Food Safety and Restaurant Inspection

Introduction

This document provides a brief description of restaurant inspection data by the DC Department of Health, the key outcome variables, and results of initial predictive analysis for those key outcome variables.

Food establishments that sell or serve food to the public must apply for a health permit and be inspected for compliance with the DC Department of Health. These establishments include restaurants, school cafeterias, bakeries, mobile food vendors, and markets. Most of the permitted food service establishments in Washington, DC receive two routine unannounced food safety inspections per year. DOH follows the federal food code and only closes an establishment for critical violations that cannot be corrected while an inspector is onsite (for several hours) and poses immediate harm to residents and visitors to the District. The purpose of a food safety inspection is to ensure the food is being handled properly from preparation through serving.

This project will help in identifying common factors of establishments that can guide future training efforts for business owners and staff and further help the department allocate staff resources in a more efficient and effective manner.

The primary goals for the project are

1. Analyze the relationship between establishment features and the kind of violations found during inspections
2. Develop a predictive model to predict with a high degree of accuracy the type of violations for each establishment
3. Develop an application that tracks upcoming inspections, displays information on expiring licenses, and aids in the prioritization of inspections based on predicted violations

Data Overview

Table 1 provides a brief description of each variable used in the data set up. The data was scraped from the DC Department of Health website by a volunteer team of Code for DC as part of a separate project. These variables were merged together as a single data file, combining data from three different data files that provided details on potential inspections, potential violations, and geographic codes for each of the establishment inspected.

Table 1: Data Description

Key Variables	Description
Inspection Id	Unique identification number for each inspection
Violation Description	Description of violation found in the inspection

Establishment Name	Name of the establishment inspected
Address	Address of the establishment inspected
Telephone	Phone number of the establishment inspected
Email	Email of the establishment inspected
Inspection Date	Date of inspection
License Holder	Name of license holder
License Number	License number
Establishment Type	Type of establishment including bakery, restaurant, hotel etc.
Risk Category	Categorical variable with five risk categories for establishments - 1 through 5 (1 being least risk prone and 5 being the most risk prone)
Inspection Type	Type of inspection including routine, follow up etc
Total Violations	Number of violations found during inspection
Priority Violations	Number of priority violations found during inspection
Priority Violations Corrected on Site	Number of priority violations corrected during inspection
Priority Violations Repeated	Number of priority violations repeated
Priority Foundation Violations	Number of priority foundation violations found during inspection
Priority Foundation Violations Corrected on Site	Number of priority foundation violations corrected on site during inspection
Priority Foundation Violations Repeated	Number of priority foundation violations repeated
Core Violations	Number of core violations found during inspection
Core Violations Corrected on Site	Number of core violations corrected on site during inspection
Core Violations Repeated	Number of core violations repeated
Critical Violations	Number of critical violations found during inspection
Critical Violations Corrected on Site	Number of critical violations corrected on site during inspection
Critical Violations Repeated	Number of critical violations repeated
Noncritical Violations	Number of noncritical violations found during inspection
Noncritical Violations Corrected on Site	Number of noncritical violations corrected on site during inspection
Noncritical Violations Repeated	Number of noncritical violations repeated
Latitude	Latitudinal coordinate of the establishment inspected
Longitude	Longitudinal coordinate of the establishment inspected

Table 2 provides a summary statistics of final data set used for analysis. The data has 166,581 unique inspection observations.

Table 2: Summary Statistics

Variable	N	Mean	SD	Min	Max
Inspection Id	165881	--	--	--	--
Inspection Date	165881	--	--	--	--
Establishment Type	165881	--	--	--	--
Risk Category	165820	--	--	1	5
Inspection Type	165817	--	--	--	--
Total Violations	165881	9.188	7.259	0	45
Priority Violations	73333	1.272	1.675	0	11
Priority Violations Corrected on Site	73333	0.269	0.615	0	6
Priority Violations Repeated	73333	0.046	0.284	0	5
Priority Foundation Violations	73333	2.228	2.424	0	12
Priority Foundation Violations Corrected on Site	73333	0.331	0.676	0	5
Priority Foundation Violations Repeated	73333	0.067	0.383	0	6
Core Violations	73333	3.475	4.134	0	24
Core Violations Corrected on Site	73333	0.384	0.833	0	8
Core Violations Repeated	73333	0.071	0.428	0	8
Critical Violations	92548	3.068	2.543	0	16
Critical Violations Corrected on Site	92548	1.069	1.372	0	9
Critical Violations Repeated	92548	0.072	0.400	0	7
Noncritical Violations	92548	5.615	4.988	0	30
Noncritical Violations Corrected on Site	92548	1.381	1.855	0	13
Noncritical Violations Repeated	92548	0.059	0.406	0	12

