Using Health Inspection Scores to Predict NYC Restaurant Yelp Ratings

Omar Smiley

Problem Statement

Every restaurant in New York City is scheduled for an unannounced inspection at least once a year. Inspectors from the NYC Department of Health and Mental Hygiene (DOHMH) conducts checks for compliance with city and state food safety regulations and marks points for any condition that violates these rules.

Can the DOHMH inspection results and demographics data be used to predict whether a restaurant has a favorable or unfavorable Yelp review?

Overview

- The Health Department inspects **approximately 27,000** restaurants in New York City to monitor their compliance with food safety regulations.
- Inspectors observe how food is prepared, served and stored and whether restaurant workers are practicing good hygiene.
- **Sanitary violations** are issued when the safety of the food being prepared and served is threatened. Sanitary violations displayed in red text are the most critical violations. Examples of sanitary violations include food being held at an unsafe temperature and evidence of mice. Sanitary violations are scored and contribute to a restaurant's grade. ***No other violations are scored or contribute to the grade.***
- Examples of non-scored violations include failing to display the Health Department-issued permit and not posting the restaurant's letter grade.

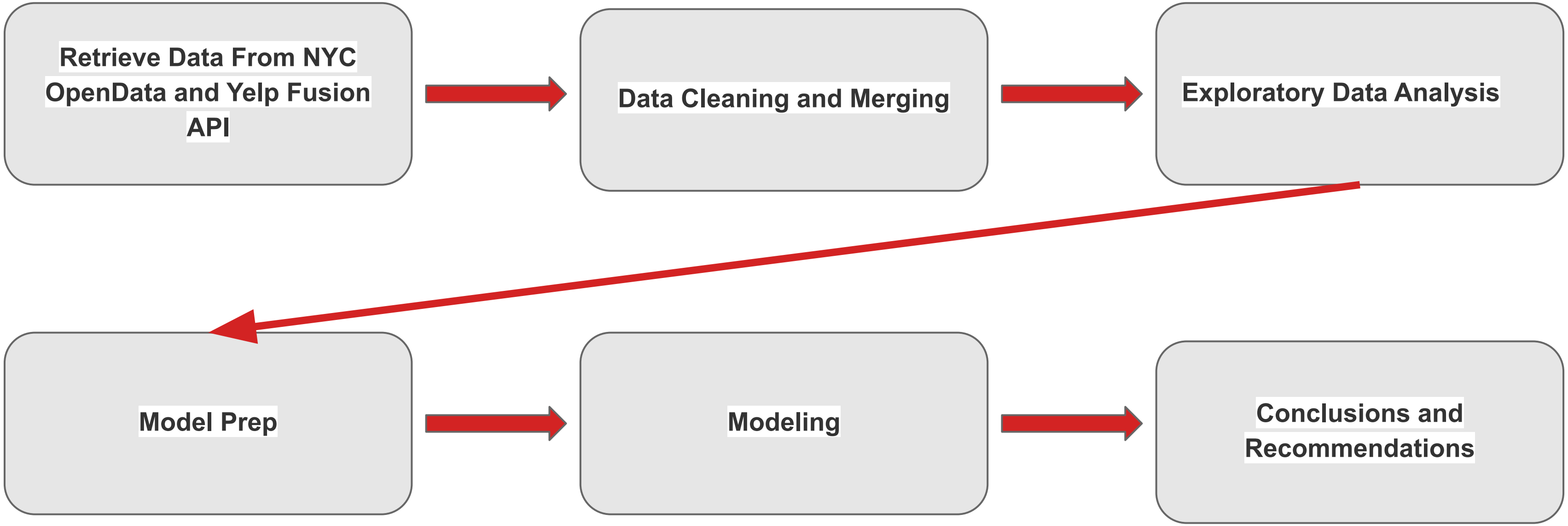
What do inspectors look for?

- Food temperatures
- Equipment maintenance
- Pest control measures

What is "Grade Pending"?

"Grade Pending" means that on both the initial inspection and reinspection, the restaurant received 14 or more points. Following the reinspection, the restaurant can post "Grade Pending" or the letter grade while they have the opportunity for an administrative hearing to determine the final grade.

