Predicting Restaurant Health Violations Using Yelp Reviews: A Machine Learning Approach

Harlan Hutton, Surabhi Ranjan, Rajashekar Vasantha, Ted Yap
Center for Data Science, New York University
December 5, 2020

Abstract — **The New York City Department of Health and Mental Hygiene (DOHMH) conducts at least one random inspection of every NYC restaurant per year, creating potential for missed opportunities to improve the health and hygiene of establishments with food safety issues and increased redundancy of inspecting clean restaurants that are following the guidelines satisfactorily. This project aims to identify restaurants who may be in violation of health and safety code using a classification model that learns restaurant inspection data and text data from Yelp consumer reviews.**

I. Business Understanding

A. Brief Introduction to the DOHMH

The DOHMH conducts at least one inspection per NYC's roughly 27,000 restaurants every year to monitor compliance with city and state food regulations. There are about 100 health inspectors who work at the DOHMH, meaning a single inspector could conduct inspections for 270 unique restaurants a year (NYC Health).

In July of 2010, the DOHMH implemented a letter grading system in addition to the inspection score, where lower scores indicate better compliance. A grade of A represents a score of 0 to 13, B represents a score of 14 to 27, and C represents a score of 28 or more. Restaurants are required to display this grade for patrons. Getting an A means that a restaurant is less likely to be inspected in the future. Scores are calculated by violations, of which there are two kinds: critical and general. Critical violations are more likely to contribute to food-borne illness and are therefore worth more points than general violations. The most drastic action a health inspector can take after inspection is closing a restaurant. Conditions that call for a closing are 1. A public health hazard that cannot be corrected by the end of the inspection or 2. A score of 28 points or more on 3 consecutive inspections (NYC Health). The grade, score, and violations all impact the overall action that the DOHMH decides to take, ranging from the most negative extreme of closing the restaurant to the most positive extreme of no action taken.

B. Brief Introduction to Yelp

Yelp is a social networking platform started in 2004 that connects consumers with local businesses. Its most notable feature is its consumer review system through which users can leave

businesses one to five star ratings as well as text descriptions or photos of their experiences. Additionally, users can read and interact with other users' reviews. As of September 30th, 2020, the Yelp app is on 32 million unique devices worldwide. 11% of advertising revenue comes from restaurants and 18% of reviews are of restaurants. In terms of Yelp users, 30% are 18 to 34, 37% are 35-54, and 32% are older than 54. 19% did not attend college, 61% hold college degrees, and 20% hold graduate degrees. 23% make $0 to $59,000 a year, 24% make $60,000 to $99,000, and 53% make over $100,000. In terms of reviews, 51% are five stars, 18% are four stars, 8% are three stars, 7% are two stars, and 17% are one star (Yelp).

C. The Problem

Health inspectors complete 3 to 4 inspections a day on average, with a single inspection taking anywhere from an hour to several hours depending on the conditions (Krishna). Much time is spent commuting, as an inspector can be assigned to restaurants anywhere within the five boroughs. There are many problems with this current process: the random assignment of inspections is inefficient when taking into account commute and shift schedules, the conditions of a restaurant can change drastically from the time of a complaint to time of inspection, and there is a lot of tension and fear between inspectors and restaurants, stemming from the fact that a restaurant's reputation is in the hands of a single person's opinion at a single point in time. Using consumer reviews to flag restaurants potentially in violation of health code makes this process more efficient by acting as a filter for the random inspection system. The DOHMH currently uses past scores to determine likelihood of next inspection, so using machine learning to predict likelihood of a future violation can further refine the pool of potential restaurants to be inspected. Further, involving consumers in the review process of a restaurant can add accountability.

