

# What's in an A?



Using Classification 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

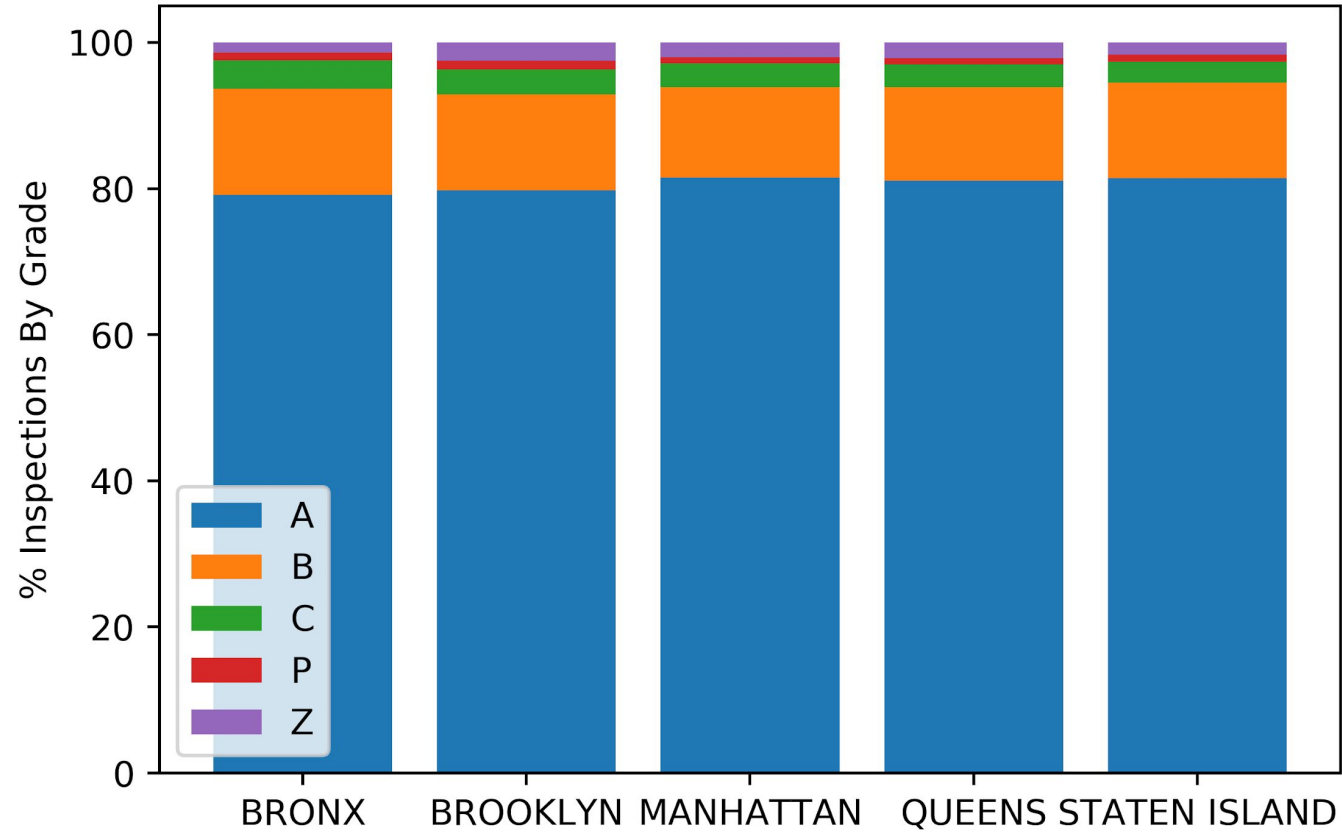
The logo for NYC OpenData. It features the letters "NYC" in a bold, blue, sans-serif font. To the right of "NYC" is the word "OpenData" in a dark grey, sans-serif font. The entire logo is set against a white rectangular background.The logo for the United States Census Bureau. It features the words "United States" in a blue, sans-serif font, positioned above the word "Census" in a larger, bold, blue, sans-serif font. Below "Census" is a thick blue horizontal line, and to the right of the line is the word "Bureau" in a blue, sans-serif font. The entire logo is set against a white rectangular background.

