Project Team Members: Aryan Jain and Mallory Sico

Milestone 3 Report

Methodology

ML Modeling: From Milestone 2, we suggested several restaurant configurations based on location, cuisine, price range and serving type analysis. For the third milestone, the approach was to model the ratings based on these four categories.

The approach to modeling the restaurant ratings was done in two steps. The first step was regression analysis to attempt to predict actual ratings. The data was pre-processed for the regression model by eliminating any categories that had fewer than 50 restaurants. A linear regression was used as the baseline for this analysis with a mean absolute error (MAE) of 0.33. This was compared with decision tree models, both default and with optimized hyperparameters.

The second step was to convert the ratings into binary classes, 1 class for restaurants in the top 90% of ratings (4.2) and 0 for all others. Preprocessing for this step was done by oversampling the minority class due to the 90/10 split of the classes. The binary classification was modeled using logistic regression, linear regression, decision tree classifier, ridge regression, and a catboost model. F1-score was the evaluation metric for determining the best model. Additionally, the data was condensed to specific categories to determine best model fitting for those categories.

Reviews Analysis: Outside of the four initial categories, there are likely many other factors that affect restaurant rating. The dataset includes written reviews for each restaurant and the rating associated with that review. This data was used to get a better understanding of the additional factors that are important to customers and therefore should receive special consideration from those looking to open a restaurant.

To analyze the review data, the reviews were separated into 5 stars and 1 star datasets. The words from the 5 star and 1 star datasets were then put into word clouds to visualize the most important words for each set. Based on the counts of these words on the reviews, we can calculate an estimate of what is important to people in Bangalore. Based on domain knowledge and the wordcloud, the words were divided into 5 categories. Following the categorization, the ratio of the topics was extracted for all 5 star and 1 star reviews using the following formula: (Count of word in all reviews of restaurant)/(Number of reviews of restaurant).

In this analysis, we assume that all reviews that are rated 5 stars are positive. That may or may not be the case. To correct for this, a sentiment analysis is performed through pre-trained language models, Textblob and VADER, to get sentiment scores and choose one model based on the results and online research. Following this, two approaches were taken to determine the most important category of words. These approaches were 1) direct correlation with ratings and 2) using feature importances. The first approach was performed by finding the correlation between ratings and the number of reviews associated with 5 or 1 stars. Additionally, a correlation matrix shows the correlation between the categories. The second approach was performed using OLS and Ridge Regression models to find the model coefficients of the categories.

Results

ML Modeling: The original baseline model was a linear regression model with MAE of 0.33. After eliminating categories that had too few restaurants, MAE improved to 0.28. After optimizing hyperparameters for the decision tree, the best MAE for the model still did not improve over the linear regression model.

After pre-processing for binary classification, the metrics from the logistic regression, linear regression, decision tree classifier (default and optimized), ridge regression (default and optimized), and catboost are shown in Table 1 for the full data set. Logistic regression was used as the baseline for binary classification with a baseline f1_score of 0.426. Catboost improved on

this up to f1_score 0.48, but this is still not sufficient for a strong prediction. In earlier milestones, we saw that price is an important confounding variable, so the same models were trained on dataframes that were limited by price range. The best performing model was the catboost model on the price range $$$$ category as shown in Table 2. The f1_score for that model was 0.831. The confusion matrix for both the overall data and the price range $$$$ catboost model validation predictions are shown in Figures 1 and 2. Catboost was chosen as the best model and when predicting the test data, the test score was 0.793 for all data with f1-score of 0.479. When testing only the price range $$$$ data, the test score was 0.511 and the f1-score was 0.614. The milestone 2 suggestions and their binary predictions are shown in Table 4.

Reviews Analysis: From the word clouds, shown in Figures 3 and 4, The words were divided into the 5 categories with the associated popular words in 5 star reviews as shown in Table 3.

Figure 5 shows the sentiment scores from VADER and TextBlob. We can see that VADER provides better estimates for more results (stark difference in the median line between the 2 models) as we assume all reviews to be positive. Therefore we will proceed with Vader for our thresholding criterias. Based on research regarding the two processes, the general consensus seems to be that VADER is better suited for "social media sentiment analysis" and our reviews would more fall under that category. We also decided to use compound scores (instead of positive or negative scores) as we are gonna be thresholding on that.

To determine the most important features, the first approach was direct correlation with ratings. The correlation coefficients between ratings and the number of 5 or 1 star ratings per category are shown in Table 3.

