Restaurant Inspections in NYC

Eric He, Aparajita Taneja, Jinsan Kim, Marcos Galante

The Question

Can a restaurant inspection's grade be predicted using the restaurant's identifying information and inspection violations?

26,216

Unique Restaurants in NYC

147,623

Restaurant Inspections Since 2012

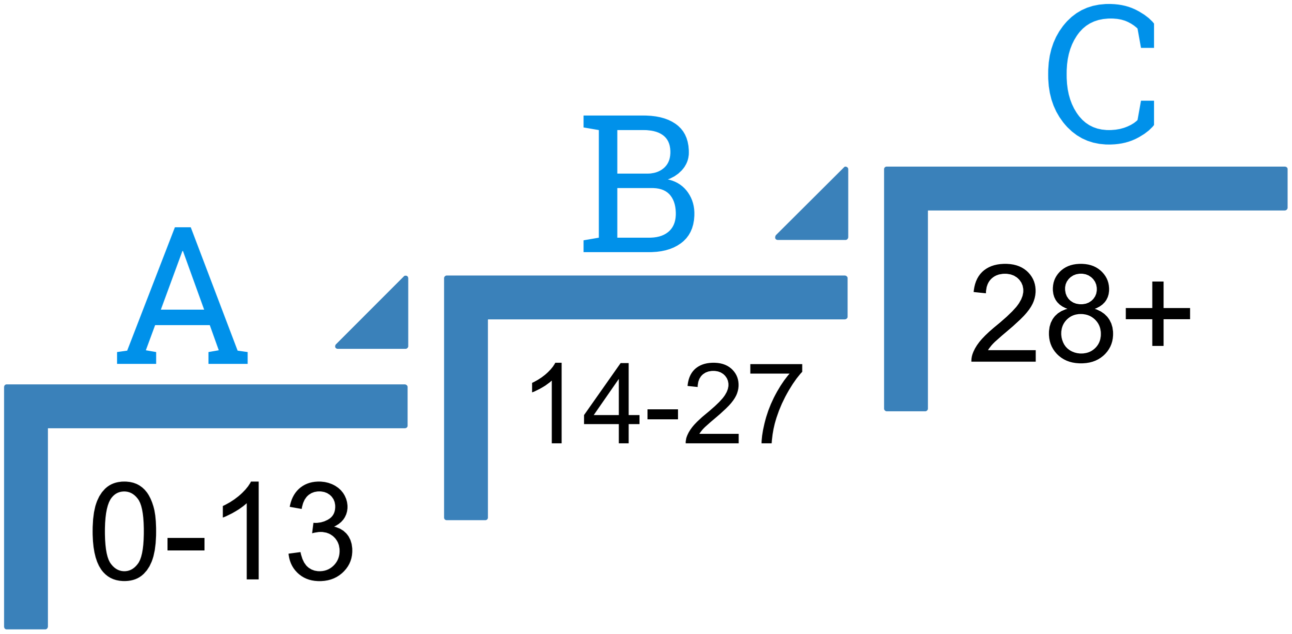
428,405

Total Restaurant Code Violations

Source: https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j/data

Data Example: The Golden Unicorn

Type: Chinese
Borough: Manhattan
ZIP Code: 10002
Violation Code: 06A
Description: “Personal cleanliness inadequate.”
Flag: Critical Violation
Score: 12
Grade: A



