Lab 2 - Predictive Modeling,
Stat 415, Spring 2023

Wing Yan Yau

2nd May 2023

1 Exploratory Data Analysis

For the *Reviews* data set, the number of missing values for each columns can be summarise as follow:

Review Text : 550	Birth Year: 2
Marital Status: 35	Has Children?: 38
Vegetarian?: 1350	Weight (lb): 97
Height (in): 5	Average Amount Spent: 2
Preferred Mode of Transport: 7	Northwestern Student?': 1

We originally planned to drop away the column 'Vegetarian?' for its overwhelmingly large amount of missing values, considering that the total number of rows (or data point) in the data set is merely 1433. Yet after a more careful investigation, among the reviews with this column of information filled, there seems to be an interesting pattern: the average rating done among non vegetarian, which is around 4.14, is much lower than vegetarian which is around 3.14, while the overall average is 3.77. Hence we decided to keep it. For columns with less than 10 missing values, we opted to remove the corresponding rows entirely. For the remaining columns, we addressed the missing values by replacing them with 'missing' for string columns, and 0 for float or integer columns. We also convert manually the problematic 'Review dates' in this data set for 15 rows to the closest proper date format, and changed the spelling of a data point in column 'Marital Status' for consistency and avoid potential problems when doing one-hot encoding.

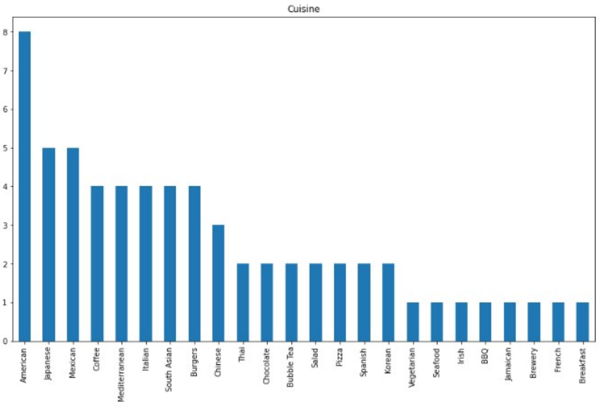

Figure 1: Bar chart for amount of restaurants of different cuisine types

As for the *Restaurants* data set, it is worth noting that there is no missing data in it. Despite variations in the number of restaurants across different cuisines, the differences are within the same order of magnitude (Figure 1.) Hence, we think this data set to be reasonably balanced, as is the case with other columns.

We applied K-means, DBSCAN, and agglomerative clustering techniques to the user demographic data. To prepare the data for clustering, we transformed the categorical variables into the one-hot encoding format. After evaluating the corresponding silhouette coefficients using the elbow method, we determined the optimal parameters: 2 clusters for K-means and an epsilon value of 2 for DBSCAN. Both K-means and DBSCAN produced 2 clusters, indicating agreement between the two methods. However, the DBSCAN algorithm resulted in an imbalanced cluster set. One cluster, in particular, was dominated by reviews from a single reviewer, with only 8 data points compared to 1427 in the other cluster. On the contrary, both K-means and agglomerative clustering produced more reasonable clusters. Between the two, agglomerative clustering revealed a clearer pattern, which is summarized in the table below. The numerical values in the table represent the average characteristics of the data points within each cluster, i.e. the values are in fact percentages (due to the one-hot encoding manipulation), indicating the proportion of data points matching that attribute. The clustering analysis emphasizes distinctive demographics within the clusters. One cluster primarily consists of married couples with children who also own a car, while the other cluster is composed of single individuals without children.

	Cluster 0	Cluster 1
Birth Year (nearest integer)	1990	1970
Marital Status: Married	0.01	0.80
Marital Status: Single	0.99	0.10
Marital Status: Widow	0.00	0.06
Has Children?: No	0.98	0.32
Has Children?: Yes	0.02	0.64
Preferred Mode of Transport: Car Owner	0.47	0.75
Preferred Mode of Transport: On Foot	0.43	0.17
Preferred Mode of Transport: Public Transit	0.11	0.08

2 Popularity matching

The most highly rated restaurants get an average rating of 5, including Fonda Cantina, World Market, Evanston Games Cafe, LeTour and La Principal. The average and median review scores are 3.8 and 3.92 respectively. The restaurant that has received the largest quantity of reviews is Campagnola with 48 ratings obtained, while the median number of reviews received for all is 23.