Workflow and Methodology



Retrieve Data From NYC OpenData and Yelp Fusion API

Data Cleaning and Merging

Exploratory Data Analysis

Model Prep

Modeling

Conclusions and Recommendations

Data Collection and Cleaning

Data Collection

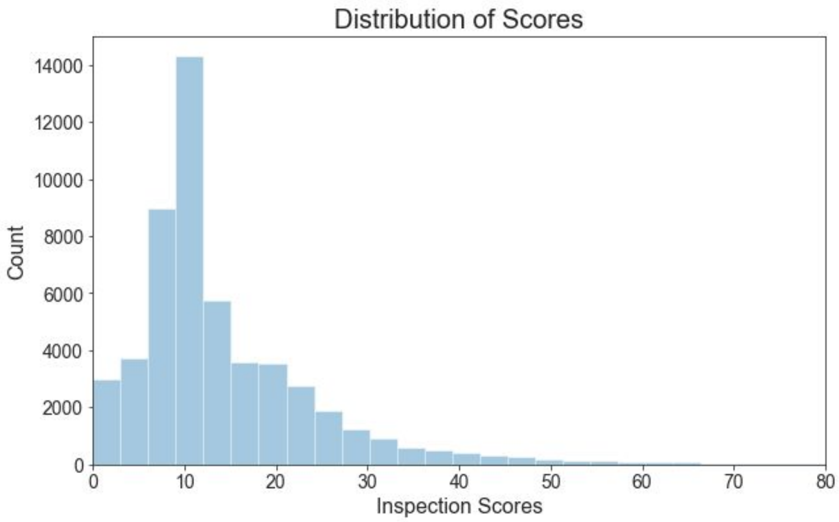
- DOHMH New York City restaurant inspection results retrieved from NYC Open Data
- Queried the Yelp Fusion API using the name and address for the restaurant provided in the NYC inspections dataset
- Utilized Levenshtein distance for name matching
- API Limits - 5000 API calls per day

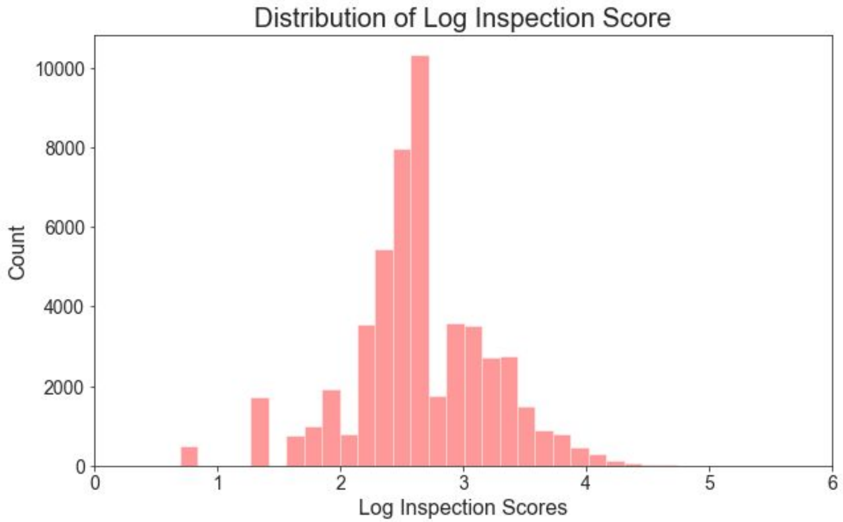
Data Cleaning/Processing

- Cuisine description - Merged similar categories (i.e., “Ice Cream”, “Gelato, Yogurt”)
- Merge demographics by zip codes
- Impute missing values (“most frequent”)
- Flag chain restaurants (more than 50 locations)

