STAT 628: Module 3

Yinqiu Xu, Xiaofeng Wang, Milica Cvetkovic

2020-11-29

Introduction

Yelp is a powerful platform for business owners to collect information about their business and data about customer satisfaction. In this project, we analyzed the customer's reviews and various other attributes of restaurants, trying to extract useful information and give suggestions to businesses. Our analysis focuses on all the pubs that have a full bar service in Wisconsin. Among these pubs, our specific goals are to analyze the most important facilities and services that customers prefer and then provide suggestions to pubs to improve their star ratings. The rating of a restaurant is measured in stars from 0 to 5. We want to answer the following questions:

1. What facilities will affect businesses' rating? How will the rating be affected?
2. How do the work hours influence the rating?
3. What additional services could improve the rating? For example, parking, valet, garage, etc.

For this project, we used a real data set from Yelp. We first conducted an exploratory data analysis. Then, we built a multiple regression model by stepwise selection to predict the rating of a business. In the following sections, we cover the details of our analysis and the model.

Preprocessing

Data and Sample Size

The Yelp dataset is released by Yelp to encourage students to do research on it. All the data are stored in four json files which contains information about business, written reviews, which are longer comments and the tips, which are shorter comments, and the data about the user who wrote the review or the tip. We mainly focused on the following features: stars, open hours and facilities/attributes data in business json file, and the content of reviews in review json file. After filtering all the open restaurants with a full bar service in Wisconsin, we got 466 pubs with 69 attributes and its stars ratings, 50569 reviews, and the corresponding business id.

Clean Attributes Variables

To obtain attributes of each pub, we separated BusinessParking, GoodForMeal, and Ambience into several binary variables. Some redundant characters are deleted from levels of attributes factor(I don't know what this means). For example, "u'free'" is the same as "free" in attributes Wifi. We deleted the redundant "u'". Also, missing values of both nominal and ordinal variables are interpolated by their mode.

Predictors

A pub's work hours are the predictors that calculated by hours.Monday-hours.Sunday attributes. (I don't understand this sentence) To get the information about reviews' sentiment of each pub, we created a new predictor called positive review ratio. We did this by dividing the number of positive reviews by the number of reviews of each pub. The sentiment analysis is used to evaluate the polarity score of a review. First, we parsed each review into single words and deleted stopwords like he, she, the. Then, we counted the number of positive words and negative words in each review. Finally, we calculated the sentiment of each review by subtracting the number of negative words from the number of positive words. The sentiments of each word are defined and stored in a lexicon which is provided by the *tidytext* library in R. After conducting sentiment analysis of each review, we grouped the reviews by business id to get the number of positive reviews. The two separate datasets, the review dataset and the business dataset, were merged by business id column.

EDA/Statistics Tests

Plots

The plots in this section show the proportion of a specific attribute of a business to the stars rating. Note that most of attributes are discrete. We plotted a histogram showing the distribution of the stars for each attribute.

