Summary

Introduction

Yelp is known for its features of helping users looking for their desired restaurants, grocery stores, etc. Besides helping customers, it also allows users to write their comment toward a certain business, and these comments can help business owners to better improve their business' performance. In this project, we will exploit the factors that can help East Asian restaurant owners in CA to enhance their business' performance from multiple perspectives.

Data Pre-Processing

1. **Data Extraction:** We are focusing on the business and review data in this project. The original business dataset has 150346 rows of data and 14 features; the review dataset has 6990280 rows and 6 features. After extracting all CA east Asian restaurants from the business dataset based on some keywords and all the object-like features, we combined it with the review dataset and finally got a dataset of 14825 rows of data and 87 features. Additionally, the dataset was splitted to modeling dataset and NLP dataset according to each teammates' requirements.
2. **Data cleaning:** We extracted information in the attributes so that more features can be applied in future modeling processes. Then, according to keywords in the original category and business name, we categorize them to their belonging cuisine type (Chinese, Japanes, Korean, Asian Fusion). For each business's schedule, we computed the daily and weekly working hour so it can be better used in the model. Additionally, we unify types in each feature (turn mixed-type feature to single type)
3. **NLP:** To start with, we built multiple word clouds to observe some words that are too big and not informative, and add these words to stop word list. These steps make some small but important information more prominent. After removing stop words in each review text, we also apply the regular expression package in python to remove symbols and punctuations for cleaner text. In addition, we apply lemmatization techniques to recover words to its normal form in order to pass more information.
4. **EDA:** First, we calculated a new column named 'Total_hour' to get the weekly open hours for each restaurant in order to eliminate the difference of opening hours on different weekdays and removed the samples with 'Total_hour' being 0. This helps us analyze whether business hours are related to ratings. Then, for each restaurant, we use the most informative observation (with the features from both business and review data) to fill the missing data in other observations. Finally, we selected features with less than 1000 missing data to do the following analysis (we also did the analysis based on the features with less than 3000 missing data, but it showed poor performance.). Based on our cleaned data (now contains 14 features, 13 of those are categorical data describing whether or not the restaurant provides this service with 1 indicating yes and 0 indicating no; and 1 continuous feature describes the weekly open hour), we figured out that positive words (like nice, comfortable, warm and etc.) tend to appear more frequently in higher-rated reviews; while negative words appear more frequently in lower-rated reviews. Figure 1 shows some examples:

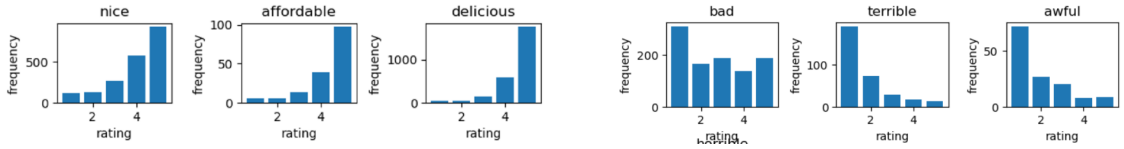


Figure 1: Frequency of Positive words(left) VS Negative words(right)

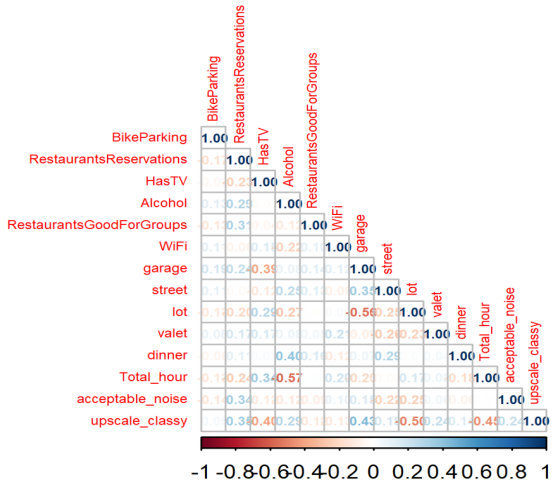


Figure 2: Correlation Coefficients between each two features

And through the correlation matrix (See Figure 2), there is no strong correlation between our 14 features since all the values in the correlation matrix are lower than 0.6. Only the correlation between the weekly opening hour and whether serving alcohol is close to 0.6 which is 0.57. This is probably because restaurants that serve alcohol tend to be open late and may have longer business hours.