Using a simple recommendation engine wherein a user can input a cuisine type and receive a recommendation based on popularity score, the recommendation for Spanish food, Chinese food, Mexican food, and Coffee are 5411 Empanadas, Joy Yee Noodle, Chipotle and Brothers K Coffeehouse respectively.

A more sophisticated model that integratedly consider both the popularities and ratings of the restaurants is a shrinkage estimator. Restaurant that has low popularity and low ratings benefits the most and restaurant that has low popularity but high ratings is hurt the most by it.

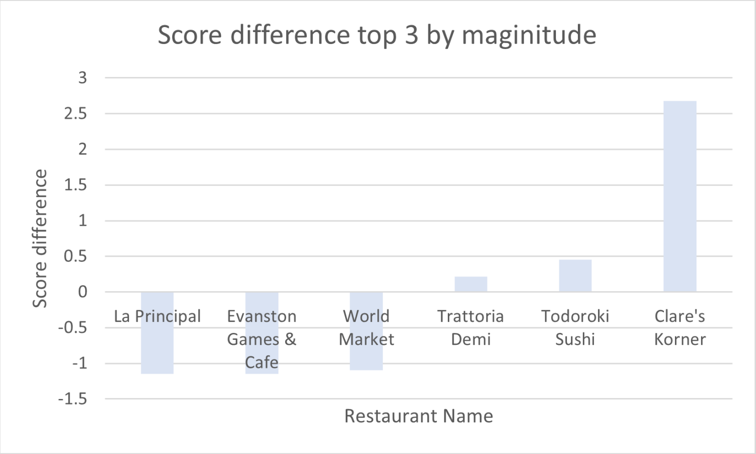

Figure 2: Changes in review scores due to shrinkage estimation (top 7 for both ends)

3 Content based filtering

We computed the euclidean and cosine distance between every restaurant. We wrote a script (cbf1) that take a user, find out the restaurants s/he likes (with ratings ≥ 4) , and loop through the have-not-tried restaurants to kind the top 10 ones with lowest average distance to her/his liked one. We also wrote one (cbf2) to take a restaurant and output the top 10 most similar ones. Here are the results for the test examples:

Script	cbf1 (users based)	cbf2 (restaurants based)
Test Sample	Jennifer Armagost	Taste of Nepal
Recommendations	Sweet Green 5411 Empanadas Hokkaido Ramen Cross Rhodes Elephant & Vine Zentli Evanston Chicken Shack Tealicious Kung Fu Tea Claire's Korner	Mt. Everest Restaurant Mumbai Indian Grill Shangri-La Evanston Tapas Barcelona Lao Sze Chuan Tomo Japanese Street Food Kuni's Japanese Restaurant Taco Diablo Kabul House Table to Stix Ramen

Table 1: Results of the 2 content based filtering methods

4 Natural language analysis

To investigate in the language part of the data set, we used 3 metrics to measure the 'distance'(or similarity) between restaurants, namely the Jaccard, TF-IDF vectors and BERT embedding using the restaurants description together with their cuisine type.

Note that the TD-IDF vectors here are composed only considering the 100 most popular words. We then compared the recommendations made by referring to the 3 distances. To be more precise, we computed the average distance between a user's liked restaurants if s/he has more than 1 of them (with rating ≥ 4), and take the mean (like_dist_mean) of it among all users. We then computed the mean (metric_mean) distance between every restaurants in the 3 metrics, and divided like_dis_mean by metric_mean for corresponding metric, where we called it the recommendation coefficient The smaller the value, the better the metric is concluded to perform.

The intuition behind is that, if the restaurants a user like is having small distance in the perspective of the metric, which is why we need metric_mean, the more possible that it can get recommended by the model. Or in another words, the model is giving better recommendations. The result we got is that BERT is the best model, followed by TF-IDF and the Jaccard gives worst recommendations.

Metrics	BERT	TF-IDF	Jaccard
Recommendation Coefficient	0.26	0.28	0.33

Table 2: Table to recommendation coefficient for the 3 metrics

Another possible, and probably better way to compute such reference value is to find the percentile of like_dis_mean among the set of distances between restaurants instead. As it resembles the focus on ranking instead of absolute values in recommendation problem in a more appropriate way.

