Zomato Restaurant Clustering and Sentiment Analysis

Midhun R
Data Science Trainee,
AlmaBetter, Bangalore

Abstract:

Every business needs to keep up with their customer demands, or they risk losing loyal customers. Clustering of products can help customers in finding similar products effectively and by identifying the customers' emotions, you can get a better idea of their experience and provide better customer service, which eventually leads to a decrease in customer churn.

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups.

Sentiment analysis is the scanning of words written or spoken by a person to determine the emotions they're most likely feeling at the time to give businesses a better read on their customers.

In this project, I have attempted to analyze the metadata and reviews of popular restaurants in Hyderabad and build machine learning models to cluster the restaurants into different segments based on cuisines and analyze the sentiments of the reviews given by the customers.

Keywords: EDA, Feature Engineering, Modelling, Regression, XGBoost, Naïve Bayes, SVM, K Means, DBSCAN

01. Problem Statement

Zomato is an Indian restaurant aggregator and food delivery start-up founded by Deepinder Goyal and Pankaj Chaddah in 2008. Zomato provides information, menus, and user-reviews of restaurants, and also has food delivery options from partner restaurants in select cities.

Two datasets are given: one with metadata of 105 restaurants and the other with reviews for these restaurants. The main objective is to understand the existing data and analyze their trends and patterns, so that machine learning models can be built, one for segmentation of restaurant types and another for sentiment analysis of reviews.

02. Introduction

In this day and age, when data science has transformed into an invisible force that influences the decisions of the modern world, food industry has also become beneficiary of data science.

Nowadays, having a proper understanding of data and connecting all data sources effectively is paramount in generating competitive advantage, providing superior customer value, and ultimately orientating the future of any business. Over the last few years, companies have begun to deploy

clustering and sentiment analysis for better customer experience.

India is quite famous for its diverse multi cuisine available in a large number of restaurants and hotel resorts, which is reminiscent of unity in diversity. Restaurant business in India is always evolving. More Indians are warming up to the idea of eating restaurant food whether by dining outside or getting food delivered. The growing number of restaurants in every state of India has been a motivation to inspect the data to get some insights, interesting facts and figures about the Indian food industry in each city. So, this project focuses on analyzing the Zomato restaurant data for each city in India.

The Project focuses on Customers and Company, the sentiments of the reviews given by the customer in the data should be analyzed and useful conclusion in the form of visualizations should be made. Also, clustering of restaurants into different segments should be carried out.

This could help in clustering the restaurants into segments. Also, the data has valuable information around cuisine and costing which can be used in cost vs. benefit analysis. Data could be used for sentiment analysis. Also, the metadata of reviewers can be used for identifying the critics in the industry.

03. Approach

The sequence of steps taken to solve the task are as follows:

1. Understanding the business task.
2. Reading data from files given and providing a summary.
3. Data cleaning, which involves removing irregularities in the data.
4. Exploratory data analysis, to find which factors affect sales and how they affect it.
5. Feature engineering, to prepare data for modelling.
6. Text Processing, to convert text to numeric data for modelling.
7. Modelling data (for both clustering and sentiment analysis) and comparing the models to find out the most suitable one for forecasting.
8. Conclusion.

04. Business Task

Build machine learning models to cluster the restaurants into different segments based on cuisines and analyze the sentiments of the reviews given by the customers.

05. Data Summary

After drive was mounted, data from csv files were read and stored in pandas dataframes. After loading the datasets, the datasets have been explored thoroughly by looking into its head, brief summary, number of records and features, etc.

The restaurant metadata contains 105 records across 6 features with information about each restaurant and the review data contains 10000 records across 7 features with information regarding the reviews given to these restaurants by various reviewers. The common feature among them is restaurant name. The restaurant metadata doesn’t have any duplicate rows, but reviews dataset have 36 duplicate rows. Two columns in restaurant metadata and five

columns in review dataset have missing values. Two columns in restaurant metadata and three columns in review dataset require conversion of datatypes.

The features in the restaurant metadata were identified:

1. Name: Name of Restaurants.
2. Links: URL Links of Restaurants.
3. Cost: Per person estimated Cost of dining.
4. Collection: Tagging of Restaurants with respect to Zomato categories.
5. Cuisines: Cuisines served by Restaurants.
6. Timings: Restaurant Timings.

The features in store dataset were identified:

1. Restaurant: Name of the Restaurant.
2. Reviewer: Name of the Reviewer.
3. Review: Review Text.
4. Rating: Rating Provided by Reviewer.
5. Metadata: Reviewer Metadata - No. of Reviews and followers.
6. Time: Date and Time of Review.
7. Pictures: No. of pictures posted with review.