Building Model

We set the rating given in each review as our response variable and all 14 features as our explanatory variables (13 features are categorical data and 1 feature 'Total_hour' is continuous data). Since there are 5 categories in the rating (1,2,3,4 and 5) and the response categories are ordered (from the lowest 1 to the highest 5), we build **cumulative logit models** for our ordinal responses. For the explanatory variables x, the model

$$logit[P(Y \le j)] = \log[\frac{P(Y \le j)}{1-P(Y \le j)}] = \log[\frac{\pi_1 + \ldots + \pi_j}{\pi_{j+1} + \ldots + \pi_J}] = \alpha_j + \beta x,\ j = 1, 2, \ldots, J-1$$

$$P(Y \le j) = \exp(\alpha_j + \beta x)/[1 + \exp(\alpha_j + \beta x)],\ j = 1, 2, \ldots, J-1$$

has parameter β describing the effect of x on the log odds of response in category j or below. Since we assume that the effect of x is identical for all $J-1$ cumulative logits, β does not have a j subscript. Table 1 shows output for our final model with 12 features after we removed the insignificant features:

Table 1: Model with 12 features

variable	coefficient &significance level	variable	coefficient &significance level	variable	coefficient &significance level	variable	coefficient &significance level
Intercept 1	-3.843***	BikeParking1	0.621***	lot1	-0.143**	dinner1	0.268**
Intercept 2	-3.204***	Reservations1	0.133**	valet1	-0.683***	Total_hour	0.008***
Intercept 3	-2.582***	Alcohol1	-0.416***	garage1	-0.877***	acceptable noise1	1.301***
Intercept 4	-1.632***	WiFi1	-0.267***	street1	0.269***	upscale classy1	-0.176**

Significant level: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Through the following Table 2, in the goodness of fit test both p-values of two models are larger than 0.1 which indicates that we can't reject the null hypothesis that the model fits the data well at significance level $\alpha = 0.1$; and in the ANOVA table p-value is also larger than 0.1 which indicates that we can't reject the null hypothesis that those two $\beta_i = 0$.

Table 2: Models' goodness of fit test and ANOVA test for two dropped features

	Goodness of Fit test			ANOVA				
	χ^2	residual df	p-value	residual deviance	residual df	df	deviance	p-value
12features Model	$4.29 * 10^{-4}$	42740	0.238	28770	42740			
14features Model	$4.3 * 10^{-4}$	42738	0.1811	28767	42738	2	2.4008	0.3011

Through the above Table 1, we can find that restaurants that provide alcohol, Wi-Fi, and parking places and have an elegant environment tend to have higher ratings. For example, if the restaurant provides alcohol, the estimated odds that this restaurant will receive a lower rating rather than a higher one (i.e.,$Y \leq j$ rather than $Y > j$) equal $\exp(-0.416) \approx 0.65$ times the estimated odds for those restaurants that don't serve alcohol. It is worth noting that the length of business hours does not make a huge difference to the rating. The estimated odds of receiving a higher rating over receiving a lower one is close to 1 ($\exp(0.008)$) with 1 increase in the weekly open hour.

What's more, we further conducted statistical analyses similar to the above analysis on each category of East Asian restaurants (like Chinese, Japanese and Asian Fusion restaurants) and found something new. For example, we noticed that for Japanese restaurants serving alcohol doesn't affect the rating, since all Japanese restaurants serve alcohol. In contrast, those who offer Wi-Fi and garage are more likely to receive higher ratings (The estimated odds that this restaurant will receive a higher rating rather than a lower one equal $\exp(0.839) \approx 2.31$ and $\exp(0.801) \approx 2.23$ times the estimated odds for those restaurants that don't serve these respectively). Those are what we couldn't figure out when we analyzed all East Asian restaurants. More details on different categories of East Asian Restaurants will be shown in ShinyApp.