A

0-13

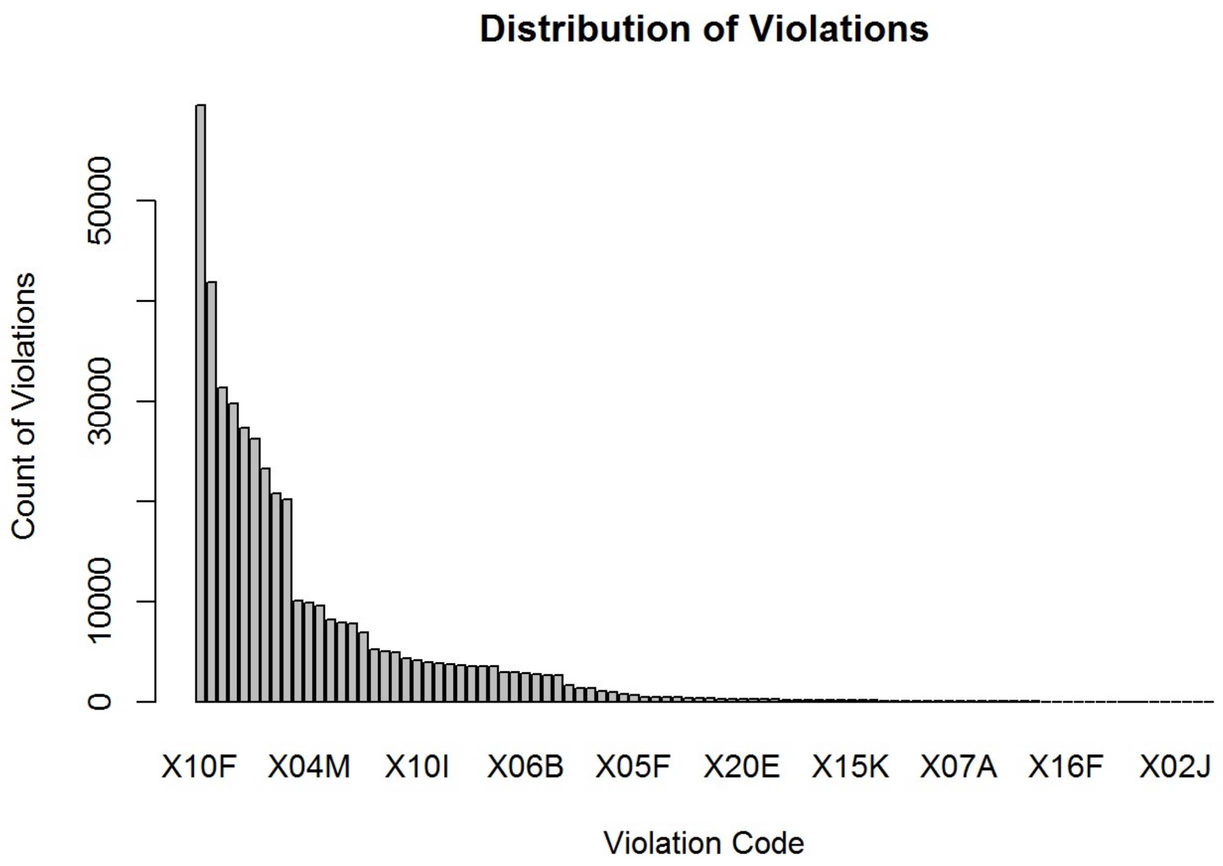
B

14-27

C

28+

Not all violations are created equal



The four most frequent violations

10F **59,491 occurrences**

Non-food contact surface improperly constructed. Unacceptable material used.

08A **41,905 occurrences**

Facility not vermin-proof. Harborage or conditions conducive to attracting vermin to the premises and/or allowing vermin to exist.

02G **31,317 occurrences**

Cold food item held above 41º F (smoked fish and reduced oxygen packaged foods above 38º F) except during necessary preparation.

