What's in an A?

Using Classification and Data Exploration to Predict NYC's Restaurant Ratings

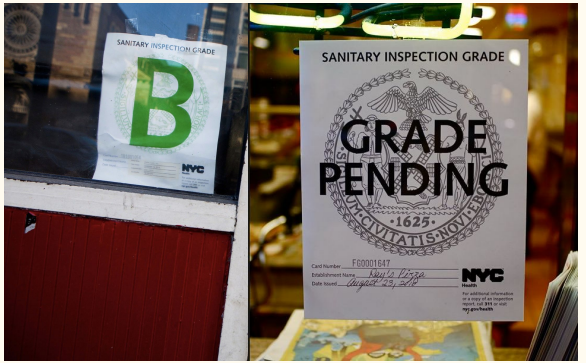
Brian Newborn - May 2018

The Question

Can we predict whether a restaurant will get an A on their Health inspection? And, can we quantify the cost of possible government interventions to boost ratings?

What is a restaurant grade?

- Yearly, at least
- “Point” system
  - A = 0-13
  - B = 14-27
  - C = 28+
- Mandatory closing if a “public health hazard” is found
- While this seems like a mainstay of NYC culture, it has only been around since 2010

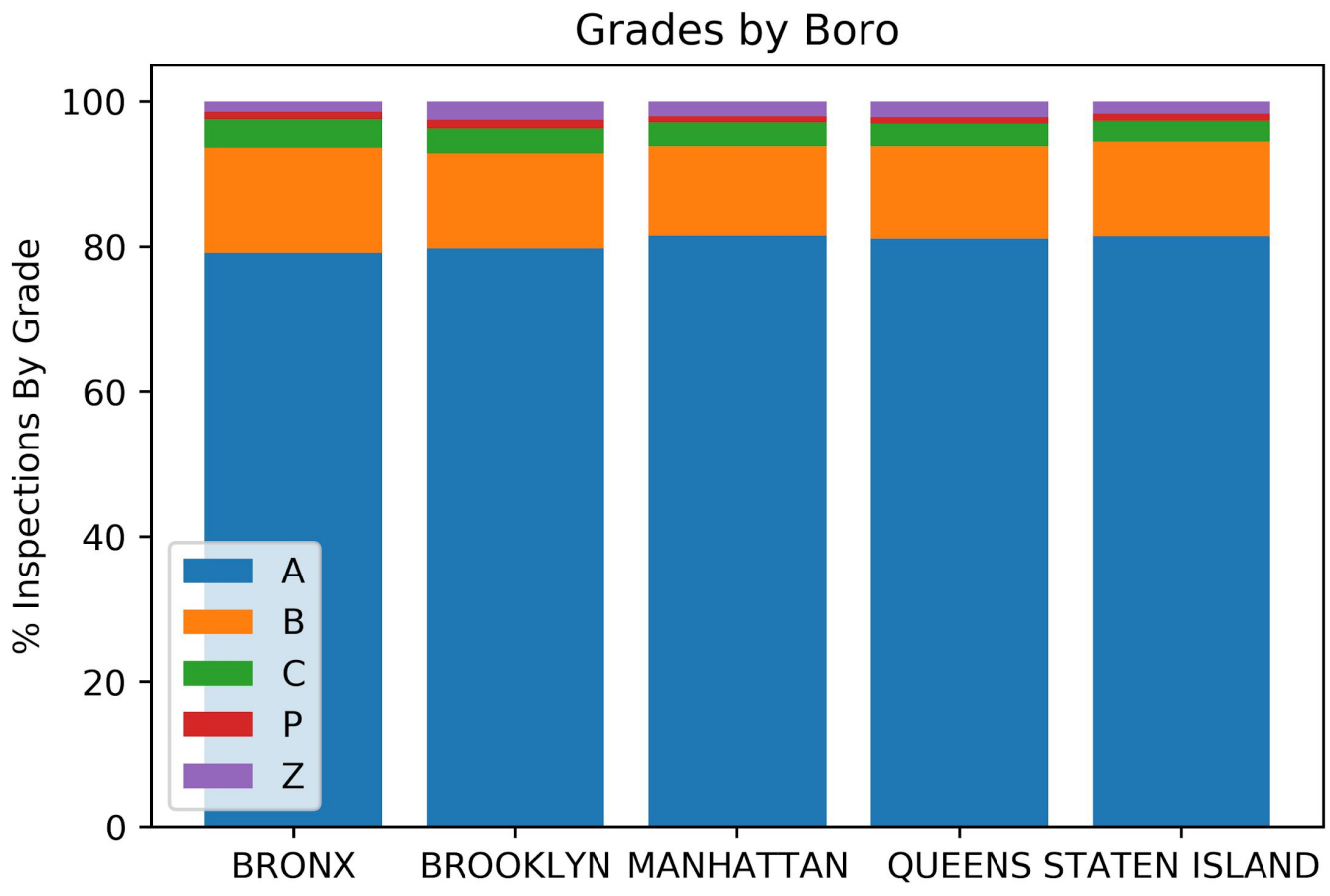


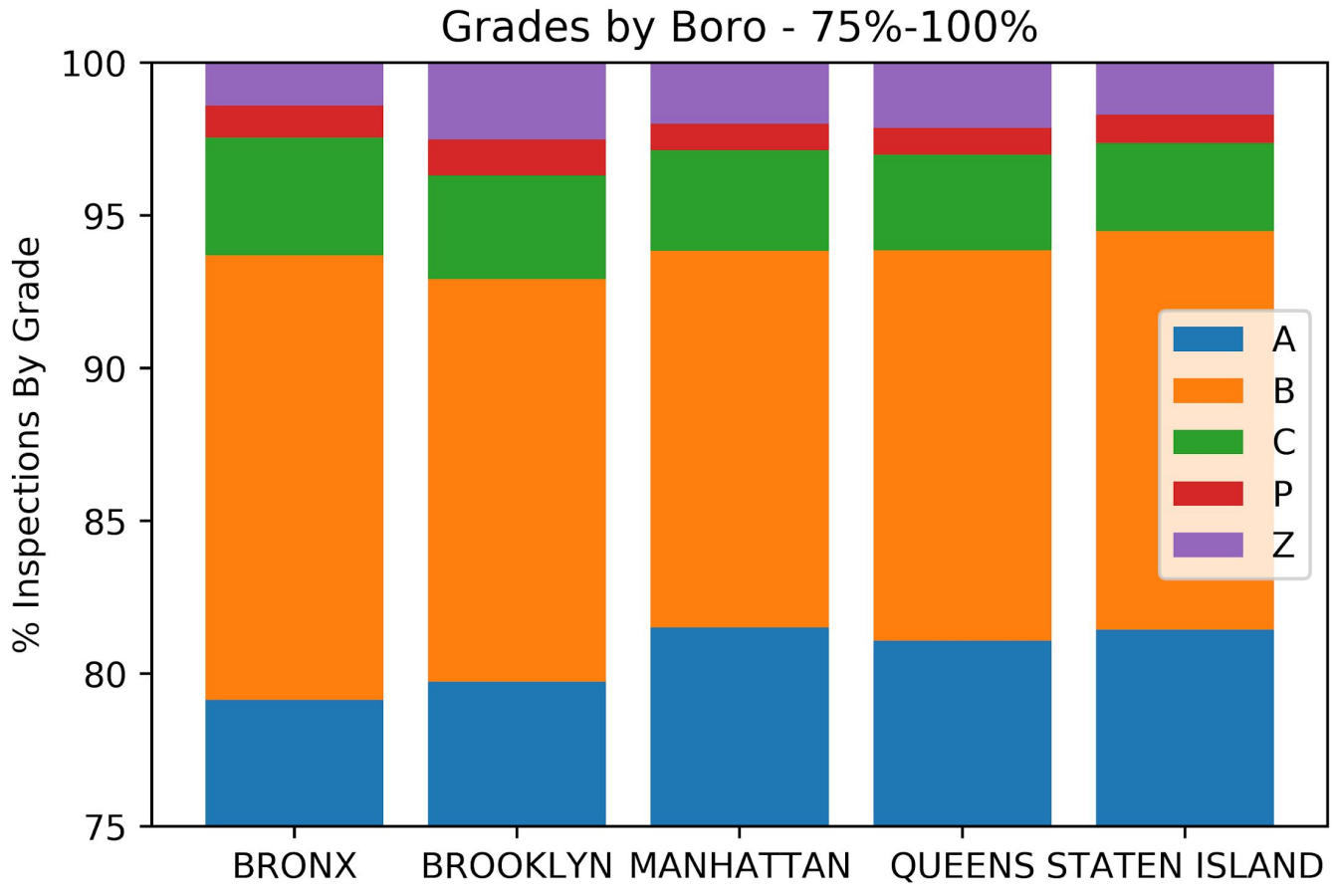
Data

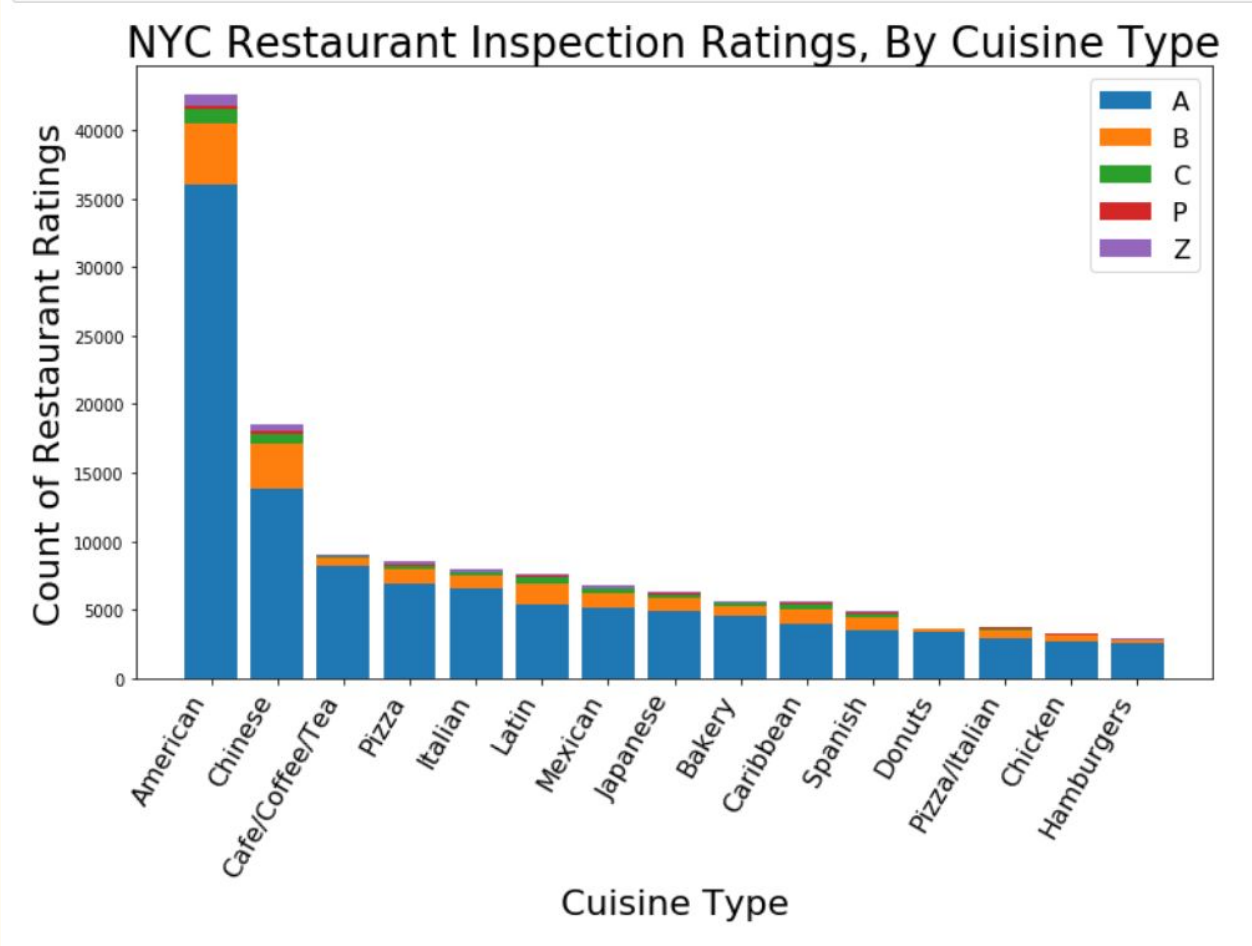
- The city posts all grades and reviews online via NYC OpenData
- Joined on US Census Data for income statistics

What’s in the Data

- Vast majority of restaurants get “A” ratings
  - ~90% Restaurants have “A” as of 2018
- Improvement in “A” rate year over year
  - NYC’s Public Health statistics suggest a decrease in foodborne illnesses, as well


Grades by Boro
% Inspections By Grade
100
80
60
40
20
0
A
B
C
P
Z
BRONX
BROOKLYN
MANHATTAN
QUEENS
STATEN ISLAND


Grades by Boro - 75%-100%
% Inspections By Grade
100
95
90
85
80
75
BRONX
BROOKLYN
MANHATTAN
QUEENS
STATEN ISLAND
A
B
C
P
Z


NYC Restaurant Inspection Ratings, By Cuisine Type
Count of Restaurant Ratings
0
5000
10000
15000
20000
25000
30000
35000
40000
A
B
C
P
Z
American
Chinese
Cafe/Coffee/Tea
Pizza
Italian
Latin
Mexican
Japanese
Bakery
Caribbean
Spanish
Donuts
Pizza/Italian
Chicken
Hamburgers
Cuisine Type

