What's Cooking?

An Exploratory analysis of New York City's restaurant inspection data

Shaival Dalal (sd3462)
Vaishali Pari (vp1096)

CONTENTS

Background 2
Data Details 2
Problem Statement 4
Analysis Approach 5
Data Exploration 6
Model Motivation 8
Model Evaluation 9
Limitations 10
Result and inferences 10
Changes from original proposal 10
Works Cited 11

TABLE OF FIGURES

Figure 1. Restaurant grades awarded by the DOHMH 4
Figure 2. Distribution of grades across the city 6
Figure 3. Confusion Matrix for Random Forest 9

TABLE OF TABLES

Table 1. Snapshot of the most hygienic restaurant for every cuisine. 7
Table 2. Snapshot of most common health code violations 8

Background

The 21st century has seen more individuals preferring to eat outside than their previous generations owing to easier access, availability of choices and an increase in disposable incomes. An individual's evaluation of the dining experience comprises personal factors such as dietary restrictions, hygiene and different food cultures. (Wijaya, King, Morrison, & Nguyen, 2013) It is important to uncover the impact of restaurant hygiene on New Yorkers as it can directly affect the health and well-being of individuals. It is also interesting to take a note of the impact food safety ratings have on the social desirability of the restaurant for it can reveal how individuals trade-off hygiene with other personally desirable factors.

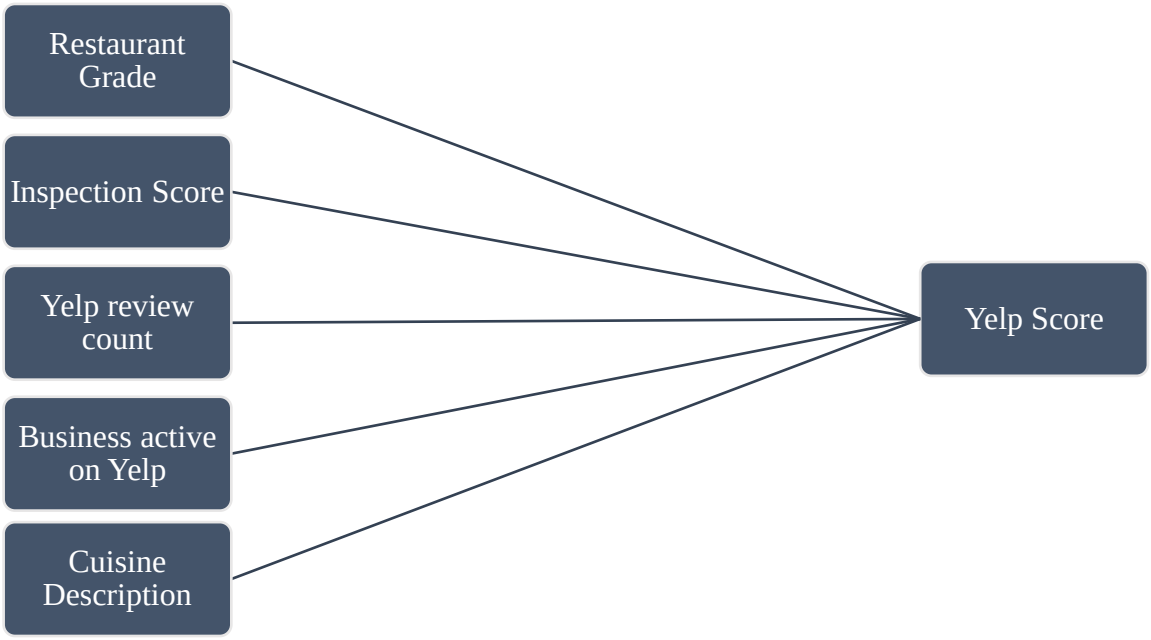
The New York City Health Department conducts unannounced inspections of restaurants at least once a year where inspectors check for compliance in food handling, food temperature, employee hygiene, and vermin control. Each violation of a regulation gets a certain number of points and at the end of the inspection, the inspector totals the points, and this number is the restaurant's inspection score—the lower the score, the better the grade.

