Predicting NYC Restaurant Sanitation Grades

Cartney Thompson
Final Project Report
IST 565 - Summer 2018

September 2018

1 Introduction

Living in New York City, residents become accustomed to signs in restaurant windows highlighting letter grades. What do these letter grades mean? Since July 2010, the Health Department has required restaurants to post letter grades showing sanitary inspection results. Restaurants with a score between 0 and 13 points earn an A, those with 14 to 27 points receive a B and those with 28 or more a C. (NYC Department of Health, 2012). Choosing a restaurant based on the sanitation grade is important to many consumers (NYC Department of Health, 2012). This report analyzes import variables for determining a restaurant's health inspection grade, and predicts the grade based upon those features. Modeling restaurant grades based on inspection results helps the city allocate resources for intervention programs, food preparation education, and prevention of food borne illnesses to the general public. A restaurant's sanitation grade, according to the New York Department of Health, helps achieve three goals:

- Inform the public about a restaurant's inspection results in a simple, accessible way.
- Improve sanitary conditions and food safety practices in restaurants
- Reduce illnesses associated with dining out

Research has demonstrated that most bacterial, viral and contaminant-based food borne illnesses occur because of poor hygiene, improper storage and handling, and inadequate cooling and heating of food (NIH, 2017). The New York Department of Health requires restaurants to follow food safety rules that are grounded in science and based on federal and state guidelines and laws. Before letter grading, restaurants were motivated to practice food safety by their own desire to maintain healthful conditions and by the threat of fines for violations found at the time of inspection. Grading introduced a potentially more significant incentive: recognition with an A grade for excellent food safety practices. (New York Department of Health, 2012).

2 The Dataset

The objective of this analysis is to predict if the letter grade a restaurant receives based upon violation types from a NYC Department of Health inspection. Models will be built utilizing NYC Restaurant Inspection dataset for the years 2010 - 2017 from Kaggle's "NYC Restaurant Inspections" dataset (Kaggle, 2017). The dataset includes restaurant inspections conducted in New York City from 2010 up until August of 2017. Within the dataset, there are 18 total attributes associated with each inspection administered by the New York Department of Health (NYC Department of Health, 2012).

These attributes include; information about the restaurant (ID, Name), it's location (NYC Borough and address), the type of cuisine offered, date of inspection, type of inspection, inspection violations (descriptions, codes, and whether the code is critical), examination score, grade, and additional time stamps including date the grade was recorded.

Before processing the data, we need to get an understanding of what our data looks like. In addition to the 18 attributes, there are 399,918 rows of data. Each row contained one inspection violation. Of those roughly 399,918 rows of inspection violations, 376,707 rows contained a grade or a score (which we can interpret into a grade). 134,836 of the rows were unique inspections (where the restaurant includes a score or grade and a valid inspection date).

3 Data Cleaning

Based on the number of attributes offered in the data set, a fair amount of data cleaning is necessary. As mentioned above, the original data set offered 18 attributes. Many of these attributes were removed. This included removal of address information, phone numbers, and restaurant IDs as inclusion of their unique values would likely lead to model over-fitting. The data set did include a fair number of null values that were removed. This included data without grades and scores, inspection dates and inspection actions (due to an establishment not being expected yet).

4 Data Exploration

After cleaning our data set, we are now down to three features in addition to our predictor variable grade (see figure 1).

- Borough location of the restaurant
- Violation Code
- Restaurant Cuisine

Before preprocessing our data for modeling, we explore our data set. Below is a basic analysis of the percentage of inspections by grade:

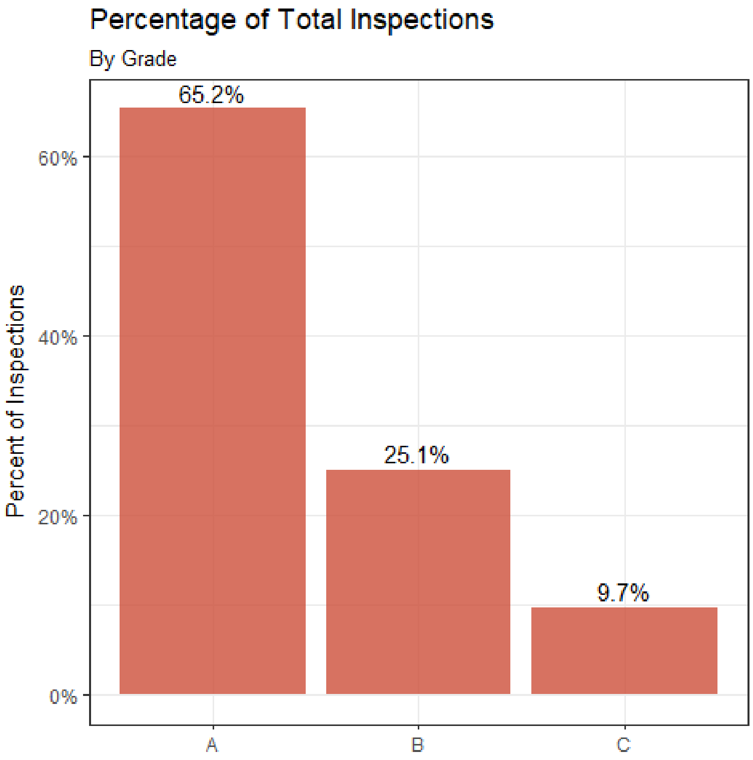


Figure 1: Percentage of Total Inspections by Grade

From the chart above, majority of restaurants receive A's or B's in during their inspections with roughly one in ten receiving the lowest grade. In addition to above, I wanted to get a high level understanding of some of our features that were part of the data set.

