HW3–Recommendation System Part II

May 13, 2025

1 Introduction

In this project, we extend our exploration of recommendation systems by focusing on collaborative filtering and predictive modeling techniques. The goal is to uncover meaningful insights and develop personalized recommendation strategies based on the Evanston restaurants and reviews dataset. The core of the project lies in implementing a content-based collaborative filtering system using linear modeling approaches.

2 Data Preprocessing

For the data preprocessing, we first removed duplicate rows to eliminate redundancy and prevent bias in the results. We then merged the restaurant and review datasets, retaining only the entries corresponding to restaurants that appear in both sources. To ensure consistency, we standardized string formats across key columns. Regarding missing values, we dropped the ”Vegetarian?” feature due to its high missingness rate of 93.77%. For categorical variables such as ”Marital Status” and ”Has Children?”, we applied mode imputation. For numerical variables like ”Weight (lb)” and ”Height (cm)”, we used the median for imputation to mitigate the influence of outliers.

3 Collaborative Filtering

3.1 User feature Matrix

First, I selected a set of demographic features and categorized them into numerical and categorical features, as shown separately in Table 1. To ensure each reviewer is only counted once—despite possibly having reviewed multiple restaurants—the code removes spaces from ’Reviewer Name’ and drops duplicate entries based on this cleaned identifier. Categorical variables are encoded using one-hot encoding to convert them into numeric form. Finally, the ’Reviewer Name’ column is dropped to produce the final feature matrix, where each row represents a unique reviewer and each column corresponds to a specific numeric or one-hot encoded demographic attribute. The number of unique users is 1,086 and the dimensionality of the resulting feature vectors is 16.

Feature Type	**Examples**
Numerical Features	Birth Year, Weight (lb), Height (cm)
Categorical Features	Marital Status, Has Children?, Average Amount Spent, Preferred Mode of Transport, Northwestern Student?

Table 1: Demographic features divided by type.

3.2 Recommendation System I

To compute how similar two users are, I applied cosine similarity to their feature vectors. Cosine similarity measures the angle between two vectors, and is especially useful because it is scale-invariant, so differences in absolute values (e.g., weight in pounds vs. height in cm) do not distort the similarity. And it works well with high-dimensional sparse vectors. The function recommend_by_cosine_similarity accepts a target user and performs a series of steps to generate personalized recommendations. First, it computes the cosine similarity between the target user and all other users based on their demographic feature vectors. It then identifies the top-k most similar users, excluding the target user themselves. For each of these top-k users, the function retrieves the restaurant or restaurants they rated the highest. Finally, it returns a dictionary that maps each similar user to their favorite restaurant(s) along with the corresponding rating. For the user 'Timothy Mace', the top 3 similar user and their favorite restaurants recommendations are shown in table 2. This algorithm not necessarily always suggest more than one recommendation for every user in the dataset. If a similar user has rated only one restaurant highly, or multiple users recommend the same place, the output may contain a single or small number of suggestions.

Similar User	**Favorite Restaurant**	**Rating**
Enid Egan	Burger King	5
Anthony Grieco	Union Pizzeria	5
Pankaj	Tapas Barcelona	5

Table 2: Top-3 most similar users to Timothy Mace and their top-rated restaurants.

3.3 Recommendation Systems II

To identify users with similar reviewing behavior, each user is represented as a vector where each entry corresponds to their rating for a specific restaurant. Since most users have not rated all restaurants, the resulting vectors contain many missing values.

To address this, missing entries are first filled using mean imputation, where each blank is replaced with the average rating for that restaurant across all users. This provides a complete, albeit coarse, initial matrix necessary for further processing. Next, Truncated Singular Value Decomposition (SVD) is applied to this matrix. SVD reduces the dimensionality of the data by projecting it into a lower-dimensional latent space, capturing the underlying structure in user preferences. The matrix is then reconstructed using the top components, resulting in a low-rank approximation that estimates the missing ratings more accurately than simple mean imputation. The final output is a fully completed user–restaurant rating matrix, where each user now has predicted ratings for all restaurants, even those they have never reviewed. These vectors can then be used to compute similarities between users based on their estimated preferences, rather than their demographic profiles.

The function computes the cosine distance between a given user and all others in the dataset. For a given target user, the function first retrieves the vector of predicted ratings, then compares it with all other users using cosine_distances. The top-k most similar users (those with the smallest distances, excluding the target user) are selected.

For each of these similar users, the algorithm identifies their favorite restaurant(s)—that is, the one(s) they rated the highest. These are compiled into a recommendation set for the target user.

