"Restaurant Pandora": A Restaurant Recommendation System Based on Yelp Information

Linxuan Shi, Peiying Yu, Ye Yuan

December 2018

1. Introduction

The Yelp website is a user-based local business review and social networking site. The site devotes to individual locations, such as restaurants, where Yelp users can submit a review of their products or services using a one to five-star rating system. Businesses can also update contact information, hours and other basic listing information or add special deals.

Suppose a student at Columbia would like to explore some new restaurants near campus. It might be difficult for the student to choose a restaurant that matches his/her preference. The restaurant the student finds on yelp might have a good yelp rate, but it might not match the student's personal preferences. For instance, the restaurant that yelp shows offers correct food type the user wants but might be too expensive or doesn't have some certain feature that the user desires. Each restaurant has pros and cons from each user's own perspective, but yelp rating only gives a general picture of the restaurant based on all the users who have commented. Therefore, we would like to provide a convenient and more flexible restaurant recommendation system which considers the specific user's historical ratings and matches the user's preferences better.

When the user is searching for restaurants, our recommendation system will suggest top 10 restaurants which match the user the most based on his/her past dining history ratings.
In general, we are aiming at how to recommend restaurants to specific users based on their personalized preferences according to official evaluations from credible sources and historical user ratings on those restaurants in yelp dataset.

In order to obtain a better understanding the rating of a restaurant, we also desire to figure out which restaurant attributes would have impact on the restaurant's Yelp rating. Then, we will be able to make our recommendation more attractive and convincing to the user. For instance, we can push a notification recommending a restaurant to the user along with maybe one or two significant attributes we've found significant that might affect the decision of the user.

2. Dataset

Data Acquisition

While Yelp has a lot of data, most of the data was locked in the site itself. Yelp provides its own API, which provides some data, but it is extraordinarily limited compared to the data on the site. So, we devised a web scraper to efficiently traverse each restaurant website on Yelp and scraped restaurant information from the Yelp website.

The scraper was developed using python with Beautifulsoup and urllib. Beautifulsoup is a powerful python library that can analyze HTML and parse certain tags and their information. urllib was used to open the sites for data processing so that Beautifulsoup could then strip the essential data.

With the scraper, we obtained information for 2,500 restaurants in NYC with 90 restaurants in the Morningside heights neighborhood. For each restaurant, we scraped 37 features. The key features we scrape for each restaurant are: restaurant name, restaurant rating (yelp rating ranging from 0 to 5 with 0.5 increment), hygiene score (A, B, C), restaurant neighborhood (Morningside height, East Village, Chelsea, etc.), noise level (Average, Loud, Quiet), parking (Garage, No Parking, Private lot, Street), Wi-Fi (Yes, No, Paid), price range (under $10, $11-$30, $31-$60, above $60), waiter service (Yes, No), discrete number of reviews and so on.

After filtering out all the restaurants in Morningside heights in the dataset, we also scraped all the reviews from each Morningside heights restaurant. In the review section on Yelp, the two feature we scraped are reviewer's username and reviewer's rating for these specific restaurants.

Data Cleaning

To store the scraped data, we used csv file format. We used R to aggregate 90 Morningside restaurants review information. In the aggregated dataset, each row represents a unique Yelp user, and each column represents a Morningside heights restaurant. If the user has a rating for a restaurant, we would fill the corresponding position with his or her rating. Otherwise, NA is filled. Since Morningside heights is a small community, we found a lot of users rated several restaurants in the Morningside heights.

For the restaurant information dataset, we dropped the rows where the restaurants have no Yelp rating.

3. Prediction System

For our prediction system, we utilized the dataset of the restaurants with all kinds of features to find out which features are significantly correlated with the rating of the restaurant. However, all of the features except for the number of reviews are categorical instead of numerical. Besides, there would be too many dummy variables if all of the features are taken into consideration. Hence, we handpicked to omit some of the features that we believed is obviously insignificant to the rating of a restaurant (e.g. accept Bitcoins or not). As a result, we have 22 features in total for the prediction model. To perform test and cross validation process, we excluded the 90 restaurants at Morningside heights to be our test dataset and rest becomes train-and-validation set. After applying the best subset selection method, we found out the optimal number is 15 subsets. The adjusted R-square is still very low which means our model does not work well. However, it is possible that all these features we took into consideration are simply not correlated to the rating of restaurants.