Model Inputs

All restaurant inspections 2014-2018

- Cuisine Type (Chinese, Deli, American etc.) - almost 100 labels
- Month of inspection
- Measurement of Neighborhood Wealth - per ZIP Code
- Longitude/Latitude
- Boro

“A Ratings” (87%)

vs

“Not A Ratings” (13%)

Measurement of Success - CBA

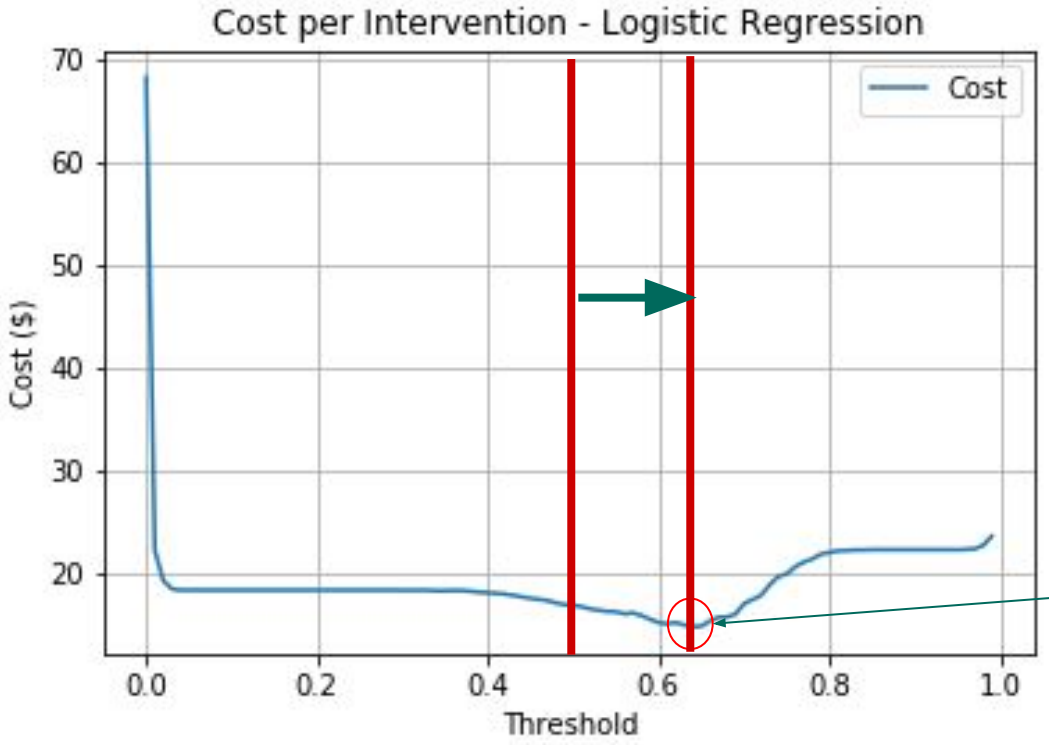
Cost of Intervention = $100

Cost of 2nd Inspection and Appeals Process = $250

Savings of Preventing B/C/Failure = 250-100 = $150

CB = $250*(Rate of False Negatives) + 100*(Rate of False Positives) - $150*(Rate of True Positives)

Logistic Regression Cost Effectiveness



Raising Threshold does the following:

- Greatly reduces false positives
- Increases false negatives at a lesser rate

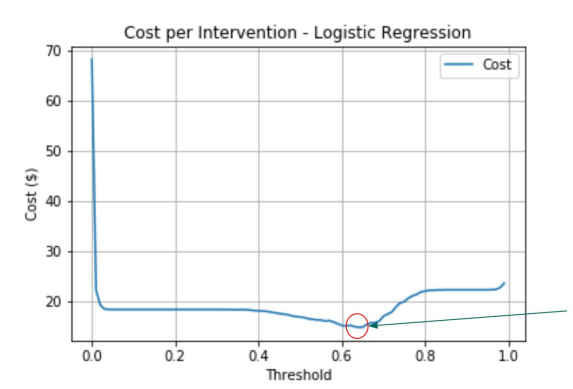
Minimum Cost

-Threshold = .62

-Cost Per Intervention = $15.57

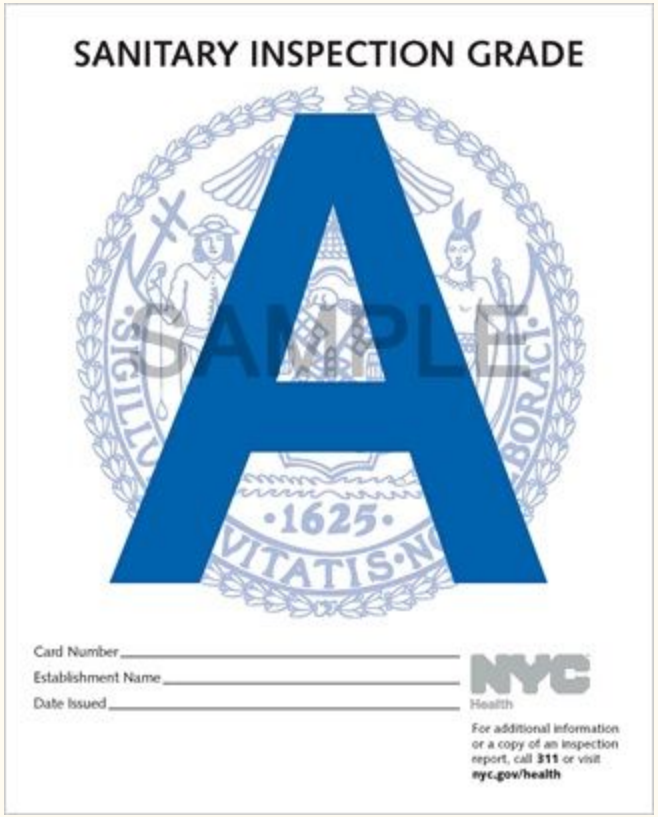
Takeaways

- By increasing our threshold to .62, the DOHMH can achieve a cost of apx. $15 per intervention
- This method increases our predicted A restaurants but dramatically reduces false positives enough to lower costs



True A	*Predicted A Correctly*	*False Positives*
True Not A	*False Negatives*	*Predicted Not A Correctly*
----------	**Predicted A (0)**	**Predicted Not A (1)**

Thank You!


SANITARY INSPECTION GRADE
A
SAMPLE
Card Number
Establishment Name
Date Issued
NYC
Health
For additional information or a copy of an inspection report, call 311 or visit nyc.gov/health

Thank You!

Further Reading & Sources

- How We Score and Grade - NYC DOH
- Restaurant Grades Home Page - NYC DOH
- NYC Open Data
- 2012 - Bloomberg Links Drop in Food Poisoning to Restaurant Grades
- 2017 - NYTimes: What do Restaurant Grades Mean?
- 2017 - Crain's: More restaurants are passing their health exams
- 2012 - Blog: "NYC Dining: The Cost of an A"
- 2015: NYU/New School - Is Public Grading Worth the Costs? An Evaluation of New York City’s Restaurant Grades Policy
- 2014 - Harvard Data Science: Big Data Analysis of NYC Health Inspections

Logistic Regression

- Changing sampling style (SMOTE) did not improve model
- Low **Precision** for Non A restaurants - 0.19
- High **Recall**

Feature	Approximate Coefficient Value
Cycle: Initial Inspection	-2
Cycle: Reopening Inspection	+1.2
Cycle: Re-inspection	+.75
Pre-Permit: Initial Inspection	-0.5
Pre-Permit: Re-Inspection	+0.3
American Restaurant	-0.3
Chinese Restaurant	+0.2

Accuracy Score = ~.89 both with over and regular sampling

Baseline = .87

Random Forest

- Oversampling non-A inspections for training boosted results
- Highest **Precision** for Non A > .40