As a demonstration, the algorithm was applied to the user Sarah Belle with top_k = 3. The output shows that her three most similar users are: *Steven Johnston*, *Johnny Mcginnis* and also *Steven Rusert.* All of them like the restaurant named as *Mumbai Indian Grill.*

Most Similar User	**Top Rating**	**Top Restaurants**
Steven Johnston	5	Mumbai Indian Grill
Johnny Mcginnis	5	Mumbai Indian Grill
Steven Rusert	5	Tapas Barcelona, Mt. Everest Restaurant, Mumbai Indian Grill, Taste of Nepal

Table 3: Top-3 most similar users to Timothy Mace and their highest-rated restaurants.

4 Predictive Modeling

4.1 Linear Regression

To enhance the input features, the age of each reviewer was calculated by subtracting their birth year from 2025, and an optional Body Mass Index (BMI) variable was computed from weight and height.

$$\text{Age} = 2025 - \text{Birth Year}$$

$$BMI = \frac{Weight(lb)}{Height(m)^2}$$

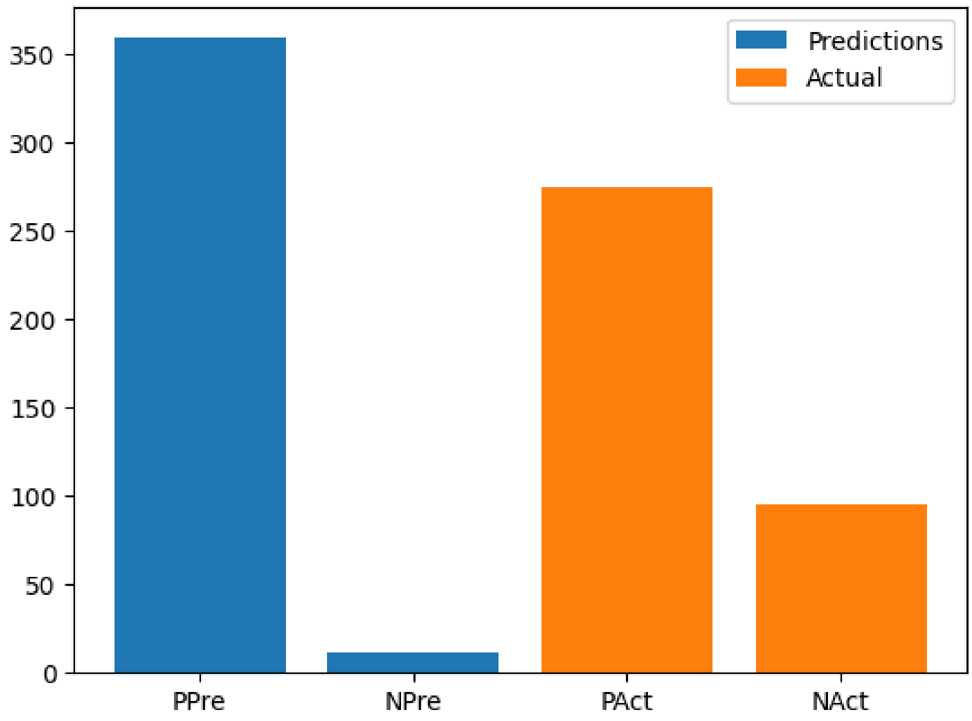

Figure 1: Logistic y predict vs actual

The selected predictors included both categorical variables (e.g., cuisine type, marital status, transportation preference) and numerical attributes (e.g., average cost, age, BMI). The data was then split into training and test sets using a random seed of 42, with the test set comprising 20% of the entire dataset. The model achieved a mean squared error (MSE) of **2.0106**and R^2 of **0.0826**.

4.2 Logistic Regression

I divided the attitude based on the users' rating for this restaurant. If it is greater than or equal to 3, we consider it positive(attitude = 1) else negative(attitude = 0). I set the same seeds and test size as the linear regression model and predict the probability of y to be positive(y=1). The MSE is equal to **0.2595**, which is much smaller than the regular linear regression model. And the log loss is equal to **0.5584**. I evaluated the model's performance on the test set and visualized the results. While the model captured the general pattern that positive reviews are more frequent than negative ones, it frequently misclassified negative reviews as positive. This misclassification led to an inflated overall positive rating across restaurants(shown in figure 1). To evaluate the model's ability to predict review sentiment using user demographics and restaurant attributes, I selected the third sample from the test set. This user reviewed a chocolate-themed restaurant with an average cost of 20, dined at a place open after 8 p.m., was married, had children, had a BMI of approximately 111.3, was 82 years old, traveled by car, and was not a Northwestern student. Based on these features, the logistic regression model predicted a high probability of positive sentiment—approximately 85.6%—resulting in a predicted label of 1 (positive). This prediction matched the actual label, indicating that the model was able to accurately infer review sentiment from structured features. While this is just a single instance, it demonstrates the model's capacity to generalize unseen data when individual demographic and contextual factors are provided.

