Predicting Restaurant Inspections

Shehzad Bashir

DAT 7

Draft Paper

July 27th, 2015

Introduction

According to the center for Disease Control and Prevention, 48 million Americans (1 in 6 people) get sick, 128,000 are hospitalized, and 3,000 die from foodborne diseases.[1] An estimated 75% of the outbreaks came from food prepared by caterers, delis, and restaurants.[2] Currently, in most cities, health inspectors are sent to restaurants in a mostly random fashion. Since cities only have a limited number of health inspectors, this inspection approach leads to time wasted on spot checks and clean restaurants, and missed opportunities to improve health and hygiene at places with safety issues.

Yelp connects people with local businesses and along the way gathers data about customers' experiences at those businesses via business reviews. Each year, millions of people cycle through and post Yelp reviews about their experiences at these same restaurants. The information in these reviews has the potential to improve the City's inspection efforts, and could transform the way inspections are targeted.

This study is based on a Data Science contest "Keeping in Fresh", co-sponsored by Yelp in collaboration with the City of Boston, DrivenData.org and Harvard University economists to predict the future health score that will be assigned to a business at their next health inspection. I don't outright believe there is a direct correlation between a restaurant with potential safety violation and a user's review, which could vary based on factors such as food preference, service, time of the day, and other non-health related issues. However, can Yelp reviews and Boston's historic health inspection data be used to make the process of sending health inspectors to restaurants more efficient? I will attempt to answer this questions using predictive machine learning models to estimate the number of violations a restaurant is likely to get at their next inspection using historic health violation and Yelp reviews. I will then test and validate models to determine the accuracy of my predictions.

Scope Limitation

Note - due to the unfitting timeline of the competition by DriveData.org, I will not be making a submission using the submission file as I will not be able to validate my model. Therefore I will use historic data provided by DrivenData.org and split it into training and test data and attempt to build the best model that predicts the total number of violation a restaurant is likely to receive during an inspection (regardless of the severity of the violation).

Data Available

Key component to success on any data science project hinges on availability of the data. The goal of this competition is to use data from social media to narrow the search for health code

[1] http://www.cdc.gov/foodborneburden/
[2] http://www.cdc.gov/mmwr/preview/mmwrhtml/ss6202a1.htm

violations in Boston. The primary datasets available through DrivenData.org for this competition were historical hygiene violation records from the City of Boston and Yelp's consumer reviews. The goal is to figure out the words, phrases, ratings, and other features and patterns that predict violations using the Yelp and Boston city's historic health violation data. Tables below represent the files provided from each data set, mapping ID file, and the format of each file.

Yelp	Format
Business	JSON
Review	JSON
Check-in	JSON
Tips	JSON
User	JSON

ID Mapping File

Yelp ID to Business ID (CSV)

Boston	Format
Training Data	CSV
Submission Data	CSV
Unique Business IDs	CSV

Yelp Data

As shown in Figure 1, Yelp data includes business descriptions, restaurant reviews, check-ins, restaurant tips, and user review history. All of the Yelp data is structured as one JSON object per line in the file. More details about each of the dataset can be found on Yelp's Dataset Challenge website. The date range of the Yelp data is from XXX 2007 - March 2015.

Boston City Data

The City of Boston records health violations at three different levels *, **, and ***, which can be thought of as "minor", "major", and "severe" violations. The training dataset provides historic records of violation along with their severity. The date range of the historic inspection data is from October 2006 – March 2015.