Exploratory Data Analysis

Health Inspection Scores



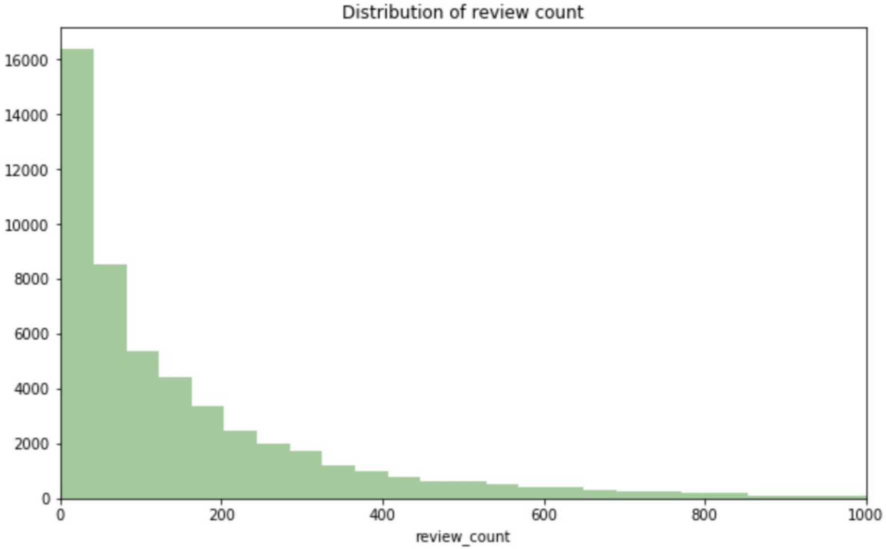


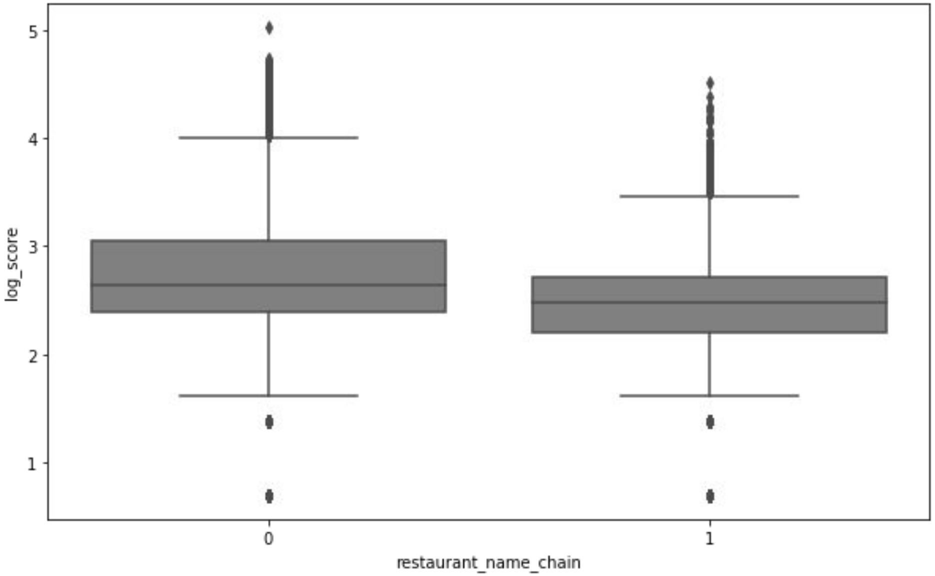
Grading Scale for Sanitary Violations

A	0 to 13 points
B	14 to 27 points
C	28 + points

Exploratory Data Analysis

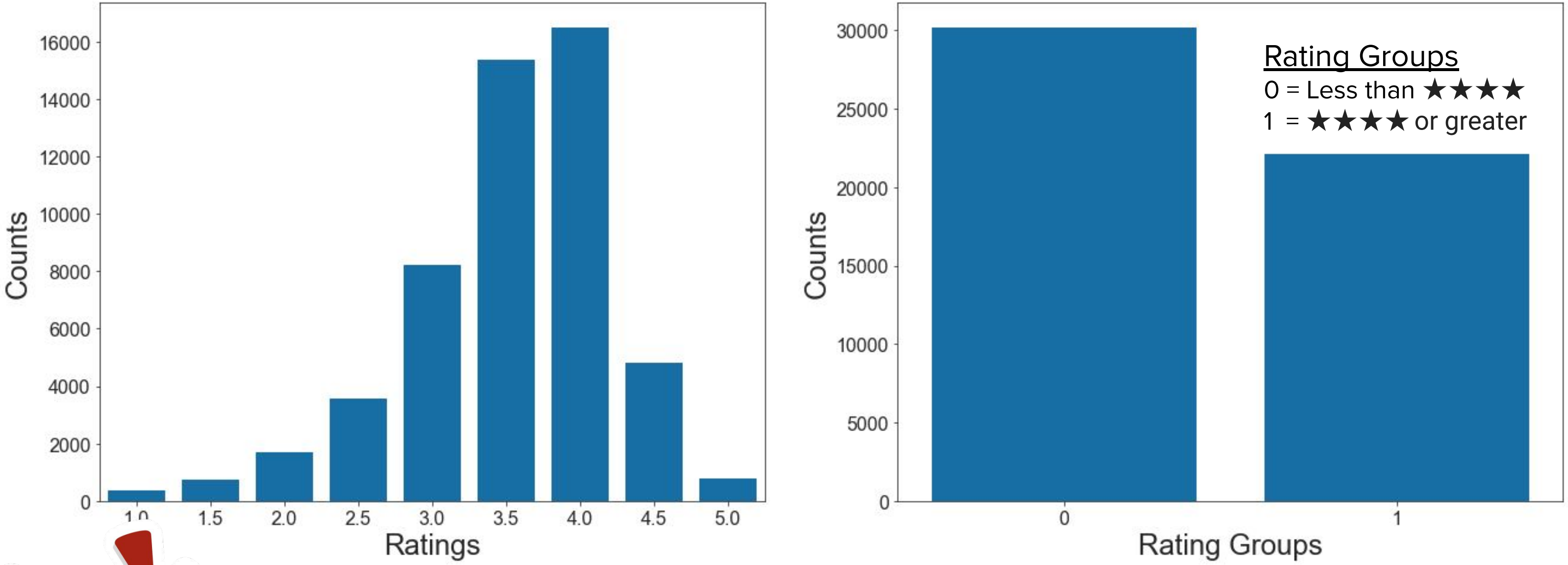
Review Counts & Chain Restaurants





Exploratory Data Analysis

Yelp Rating Distributions

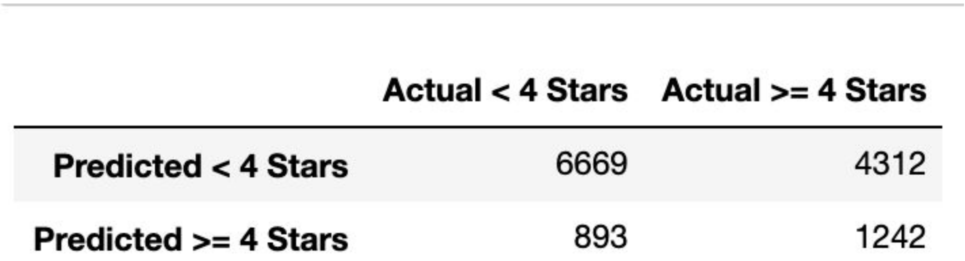


Model Evaluation - Logistic Regression

	Baseline	Logistic Regression

Accuracy	Train = 0.576543	Train = 0.6106747998475028
		Test = 0.603156450137237

Confusion Matrix



	Actual < 4 Stars	Actual >= 4 Stars
Predicted < 4 Stars	6669	4312
Predicted >= 4 Stars	893	1242

Correctly classifying:

- **1242 / 2135** Actual >= 4 Star Ratings
  True Negative Rate: 58%
- **6669 / 10981** Actual <4 Star Ratings
  True Positive Rate: 61%

Conclusions and Recommendations

Conclusions

- Logistic regression model allows us to properly predict 60% of observations as having a favorable Yelp average rating of 4 or higher.
- Using lasso regression for feature selection and interpretation:
    - Restaurants with larger number of Yelp reviews (log_review_count) less likely to have a poor rating (less than 4)
    - Chain restaurants (restaurant_name_chain) are more likely to have a poor rating
    - Wine bars, pubs and italian restaurants have relatively better ratings compared to Chinese and “Chicken Wing” restaurants
    - Restaurants more likely to have worse ratings if violation code 4D (employee not washing hands after bathroom) is noted by inspection

Recommendations

- Further analysis of additional DOHMH inspections in more years
- Look into incorporating further features to improve predictions
- NLP analysis of Yelp restaurant reviews submitted by users
- Geographic mapping of using Bokeh or similar package

Resources

- Yelp Fusion API Endpoint Documentation, https://www.yelp.com/developers/documentation/v3
- NYC OpenData, “DOHMH DOHMH New York City Restaurant Inspection Results https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j
- Python Timer Functions: Three Ways to Monitor Your Code https://realpython.com/python-timer/#creating-a-python-timer-class
- How to Use the Yelp’s Fusion API https://medium.com/@morgannegagne/how-to-use-the-yelp-fusion-api-70e62f96b0ab
-