Improving Success Rates of Restaurants

Final Report, Capstone Project 1

1. Problem Statement

Restaurants are a big part of modern life and they make up a significant portion of small businesses everywhere. They are known for high turnover and failure rates, especially within the first year of business.

Failed restaurants are costly to the restaurateurs, to the industry and to the diners. When restaurants fail as businesses, the restaurateurs themselves and their financial sponsors bear economic burden, the food services industry lose potential growth from the contribution of new ideas and economy, and the diners lose out on opportunities for new enjoyable experiences. This study in data science aims to help restaurateurs to incorporate more success factors into their strategies when opening new restaurants.

2. Dataset

In order to make inferences about restaurants in general, a decent sample of restaurant data is needed for analysis. Yelp provides just the dataset needed at https://www.yelp.com/dataset. The dataset is downloaded and unpacked into 3 files that are in json lines format. Each file is gigabytes in size.The first file contains 192,609 businesses on Yelp with over 1.2 million business attributes like hours, parking, availability, and ambience. The second file includes the aggregated check-ins over time for each of the 192,609 businesses. The third file contains the reviews of the 192609 businesses. The file containing the reviews is over 5 gigabytes in size.

2.1 Data Wrangling

- Read files that are in json lines format.
- Format json strings.
- Create pandas dataframes from json strings.
- Filter businesses into restaurants dataframe.
- Filter reviews for restaurant reviews by restaurant standing.
- Filter restaurants outliers - those with less than 10 reviews.
- Triage restaurants into those of good (4-5 stars), moderate (2.5-3.5 stars) and poor (1-2 stars) standing into dataframes.
- Triage check ins and reviews by restaurant standing (good, moderate, poor) into 6 dataframes.
- Parse and reshape date columns of restaurant check in dataframes.
    - Split original date string by comma
    - Strip each split date string of leading whitespace
    - Convert each date string to datetime value
    - Insert each date value with its business id into a dictionary
    - Create list of the dictionaries containing business ids and dates
    - Create dataframe from list

- Reset index for all resulting dataframes before using for analysis.

3. Initial Findings

Through exploratory data analysis, significant relationships were found between restaurant category and restaurant standing, and between restaurant standing and rating volatility. Bootstrap inference was utilized to confirm the significance of relationships between variables.

3.1 Restaurant Ratings

Yelp restaurant dataset contains significantly more restaurants in moderate and good standing than those in poor standing. Most restaurants have 3-4 stars in rating, with 1 star ratings being the rarest.

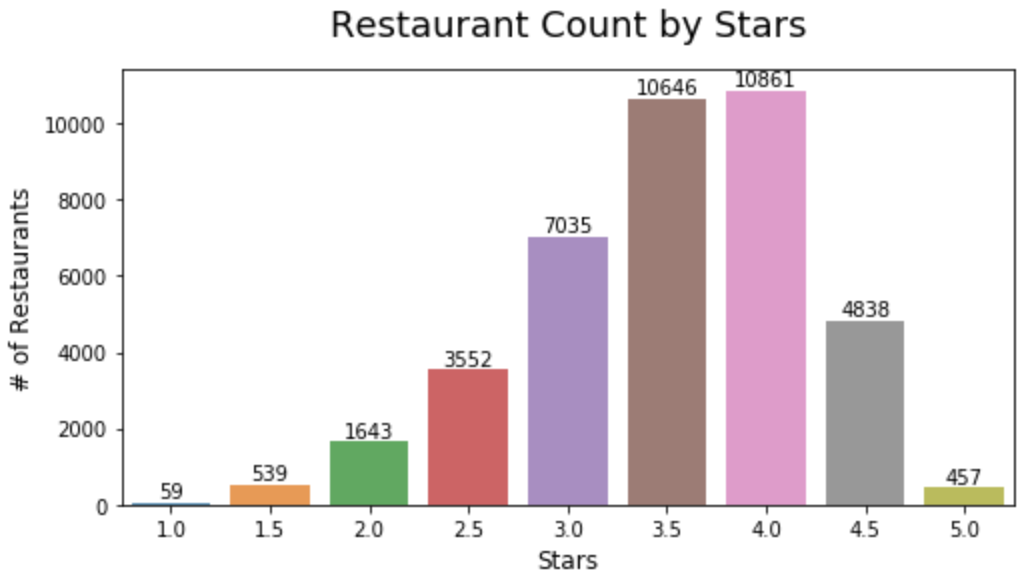

Figure 1.1 *Ratings distribution of yelp restaurants*

When triaged into 3 standing categories of good (4-5 stars), moderate (2.5-3.5 stars), and poor (1-2 stars), the distribution shows that most restaurants are average and above average (figure 1.2).