In order to be able to identify restaurants across the Boston and the Yelp data, DrivenData.org provided an ID mapping that correlates restaurants in both data sets. Particular restaurants in the Boston dataset may match multiple Yelp businesses (since sometimes restaurants change names or move). Table below provides a sample of what the training data looks like.

id	date	restaurant_id	*	**	***
589	2/2/2010	KAoKWjog	3	0	1
28589	12/10/2009	p038M4om	2	0	0
31170	7/16/2008	B1oXymOV	4	0	0
2600	1/30/2015	m0oWJl3G	1	0	3
1016	3/19/2012	rJoQwlEV	0	0	0
10904	2/10/2010	rNo6Ag36	0	0	0
6805	12/30/2012	V430G4EB	0	0	0
30877	9/17/2010	lnOReVoN	1	1	0
284	3/18/2010	njoZ9y3r	6	0	0

Features and Response

Response: Health Violations

The response, otherwise known as the *dependent variable* or the *y*, is the predicted total number of violations a restaurant is likely to receive. Due to scope limitations, number of violations will not be predicted by severity level but rather the total number of violations, which is the sum of minor, major, and severe violations.

Features: Yelp Reviews and Historic Health Violation Data

The features, otherwise known as the *independent variables* or the *X*, will be primarily derived from Yelp reviews and the City of Boston's historic health violation records. The following lists include features that will be explored for this study. Note some features in the lists below may not be used for analysis as they were found not found suitable for the predictive model.

Features available from Boston's historic health violation records include:

- Date: Date of previous health inspection – some restaurants have over 30 health inspection records
- *: Number of minor health violations
- ** : Number of major health violations
- *** : Number of severe health violations

Features available from Yelp dataset include:

- Type of Business
- Business Name
- Location and address
- Stars rating – rounded to half stars
- Review Count – number of reviews

- Categories – category a restaurant falls into
- Hours of operation
- User ID – unique ID of the reviewer
- Review text
- Date of review
- Check-ins
- Votes – number of votes each review received
- User: Average number of starts a reviewer gave on Yelp reviews
- User: Number of reviews a user has posted on Yelp
- User: Number of friends per reviewer
- User: Indication of whether a reviewer is an Elite Yelp member
- User: How long a user has been Yelping since

Data Cleaning and Pre-processing

As mentioned previously, there were two primary datasets provided by DrivenData.org for this particular competition – Yelp data, and City of Boston's historic violation records. Prior to conducting any analysis, here are the steps I took to process and clean the data.

1. First, in order to work with the data in a consistent format, I converted Yelp data (which was provided in JSON format) to CSV files for uniformity with the Boston city data.

2. Second, I split the data into two datasets: test and validation. Because I will not be using the 'submission' data required for the competition, I will use the training dataset provided by DrivenData and split that dataset by date. I will use October 2007 – March 25th, 2014 as training data and March 26th, 2014 – March 25th, 2015 data for validation once I've tested my model.

3. Merged datasets using multiple joins from each of the Yelp files, cleaned and filtered/dropped unnecessary variables for ease of use

---

Following is an outline of how I will structure the rest of the paper!!

Feature Engineering

> ***Feature Engineering*** is the process of using domain knowledge of the data to create features that make machine learning algorithms work. It is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.

New Feature: After split, combined to create one record per restaurant and averaged the number of violation, then created new feature to sum violations from each level

Data Exploration (What you learned from exploring the data, including visualizations)

- Heatmap, trends, other exploratory data
- Initial features explored and found what?
- Feature Selection (how did you choose which features to use in your analysis)

> ***Natural Language Processing (NLP)*** is a branch of artificial intelligence that deals with analyzing, understanding and generating the languages that humans use naturally in order to interface with computers in both written and spoken contexts using natural human languages instead of computer languages. (Source)

Natural Language Processing (NLP)

- How text processing and NLP was applied in this project

Applying Machine Learning & Model Evaluation (Details of your modeling process, including how you selected your models and validated them)

- How machine learning (model) was applied and what steps to get there
- Models tested/used = Linear Regression

Linear Regression (Best model should be # 1)

a. Model Findings
b. Parameter Tuning (How did you choose which features to use in your analysis)
c. Model Evaluation

Findings/Predictions

- Challenges and successes
- Conclusions and key learnings
- Possible extensions or business applications of the project