Zomato Restaurant Clustering and Sentiment

Vishal Raul
Data science trainees,
AlmaBetter, Bangalore

Abstract:

In today's digital era where everybody has their own opinion and so can post various feedbacks regarding different products. Reviews are critical in nature and can affect the market value of a product. Amazon, Zomato, Quora, Twitter, etc. are there few such platforms that provide space for product reviews? Since reviews are textual and diversified; a collective representation will help users to summarize their opinions. Honest food reviews and its analysis is still a challenge. 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. Introduction

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 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.To cluster data points we have used k-means clustering. K-mean clustering uses "centroids", k different randomly – initiated points in the data, and after every point has been assigned, the centroid is moved to the average of all of the points assigned to it. After clustering it's important to finalize the number of clusters. Sentiment analysis is one part of Natural Language Processing that is often used to analyze words based on the patterns of people in writing to find positive, negative, or neutral sentiments. Sentiment analysis is useful for knowing how users like something or not.

2. Problem Statement

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.

3. Data

Data Description:

The collected data has been stored in the Comma Separated Value file Zomato Restaurant names and Metadata.csv. Each restaurant in the dataset is uniquely identified by its Restaurant Id.

Data fields:

Use this dataset for clustering part

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

Merge this dataset with Names and Matadata and then use for sentiment analysis part

Zomato Restaurant reviews.csv

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

4. Steps involved

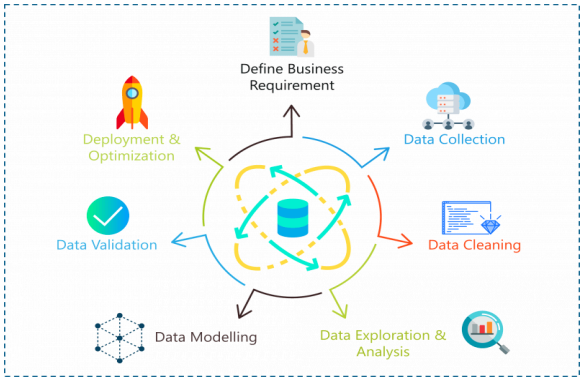


Image credit: edureka.com

- **Define Problem Statement**

  Before you even begin a Data Science project, you must define the problem you're trying to solve. At this stage, you should be clear with the objectives of your project.

- **Data Collection**

  Our dataset contains a large number of null values which might tend to disturb our accuracy hence we dropped them at the beginning of our project in order to get a better result.

- **Data Cleaning**

  Data cleaning is the process of removing redundant, missing, duplicate and unnecessary data. This stage is considered to be one of the most time-consuming stages in Data Science. However, in order to prevent wrongful predictions, it is

important to get rid of any inconsistencies in the data.

- **Data Analysis and Exploration**

  Once you're done cleaning the data, it is time to get the inner Sherlock Holmes out. At this stage in a Data Science life-cycle, you must detect patterns and trends in the data. This is where you retrieve useful insights and study the behavior of the data. At the end of this stage, you must start to form hypotheses about your data and the problem you are tackling.

- **Data Modelling**

  This stage is all about building a model that best solves your problem. A model can be a Machine Learning Algorithm that is trained and tested using the data. This stage always begins with a process called Data Splicing, where you split your entire data set into two proportions. One for training the model (training data set) and the other for testing the efficiency of the model (testing data set).

- **Optimization and Deployment**

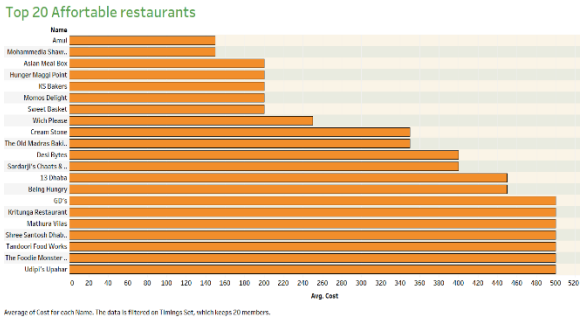
  This is the last stage of the Data Science life-cycle. At this stage, you must try to improve the efficiency of the data model, so that it can make more accurate predictions. The end goal is to deploy the model into production or production-like environments for final user acceptance. The users must validate the performance of the models and if there are any issues with the model then they must be fixed in this stage.

