Zomato Restaurant Clustering and Sentiment Analysis

By Rahul Gupta
Data Science Trainee
AlmaBetter, Bangalore

Abstract:

India is well-known for its unique multi-food cuisine, which is offered in a huge number of restaurants and hotel resorts and symbolizes unity in variety. In India, the restaurant industry is changing rapidly. More People are appealing to the concept of eating restaurant meals, whether they dine outside or have food delivered to their homes. The increasing number of restaurants in every Indian state has encouraged an analysis of the information to gain some insights, noteworthy facts, and statistics about the Indian food sector. As a result, the purpose of this study is to analyze Zomato restaurant data in Hyderabad. Zomato is a restaurant aggregator and food delivery service based in India. With the use of unsupervised and supervised machine learning algorithms, the work here clusters restaurants into distinct segments and evaluates the sentiments in customer reviews. The analysis also resolves several business cases that can directly assist customers in locating the best restaurant in their area, as well as the company's growth and development in areas where it is currently underperforming.

Keywords: Cost-Benefit Analysis, Clustering, K Means Clustering, Sentiment Analysis

1. Problem Definition and Methods:

1.1 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.
The Project focuses on Customers and Company, you have to analyze the sentiments of the reviews given by the customer in the data and make some useful conclusions in the form of Visualizations. Also, cluster the zomato restaurants into different segments. The data is visualized as it becomes easy to analyze data at instant. The Analysis also solves some of the business cases that can directly help the customers finding the Best restaurant in their locality and for the company to grow up and work on the fields they are currently lagging in. 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.

1.2 Business Problem Analysis

Indian cuisine encompasses a wide range of regional and traditional dishes from across the

Indian subcontinent. You may discover something distinctive to love in each state. Aside from traditional North and South Indian cuisine, food culture has been greatly influenced by and evolved around various cultures. It would be an understatement to say that Indians are foodies. In India, the restaurant industry is thriving, and people enjoy celebrating tiny milestones in their lives with nice food and a pleasant ambience. Zomato is a website that connects individuals with restaurants. Zomato is an Indian restaurant aggregator that offers information about restaurants, menus, and user ratings, as well as food delivery alternatives. They essentially receive orders on behalf of the restaurant and have the meal delivered to the customer.

It is critical for Zomato to examine its datasets and make suitable strategic decisions in order to ensure its success.

The problem statement here requires us to group the restaurants in order to assist customers in finding the top restaurants in their city based on their preferences and budgetary resources. This will aid Zomato in developing a strong recommendation system and a user-friendly platform for its users.

Zomato will be able to separate out the restaurants that need to be upgraded for the business to be successful utilizing a cost-benefit analysis based on the cuisines and costs of the restaurants.

Sentiment analysis is vital for understanding fields that are underperforming and need to be improved by getting a sense of how people really feel about a particular restaurant. To find industry critics and, in particular, to work on their reviews in order to establish a laudable reputation.

1.3 Algorithms and Methods

There are two datasets to work with in this problem statement:

- Zomato Restaurant Names and Metadata
- Zomato Restaurant Reviews

The project is divided into two sections, the first one being the clustering of restaurants. Clustering is the process of separating a population or set of data points into several groups so that data points in the same group are more similar than data points in other groups. To put it another way, the goal is to separate groups with similar characteristics and assign them to clusters.

1.3.1 K Means Clustering:

K-Means Clustering is an unsupervised learning algorithm used in machine learning and data science to handle clustering problems. It's an iterative technique that splits an unlabeled dataset into k clusters, with each dataset belonging to only one group with similar qualities. It's a centroid-based approach, which means that each cluster has its own centroid. The main goal of this technique is to reduce the sum of distances between data points and the clusters that they belong to. The technique takes an unlabeled dataset as input, separates it into a k-number of clusters, and continues the procedure until no better clusters are found. In this algorithm, the value of k should be predetermined.

The k-means clustering algorithm primarily accomplishes two goals:

- Iteratively determines the optimal value for K center points or centroids.
- Each data point is assigned to the k-center that is closest to it. A cluster is formed by data points that are close to a specific k-center.

