Restaurant Inspection Forecasting

Messy Data & Machine Learning

December 13, 2019

Shannon Kay, Jaejin Kim, & Jessica Spencer

I. Introduction

In an ideal world, restaurant inspections would be very straightforward evaluations. Is your restaurant clean? Is the food prepared up to code? If the answer is yes, the restaurant should pass. However, understanding what criteria restaurants need to meet to ensure that they pass inspections is not as simple in practice. We would hope that inspections depend upon adhering to the jurisdiction's health code in a manner that prioritizes consumer health and safety. However, health inspectors are human, which means that bias may occur with respect to restaurants of certain types, neighborhood, etc. A well performing predictive model can be very helpful not only for the restaurant owners but also for inspectors by optimizing order of inspections to deliver the result faster. It can also be utilized by the officials in the health department of the city government, by identifying features other than hygiene that affects the result of an inspection.

Additionally, individual cities strive to standardize the ways they report information, which makes data analysis within cities possible more than ever before. As open data evolves from novelty into common practice, we wonder whether available city-level data will allow us to make comparisons across cities in areas where records were previously not considered comparable. We believe that as cities standardize their data, it will become easier to identify common features and thus analyze the differences between cities.

We decided to test this theory using food inspection data in Chicago and New York. Our research question is: "Is there a difference in the pass/fail rate for restaurant inspections between Chicago and NYC when controlling for characteristics intuitively predictive of whether a restaurant passes or fails?"

II. Literature Review

In New York City, random food inspections are done at least once a year. A good grade on an inspection is an A, or a score between 0-14, while a C, a score greater than 28, documents that the restaurant had more substantive health and safety violations. These grade cards must be posted where people can see them when entering the restaurant. If a restaurant is given a bad grade on its first inspection, the grade is not recorded, and the restaurant is given another opportunity to fix any violations. If a restaurant receives a C, it is inspected more often going forward. A restaurant is closed down at the inspectors' discretion when there have been too many health and safety violations for it to remain open to the public. For example, we found two restaurants in the dataset that were shut down with a score of 54, and 61. A critical violation is defined as any violation that poses a significant risk to consumer health.

In Chicago, a random food inspection is performed at least once a year by the Department of Public Health's Food Protection Program. Restaurant inspections in Chicago are informed by categories. Category I facilities have a higher risk of leading to food-borne illnesses and thus receive more inspections. Category II has only a median risk, and Category III has low risk. Restaurants with higher potential risk are subject to more frequent inspections, and when violations that may pose a significant risk to consumer health are found, the restaurant may be closed down at the inspectors' discretion either temporarily or permanently.

III. Data Sources

1. Chicago Food Inspections Data
2. NYC Restaurant Inspections Data
3. IRS Individual Income Tax ZIP Code Data

IV. Data Processing:

Selecting Data

Our driving research question was how similar the pass/fail rates for restaurant inspections in Chicago and New York City are, and whether these rates vary based on the characteristics of the neighborhood the restaurant is located in. While NYC Open Data has a dataset of Demographic Statistics by Zip Code through the Department of Youth and Community Development that contained information about gender, ethnicity, citizenship and public assistance percentages by zip code, we were unable to find an equivalent dataset for Chicago. Our search for some indicators that might inform us about the locale of the neighborhood led us to the IRS Individual Income Tax ZIP Code Data. This dataset provides tax return information by zip code, and was the only data source we could find that would link to both the Chicago and NYC datasets uniformly. Since we believed neighborhood socioeconomic status would be an important predictor of restaurant inspection pass/fail rates, we chose to integrate this data into our analysis.

Cleaning

Creating a dataset that combines Chicago and New York City restaurant inspections with the IRS Individual Tax Zip Code Data involved multiple layers of data cleaning. We first chose which years to analyze. Chicago's dataset contains inspections from 2010-2019, while the New York City data has inspections from 2013-2019. Since the data in NYC 2013 and 2014 is sparse, 2015-2019 provided the best available inspection data for both cities. However, the IRS Individual Tax Zip Code Data is only available through 2017, so we chose to use data from 2015-2017.

The Chicago dataset contained entities beyond restaurants, such as wholesale retailers, grocery stores, daycares, schools, and long-term care facilities. We chose to retain any inspections where the facility type included bakery, cafe, restaurant, tavern, deli, ice cream, or paleteria. We removed inspections with results recorded as "not ready", which were generally restaurants that were preparing to open, and "out of business," which meant that restaurant had closed. This produced a dataset of restaurants that were currently serving consumers and that had either passed or failed their inspections. We also limited the inspection type to canvass (the routine yearly inspections in Chicago), complaint, license, and suspected food poisoning.