06. Data Cleaning

Data cleaning is done to ensure that the dataset is correct, consistent, and usable.

There are no duplicate rows to remove in restaurant metadata, but it was found that there are 36 duplicate rows in review dataset. They were removed.

Two out of 6 columns in the restaurant metadata had missing values in them.

```
Collections: 54(51.43%)
Timings: 1(0.95%)
```

More than half of the observations of restaurant metadata have missing values in 'Collections'. Imputing missing values in this feature is possible only by collecting more data. So, this feature was removed since it will lead to inaccurate data analysis.

There is only 1 missing value in 'Timings' and it was imputed with its mode.

Five out of 7 columns in the review dataset have missing values in them.

```
Reviewer: 2(0.02%)
Review: 9(0.09%)
Rating: 2(0.02%)
Metadata: 2(0.02%)
Time: 2(0.02%)
```

Since this is a very small number of data, these observations were removed from the dataset.

The datatype of 'Cost', 'Cuisines' (in restaurant metadata), 'Rating', 'Metadata' and 'Time' (in review dataset) was converted to int, list, float, dictionary and datetime respectively.

07. Exploratory Data Analysis

Dataset, after cleaning, is subjected to exploratory data analysis, which visualizes the data and identifies trends and patterns.

The following observations were made after EDA:

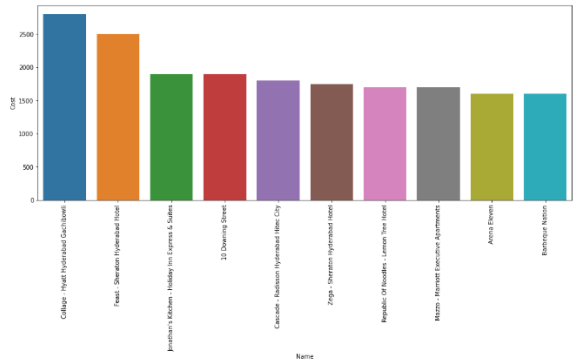


Figure 1: Most expensive restaurants

Collage - Hyatt Hyderabad Gachibowli is the most expensive restaurant followed by Feast - Sheraton Hyderabad Hotel.

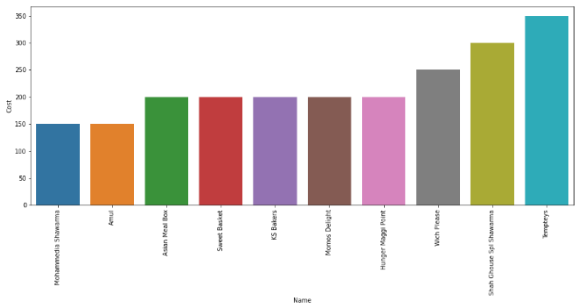


Figure 2: Least expensive restaurants

Mohammedia Shawarma and Amul are the most affordable restaurants among the 105 restaurants in the dataset.

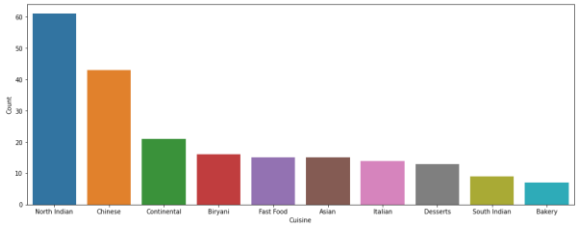


Figure 3: Most popular cuisines

North Indian is the most popular cuisine by availability, followed by Chinese.

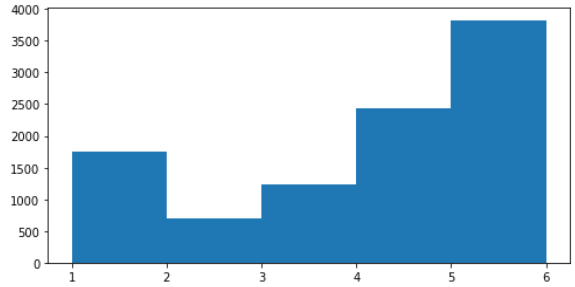


Figure 4: Distribution of ratings

It is found the interval between 4 and 5 has the most frequency of Rating.

