Springboard–Foundations to Core Program
Capstone Project 2 Proposal
By Anita Durg
October, 2022

(1) What is the business problem?

The City of Boston is a home to thousands of restaurants but has only a handful of health inspectors to monitor and improve food safety and public health. Health inspections are usually random, which can increase time spent at clean restaurants that have been following the rules carefully — and missed opportunities to improve health and hygiene at places with more serious food safety issues.

Goal is to predict the number of health violations for all three levels (*(one star)- "minor", ** (2 stars) - "major", and ***(3 stars) - "severe") during an inspection of a restaurant in the city of Boston on a specific date using Yelp data.

(2) Who are the intended stakeholders, and why is this problem relevant to them?

- City of Boston
- Public

The problem is relevant to the City of Boston as it takes pride in engaging, empowering, and improving life for residents in the City through technology. It is the responsibility of the city to ensure public health and safety. Our predictive model will help City of Boston as it can substantially improve the City's inspection efforts and can change the way inspections are organized

(3) Where are the datasets available from?

Key Data Source - Yelp, cityofboston.gov, DrivenData.

All Historical Violations:

https://drivendata.s3.amazonaws.com/data/5/public/AllViolations.csv

Restaurant ID Mapping:

https://drivendata.s3.amazonaws.com/data/5/public/restaurant_ids_to_yelp_ids.csv

Yelp Reviews:

https://www.yelp.com/dataset

(4) What data science approaches do you anticipate you will use to model the business problem as a data science problem?

The business need can be addressed by building classification models. Here, the Logistic regression algorithm might be appropriate. Random Forest Classifier model might work for this problem. RFC is known as an accuracy focussed algorithm. It might capture more patterns from the reviews and help in predicting the numbers for the three categories correctly. In general, several algorithms will be used and the corresponding models will be evaluated and ranked with respect to the appropriate performance evaluation metrics.

(5) How do you anticipate that the intended clients will use the results of your CP2 to address the original business problem?

Models will be built so that they will predict the class to which a given input belongs, where the classes are defined according to the number of health violations for all the three levels like one-star, two-stars, and three-stars for a restaurant for a specific date. The input will be a vector of features to be engineered using the available data, although we anticipate that the textual data will contribute the most to these features. The result generated by models will guide the city of Boston to decide where to send their health inspectors.

In addition, we will explore interpretability approaches to establish connections between the features and the output of the models.