The second approach, feature importance was evaluated using the correlation matrix shown in Figure 6. There is a high correlation between “vibe” and “service”. “Service” and “group” are also highly correlated. The feature coefficients are shown in Table 3 for OLS and Ridge Regression models.

Analysis

ML Modeling: Because the binary classification model for all data performed poorly, we did not use that model to predict the ideal restaurant configurations as new data points. Instead we used the model for just $$$$ price range. For ideal restaurants that were suggested in milestone 2, these were input into the model for prediction. The model predicted that two of the three would be in the 90th percentile, but the model test score shows that the performance does not generalize all that well. It is likely that there are other factors influencing the ratings. Because of this, the reviews analysis was performed to give more insight into important factors.

Reviews Analysis: From this analysis, we make an interesting insight that "delivery" is one of the biggest culprits in getting the 1 star ratings. This is something that is generally not in the hands of the restaurant and more often than not is a fault of Zomato (the delivery partner). This tells us that when setting up the restaurant ensure that you have **great packaging and that you do not deliver very far** (This would lead to the food being served cold).

From the correlation analysis and the feature importances from linear models, we see that vibe is the dominant feature with service also contributing more than the remaining features. The takeaway from this analysis is that Vibe matters the most to customers when they decide to give a 5 star rating followed by Service.

Project Team Members: Aryan Jain and Mallory Sico

Table 1 : Models for Binary Classification for All Data

	train_score	val_score	auc_score	precision	recall	f1_score
LinearRegression	0.555	-0.266	0.264	0.137	0.550	0.706
LogisticRegression	0.865	0.839	0.160	0.128	0.566	0.719
Decision Tree Classifier	0.97	0.828	0.171	0.139	0.558	0.709
Optimized_ Decision Tree	0.921	0.837	0.162	0.125	0.525	0.699
Ridge	0.551	-0.263	0.264	0.134	0.550	0.707
Optimized Ridge Classifier	0.860	0.837	0.162	0.129	0.558	0.714
catboost	0.944	0.787	0.124	0.086	0.550	0.731

