Suggestions for Ice cream shop oweners based on Yelp Review

¶

STAT 628 Module 3 Group 3

1. Introduction

In this project, we would use Yelp data to analysis ice cream business in the U.S. To be more specific, we mainly choose 7 target cities (Las Vegas, Phoenix, Charlotte, Pittsburgh, Madison, Cleveland, Urbana-Champaign) and select 455 ice cream shops in these cities. We have 45k reviews in total for our analysis. There are 3 goals we would like to achieve for the project, which are:

- **Influencing factors:** to find influencing factors of success for ice cream shops
- **Recommendation:** to provide recommendation for customers, such as recommended shop, flavor, topping in a certain city
- **Score prediction:** to construct score-prediction model for shop owners to help them rearrange their facilities/services and improve their scores

2. Advice for Ice-cream shop owners

This section is to find the commen features of successful ice-cream shops. In other word, we hope to analyze users' reviews and find the factors that they concerned most with the hope to provide suggestions to shop owner.

2.1 NLP on reviews

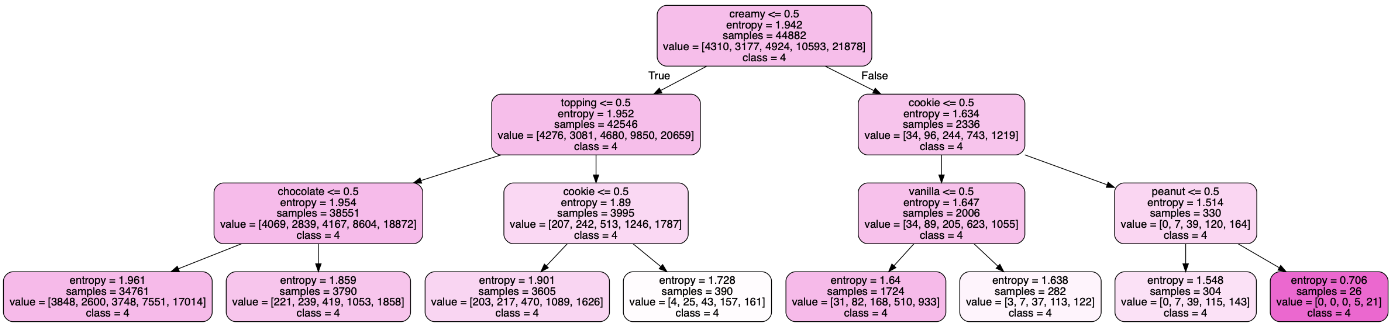
To deal with text data, we mainly adopt the following NLP method to analyze.

- Word tokenization: split sentences into lists of word
- Filtering: get rid of stopwords, punctuations which do not have real meaning. We also convert synonyms, abbreviations into uniform word
- POS classification: classify words into noun, adj/adv, verb
- Word frequency: calcute word frequency for each POS class

2.2 Decision tree

After data preprocessing, we further choose words from 3 POS classes above and further divide them into more detailed dimensions. Note that we only choose top words for each dimension based on Information Gain with respect to rating *Decision Tree*. That is to say, these words have more power of distinguishing scores of reviews so that we put more attention on them.

For each kinds of words, we fit a classification tree for 3-4 layers, and only extract the words at these layers as our high information words. The following graph is this procedure for ice cream properties, and we only choose words which are criterion at each node.



After repeating this procedure for words of each type, we get:

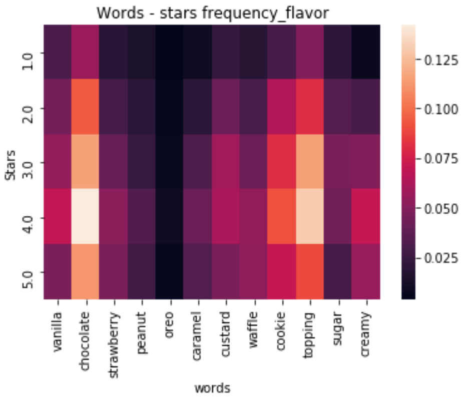
- Nouns:
  - About ice-cream: topping, cookie, waffle, crunch, sugar, diary, homemade, custard
  - Ice-cream flavor: vanilla, chocolate, strawberry, peanut, oreo, caramel
  - About service: price, quality, wait, employee, texture, fan
- Adj/Adv: delicious, friendly, small, fresh, creamy, hard
- Verb: parking, located

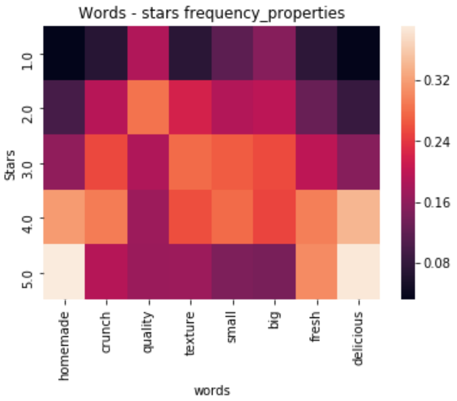
Then we categarize these words in a more informative way:

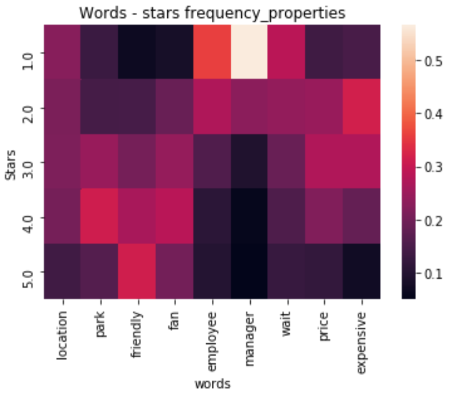
- flavor : vanilla, chocolate, strawberry, peanut, oreo, caramel, custard, waffle, cookie, sugar, creamy
- properties : homemade, crunch, quality, texture, small, big, fresh, delicious
- service : location, park, friendly, fan, employee, manager, wait, price, expensive

2.3 Heatmap

To make the information each words contain clear, we must connect them with ratings so that we can give suggestion. Therefore the most concise way is to use heatmaps. Because the ratings with stars 4, 5 take a high proportion, we have rescale the rating by the total number for each stars. And we split the big heatmap to 3 submap in three aspects.


Words - stars frequency_flavor


Words - stars frequency_properties


Words - stars frequency_properties

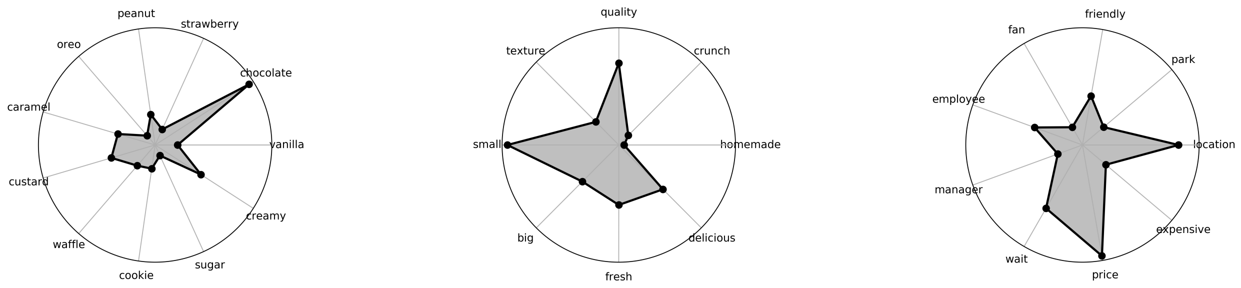
It's clear some words have high correlation with the rating, and it's still not neat enough to extract the information, so we decided to do inference for different cities separately.

2.3 Radar graph

In this section, we plot the rador graph for each city and also each aspect, thus to give corresponding suggestions.

Specific, for flavor part, we plot the words proprotion associated with the postive rating(4,5) to show the preference, while for the property and service, we plot the words proprotion associated with the negative rating(1,2,3) to point out the aspects requiring improvement.

The follow is the graph for Madison. Flavor - Property - Service



3. Stars prediction

This section is designed to predict the star of a business based on the attributes it have. We hope to give suggestions for new shop owner on location, open hour, rules, facilities, etc.

3.1 Data pre-processing for 'business' data set

(1) Data selection
There are 19209 business record in ‘business' dataset. We select opening business with category contains 'Ice Cream' and review count greater than 10 in target cities. Now the sample size is 453.

(2) Data cleaning
The main variables need to be cleaned are 'attributes' and 'hours'. Variable 'attributes' lists the information of various attributes and 'hours' list business time from Monday to Sunday.

a. For 'attributes', there are total 32 attributes are mentioned. We find different business offer different attributes information. For example, information about WiFi is provided by 426 business while information about smoking is only provided by 4 business. This introduces missing values in every attribute. We deal with it by the following 3 methodds:

- Drop the variable. If less than 100 business provide status of this attribute, we delete this attribute since missing value of this variable contributes too much to be handled.
- Randomly replce missing value with non-missing value. We do t-test for each attribute to test if the mean stars of business that tell information about this attribute differs from that of business don't tell. If there is no difference, it is reasonable to consider that this variable's missing value are missing at random regardless of stars. That is to say unobserved true value of this attribute shares the same distribution of observed value. Thus, we replace missing values with values that are sampling from distribution of non-missing values. Take 'Bikeparking' as an example. There are 91 FALSE, 343 TRUE and 19 NA. We assign FALSE and TRUE to NA randomly and proportionally, that is assigning 4 FALSE and 15 TRUE to replace NA.