Accuracy Score: .91-.92 with oversampling

Logistic Regression = .89

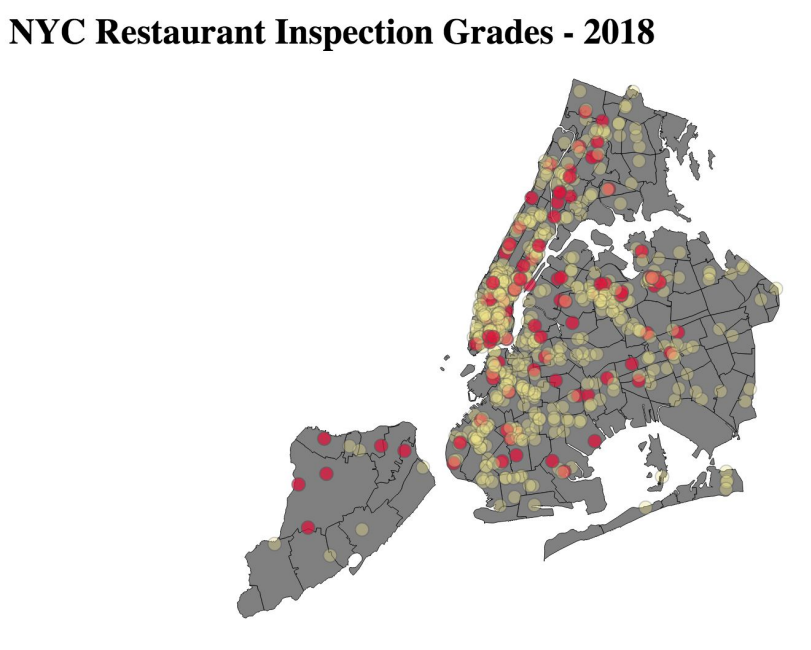
Baseline = .87

Appendix

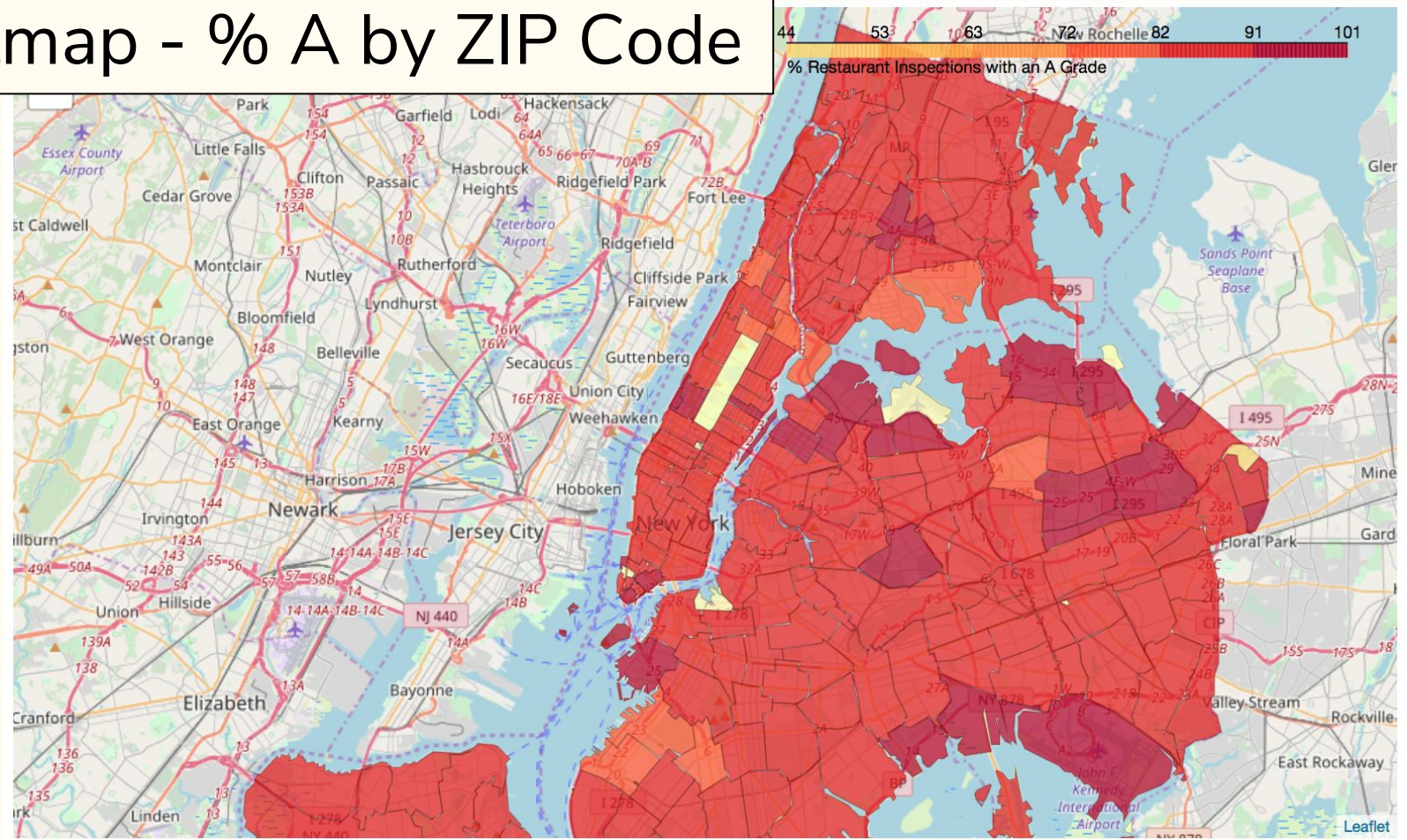
1) D3 2018 Inspections Map
2) Folium Heatmap
3) Logistic Regression ROC/AUC
4) Logistic Regression Confusion Matrix
5) Random Forest ROC/AUC
6) Random Forest Confusion Matrix