The K-means clustering algorithm's performance is dependent on the very efficient clusters it creates. However, determining the ideal number of clusters is a difficult process. There are several

methods for determining the best number of clusters, but we will focus on the most appropriate approach for determining the number of clusters or K value. The procedure is as follows:

Elbow Method

One of the most prominent methods for determining the ideal number of clusters is the Elbow approach. This approach makes use of the WCSS value notion. Within Cluster Sum of Squares (WCSS) is a term that describes the total variations within a cluster. The sharp point of bend or a point of the plot looks like an arm, then that point is considered as the best value of K.

The Curse of Dimensionality

When we have too many features, it becomes more difficult to cluster observations having too many dimensions causes every observation in the dataset to appear equidistant from every other observation. This is a serious concern since clustering requires a distance measure like Euclidean distance to estimate the similarity between observations. If all of the distances are roughly identical, all of the observations appear to be similarly similar (and equally dissimilar), and no meaningful clusters can be established.

1.3.2 Principal Component Analysis

Principal Component Analysis, or PCA, is a dimensionality-reduction approach for reducing the dimensionality of large data sets by transforming a large collection of variables into a smaller one that retains the majority of the information in the large set.

Naturally, reducing the number of variables in a data set reduces accuracy; nevertheless, the idea of dimensionality reduction is to exchange some accuracy for simplicity. Because smaller data sets are easier to study and interpret, and because machine learning techniques can analyze data more easily and quickly without having to deal with unnecessary factors.

PCA's basic concept is to reduce the number of variables in a data collection while retaining as much information as feasible.

Principal components are new variables that are created by combining or mixing the basic variables in a linear way. The new variables (i.e., principle components) are uncorrelated as a result of these combinations, and the majority of the information from the initial variables is squeezed or compressed into the first components. For instance, 10-dimensional data gives you ten principal components, but PCA seeks to place as much information as possible in the first component, then as little information as possible in the second, and so on.

Sentiment Analysis, the second half of the project, is carried out using supervised machine learning methods like Logistic Regression and Random Forest.

1.3.3 Logistic Regression

Logistic regression is a statistical analytic approach for predicting a binary outcome, such as yes or no. A logistic regression model analyses the relationship between one or more existing independent variables to predict a dependent data variable. Except for how they are employed, Logistic Regression is very similar to Linear Regression.

Instead of fitting a regression line, we fit a "S" shaped logistic function in logistic regression, which predicts two maximum values (0 or 1). Because of its capacity to generate probabilities and classify fresh data, Logistic Regression is a key machine learning technique.

The sigmoid function is a mathematical function for converting anticipated values into probabilities.
It maps any real value into another value within a range of 0 and 1.
The logistic regression's value must be between 0 and 1, and it cannot exceed this limit, resulting in a "S" curve. The Sigmoid function, often known as the logistic function, is the S-form curve.
The concept of the threshold value is used in logistic regression to describe the probability of either 0 or 1. Values over the threshold value tend to be 1, while those below the threshold value tend to be 0.

1.3.4 Random Forest

Random forest is a supervised machine learning algorithm that is commonly used to solve classification and regression problems. It creates decision trees from various samples, and uses the majority vote for classification and the average for regression.
One of the most essential characteristics of the Random Forest Algorithm is that it can handle data sets with both continuous and categorical variables, as in regression and classification. For classification challenges, it produces better results.

2. Methodology and Results:

2.1 Data Summary

2.1.1 Restaurant Names and Metadata

- Name : Name of Restaurants
- Links : URL Links of Restaurants
- Cost : Per person estimated Cost of dining
- Collection : Tagging of Restaurants w.r.t. Zomato categories
- Cuisines : Cuisines served by Restaurants
- Timings : Restaurant Timings

2.1.2 Restaurant Reviews

- Restaurant : Name of the Restaurant
- Reviewer : Name of the Reviewer
- Review : Review Text
- Rating : Rating Provided by Reviewer
- MetaData : Reviewer Metadata - No. of Reviews and followers
- Time: Date and Time of Review
- Pictures : No. of pictures posted with review

These were the two datasets that were given in order to finish the project's analysis. The first dataset contained information about Hyderabad's distinct restaurants, including costs, links, cuisines, collections, and timings. The only essential factors included in clustering from this dataset were Name, Cost, and Cuisines.The next dataset was mostly utilized for sentiment analysis of customer reviews.