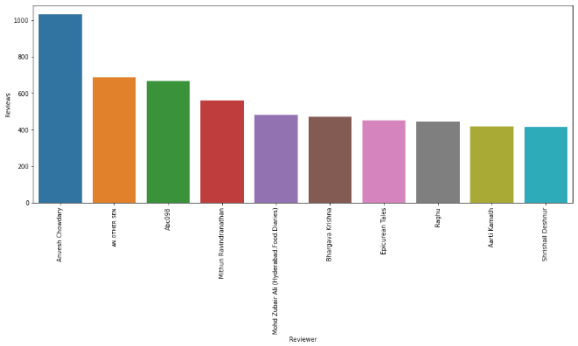


Figure 5: Most experienced reviewer

Anvesh Chowdary has written the greatest number of reviews with over 1000 reviews.

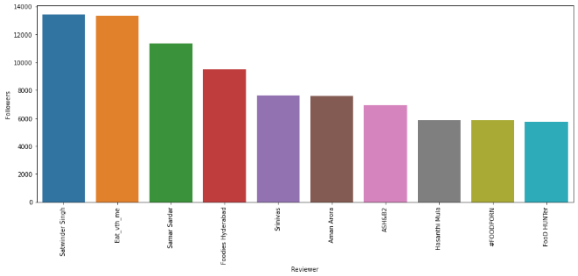


Figure 6: Most popular reviewer

Satwinder Singh is the most popular reviewer based on the number of followers.

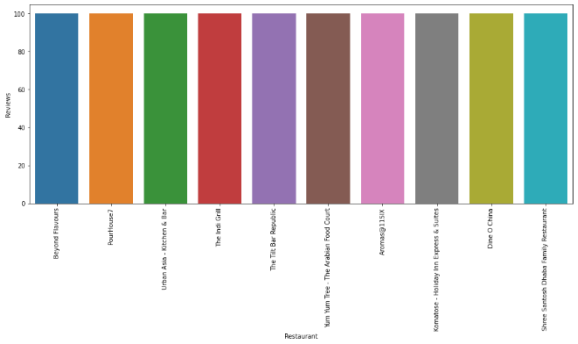


Figure 7: Most reviewed restaurant

It is observed that all the 10 most reviewed restaurants have the same number of reviews, which is 100. So, the distribution of number of reviews was check.

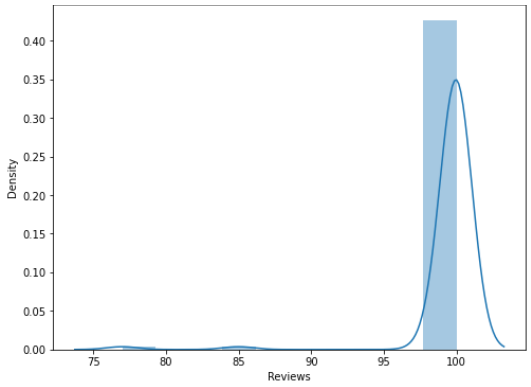


Figure 8: Distribution of the number of reviews each restaurant has

The distribution of the number of reviews each restaurant has is negatively skewed with a long tail and a narrow curve. Almost all restaurants have 100 reviews.

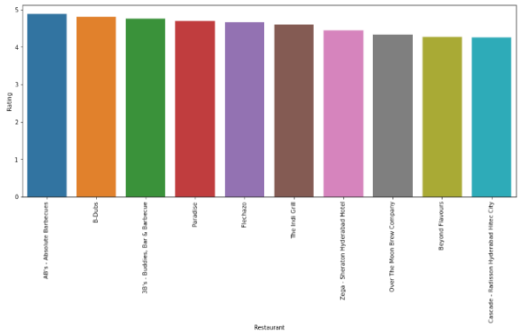


Figure 9: Highest rated restaurants

AB's - Absolute Barbecues has the highest average rating followed by B-Dubs and 3B's - Buddies, Bar & Barbecue.

Figure 10: Review wordcloud

Most frequent words used by the reviewers are place, good, food and taste.

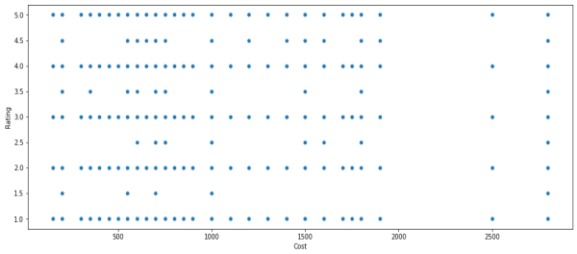


Figure 11: Rating vs cost

It is observed that there is no relationship between the rating of restaurants and the cost of food.