4.3 Coefficient Analysis

My favorite model is Logistic regression model because it has higher accuracy scores. I changed the penalty into L1 penalty and did the regression again. By changing the penalty, we found the predictive ability is a little bit lower than the previous logistics regression model using the default penalty L2. MSE is higher than the model with default penalty L2(0.2621 > 0.2595). Log Loss is 0.5557 slighly lower than the 0.5584. But the model is much better than the linear model which has the MSE equal to 2.0106. Since the L1 penalty has the power to select

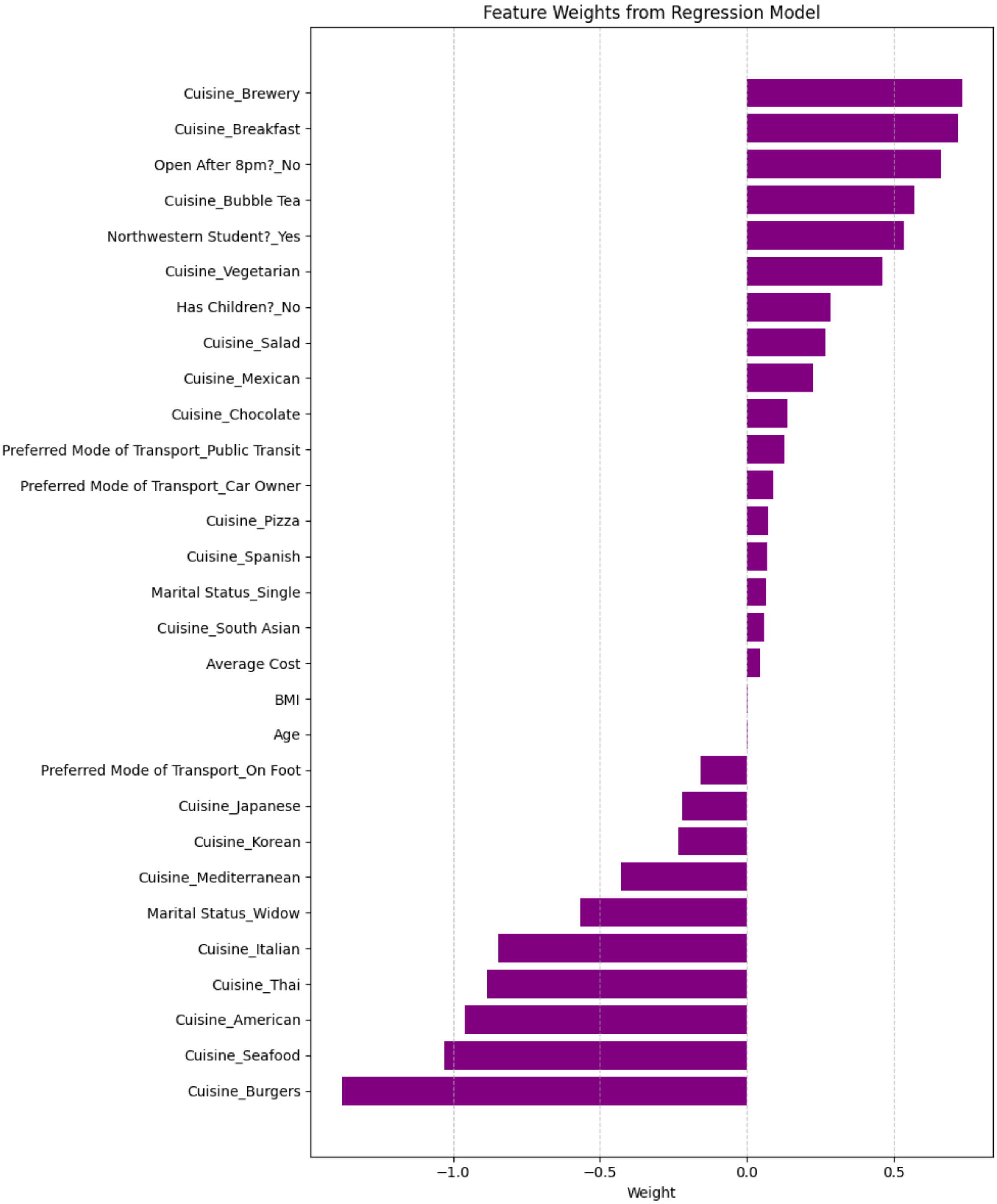

Figure 2: Feature weights from the regression model

features, we output the weight of each output and displayed below at table 5 and figure 2. For the positive weights, Cuisine_Brewery has a large positive weight (0.733), meaning reviews associated with brewery-type restaurants tend to be more positive. Northwestern Student?_Yes (0.535) suggests that Northwestern students tend to give higher ratings. Good restaurants are usually closed before 8pm in the evening. For the negative weights, Cuisine_Burgers has a large negative weight (-1.38), suggesting burger restaurants are more likely to receive lower ratings. Marital Status_Widow (-0.567) implies this demographic tends to give slightly lower ratings. I also found that features like Age, BMI, and Average Cost have relatively small coefficients, suggesting they have a weaker influence on users' attitudes toward restaurants.