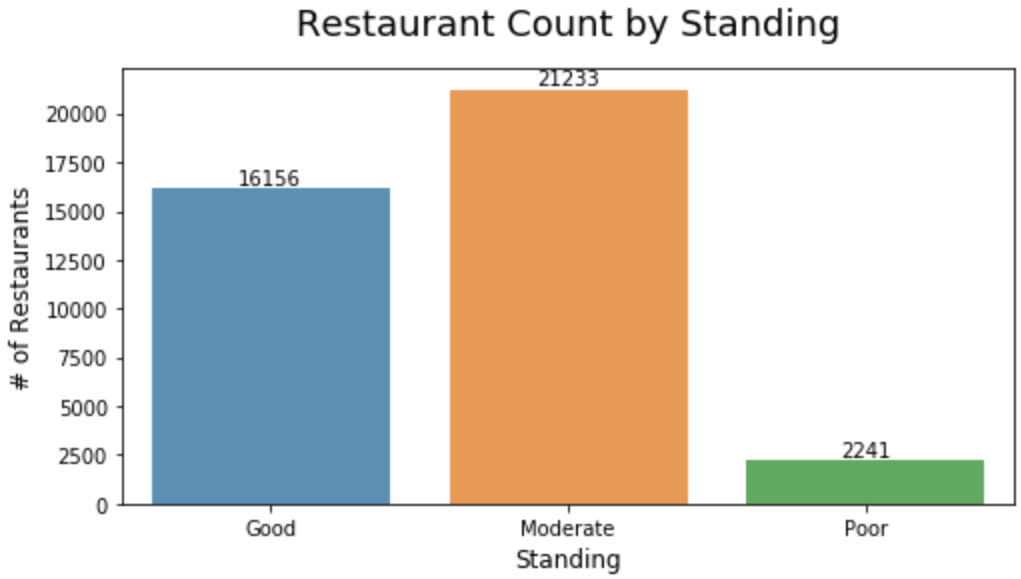

Figure 1.2 *Distribution of yelp restaurants triaged into 3 standing categories*

The restaurants average 3.5 stars in rating and 0.71 in standard deviation (figure 1.3).

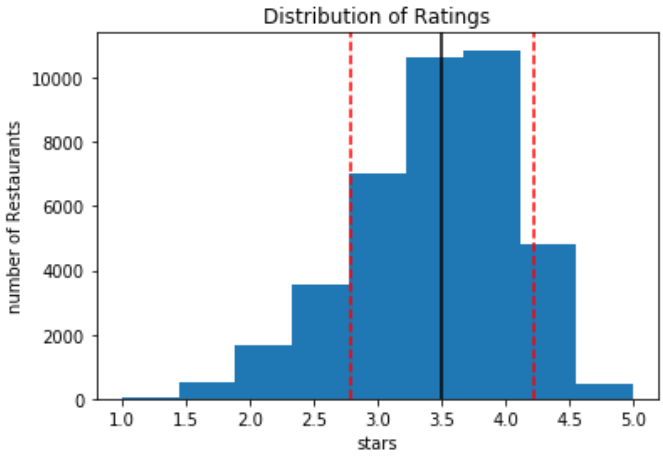

Figure 1.3 *Mean and standard deviation of yelp restaurant ratings*

This is good news for future restaurateurs. It is easier to perform well than to perform poorly!

3.2 Restaurant Categories

Yelp has over 7000 categories for restaurants. Restaurants of type Vegetarian, Vegan, Food Truck, Cocktail, Dessert, Tacos, Juice, Coffee, and Ramen have standing distributions that differ from that of the larger population. These categories contain more good restaurants than moderate and poor restaurants (figure 2.1).