2.2 Data Cleaning and Preprocessing

Both datasets required little cleaning; all that was required was to remove certain null values, convert values to acceptable data types, and select only the most significant features. Features like Link, Collections, and Timing, for example, don't help distinguish across instances.

2.3 Feature Engineering

The process of selecting, modifying, and transforming raw data into meaningful numerical features that machine learning algorithms can exploit is known as feature engineering. For example, every restaurant's multiple cuisines were represented as strings, and it was necessary to categorize and generate dummy variables for each

cuisine provided. Many of the cuisines were misspelled due to the addition of an extra space at the beginning of the string. For example, North Indian food was divided into two categories: 'North Indian' and ' North Indian'. It's also worth noting that a number of categories were created that were unneeded. For example, the dataset included both 'Chinese' and ' Momos' as different cuisines.

For the restaurant dataset, new features such as the total number of cuisines served and the average rating of the restaurant were generated by grouping in the customer ratings.

Similarly, the customer reviews dataset's reviews and followers were provided in string format, and they were separated to obtain new features such as reviews and followers.

2.4 Exploratory Data Analysis

Exploratory data analysis is a crucial part of data analysis. It is looking through and assessing a dataset to find patterns, trends, and conclusions that may be used to make better data-related decisions. The results are generally summarized using statistical graphics and other data visualization tools. To study the data, pandas is used, while matplotlib and seaborn are used to visualize it.

The following are some essential results from the analysis:

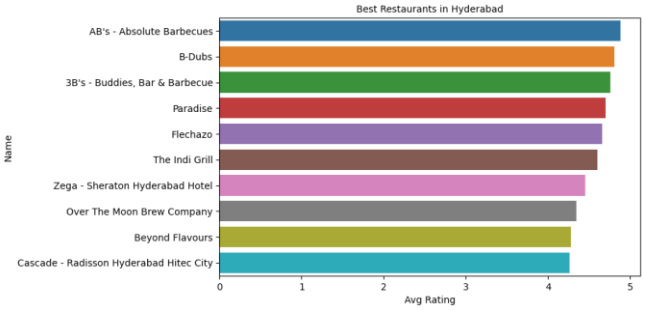
- Best restaurants in the City
- The Most Popular Cuisines in Hyderabad
- Restaurants and their Costs
- Cost-Benefit Analysis

2.4.1 Best restaurants in the City

Food, ambiance, cost, location, ratings, and other considerations all have a role in selecting a decent restaurant, but the three most significant are cuisine, cost, and reviews. When looking for a nice restaurant, the first thing that comes to mind is whether or not the cuisine you choose is accessible, and if so, whether or not the taste is satisfactory. The second consideration is value for money; it is critical that you receive exactly what you paid for. Reviews are put in place to aid in the above-mentioned judgments. They offer you a sense of what the restaurant is like based on the experiences of people who have visited it multiple times.

To aid in decision-making, the dataset includes the following features: Name, Cost, Total Cuisines, and Average Ratings. The best restaurants in the city would be those with reasonable prices, great ratings, and a large variety of cuisines.

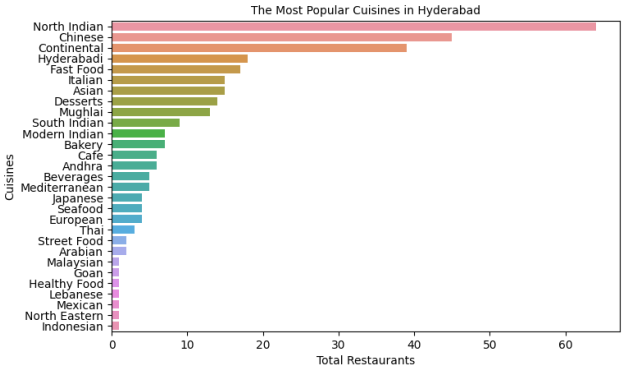
This is a plot of the sorted data, and these are the best restaurants based on the factors indicated above.



