Module3 Summary

Yutong Zhang, Jiaqi Xia, Midhun Satheesh

1. Introduction

a. Background Information

In this project, we used the Yelp dataset, focusing on fast food restaurants in the US. We wish to give some suggestions to owners of fast food restaurants to improve their star ratings based on closed fast food restaurants and open fast food restaurants with low star ratings (stars ≤ 2). Our work is divided into two parts: attribute analysis and review analysis. For the first part, we conducted a **logistic regression** to find the attributes causing the closure of a fast food restaurant. For the second part, we cleaned the reviews and conducted a **Chi-Square Test** to test the independence between customers' attitudes and different food. Finally, we combined our findings from these two parts and gave our conclusion and suggestions for fast restaurants.

b. Data Cleaning

The Yelp Dataset consisted of 150,346 entries of businesses and 6,998,280 entries of reviews. Our goal was to explore the attributes and reviews of the fast food restaurants in the US, so we only use the business.json file and review.json file. We conducted the following steps to filter the data:
(1) selected all the states in the US and kept the businesses in the US for our analysis, (2) extracted businesses containing "Fastfood" in the "categories" variable, and (3) selected the fast food restaurants with the low star ratings and put them into two files, one for open and the other one for closed restaurants. Finally, we merged these two files with the corresponding reviews and got our final dataset, which contains 910 restaurants and 28,272 reviews for closes, 2,238 restaurants and 57,354 reviews for opening respectively.

c. Exploratory Data Analysis (EDA)

i. Barcharts to compare star ratings

Barplots can provide us with intuitive visualization. We plotted the distribution of star ratings for each selected attribute and compared the differences in distribution between closed and open fast food restaurants with low star ratings. Figure.1 shows one example:

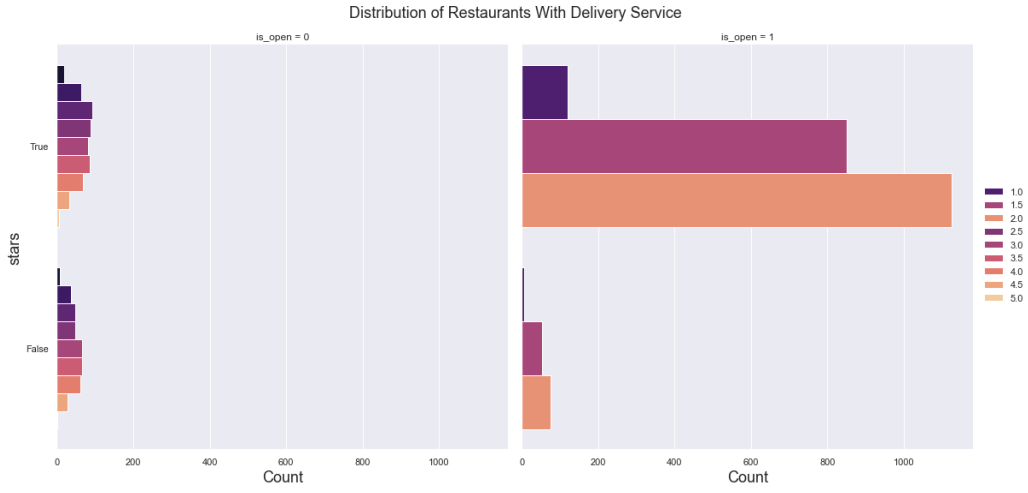

Figure.1

We can conclude from this plot that most open fast food restaurants provide delivery service while only half of the closed delivery service provide delivery service. One suggestion we can give fast food

restaurants is that if a fast food restaurant provides delivery service, there is a high possibility for it to survive.

ii. Word Cloud

After the initial data cleaning, we cleaned the review for each business by following these steps: (1) tokenization, (2) removing stopwords, (3) stemming, and (4) lemmatization. Here, Figure. 2 shows two word clouds for two scenarios for closed restaurants. The first is one about all the positive words from the reviews while the second one is from the negative words.

Figure.2

2. Attribute Analysis

a. Goal

Our goal is to explore which attributes affect whether a restaurant is closed or not.

b. Data Preprocessing

We analyzed the attributes provided by fast food restaurants using the "business.json" file. It turns out to be 38 attributes in total. However, not all businesses provide all 38 attributes, so we only selected attributes that are provided by at least 80% of businesses in order to avoid bias in our analysis, which are: **RestaurantsDelivery, RestaurantsTakeOut, RestaurantsPriceRange2, BusinessAcceptsCreditCards, RestaurantsReservations, GoodForKids, RestaurantsGoodForGroups.** After attribute selection, for rows without these attributes, we infused all missing values in the RestaurantsPriceRange2 column with 1; for other missing values, we replaced them with FALSE.

c. Logistic Regression

i. Logistic Regression model

We build a logistic regression, setting the attributes we selected before as the explanatory variables, and the status of fast food restaurants as the response variable. We set the closed fast food restaurants to 0, and the open fast food restaurants with low star ratings to 1.

The p-values of **BusinessAcceptsCreditCards** and **GoodForKids** are greater than 0.05, which means that there is no statistically significant relationship between the status of fast food restaurants and these two variables. So we drop these two variables and build a reduced model as our final model:

$$P(Y = 1 \mid X) = logistic(1.76 + 3.22*Delivery - 1.13*TakeOut - 1.88*Price - 1.45*Reservations - 0.29*GoodForGroups)$$

$$= \frac{exp(1.76 + 3.22*Delivery - 1.13*TakeOut - 1.88*Price - 1.45*Reservations - 0.29*GoodForGroups)}{1+exp(1.76 + 3.22*Delivery - 1.13*TakeOut - 1.88*Price - 1.45*Reservations - 0.29*GoodForGroups)}$$

ii. Interpretation

In our final model, we have $\widehat{\beta_0} = 1.76$ (intercept), $\widehat{\beta_1} = 3.22$ (delivery), etc. Then, for instance, among restaurants with identical selected attributes except for delivery service, if the restaurant provides delivery service, the odds of avoiding closure will increase by 25.03%.

iii. Prediction

Once we fit our logistic regression model, we can also predict probabilities of Y=1 given a set of covariates. For example, the probability of remaining open for a fast food restaurant with delivery service, takeout service, good for groups, and low price (RestaurantsPriceRange2=1) is estimated to be about 84.22%.