# What's in the Data

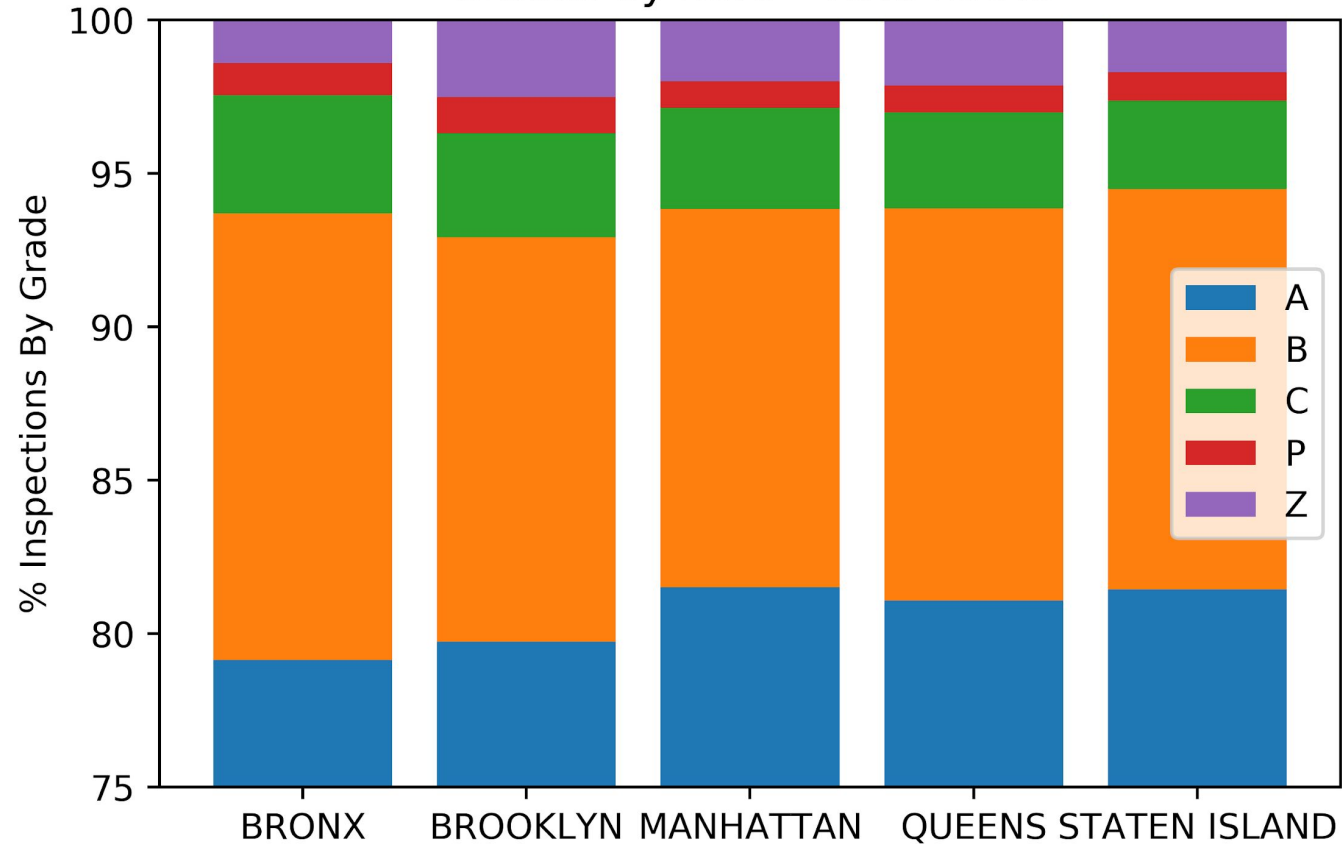


- Vast majority of restaurants get “A” ratings
  - ~90% Restaurants have “A” as of 2018
- Improvement in “A” rate year over year
  - NYC's Public Health statistics suggest a decrease in foodborne illnesses, as well