2.4.2 The Most Popular Cuisines in Hyderabad

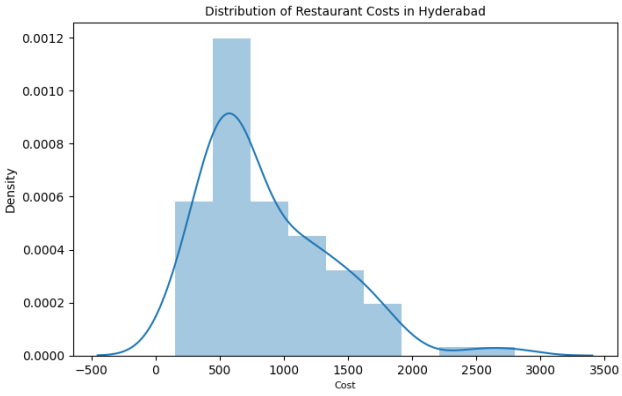
The most popular cuisines are those that are offered by the majority of restaurants in Hyderabad. Here's a plot of the various cuisines served in Hyderabad, along with the total number of restaurants that serve them. Despite its location in South India, North Indian cuisine is the most

popular in restaurants, followed by Chinese and Continental cuisines. The variety of cuisines available in Hyderabad demonstrates the city's numerous dining options.



2.4.3 Restaurants and their Costs

The cost per person in Hyderabadi restaurants ranges from 150 INR to 2800 INR. The cheapest restaurant is Mohammedia Shawarma, while the most expensive is Collage - Hyatt Hyderabad Gachibowli.



Top 5 Cheapest Restaurants

	Name	Cost
89	Mohammedia Shawarma	150.0
23	Amul	150.0
54	Asian Meal Box	200.0
101	Sweet Basket	200.0
59	KS Bakers	200.0

Top 5 Costliest Restaurants

	Name	Cost
92	Collage - Hyatt Hyderabad Gachibowli	2800.0
56	Feast - Sheraton Hyderabad Hotel	2500.0
21	Jonathan's Kitchen - Holiday Inn Express & Suites	1900.0
18	10 Downing Street	1900.0
91	Cascade - Radisson Hyderabad Hitec City	1800.0

2.4.4 Cost-Benefit Analysis

Every time you engage in a company endeavor or make a business choice, you must consider whether the option is worthwhile. A Cost-Benefit Analysis is a method of evaluating the value of a choice by estimating the costs of implementing it and comparing them to the benefits of doing so. If the expected benefits outweigh the costs, you'll profit from the decision; if not, it's time to devise a better strategy.

Zomato is an online food delivery service and a search engine for Indian restaurants. Zomato is a food delivery service that focuses on internet ordering, restaurant reservations, and reward programmes. Restaurant chains who want to reach a wider audience, as well as app users who simply want to try out local eateries and cuisines, are the company's target clients. Here is a simple cost-benefit analysis that can be performed based on the limited information available

Costs

When calculating costs, start with direct costs, which are expenses directly tied to the production or development of a product or service (or the implementation of a project or business decision), which in Zomato's case is essentially the mobile app. Maintaining the application, conceptualizing strategies, including restaurants, marketing, food delivery partners, and customer service, necessitates the participation of a large staff. The employees' pay would be a direct cost.

Utilities, rent, partners, advertisements, and other indirect costs are examples.
Other expenses are difficult to quantify, such as negative platform reviews that cause customers to avoid using the app altogether, a poor social network presence, and so on.