5. Exploratory Data Analysis:

5.1. Restaurants and their Costs

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

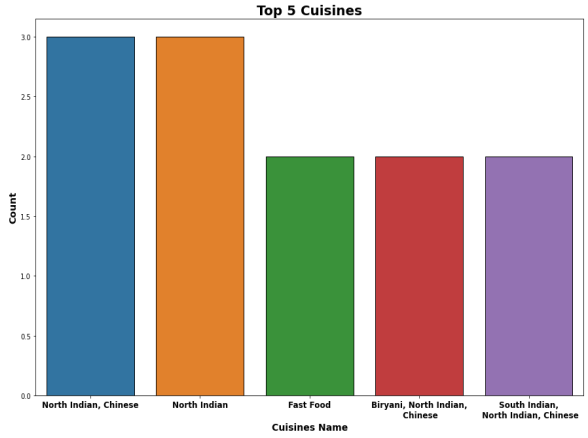
Graph 1. Restaurants and their Costs



5.2. Cuisines

Indian cuisine consists of a variety of regional and traditional cuisines native to the Indian subcontinent. North Indian cuisine is the most popular in restaurants, followed by fast food and Biryani. The variety of cuisines available in the restaurant has numerous dining options.

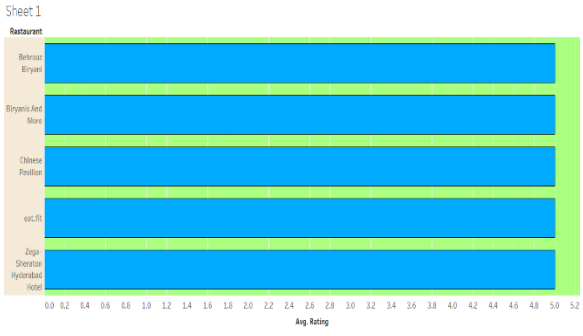
Graph 2. The Most Popular Cuisines



5.3. Best Restaurants

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

Graph 3. Best Restaurants



6. Data Preprocessing

6.1. Stemming and Lemmatization

Stemming just removes or stems the last few characters of a word, often leading to incorrect meanings and spelling. **Lemmatization** considers the context and converts the word to its meaningful base form, which is called Lemma. Sometimes, the same word can have multiple different Lemmas. We should identify the Part of Speech (POS) tag for the word in that specific context. Here are the examples to illustrate all the differences and use cases:

1. If you lemmatize the word '**Caring**', it would return '**Care**'. If you stem, it would return '**Car**' and this is erroneous.
2. If you lemmatize the word '**Stripes**' in **verb** context, it would return '**Strip**'. If you lemmatize it in **noun** context, it would return '**Stripe**'. If you just stem it, it would just return '**Strip**'.
3. You would get same results whether you lemmatize or stem words such as **walking, running, swimming**... to **walk, run, swim** etc.
4. Lemmatization is computationally expensive since it involves look-up tables and what not. If you have a large dataset and performance is an issue, go with Stemming. Remember you can also add your own rules to Stemming. If accuracy is paramount and the dataset isn't humongous, go with Lemmatization.

6.2. Top Vocab in collections and cuisines

Next, we are trying to retrieve words which are used in our data set most of the time. We can see in the left image that words like venu, veggie, unknown, shawarma etc. are

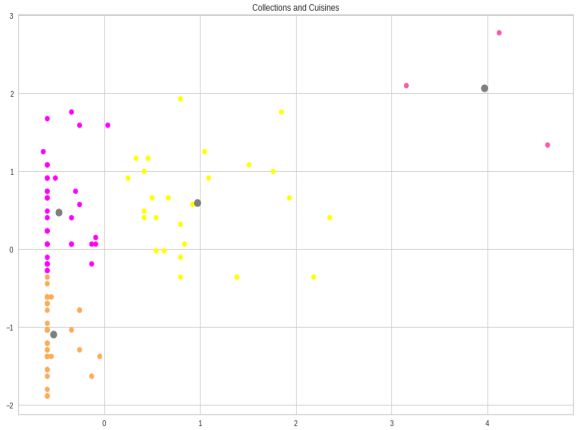
used in Collection. On the right side of the image are sushi, Spanish, south, seafood, salad, pizza, north, Momo, Mexican, kebab etc. words used in cuisines.

Graph 5. Vocab

7. Restaurant Clustering

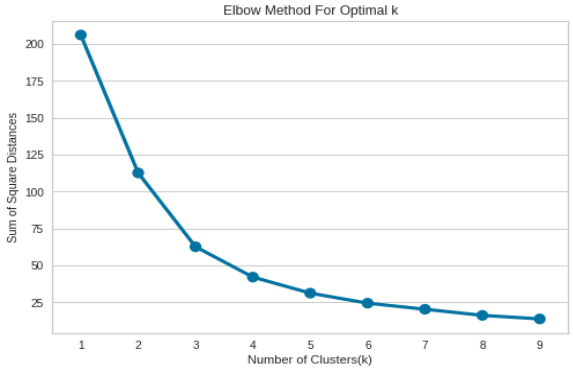
7.1. KMeans Clustering

KMeans clustering is a method of vector quantization, originally from signal processing that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster. K = 4, we draw k means clustering. Here we can see 4 black dots as the centroid of clusters.