D. Literature Review

The DOHMH has done previous work with Columbia University to identify unreported cases of foodborne illnesses using Yelp reviews. Data mining software was created to flag reviews containing potential descriptors of illness, which were then manually reviewed by epidemiologists who decided which ones to send to the DOHMH for further review. After inspections of the restaurants and interviews with the Yelp users who posted the original reviews, the DOHMH identified three outbreaks that accounted for 16 unreported illnesses. The project provided evidence that "by incorporating website review data into public health surveillance programs, health departments might find additional illnesses and improve detection of foodborne disease outbreaks in the community. Similar programs could be developed to identify other public health hazards that reviewers might describe, such as vermin in food

establishments" (Harrison). Previous research also shows that publishing inspection scores in the media provides information to customers and also influences them, and those customers in turn influence restaurant management decision making. In their study "The Impact of Publishing Foodservice Inspection Scores," Almanza, Ismail, and Mills found that the sharing of inspection scores in media caused future inspection scores to increase and consumer complaints to decrease (Almanza). Finally, research has shown that millennials use social media for dining information seeking and sharing more than any generations before them (Newkirk).

II. Data Understanding

A. Data Collection & Extracted Features

The DOHMH New York City Restaurant Results dataset obtained from NYC OpenData contains inspection results for currently open restaurants up to three years from the most recent inspection and has 400,000 rows and 26 columns. Every row represents a violation citation from an inspection, so more than one violation in a given inspection results in additional records with repeated inspection description values. The 26 columns contain information about the:

- Location of the restaurant: address, building, borough, GPS coordinates, phone number, community board, council district, census tract
- Inspection itself: date, score, grade, grade date, record date, type of inspection, action taken, violation code, violation description, critical flag
- Type of restaurant: name, individual CAMIS identification number, cuisine type, BIN number, BBL number, NTA code

The Yelp reviews dataset was obtained using a scraper built by us that pulled 100,000 reviews for 2000 unique restaurants. The data scraped from the Yelp website includes the user review, the date of the review and the rating given by the user.

B. Potential Biases in the Data

The nature of a Yelp review introduces voluntary response bias: people are more likely to leave Yelp reviews when they have strong feelings about a restaurant experience, which is evident in Yelp's report that 51% of reviews are five stars. Additionally, the users who posted the reviews in this dataset may not be representative of the general population, as the majority of Yelp users are under 55 years old and college educated with salaries of over $100,000.

Yelp has also struggled with "astroturfing" in the past, a term that has come to represent the process of companies purchasing fake reviews to boost their ratings on review sites. In 2013, a study conducted by Harvard found that about 20% of Yelp reviews were fake. Following the release of this study, The New York Attorney General completed a sting operation that found 19 companies guilty of commissioning fake Yelp reviews from freelancers at a rate of $1 - $10 per review (D'Onfro). Additionally, Yelp now offers advertising packages for businesses wanting to increase their traffic on Yelp ranging from $400 to $2,250 a month (Leffler). It is important to note that the reviews in the Yelp dataset may be sponsored or fake.

The DOHMH dataset introduces survivorship bias in that it does not include inspection results for restaurants that have closed permanently, so the model learns the data of restaurants that succeeded rather than failed.

C. Data Leakage

Because each row in the DOHMH dataset represents a violation rather than an inspection, there are multiple rows for a single inspection. This presents potential for data leakage when splitting the dataset into training and test sets, as the violations from a single inspection can end up in both sets. There is also the possibility of a discrepancy between scores and grades, as restaurants can dispute their initial inspection scores through an adjudication process which can take months. This means a score given today may be revised in subsequent weeks or months.

III. Data Preparation

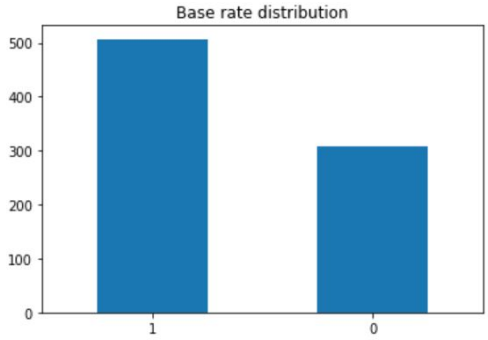
A. Data Cleanup

To clean the DOHMH dataset, we removed rows that had inspection dates with incorrect date formats, added the inspection date as a column, removed rows without a borough, split inspection types, created a single row for restaurant and inspection date, and changed the "action taken" column to a binary variable. We wanted to limit our analysis to routine inspections noted as "cycle inspections" in the dataset to prevent any data leakage, so we filtered out pre-permit inspections and re-inspections. The action column, which contains information about what action the DOHMH decided to take based on the number and type of inspection violations, originally contained 5 options:

1. No violations were recorded at the time of this inspection - Restaurant is up to standard and can continue operating.

2. Establishment Closed by DOHMH - Restaurant violates hygiene standards and has been instructed to shut down.
3. Establishment re-opened by DOHMH - Restaurant opened on re-inspection after being shut down for violations.
4. Establishment re-closed by DOHMH - Restaurant shut down again after a re-inspection.
5. Violations were cited in the following area(s) - No specific action taken by DOHMH.

We decided to build a binary classifier with a very clear distinction between restaurants that will be shut down based on a health inspection and restaurants who maintain high hygiene standards. To do this, we consider the "action" column as our target variable, and the actions "No violations were recorded at the time of this inspection" and "Establishment Closed by DOHMH" as our target outputs which are encoded as 0 and 1 respectively. The graph below demonstrates the distribution of our target variable (post data cleaning that has been elaborated upon further) and we see that the distribution is 60-40 which indicates that our data is not significantly imbalanced.



Next, we transformed the non-numerical features to numerical (see section IV.C). With every inspection of a restaurant treated independently, the final DOHMH dataset contains 2,161 violations of 1,826 unique restaurants, with the number of violations per restaurant ranging from 1 to 25.

To clean the Yelp reviews dataset, we first converted the text to lowercase and removed punctuation, removed stop words, dropped rows with null values, and added features for the sentiment of the review with or without stop words. We discovered that removing stop words decreases the range of the sentiment of the reviews and decided to not remove stop words in order to encourage more diversity, or greater difference of values between positive and negative sentiments. To combine datasets, we found all reviews from the past 6 months leading up to that specific inspection date for each restaurant using its unique business ID (see section III.B). We then pulled the 10 most recent reviews, combined them into one, calculated their average user-rating and sentiment score, and merged them with the DOHMH

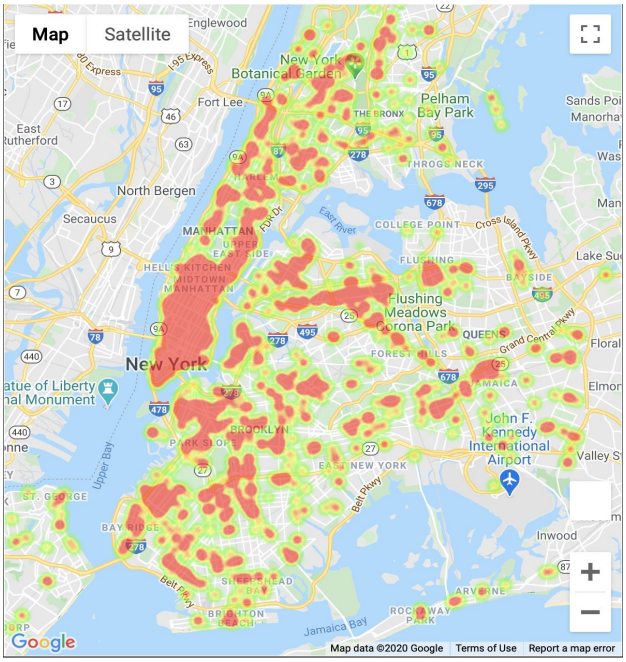
dataset; creating a final combined dataset of 816 rows with each row corresponding to a unique restaurant+inspection date identifier (see section III.D).