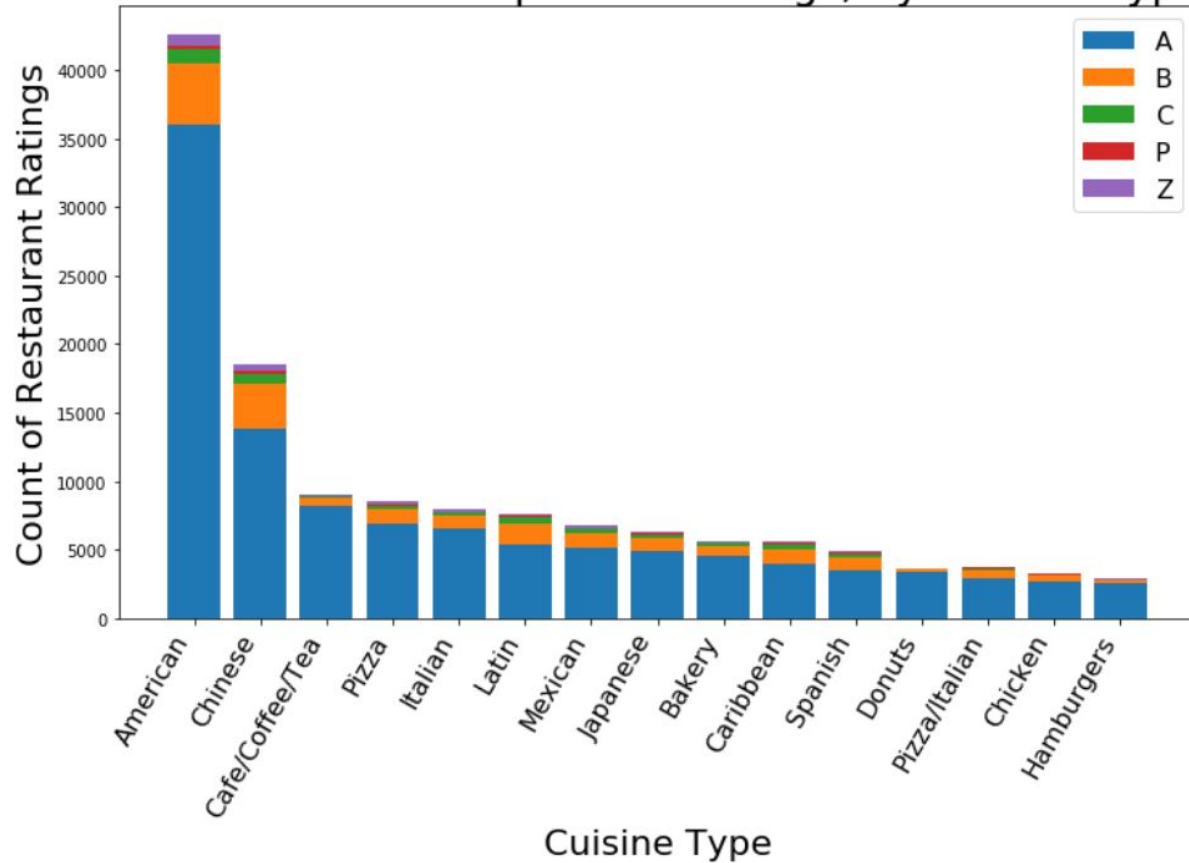
Grades by Boro



Grades by Boro - 75%-100%



# NYC Restaurant Inspection Ratings, By 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



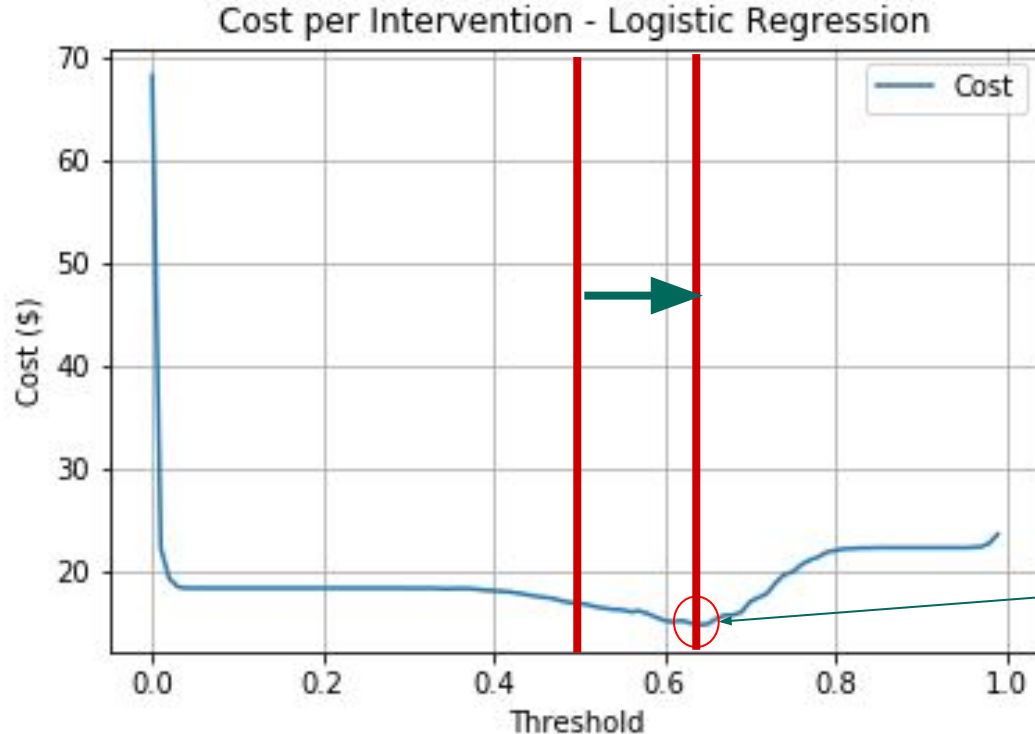
Cost of Intervention = \$100

Cost of 2nd Inspection and Appeals Process = \$250

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

CB =  $\$250 * (\text{False Negatives}) + 100 * (\text{False Positives}) -$   
 $\$150 * (\text{True Positives})$

# Logistic Regression Cost Effectiveness



Raising Threshold does the following:

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

## Minimum Cost

-Threshold = .62