7.2. Elbow Method (Clustering)

In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use. The Elbow method gets the perfect number of clusters. We got 4 numbers of cluster best suites to our data frame.



7.3. Silhouette score method

Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. The Silhouette score is calculated for each sample of different clusters Computing the mean Silhouette Coefficient of all samples.

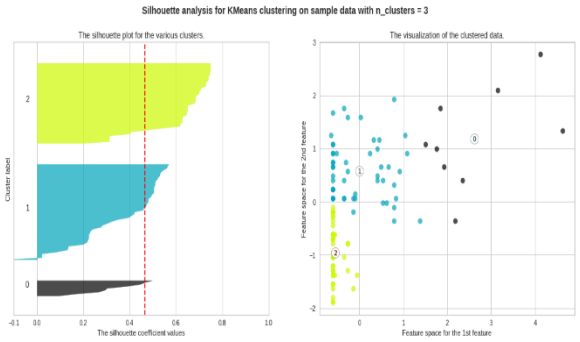
The Silhouette Coefficient is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. The Silhouette Coefficient for a sample is (b - a) / max (a, b). To clarify, b is the distance between a sample and the nearest cluster that the sample is not a part of. Note that the Silhouette Coefficient is only defined if the number of labels is 2 <= n_labels <= n_samples - 1.

This function returns the mean Silhouette Coefficient over all samples. To obtain the values for each sample, use silhouette_samples.

The best value is 1 and the worst value is -1. Values near 0 indicate overlapping clusters. Negative values generally indicate that a sample has been assigned to the wrong cluster, as a different cluster is more similar.

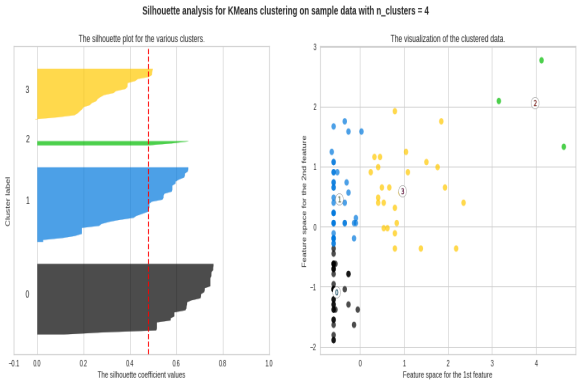
For **n_clusters = 3**

The average silhouette score is: 0.4655



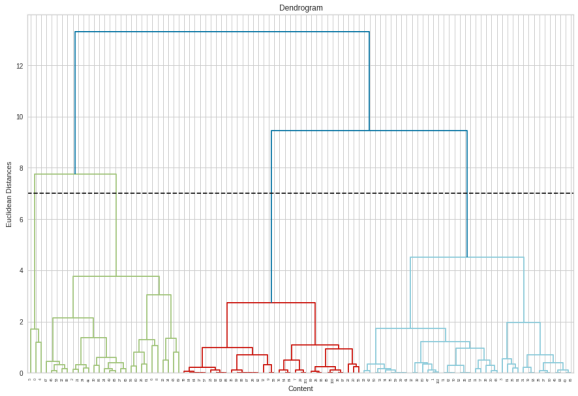
For n_clusters = 4

The average silhouette score is: 0.4799



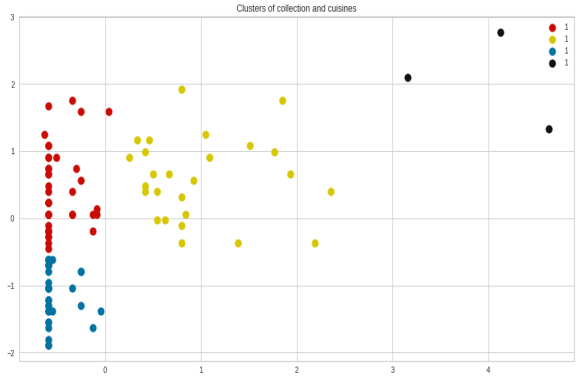
7.4. Dendrogram clustering

A dendrogram is a tree-structured graph used in heat maps to visualize the result of a hierarchical clustering calculation. The result of a clustering is presented either as the distance or the similarity between the clustered rows or columns depending on the selected distance measure.Hierarchical clustering can be represented by a dendrogram. Cutting a dendrogram at a certain level gives a set of clusters. From abow dendrogram cutting at y=7 gives **n_cluster=4**.



7.5. Agglomerative hierarchical Clustering

Also known as bottom-up approach or hierarchical agglomerative clustering (HAC). A structure that is more informative than the unstructured set of clusters returned by flat clustering. This clustering algorithm does not require us to pre-specify the number of clusters. Bottom-up algorithms treat each data as a singleton cluster at the outset and then successively agglomerates pairs of clusters until all clusters have been merged into a single cluster that contains all data.