Seeing the Data

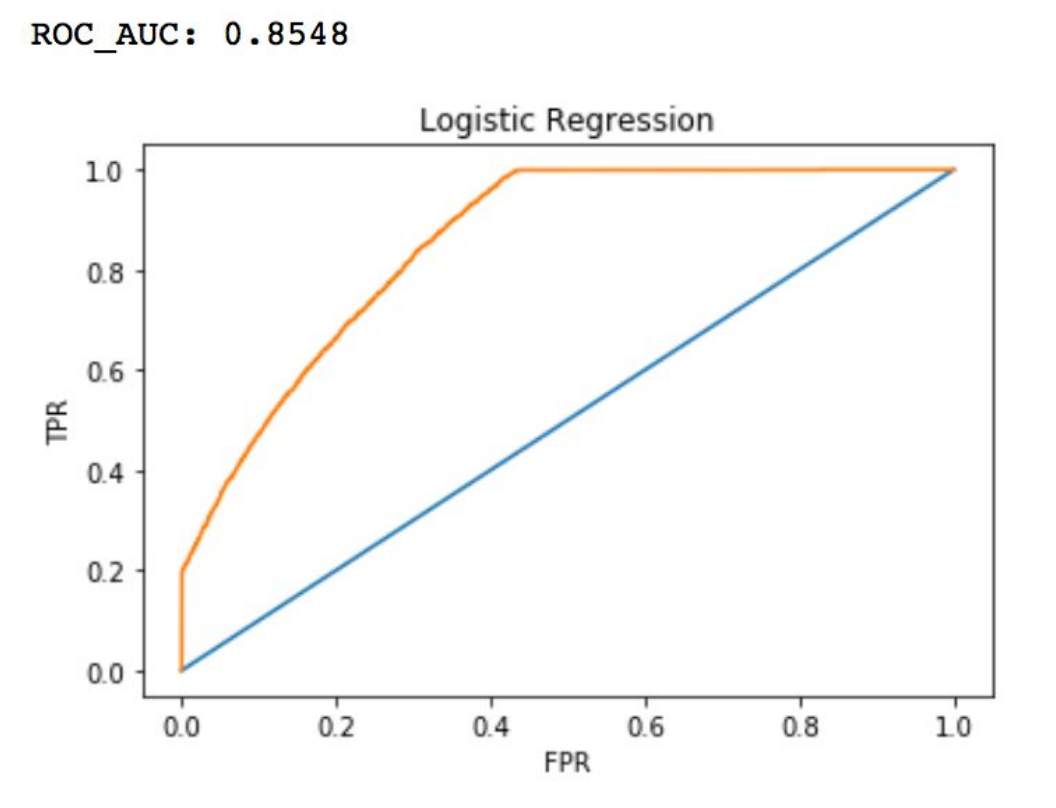
We also put together an interactive, by inspection map, for review.