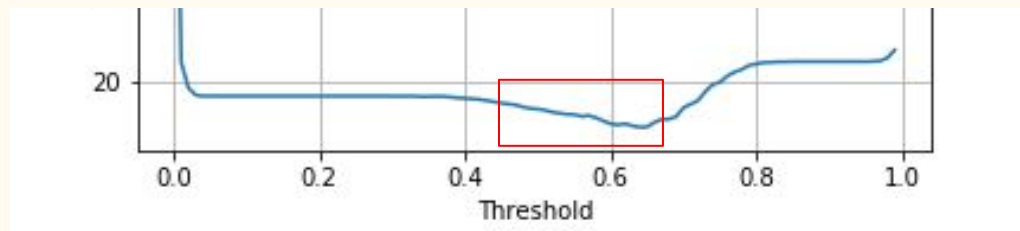
-Cost Per Intervention = \$15.57

Threshold = 0.5

True A	54.1%	33.1%
True Not A	.1%	11.9%
----- -----	Predicted A (0)	Predicted Not A (1)

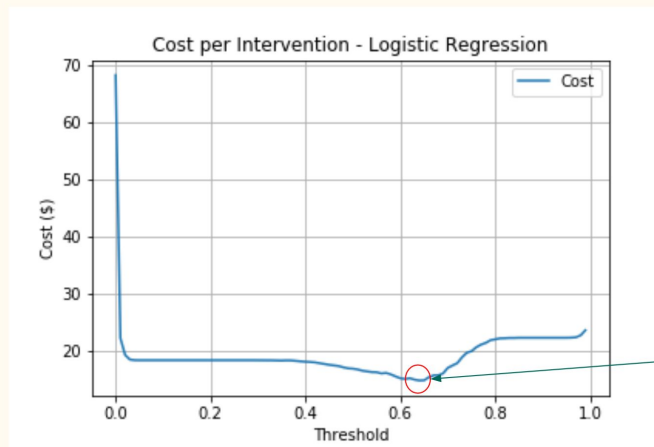
Threshold = 0.62

True A	64.3%	23%
True Not A	3%	9.8%
----- -----	Predicted A (0)	Predicted Not A (1)



# Takeaways


- Raising the bar for the model to predict a “Not A” result, the DOHMH can achieve a **cost of apx. \$15 per intervention**
- Increases our predicted A restaurants but dramatically reduces false positives
- This will help the Department target as good of a swath of restaurants as possible for their outreach program