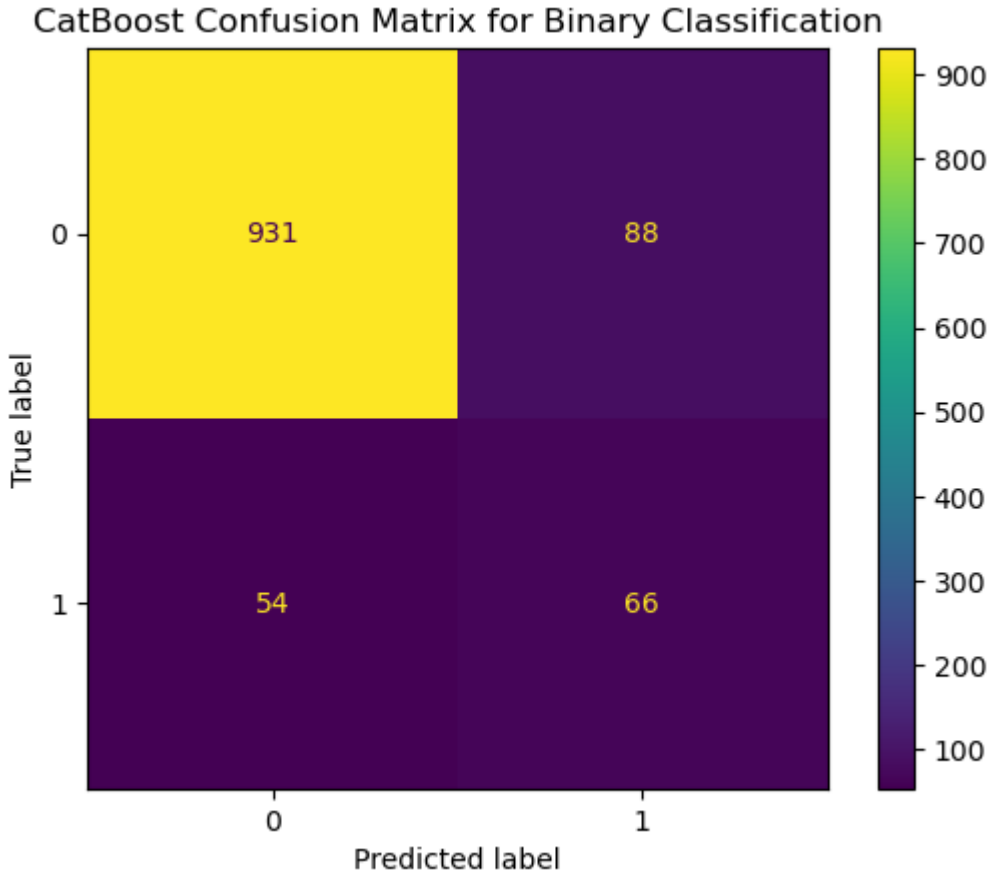

Figure 1: Confusion Matrix for All Data

Table 2: Models for Binary Classification for Only Price Range $$$$

	train_score	val_score	auc_score	precision	recall	f1_score
LinearRegression	0.264	0.191	0.389	0.312	0.655	0.671
LogisticRegression	0.735	0.688	0.311	0.265	0.637	0.686
Decision Tree Classifier	0.978	0.614	0.385	0.406	0.637	0.615
optimized_Decision Tree	0.636	0.655	0.344	0.625	0.965	0.670
Ridge	0.260	0.225	0.384	0.312	0.655	0.671
Optimized Ridge Classifier	0.728	0.680	0.319	0.265	0.620	0.677
catboost	0.859	0.501	0.172	0.250	0.913	0.831