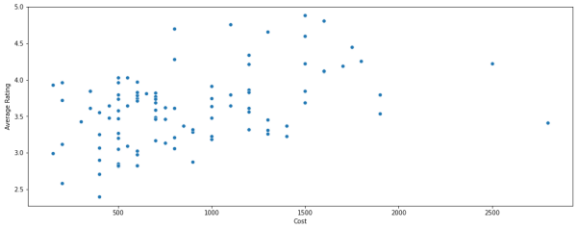


Figure 12: Average rating vs cost

It is observed that there is some linear relationship between the average rating of restaurants and the cost of food.

08. Feature Engineering

Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in machine learning.

'Links' and 'Timings' are not useful in the clustering of restaurants. So, they are removed from the restaurant metadata and a new dataframe was created to be used for clustering.

Average rating of each restaurant also needs to be considered for clustering. So, the average of 'Rating' was calculated and then merged it with the clustering dataframe.

'Cuisines' are stored in the format of a list. It is necessary to convert them to numerical

data as algorithm works only on numbers. So, it was encoded using multilabel encoder since list is iterable.

It was found that there are no misspelled cuisines in the list. Multiple cuisines were combined to reduce the number of dimensions.

Input text data and target sentiment data are the only features necessary for sentiment analysis. So, 'Tokenized_Review' and 'Rating' are retained from review dataset to create a new dataframe for sentiment analysis.

Since Rating is a continuous numerical data, it was converted to categorical data to use as the target feature for classification. The value of 1 represents positive review and 0 represents negative review.

09. Text Processing

Lemmatization is used for text normalization since meaning of words is more crucial than the getting base words to determine which class the text data belongs to.

TF-IDF is used for feature extraction from text since just the importance of words also needs to be considered.

Wordcloud images were generated for both positive and negative reviews.

Figure 13: Positive review wordcloud

Figure 14: Negative review wordcloud

10. Modelling

For restaurant clustering based on cuisines, feature scaling was performed on the data as the Euclidean distances between points are considered for K means and DBSCAN algorithm. Since the clustering dataframe has a large number of features, dimensionality reduction was also performed using principal component analysis (PCA) with 3 principal components.

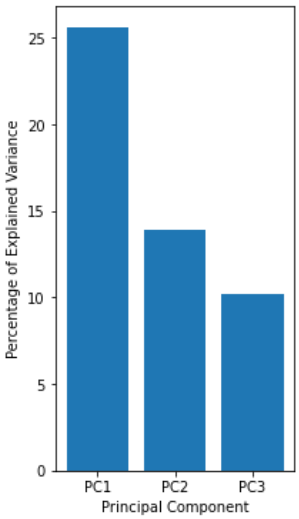


Figure 15: Scree plot

It is observed that the cumulative variance contributed by the 3 principal components is close to 50%.

Clustering was done with this data using 2 different algorithms:

1. K Means
2. DBSCAN

PCA plot for both models were generated.