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**



**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](http://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**

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

*Baseline = .87*

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

# 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) Logistic Regression Coefficients
- 2) D3 2018 Inspections Map
- 3) Folium Heatmap
- 4) Logistic Regression ROC/AUC
- 5) Logistic Regression Confusion Matrix
- 6) Random Forest ROC/AUC
- 7) Random Forest Cost
- 8) Random Forest Confusion Matrix

# Logistic Regression

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

**0 - “A”  
Grade**

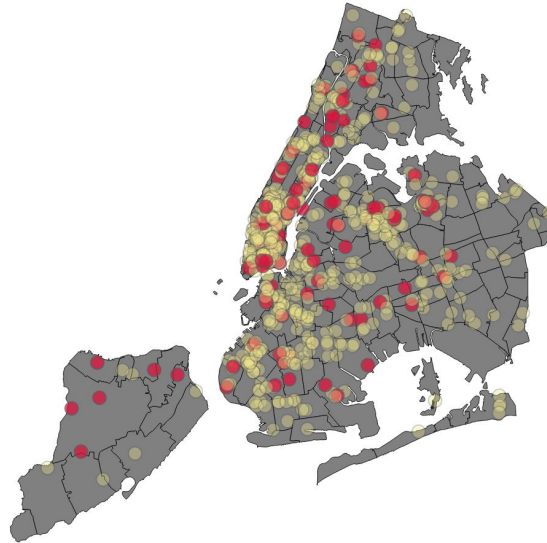
**1 - “Non A”  
Grade**

---

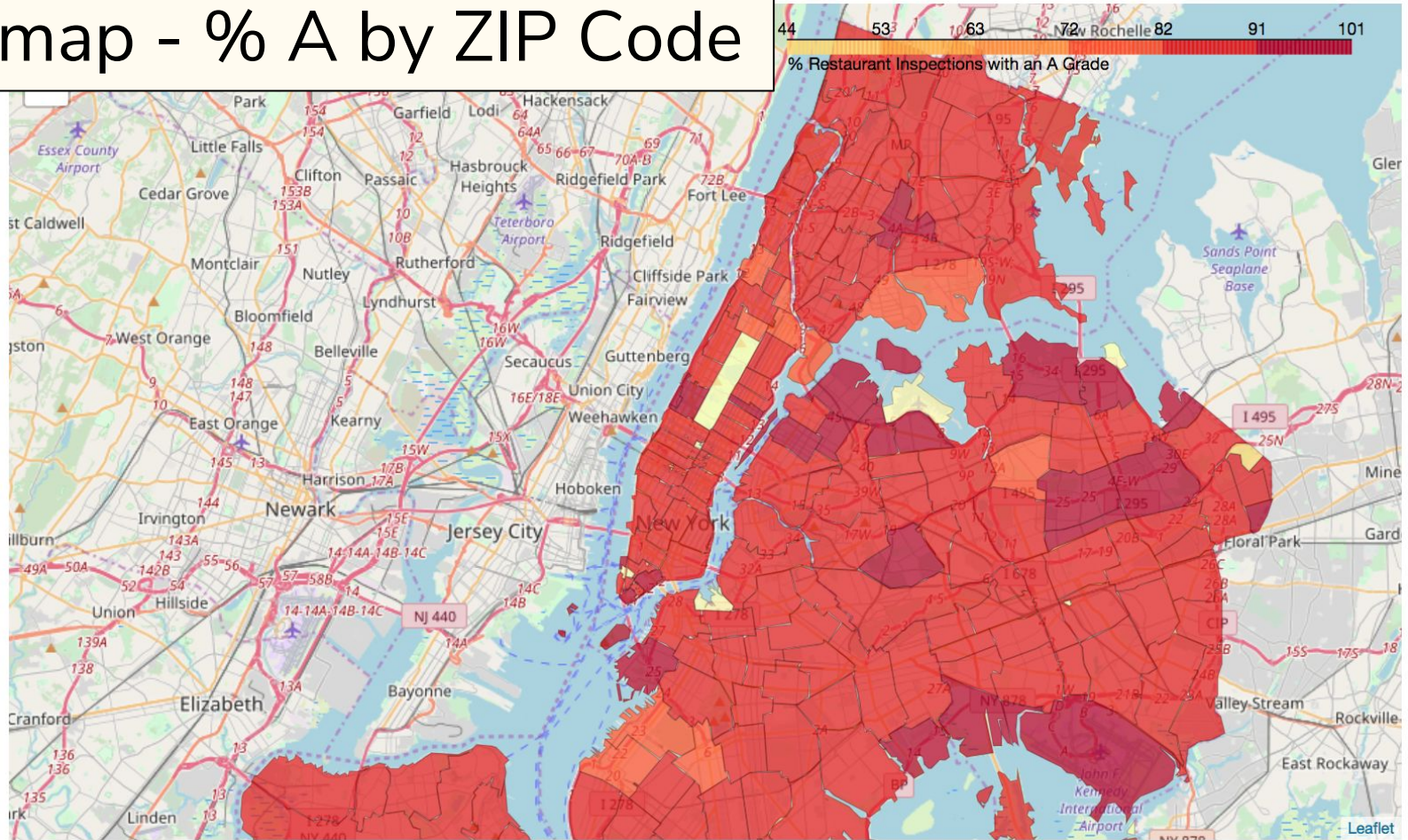
# Seeing the Data

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

**NYC Restaurant Inspection Grades - 2018**



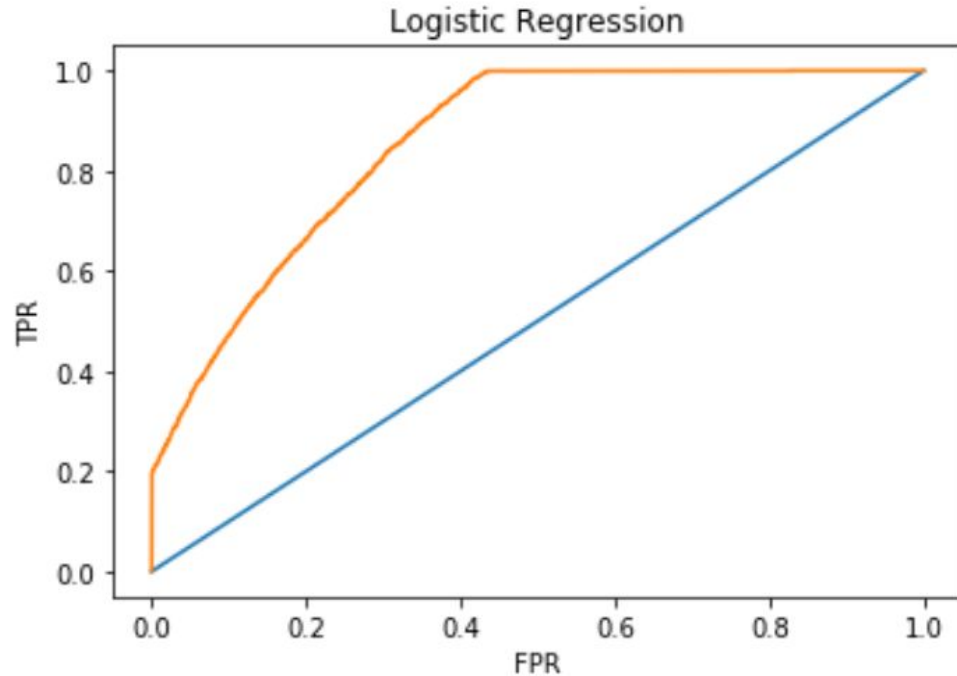
# Heatmap - % A by ZIP Code



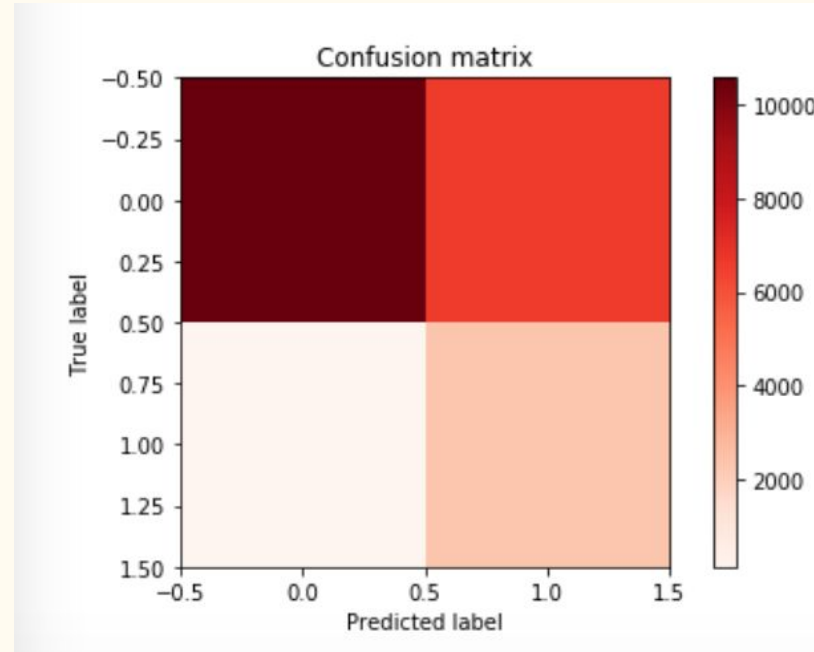


# Logistic Regression ROC/AUC

ROC\_AUC: 0.8548



