Geovanna K. Meier

Predicting Critical Health and Safety Violations at Food Establishments

The Issue and Why it Matters

As a destination that attracts millions of tourists every year, Southern Nevada, and particularly Las Vegas, offers many possibilities when it comes to dining options. With food offerings that range from family owned locales to Michelin star restaurants, the choices may seem endless. This can be a great experience for visitors, but a challenging task for the Southern Nevada Health District, as it must inspect every establishment to ensure that it meets the food and safety requirements to protect the public's health.

Since these are unannounced inspections that happen at least once a year, it's vital to prioritize inspections at restaurants that are most likely to have food and safety violations to avoid foodborne outbreaks and maximize the effectiveness of inspections.

Approach

Using publicly available data from two sources, restaurant inspection results and Yelp's data on businesses and consumers' ratings, the objective is to predict critical health and safety violations, using the inspection results as the predictor variable. This will be accomplished by identifying patterns between business attributes, consumer ratings, and inspections results.
The results, can provide actionable insights to local officials to decide when to adjust the rank and frequency of inspections based on locations that are most likely to have critical violations.

Open Data

1. The Southern Nevada Health District Restaurant Inspections is a publicly available dataset that provides inspection results of food establishments. Each record represents a single establishment, with the name, location, date of the inspection, and inspection grade, among other attributes:

Serial Number	**Zip***	**Inspection Demerits**
Permit Number	**Current Demerits**	**Inspection Grade***
Restaurant Name	**Current Grade***	**Permit Status**
Location Name	**Date Current**	**Inspection Result**
Category Name*	**Inspection Date***	**Violations**
Address	**Inspection Time**	**Record Updated**
City	**Employee ID**	
State	**Inspection Type**	

Geovanna K. Meier

Target variable:
Once an inspection is conducted, the establishment receives an inspection grade based on the number and type of violations. Each violation is assigned a value, called demerit, that results in a grade. 'A's are given to establishments that are compliant (0 – 10 demerits), while 'B' (11 – 20) and 'C' (21-40) are downgrades and indicate critical or major violations. Hence, the objective is to predict restaurants that are likely to be downgraded.

2. The Yelp Academic Dataset includes two files, the businesses description and the customer reviews. Each file is a JSON object file per line file. The business description contains:

business_id	**state**	**review_count ***
name	**postal_code**	**is_open**
neighborhood	**latitude**	**attributes ***
address	**longitude**	**categories ***
city	**stars***	**hours**

The reviews file contains customers' reviews of a particular restaurant including:

review_id	**stars**	**useful**
user_id	**date**	**funny**
business_id	**text**	**cool**

How to improve decision making

After downloading the.csv file for the restaurant inspections and the JSON files for the Yelp data, the datasets will be merged. Since the inspection results don't include a restaurant id to match the business id from Yelp, the datasets need to be merged using a combination of restaurant name, zip code, and date.

For the target variable (inspection grade), a separate variable will be created to have two levels. One level will combine the B, C, and other conventions denoting a downgrade or non-compliant and a second level will have the A-grade or compliant.

Since there are other variables in both data sets that have categorical variables with multiple levels (i.e. category names, attributes, categories), similar levels would be grouped and changed to binary variables or left as levels in the same variable.

Geovanna K. Meier

For the analysis, the variables marked with an * are being considered. The idea for using the date, is that the month of the inspection and location can influence food handling and preparation. For instance, summer months may see higher food violations as higher temperatures can increase the growth of bacteria if food is not prepared or consumed promptly. For the Yelp data, feature engineering will be used to obtain specific information on the establishment, such as cuisine type, using the attributes variable.

A logistic regression and/or random forest will be used for the predictive model.

Deliverables

The deliverables will include a Python script and a Jupyter notebook including visualizations for the EDA, Python code and write up with the findings.