Then we cleaned the NYC data to ensure that we were looking at similar inspections in both cities. We determined that cycle-compliance, cycle-initial, cycle-re-inspection, and cycle-second-compliance were the most comparable to the Chicago food inspections. We excluded pre-permitting inspections, as we removed "not ready" results from the Chicago data, and eliminated administrative or particularly specific inspections that did not clearly map across datasets, such as those pertaining to the Calorie Posting, Trans-Fat, or the Smoke-Free Air Act. We also chose to filter out inspections with scores that were not available or that were less than 0, because there should not be negative scores, and the score variable determines whether a restaurant passes or fails in NYC.

Feature Engineering

For some variables, standardizing the variable across both cities was as simple as renaming them. Both cities had a unique ID for each restaurant. In Chicago, this ID was originally called "License" and in NYC it was originally "CAMIS". We decided to term the unique identifier Restaurant_ID. We standardized `AKA Name,` which was the restaurant name minus business abbreviations like LLC and INC in Chicago, and "DBA," which was the business name of the restaurant in NYC, as Name. Zip codes were stored in both datasets, which we standardized as Zip_code. Creating a usable date simply involved changing character strings to date format using the lubridate package, and renaming those Date. Once this reformatting was complete, we generated Year, Month, and Weekday. While the names of Inspection Types were different between cities, this column exists similarly in both places. We did not change the types of inspections listed in either city, because this characteristic was used to filter the comparable inspections for the final dataset but was not used as a predictor in any of our models.

Considering the Chicago and NYC restaurant inspections are documented differently, we had to do some feature engineering to create symmetrical inspection predictors across both cities.

After careful consideration, we chose to look at the number of violations per inspection, the presence of critical violations (those which indicate the most risk for consumer health), and inspection outcome. Initially, the Chicago data had one inspection per row, while the NYC data had one violation per row. Violations in Chicago were stored in a character string with separate violations separated by a vertical bar, so we calculated number of violations per inspection by unnesting them, using the vertical bar as a delimiter, and then counting the number of violations per inspection. In New York, we determined an equivalent predictor, violations per inspection, through grouping by Name and Inspection Date and then counting the number of rows that appeared for each unique Name-Date combination.

Flags for critical violations already existed in the New York dataset as a binary indicator, so this only had to be created for Chicago. Reading the documentation for Chicago food inspections informed us that violation numbers 1-14 constitute critical violations. We separated these violation numbers from the Violations column while it was un-nested to count the number of violations, and indicated a critical flag if violation number 1-14 appeared at any given inspection.

Finally, we had to create a pass/fail outcome variable in both datasets. Inspection results in Chicago are "pass," "pass with conditions," "fail," or "no entry." We re-coded the results into a binary pass/fail outcome where "pass" or "pass with conditions" were coded as pass, indicated by a 0, and "fail" or "no entry" were coded as fail, indicated by a 1. Inspections in NYC receive an "A" for a score of 0-14, a "B" for 15-27, and a "C" for higher than 28. We converted this into a binary outcome to match that created in Chicago by considering a score of less than 28 a passed inspection, or a 0, and a score greater than 28 a failed inspection, or a 1. Once these pre-processing steps were completed, we selected Restaurant_ID, Name, City, Zip_code, Date, Year, Month, Weekday, Inspection_Type, Number_Violations, critical_flag and outcome from each and exported these tol a CSV file for each city.

To integrate a measure of socioeconomic status, we calculated the average income by zip code using the IRS Individual Income Tax ZIP Code Data. IRS data was provided by year and state, so we read in files for Illinois and New York states for 2015, 2016, and 2017. Rather than cleaning the entire file, we selected only relevant columns, as there was considerable information about tax return data that was irrelevant to our analysis. The final columns selected reflect the relevant information needed to compute both average Adjusted Gross Income and Average Total Income. After cleaning the dataset to one column per zip code, we intended to compute these averages by zip code, as that was our preliminary indicator for location. However, we quickly realized that this variable would have a large number of levels if we used it as a categorical predictor in our analyses.

To create slightly larger categories, we decided to group the zip codes into major neighborhoods for both cities. We used a few different sources to inform which zip codes were a part of which neighborhoods, as well as our own discretion, because zip codes do not always line up with neighborhood boundaries. This worked well since all three datasets had zip code as a common key. After identifying the zip codes that fall within major neighborhoods in each city, we filtered the IRS data (which was provided by state and year) from all of the zip codes in the state down to the ones for Chicago and New York City, and then created the neighborhood variable. We validated this with a list of the unique zip codes appearing in both restaurant datasets. We chose to eliminate some rows with zip codes outside of New York City, such as the zip code for Riker's Island and La Guardia Airport. Once the neighborhood variable was complete, we grouped by neighborhood and year, summed the adjusted gross income and total income for all of the zip codes in the neighborhood, and divided by the corresponding number of individual returns for each.