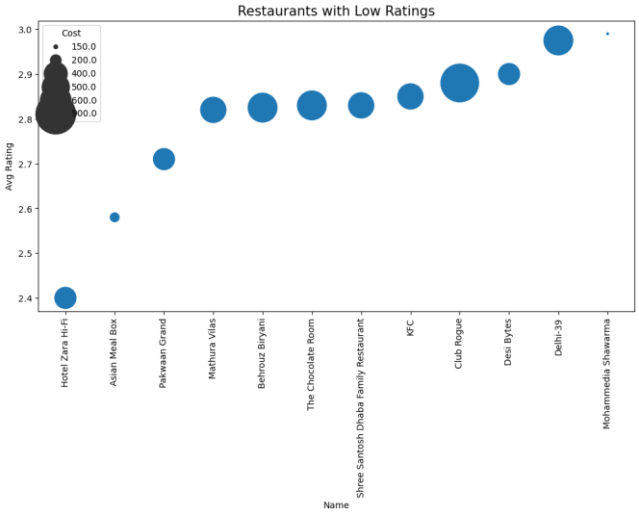
Benefits

Advertising is the primary source of revenue. More restaurants are promoting themselves on Zomato's feed in order to acquire exposure and attention from a huge number of Zomato subscribers and customers.
Zomato charges restaurants a commission based on the number of orders placed through its food delivery service. The company makes money by charging restaurants a commission for each delivery, which is split between the delivery partners and the company. Due to the high level of competition and the need to offer large discounts, internet meal delivery represents a small fraction of total revenue compared to other revenue streams.

Comparison

The information we have includes the pricing per person, the cuisines available at the restaurant, and the restaurant's average rating. Zomato will have an issue if a restaurant has a poor rating, a high per-person cost, and a limited selection of popular cuisines. Negative reviews are an intangible cost to the business, and as a result, the business will begin to lose everyday application users. The app's users are a valuable asset to the company; because of their enormous viewership, Zomato receives advertising from various restaurants.
Overall, it's critical to identify which restaurants Zomato has to improve in order to improve its overall customer experience, and if improvement tactics fail, they must delist those restaurants.

Here's a scatter plot of the restaurants having the lowest Average Rating according to their per-person Cost.



Mohammedia Shawarma has the highest rating among these restaurants and the lowest price, hence it seems profitable enough but some restaurants like Club Rouge have low rating yet high per-person dining cost, this will not generate significant revenue and needs improvement.

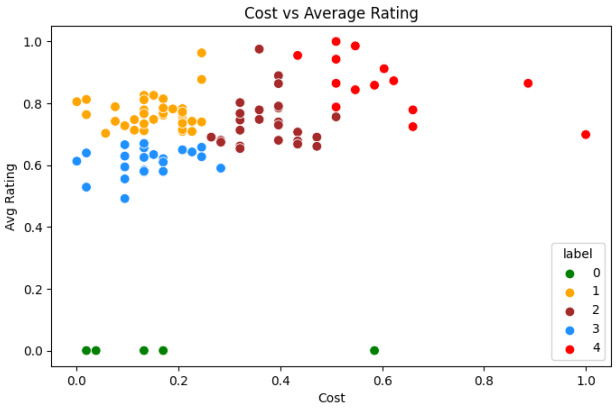
2.5 Restaurant Clustering

Approach 1 Here's a scatter plot of the restaurant clusters formed by K Means Clustering on the basis of just two input variables Cost and Average Rating.
The clusters are fairly distinct from one another. Because there were just two input variables, they were easy to separate and interpret.

- Restaurants with the label 0 were in the names dataset but were not reviewed.
- Restaurants with favorable reviews and inexpensive prices are labeled as label 1.
- Label 2 restaurants are fine dining establishments with good reviews and reasonable prices.
- Restaurants in the Label 3 category are modest eateries with low prices and

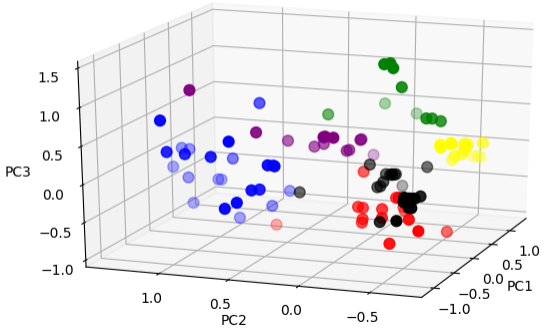
average reviews.

- Label 4 restaurants are those that are both pricey and have above-average reviews.



Approach 2 Here's a 3D scatter plot of restaurants clustered on the basis of three principal components.
The data points inside the clusters shared a lot of commonalities.