# Logistic Regression Confusion Matrix



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

*# True A*

*# False Predicted Non A*

*# False Predicted A*

*# True Predicted Non A*



19747

# Logistic Regression Confusion Matrix

Threshold at 0.51

```
[[10698  6539]
 [   168  2342]]
```

Threshold at 0.61

```
[[12702  4535]
 [   584  1926]]
```

True A	54.1%	33.1%
True Not A	.1%	11.9%
----- -----	Predicted A (0)	Predicted Not A (1)

# True A

# False Predicted A

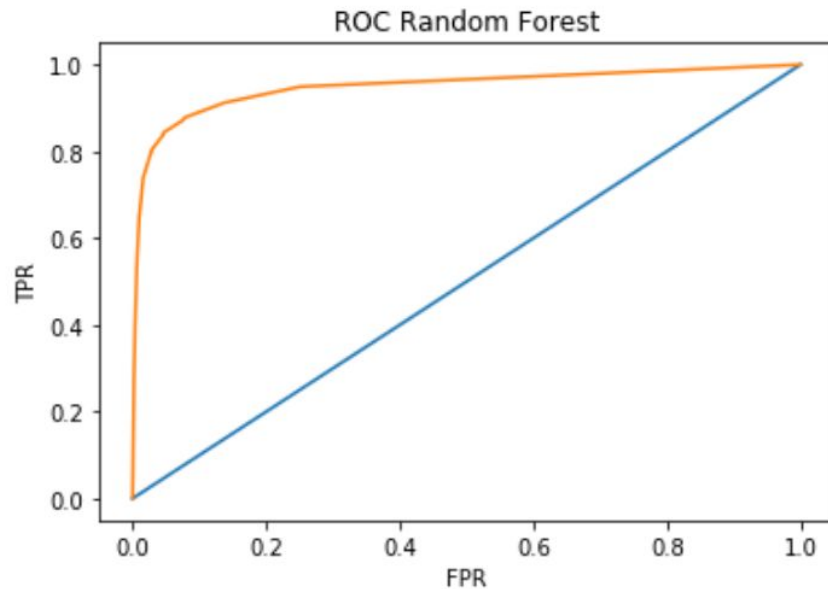
True A	64.3%	23%
True Not A	3%	9.8%
----- -----	Predicted A (0)	Predicted Not A (1)