The last step in data cleaning was to join all three files together. We first combined the Chicago and NYC datasets into restaurant_data with a row bind, as these datasets had the same 12 columns and contained one row inspection. We then joined this dataset to the IRS demographics by zip code, which added columns for Neighborhood, average Adjusted Gross Income, and Average Total Income using zip code and year as the key. Pre-processing steps for all of the data can be found in the preprocess.R script.

V. Methods

We began with three distinct binary classification methods. The first method we used was binary logistic regression, where we ran two different models. The baseline model used only city as a predictor to spot any overwhelming difference between Chicago and New York. However, after this initial model we decided to use neighborhood *instead* of the city predictor; the two variables are too highly correlated to include both. The second logistic regression model included all of the predictors resulting from the variable standardization and feature engineering described above: Neighborhood, Weekday, Month, critical flag, Number of Violations per Inspection, and average Adjusted Gross Income. We then ran random forest model and a lasso model with this set of predictors.

To evaluate these models, we have decided to use Area Under the Curve, or AUC. We randomly shuffled our data into a training set (60% of the data), a validation set (20%), and a testing set (20%), then trained all models on the training set, and predicted on our validation set to get the AUCs. The AUC is the probability that the model will score a randomly drawn positive sample higher than it would a randomly drawn negative sample. In other words, this is the number of predicted true positives over the predicted false negatives, or the probability of correctly ranking a "positive"-"negative" pair. AUC favors models that are discriminative, rather than representative, which means that will allow us to choose a model that predicts out-of-sample data well. Since our aim is to develop a model that predicts well on out of sample data in order to inform restaurant owners, inspectors, and even policy-makers in a helpful manner, AUC is the model-evaluation method best suited to our analysis. We then trained the model with the best AUC on both the training and validation sets, and observed our prediction rates on our test set.

VI. Results

All of the models we ran had high AUC values, but the random forest model performed best on this measure, with 0.921. Logistic regression came close, with 0.864 AUC, and the Lasso model .862. Pictured in the following section are the ROC curves for these models:

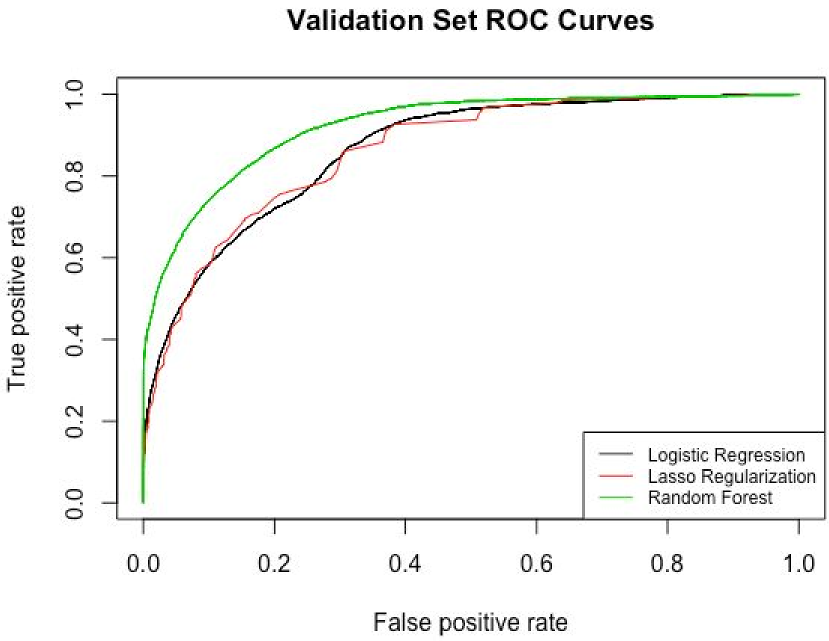

Figure 1: Validation set ROC Curves

Interestingly, the Lasso model drops weekday, month and average AGI. The coefficients retained for neighborhood are all neighborhoods in Chicago, which suggests that City may be a better predictor of inspection failure than neighborhood. Lasso also retains coefficients for number of violations and critical flags.