Data Details

To conduct this study, we source data available from publicly accessible datasets provided by the City of New York and Yelp.

- NYC Restaurant Inspection data from: https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j
- Social Data from Yelp Fusion API: https://yelp.com

Attributes	Description	Data Type
CAMIS ID No	New York City-Wide Agency Management Information System (CAMIS) ID Number. This is a unique identifier for the restaurant.	Numeric
Business Name	Name of the restaurant.	Text
Borough	Borough in which the restaurant is located	Text
Building	This field represents the building number for the restaurant.	Text
Street	This field represents the street name at which the restaurant is located	Text
Zipcode	Zip code as per the address of the restaurant	Numeric
Business Phone	Business phone number	Numeric
Cuisine Description	This field describes the restaurant cuisine.	Text
Inspection Date	This field represents the date of the inspection	Text
Action	This field represents the actions that is associated with each restaurant inspection.	Text
Violation Code	This field represents each violation associated with a restaurant inspection.	Text
Violation Description	This field describes the violation codes.	Text
Critical Flag	Critical flag. Critical violations are those most likely to contribute to foodborne illness.	Text
Inspection Score	Total score for a particular inspection; updated based on adjudication results.	Numeric
Restaurant Grade	This field represents the grade associated with this inspection.	Text
Grade Date	The date when the grade was issued to the restaurant.	Text
Record Date	The date when the web extract was run to produce the data set from New York City Department of Health and Mental Hygiene.	Text
Inspection Type	A combination of the inspection program and the type of inspection performed.	Text



Predictor Variables **Target Variable**

Inspection Score is translated to Restaurant Grade based on a range specified by the Department of Health and Mental Health (DOHMH). Below is a way using which the DOHMH translates scores to grades.

Score	Grade
0-13	A
14-27	B
28 or more	C

Problem Statement

Predict Yelp Restaurant Rating by placing an emphasis on restaurant grade and inspection score among other predictors.

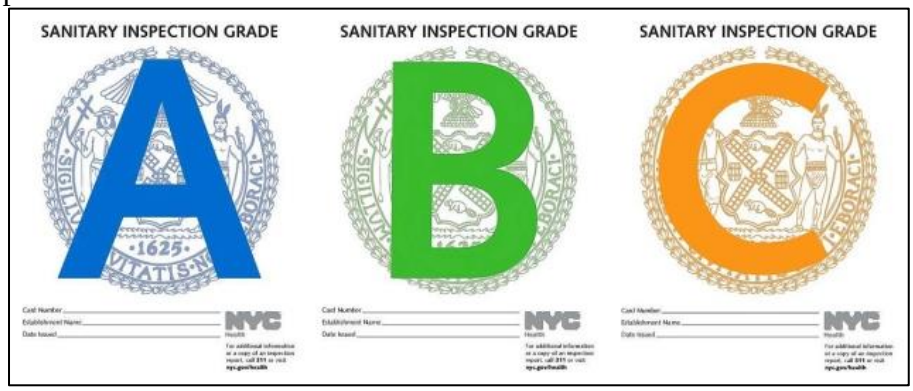


Figure 1. Restaurant grades awarded by the DOHMH

Analysis Approach

1. **Data Collection:**
    - We collect data from NYC Department of Health and Mental Hygiene related to inspections performed and violations reported
    - We collect data from Yelp using the Fusion API for restaurant price and rating
2. **Data Cleaning:**
    - We analyse the data to identify missing and malformed values and we remove them.
    - We examine the date of the records to find and eliminate records containing dates in the future
3. **Data Exploration:**
    - We plot graphs to explore the data to help us understand the data better and draw additional inferences.
    - We analyse the highest rated restaurants based on their inspection score for every cuisine
    - We find out the most common violation code by calculating frequencies for every borough and every cuisine
    - Using Yelp's APIs, we can also fetch a restaurant's social standing in the form of ratings it receives from its customers and correlate it with the inspection score to identify if any trends exist
4. **Model Building:**
    - We develop models to fit the data and predict the Yelp score by placing emphasis on the restaurant grade and inspection score.
5. **Model Evaluation:**
    - We evaluate the model on the basis of accuracy and by referring to the confusion matrix