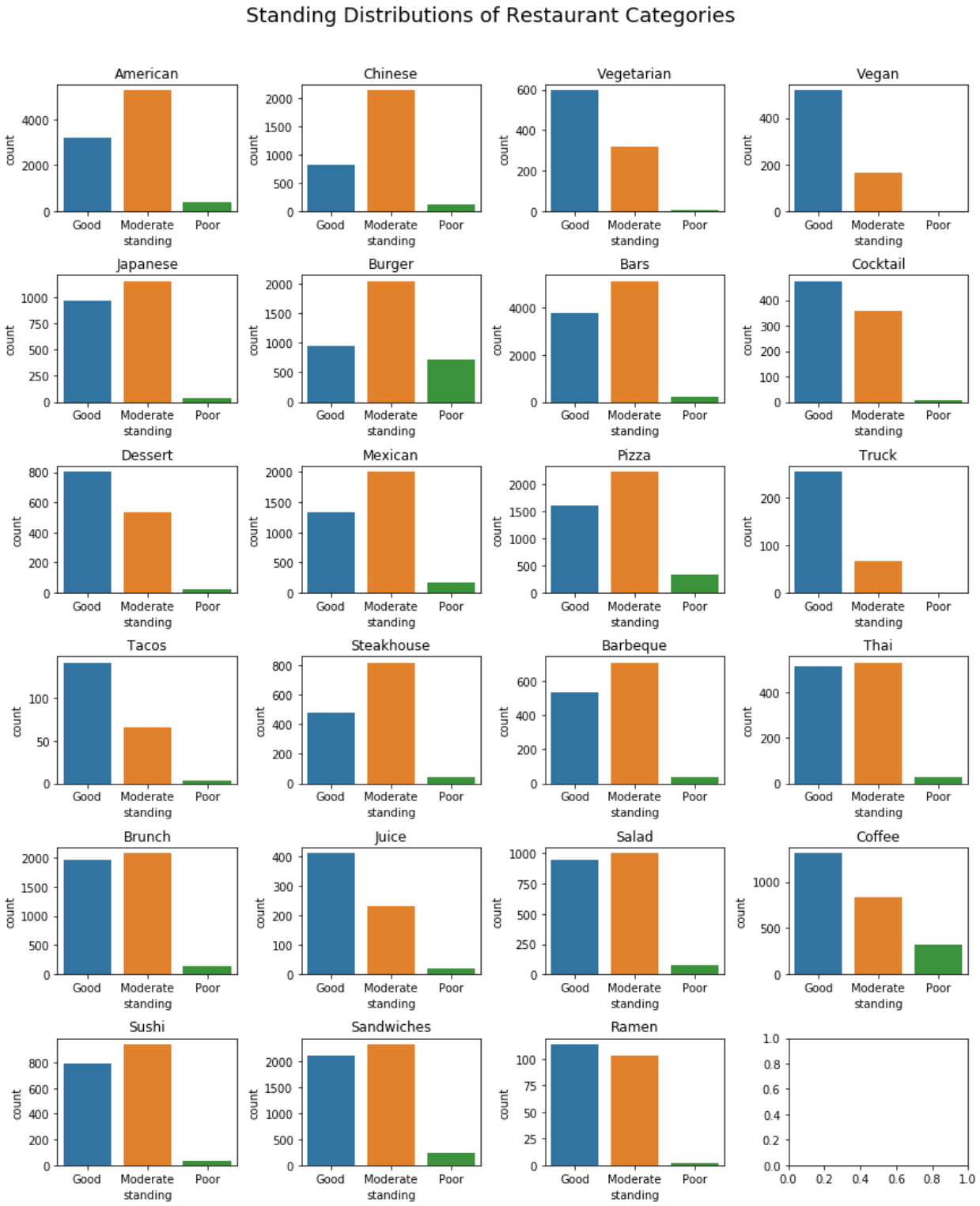

Figure 2.1 *Distribution of yelp restaurant standing by category*

Using bootstrap inference to perform a t-test on the hypothesis of whether mean ratings differ by category, significant relationship is found between restaurant category and restaurant standing (figure 2.2). New restaurateurs might benefit from knowing restaurants in what categories are easier to start up and manage. Perhaps keeping it simple with a food truck or focusing on beverages is the way to go for a new restaurateur in the business.