B. Record Linkage

The DOHMH and the Yelp dataset could not be linked directly as they were obtained from different sources. This meant that the naming convention for the restaurant names and addresses were slightly different for each of the above mentioned datasets. To overcome this, we used the *Business Search API* provided by Yelp to determine the restaurants in the Yelp dataset that correspond to the restaurants in the DOHMH dataset (documentation).

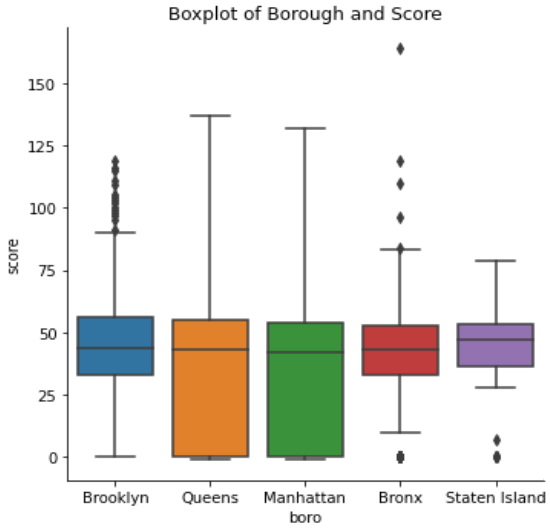
C. Exploratory Data Analysis

Some interesting findings during our exploratory data analysis are in the below graphs. In order to look at the distributions of the features, we used score as a proxy for action taken, as action taken is directly dependent on the score assigned to a restaurant. We looked at a heatmap using the Python library "gmaps" and the GPS coordinates of each restaurant with inspection scores as weights. Red indicates a higher score with the maximum being a score of 164 and green indicates a lower score with the minimum being 0. A higher score assigned to a restaurant indicates more violations in both severity and number. This provides evidence that borough alone is not indicative of inspection outcome.

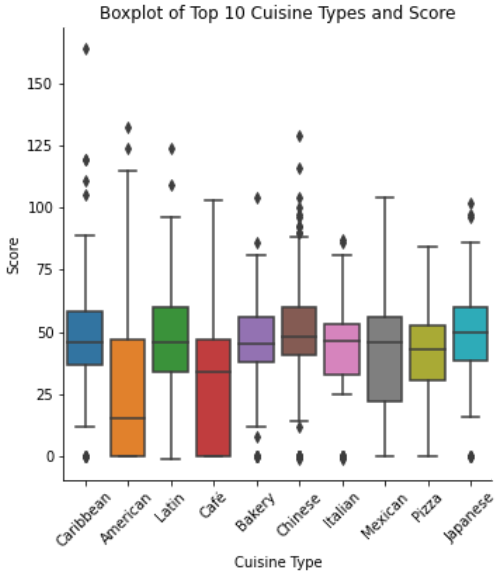


Next, we looked at the distributions of each feature. For the borough feature, Manhattan was the most prevalent with 804 restaurants, while Staten Island was the least with 43 restaurants. This supports our findings in the heatmap that the distribution of inspection scores does not vary significantly across

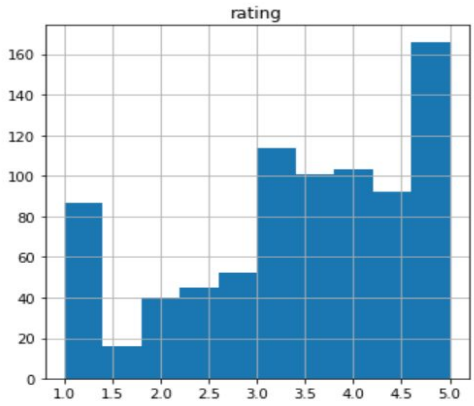
boroughs, although there is some skewness in the scores of Queens and Manhattan. This graph was an early indicator to us that borough by itself may not be an adequate predictor of inspection outcome.



For cuisine description there are 69 unique categories ranging from “American”, “Portugese,” to “Pancakes/Waffles.” Looking at the boxplot of the ten most prevalent cuisines, we see that the distribution of scores does differ, indicating that cuisine category could be a good predictor of inspection outcome.



