STAT 628 Module 3 Summary

Marwan Lloyd, Xingrong Chen, Yuling Lyu

1   Introduction

To some extent, ratings and comments reveal why people like or dislike a coffee shop. In this analysis, we focus on using data from Yelp which contains information about coffee shops in Pennsylvania (PA) and comments they receive from the customers to figure out what makes some coffee shops more attractive than its competitor. Based on our analysis, we established a set of recommendations that might improve their business or at least increase their ratings on Yelp.

2   Data Cleaning

The data we are using consists of four JSON files which contain reviews people made on Yelp and business information of restaurants all across the U.S. To study comments people made on coffee shops, we first converted JSON files to data tables and join the tables together based on unique identifiers, then drop the columns we are not interested in and filter out all non-coffee businesses, finally, we narrow down our analysis to locations in PA , since the remaining data we got after the first two steps' cleaning are mostly in there.
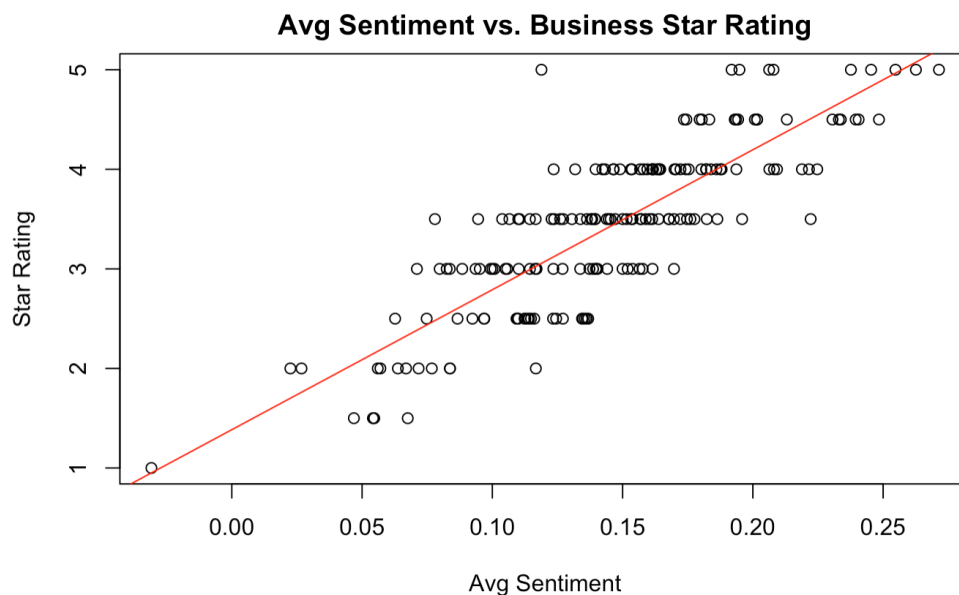

Figure 1: Scores v.s. ratings

3   Data Analysis

Sentiment analysis is used to let computers simulate human reading, i.e. text reviews and respond with whether there was a positive or negative impression of the reviews and divides reviews into

1

three categories (positive, negative, and neutral). To conduct sentiment analysis, we used the SentimentAnalysis package in R which utilizes the existing dictionary QDAP to give every review a sentiment score.

Figure 2: Sentiment Score Distribution

According to the above plot, we got a sense that coffee shops that receive high ratings do not necessarily receive stronger positive reviews, but the high ratings do indicate that they receive fewer strong negative reviews.

In addition, we fitted ratings by average sentiment scores, the model shows that the average sentiment score is a good predictor for ratings. On average, a 0.1 increase in average sentiment scores will increase the ratings by 1.4. And adjusted R^2 shows that 70% of variation in ratings is explained by the average sentiment score. Thus, we expected review content and sentiment scores would help us explain the ratings well. So based on the sentiment score, we categorize the reviews into positive, neutral, and negative. In each of the categories, we exploited the top-mentioned ideas in the comments to infer what might be the reasons why people made positive or negative reviews.