Data Exploration

We can answer various questions by analysing the data collected from the above sources. The questions are:

1. What is the distribution of grades across the city and in all the five boroughs?

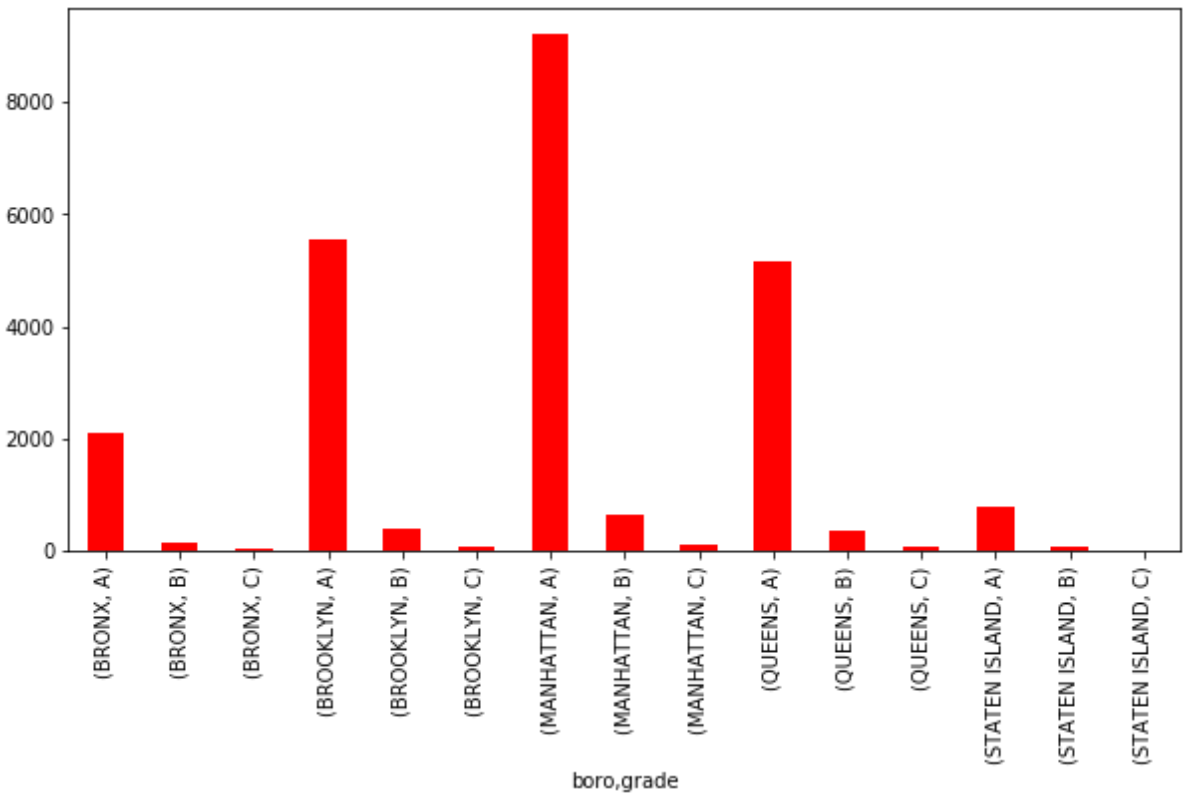


Figure 2. Distribution of grades across the city

2. Which is the most hygienic restaurant for every cuisine (bakery, Asian, Mexican etc.)

dba	cuisine_description	score	grade	
171881	AFGHAN KEBOB HOUSE	Afghan	4.0	A
335707	HONEY BEE'S KITCHEN	African	4.0	A
455	SONNY'S HEROS	American	0.0	A
361140	HOT DOG CONCESSION	Armenian	0.0	A
264011	ZHIQING ACTIVITY CENTER	Asian	0.0	A
356557	TWO HANDS TRIBECA	Australian	3.0	A
77750	EAT A BAGEL (On the Guy V. Molinari Ferry)	Bagels/Pretzels	2.0	A
64641	MONA'S BAKERY & CAFE	Bakery	0.0	A
166574	CURRY HUT	Bangladeshi	4.0	A
227314	BED STUY	Barbecue	0.0	A
388412	ORTZI RESTAURANT AND BAR LOCATED INSIDE LOUMA ...	Basque	7.0	A
136743	STAND 140	Bottled beverages, including water, sodas, jui...	0.0	A
100935	FAVELA	Brazilian	4.0	A

Table 1. Snapshot of the most hygienic restaurant for every cuisine.

3. Most common health code violation of restaurants

	Count of violations	
violation_description	violation_code	
Non-food contact surface improperly constructed. Unacceptable material used. Non-food contact surface or equipment improperly maintained and/or not properly sealed, raised, spaced or movable to allow accessibility for cleaning on all sides, above and underneath the unit.	10F	37443
Facility not vermin proof. Harborage or conditions conducive to attracting vermin to the premises and/or allowing vermin to exist.	08A	16854
Food not protected from potential source of contamination during storage, preparation, transportation, display or service.	06C	14644

Table 2. Snapshot of most common health code violations

Model Motivation

The target variable in our project is the Yelp score which is sourced by gathering the collective ratings of thousands of Yelp users. Yelp aggregates the score and round it to the nearest 0.5 points. The resultant variable, Yelp score, is a discrete set of values that represent a restaurant's rating based on Yelp users' experiences.