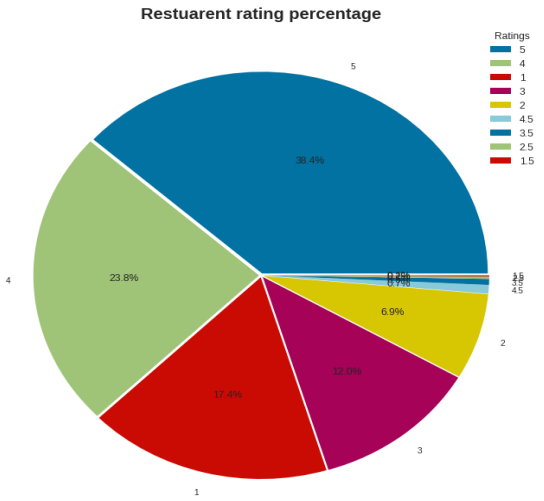
8. Sentiment Analysis

Sentiment Analysis is the process of computationally determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, deriving the opinion or attitude of a speaker.

Natural Language Processing is one part of Artificial Intelligence and Machine Learning to make an understanding of the interactions between computers and human (natural) languages. Sentiment analysis is one part of Natural Language Processing that is often used to analyze words based on the patterns of people in writing to find positive, negative, or neutral sentiments. Sentiment analysis is useful for knowing how users like something or not. Zomato is an application for rating restaurants. The rating has a review of the restaurant which can be used for sentiment analysis. Based on this, writers want to discuss the sentiment of the review to be predicted. The method used for preprocessing the review is to make all words lowercase, tokenization, remove numbers and punctuation, stop words, and lemmatization. Then after that, we create a word vector with the term frequency-inverse document frequency (TF-IDF). The data that we process are 10,000 reviews. After that, make positive reviews that have a rating of 3 and above, negative reviews that have a rating of 3 and below, and neutral ones that have a rating of 3. We use Split Test, 70% Data Training and 30% Data Testing.

8.1. Restaurant Rating Percentage

Even if majority ratings are good, we still have considerable count of poor ratings.The customers with a good number of followers who have given more reviews with constantly low ratings to understand the fields that need to be worked on.



* Logistic Regression

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on.

classification_report for LogisticRegression

	precision	recall	f1-score	support
average	0.89	0.36	0.51	948
bad	0.89	0.90	0.90	1810
good	0.88	0.98	0.93	4708
accuracy			0.88	7466
macro avg	0.89	0.75	0.78	7466
weighted avg	0.89	0.88	0.87	7466

*Random Forest Regression

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single mode

Classifiation report RandomForestClassifier

	precision	recall	f1-score	support
average	0.99	0.97	0.98	948
bad	0.99	0.98	0.99	1810
good	0.99	1.00	0.99	4708
accuracy			0.99	7466
macro avg	0.99	0.98	0.99	7466
weighted avg	0.99	0.99	0.99	7466

9. Word Cloud

Word Clouds are visual displays of text data – simple text analysis. Word Clouds display the most prominent or frequent words in a body of text (such as a State of the Union Address). Typically, a Word Cloud will ignore the most common words in the language (“a”, “an”, “the” etc.).

Word cloud (good)

Word cloud (average)

Word Cloud (bad)

10. Conclusion:

Clustering Analysis

- Top 5 cuisines are 'North Indian, Chinese','North Indian','Fast Food','Biryani, North Indian, Chinese','South Indian,North Indian, Chinese'.
- Bottom 5 cuisines are 'Biryani, South Indian,North Indian,Chinese, Andhra','Italian, North Indian,South Indian, Asian','North Indian,Fast Food','Arabian, North Indian, Chinese','Thai, Asian, Chinese, Malaysian'
- In the collection the top vocab used is week,visit,venues,veggie,unknown etc.
- In cuisines the top vocab used is wrap, thai,sushi,wraps, street,spanish,south etc.
- silhouette score at **n_clusters = 4**, we get highest silhouette score is 0.47229
- From the elbow method we get 4 numbers of clusters, which is best among all.
- Applied agglomerative hierarchical clustering from this we find 4 numbers of clusters that fit our model.
- By applying different clustering algorithms to our dataset. we get the optimal number of clusters is equal to 4.

Sentiment Analysis

- Categorize rating in 3 types i.e. good, bad and average. 4500+ good, 1700+ bad and 900+ average ratings given by customers.
- By using logistic regression and random forest regression model on

reviews dataset, we get 89% accuracy at logistic regression & 99% accuracy at random forest regression.

11. References:

1. **https://www.kaggle.com/**
2. **https://www.analyticsvidhya.com/**
3. **https://www.geeksforgeeks.org/**
4. **https://learn.almabetter.com/**
5. International Journal of Advanced Trends in Computer Science and Engineering.