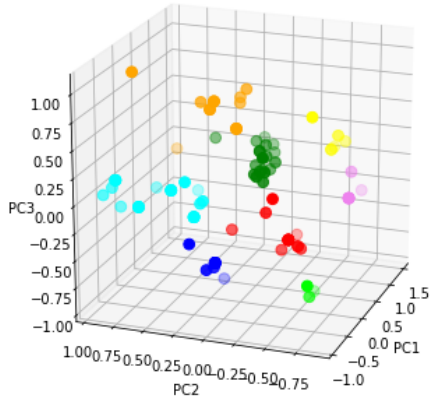


Figure 16: PCA plot of K means clustering restaurants based on cuisines

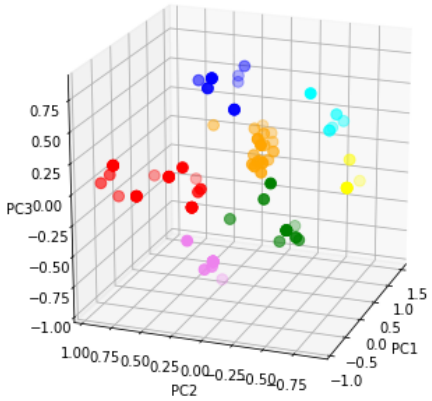
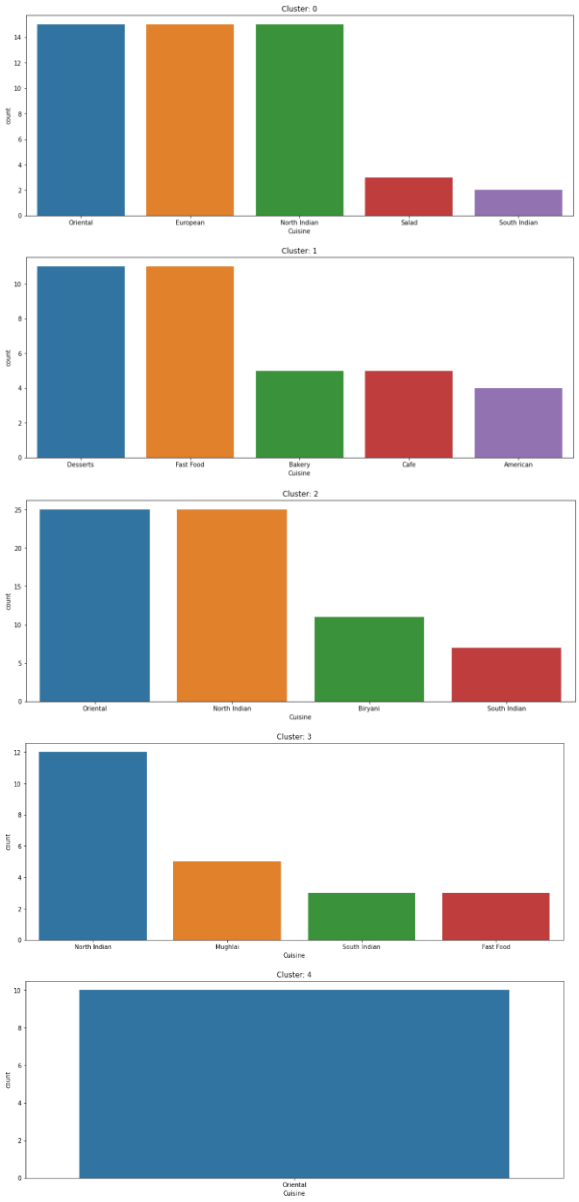


Figure 17: PCA plot of DBSCAN clustering restaurants based on cuisines



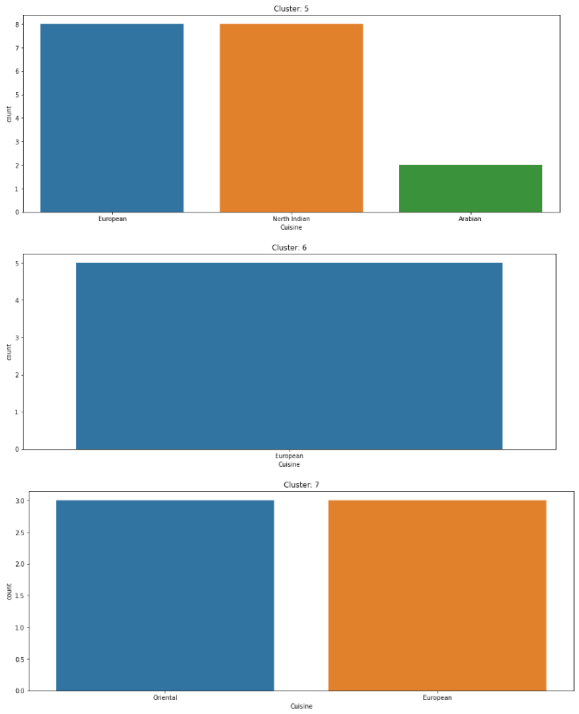
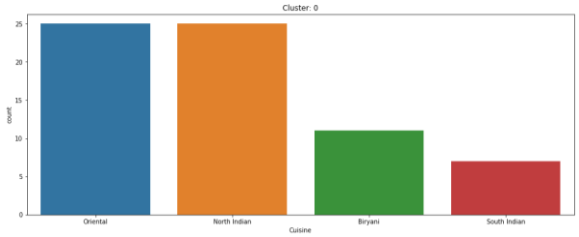


Figure 18: Top cuisines in each cluster of K means clustering restaurants based on cuisines

- Cluster 0: Seems like multicuisine restaurants.
- Cluster 1: Dominated by cuisines which are served and consumed quickly.
- Cluster 2: Dominated by cuisines which are found to be most popular during EDA.
- Cluster 3: Dominated by cuisines commonly eaten in North India.
- Cluster 4: Dominated by Oriental cuisine.
- Cluster 5: Seems like multicuisine restaurants.
- Cluster 6: Dominated by European cuisine.
- Cluster 7: Seems like multicuisine restaurants.