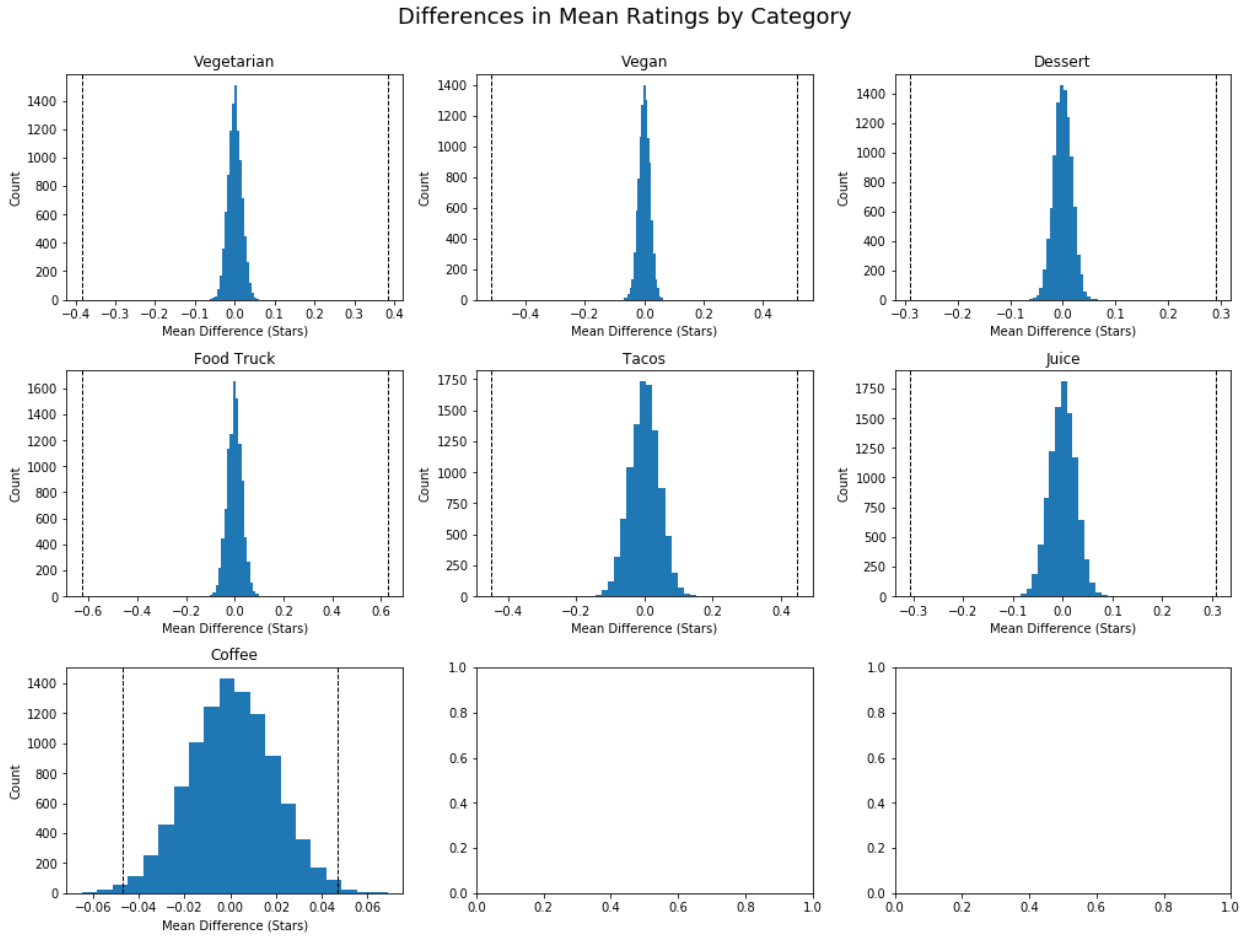

Figure 2.2 *Test statistics of differences in mean ratings by category*

3.3 Performance Volatility

Moderate restaurants fluctuate more widely in performance over time than good and poor restaurants (figure 3.1). Poor restaurants fluctuate more widely in performance over time than good restaurants. Good restaurants perform with higher consistency over time.

