Predicting Restaurant Inspections

DAT 7   
Shehzad Bashir

July 13th, 2015

Shehzad Bashir

DAT 7

Final Paper

August 12th, 2015



# Introduction

According to the center for Disease Control and Prevention foodborne diseases account for 48 million Americans (1 in 6 people) illnesses, 128,000 are hospitalizations, and 3,000 deaths.[[1]](#footnote-1) An estimated 75% of the outbreaks come from food prepared by caterers, delis, and restaurants.[[2]](#footnote-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 health inspection efforts, and could transform the way inspections are targeted.

This study is based on a Data Science contest “[Keeping it Fresh](http://www.drivendata.org/competitions/5/)”, 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. There may or may not be a direct correlation between a restaurant with potential safety violation and a user’s review, as reviews 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 attempt to build the best model that predicts the total number of violations 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 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.



|  |  |
| --- | --- |
| **Boston** | **Format** |
| Training Data | CSV |
| Submission Data | CSV |
| Unique Business IDs | CSV |

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



ID Mapping File

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](http://www.yelp.com/dataset_challenge) website. The date range of the Yelp data is from September 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.



# Features and Response

Prediction

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 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 rolled up the dataset by ‘restaurant ID’ to create one record for each restaurant.
3. Merged dataset using multiple joins from each of the Yelp files, cleaned and filtered/dropped unnecessary variables for ease of use

**Feature Engineering**

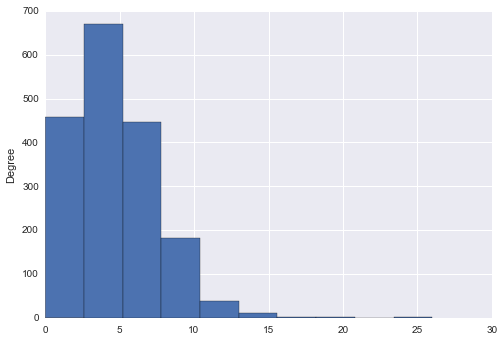
***Feature Engineering*** is the process of using domain knowledge of the data to create [features](https://en.wikipedia.org/wiki/Feature_%28machine_learning%29) that make [machine learning](https://en.wikipedia.org/wiki/Machine_learning) algorithms work. It is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.

Rolling up by ID, step 2 of the data cleaning process, required defining how to handle other features that cannot be simply rolled up for restaurants with multiple records (more than one restaurant inspection). Therefore, for restaurants that had more than one inspection I averaged the total number of violations for each restaurant, by severity. In addition, I created a new feature in the dataset, which simply calculated the sum of the total number of violations for severity levels. The total number of violations is the ‘response variable’ of this data.

In addition to the total violations feature, a new feature indicating sentiment polarity was developed for enhancing model accuracy. This feature will be discussed further in the Natural Language Processing part of this report.

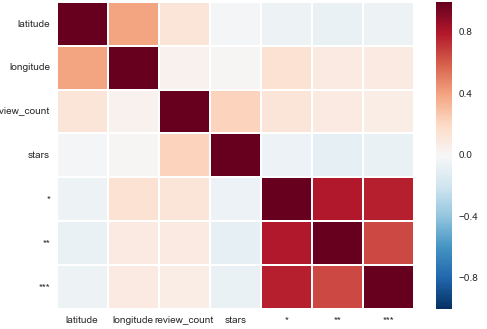
Data Exploration

To begin data exploration, it is important to understand the response variable. I began by looking at a simple histogram of what the range of total violations looks like. As noted in the histogram, the majority or the restaurants have anywhere from 0 to 10 total violations. In addition, there are few outliers in the dataset that have high number of violations, indicating a very poor health inspection record.



