Team Members: Mallory Sico and Aryan Jain

Milestone 1 Report

Motivation

Main question: Where (location: city and area) and what type of (cuisine type, budget range, restaurant type) restaurant should we open in Bangalore?

Overarchingly, we are interested in this because: We want to see if we can use data science to accurately predict the success of a restaurant based on ratings and reviews of restaurants from 5 years ago. This will be interesting to compare with the restaurants that are currently open to see if the same success metrics would apply today.

In initial conversations about answering these questions, we each came up with a preliminary intuitive approach of how we would use this data. Both of our initial intuition was to focus on location first, either to find the part of the city with the most restaurants or the highest ratings. The next step would be to look at the other three variables (cuisine, budget range, and restaurant type) to find the ones that appear the most in the dataset or that have the highest ratings. This intuitive approach produced the following results:

Aryan:
Location: BTM
Cuisine: Asian, Chinese, Thai, Momos
Type: Buffet
Price Range: $$$$

Mallory:
Location: Church Street
Cuisine: North Indian
Type: Delivery
Price Range: $$

From this result, it is clear that there are many ways to approach this question which will result in different answers. This approach does not consider underlying relationships in the data that can skew the results. These are the problems that we aim to resolve in milestone 2 and 3 of the project.

Dataset

The dataset for this project comes from Kaggle[1]. It is the “Zomato Bangalore Restaurants” dataset. It includes information about restaurants in Bangalore, including URL, address, restaurant name, ability to order online, ability to book a table, rate, votes, phone number, location, restaurant type, most liked dish, cuisines, approximate cost per person, review highlights, type of food it is listed as on website and location it is listed in on website.

The main metrics that we will focus on are ratings, location identifiers, restaurant types, cuisines, and approximate cost per person. These are appropriate to answer our question because in an intuitive sense, they capture characteristics that are important in having a successful restaurant from both the owner’s and the consumer’s point of view.

Success Metric: The ratings will be used as the success metric for this project. Initially, the number of votes was considered as a success metric. However, upon exploration of the data, the distribution of votes is heavily skewed as more than 50% of the data had fewer than 50 votes as shown in Figure 1a. Additionally, older restaurants and restaurants with higher ratings are likely to have more votes. Due to

these issues, the ratings were chosen as the sole success metric. The ratings distribution (Figure 1b) is more evenly distributed and ratings are not highly dependent on the age of the restaurant.

Methodology

Milestone 1: In this phase, we have based our results on EDA and come up with an intuitive solution to our business problem. We began exploring correlations in the dataset as shown in Figure 2. Our thought process for the upcoming milestones is highlighted below.

Milestone 2: In this phase of the project, we will explore the following methods. Multicollinearity: As already identified, there is high collinearity between the votes and the ratings. The votes will not be used, but this is just one example of correlation that we would want to correct for in relevant metrics before making a decision.

Confounder Identification: We have previously seen the effect of Simpson's Paradox in class and how it can lead to misleading results. We want to adjust for confounders, so that we can answer questions such as "Are Italian restaurants rated higher or is it because they are always high-end establishments?"

Hypothesis Testing: We will check the significance of the results from questions such as "Is cuisine X really better than cuisine Y?"

Additionally, we will look to extract some more features, such as type of phone number (landline vs personal), if possible to see if they have any effect on ratings.

By the end of milestone 2, we aim to have a data-driven ideal configuration. We intend to use our preliminary intuitive approach as a starting point and will update it based on the results obtained throughout this cycle.

Milestone 3: In this phase of the project, we will build an ML model (linear regression and decision trees), to predict ratings and analyze feature importance. Using linear regression, we will predict ratings of the ideal configuration from milestone 2.

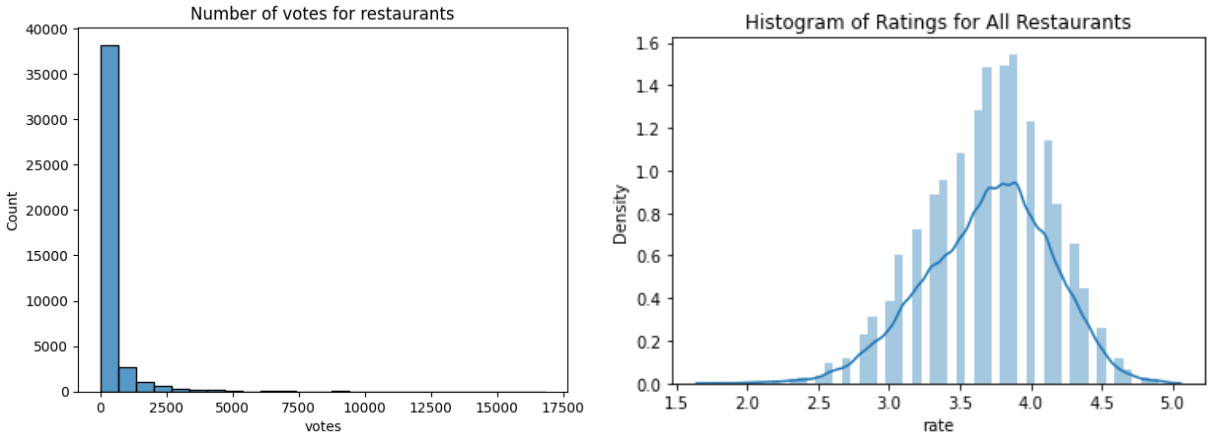
We will then use a decision tree to help us extract the ideal process for selecting features (For example: "Do we first select a location and then choose the other parameters or is cuisine the biggest differential in the success of a restaurant?"). From this decision tree, we will re-estimate the ideal configuration for the restaurant and predict its rating.

The evaluation metric will be the rating. A successful rating will be determined as being in the 90th percentile of the core four factors of location, cuisine, restaurant type and budget range.

Link to Data:

1) https://www.kaggle.com/datasets/himanshupoddar/zomato-bangalore-restaurants

Team Members: Mallory Sico and Aryan Jain



(a) (b)

Figure 1: (a) Number of Votes for Restaurants (b) Histogram of Ratings for All Restaurants

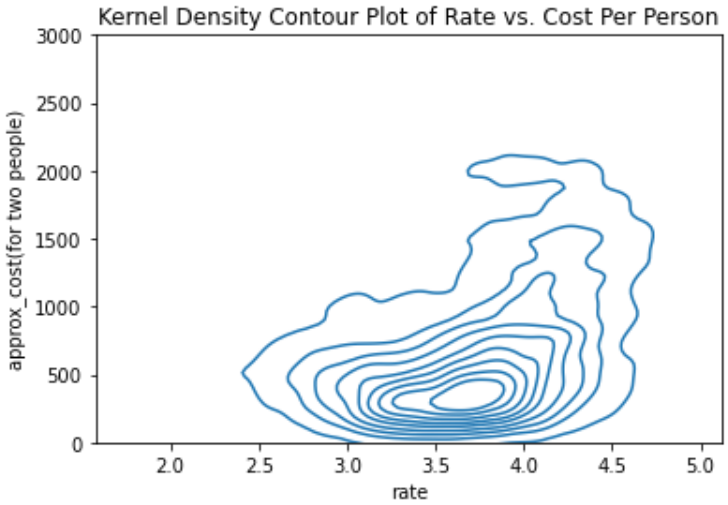


Figure 2: Kernel Density Contour Plot of Rate vs. Cost Per Person