3. Review Analysis

a. Goal

In the review analysis, we aim to study the types of food offered by closed fast-food restaurants and the satisfaction level of customers with these foods and to explore which foods in restaurants make customers unsatisfied, which leads to the closure of restaurants.

b. Data Preprocessing

We combined the review data set and business data set by business_id. Then, we conducted the following steps to clean the review data:
(1) word tokenization to convert text into words; (2) convert word abbreviation into full words. For example, We converted “wasn't” to “was not”, “can't” to ”can not”, etc; (3) checked encoding and made it consistent for all reviews; (4) lowercase all reviews and remove numbers, punctuations, and stopwords in the reviews (we loaded the "Stop Words" data from the tidytext package); (5) picked the most effective sentiment shifters, find synonyms for these phrases, and replaced the negations with their synonyms. For example, we replaced “not worth” with “expensive” and “never disappointed” with “satisfied”; (6) split the words in reviews into positive and negative ones by calculating the compound scores and add two columns in the review data frame; (7) calculate the average stars for each business_id and combine the positive and negative review words for each business_id.

An example is shown in Table.1:

idx	name	stars	positive	negative
2	Hardee's	1.78	['love', 'delicious']	['horrible', 'negative', 'bad', 'angry', 'negative', 'pathetic', 'with']

Table.1

c. Sentiment Analysis for Informative Nouns

We counted the frequency of all words in the review texts and sorted them from largest to smallest. Then we manually selected the top 6 fast food nouns, which are cheese, hot dog, pizza, salad, taco, and burger, according to the word's frequency. Then, we counted the occurrence frequency of each food noun in the positive and negative review words in Table.2. And we calculated the proportion of food nouns appearing in positive review words to food nouns appearing in all review words and defined the result as “customer attitude score”. The attitude score ranges from 0 to 1, with 0 indicating that the customer is completely unsatisfied and 1 indicating that the customer is completely satisfied. The result is shown in Table.2:

name	**pos_count**	**neg_count**	**attitude_score**
cheese	37196	9858	0.790
hot	39944	10330	0.795
pizza	16565	3347	0.832
salad	32736	7674	0.810
taco	18182	4901	0.788
burger	27139	7487	0.784

Table.2

d. Tests and Results

The p-value for the chi-square contingency table test is less than $2.2\times10-16$, which means that customers have different attitudes towards different kinds of food. Pizza has the highest customer attitude score equal to 0.832, while burger has the lowest one equal to 0.784.

4. Suggestions and Conclusion

a. Conclusion

- From attitude analysis, {RestaurantsDelivery, RestaurantsTakeOut, RestaurantsPriceRange2, RestaurantsReservations, and RestaurantsGoodForGroups} are significant attributes that will affect the closure of fast food restaurants.
- From review analysis, among all kinds of fast food, only 78.4% of reviews about burgers are associated with positive words, while above 83.2% of reviews about pizza are associated with positive words. So fast food restaurants need to put in more attention to improving the quality of burgers the most.

b. Recommendations for Fast Food Businesses

- Provide delivery service:
  Why? If the fast food restaurant provides delivery service, the odds of avoiding closure will increase by 25.03%.
- Introduce competitive pricing and great deals:
  Why? Food with a lower price range will improve the odds 6.55 times to let restaurants remain open.
- Cancel the reservation needs:
  Why? Canceling the reservation will increase the probability of fast food restaurants remaining open by 4.26 times.
- Provide more space for the people who come individually to follow the trend of the fast pace of modern society.
  Why? Fast food restaurants which are suitable for eating alone tend to have a higher probability(95.67%) to remain open than those which are unsuitable for eating alone (94.31%).
- Improve the quality and taste of burgers
  Why? Burger has the lowest customer attitude score equal to 0.784 among the food menu.

c. Limitations

- For the study of missing values, the error will be reduced if we apply some machine learning methods to the imputation of missing values.
- Since our recommendations are only based on the proportion of attributes, they are not comprehensive enough for fast food restaurants. Our recommendations limit how to help a fast food restaurant avoid closure. In the future, we would like to expand our model to help fast food restaurants improve their star ratings.

d. Future work

- Introducing Aspect Based Sentiment Analysis (ASBA) and Natural Language Inference (NLI) models for a comprehensive study. Using advanced transformer bases NLP models like GPT-3.

5. Shiny App

https://msatheesh.shinyapps.io/Module-3/

6. Contributions and References:

a. Contributions

- **Jiaqi Xia**: code (data cleaning, barplots in EDA, attribute selection of attribute analysis, review analysis); summary(background information, data cleaning, review analysis); slides
- **Yutong Zhan**g: code (data cleaning, logistic regression of attribute analysis, review analysis); summary(data cleaning, attribute analysis, review analysis); slides
- **Midhun Satheesh**: code (wordclouds in EDA, data processing in Review Analysis), Shiny app

b. References

[1] Python | NLP analysis of Restaurant reviews
https://www.geeksforgeeks.org/python-nlp-analysis-of-restaurant-reviews/.
[2] Amazon Product Review / Sentiment Analysis Using R
https://github.com/ozlemuysal/Amazon-Product-Review-Sentiment-Analysis-R.