We choose a supervised learning approach and utilize classification based machine learning models. The following models were used and evaluated for best predictive performance:

- Logistic Regression

  It is a simple model that focuses on prediction of probabilities. It forms a sigmoid curve, with values ranging from 0 to 1. This model is easy to understand as coefficient weights are indicative of a variable's impact on the response variable

- Decision Tree

  Decision Tree works by creating partitions to separate data. Each decision tree starts by a parent node that represents a feature and the tree creates successive splits based on values of other features to obtain a purer child node.

  Decision Trees take into consideration interactions between variables and they are easy to interpret.

- Random Forest

  Random Forest is an ensemble learning method which focusses on developing multiple weak classifiers in the forms of decision trees. The algorithm chooses a random subset of features to develop a weak model and repeats the process multiple times. In the end, it predicts by adopting a voting mechanism based on every generated tree's decision. Random Forest allows us to eliminate bias that creeps in the model resulting in a powerful model

- K-Nearest Neighbours

  K-Nearest Neighbours is an algorithm that assigns a class based on proximity to records featuring similar characteristics. It follows a voting method where it takes into consideration the votes given by k-nearest neighbours. This algorithm yields good results when the number of records are much greater than the number of features which allows us to ignore the negative effects of high dimensionality i.e. curse of dimensionality.

- Support Vector Machine

  Support Vector Machine works by finding the best hyperplane that separates data as far as possible, into distinct classes. SVM can construct non-linear hyperplanes which help fit our data. However, care needs to be taken as the time complexity of SVM is quadratic in nature.

Model Evaluation

We evaluate the models based on their predictive power. We use accuracy as a metric to decide a model's power. We can also look at the confusion matrix to determine the False Positive Rate (fpr), and the True Positive Rate (tpr).