We also analyzed the distribution of Yelp ratings for these restaurants and we see that there is a higher number of extreme reviews supporting our voluntary response bias hypothesis that people are more likely to leave Yelp reviews when they have strong feelings about a restaurant experience.



D. Feature Engineering

Once the reviews scraped from the Yelp website were linked to the DOHMH dataset, we had to decide what transformations to perform on the reviews before they were fed into the model. One of the decisions we had to make was to choose the number of reviews prior to the inspection date to be included as a feature. We tried 10 and 50 reviews and found that although using 50 reviews gave us a better performance on the train set, it also caused the feature space to blow up when using sentiment scores per review as features and reviews as tokenized words. Due to this, we use the most recent 10 reviews.

Next, we used TextBlob to generate sentiment scores for each review (see documentation). Our first approach was to use the average sentiment scores of the latest 10 reviews as a feature for prediction (*Feature Group 1*). However, in order to see if the reviews individually contribute to the inspection result, we treat each review's sentiment score as a feature (*Feature Group 2*).

To see if the words present in the reviews had prediction power, we combined the latest 10 reviews into a single string, tokenized it using TF-IDF tokenizer (n_gram = 2) (*Feature Group 3*). We also retained the average sentiment score and ratings of the latest 10 reviews as a feature. Since this resulted in a very large number of features (142,315 total feature columns after tokenizing the review columns and one-hot-encoding the other categorical features), we applied Principal Component Analysis (PCA) to reduce the dimensionality to 10 dimensions.

IV. Modeling & Evaluation

A. Evaluation Metric

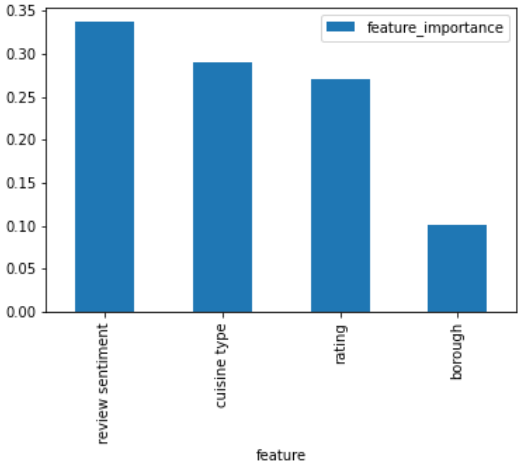
In the context of inspections, a false positive means our model incorrectly identifies a restaurant as likely to fail a health inspection, when in actuality it may pass. In contrast, a false negative means we fail to identify a restaurant that will actually fail the inspection. Such a restaurant will continue to operate,

putting the health of individuals who dine there at risk. False positives are more expensive for the DOHMH because an inspection costs time and money. However, false negatives are more expensive for the general public in terms of potential health threats. Since the F1 score takes the harmonic mean of precision and recall, we can take into account both errors and assign recall a higher weight using the weighted beta F1 score (with beta=1.5). In addition, we used AUC score as a simple measure of performance to efficiently compare baseline models.

B. Feature Selection

Our goal is to identify whether or not the use of Yelp reviews helps to improve the performance of our model. To start off, the baseline model uses only the DOHMH dataset without the Yelp reviews, as this more closely mimics the current assignment of inspections. To improve upon the baseline model, we use three different methods to merge Yelp reviews with DOHMH dataset. The features used in these three methods are outlined below. We hypothesized that users who visit restaurants are mindful of where they dine and if they find something amiss, they are likely to post on Yelp about it. To test this hypothesis, we checked if the ratings users gave and their review sentiments (*see Appendix VII*) were correlated to the result of DOHMH inspections. As mentioned in section III.D, we looked at both average review sentiment and 10 most recent review sentiments.