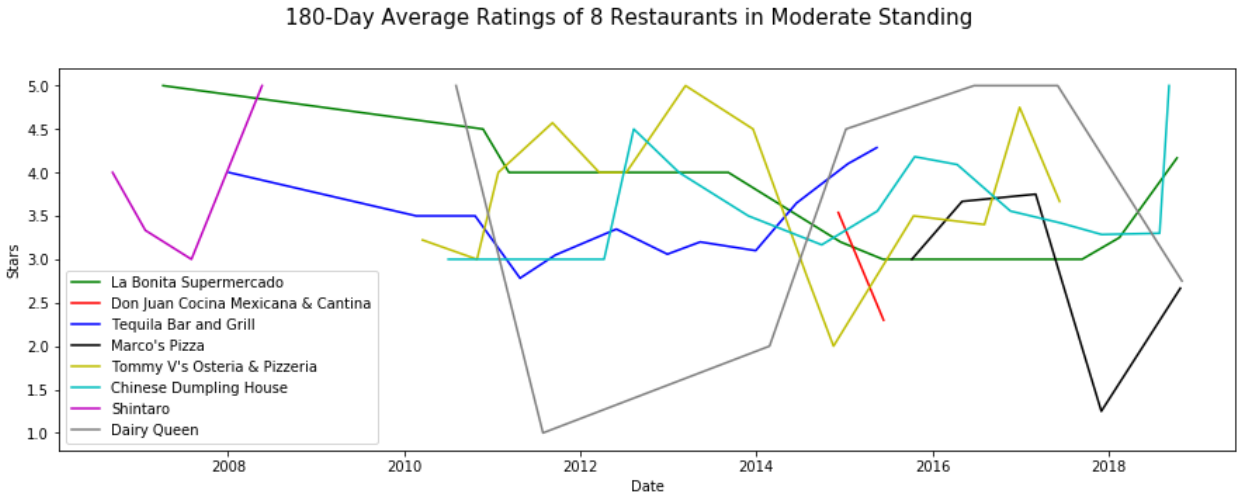

Figure 3.1 *Changes in the ratings of 8 yelp restaurants in moderate standing*

Using bootstrap inference to perform t-tests, the resulting p-values of 0 suggests there are significant differences in the standard deviations of ratings depending on restaurant standing (figure 3.2). These differences are most significant between moderate restaurants and those of other standings.

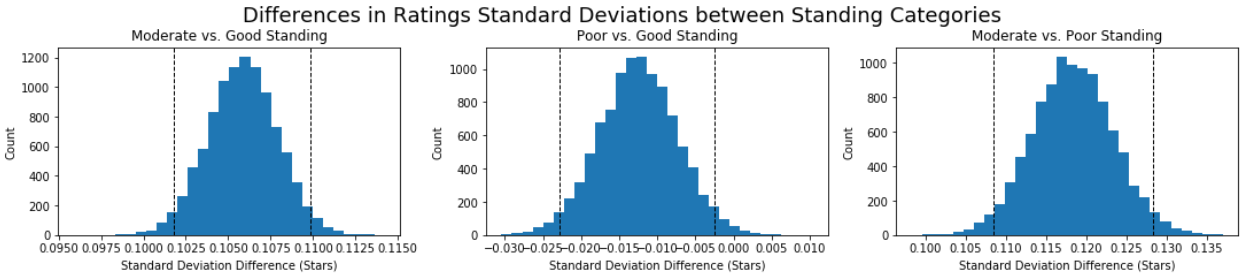

Figure 3.2 *95% confidence intervals of differences in rating standard deviations between standing categories*

The larger variances in ratings of moderate restaurants suggests to future restaurateurs the importance and advantage of starting on the right foot, and the difficulty of improving consistency in performance once the restaurant is up and running.

3.4 Price and Features

Price, and other features such as alcohol, reservations and outdoor seating make no difference in ratings.

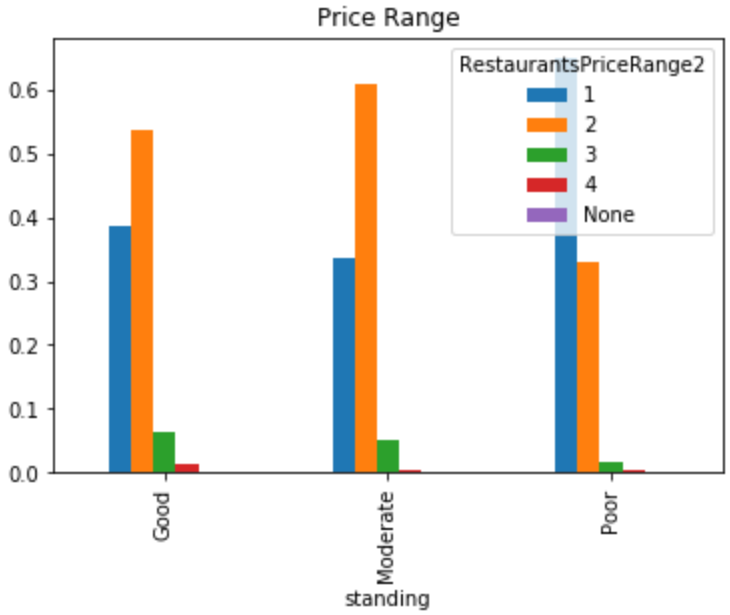

Figure 3.3 *Normalized distribution of price range is similar across the standing categories*

Using bootstrap inference for the hypothesis test, p values greater than 0.05 suggest that none of the restaurant attributes tested make a difference in restaurant ratings.