- Impute missing value by KNN. If there is significant difference between stars of business that provide a given attribute's information or not, we can conclude that missing values are not missing at random. Thus we use KNN model to predict the true value of missing value based on stars and review counts.

Now there is no missing value in selected attributes. However, some attributes have multiple categories. We attempt to reduce categories for the convinence of later analysis. F-test shows significant difference exists among multiple groups. Then we perfom Tukey test to test difference between pairwise groups. If the difference is not significant, we combine the two categories into one.
b. For 'hours', we reduce 7 days to 2 groups: weekday and weekend. If the shop is close, then there is missing value for open time and close time of this day. There are only 4 business close on weekdays so we ignore it. For each business, we compute weekday open time by mode of open time from Monday to Friday. Weekday close time is likewise. There are 22 business close on weekends, so we add a varibale to show business opening time. If it closes on weekend, then the opening time is 0. Otherwise it is calculated by counting hours from open time and close time. For each business, we compute weekend open time by mean of open time on Saturday and Sunday. Weekend close time is likewise. Too many different time points lead to complicated analysis, so we find the best cut points for open time, close time and opening time by a proposed algorithm[1]. Since cut points of open time for both weekdays and weekends are insignificant, we drop this two variables. Optimal cut points of close time for both weekdays and weekends are 23:00 and for business time is 11 hours. Hence, these three variables can be converted to two-categorical variables.

3.2 Computing user weight for 'user' data set

The purpose of this section is to classify users into trustful and not so trustful customers group. We prefer to value more of reviews from trustful users since their comments may be more justify and meaningful. Consequently, we construct a logistic regression model to predict the validity of a customer.

- Dependent Variable: Elite (title given by Yelp indicating the user is active and faithful)
- Independent Variables: number of fans, number of friends, review count, userful count, starting year, cool/funny count, compliment note count, compliment writer count
- Threshold: 0.1 Note that faithful users are not necessary "elite" customers, while "elite" customers are indeed faithful. So we would use a lower threshold to find out all potential faithful users.
- AUC:0.962; Recall: 98.1%; Accuracy: 81.8%

Based on this prediction model, we figure out 38.1% faithful users and give their reviews higher weight in latter analysis. Specifically, weight for "elite" user is 2, weight for "non-elite" is 1 and we can get the weighted sum of stars for each shop.

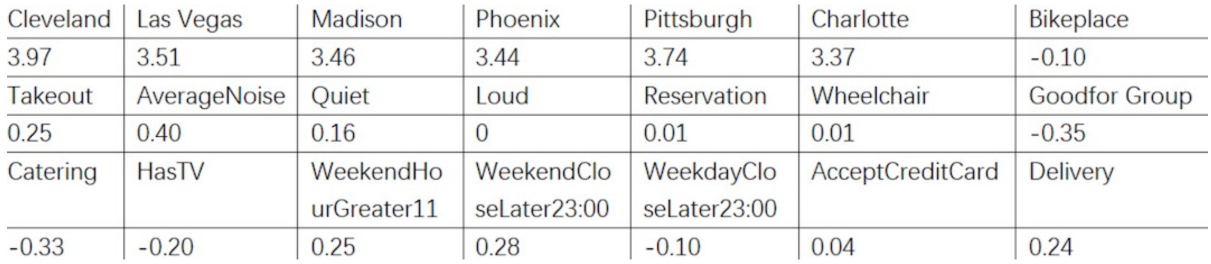
3.3 Build prediction model

(1) Weighted ratings of business
Given the fact that user's contribution and liablility differ from each other, the stars they give to a business in review should be viewd in a weighted way. We use users' weight to compute the weighted star ratings of business. For a given business, we take the average weighted stars in corresponding reviews.
(2) Multivariate linear regression model
Since linear regression is the most interpretable and easiest model, we start with building mutivariate linear regression model between stars and various attributes. We use 5-fold cross validation. When testing the model, we first predict the stars by model in validation dataset. Then we round both the predicted result and the observed stars to integer and count how many samples have the same predicted and observed value. The main reason to do this is that we prefer to provide integer stars to businees rather than float. The mean 'misclssification rate' is 31%.
(3) Classification tree model

Then we build classification tree model to compare with linear regression model, which is less interpretable but more reasonable for classification. Before training model, we convert continous stars to categorical ones. Then 1000 times 5-fold cross validation is performed. The mean misclassification rate is 29%.

