Simon Li A13549622

Timothy Lue A13559509

What makes the best restaurant rating?

Abstract

We want to answer the question, what makes the best restaurant rating? To do so, we want to the explore the relationship between the type of the restaurant and its ratings. We will analyze whether certain categories or combination of categories can predict the success of the restaurant. Our analysis approach is to find a set of common categories that restaurants share, such as food type and origin, and use them to create a model to predict the rating a restaurant. We are using the Yelp Dataset to conduct this analysis. We run multiple linear regression on common categories of restaurants and their ratings. We found that some categories, such as Fast Food, have a statistically significant negative correlation to ratings, where as other categories, such as Cafes, have a statistically significant positive correlation to ratings.

I. Introduction

What makes the best restaurant rating? Restaurant rating sites allow users to post reviews and ratings about their experience at a particular restaurant. They have become a popular thing to check during the last decade, and today we can check a restaurants rating directly from our phone. Sites like Yelp have become a popular thing to check for new customers to determine if they wish to dine at a restaurant that they have never visited before. Thus it becomes essential for restaurants to strive toward getting high ratings if they want new customers. Through this research project, we are trying to determine what aspects of a

restaurant influence their rating in hopes of helping restaurants better understand and work toward higher ratings that bring in more customers.

II. Data Set

We are using the data set from Yelp, more specifically business.json as of 11/15/2018. The dataset we are using contains information about 188593 businesses on Yelp. Because we are only interested in observing businesses that are restaurants, we narrowed down the number of businesses to 41342 by only checking those that had the keyword “Restaurant”.

The data seems to be collected through compiling all the businesses that Yelp has on their site into a database that records the businesses’ information alongside their corresponding rating. For instance, the data set we are using contains business data including location data, attributes, and categories. Since the ratings are typically posted by customers who have visited the restaurant and not by any respected organization, the reliability of the rating may be a possible source of noise. Another possible source of noise may come from the promotions that some restaurants tend to hold. For example a restaurant may offer a free drink to those who place a review and those who do so might not provide the most accurate rating based on the food and service. The relevant predictor from this data set that we are using is the food type that restaurants are serving in relation to rating score. Using that information, we will be able to predict whether a restaurant having a certain attribute will tend to have higher ratings.

III. Methods

Because we want to see whether having a certain category is correlated to the rating of a restaurant, we believe that multiple linear regression is the most appropriate method. We plan

on identifying common categories, then create a categorical variable that represents whether each restaurant has this category. For example, “Pizza” is a common category that many restaurants have. We added a variable called “Pizza”, and set its value to 1 for restaurants who are in this category, and 0 to restaurants that do not. We then run multiple linear regression on this data, with a list of 0 or 1 values representing whether a restaurant is of some category as the predictor, and the rating as what we want to predict.

IV. Hypothesis

We hypothesize that the category of a restaurant will account for some statistically significant, but not a majority of the variance. While we expect the category of the restaurant to have some correlation to its rating, there are still many other factors that can influence the rating, many of which are difficult to quantify.

For specific categories that will have a significant positive or negative correlation to the rating, we believe that restaurants with fast food related categories will typically have lower ratings. This is because we believe that ratings are reflective of the quality of food and service at restaurants. For fast food restaurants, we typically expect convenience and low price over quality. For categories that have positive correlations, we are less confident about our predictions. We want to say that niche categories, perhaps with a focus on health or sweets, should have higher ratings.