04L **29,735 occurrences**

Evidence of mice or live mice present in faculty’s food and/or non-food areas

We regressed inspection score on inspection violations to see which violations were more severe

```
## Coefficients: (1 not defined because of singularities)
##                 Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)     -0.17500  0.02324    -7.530   5.10e-14
## X02A            10.35111  0.21174    48.885   < 2e-16
## X02B            7.81572   0.02951    264.892  < 2e-16
...
## X22F            0.79420   0.25346    3.133    0.001728
## X22G            -0.79789  0.65068    -1.226   0.220111
## ---
## Residual standard error: 3.903 on 142941 degrees of freedom
## (4588 observations deleted due to missingness)
## Multiple R-squared: 0.8529, Adjusted R-squared: 0.8528
## F-statistic: 8914 on 93 and 142941 DF, p-value: < 2.2e-16
```

Selected severe violations

07A **Associated with score increase of 29.46**

Duties of an officer of the Department interfered with or obstructed.

29.29

06H

Records and logs not maintained to demonstrate that HACCP plan has been properly implemented.

05E **23.21**

Toilet facility not provided for employees or for patrons when required

04F

21.54

Food, food preparation area, food storage area, area used by employees and patrons, contaminated by sewage or liquid waste.

The Modeling

Four models used

- ANOVA/Linear Regression
- Naïve Bayes
- Classification Tree
- Random Forest

Predictors

- Violations
- Restaurant Type
- Zipcode
- Borough

Linear Regression used numeric scores and converted to grade classifications; others used categorical grade classifications from the get-go.

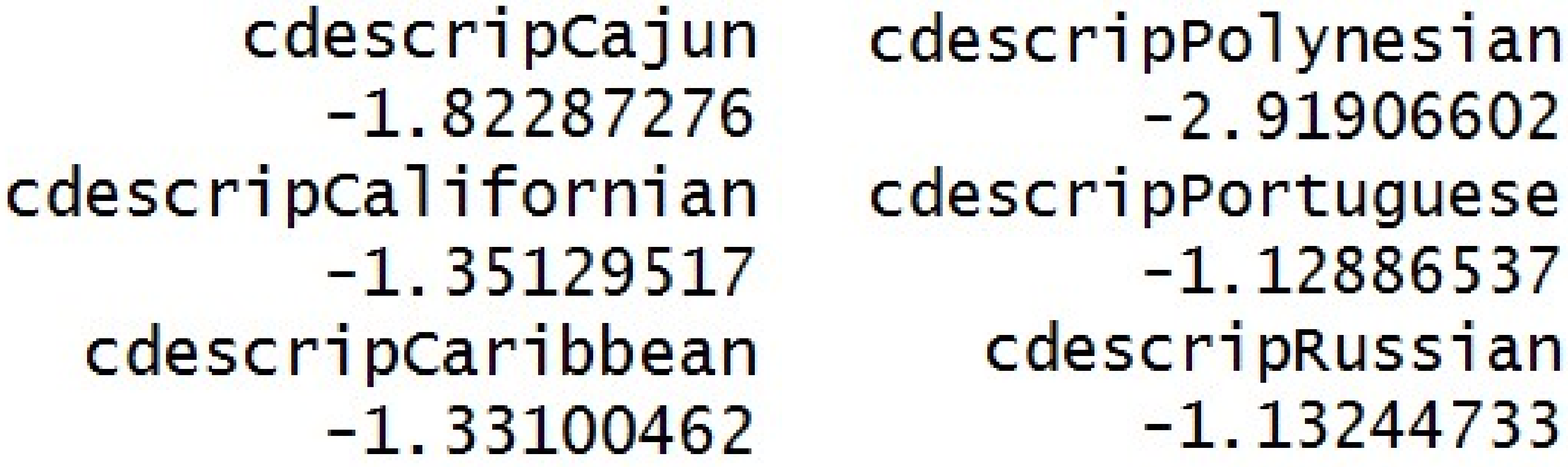
Method: ANOVA/Linear Regression

- Best at predicting B and C scores
- Zipcode was not a statistically significant variable, so it was removed

