Data Analysis based on Yelp Dataset

With the development of modern technology and the Internet, more and more people are using online review software to rate and comment on businesses. As the largest online review app in the United States, Yelp has collected ratings and reviews from more than 150,000 businesses.

In this project, we only focus on analyzing elements that impact bar ratings in the US. Many factors can affect the extent of overall comments in bars as social spaces. By analyzing factors together with open hours, certain attributes, and customers' comments, we sought to determine the ability effect of those elements on bar rating. We will provide facts-driven insights for bar proprietors and managers to optimize their commercial enterprise approaches and improve ratings. Following the same logic and steps of the process, we can also apply our model to other business categories in the future.

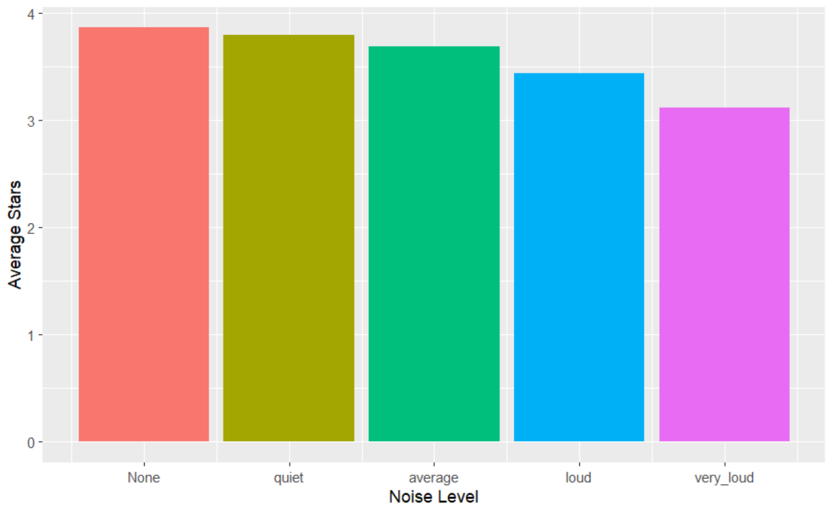
1. Exploratory Data Analysis

The reason we choose bars is because bars are social places and would provide important insights into consumer behavior and create a safe environment, which is important for understanding people's choices and preferences in social settings. We consider any business successful if it has a mean rating of 3.5 or greater over 100+ reviews. Any business fitting this criteria is 150% more likely to remain open as compared to businesses that don't fit the criteria based on our logistic regression modeling.

2. Key Findings About Businesses

2.1 Suggestions Based on Business Attributes

At first, in our data exploration, we find the noisier one bar is, the lower average rating it has according to figure below. Thus, we decided to explore how attributes impact ratings and which attributes should be improved so as to make business more successful.



To do this, we need to figure out which attribute is the most significant. In order to get general results, we chose attributes of which more than 70% bars have. For missing value, we simply assumed this bar doesn't have this attribute. Setting attributes as our feature vectors and rating as our output, then we applied four basic machine learning models to do predictions and checked which model had the highest score.

From the figures below, we can see decision tree performs best no matter whether we did standardization. So, we use Jacknife method based on decision tree model to look for the most significant feature. From figure below, we can see that score reduced the most when we remove \`NoisyLevel\` feature no matter whether this bar is opened or not, followed by \`PriceRange\`and \`Alcohol\`, which means these three features are the first three most significant attributes of bars. Thus, if one bar wants to be successful, the first most important task for the owner is to provide a comfortable eating or drinking atmosphere, choose a proper price range and think highly of alcohol quality and types.