Baseline	Cuisine Type, Borough
Feature Group 1	Cuisine Type, Borough, Average Rating, Average Review Sentiment
Feature Group 2	Cuisine Type, Borough, Average Rating, 10 Most Recent Review Sentiment
Feature Group 3	Cuisine Type, Borough, Average Rating, 10 Most Recent Reviews (Tokenized Words) with PCA



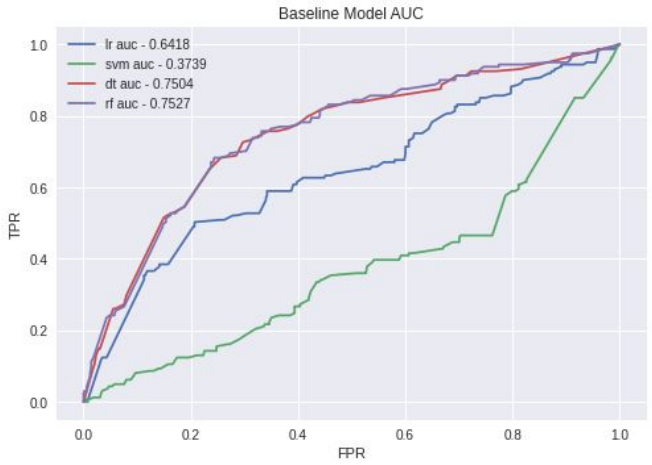
Looking at feature importance (using random forest) for Feature Group 1, we found that "review sentiments" was the most important feature in predicting restaurant violations.

C. Baseline Modeling and Results

Our baseline dataset includes two features, borough and cuisine description, with the binary target variable "action." We wanted to use both linear and non-linear models, so we trained logistic regression, support vector machine (kernel = linear), decision tree, and random forest models with default

parameters. For our baseline and feature group 1 models, we have label-encoded our features as we wanted to rank feature importance and also not increase the dimensionality of our dataset.

Our dataset is small and logistic regression is more suited to such datasets. It also has higher interpretability, is a simpler linear model than SVM, and obtains a higher AUC (specific to our baseline model). The random forest and decision tree classifiers help us identify any non-linear relationships missed by the logistic regression model. The other side of the story is F1 scores, which lead us to believe that the logistic regression and SVM models are not learning from the data and are predicting one label (precision = recall = 0) whereas the decision tree and random forest models are overfitting on the training data (graph included below).



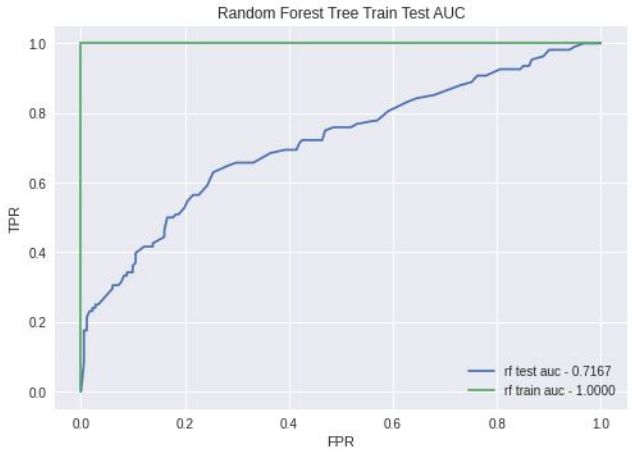
Baseline Model	Weighted F1 Score
Logistic Regression	0.0
Support Vector Machine	0.0
Decision Tree	0.444
Random Forest	0.456

D. Model Selection Approach

To improve the baseline model, we followed two approaches -

1. Use different feature groups as outlined in section IV.B.
2. Use different models and perform hyperparameter tuning with cross validation to select the best performing model.

The hyperparameter tuning range was chosen based on bias-variance analysis. For example, the below plot shows the train and test AUC scores (for feature group 1) when default parameters of the random forest model were used, clearly indicating overfitting. This analysis was performed for each model and the hyperparameter space was designed based on the results of the analysis.



The best hyperparameters were chosen using 5-fold cross validation.

The table below shows the average weighted F1 score evaluated with 5-fold cross validation for different models.

