ORIE 4741 Project Proposal

Yuyang Chen(yc2324), Xinyun Tang(xt222), Xiaoxi Zhang(xz577)

1 The Problem

Restaurant inspections are conducted by Department of Health in most cities for food security control. Example inspections include food handling practices, product temperatures, personal hygiene, facility maintenance and pest control. Although data is open to the public, general population only realize what restaurants have health issues after they fail inspections. We are interested to see how restaurant review data from Yelp can help us to predict whether a restaurant can pass an inspection. Furthermore, we can even use the same set of data to determine a healthy level for restaurants in those cities with more specified health policies.

We believe our project can help to ensure people can eat healthy and safely while dining out. Also, we can better allocate the health department resources on inspection work, along with the combination of public data on Yelp.

2 Datasets description

2.1 Yelp Dataset

The independent variables in our problem come from `https://www.yelp.com/dataset/challenge`. This dataset contains extensive description about business owners on Yelp.com, including their overall ratings, and customer reviews. The key features we are interested in are the type of cuisine a restaurant sells, their overall rating, and features extracted from customer reviews (e.g. whether people had food poisoning and whether the restaurant is expensive, etc.).

2.2 Inspection datasets

The datasets contain information about restaurant address, category, inspection date, violation type or code, description of violations, risk level and the inspection score. Date information enables us to analyze recurrent inspections on certain restaurants and to answer whether previous inspection act a positive or negative predictor of inspection results and whether a seasonal trend exists. We will use address information to tell whether specific neighborhoods have higher risks of violations. Useful text information of violations can be extracted from violation types and descriptions so that we can look for specific texts from Yelp reviews to form our feature space. We plan to use inspection datasets from San Francisco, New York City and Chicago. They adopt different grading systems for restaurant inspections, which are continuous scores, level scores and pass or fail, respectively. This difference enables us to build different models, such as regression, clustering and classification for predicting the inspection results of restaurants and giving expected healthy level.