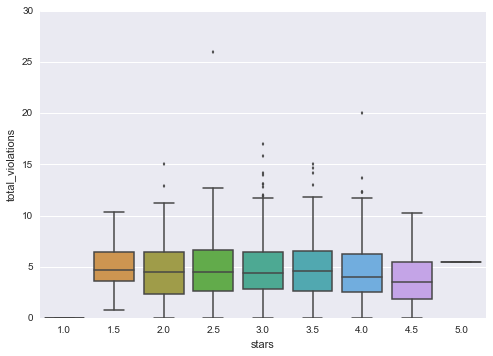
Number of Health Violations

A correlation heatmap indicates that, unsurprisingly, there is a strong relationship among the different levels of severity, indicating that if a restaurant gets one level of health violation, they are likely to get other levels of violation as well. The below heatmap includes some of the features that were explored in the dataset.



In addition, I wanted to look at the relationship between a Yelp rating of the restaurant and the total number of violations a restaurant received. The following box plot indicates that generally there is very low variance in the number of violations for restaurants that had anywhere from 2 to 3.5 star rating. Restaurants with a rating of 4 or 4.5 tend to have a low number of health violations, indicating a positive consumer experience and arguably a cleaner restaurant.

**Total Health Violations and Yelp Rating**



Applying Machine Learning and Model Evaluation

### Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. There are two main categories of machine learning: supervised learning and unsupervised learning. Supervised learning predicts an outcome based on input data, whereas unsupervised learning extracts structure from data. There are two categories of supervised learning:

### Regression – Outcome we are trying to predict in continuous

### Classification – Outcome we are trying to predict is categorical

### My response variable is the total number of violations, therefore, for this study, I will only be applying supervised regression methods. There are many different models that can be applied to predict continuous response, however for this study I will limit to two of the more common models, which are linear regression and random forests. Linear regression models the relationship between a scalar dependent variable (*y)* and one or more explanatory (independent) variables (*X*). Random forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.[[3]](#footnote-3)

### Baseline Model (Model # 1) – Linear Regression

### My first model is a linear regression baseline model to evaluate my predictions based on the features available for assessment without conducting sentiment analysis. The following features were used for the baseline model:

### Number of Yelp reviews a restaurant received

### Length of the review text

### Restaurant’s Yelp star rating

### Model Evaluation Metric

### The evaluation metric used for this and all models used in this study is the Root Mean Squared Error (RMSE). The Root Mean Square Error (RMSE) (also called the root mean square deviation, RMSD) is a frequently used measure of the difference between values predicted by a model and the values actually observed from the environment that is being modelled. [[4]](#footnote-4) The RMSE of a model prediction with respect to the estimated variable X*model* is defined as the square root of the mean squared error:

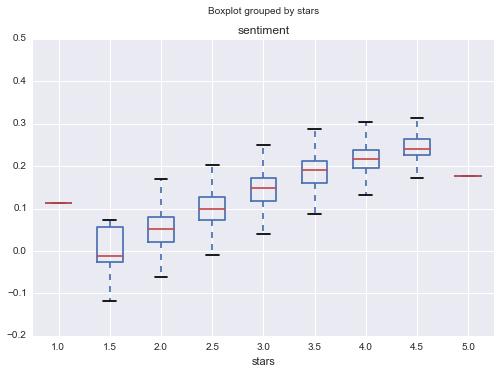


Using cross validation method to test my prediction, the baseline model RMSE score was **2.85**. Cross-validation is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set.

### Applying Natural Language Processing (NLP)

***Natural Language Processing (NLP)*** is a branch of [artificial intelligence](http://www.webopedia.com/TERM/A/artificial_intelligence.html) that deals with analyzing, understanding and generating the languages that humans use naturally in order to [interface](http://www.webopedia.com/TERM/I/interface.html) with computers in both written and spoken contexts using natural human languages instead of computer languages. ([Source](http://www.webopedia.com/TERM/N/NLP.html))

### In an effort to improve on the baseline model, my next effort was to add a feature that takes into consideration sentiment polarity of the Yelp reviews for each restaurant. A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level — whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. A sentiment polarity score ranges from -1.0 (negative sentiment) to 1.0 (positive sentiment). The boxplot below represents the sentiment polarity score by Yelp star rating of a restaurant, representing a clear correlation among the polarity score and the star rating, indicating positive reviews for restaurants with higher rating.

