Recommender for Restaurants in Mexico

Data and Strategy

This project aims to create a recommendation system for various restaurants in the surrounding cities of Mexico City. Data was gathered from an open content directory called Chef Moz. This data contained 1160 ratings from 133 users who rated a total of 87 restaurants. These restaurants were located in three cities in central Mexico, San Luis Potosi, Ciudad Victoria, and Cuernavaca. The data was gathered from https://www.kaggle.com/uciml/restaurant-data-with-consumer-ratings?select=README.

The goal is to be able to input a specific user's ID number, and create recommendations for that user. We also will be able to input a restaurant and find its similar restaurants.

The two methods that will be used are collaborative filtering and content-based filtering. Collaborative filtering works based on the assumption that people who agreed in the past will agree in the future, and they will like similar kinds of items as they liked in the past. Therefore, this method incorporates both the ratings of the target user and similar users to them. Content-based filtering works based on both a description of the items, restaurants for this project, and also the profile of the user's preferences. This method treats the recommendation as a user-specific classification problem and learns a classifier for the user's like and dislike based on product features.

The data was separated into three dataframes: one for the restaurant ratings, 1160 rows, one for user information, 410 rows, and one for restaurant information, 230 rows.

The ratings included overall restaurant ratings, as well as service and food ratings, all ranging from 0 to 2. Each user rated at least 3 restaurants and 8 restaurants on average. The average overall restaurant rating was 1.20. Service was rated slightly lower overall compared to food.

User information was gathered and analysed to see what type of people were rating the restaurants. User preferences such as ambience, cuisine, budget, transportation and dress as well as attributes such as marital status, age, birth year, kids, drink level and even personality, height and weight were examined. 17 missing values were filled in based on inferences. For example, for missing 'budget' information, 'medium' was filled in and for missing 'transport' information, 'on foot' was filled in. Most reviewers were in their 30s, preferred a family ambience, Mexican food, and had no preference on dress. We did see relationships between attributes such as married users had kids, single users were more likely to be social drinkers, and car owners were non-smokers.

For the 87 restaurants, we looked at city, dress code, price, cuisine, ambience, price, accessibility, alcohol options, smoking areas, parking options, and whether they were franchises or not. Restaurants were mostly Mexican, bars/breweries, Japanese, fast food, pizzerias, American, or other international restaurants. They mostly had a familiar ambience and were priced mid-range. The majority of restaurants were in San Luis Potosi.

Recommenders

First, a simple recommender was created that weighted ratings based on how many users rated a particular restaurant. Then, recommendations were pulled from 50% of top-rated restaurants overall. This simple recommender can then be used to find top restaurants according to certain criteria, such as city, ambience, and cuisine.

Second, user-based collaborative filtering along with predictive algorithms were used to create recommendations. After testing ten different algorithms including KNN-based and Matrix-Factorization-based, it was found that Single Value Decomposition (probabilistic matrix factorization), SVD++ (which takes into account implicit ratings), KNN with Means (using mean ratings of each user), and KNN with ZScore (takes into account the z-score normalization of each user) were the best. This valuation was based on root mean squared error and mean absolute error scores of cross-validated testing. The SVD algorithm was used to create the recommender. After tuning the algorithm's hyperparameters, the best rmse was .746 and the best mae was .590. To increase the quality of this recommender, it may be beneficial to gather more user ratings and for more restaurants in the future. This recommender employs specific functions and cosine distances between restaurant vectors. This recommender may now be used to find similar restaurants to a given restaurant. Also, it can gather recommendations for a specific user by finding this user's top rated restaurants and finding similar restaurants to those.

Third, a content-based recommender was built that was based on vectorized attributes of each restaurant. We used jaccard similarities and cosine distances between restaurant vectors to find similar restaurants to any given restaurant.

In the future, ideally more ratings data would be gathered for these users to create more robust recommendation systems. With more recommendations, these systems will output a higher quantity of restaurants with a higher accuracy. With more output, these restaurants could then also be compared to our user profile information (not just their ratings) to create even more tailored recommendations.