A Research on Yelp.com – What Affects Restaurants' Ratings?

Fangfang Wan
Masters in Computational Social Science, Social Science Division, University of Chicago (https://macss.uchicago.edu/)
Email: fangfangwan@uchicago.edu Phone: (+1) 607-761-8398

Background and Introduction

- Previous literature indicates strong positive correlation between a business's customer review and revenue, both in traditional "viva voce" (oral) form (Mahajan et al., 1984) and more recent online reviews (Chevalier and Mayzlin, 2006).
- Geological location is the main factor that I will further analyze in this research. "Contagion effect" is a concept in psychology (Hartfield et al., 1993), and it has been deeply investigate in behavioral economics, especially in behavioral finance (Hirshleifer and Teoh, 2003).

Methods

- Data collection – web scraping with *BeautifulSoup* library in Python. Collected data of top 1000 restaurants in Chicago.
- Model building
  - Basic Model: linear regression with *sklearn* in Python with star rating as dependent variable
  - Build restaurant pairs to analyze contagion effect (about ½ million pairs
    - Compute cosine similarity of each pair
    - For each pair, compute the correlation between 1. star rating difference and geographical distance (calculated from latitude and longitude), and 2. cosine similarity and geographical distance
- Validation – K-fold cross validation

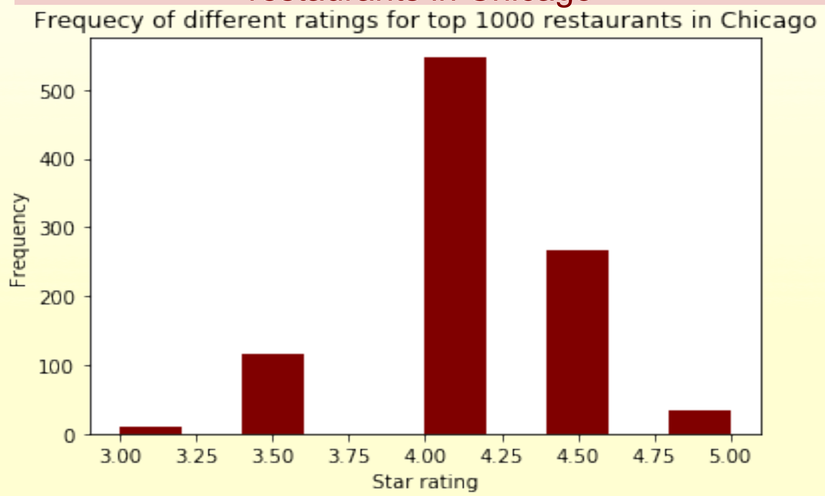
Research questions

- What affect a restaurant's Yelp.com rating stars?
- Is there "contagion effect" in catering industry?

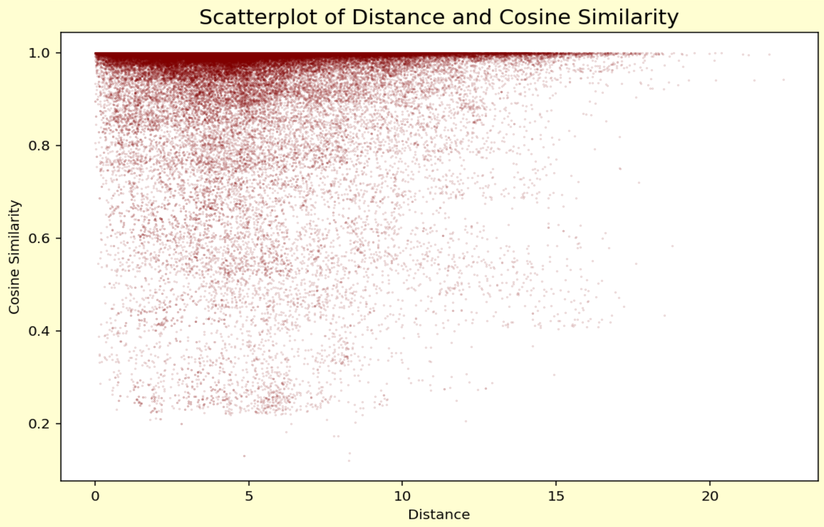
Results

- In linear regression, I include number of reviews, operating hours throughout a week, neighborhood dummy variables and about 20 attribute dummy variables (including ambience, parking availability, loudness, category, etc.)
- Features to notice:
  - Restaurants in some neighborhood performs better, (e.g. West Town)
  - Higher price is associated with lower rating.
  - The louder the restaurant, the lower the ratings.
  - R-squared: 0.325, which is reasonable, since there are some outside features that are not included in Yelp.com but can affect star ratings (e.g. taste of customers).
- Concerning contagion effect, the correlation between geographical distance in km and star rating in each restaurant pair is 0.051.
- To analyze the reason, I compute the cosine similarity of each pair, and found out that the correlation between cosine similarity and geographical distance in km is -0.055
- K-fold Cross validation with K=4 shows that the linear regression yields mean squared error of 0.1.

Frequency of different ratings for top 1000 restaurants in Chicago



Distance and cosine similarity



Conclusion

- There are some attributes of a restaurant that are associated with its rating, for example, neighborhood, what type of food, operating hours, etc.
- There exhibits some contagion effect among restaurants in Chicago. Restaurants that are closer in geographical distance are more similar, and they also have closer Yelp.com ratings.
- There are still some factors which affect the ratings, but they are hard to collect and/or quantify (e.g. taste of customers)

Limitations

- Omitted Variable Bias – some factors cannot be easily collected and/or quantify
- Only included top 1000 restaurants in Chicago. Other restaurants in Chicago and restaurants in other areas may exhibit different behaviors.

Acknowledgements

Thanks Dr. Benjamin Soltoff for providing help and suggestions on my research project.

References

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. Journal of marketing research, 43(3), 345-354

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1993). Emotional contagion. *Current directions in psychological science*, *2*(3), 96-100.

Hirshleifer, D., & Hong Teoh, S. (2003). Herd behaviour and cascading in capital markets: A review and synthesis. *European Financial Management*, *9*(1), 25-66.

Mahajan, V., Muller, E., & Kerin, R. A. (1984). Introduction strategy for new products with positive and negative word-of-mouth. Management Science, 30(12), 1389-1404