Index	**Feature**	**Weight**
1	Cuisine_Burgers	-1.379825
2	Cuisine_Seafood	-1.028673
3	Cuisine_American	-0.962295
4	Cuisine_Thai	-0.884677
5	Cuisine_Italian	-0.847409
6	Marital Status_Widow	-0.567004
7	Cuisine_Mediterranean	-0.429681
8	Cuisine_Korean	-0.232179
9	Cuisine_Japanese	-0.219727
10	Preferred Mode of Transport_On Foot	-0.155682
11	Age	0.002593
12	BMI	0.002866
13	Average Cost	0.045222
14	Cuisine_South Asian	0.057750
15	Marital Status_Single	0.067353
16	Cuisine_Spanish	0.069107
17	Cuisine_Pizza	0.072010
18	Preferred Mode of Transport_Car Owner	0.090199
19	Preferred Mode of Transport_Public Transit	0.129375
20	Cuisine_Chocolate	0.138760
21	Cuisine_Mexican	0.227989
22	Cuisine_Salad	0.267828
23	Has Children?_No	0.286672
24	Cuisine_Vegetarian	0.461732
25	Northwestern Student?_Yes	0.535825
26	Cuisine_Bubble Tea	0.570746
27	Open After 8pm?_No	0.662238
28	Cuisine_Breakfast	0.718925
29	Cuisine_Brewery	0.733206

Table 4: Feature weights from the regression model

Coffee Example I used the demographic features that mentioned above and did the linear regression model for reviews samples of all coffee shops in the dataset, which contains 66 records in the dataset. I computed the mean scores of each coffee shops and did the regression again. The Brothers K Coffeehouse got the predicted average rating equal to 4.35, Philz Coffee got the average rating equal to 4.54 and Pâtisserie Coralie got the average rating equal to 4.25. Then I examined the weights produced by the linear model and showed in the table 4 below. The weights from the linear model reveal clear patterns in coffee shop preferences based on demographic features. Individuals without children(0.6853), Northwestern students(0.4614), and those who prefer coffee shops open after 8 p.m. (0.5566) show strong positive associations with coffee preference, as indicated by relatively large positive weights. Conversely, individuals who walk as their primary mode of transport, those with children, and non-students show strong negative associations, suggesting they are less likely to prefer coffee (–0.695, –0.685, and –0.461, respectively). Married individuals and car owners also show mild positive associations, while features such as age, weight, and height have insignificant influence on preference. The features Average Cost has no effect on coffee shop preferences and can be excluded when predicting coffee shop ratings. These results suggest that coffee preference is more common among students, childless individuals, and those who drive or value late-night availability, while walkers, parents, and non-students tend to show lower preference for coffee.

Feature	**Weight**
Average Cost	0.0000
Age	0.0081
Weight (lb)	0.0050
Height (cm)	0.0052
Open After 8pm?_Yes	0.5566
Marital Status_Married	0.1437
Has Children?_No	0.6853
Preferred Mode of Transport_Car Owner	0.1082
Preferred Mode of Transport_On Foot	-0.6953
Preferred Mode of Transport_Public Transit	0.0044
Northwestern Student?_Yes	0.4614

Table 5: L1-Regularized Logistic Regression Feature Weights

5 Text Embedding

In this step, we leverage the semantic information in the review text to predict sentiment using text embeddings. Specifically, we use Sentence Transformers to convert each review in the 'Review Text' column into a fixed-length embedding vector that captures its semantic meaning.
The model 'all-MiniLM-L6-v2' from the SentenceTransformers library is used to encode the reviews. These embeddings are then used as input features (X) for a logistic regression model, with the target variable (y) being the binary sentiment label (attitude), which indicates whether a review is positive or negative.
The data is split into training and test sets, and logistic regression is trained with an L1 penalty to compare performance fairly with earlier models that used L1 penalty. After training, the model's performance is evaluated using Mean Squared Error (MSE) and accuracy. The MSE value is equal to 0.1391 while the accuracy is equal to 0.8041. The MSE value is lower than the linear model(2.0106) and logistic regression model(0.2622), which means the textual information contains much more useful information than we thought before. And sentence-transformer model can leverage the power of the digging out useful information from the natural language.

6 Conclusion

There are many interesting things in the dataset. I chose to analyze the trend for the northwestern students. I analyzed the mode of five different categorical features among Northwestern students. The results show that most Northwestern students prefer Mexican cuisine, such as Chipotle and Chili's. In terms of lifestyle, the majority are single, do not have children, and prefer walking as their main mode of transportation. Additionally, they tend to favor restaurants that are open after 8 PM, indicating a preference for late dining options.