Average Weighted F1 Scores on Cross Validation Set						
	Logistic Regression	Support Vector Machine	Decision Trees	Random Forest	Gradient Boosting	XGBoost
Feature Group 1	0.3250	0.5232	0.5513	0.4598	0.5614	**0.6008**
Feature Group 2	0.6595	0.6627	0.6940	0.5934	0.7012	**0.7200**
Feature Group 3	0.8014	0.8111	0.7325	0.8301	0.8380	**0.8417**

In addition to the models described earlier, we looked at boosting-based ensemble models. We chose to use Gradient Boosting Classifier and XGBoost for this. We chose XGBoost as it allows for parallelization and also has a better tree pruning algorithm compared to vanilla gradient boosting (Chen). Since XGBoost outperformed all other models across different feature groups, we then performed hyperparameter tuning to further improve the performance of the model. We performed grid search over a range of values for max depth, number of estimators, and learning rate, since these hyperparameters are often the most sensitive to the performance of the model. The table below shows the weighted F1 score corresponding to different hyperparameters of the XGBoost model.

Max Depth	Number of Estimators	Learning Rate	Weighted F1 Score
5	80	0.1	**0.8556**
10	100	0.01	0.7966
3	70	0.001	0.7515

Additionally, we compared performance after oversampling the minority class and found that it resulted in lower weighted F1 scores.

V. Model Deployment

A. From Results to Deployment

These results can be deployed by the DOHMH as long as there is continuous access to Yelp reviews as they appear. We can build an automated system on a cloud platform where the model is re-trained and evaluated whenever new data is available. Using social media to help find restaurants in violation of hygiene standards can be an important tool to ensure efficiency and accountability, as restaurants can be flagged closer to real time rather than at random or based on previous inspections. When successful, this model disrupts the traditional inspection timeline for the better. An aspect that may be tricky to navigate in deployment is the criteria set for reviews, which is an "art" that comes into play in most text classification algorithms. We discuss mitigation of this in section V.C.

B. An Ethical Aside

An interesting point to make about using customer reviews to help flag restaurants potentially in violation of health and safety standards is that allowing a customer to give his or her input into what is right or wrong, healthy or unhealthy, safe or unsafe means that we also allow a customer to apply personal norms and social code to a government mandated process. These norms and social codes can vary across gender, race, ethnicity, age, or region. In their article "Conflicting Social Codes and Organizations: Hygiene and Authenticity in Consumer Evaluations of Restaurants," Lehman, Kovács, and Carroll tell a story about a Chinese restaurant that was found in violation of health code for hanging ducks at room temperature for an extended period of time. There was backlash from the restaurant and its community because this style of preparation is a Chinese tradition of over 4,000 years, and many thought the health department was ignorant for their decision. This story points out the fact that inspectors themselves apply their own biases to the inspection process, so allowing customers to also have input into a potential health inspection may not only result in more human error, but even highlight issues

Americans have in understanding cultures different than their own (Lehman).

C. Risks and Risk Mitigation

Based on previous work done in combination with our own, we believe this is a good tool for this problem but would need manual review as well, like the DOHMH and Columbia University did with epidemiologists who reviewed flagged restaurant reviews.

VI. Conclusion

A. Limitations

Both the data obtained from the DOHMH and Yelp came with their own limitations and biases outlined in section II.B, such as the selection bias of Yelp reviews and survivorship bias of inspection data. However, further issues we encountered throughout the development of the model forced us to make limiting decisions like focusing our analysis to only the most extreme ends of action taken by the DOHMH and routine inspections rather than pre-permit or re-inspections.

B. Future Work & Final Remarks

The current work included only a subset of restaurant reviews. Future work could include the complete dataset. As a better alternative to TF-IDF embeddings, we can use pre-trained word embeddings like BERT or GloVe(GloVe documentation, BERT documentation). Using word embeddings from one of these pretrained models, a deep learning based approach can be adopted to derive contextual information from the reviews. Future work could include expanding data collection to not only other sites used for reviews, like Zagat or Google Reviews, but also to other social networking sites like Twitter and Facebook. Another aspect to consider in future work would be extending this to a multi-class classification problem by including the other levels of action taken by the DOHMH in our target variable.

