**Zomato Restaurant Clustering and Sentiment Analysis**

**Ruchika Nayak and Om Prakash Pradhan**

**Data science trainees,**

**AlmaBetter, Bangalore**

**Abstract:**

This project focuses on Customers and Company. Our goal is to cluster the restaurants based on various features and analyze the sentiment of people based on the reviews they have written on Zomato app and also to make some meaningful conclusion from the data in the form of visualization. The aim of clustering is to segregate groups with similar traits and assign them into various clusters.

To achieve our first goal, we have used Silhouette Score, Elbow method and Dendrogram to find out the optimal number of clusters followed by K-means clustering and Hierarchical clustering algorithms to cluster the restaurants. For sentiment analysis we have used various NLP techniques to clean our data and to classify the sentiments into positive and negative we have used various binary classification algorithms.

***Keywords:* *Clustering, sentiment analysis Silhouette score, Elbow method, Dendrogram, K-means clustering, Hierarchical clustering, NLP***

**1. Problem Statement**

Zomato is an Indian restaurant aggregator and food delivery start-up which provides information, menus and user-reviews of restaurants, and also has food delivery options from partner restaurants in select cities. 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. It also focuses on customers and company, and we have to analyze the sentiments of the reviews given by the customer in the data and made some useful conclusions. Also, to cluster the Zomato restaurants into different segments.

**2.Introduction**

### Zomato has either invested or acquired in quite a few startups internationally to help grow its, the recent one Grofers. This helps Zomato focus mainly and food delivery and lets Grofers handle daily usage products.

### In this project we have performed EDA, clustering and sentiment analysis on the given dataset to analyze the popularity and feelings of the people who visits and taste the food of Zomato.

## **3.Zomato Restaurants**

Zomato is an Indian multinational restaurant aggregator and food delivery company founded by Deepinder Goyal and Pankaj Chaddah in 2008. Their technology platform connects customers, restaurant patterns and delivery partners, serving their multiple needs. Customers use their platform to search generated reviews and view and upload photos, order food delivery, book a table and make payments while dining-out at restaurants. On the other hand, they provide restaurant partners with industry-specific marketing tools which enable them to engage and acquire customers to grow their business while also providing a reliable and efficient last mile delivery service.

**4.Steps involved:**

* **Reading and understanding data**

We have two dataset for our project. One dataset is **Zomato Restaurant names and Metadata** and the other one is **Zomato Restaurant Reviews**. After reading the datasets which is in csv format we inspected the rows and data type of each columns.

* **Null values and Duplicates**

Our dataset have some null values and duplicate rows. So we drop those rows and columns from the datasets.

* **Exploratory Data Analysis (EDA)**

After loading the dataset,

* In EDA of Names and Metadata set we made some changes in our columns like in Cost column we convert it to integer type. Then we used bar plots to visualize the Top 10 popular cuisines and Top 10 popular collections.
* In EDA of Restaurant reviews dataset we also made some changes like we separate the metadata column into two parts i.e. Reviews and Followers. Then we visualize some features using barplots, like Top 10 costliest and cheapest restaurants, Top rated and worst rated restaurants, most reviewed and most rated restaurants, restaurants with most number of positive and negative reviews.
* **Feature engineering**

Both the dataset is merged for clustering and sentiment analysis. Then some of the columns are removed which are not so important for our project.

**5**. **Clustering:**

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups(clusters). It is a common technique for statistical data analysis.

* **Data preparation for clustering**

From the merged dataset the metadata column has been separated the reviews and followers into two separate columns. Then we have taken some selected columns for the clustering like ‘Name’, ‘Cost’, ‘Mean Rating’, ‘Mean Followers’. We have aggregated the ‘Rating’ and ‘Number of followers’ column to make it a single value for each restaurant.

* **Fitting different algorithms:**

For clustering we have used various algorithms like:

1. K-means clustering
   1. Silhouette Score
   2. Elbow method
2. Hierarchical clustering
   1. Dendogram
   2. **Algorithms:**