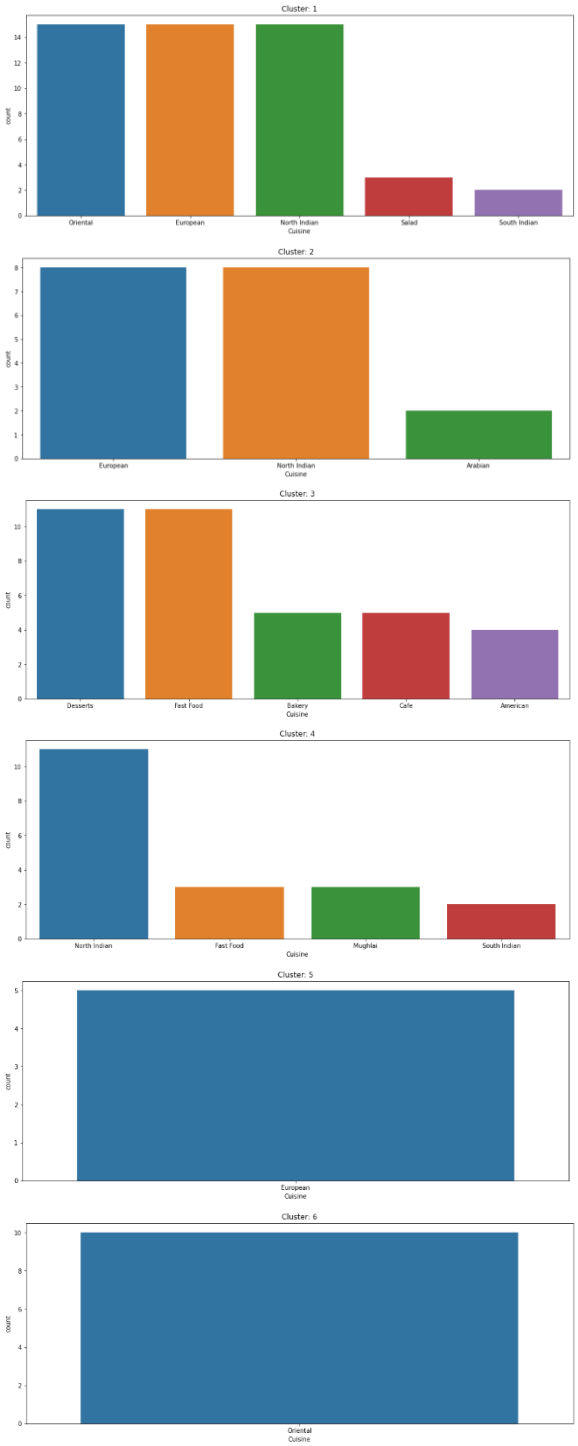


Figure 19: Top cuisines in each cluster of DBSCAN clustering restaurants based on cuisines

- Cluster 0: Dominated by cuisines which are found to be most popular during EDA.
- Cluster 1: Seems like multicuisine restaurants.

- Cluster 2: Seems like multicuisine restaurants.
- Cluster 3: Dominated by cuisines which are served and consumed quickly.
- Cluster 4: Dominated by cuisines commonly eaten in North India.
- Cluster 5: Dominated by European cuisine.
- Cluster 6: Dominated by Oriental cuisine.

Silhouette coefficient was used to compare the two clustering models find the out the best among them. Higher the silhouette score, better the model.

```
K Means: 0.676246
DBSCAN: 0.646559
```

Figure 20: Silhouette coefficients of both the models

Both the models clustered the restaurants based on cuisines similarly, except for 1 missing repeated cluster in the K means model. The model built using K means algorithm has higher silhouette coefficient than that built using DBSCAN algorithm by only 4.6%

So, either of them can be chosen for clustering the restaurant dataset based on cuisines, depending on the number of clusters preferred and whether or not outliers be included.

For restaurant clustering based on cost and ratings, feature scaling was performed on the data as the Euclidean distances between points are considered for K means and DBSCAN algorithm

Restaurant clustering was done with this data using 2 different algorithms:

1. K Means
2. DBSCAN

Scatter plots of the clusters created by both the models were generated to analyze the clusters.

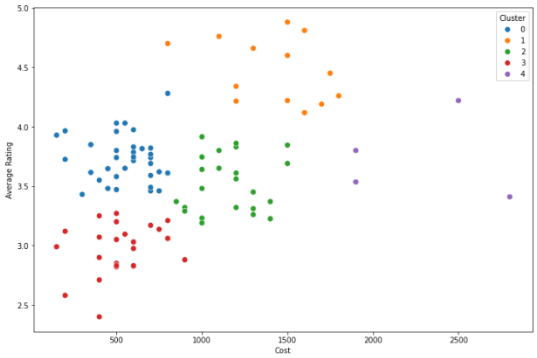


Figure 21: Scatter plot of clusters based on cost and rating created by K means model

- Cluster 0: Affordable and average-rated restaurants.
- Cluster 1: Medium-priced and high-rated restaurants.
- Cluster 2: Medium-priced and average-rated restaurants.
- Cluster 3: Affordable and low-rated restaurants.
- Cluster 4: Expensive restaurants.