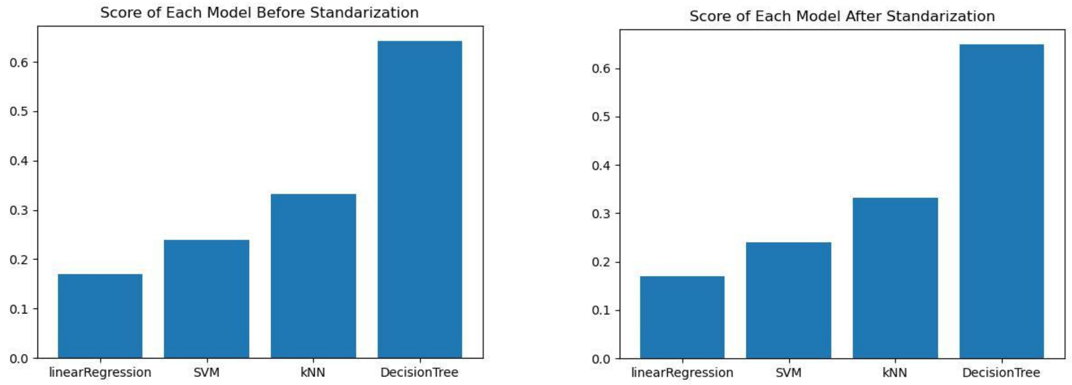


Figure 2.1: Score of Machine Learning Model Respectively

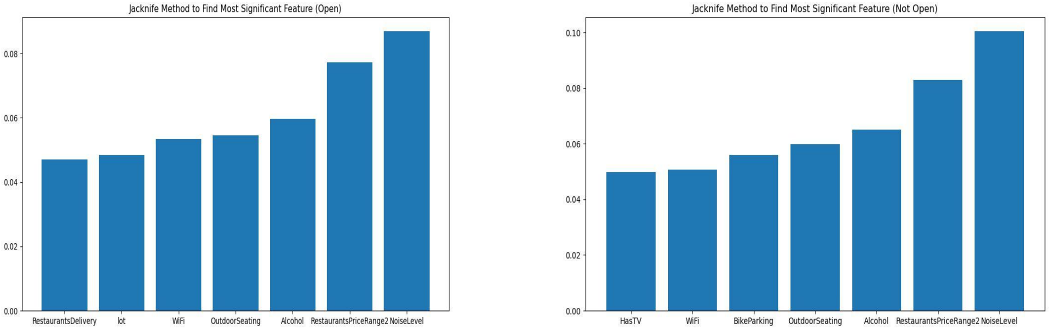


Figure 2.2: Reduced Score [Opened Bars (Left), Closed Bars (Right)]

Moreover, we can find the `Bikeparking` feature becomes less significant while the `lot` feature becomes more significant due to the development of transportation. We also noticed that Wifi` and `Delivery` features become more significant, which reflects two trends below:

- With the development of technology, people have become more dependent on smartphones. So, they probably are more prone to choose a bar that could provide wifi while eating, especially when reception is poor.
- As the speed of our daily life becomes higher and higher, people become more willing to wait for others to deliver meals instead of coming to bars and taking food or alcohol out, which could help people squeeze more time into work or study.

Therefore, bars should also sensitively spot changes in time and then make improvements to catch up with the tendency in time if they want to be more successful.

2.2 Suggestion Based on Review Words

To be more practical, we used both the word cloud chart and the LDA model to analyze reviews and then make specific suggestions against one bar. The reasons why we chose these two methods are below:

- As a visualization method that allows readers to see which words appear more frequently in a bunch of text, a word cloud chart helps us extract general customers' comments quickly.
- Since the basic idea of the LDA model is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words, we divide all reviews into three topics. One is positive, one is neutral, and the other one is negative. We assume that the more negative words this topic has, the more negative this topic is, and vice versa. Thus, we can

classify which result belongs to which topic according to the beta value of emotional words, such as "bad", "happy", "delicious" and so on.

Take Elsa's Mexican Restaurant as an example. The figures below show the results of the word cloud and LDA model used to analyze reviews of this bar.

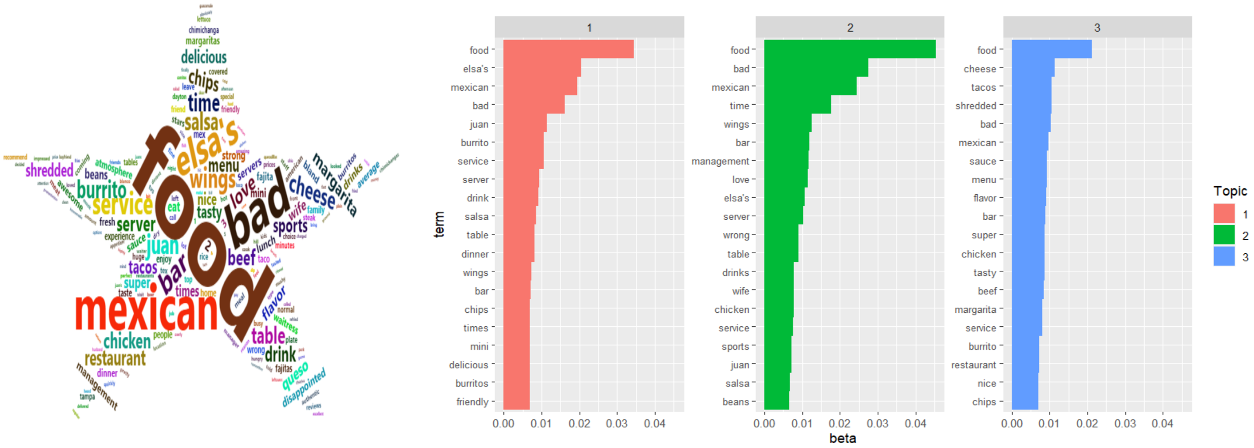


Figure2.3: Word Cloud (Left) and LDA Result (Right) of Elsa's Mexican Restaurant's Reviews

In the word cloud chart, we can see that "bad" appears pretty frequently, which means that this bar needs to reconsider their food fresh, food tastes or customer service and make relevant improvements. Then, we can extract specific suggestions from the results of the LDA model.