	Actual	A	B	C
Predicted	A	7423	156	44
	B	1568	3287	250
	C	1	343	1058

Method: ANOVA/Linear Regression

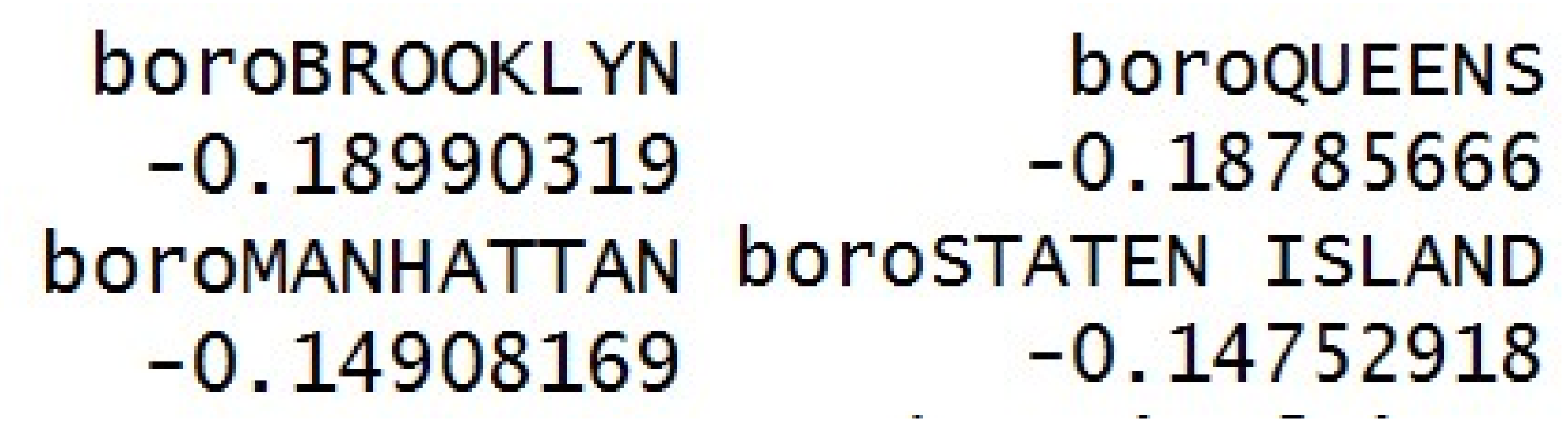
- All restaurant classifications had negative coefficients, with Polynesian restaurants having the smallest coefficient of -3
- Only minor effects associated with borough



```
     cdescripCajun   cdescripPolynesian
       -1.82287276          -2.91906602
cdescripCalifornian  cdescripPortuguese
       -1.35129517          -1.12886537
  cdescripCaribbean     cdescripRussian
       -1.33100462          -1.13244733
```

Method: ANOVA/Linear Regression

- All restaurant classifications had negative coefficients, with Polynesian restaurants having the smallest coefficient of -3
- Only minor effects associated with borough



```
  boroBROOKLYN         boroQUEENS
   -0.18990319        -0.18785666
boroMANHATTAN boroSTATEN ISLAND
   -0.14908169        -0.14752918
```

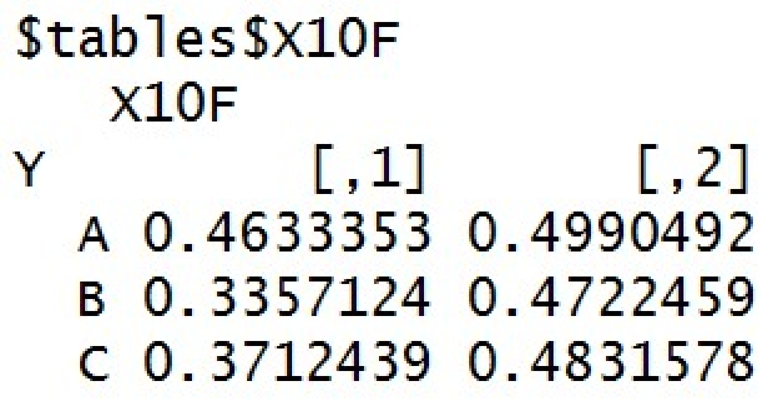
Method: Naïve Bayes

- **Strong tendency to predict A scores**
- **Possible model improvements with different classification percentage**