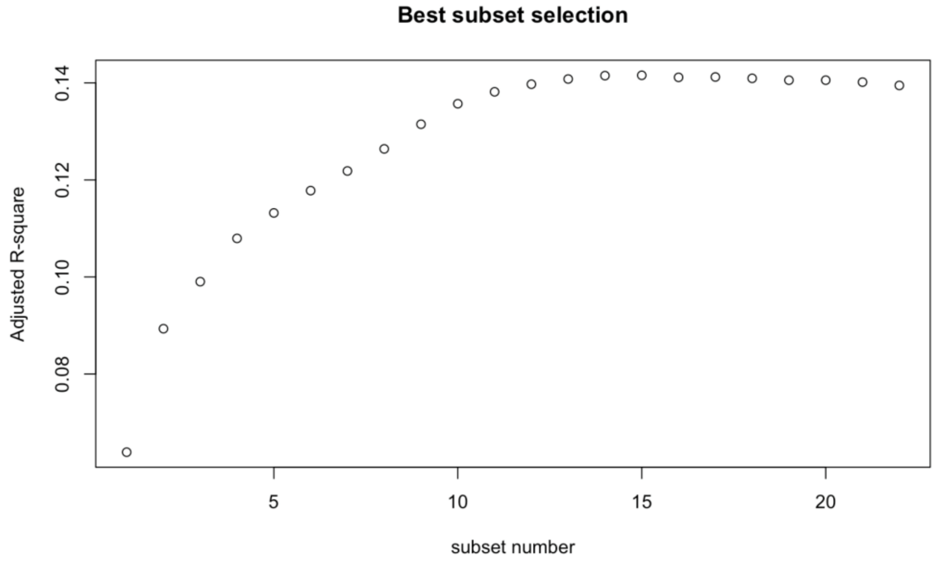


Figure 1: Adjusted R-square for best subset selection

4. Recommendation System:

For our recommendation system, we employed the user-based collaborative filtering. It works by looking for users who share the same rating patterns with the target user. We used the ratings from those like-minded users to make the predictions. We scraped the clients' ratings of 90 restaurants near Columbia University. Then, we used a join function to merge all tables together to make it like the table we used in the class to rate the movie. The whole table includes thousands of clients and we selected only the top 100 individuals who rate the largest number of restaurants.

For the first step of collaborative filtering system, we needed to decide what type of users we will use to make prediction and so we used KNN algorithm to do this. The input is a user corresponding to the business he liked and rated with high ranking. We then calculated the distance for pairwise comparison between the target person and all other users. For this purpose, we calculated the distance by summing up the distances of each business. If the business is rated by only one person of the two, then the distance should be 0. After that we sorted the neighbours by similarity and choose the top k of them for afterwards process. Moreover, we practiced multiple values for k in order to select the optimal KNN model.

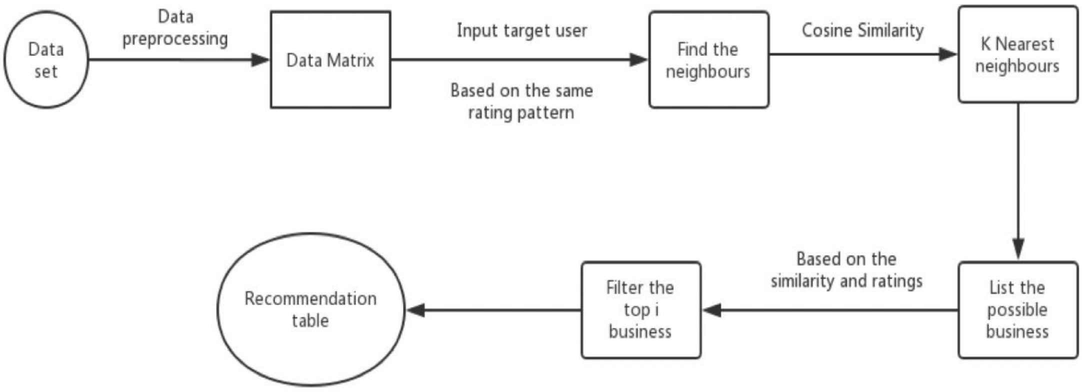


Figure 2: Methodology for Recommendation System

5. Result

For the prediction system, we chose the 90 restaurants at Morningside heights as our test set to find out their predicted rating, along with their differences to the actual ones. As we can see from below, the MSE is 0.206, which is relatively low given that rating ranges from 1 to 5. Even the features are insignificant to rating, it would be paradoxical that the MSE is low. One explanation to this unconventional phenomenon we find out is that all the rating of these 90 restaurants are very close to each other, ranging from 3 to 5, and most of them are 4. And the intercept is a major part of the predicted rating. Hence the MSE is relatively low on a scale of 1 to 5.