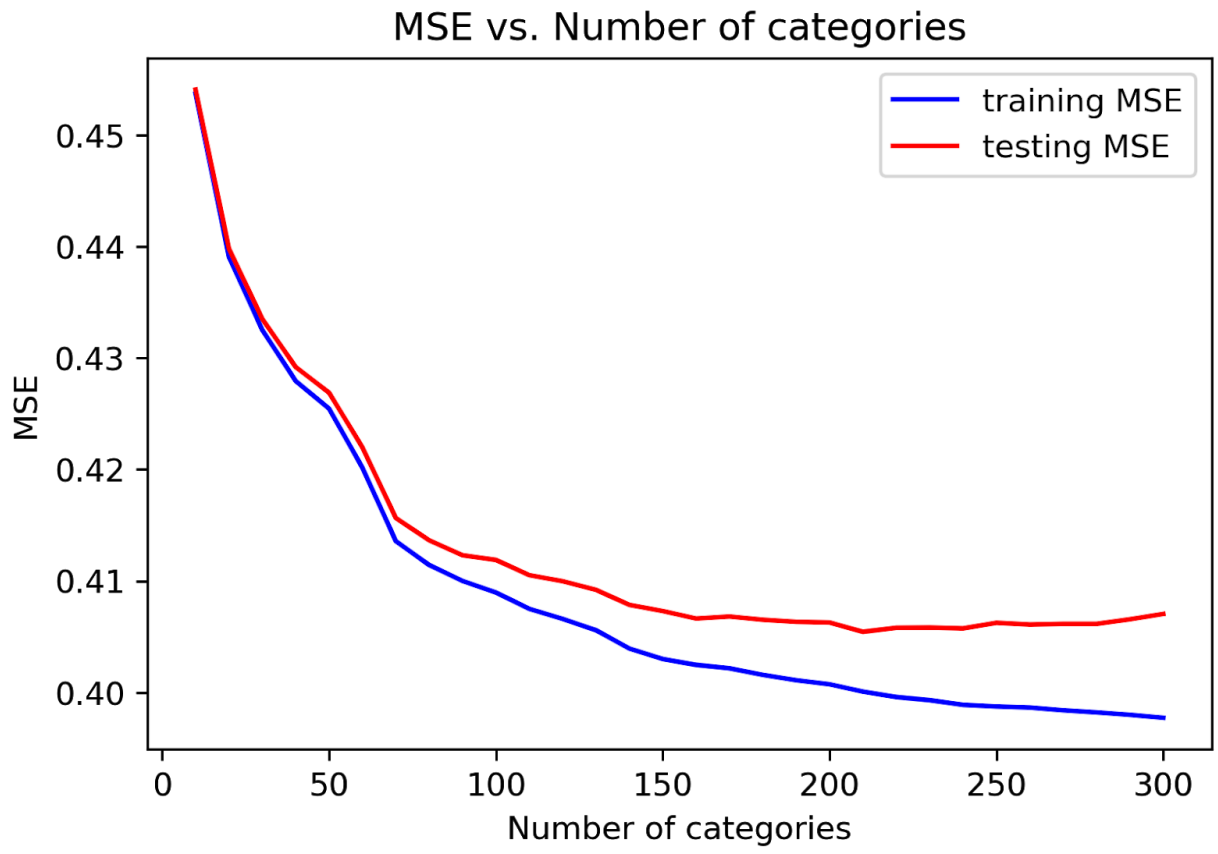
We also hypothesize that the more category we use as predictors, the more accurate our model is. However, we believe that after a certain number of category is used, increasing the number of predictors used does not reduce the MSE further, and may even reduce it.

V. Model Selection

One variable in our model is the number of categories to use as predictors. We hypothesized that as the number of categories increase, our model will improve initially, but we are unsure about whether using too many predictors will worsen the model's accuracy.