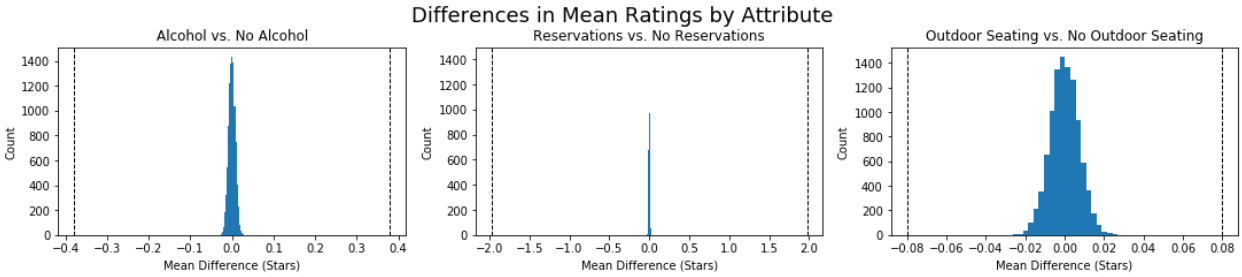

Figure 3.4 *Differences in mean ratings between different features and the test statistics*

3.5 Location and Density

Restaurants in Las Vegas show similar distributions regardless of the restaurant density in the surrounding area.

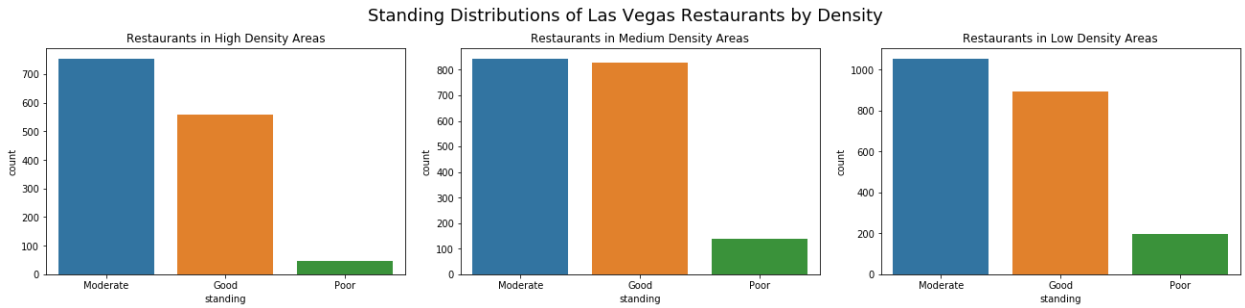

Figure 3.5 *Distribution of restaurants are similar in high, medium and low restaurant density areas of Las Vegas*

4. Predicting Closure

Through predictive modeling using machine learning algorithms, an accuracy of 0.75 was achieved in predicting whether a restaurant in the yelp dataset is open or closed. The average change in star ratings over time, review frequency and count as one principal component, and restaurant density within one square kilometer are the important predictors for restaurant closure.

4.1 Feature Correlation

The correlation heatmap of the features shows that none of the features are highly correlated. Some features that don't make sense for the prediction are excluded, such as the length of time between first and last reviews, latitudes and longitudes.