Restaurant	Predicted	Actual	Restaurant	Predicted	Actual	Restaurant	Predicted	Actual
Amelie	3.914973	4.5	Doaba Deli	3.914973	4.5	Kikoo Sushi	3.914973	4.5
Amy Ruth's	3.914973	4	Dun Huang	3.785996	4	Kingston	3.914973	4
Antojitos Mexicano El Taco Taco	3.914973	4	e's BAR	3.914973	4	Kitchenette Uptown	3.914973	3.5
Artopolis	3.914973	3	El Puerto Seafood	3.914973	4	Koko Wings	3.914973	3.5
Atlas Kitchen	3.785996	4	Flat Top	3.914973	4	Koronet Pizza	3.914973	3.5
Awash Ethiopian Restaurant	3.914973	4	Flor de Mayo	3.785996	4	La Diagonal	3.914973	4
Babbalucci	3.914973	4	Friedman's	3.914973	3.5	La Piccola Cucina	3.914973	4
Bar314	3.914973	5	Go! Go! Curry!	3.914973	3.5	La Savane	3.914973	4.5
Belle Harlem	3.914973	5	Grain House	3.785996	4	Le Monde	3.914973	3
Bettolona	3.914973	4	Greek Taverna	3.914973	3.5	Lido	3.914973	4
BLVD Bistro	3.914973	4	Harlem Ale House	3.914973	5	Lolo's Seafood Shack	3.914973	4
Broadway Restaurant	3.914973	4	Harlem Taco & Bowl	3.914973	4.5	Loui Loui	3.785996	4
Buceo 95	3.870583	4	Hex & Company	3.914973	4	Malaysia Grill	3.914973	4
Cantina Taqueria & Tequila Bar	3.914973	4	Hula Poke	3.914973	4	MAMA's TOO!	3.914973	4
Carmine's Italian Restaurant	3.870583	4	Infamous Chicken	3.914973	4	Manna Korean Food	3.914973	5
Casa Mexicana	3.914973	4	Isola on Columbus	3.914973	4	Marlow Bistro	3.914973	4
Chapati House	3.914973	4	Jin Ramen	3.914973	4	Massawa	3.914973	4
Community Food & Juice	3.914973	3.5	JJ's Place	3.870583	5	Max Caffe	3.914973	3.5
Dig Inn	3.914973	4	Junzi Kitchen	3.785996	4	Max Soha	3.870583	4
Dive 106	3.914973	4.5	KALBI	3.914973	4.5	Mel's Burger Bar	3.914973	3.5

Figure 3: Predicted Yelp ratings given by the prediction system

For the recommendation system, we chose the top 100 individuals who rate the largest number of restaurants. We used KNN to predict their ratings based on their previous ratings and selected top closest customers to predict. We tried k from 2 to 10 to discover the optimal value of k, and k equals to 4 gives the smallest MSE (0.658), which is acceptable comparing to their ratings range from 1 to 5.