(4) Compare models

From above, we can see classification tree model has slightly lower misclassification rate. However, it is hard to interprete how attributes affect the stars to business owners. Overall, we choose linear regression model for predicting stars. The coefficients are listed:



Cleveland	Las Vegas	Madison	Phoenix	Pittsburgh	Charlotte	Bikeplace
3.97	3.51	3.46	3.44	3.74	3.37	-0.10
Takeout	AverageNoise	Quiet	Loud	Reservation	Wheelchair	Goodfor Group
0.25	0.40	0.16	0	0.01	0.01	-0.35
Catering	HasTV	WeekendHourGreater11	WeekendCloseLater23:00	WeekdayCloseLater23:00	AcceptCreditCard	Delivery
-0.33	-0.20	0.25	0.28	-0.10	0.04	0.24

From the reult we can see ice-cream shops in Phoenix have the lowest average stars and Cleveland hve the highest. Not being nosiy, providing reservation, wheelchair, creditcard, delivery and take-out service will help promote stars, while goodforgroup， catering， TV are not helpful for improving stars. At weekend, businees time longer than 11 hours and closing door after 11:00 pm is suggested.

4. Conclusion

- Overall:
  (1) It is unnecessary to have caterings nor TVs.
  (2) The shops shouldn't be too noisy, thus noisy group activities should not be allowed.
  (3) Basic service such as reservation, wheelchair, take out, delivery, credit card acception are expected to provide.
  (4) It is better to close after 11:00 pm at weekend while it is not necessary to do so at weekdays.
  (5) Provide different sizes of ice-cream. Fresh and delicious are vital.
- Specified for city:
  Charlotte:
  (1) Flavor: The business owners should emphasize on the flavor of chocolate and strawberry, making the ice cream creamy, and being innovative in developing new ice cream accompanied by cookies and waffle.
  (2) Service: The business owners should know the location and long waiting time for waiting are the major concerns for ice cream business, so the new ice cream store should address such problems. Also the atitude of employee are crucial for having good comments.
  Cleveland:
  (1) Flavor: The business owners should pay much attention to the flavor of chocolate and caramel, and also make ice cream creamy.
  (2) Service: Price, waiting time and service atitude are the major concerns for ice cream business, so new ice cream stores should address such problems by providing higher quality service, and also launching more affordable items.
  Madison:
  (1) Flavor: The business owners need to notice that the flavor of chocolate and custard are most popular, and peanut catches lots of attention too. They also need to make the ice cream creamy enough to earn high ratings.
  (2) Service: Location and price are two most apparent problems. They should launch some new items with cheaper price, and it's also worth trying to open new stores in good locations.
  Phoenix:
  (1) Flavor: Choclate flavor is recommended. Also, try to develop new ice cream accompanied by cookies.
  (2) Service: The ice cream business in Phoenix has much potential for improvement and actually lots of customers are not satisfied with the service. Therefore, business owners can open new stores in center

customers are not satisfied with the service. Therefore, business owners can open new stores in center locations and offer high-quality service to attract more potential customers.
Pittsburgh:
(1) Flavor: The business owners should do well in the flavor of chocolate, and be innovative in developing new ice cream accompanied by waffle.
(2) Service: Do something to avoid long waiting time and high price.
Las Vegas:
(1) Chocolate, strawberry and custard are popular. Make them creamy and add toppigs.
(2) Good location and well-trained staffs should be paid much attention to.

5. Pros and Cons

Pros:
(1) We deal with missing values in different ways under different conditions.
(2) We recompute business stars with users' weight rather than common average, which is more reasonable.
(3) We use both basic statistics test and advanced tree model buiding in vaiable selection depend on differnt steps.
(4) Spatial anaysis would distinguish ice cream preference in different regions, and offer more specific suggestions.
Cons:
(1) We are not able to quantify the effect of each word or feature to the rate of review.
(2) The sample size of business is not large enough compare to the numbers of ccovariates. Some categories of some attributes have too small sample size to guarantee the predition precise.

6. Appendix

References

[1] Lopez-Raton, M.(2014).OptimalCutpoints: An R Package for Selecting Optimal Cutpoints in Diagnostic Tests. Journal of Statistical Software 61(8), 1–36.
[2] Data resources: https://uwmadison.box.com/s/bp36qfdw9twqf6po4tft6iktdfpzr0k0 (https://uwmadison.box.com/s/bp36qfdw9twqf6po4tft6iktdfpzr0k0)
[3] https://cran.r-project.org/web/packages/rpart/rpart.pdf (https://cran.r-project.org/web/packages/rpart/rpart.pdf)

Contribution: Each member in our group contributes much to this project and we all participate in slides design, report compiling. The table is our duty for this project.

Member	Contribution
Ke Chen	Transform json to csv, NLP on reviews, user-weight computation
Chen Hu	Attribute data preprocessing, Stars prediction
Nan Yan	Shiny, give suggestion, manage Github
Richard Yang	Regression tree, Spatial analysis, all other graphs