5 Borough

Not surprisingly, there are five boroughs in our data set: Manhattan, Brooklyn, Queens, Bronx, and Staten Island. With regards to grade distribution by borough, see 2 figure below. Staten Island has a smaller percentage of its restaurants receiving A grades, however the distribution does not vary by borough location.

6 Violation Code

There are overwhelming number of violation codes. In total, there are more than 98 violation codes in the data set. Luckily NYC Department of Health provides categories

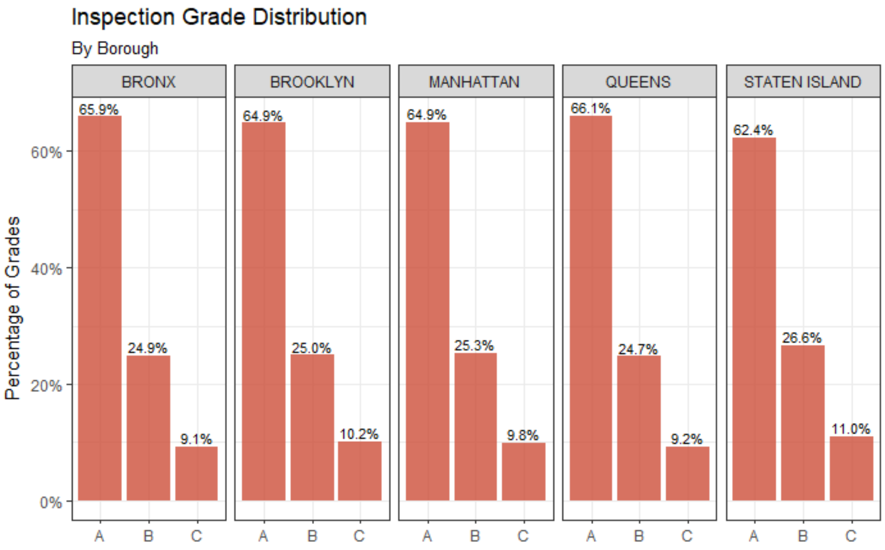


Figure 2: Inspection Grade Distribution

for each violation code By grouping the violation codes, we now have 11 unique categories. Below is a percentage breakout of total violations by category (figure 3).

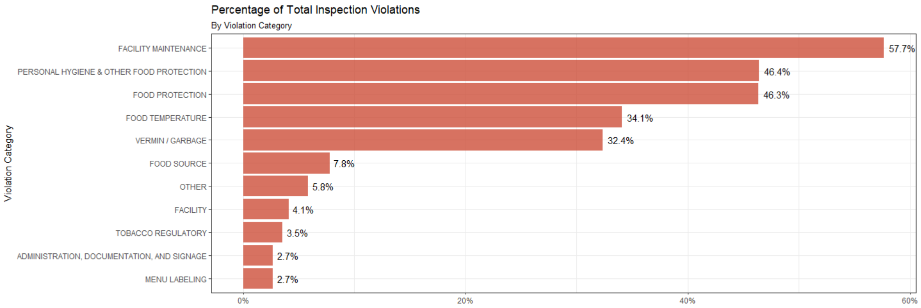


Figure 3: Percentage of Total Inspection Violations by Category

More than half of inspection violations come from facility maintenance - meaning general building cleanliness or sanitation issues. After, Facility Maintenance, Personal Hygiene, Food Protection, Food Temperature, and Vermin/Garbage (for New York City, think rats) make up the most common inspection violations.

7 Cuisine Description

Similar to violation codes, restaurant cuisines have a large number of values. Within the data set, there are more than 84 unique cuisine types. Unlike violation codes, there isn't a restaurant cuisine hierarchy. When exploring the relationship between restaurant cuisine and grade, restaurant cuisines that would have smaller menus tend to dominate the types of restaurants that get high grades. Restaurants with ethnic cuisines tend to have more restaurants in the C grade.

8 Data Preprocessing

For consistency, we want the data pr-processing to be similar across all algorithms for comparison's sake. The following steps taken involved the following:

- Re-Level Grades to three levels, "A", "B", "C"
- One Hot encoding violation codes
- Removal of attributes that had no importance to prediction. Utilizing information gain, we eliminate attributes with low values. We will eliminate Boro, Tobacco.Regulatory, Other, Food Source, Cuisine Description, and Menu Labeling from our model.