# False Predicted Non A

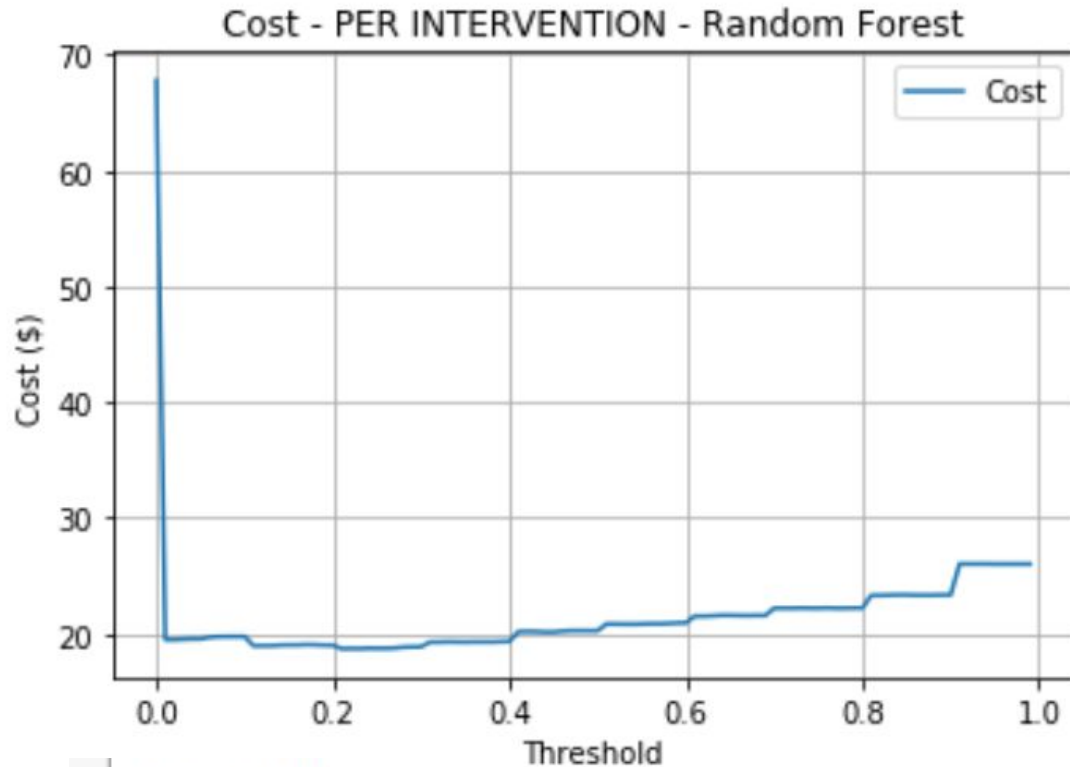
# True Predicted Non A

# Random Forest ROC/AUC

ROC\_AUC: 0.9487



# Random Forest Cost

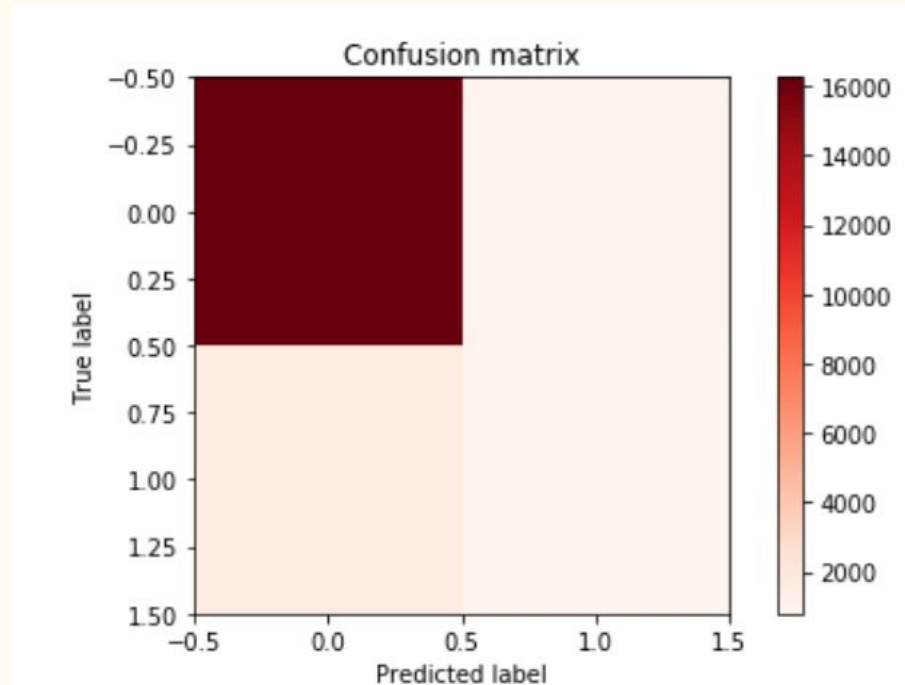


17.6710

\*\*\*\*\*

```
[[13308 3946]
 [ 824 1669]]
```

# Random Forest Confusion Matrix



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

*# True A*

*# False Predicted Non A*

*# False Predicted A*

*# True Predicted Non A*