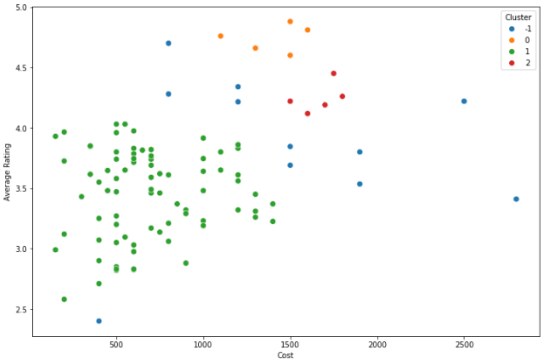


Figure 22: Scatter plot of clusters based on cost and created by DBSCAN model

- Cluster 0: Medium-priced and high-rated restaurants.
- Cluster 1: Low and medium budget restaurants.
- Cluster 2: Medium-priced and above average-rated restaurants.

There are some outliers as well, especially the expensive restaurants.

Silhouette coefficient was used to compare the two clustering models find the out the best among them.

```
K Means: 0.461043
DBSCAN: 0.392883
```

Figure 23: Silhouette coefficients of both the models

The K means algorithm was more successful than DBSCAN algorithm in creating accurate and distinguishable clusters and also the model built using K means algorithm has higher silhouette coefficient than that built using DBSCAN algorithm by 18%. So, the model built using K means algorithm should be selected for clustering based on cost and ratings.

After the model for restaurant clustering was created, the training of models for sentiment analysis was carried out.

Input and target data, which in this case were vectorized corpus and sentiment respectively, were split into training and test data with 25% test data. Training and test data of independent features were scaled using standardization.

The dataset is moderately imbalanced. If we train the models without fixing this problem, the model will be completely biased. So, an over sampling method called SMOTE was used to balance the dataset. Undersampling was not employed as the dataset is already small and undersampling will lead to loss of information.

Model training was done using 4 different algorithms:

1. Logistic regression
2. Random forest
3. Naïve Bayes
4. Support Vector Machine

Evaluation metrics like Accuracy, Precision, Recall and F1 Score were calculated for each model.

	Model	Accuracy	Precision	Recall	F1Score
0	Logistic Regression	77.059060	77.225251	77.059060	77.130006
1	Random Forest	85.777421	85.682893	85.777421	85.669634
2	Naive Bayes	77.380474	77.435115	77.380474	77.406250
3	Support Vector Machine	79.509843	79.564589	79.509843	78.808569

Figure 24: Classification model metrics

F1 Score was used to compare different models, since the dataset is imbalanced, and find out which one is better. Higher the F1 Score, better the model.

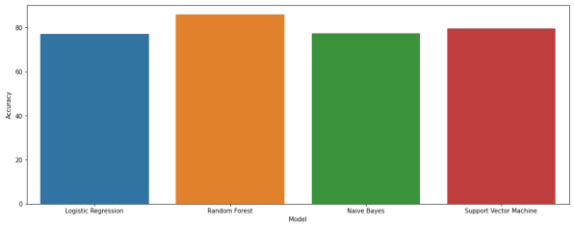


Figure 25: Accuracy comparison

The model built using the random forest algorithm has the highest F1 Score. So, it can be used for sentiment analysis after performing hyperparameter tuning.

Best parameters and accuracy were found after hyperparameter tuning.

```
The best parameters:
min_samples_split=10
min_samples_leaf=2
max_leaf_nodes=100
max_features=auto
max_depth=None

Best F1 Score: 86.4619%
```

Figure 26: Best parameter values and accuracy of the hyperparameter tuned model

Confusion matrix and ROC curve were also plotted.