According to the beta value of the word "bad", we can say topic 3 is the positive topic and topic 2 is the negative topic. On a positive note, "cheese" and "tacos" are mentioned more, which means they are two of the most delicious foods of this bar. That is to say, when a person visits this bar for the first time, he can order a taco without concerns. Oppositively, "wings" are mentioned more in negative reviews, which means this bar should improve their present ways of making wings.

2.3 Suggestion Based on Open hours

To analyze how open hours affect bars' ratings, we did a series of steps. First, We divide each day into 48 time periods with each half-hour as a period, covering from Monday to Sunday. Next, we used the open hours information of each bar, mapped it to the corresponding period, and marked the open or closed status of each period with 0 or 1. Furthermore, we classified the date of each review and assigned it to different periods from Monday to Sunday. Then, we calculated the average ratings for each period from Monday to Sunday, respectively.

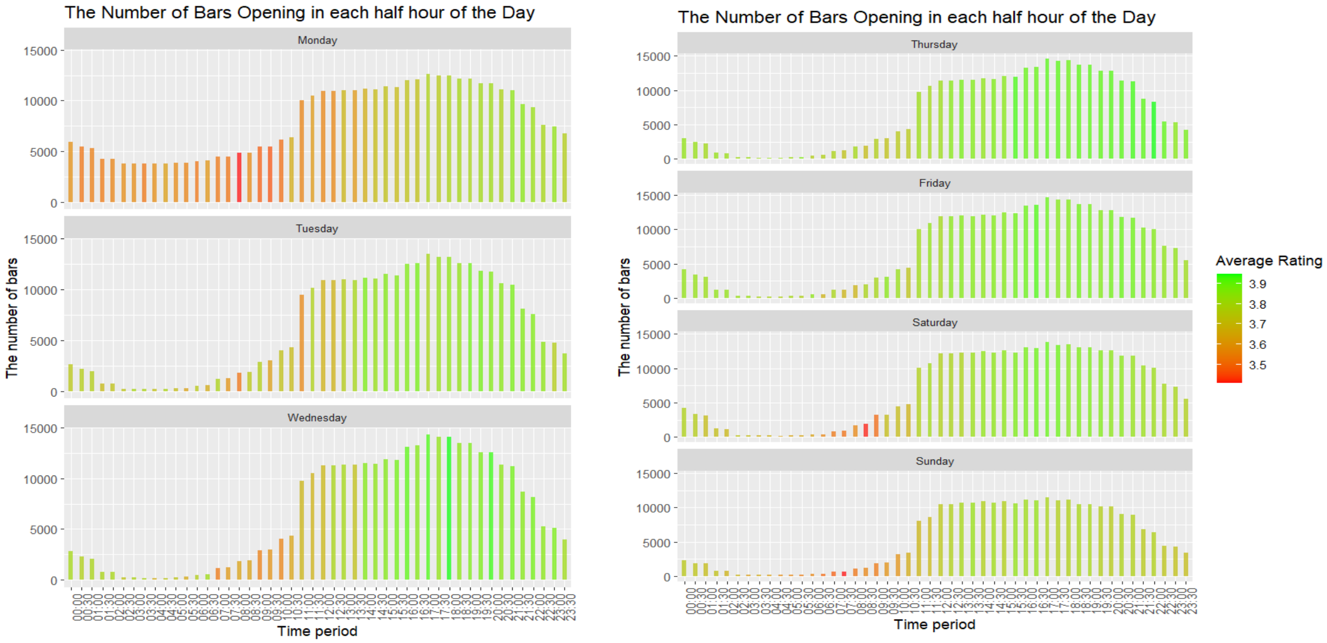


Figure 2.4 The number of bars opening in each half hour of the day in a week

After analyzing the relationship between opening hours and ratings, we observed some interesting findings. In Figure 2.4, we can see that afternoon and evening have more bars opening and higher ratings. Also, there are many bars open after Monday midnight but with a lower average rating. After that, we build a decision model which shows as the following plot :