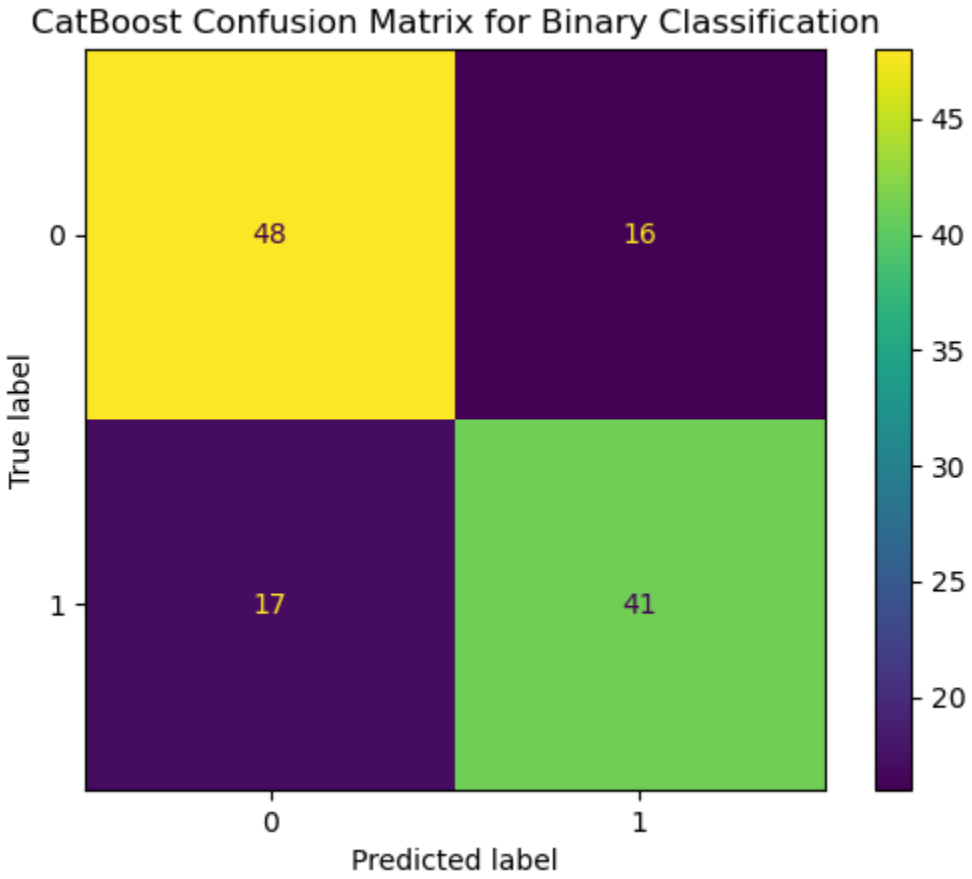

Figure 2: Confusion Matrix for Data Limited to Price Range $$$$

Project Team Members: Aryan Jain and Mallory Sico

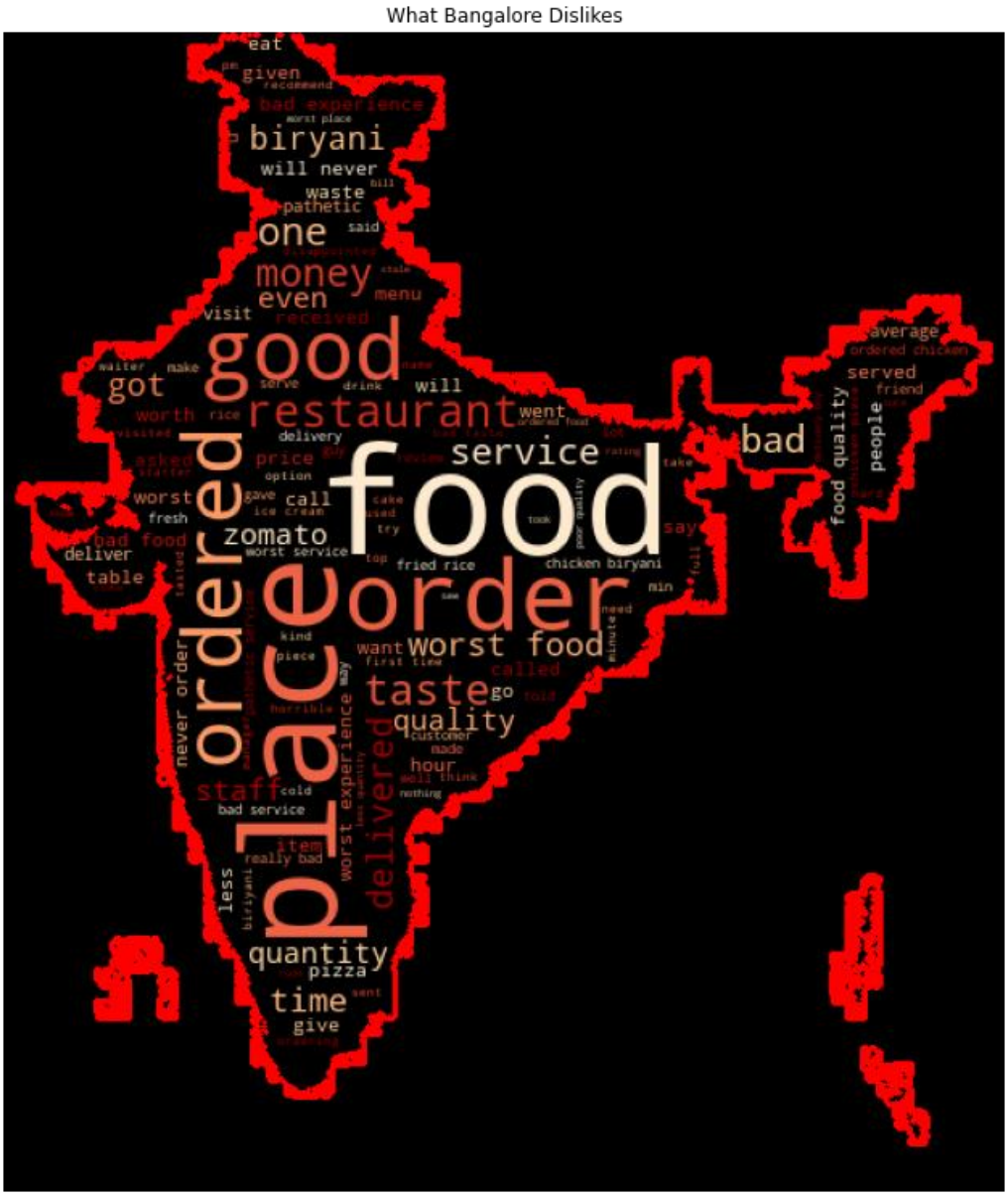

Figure 3: Word Cloud for 1 Star Ratings

Project Team Members: Aryan Jain and Mallory Sico

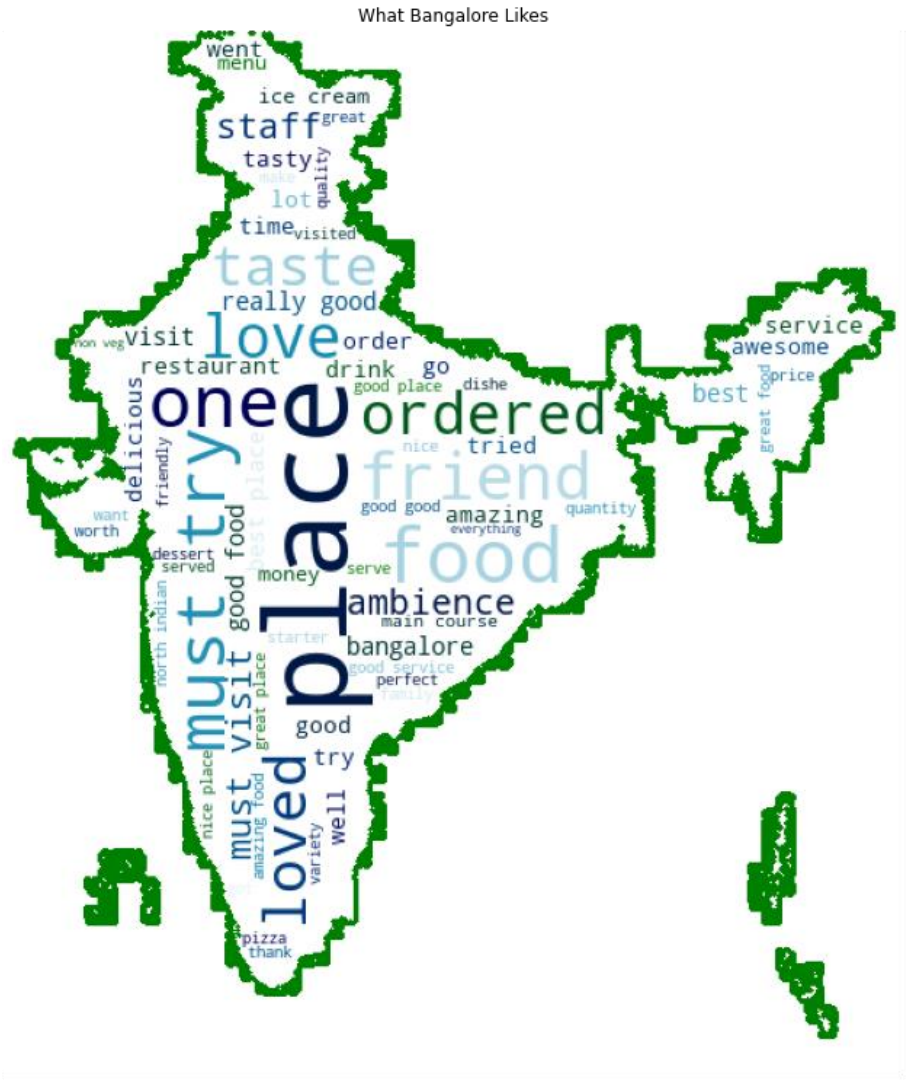

Figure 4: Word Cloud for 5 Star Ratings

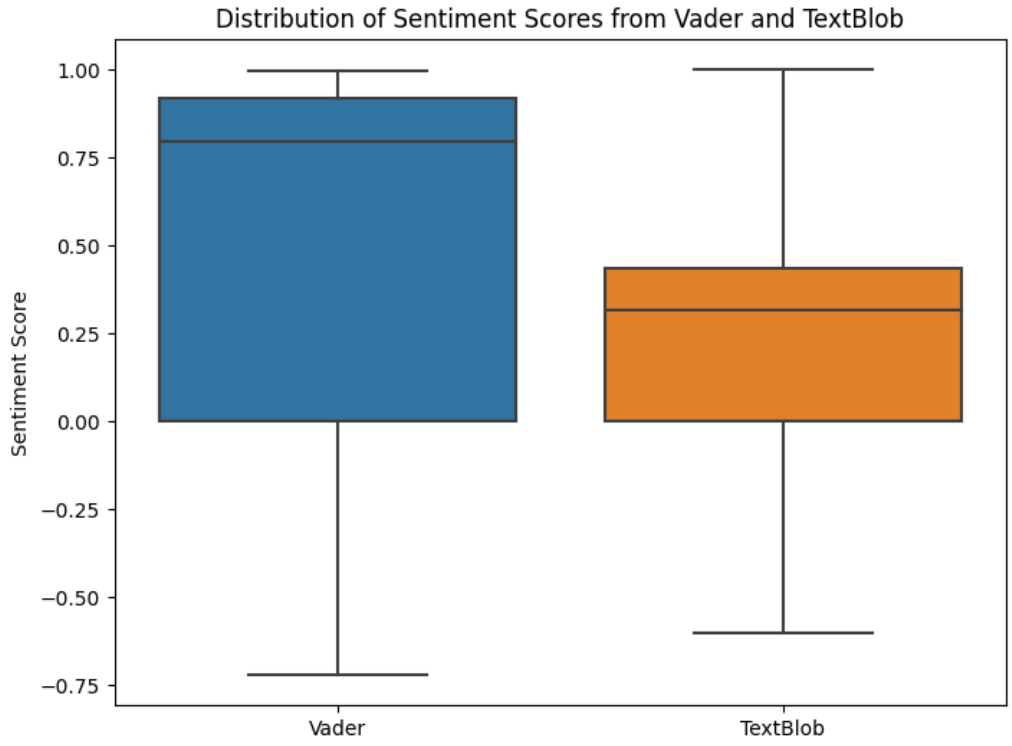

Figure 5: Distribution of Sentiment Score from Vader and TextBlob