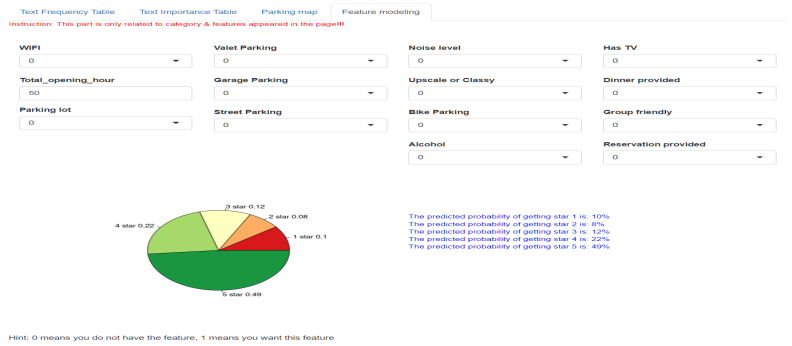
Shiny

Our shiny app has four main purposes, which shows business owners their restaurant's Review text frequency; Review text importance; Interactive cuisine/category map of CA; and the predicted probability pie chart for 5 rating categories.

The text frequency & importance table were built according to different techniques. The frequency table was purely depending on the frequency of a certain text, whereas the importance table was based on the TF-IDF algorithm.

The cuisine map view was a HTML file which was built by using python 'plotly' package. Although it is an interactive graph, I can't be updated in real-time.

The last part is our model. Instead of focusing on a single restaurant, we built a model for each category of restaurant (Chinese, Japanese and Asian Fusion). After selecting a specific restaurant category, the business owner could find out the probability of getting a certain star when adding the features they need.



Limitation:

1.Modeling limitation:

a. Since some services are provided in all restaurants, they will be removed from the explanatory variables when fitting the model (or there will exist collinearity). And we may ignore the effect of these features on the rating (like the effect of Alcohol on Japanese restaurants).
b. For Korean restaurants, there exists collinearity between most of the features. Therefore, building a model to analyze the effect of each feature on the performance of Korean restaurants might not be the best choice.
c. The quantitative effect of each feature on the rating can be described in detail through the multi-category logit model, but the prediction accuracy and classification speed of the model may not be as good as other machine learning models like KNN, decision tree etc.

2. Shiny limitation:

a. Due to python virtual environment setting, some plotting functions that were builts using python are not able to show in the shiny server. Thus, in the text frequency and importance, we applied table view instead
b. The cuisine map feature can be more informative if we can combine it with parking map.
c. For the modeling part, not all the modeling were trained using the 14 features, thus, it might be more reasonable to shadow features that were not used to train in the model

Conclusion

First, when we preliminarily analyze customers' reviews through word clouds based on our cleaned data, we found that the more positive words in the review, the more likely it will give a higher rating. Therefore, paying attention to customers' feelings and providing them a positive, elegant and comfortable experience will help restaurants get high ratings.

Second, by using the text frequency and text importance technique to analyze text can provide different points of views for the business owner. For instance, some words appears to be the 10th position in text frequency plot, but they might be top 5 in the TF-IDF plot/table

Third, for the restaurants in different categories, they need to focus on different features. For example, if the Chinese restaurant provides dinner and alcohol, it will be more likely (the odds of getting a higher rating will be 1.46 and 1.81 times respectively) for them to get a higher rating. However, for Japanese restaurants, providing dinner isn't a very significant factor. And for all the restaurants, the weekly open hour is not a factor that would make a huge difference in ratings, so restaurants don't have to work unnecessary extra hours to earn high ratings.

Reference:

1. Agresti, Alan. An Introduction to Categorical Data Analysis (Wiley Series in Probability and Statistics). United Kingdom: Wiley, 2007.

Contributions

Only two of the group members were involved in the project. We had group meetings every week for about 3 hours, and the project took us around 10+ hours of group meetings plus individually contributed time. Although we only had two teammates who pretty much did everything, we actively helped each other on their responsible sections like providing ideas, dealing with data, and complementing each other.

The following were the task perform by each person:

1. **Shunyi Huang** - Data Pre-processing, NLP, EDA on review data, Shiny, report, PPT
2. **Yuqian Chen -** Data pre-processing, EDA, Modeling analysis, Shiny of model part, report, PPT
3. **Qizhou Huang** - Only contribution