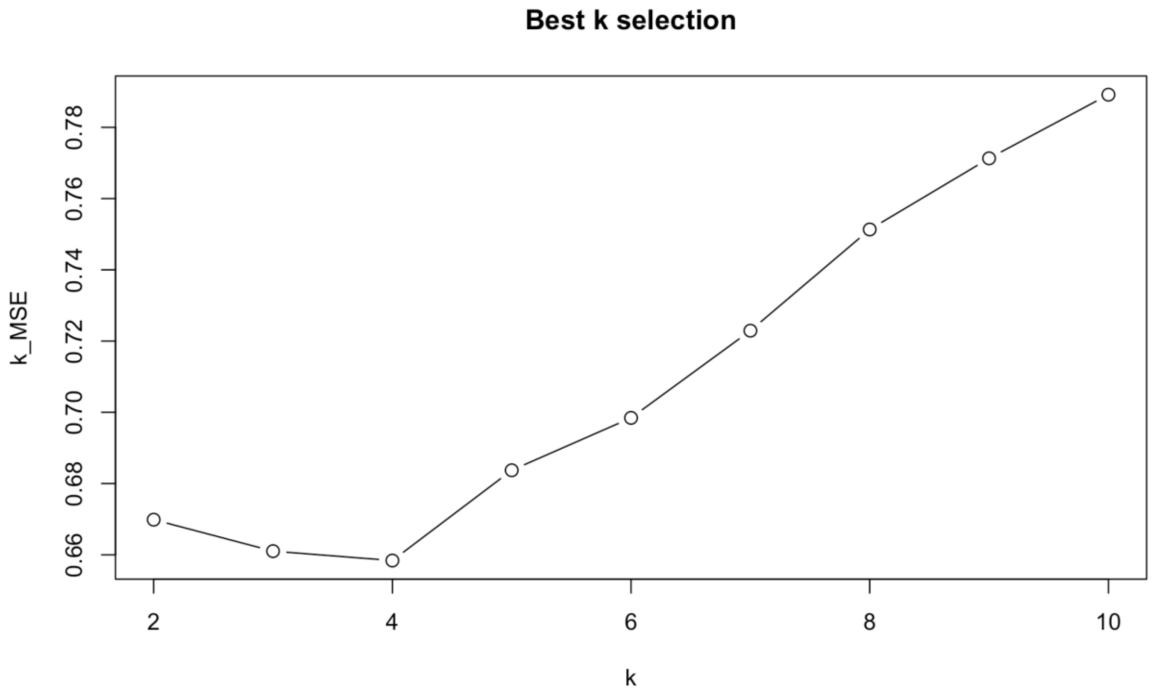
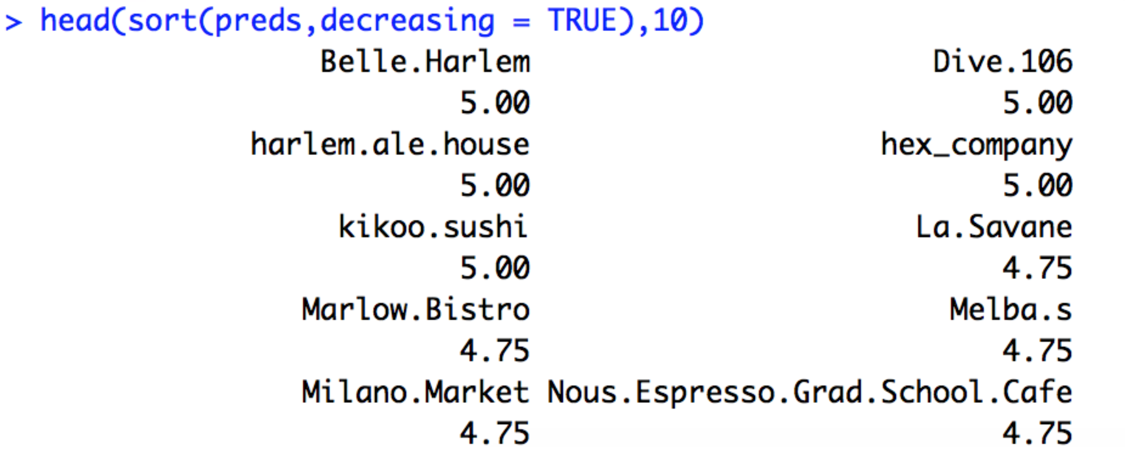


Figure 4: Model selection for KNN

We have also ranked several restaurants we tried before and rated them ourselves. And then chose the top 5 restaurants recommended to have a try. 4 of these 5 suit our taste.



```
> head(sort(preds,decreasing = TRUE),10)
                 Belle.Harlem                     Dive.106 
                         5.00                         5.00 
             harlem.ale.house                  hex_company 
                         5.00                         5.00 
                  kikoo.sushi                    La.Savane 
                         5.00                         4.75 
                Marlow.Bistro                     Melba.s 
                         4.75                         4.75 
                Milano.Market Nous.Espresso.Grad.School.Cafe 
                         4.75                         4.75 
```

Top 10 Restaurant Recommended

Next time, we will expand our dataset and try all the 5 stars the recommendation system predicts for us.

6. Conclusion

To sum up, we used best subset selection first to build our predict for restaurants' rating. However, when we selected the best 15 features out of the 22 total features, the performance of fit is not good. Then we tried cross-validation regression to build the model and the best model still shows that all the features are not significant or have very small coefficient.

So, the result of prediction shows that rating of a restaurant has little relationship to most features that Yelp provides to its user. Most influential feature that would affect rating such as quality of food or serving time of food are not taken into consideration of the developers of Yelp.

As for the recommendation system, we applied a customer-based KNN method to build the system. It finds the 4 closest reviewers based on our previous ratings. And then we calculate the average ratings of the 4 reviewers to each restaurant. The recommendation result provided us with satisfying result and we believe it can continuously perform.

7. Improvement & expansion

Currently, the recommendation system of our project only focuses on the Morningside Heights and areas around Columbia campus. But we believe it can also apply to a larger scale of area, entire Manhattan for instance, or even any other places.

However, our project may only be functional at metropolitan areas since yelp's major user communities are located in these areas. Insufficient reviews and data of rural area businesses will lead to inaccurate results at these locations. Besides, the dataset would become much larger if we want to expand our method to vast locations, the processing time of generating each customer's potential ratings will also increase drastically. Therefore, it is inevitable for us to optimize our algorithm or even design a new one if necessary.

For any questions regarding this report, please contact:

Peiying Yu at py2244@columbia.edu

Linxuan Shi at ls3510@columbia.edu

Ye Yuan at yy2818@columbia.edu