- Cluster 0 - The eateries in this cluster primarily serve continental and fast meals. The average rating is 3.42, and the average cost is 942 INR, including a 2500 INR outlier and a 600 INR median cost. This means that the eateries in this cluster, with the exception of one, are all rather inexpensive.
- Cluster 1 - The restaurants in Cluster 1 specialize in North Indian cuisine, as well as other complementing cuisines. The average cost is 823 INR and the average rating is 3.63. The prices of these restaurants are slightly higher than those in cluster 0.
- Cluster 2 - Restaurants in Cluster 2 serve a variety of popular cuisines, including North Indian, Chinese, and complimentary. The average rating is 3.77, which is higher than the other two clusters, and the average price is 1331 Indian rupees. These establishments are fine dining restaurants.
- Cluster 3 - The restaurants in this cluster serve a variety of foreign cuisines, including Chinese, Thai, Asian, and seafood, among others. The average rating is 3.18, owing to the fact that these cuisines aren't particularly popular in Hyderabad, and the average cost is 890 INR.
- Cluster 4 - Cluster 4 consists primarily of small eateries, bakeries, and cafes. The average cost is 406 INR and the average rating is 3.14.
- Cluster 5 - Popular cuisines such as North Indian, Chinese, and notably Hyderabadi are accessible at restaurants in cluster 5. The average cost is 674 INR, and the average rating is 3.24. These are casual dining establishments with lower prices and ratings per person than cluster 2.



2.6 Sentiment Analysis

Sentiment analysis is a machine learning technology that looks for polarity in texts, ranging from positive to negative. Machine learning tools learn how to detect sentiment without human input by training them with samples of emotions in text. Sentiment analysis models can be trained

to understand things like context, sarcasm, and misapplied words in addition to simple meanings. To command and train machines to perform sentiment analysis, a variety of techniques and complicated algorithms are used.

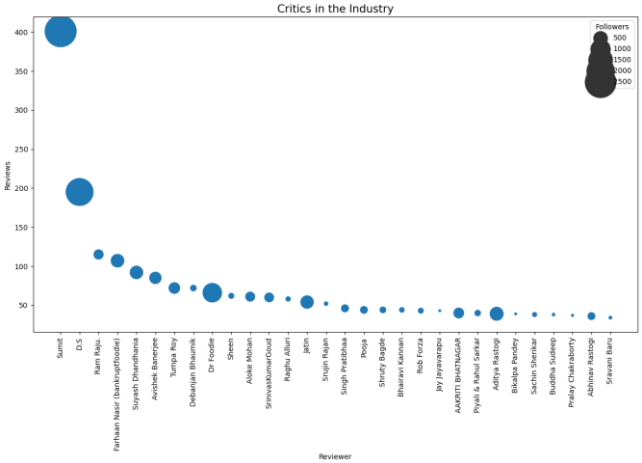
Positive Word Cloud

Negative Word Cloud

2.6.1 Critics in the Industry

Customers have all the power they need to build or break a firm in today's Internet-driven, social-media world. If clients have a good experience, they tell their friends, family, and acquaintances about it, which leads to new business. All of this word-of-mouth marketing is free, and when it's posted on a public platform, it's shared with anybody who uses that platform. Customers will complain if you are unable to provide a pleasant client experience for any reason. To decrease the bad marketing impact, Zomato, like any other business, needs to focus on the criticism, particularly with those reviewers who have been following the most. In order to grow a loyal consumer base, every industry must fight the complaints and critiques it receives.

Here, an attempt has been made to group consumers with a large number of followers who have left more reviews with consistently low ratings in order to identify the top critics and the areas that need to be improved.



2.6.2 TF-IDF

Term Frequency Inverse Document Frequency of records is abbreviated as TF-IDF. It's the process of determining how relevant a word in a series or corpus is to a text. The meaning of a word grows in proportion to how many times it appears in the text, but this is offset by the corpus's word frequency (data-set). It's frequently used in information retrieval, text mining, and user modeling searches as a weighting factor.

2.6.3 Model Evaluation

Here are the results for the two models trained by inputting the review text, Logistic Regression and Random Forest.

In the business challenge, correctly anticipating negative sentiments is critical, but it is even more critical for the models to limit the amount of false positives. False positives suggest that the reviews

were genuinely unfavorable but were classified as positive, resulting in the loss of a complaint to address.