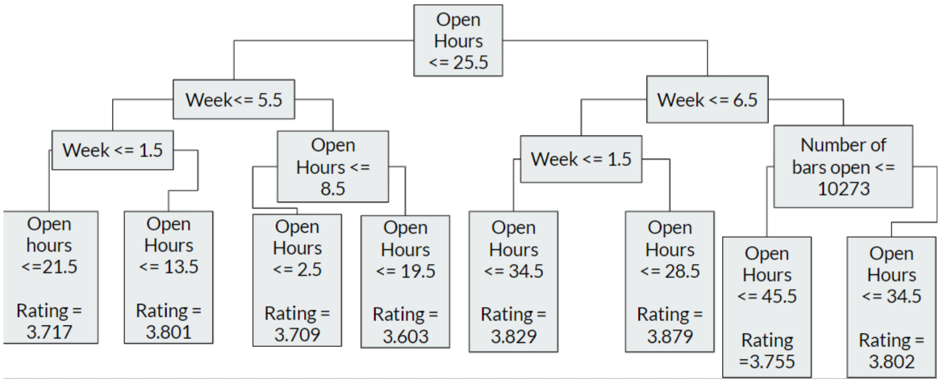

Open Hours <= 25.5
Week<= 5.5
Week <= 6.5
Week <= 1.5
Open Hours <= 8.5
Week <= 1.5
Number of bars open <= 10273
Open hours <=21.5 Rating = 3.717
Open Hours <= 13.5 Rating = 3.801
Open Hours <= 2.5 Rating = 3.709
Open Hours <= 19.5 Rating = 3.603
Open Hours <= 34.5 Rating = 3.829
Open Hours <= 28.5 Rating = 3.879
Open Hours <= 45.5 Rating =3.755
Open Hours <= 34.5 Rating = 3.802

Figure 2.5 Decision tree

As we can see from the above decision tree plot, “week” means week days, the number corresponding to Monday through Sunday with a representation of 1 to 7. For example, “week” less than 1.5 means open on Monday. Open hours means the starting hours, and number corresponding to each half hour of the 24 hours in a day. For example, “open Hours ” <= 2.5 means the bar opens before 1:15 am. The model score is 0.85 which is high enough to say this is a good model. We suggest that the bar managers take this model as a reference when setting the open hours.

We also mapped the mobility data grouped by state to evaluate the top 3 busiest days in terms of traffic which coincides with higher ratings. These days have around 9% - 16% more trips taken than other days. Based on the operating days for a business we check if their business is open on high traffic days.

3. Conclusion:

Overall, for all of bars in the U.S., we found customer would think highly of noise level and price range when they visit a bar. Also, since the afternoon and evening were times when more bars were open and had higher rating, we recommended that bars consider opening on these busy times and days to obtain more customers and better ratings.

As mentioned before, our final target is we want to apply our model to all categories of business. So, for any category of business (denoted as $\mathcal{H}$), based on our assumptions, we can first analyze which attributes are significant and what time has more customers through decision tree model. Then we can conclude which aspect $\mathcal{H}$ needs to be improved according to beta values of typical emotional words through LDA model.

4. Contributions

Contributions	Feiyun Yan	Xingyu Tang	Kai Shukla
Presentation 1	Responsible for slides 4-7. Reviewed/edited slides and provided feedback	Responsible for slides 12-16. Reviewed/edited slides and provided feedback	Responsible for slides 2-3, 8-11. Reviewed/edited and provided feedback on all

	on all slides.	on all slides.	slides.
Presentation 2	Reviewed/edited slides and provided feedback on all slides.	Reviewed/edited slides and provided feedback on all slides.	Reviewed/edited slides and provided feedback on all slides.
Summary	Responsible for introduction, analysis for open hours. Reviewed/edited and provided feedback on whole document.	Responsible for analysis for attributes and review words. Reviewed/edited slides and provided feedback whole document.	Responsible for one of EDA parts. Reviewed/edited slides and provided feedback whole document.
Code	Responsible for EDA and analysis for open hours code. Reviewed code for analysis section.	Responsible for data cleaning, feature engineering of bars attributes and bars reviews, building proper model to analyze.	Responsible for analysis for analyzing relationship between rating and zipcode. Did data exploration of open hours in IL state.
Shiny App	Reviewed and provided feedback on Shiny App.	Build a small shiny related to reviews. Reviewed and provided feedback on Shiny App	Responsible for Most part of Shiny app.

5. References

[1] https://krisrs1128.github.io/stat992_f23/website/docs/2022/06/02/week12-2.html