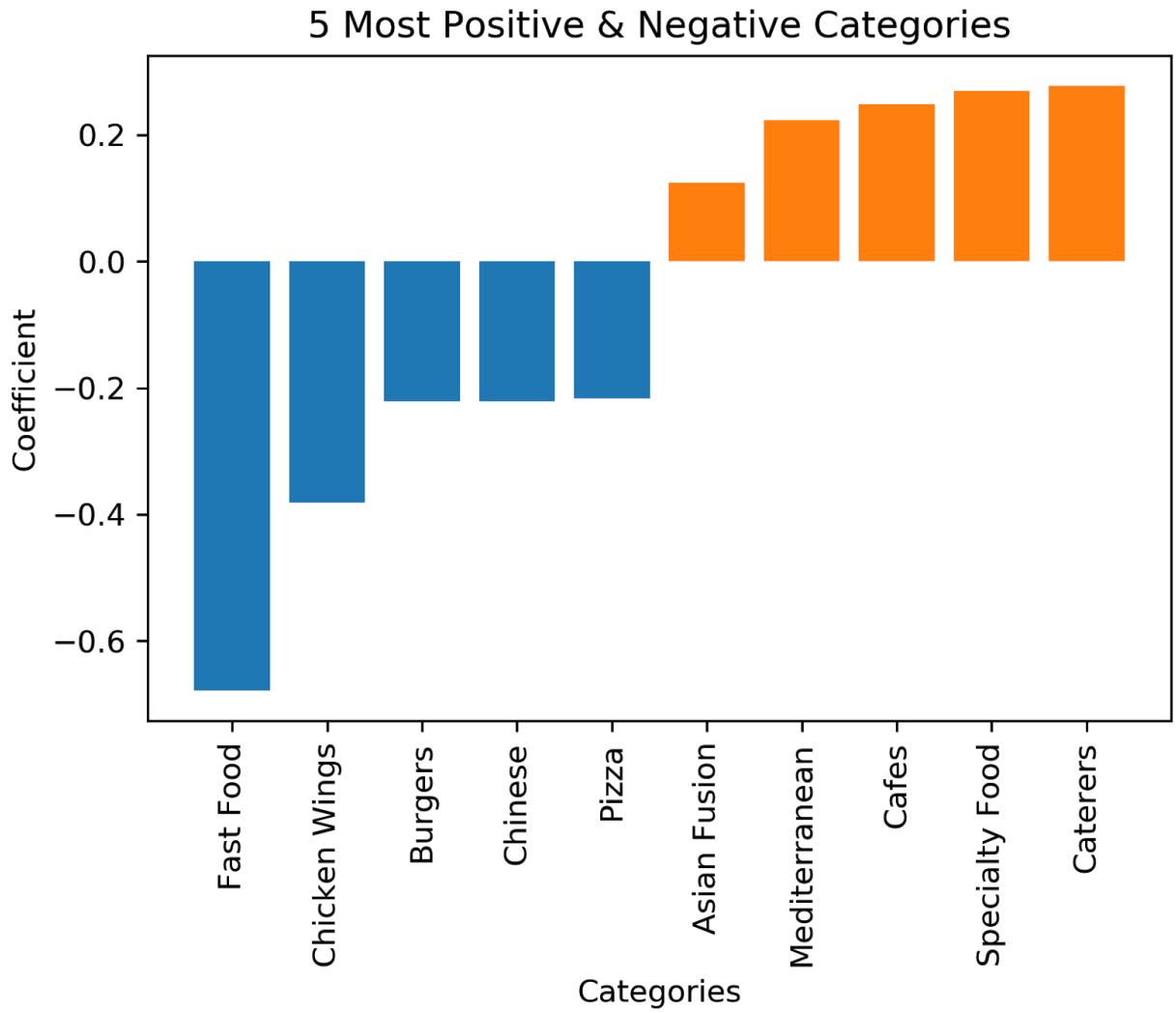
To determine this, we used K-fold cross validation to measure the testing MSE of our model when an increasing number of category is used as predictor.

VI. Model Estimation



Using K-Fold cross validation, we can see that the testing MSE of our model decreases rapidly as we increase the number of categories initially. As we reach around 200 categories,

the testing MSE starts plateauing, and starts increasing as we reach 300 categories. This means that the optimal number of categories to use is around 200.



This graph shows the 5 categories that have the most positive/negative correlation to the restaurant's rating according to our model. For this model we used the 200 most common categories as predictors.

VII. Conclusion

Our model has a r-square value of around 0.33, which means that about 33% of the variance in the ratings can be explained by our model. This is roughly what we expected, as we expect that the category of a restaurant should be correlated to its rating, although there should be other factors that have significance.

The most positive and negative categories all have p-values close to zero, which means that we are confident that the these categories do have a statistical significant correlation to the ratings.

Our results above confirmed some of our hypotheses, such as the hypothesis that the category of a restaurant will account for some statistically significant. Most notably, the category “Fast Food”, and categories that people generally associate with fast food, such as pizza, burgers, and chicken wings, have the most negative impact on ratings. This result makes sense because we expect low price and convenience instead of quality from these types of restaurants. In contrast, the categories “Caterers” and “Cafes” have the most positive impact on ratings. This result also makes sense since we expect these places to be service oriented, and service is typically important to a customer’s rating of a restaurant. Casual relaxation at cafes that help make customers’ experiences there positive. Other possible extensions of this research is to incorporate the price range of a restaurant, hours of operation, or location into account. We suspect that there is a high likelihood of a correlation between the price range and ratings. Howerever, despite seeing the price range represented as the number of dollar signs on yelp’s website, this variable was not available in their dataset, thus we were unable to incorporate it into our current Research Project.

VIII. Data Set Link

Yelp Dataset, business.json, https://www.yelp.com/dataset