Random Forest performs better in terms of decreasing False negatives than Logistic Regression, but having a higher number of false positives. This suggests that Logistic Regression is penalizing False Positives more aggressively, which is exactly what we want.

```
Results for Logistic Regression
0.8674033149171271
[[ 628  100]
 [ 164 1099]]
              precision    recall  f1-score   support

           0       0.79      0.86      0.83       728
           1       0.92      0.87      0.89      1263

    accuracy                           0.87      1991
   macro avg       0.85      0.87      0.86      1991
weighted avg       0.87      0.87      0.87      1991
```

```
Results for Random Forest
0.8779507785032646
[[ 542  186]
 [  57 1206]]
              precision    recall  f1-score   support

           0       0.90      0.74      0.82       728
           1       0.87      0.95      0.91      1263

    accuracy                           0.88      1991
   macro avg       0.89      0.85      0.86      1991
weighted avg       0.88      0.88      0.87      1991
```

3. Conclusion and Recommendations:

3.1 Conclusion:

Clustering is the process of identifying unique groupings or "clusters" within a data set. The programme constructs groups using a machine language algorithm, and items in a comparable group will have similar features in general.
One of the challenges that organizations have is figuring out how to arrange the massive volumes of data accessible into usable structures. Alternatively, divide a large heterogeneous group into smaller homogenous groupings. Cluster analysis is an exploratory data analysis tool that seeks to group things so that the degree of relationship between two objects is greatest if they belong to the same group and minimal if they don't.
This enables businesses to assist their clients in quickly locating the information they require. This analysis included all of the essential subjects in both the business and technological domains.

Some important insights to draw from the analysis includes:

- The best restaurants in Hyderabad are AB's - Absolute Barbecues, B-Dubs, and 3B's - Buddies, Bar & Barbecue.
- The most popular cuisines are the cuisines which most of the restaurants are willing to provide. The most popular cuisines in Hyderabad are North Indian, Chinese, Continental, and Hyderabadi.
- The restaurants in Hyderabadi have a flexible per person cost of 150 INR to 2800 INR. The cheapest is the food joint called Mohammedia Shawarma and the costliest restaurant is Collage - Hyatt Hyderabad Gachibowli.
- Upon conducting basic cost-benefit analysis on Zomato with a few assumptions one basis of the little business understanding that could be gathered, it can be concluded that it is important to separate out the restaurants with the lowest rating in order to improve its overall customer experience. These restaurants were small food joints or restaurants with high prices according to the food they were serving. Efforts should be made to advertise more and analyze the reviews, especially for these restaurants,

and work on them. Mohammedia Shawarma seems to be profitable.

- Restaurant Clustering was done in two approaches. First with just two features and then with all of them. K means Clustering worked well in the first approach but as we increase the dimensions, it isn't able to distinguish the clusters hence principal component analysis was done and then clustered into 6 clusters. The similarities in the data points within the clusters were pretty great.
- Critics in the Industry were identified by grouping the customers with a good number of followers who have given more reviews with constantly low ratings. Sumit, D.S, and Ram Raju are the top three critics.
- Sentiment Analysis was done on the reviews and a model was trained in order to identify negative and positive sentiments. Even though the number of false negatives is higher in the case of Logistic Regression than Random Forest, it is performing better in terms of reducing False positives. This indicates that Logistic Regression is penalizing False positives more just as we want.

3.2 Challenges:

- Because the data was provided in a raw format in string format, the project's main problem was extracting key information from the dataset in numerical form.

3.3 Recommendations:

- Negative reviews should be approached to reach a win-win solution.
- Ratings should be gathered according to category, such as packing, delivery, taste, quality, amount, and service. This would aid in identifying and addressing lagging fields.

4. References:

- Machine Learning Mastery
- GeeksforGeeks, geeksforgeeks+python&oq=●%09GeeksforGeeks&aqs=chrome.2.69i57j0i22i30l9.2572j0j7&sourceid=chrome&ie=UTF-8
- https://www.analyticsvidhya.com/blog/
- https://towardsdatascience.com/tagged/blog
- https://builtin.com/data-science/best-data-science-company-blogs-machine-learning
- scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html