For some more investigation on the metrics themselves, we also look into which restaurants has the highest TF-IDF score for some keywords. For example, Dave's Italian Kitchen has the highest TF-IDF score for the word 'cozy', and 'Lao Sze Chuan' for the word 'Chinese'.

5 Collaborative Filtering

We used the demographic data of users in the 'Reviews' data set to form users feature vectors. We then wrote a script that takes a user and outputs k recommendations made by similar users. This is done by checking users ranked by similarity score computed using their feature vectors (with each attributes normalised), and outputting the restaurants they like, i.e. they have given a rating of 4 or above, until there is k of them.

Script	cf1 (demographics based)	cf2 (reviews based)
Test Sample	Jennifer Armagost	
Recommendations	Peppercorns Kitchen Celtic Knot Public House Kung Fu Tea Claire's Korner Papa Bop Tapas Barcelona Taco Diablo LeTour Kabul House Tealicious	Picnic Union Pizzeria Sarah's Brick Oven Nakorn Hecky's BBQ Shangri-La Evanston Kung Fu Tea Joy Yee Noodle Kansaku Fonda Cantina

Table 3: Results of the 2 collaborative filtering scripts

While the above approach quantifies the similarity between users in terms of demographics, we can also achieve that by examining the reviews text they provided. In this approach, a user's feature vector is represented by a 1×63 reviews vector. Each component of the vector corresponds to the BERT embedding of their review text for a particular restaurant, or the embedding for 'missing' if no review is available. However, cf1 appears to have a higher degree of overlap with the results obtained from content-based filtering in the previous section (Table 1). This difference in performance could potentially be attributed to the ineffectiveness of the BERT embedding method. Since a significant number of 'missing' reviews are involved in the relatively small size of text data used here, the metric used to determine the similarity between users' tastes (based on BERT distance) may not perform well: It is possible that the presence of numerous missing reviews in common among users leads to the inaccurate conclusion of similar tastes, thereby affecting the overall effectiveness of the metric.

6 Predictive modeling

To start with, we wrote the most naive model attempting to use only the users' demographic data, along with the cuisine type for a restaurant, to predict the rating of a particular user on a certain restaurant, assuming the rating is simply a linear function of the parameters mentioned. The performance of such model using train/test split (with 5:2 train-to-test ratio) is unsurprisingly horrible, with its $R^2 = 0.00955$ for testing set.

To prevent over-fitting and boost features selection, we then tried adding an L1 penalty to the above linear regression model, i.e. training a Lasso model. The results are is slightly better with $R^2 = 0.0195$ for the testing set.

Has Children?: No	0.176
Average Amount Spent: Low	-0.165
Height (in)	0.006
Birth Year	0.002
Marital Status: Married	0.002

Table 4: top 10 Weightings for the linear model in predicting shop rating

We lastly investigated in if the demographic features are useful for predicting coffee scores. The R^2 of the testing set is 0.234 when we select the hyper-parameter of the L1 penalty to be 0.01. We looked into the weights of the model for intuitions in what features might be helpful or provide information in predicting the ratings. Having no children might be a significant feature for coffee shop lovers. The preferred mode of transport is seemingly useful in giving information on this problem, where we might conclude that people who take public transit are more likely to rate a coffee shop with high score while those who own a car might do the opposite. Northwestern students may tend to give higher ratings for coffee shops too.

Has Children?: No	0.822
Vegetarian?: missing	-0.769
Preferred Mode of Transport: Public Transit	0.530
Preferred Mode of Transport: Car Owner	-0.321
Marital Status: Married	-0.206
Northwestern Student?: No	-0.200
Average Amount Spent: Low	-0.188
Average Amount Spent: Medium	0.0645
Birth Year	0.0288

Table 5: Weightings for the linear model in predicting coffee shop rating

7 Discussion

Finally, as a side note, we present some interesting findings on the data set. User Melody Smith rated the restaurant Evanston Chicken Shack 15 times in 2022 mid January to early February, with every of them being the lowest score, 1. We find it amusing that even though she hated the restaurant so much, she has to put up with it that many times and keep going back, and keep writing negative reviews. Another finding is that, almost half of the review texts with rating score as 5 contain an exclamation mark in it, while for the others all are below 20% It is possible to be explained by the more general excitement found among people who just had a great dining experience.