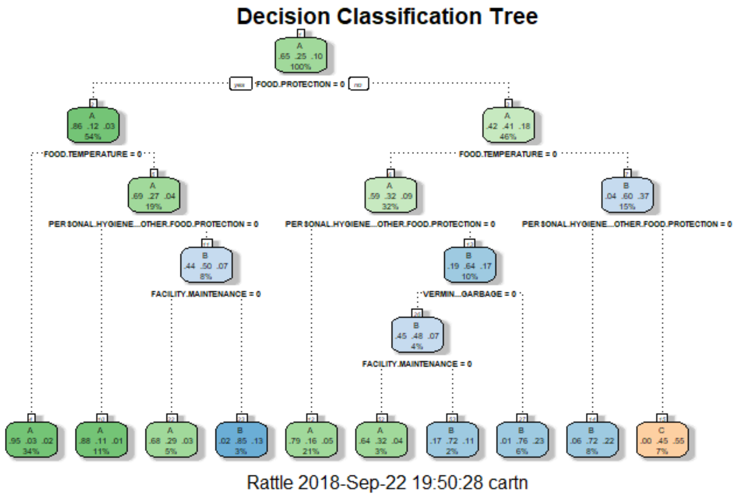
Feature	Information Gain
TOBACCO.REGULATORY	0.0001417768
BORO	0.0002522855
OTHER	0.0003572407
FOOD.SOURCE	0.0019062461
CUISINE.DESCRIPTION	0.0015675046
MENU.LABELING	0.0019215343
ADMINISTRATION..DOCUMENTATION..AND.SIGNAGE	0.0022112137
FACILITY.MAINTENANCE	0.004928505
PERSONAL.HYGIENE.OTHER.FOOD.PROTECTION	0.0282252161
FACILITY	0.0480744261
FOOD.TEMPERATURE	0.0931827704
VERMIN...GARBAGE	0.0988480326
FOOD.PROTECTION	0.1592240207

9 Experiments

For the experiments, we will utilize hold out groups to train the model, and use the test group to measure the accuracy. We will keep this same model across all experiments we run so we can compare the accuracy across all algorithms. Our default parameters used 3-fold Cross Validation across all algorithms.

9.1 Decision Tree

First we will develop a classification model via decision tree algorithm. Using the default model, and train/test validation method, we notice that the first node in our decision tree is Food Protection. Going back to our information gain table, Food Protection was the information gain was highest for this feature.


Decision Classification Tree

Rattle 2018-Sep-22 19:50:28 cartn

Figure 4: Decision Tree

When evaluating our model on the hold-out test data, the model evaluation performance was 82.0 percent accurate. On the test data, the model accuracy is roughly 82.3 percent in predicting the correct inspection grade on the test data.

9.2 Naive Bayes

Utilizing Naive Bayes classifier, accuracy on our model was utilizing the default model was low at 74.4 percent (show below). For the test data accuracy was slightly lower at 73.7 percent. Naive Bayes was by far the slowest of the three algorithms to compute.

9.3 Support Vector Machines (SVM)

Utilizing SVM classifier, again we go with the default model to compare with the other classifiers to start. The default parameters used were for linear SVM, were 3-fold cross validation. Our c value remains constant at one since all attributes have a value of 0 or 1.Our training accuracy was 80.1 percent. Predicting our test data came out to 80.4 percent.

In summary below is a table displaying the algorithm, the training accuracy, and the test accuracy performance.

Algorithm	Training Accuracy (Percent)	Test Accuracy (Percent)
Decision Tree	82.0	82.3
Naive Bayes	74.4	73.7
SVM	80.1	80.4

10 Conclusion

After this experiment, determining a restaurant's health inspection grade varies when using different algorithms. Out of all variables, Food Protection, Vermin/Garbage, and the the Food's Temperature are the strongest predictors in predicting the health inspection grade. The restaurant's location and cuisine offered are not strong predictors, although the type of cuisine (Ethnic vs. Small Item vs. Quick Serve) may play a hand and would need to be investigated further. As for which model is best at predicting a restaurant's grade based upon violations cited, Decision Trees provided the strongest accuracy and was actually the fastest to compute. This information could be extremely valuable to restaurants to help them focus on areas such as Food Protection, Vermin, and Food Temperature once inspection time comes.

References

National Institute of Diabetes and Digestive and Kindney Diseases, 2017. *Foodborne illnesses.* [online] https://www.niddk.nih.gov/health-information/digestive-diseases/foodborne-illnesses

NYC.gov, 2016. *Self-Inspection Worksheet.* [online] ://www1.nyc.gov/assets/doh/downloads/pdf/rii/self-inspection-worksheet.pdf

Kaggle, 2017. NYC Restaurant Inspections. [online] https://www.kaggle.com/new-york-city/nyc-inspections/home