Using Yelp Reviews to Predict Restaurant Health Violations

Ally Chao, Xinyu Wu

Project Overview

In this project, our goal is to use supervised machine learning and NLP techniques to understand and predict health violations of Boston-based restaurants. We will combine two datasets, the first being Yelp's restaurant review dataset, and the second being a dataset with historical Boston restaurant health violations. We plan on combining the health violations data (mild, major, severe categories) into a single metric for each restaurant that is a proxy for the severity of the health violations - thereby turning this problem into a regression task. Even without the extra considerations related to the ongoing pandemic, holding restaurants to high health standards is critical — the CDC estimates 48 million illnesses, 128,000 hospitalizations, and 3,000 deaths annually related to foodborne diseases like E. coli and Salmonella. Therefore, this project is important to both consumers who enjoy eating out without risking their health and restaurant-owners that want to attract new customers and grow their business.

Project Dataset

For this project, we will be using three datasets:

- The Yelp dataset, consisting of 3 json files that contains data on businesses and their reviews (https://www.yelp.com/dataset). We will primarily be using the dataset of Yelp reviews.
    - 6,990,280 records and 9 features (review_id, user_id, business_id, stars, date, text, useful, funny, cool)
- A simple csv "dataset" to map the restaurant_id in the violations dataset to the business_id in the Yelp dataset (https://drivendata.s3.amazonaws.com/data/5/public/restaurant_ids_to_yelp_ids.csv)
    - 1868 records (1868 restaurants) and 2 features (restaurant_id and yelp_id)
- A csv dataset containing the historical violations of data for Boston (https://www.drivendata.org/competitions/5/page/33/)
    - 34880 records and 5 features (date, restaurant, number of level 1 violations, number of level 2 violations, and number of level 3 violations)

Approach and Methodology

For feature normalization, we will scale the ratings (0-5), review count, and other quantitative measures. We will combine the mild, major, and severe health violations into a single score that is normalized. For feature selection, we will use Lasso regression and correlation metrics such as Pearson's correlation. Irrelevant features (name, hours of businesses, etc.) will be removed. We will use multiple linear regression, Lasso regression, Ridge regression, and polynomial additive regression. Other potential models include support vector regression, decision tree regression, and random forest regression. Our project will use Python, and packages including json, pandas, numpy, and scikit-learn. Data visualization will utilize matplotlib and seaborn. To evaluate our model, we'll use mean squared error and R^2 score.

Outcome

We hope to produce a model that is able to predict the health violation score (a weighted combination of the number of mild, major, and severe violations) of a restaurant based on its Yelp ratings.

Plan

We will both work together to handle the data cleaning and preprocessing as well as feature normalization and selection. Each of us will also produce two data visualizations. Xinyu will train and test the data (and report evaluation metrics) using multiple linear regression and Lasso regression, and Ally will perform the Ridge regression and polynomial additive regression.