After we selected random forest as the best predictive model, we trained that model on the combined training and validation sets (60% + 20%). We previously ran the model the training set (60% of the data). The resulting AUC was slightly higher, at 0.922. This makes sense, as models become better at predictions given a larger set of training data. Below is the ROC curve for the random forest model:

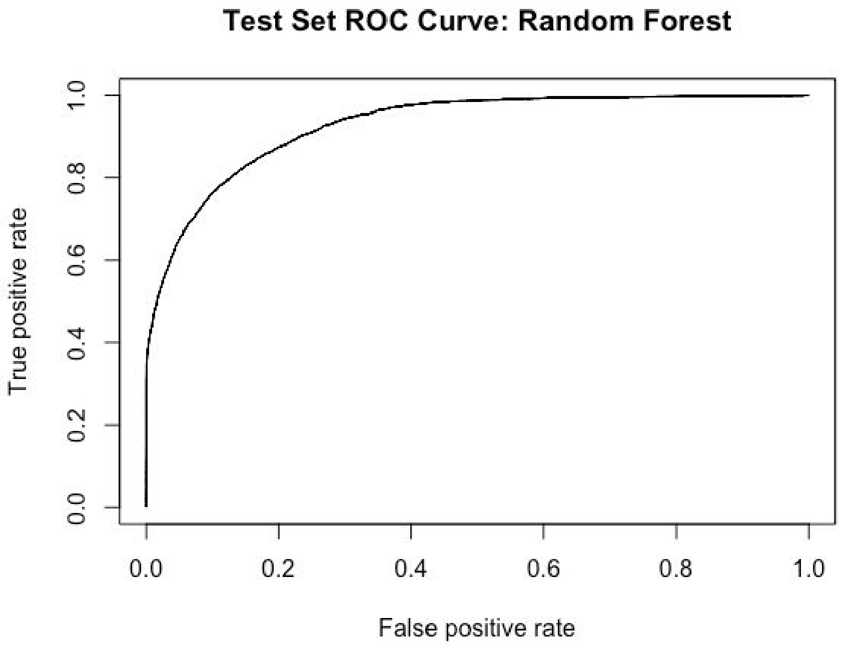

Figure 2: Test Set ROC Curve of the Random Forest model

VII. Discussion & Limitations

For the baseline logistic regression, we get an odds-ratio of .13 after exponentiating the regression coefficient. This means that inspections in New York City are about 86.95% less likely to receive a failing result than those in Chicago, the reference group. A glance at our complete dataset shows that about 46% of all restaurant inspections in Chicago result in failure, as opposed to about 10% in New York City. While this could occur for many reasons, some background research suggests that, in Chicago, "the majority of problems that lead to failed health inspections— even the ones that lead to temporary closings— are often easy and quick to fix, exposing customers to little risk" (CBS, 2015). These findings suggest that though we are able to create similar features to analyze in both cities, Chicago and New York may use different criteria to determine when restaurants fail inspections.

Using the random forest model, we were able to build a relatively well-performing predictive model that could be potentially useful in many different situations. However, there were limitations that originate from the data itself. The biggest limitation in our analysis is the lack of consistency and uniformity of inspection data cities. Though both cities have similar procedures and records for inspecting restaurants, they had few intersecting data columns, and those that did overlap may not have been evaluated uniformly. While we take note of the number of violations at each inspection and whether or not there were any critical health concerns, we cannot see the characteristics that inform these determinations or know, for example, that a critical flag in Chicago has the same meaning as a critical flag in New York. Reducing the data down to the predictors that can be created similarly in both cities unfortunately eliminates relevant information about the inspections. For example, NYC's cuisine variable would have been an interesting covariate, but Chicago did not record the cuisine type for inspected restaurants. In addition, if the City of Chicago made the demographic statistics by zip code made available as NYC did, we would have been able to include features that could affect inspectors' decision, but are not recorded during an inspection.

Though our models suggest that restaurants in Chicago are far more likely fail inspections than restaurants in NYC, we are hesitant to make any inferences about the nature of Chicago's restaurants from this analysis. More likely, the differences observed are due to differences in how the two cities inspect restaurants. Despite the availability of open data, it seems there is still considerable progress to be made before data from different cities can be integrated and analyzed reliably.

VIII. References

CBS. (2015, October 30). The Chicago Restaurants That Failed Health Inspections In 2015. Retrieved December 12, 2019, from https://chicago.cbslocal.com/2015/10/30/the-chicago-restaurants-that-failed-health-inspections-in-2015/.

Chicago Community Area and Zip Code Equivalency Files. (n.d.). Retrieved December 12, 2019, from http://robparal.blogspot.com/2013/07/chicago-community-area-and-zip-code.html.

Department of Health. (n.d.). Retrieved December 12, 2019, from https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm.

Food Establishment Inspections. (n.d.). Retrieved December 12, 2019, from https://www1.nyc.gov/site/doh/services/restaurant-grades.page.

Food Inspections. (n.d.). Retrieved December 12, 2019, from https://healthdata.gov/dataset/food-inspections.

SOI Tax Stats Individual Income Tax Statistics ZIP Code Data (SOI): Internal Revenue Service. (n.d.). Retrieved December 12, 2019, from https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-zip-code-data-soi.