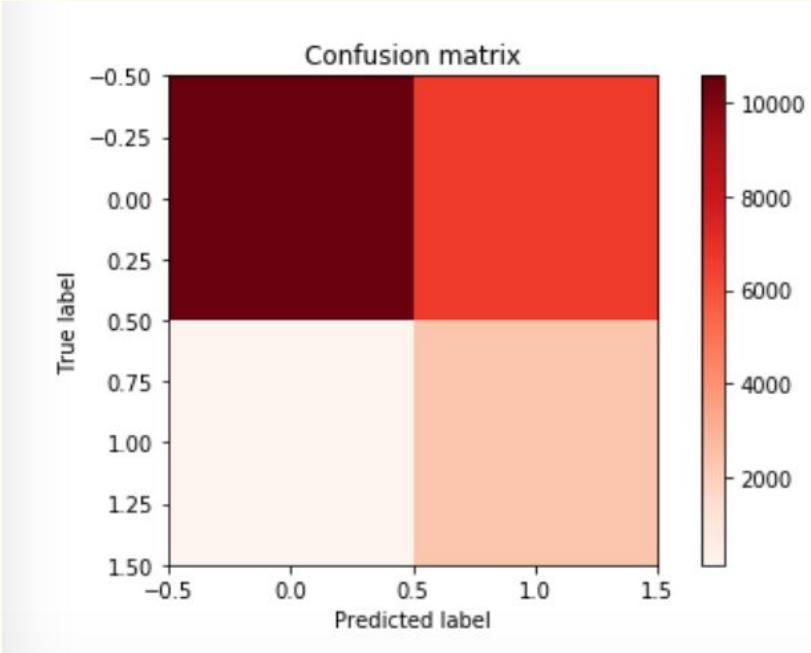
Heatmap - % A by ZIP Code



Logistic Regression ROC/AUC



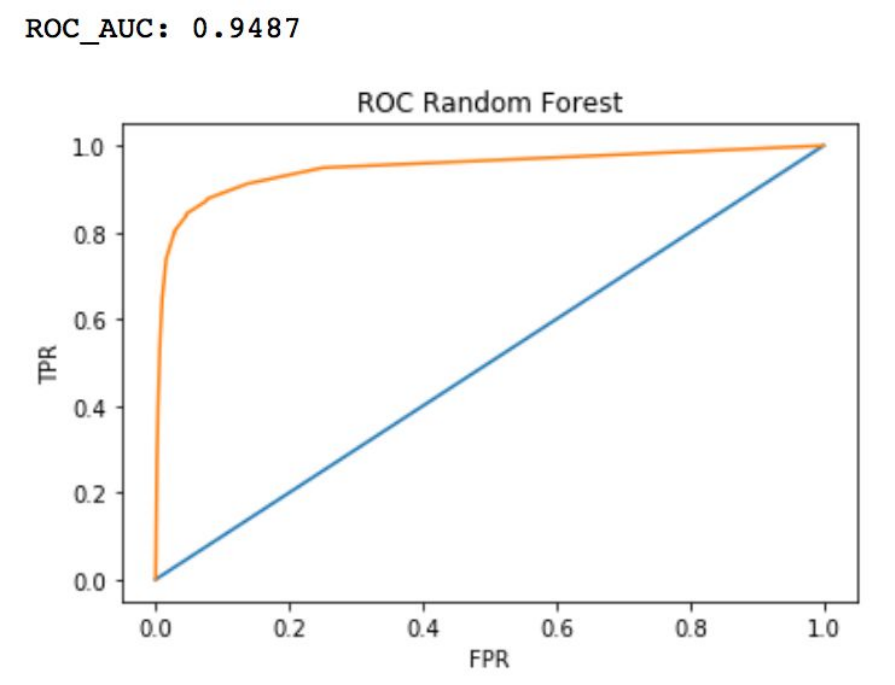
Logistic Regression Confusion Matrix



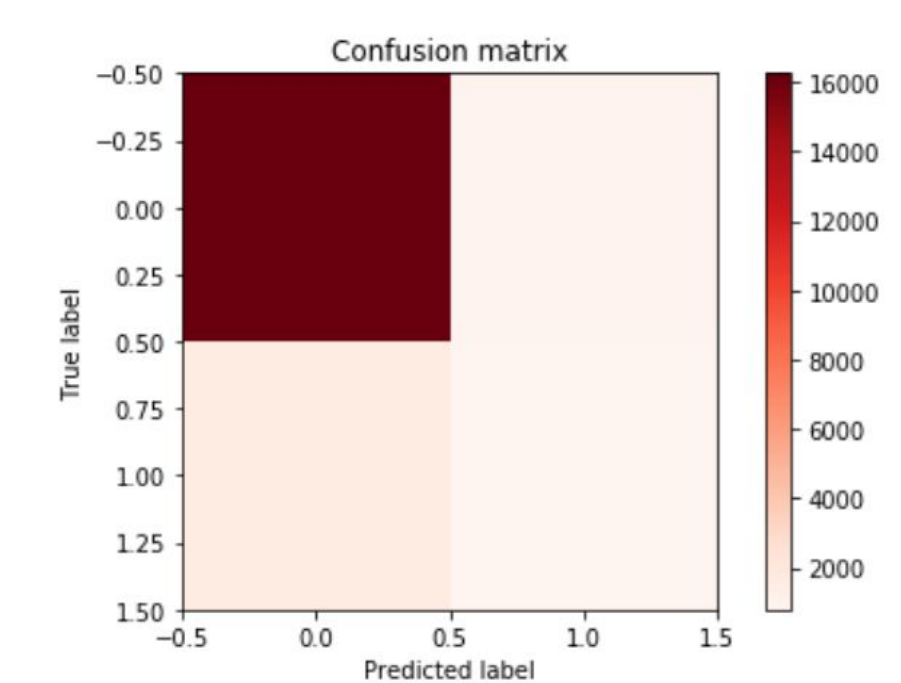

```
[[10559   6634]
 [  140   2414]]
```

```
# True A                  # False Predicted Non A
# False Predicted A       # True Predicted Non A
```

Random Forest ROC/AUC



Random Forest Confusion Matrix




```
[[16256   973]
 [ 1726   792]]
```

```
# True A                  # False Predicted Non A
# False Predicted A       # True Predicted Non A
```