**Zomato Restaurant Clustering and Sentiment Analysis**

**Souvik Bhattacharyya**

**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 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: Clustering, K Means Clustering, Sentiment Analysis***

**Problem Definition and Methods:**

**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.

**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 ambiance.

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.

**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.

**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.

**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.

**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.

**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. Classification challenges, produces better results.

**Methodology and Results:**

**Data Summary**

**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

**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 Cuisine. The next dataset was mostly utilized for sentiment analysis of customer reviews.

**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.

**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 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 was 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.

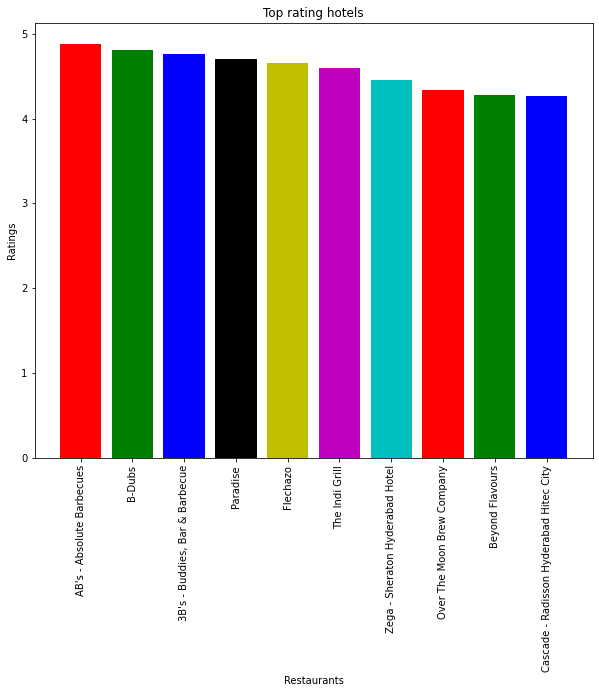
**Exploratory Data Analysis**

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**

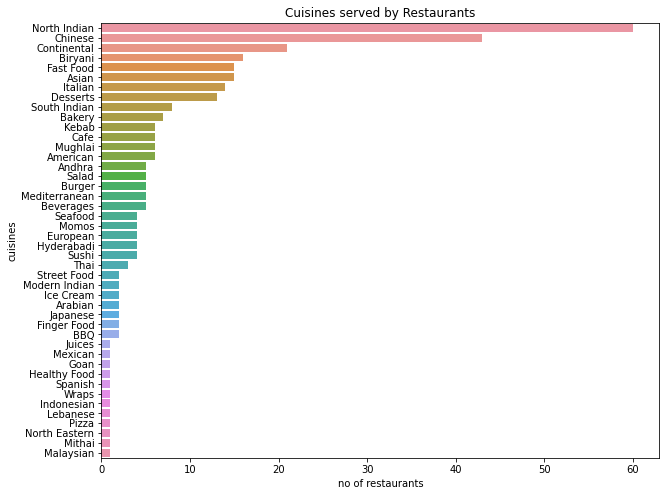
The Most Popular Cuisines in Hyderabad Restaurants and their Cost-Benefit Analysis

**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.

**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.

**Restaurants and their Costs**

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

**Restaurant Clustering**

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.

**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.

**Model Evaluation**

Here are the results for the two models trained by 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 number 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 has a higher number of false positives. This suggests that Logistic Regression is penalizing False Positives more aggressively, which is exactly what we want.

**Conclusion**

The best restaurants in Hyderabad are AB's - Absolute Barbecues, B-Dubs, and 3B's - Buddies, Bar & Barbecue include 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 food joint is Mohammedia Shawarma and the costliest restaurant is Collage – Hyatt Hyderabad Gachibowli.

Upon conducting a basic cost-benefit analysis on Zomato with a few assumptions and 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.