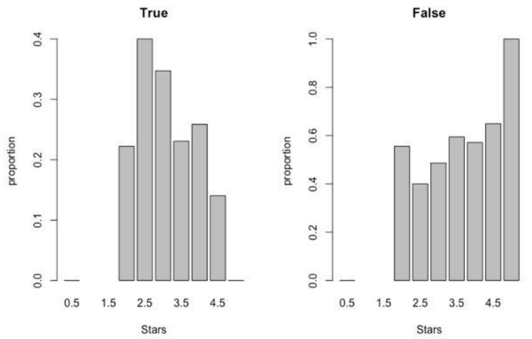

Figure 1: Delivery

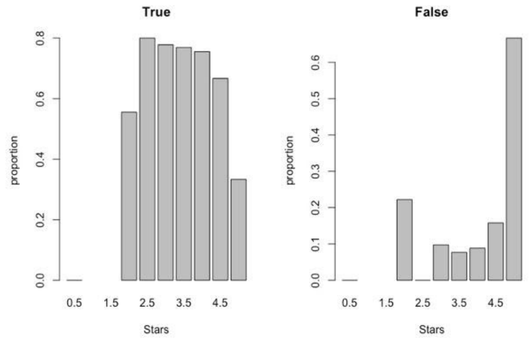

Figure 2: Takeout

From the figure 1, we can see that whether or not a business offers a delivery service may not make a big difference in the business rating. From the Figure 2, we can see that takeout service could not be a competitive power of a business. Businesses without the takeout services can still have high rating.
We did similar exploration looking into whether a bar is good for groups. We discovered the businesses that are good for groups had almost uniformly distributed star ratings. On the other hand, the bars that were not good for the groups also seemed not to have any meaningful stars distribution. In addition, we looked into the affect that the presence of TV might have on the stars rating. Interestingly, it seemed like, in Wisconsin, the businesses with the TVs received lower stars, than the businesses without the TVs. This led us to the conclusion that the presence of TVs have no significance and we did not continue any further exploration.
Then we explored the plots that showed us the frequency of each word in the reviews and how it was related to the distribution of the stars. As an example, here are the graphs showing the star distribution based on the various alcohol beverages mentioned in the reviews. It seems like the bars that offer absinthe, brandy, beer and cocktails tend to have higher star ratings.

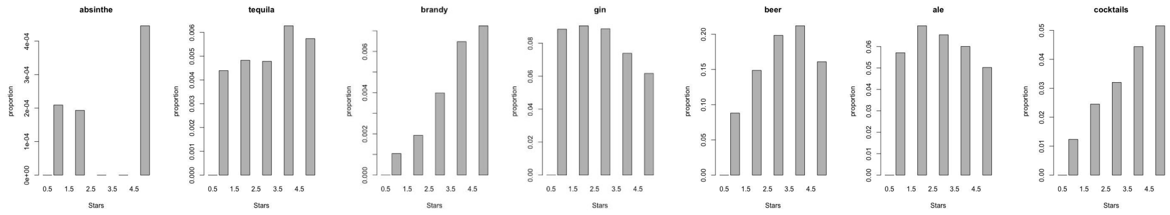

Figure 3: Figure 4: Figure 5: Figure 6:

Finally, we looked into overall distribution of the star ratings among all the restaurants and bars in Wisconsin, and the distribution of the stars among each review. The overall ratings seemed to be somewhat normally distributed for the ratings between 2 stars and 4.5 stars. It seemed like there are no really bad ratings that are less than 2 stars, and the number of ratings that are 5 stars is very small. When we look at the number of stars each reviewer gave, it seems like in general the users tend to give more positive than the negative reviews. In terms of suggesting the concrete feedback, this might be an issue if we find that we don't have enough negative reviews to suggest improvements.

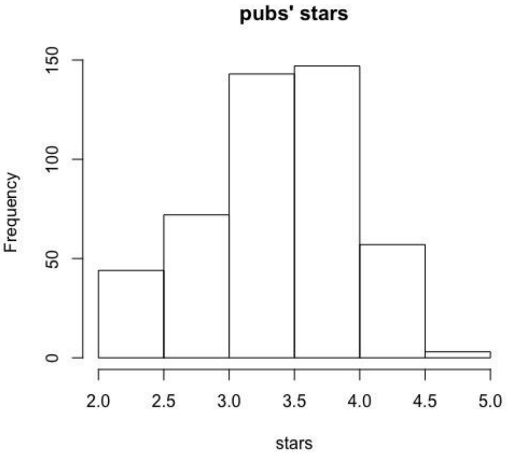

Figure 7: Stars distribution

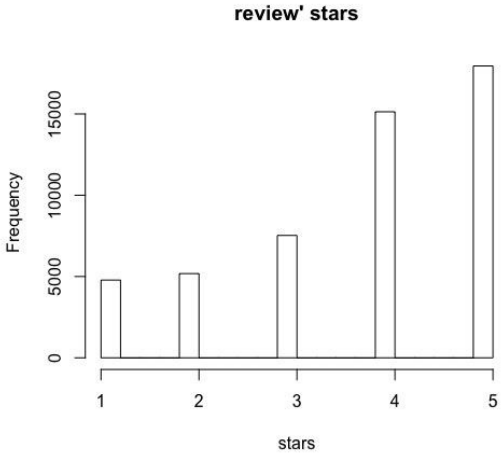

Figure 8: Stars distribution in each review

Key Findings, Statistical Analysis and Model Diagnostics

Data-Driven Business Plan

This section should be more than a page according to the professors instructions

Conclusion