4 Key Findings About Businesses

Firstly, we found that average review sentiment score is a good predictor for location star rating. In general, a 0.1 increase in average sentiment scores will increase the ratings by 1.4.

Secondly, we found that bad reviews usually focus more on service issues like customer service, and pre-food reception, while good reviews take good service as a given and focus more on added benefits like atmosphere and environment. In our application, we put together lists of the 25 most important positive and negative words from their respective reviews. Some of the most frequently mentioned positive words were "delicious, selection, hot, cute, fast, and cozy" while for negative reviews we saw words like "long, cold, rude, wrong, and slow" near the top.

	Positives	Positive Frequency	Negatives	Negative Frequency
4	hot	12	wrong	1
1	time	10	long	5
3	favorite	10	rude	1
2	delicious	9	cold	4
9	sweet	8	hard	4
15	friends	8	disappointing	4
18	fresh	6	bitter	0
7	cute	4	ill	3
16	helpful	4	miss	0
20	cool	3	limited	3

Figure 3: An example of the top words frequency

What is more, minimizing bad reviews has a better effect on raising ratings than getting more positive reviews. Coffee shops with high ratings usually receive fewer strong negative reviews, but do not necessarily receive stronger positive reviews. Thus, through the table shown above, we allow coffee shop owners to drill down into the heart of what is causing them to lose stars via negative reviews (on the right highlighted in orange), and also see what they are doing correctly (highlighted on the left in green)

5 Recommendations for Businesses

5.1 Recommendations

From our analysis above, we have the following suggestions for all coffee shop owners.

1. Customer reviews play an important role in improving coffee shops' ratings, which suggests that owners should pay more attention to them.

2. Since locations with higher reviews did not necessarily receive stronger positive reviews, while they did receive fewer strong negative reviews, coffee shop owners should work hard on minimizing bad reviews by solving customers' problems and complaints, which has a better effect on raising ratings than getting more positive reviews.

3. Good service should be taken as a basic requirement that prevents a coffee shop from receiving many complaints while providing some additional benefits such as a romantic atmosphere and a good environment that seems more effective in receiving good reviews thus raising the ratings on

Yelp. Ways to do this include:

- Ensuring adequate staffing at peak rush hour times (typically early morning)

- Early preparation of stocked items, many positive reviews appreciated abundance while negative reviews focused on emptiness and scarcity

- Ensuring staff stay up to date on professional standards and customer experience management to minimize poor customer experiences from employees

5.2 Limitations of recommendations

Since we narrowed down our analysis to locations in PA, it may lack the ability of applicability to the coffee shops that located in other areas. Besides that, due to how variable coffee locations can be for a variety of reasons (from intended audience to food served alongside the coffee), these recommendations are not as specific specific as might be found in recommendations for another food industry.

6 Conclusion

In this study, we conducted sentiment analysis and statistical testing on coffee shop ratings to help us determine what makes a coffee shop receive higher ratings on Yelp. By using the sentiment analysis, we gave every review a sentiment score and got some key findings about businesses, based on which we built a Shiny app to visualize our analysis and make it easier to understand. Finally we provided some recommendations for coffee shop owners so as to raise their ratings on Yelp. Through the determination of the most influential positive words in positive reviews, and the most impactful negative words in negative ones, we reached the conclusion that the cut-off between postive and negative reviews is receiving of food.

A   Contributions and References

Contributions:

Marwan Lloyd: Created code related to sentiment analysis and model development. Wrote the Shiny app and maintained GitHub repo. Wrote/edited parts of the summary. Contributed to data cleaning. Contributed to the slides.

Xingrong Chen: Created code related to data merging. Maintained GitHub repo. Wrote/edited the Introduction, Data Cleaning, Data Analysis, and Key Findings parts of the summary. Contributed to the slides.

Yuling Lyu: Worked on slides for the presentation. Wrote/edited the Data Analysis, Key Findings About Business, Recommendations for Businesses, and Conclusion parts of the summary.