Class imbalance
- To model counts, use a different model and performance measure
- How to make use of items that don't match from Yelp, Google Places
- Previous values seem quite useful: can me model those better (not just use one previous but all previous)

Appendix A



Score vs. Number of Critical Violations
100
90
80
Score
0
5
10
15
20
Critical Violations

Appendix B



Variable Importance Plot (Random Forest)
avg_neighbor_num_critical
days_from_open_date
days_since_previous_inspection
num_critical_previous
Median_household_income_dollars
Percent_Families_Below_Poverty_Line
rating
price
grocery
sandwiches
hotdogs
mexican
pizza
coffee
tradamerican
burgers
chicken_wings
chinese
delis
breakfast_brunch
italian
salad
bbq
seafood
bakeries
newamerican
sushi
bars
0
50
100
150
200
250