****

### Model with Sentiment Polarity Score (Model # 2) – Linear Regression

My second model included the sentiment polarity score as a feature along with the number of yelp reviews, length of text, and ratings. Adding this feature improved the model very little, bringing the RMSE to **2.82**. Note, RMSE is a loss function, therefore we want to try to minimize RMSE score to improve our model.

### Model with Sentiment Polarity Score (Model # 3) – Random Forests

A Linear Regression model provides a good foundation for predicting relationship among variable, however a Random Forest model is able to discover more complex dependencies within a dataset. Using the same features used in model # 2, I was able to improve the RMSE to **2.61**, with no parameter tuning. However, tuning the parameters of a model to a given problem at hand may be essential for good algorithm performance.

### Pre-processing for Text Analysis

### Machine learning models cannot handle standard text data, therefore prior to building a model for text analysis it is required to separate text into units such as sentences or words to give structure to previously unstructured text. One method to perform this function is to use the Term Frequency – Inverse Document Frequency (TF-IDF) method. TF-IDF computes “relative frequency” that a word appears in a document compared to its frequency across all documents. It’s particularly useful in identifying important words in each document.

### Model with text analysis (Model # 4) – Random Forests

Applying TF-IDF to text data and taking other features applied in model # 3 into consideration, I was able to improve the RMSE score to **2.51** with a new random forest model. Note building a TF-IDF increased the number of features to approximately 83,000. Therefore it becomes more relevant to take into consideration tuning parameters for performance and improving accuracy. After testing several parameters such as the maximum number of features to consider and the minimum document frequency, I was able to narrow down the number of features to 1,500 and with a minimum document-frequency of 6.

**Model Summary**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Model Type | Number of Features | RMSE |
| 1 | Linear Regression | 3 | 2.85 |
| 2 | Linear Regression | 4 | 2.82 |
| 3 | Random Forests | 4 | 2.61 |
| 4 | Random Forests | 1500 | **2.51** |

# Summary and Next Steps

In summary, while I was able to improve the RMSE of each iteration of the models applied, I believe the models can be tuned further to improve accuracy. Working with text particularly adds a new challenge in building a reliable model and tuning parameters. One of the biggest challenges may have been the performance and computing power it requires when working with a dataset that increases significantly in size once text is taken into consideration. Therefore, it becomes extremely important to balance the tradeoff between performance and the number of tests to run.

One particular challenge in this dataset was the limitation of the numerical features available (excluding text data). Limiting the scope of the project removed some features, such as date, that may be good indicators of predicting the number of health violations a restaurant is likely to have at their next inspection. Therefore I will continue to work on this study in an effort to improve accuracy. Moving forward, my goal will be to predict the number of health violations by severity, as required for DrivenData competition. In addition, I will account for features such as date and user profile data which were previously disregarded due to the scope limitation. In addition, I will continue to test different tuning parameters with each of the models above to improve RMSE score.

1. <http://www.cdc.gov/foodborneburden/> [↑](#footnote-ref-1)
2. <http://www.cdc.gov/mmwr/preview/mmwrhtml/ss6202a1.htm> [↑](#footnote-ref-2)
3. https://en.wikipedia.org/ [↑](#footnote-ref-3)
4. http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=9&ved=0CF0QFjAIahUKEwiEvYKF36HHAhXHjw0KHYxTAa4&url=http%3A%2F%2Fwww.ctec.ufal.br%2Fprofessor%2Fcrfj%2FGraduacao%2FMSH%2FModel%2520evaluation%2520methods.doc&ei=EErKVYT1F8efNoynhfAK&usg=AFQjCNGpg8UCAQI6SuRjBqCVemKgCXPZpw&sig2=X7kMsRzYdDpuMAm-kxC7Pg&bvm=bv.99804247,d.eXY [↑](#footnote-ref-4)