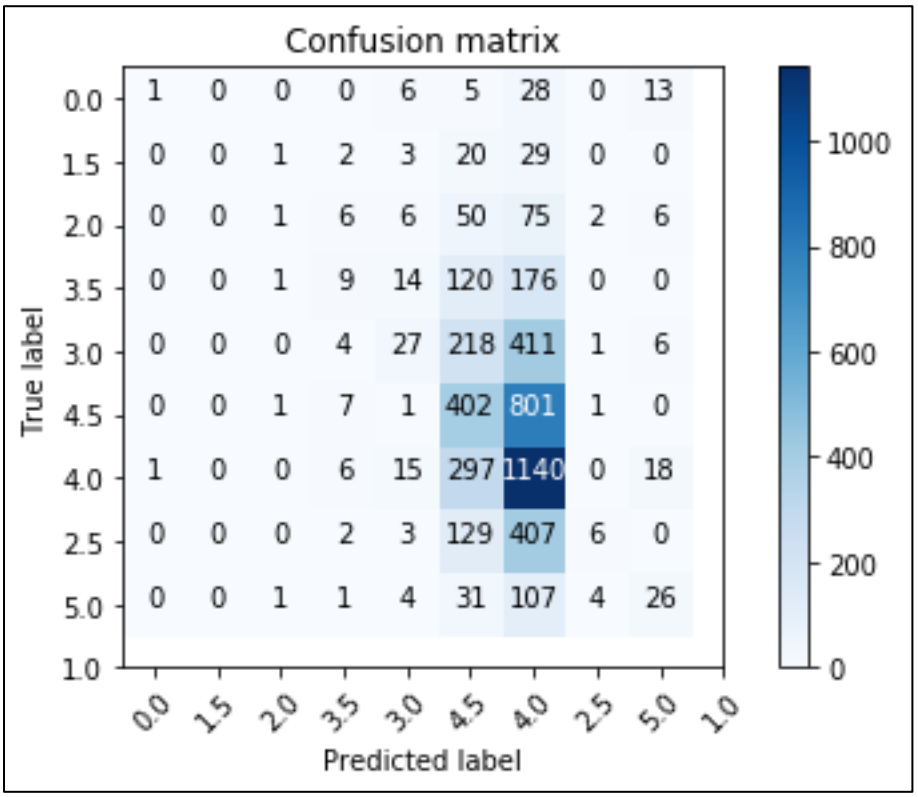


Figure 3. Confusion Matrix for Random Forest

Limitations

- To enable this study, we take into consideration the number of reviews that a restaurant has received. Merely counting the number of reviews belies the sentiment hidden within. The review may or may not be negative. Considering the sentiment of a review can have a major positive impact on the model's predictive power.
- We restrict target variable to Yelp rating only. Not all restaurants may be active on Yelp and not all customers may provide a feedback on Yelp. By combining data from multiple websites such as Facebook, Google, and UrbanSpoon, we can get a better picture of how hygiene impacts a restaurant's rating.
  We can also get to know about how customers of a particular website value a restaurant's hygiene rating.
- There is a possibility of restaurants tricking the system into receiving higher grades by placing extra emphasis on hygiene near the time of inspection. Such restaurants may add noise to our data by introducing ratings that mislead the model.
  Correct identification of such restaurants is necessary and is a limitation of our study.

Result and inferences

Based on the accuracy of all the models, we conclude that Random Forest produces the best possible result. The "No Information Rate" i.e. the percentage of value held by the majority class is 31.35%. The jump in accuracy is merely 3.3% when we use Random Forest.

By observing the Confusion Matrix and the graph, we can conclude that Random Forest is best at identifying true positives and true negatives. SVM appears to be good with classifying the majority class but fails to consider minority classes.

What this indicates about our project is that customers seldom care about the sanitary grade of the restaurant before posting their reviews and ratings on Yelp. Although hygiene should be a priority for customers, it does not appear that customers in New York city favour one restaurant over the other solely based on their sanitary grade.

Changes from original proposal

Our original project was aimed at conducting an exploratory analysis on restaurant data provided by the DOHMH. However, we were advised to draw richer inferences by incorporating predictive modelling to understand various interactions in the data.
We chose to predict the Yelp score of a restaurant based primarily on a restaurant's hygiene score awarded by the DOHMH.

Marks Distribution

Name	NetID	Marks
Shaival Dalal	sd3462	6
Vaishali Pari	vp1096	4

Works Cited

Wijaya, S., King, B., Morrison, A., & Nguyen, T.-H. (2013). International visitor dining experiences: A conceptual framework. *Journal of Hospitality and Tourism Management, 20*, 34-42.