VI. Bibliography

1. NYC Health. “Food Establishment Inspections.” *NYC Health*, https://www1.nyc.gov/site/doh/services/restaurant-grades.page.
2. NYC Health. “What to Expect When You're Inspected: A Guide for Food Service Operators.” *NYC Health*, June 2016, https://www1.nyc.gov/assets/doh/downloads/pdf/rii/blue-book.pdf.
3. Yelp. “An Introduction to Yelp Metrics as of September 30, 2020.” *Yelp Newsroom*, 30 September 2020, https://www.yelp-press.com/company/fast-facts/default.aspx.
4. Krishna, Priya. “The Life of a Restaurant Inspector: Rising Grades, Fainting Owners.” *The New York Times*, 5 June 2018,

https://www.nytimes.com/2018/06/05/dining/restaurant-health-inspector.html?action=click&module=RelatedLinks&pgtype=Article.

5. Harrison, Cassandra, et al. "Using Online Reviews by Restaurant Patrons to Identify Unreported Cases of Foodborne Illness — New York City, 2012–2013." *Morbidity and Mortality Weekly Report*, vol. 63, no. 20, 2014, pp. 441–445. *JSTOR*, www.jstor.org/stable/24855169. Accessed 4 Dec. 2020.
6. Almanza, Barbara A. PhD, RD, Ismail, Joseph PhD & Mills, Juline E. PhD (2002). "The Impact of Publishing Foodservice Inspection Scores." *Journal of Foodservice Business Research*, 5:4, 45-62, https://www-tandfonline-com.proxy.library.nyu.edu/doi/abs/10.1300/J369v05n04_04
7. Newkirk, Ryan W et al. "The potential capability of social media as a component of food safety and food terrorism surveillance systems." *Foodborne pathogens and disease,* vol. 9,2 (2012): 120-4. https://pubmed.ncbi.nlm.nih.gov/22217109/
8. D'Onfro, Jillian. "A Whopping 20% Of Yelp Reviews Are Fake." *Business Insider*, 2013, https://www.businessinsider.com/20-percent-of-yelp-reviews-fake-2013-9.
9. Leffler, Nick. "Yelp Advertising Review: Was My 3-Month Contract Worth It?" *Exprance*, 2018, https://www.exprance.com/yelp-advertising-review/.
10. Lehman, David W., et al. "Conflicting Social Codes and Organizations: Hygiene and Authenticity in Consumer Evaluations of Restaurants." *Management Science*, vol. 60, no. 10, 2014, pp. 2602–2617., www.jstor.org/stable/24550932.

Code can be found in GitHub: https://github.com/rajashekarv95/Yelpies

VII. Appendix

A. Reviews that suggest possible health violations

- dirty the food behind the counter looked old and stale expensive should have checked yelp before i ordered
- restaurant got an a but they are very dirty i usually order then pick up but today i arrived a bit early while they were still preparing my food there were no usage of gloves and i saw the guy in the apron preparing my food just picked his nose then continue touching my food i'm now wondering how they prepaid their salad that is already there that they put in everyone's food i feel so sick

- good quantity of food for the price flavor is ok slightly bland and underwhelming indian food inside is a bit dirty with some flies food was really just average and i'm not sure i'd go back
- employee do not wash hands handle money and food with same gloves they make sandwich with do not change gloves.

B. Contributions

Harlan Hutton - EDA, Over/Undersampling, Business Understanding & Research, Writeup & Formatting

Surabhi Ranjan - EDA, Data Cleaning, Feature Encoding, PCA

Rajashekar Vasantha - Problem Formulation, Yelp Reviews Scraping, Feature Engineering

Ted Yap - Yelp Reviews Scraping, Model Selection, Problem Formulation