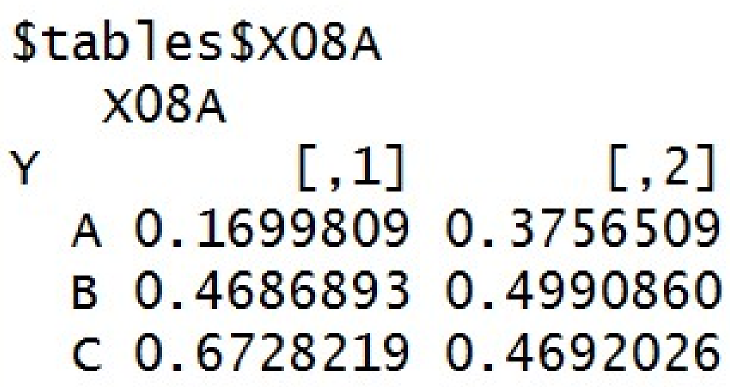
	Actual	A	B	C
Predicted	A	8459	2504	102
	B	432	1160	926
	C	101	122	324

Method: Naïve Bayes

- **Most probability estimates were very small**
- **However, “severe” violations from before continued to have the most signal**



```
$tables$X10F
   X10F
Y          [,1]      [,2]
  A 0.4633353 0.4990492
  B 0.3357124 0.4722459
  C 0.3712439 0.4831578
```

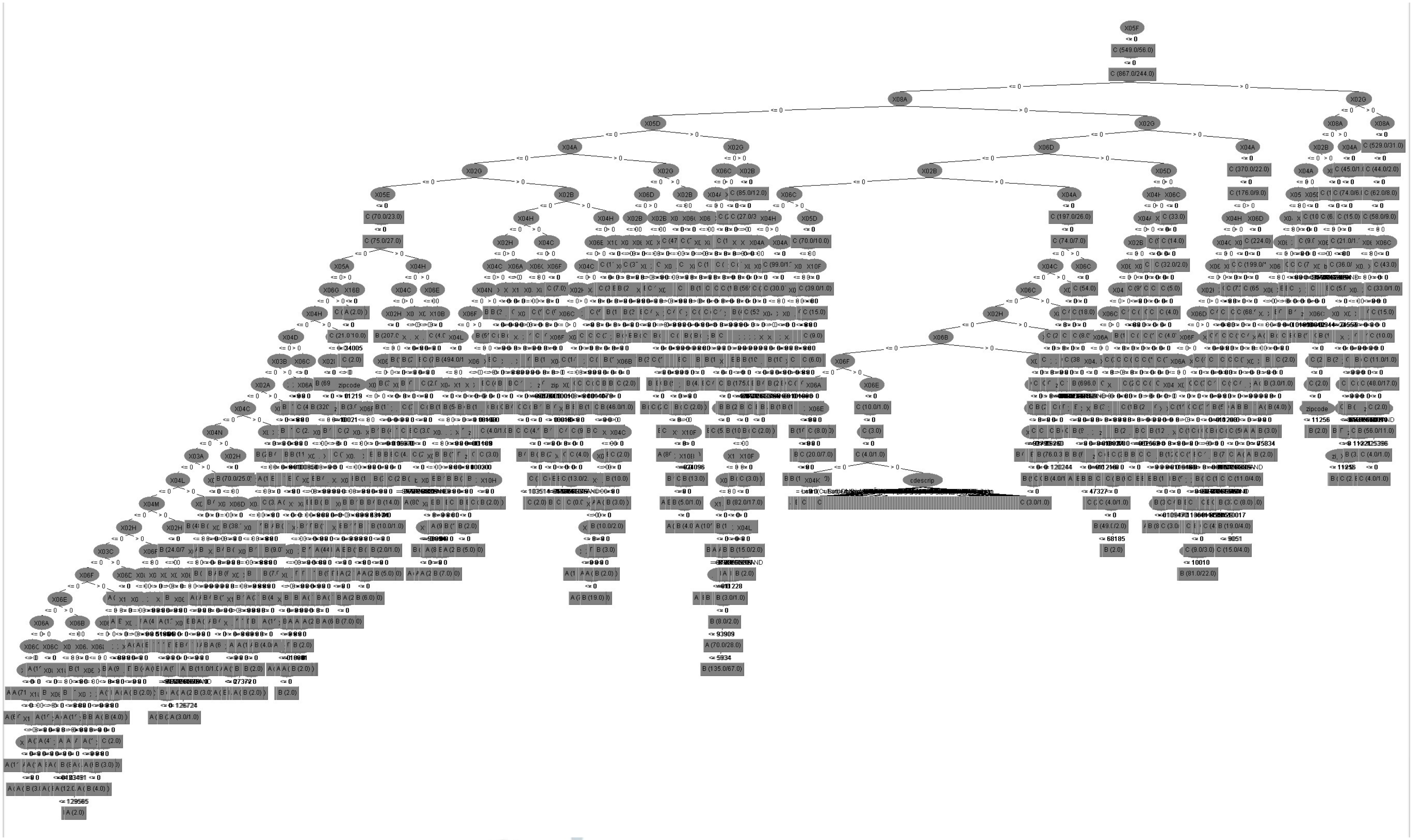


```
$tables$X08A
   X08A
Y          [,1]      [,2]
  A 0.1699809 0.3756509
  B 0.4686893 0.4990860
  C 0.6728219 0.4692026
```