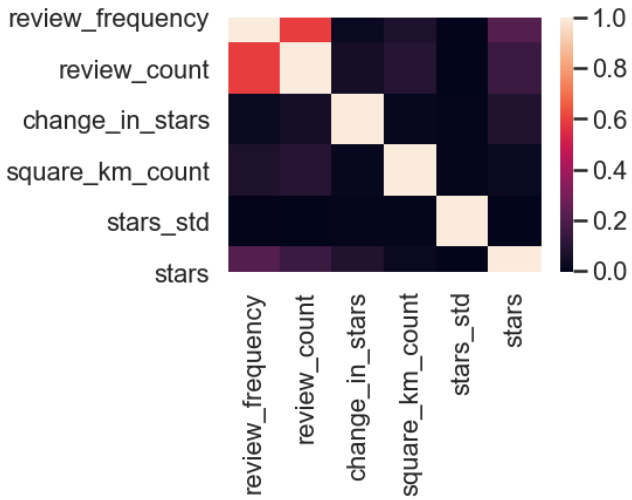

Figure 4.1 *No features are highly correlated*

4.2 Principal Component Analysis

By performing principal component analysis on the continuous features of the dataset, the dimensions of the dataset are reduced from 7 to 5. The 3 most highly correlated features - review frequency, review count, and stars are taken for PCA to be reduced to 1 principal component.

4.3 Sampling Strategy

The yelp restaurants dataset presents an imbalanced dataset when it comes to predicting whether a restaurant is open or closed. Before implementing the sampling strategy there are 24014 samples in class 1 (open) and 7690 samples in class 0 (closed). After using an over-sampling strategy from RandomOverSampling in the imblearn package, there are 24014 in class 1 (open) and 19211 in class 0 (closed).

4.4 Decision Tree

The decision tree classifier with tuned parameters provides a model with an accuracy score of 0.67 on the training set and 0.66 on the test set. With two scores that are close to each other, the model is not overfitted nor underfitted. A classification rate of 67% is considered good accuracy.

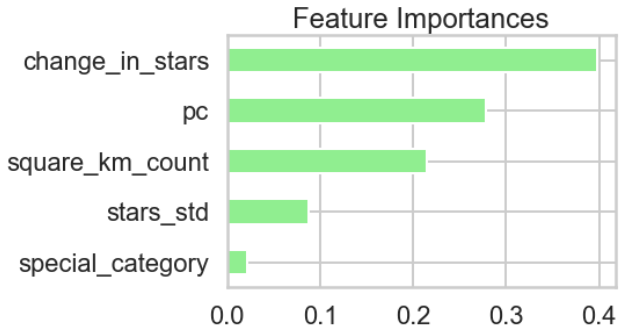

Figure 4.2 *Change in stars is the most important feature in predicting restaurant closure*

Compared to restaurants that are still open, closed restaurants have around 3 times more of a drop in star ratings over time. Closed restaurants also have higher restaurant density within one square kilometer of their locations.

4.5 Voting

Using a voting ensemble that includes decision tree, logistic regression, and K nearest neighbors, the model provides an improved accuracy score of 0.76. In particular, K nearest neighbors seems to have performed even better than the decision tree.

5. Summary

Yelp restaurant dataset contains significantly more restaurants in moderate and good standing than those in poor standing. Most restaurants have 3-4 stars in rating, with 1 star ratings being the rarest.

Yelp has over 7000 categories for restaurants. Restaurants of type Vegetarian, Vegan, Food Truck, Cocktail, Dessert, Tacos, Juice, Coffee, and Ramen have standing distributions that differ from that of the larger population. These categories contain more good restaurants than moderate and poor restaurants. Perhaps keeping it simple with a food truck or focusing on beverages is the way to go for a new restaurateur in the business.

Moderate restaurants fluctuate widely in performance over time. Good restaurants tend to perform more consistently (good) over time and poor ones perform consistently poorly. The larger variances in ratings of moderate restaurants could suggest to future restaurateurs the importance and advantage of starting on the right foot, and the difficulty of improving consistency in performance once the restaurant is up and running.

Using bootstrap inference to perform t-tests on the hypotheses around restaurants, we found some significant relationships between variables in the data. Restaurant category affects restaurant rating. Restaurant standing affects the volatility in the restaurant's ratings over time. Contrary to what one might think, price range does not seem to affect restaurant standing.

The average change in star ratings over time, review frequency and count as one principal component, and restaurant density within one square kilometer are the important predictors for restaurant closure. By balancing out the samples in the two classes closed and open, tuning hyperparameters in the classifiers, and trying out a variety of ensemble methods, we were able to predict with an accuracy of 0.75 for when a restaurant is open or closed. Interestingly, although restaurant density was not found to be related to restaurant standing through statistical data analysis, using machine learning in-depth analysis it was found to be quite influential in predicting whether a restaurant is open or closed.