The types of inspections can be of three types:

1. Routine inspections which are defined by the risk category of a food establishment. The higher the risk category, the greater the number of routine inspections in a year.
2. Follow up inspections which follow up on any pending issues found during the routine category.
3. Other inspections which are conducted by the department and do not belong to either the routine category or follow ups.

The data provides details on each inspection type and has been cleaned to categorize each inspection as one of the three categories listed above.

Furthermore, the variable Establishment Type has also been categorized into more manageable categories. Table 3 lists down the Establishment Categories and their sub-categories.

Table 3: Categorization of Establishment Types

Establishment Category	Establishment Type
Confectionary and Catering	Bakery
	Caterer
	Ice Cream Manufacturer
Grocery and Food Products	Delicatessen
	Food Products
	Grocery Store
Restaurants and Hotels	Hotel
	Restaurant Total
Marine	Marine-Food (Wholesale)
	Marine-Food Prod (Retail)
Vending and Cafeteria	Mobile Vending
	School Cafeteria
Others	Commission Merchant
	Unknown
	Unlicensed Food

Given these establishment categories, Table 4 provides a comparison of how inspections by each category of establishments are spread across risk categories. Most of the Confectionery and Catering inspections as well as Grocery and Food Products inspections belong to risk category 2 and category 3, most of the Restaurant and Hotels inspections are category 3, while the inspections of Marin establishments are mostly category 4.

Table 4: Percentage of Inspections by Establishment Category and Risk Category

Inspections by Establishment Category	Risk Category					
	1	**2**	**3**	**4**	**5**	**Missing**
Confectionary and Catering (N=1862)	12.78%	41.73%	33.67%	11.82%	0.00%	0.00%
Grocery and Food Products (N=32485)	11.23%	42.77%	40.75%	4.72%	0.52%	0.01%
Marine (N=195)	0.00%	17.95%	13.33%	49.74%	18.97%	0.00%
Others (N=385)	70.39%	4.68%	12.21%	0.00%	9.09%	3.64%
Restaurants and Hotels (N=122589)	0.44%	17.64%	68.39%	10.20%	3.31%	0.02%
Vending and Cafeteria (N=8365)	22.50%	72.95%	4.39%	0.00%	0.00%	0.17%

Initial Predictions

With these data, I ran logistic regressions to identify the key predictors for Core violations, Priority violations, Priority Foundation violations, Critical violations, and Non-Critical violations. The significant variables from the analysis were (controlling for year fixed effects):

1. Establishment category, as explained above
2. Risk category of the inspection (Range 1 through 5)
3. Type of inspection (Routine, Follow Up, Others)
4. Season of inspection (Spring, Summer, Fall, and Winter)

Given these predictors, I used Random Forest to predict Critical violations, as a test. The results of the Random Forest are provided in Table 5.

Table 5: Random Forest Results for Critical Violations

	Precision	Recall	F1-Score	Support
Critical	0.92	0.73	0.81	24338
Others	0.25	0.67	0.36	3144
Missing	0.97	0.96	0.96	22283
Average/Total	0.9	0.83	0.85	49765

Next Steps

The initial results from the model are encouraging and the F1-Score can be improved by adding other significant variables as well as allowing the model to incorporate non-linear relationships using Random Forest. Following are the next steps involved in improving the model and accomplishing the final objectives of the project:

1) Improve prediction models by including:
    a. Time since last inspection and to next inspection (both routine and follow up)
    b. Time to and since date of license renewal
    c. Geographic locations of establishments
2) Improve prediction models by including non-linear relationships
3) Make predictions for Core, Priority, Priority Foundation, and Non-Critical violations
4) Develop an app that tracks upcoming inspections, displays information on expiring licenses, and aids in the prioritization of inspections based on predicted violations