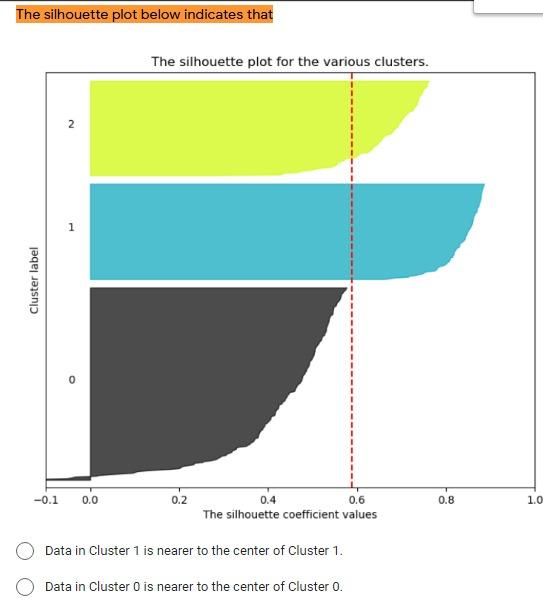
Following algorithms are used to visualize and find the optimal number of clusters:

* **K-means clustering**:

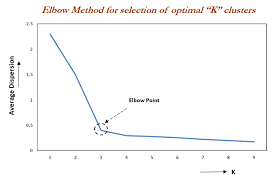
The k-means clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. k-means clustering groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process.

Techniques used to find the number of clusters are:

1. **Silhouette Score:** It determines whether there are large gaps between each sample and all other samples within the same cluster or across different clusters.

****

1. **Elbow method:** It is used to find the elbow in the elbow plot. The elbow is found when the dataset becomes flat or linear after applying the cluster analysis algorithm. The elbow plot shows the elbow at the point where the number of clusters starts increasing.

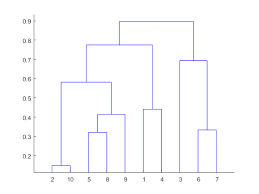
****

* **Hierarchical clustering:**

Hierarchical clustering also known as hierarchical cluster analysis, is an algorithm that groups similar object into groups called clusters It’s a method of cluster which seeks to build a hierarchy of clustering.

Technique used to visualize hierarchical clustering is:

1. **Dendrogram:** It is a visual representation of the compound correlation data. The individual compounds are arranged along the bottom referred to as leaf nodes. Compound clusters are formed by joining individual compounds or existing compound clusters with the join point referred to as a node.

****

**6. Sentiment Analysis:**

Sentiment analysis is an approach to natural language processing(NLP) that identifies the emotional tone behind a body of text. This is a popular way for organizations to determine and categorize about a product, service or idea.

* **Data Preparation for Sentiment Analysis**

For sentiment analysis we have taken some selected columns from the merged dataset, like ‘Name’, ‘Review’, and ‘Rating’ columns. Then we grouped the ‘Rating’ column into two parts, one with rating 0 to 3 as negative (0) and rating 4 to 5 as positive (1) class. Then we have removed the stop words and punctuations from the ‘Review’ column followed by stemming and vectorization.

* **Fitting different models:**

Various models used for Sentiment Analysis are:

1. Naïve- Bayes
2. Logistic Regression
3. XGBoost
4. SVM

**6.1. Models**

After dataset preparation following models are used to find out the accuracy and most appropriate model among them.

* **Naïve-Bayes:**

It is a simple technique for collection of classification algorithms based on Bayes Theorem. It’s a probabilistic machine learning model that’s used for classification task. This algorithm is mostly used in text classification and with problems having multiple classes.

* **Logistics Regression:**

Logistics regression estimates the probability of an event occurring, such as voted or didn’t vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1.

* **XGBoost:**

It is a powerful approach for building supervised regression models. is an open-source software library which provides a regularizing gradient boosting framework.  it aims to provide a "Scalable, Portable and Distributed Gradient Boosting (GBM, GBRT, GBDT) Library".

* **SVM:**

SVM or support vector machines are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. It is a classification technique used for the classification of linear as well as non-linear data. It’s the margin based classifier.

**6.Conclusion:**

So, that’s it. We reached the end of our exercise. Starting with data loading and reading the data so far we have done EDA, two dataset merging followed by clustering and then different model building.

From all these process we came to know that:

* In the EDA we have explored popular cuisines and collections, costliest and cheapest restaurants, top rated and worst rated restaurants, most reviewed and most followed restaurants and the restaurants with most positive and most negative reviews.
* The optimal number of clusters by taking 2 variables at a time are either 3 or 4. And optimal number of clusters by taking all variables at a time we get 4.
* And for sentiment analysis logistics regression and SVM are the two most appropriate models with accuracy of nearly 87%.

**7. Challenges Faced:**

* Selecting the appropriate evaluation methods for clustering.
* Selecting the optimal number of clusters.
* Selecting the optimal number of features for sentiment analysis.

**References-**

1. GeeksforGeeks
2. Wikipedia
3. Kaggle
4. Javatpoints