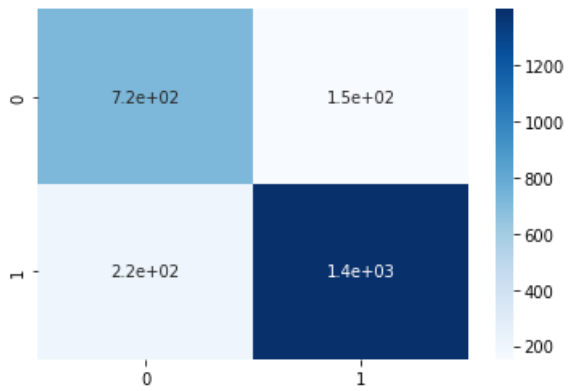


Figure 27: Confusion matrix of the hyperparameter tuned model

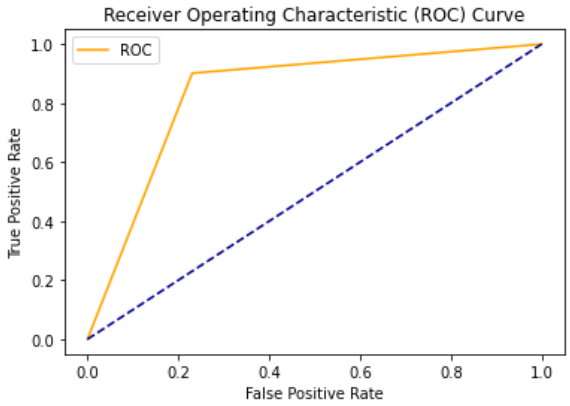
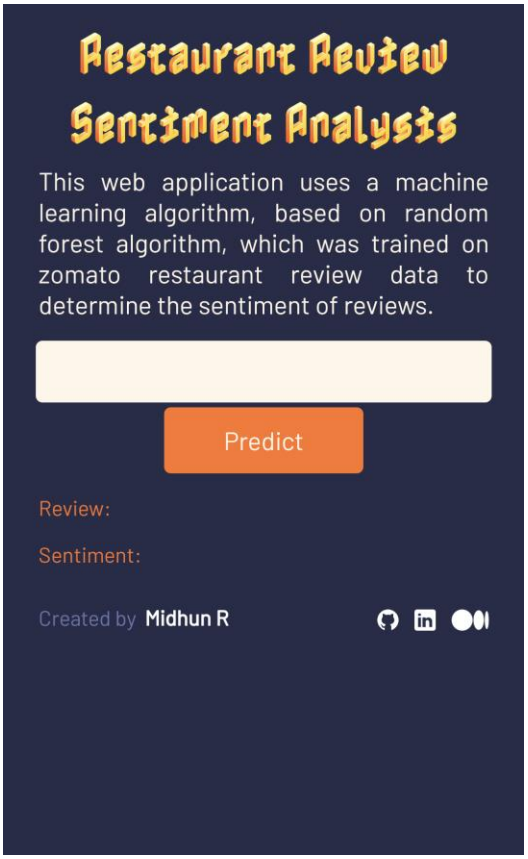


Figure 28: ROC curve of the hyperparameter tuned model



11. Deployment

- A web application is built to demonstrate the working of the trained machine learning model using a combination of HTML, CSS, and JavaScript.
- The prediction of sales using the trained ML model is carried out via a Flask API.
- This web application is deployed with AWS Elastic Beanstalk, employing CI/CD pipeline.

Link to deployed model: http://sentimenent-analysis-zomato-review.ap-south-1.elasticbeanstalk.com/

12. Challenges

Handling 2 datasets was a bit difficult as a beginner especially since one of them had 10,000 rows of text data. The inspection and cleaning of dataset was tedious and complicated since it involved conversion of objects to list, dictionaries, etc. Visualization

of data was properly carried out after providing a great amount of attention to each and every details.
Feature selection and processing was very research intensive. Training of models was a time-consuming process especially since several models had to be created to perform clustering and classification. Also in clustering, two different cases had to be carried out.

13. Conclusion

The following conclusions were drawn from EDA:

- Collage - Hyatt Hyderabad Gachibowli is the most expensive restaurant and Mohammedia Shawarma, and Amul are the most affordable ones.
- North Indian cuisine is the most popular cuisine.
- Anvesh Chowdary is the most experienced reviewer while Satwinder Singh is the most popular one.
- AB's - Absolute Barbecues is the highest rated restaurant.
- Some linear relationship exists between the average rating of restaurants and the cost of food.

The following conclusions were drawn from Modelling:

- Either of the two models, trained using K means algorithm or DBSCAN algorithm, can be chosen for clustering the restaurant dataset based on cuisines, depending on the number of clusters preferred and whether or not outliers be included.
- The model built using K means algorithm is selected for clustering based on cost and ratings.
- For sentiment analysis, the model built using random forest algorithm was chosen over others.
- If model interpretability is more important than accuracy, model built using logistic regression should be chosen. Since the difference between accuracy of these two models is less than 7%, there won't be much difference in the result.

References-

1. Scikit Learn
2. GeeksforGeeks
3. Analytics Vidhya
4. Towards Data Science