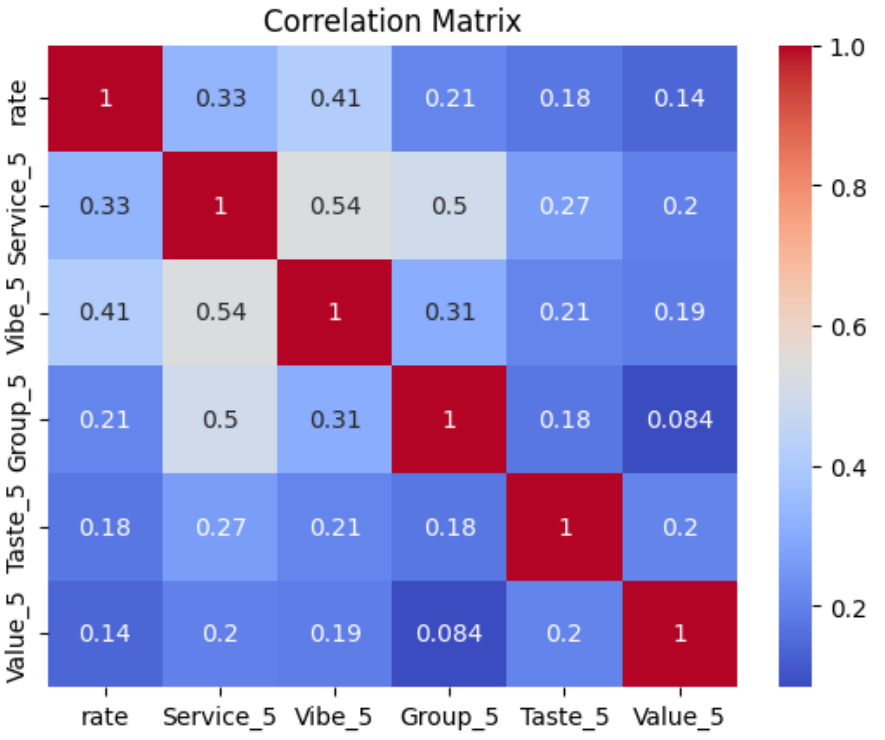

Figure 6: Correlation Matrix of Review Analysis and Ratings

Table 3: Review Analysis Category, Correlation and Feature Importance

Category	Associated Words	# Stars for Review in which category was analyzed	Correlation between ratings and number of X star ratings	Feature Importance with OLS	Feature Importance with Ridge Regression
Service	staff, service, friendly, good service	5	**0.404**	**0.163**	**0.080**
Vibe	vibe, ambience, ambiance	5	**0.509**	**0.489**	**0.329**
Groups	friend, friends	5	0.235	0.079	0.004
Taste	Taste, tasty, great food, good food, delicious	5	0.069	0.083	0.008
Value	Money, worth	5	0.151	0.108	0.014
Delivery	Delivery	1	-0.186	N/a	N/a

Table 4: Milestone 2 Suggestion Predictions

Location	Cuisine	Price Range	Prediction 1 - In the 90th percentile of ratings 0 - not in 90%
BTM	Continental	$$$$	1
Koramangala 5th Block	North Indian	$$$$	1
HSR	Cafe	$$$$	0