Method: Classification Tree

- **Strong performance at predicting A, B scores**
- **Borough, Zipcode, Restaurant Type always ended up being the last splits**

	Actual	A	B	C
Predicted	A	8858	427	89
	B	117	3207	465
	C	15	152	798


X05F
<= 0
C (549.0/56.0)
<= 0
C (867.0/244.0)
<= 0
> 0
X08A
X02G
X05D
X02G
X08A
X08A
C (529.0/31.0)
X04A
X02G
X06D
X04A
X02B
X04A
C (370.0/22.0)
C (176.0/9.0)
C (45.0/1.0)
C (44.0/2.0)
C (74.0/6.0)
C (62.0/8.0)
C (58.0/9.0)
X02G
X02B
X06C
X02B
C (85.0/12.0)
X02B
X05D
X05E
C (70.0/23.0)
C (75.0/27.0)
X05A
X04H
X06C
X05D
C (70.0/10.0)
C (197.0/26.0)
C (74.0/7.0)
X04C
X06C
C (54.0)
X02H
X06B
X06F
X06E
C (10.0/1.0)
C (3.0)
C (4.0/1.0)
cdescrip
C (3.0/1.0)
B (135.0/67.0)
A (70.0/28.0)
B (8.0/2.0)
B (81.0/22.0)
A (2.0)

Method: Classification Tree

05F

Insufficient or no refrigerated or hot holding equipment to keep potentially hazardous foods at required temperatures.

08A

Facility not vermin-proof. Harborage or conditions conducive to attracting vermin to the premises and/or allowing vermin to exist.

02G

Cold food item held above 41 F (smoked fish and reduced oxygen packaged foods above 38 F) except during necessary preparation

Method: Random Forest

- **Performed worse than the classification tree**
- **The model was quite difficult to read**

	Actual	A	B	C
Predicted	A	8850	795	141
	B	122	2919	714
	C	18	72	497

What to do next

Incorporate date and previous inspection grades in the model

Pursue a more thorough model selection process

Look into using different classification thresholds

Thanks!
Any Questions?