Unsupervised – ML – Zomato Restaurant Clustering and Sentiment Analysis

Sethupathy M
Sri harish A

Capstone project 4 – Alma Better

Abstract:

This project contains two datasets, one of which contains the information about Zomato restaurants names and Meta data and the other contains Zomato restaurants names, rating, reviews. This project contains: Exploratory data analysis, text pre-processing, choosing appropriate features, clustering algorithms, model validation, dataset merging, sentiment analysis, model for sentiment analysis, analysis of model performance. The clustering helps to group the restaurants based on cuisines that benefits the customers to choose between the similar restaurants. Another point of view is to do sentiment analysis which helps the restaurant's owner to know the sort of reputation their restaurant has earned. With KMeans Clustering and DBSCAN Clustering restaurants are clustered with the silhouette score of 0.195 and 0.107 respectively.

In this experiment by using Clustering, EDA, Sentiment analysis, we are analyzing and visualizing the different perspectives. Which gives us different insights like

- Similar set of cuisines in various restaurants.
- Similar group of restaurants based on cuisines
- Most affordable and expensive restaurants
- Commonly used words by reviewers for expressing a positive or negative sentiment
- Restaurant with a greater number of positive sentiments
- Do the expensive restaurants always get higher ratings?
- Is there any relationship between ratings and review length?
- Distribution plot for polarity

Problem statement:

By using the given two datasets perform clustering and sentiment analysis to gain insights which would help improving the customer experience and to solve some of the business cases.

Features:

Zomato Restaurant Names and Metadata

- **Name –** Name of the Restaurants.
- **Links –** Links of the Restaurants.
- **Cost –** Average cost of the meal in Restaurants.
- **Collections** – The Collections in Zomato features popular restaurants across specific themes and trends at a particular location.
- **Cuisine –** A cuisine is specific set of cooking traditions and practices, often associated with a specific culture or region.
- **Timings –** Opening and closing time of the Restaurants.

Zomato Restaurant Reviews

- **Restaurants –** Name of the Restaurants.
- **Reviewer –** Name of the Reviewer.
- **Review –** Experience of the reviewer in the restaurant expressed in words.
- **Rating –** Experience of the reviewer in the restaurant expressed in numbers.
- **Metadata –** A set of data that describes and gives information about reviewers.
- **Time –** Time at which the reviews and ratings are given.
- **Pictures –** Number of pictures taken by the reviewer.

Introduction

Zomato is one of the most comprehensive and user-friendly apps where people can search for nearby restaurants and cafes, order food online, and get it delivered at their doorstep in no time. Moreover, you can also get accurate information about restaurants as it provides menus, reviews, and ratings. Based on that, users can place orders and enjoy lip-smacking food at their homes.

Zomato was founded by Deepinder Goyal and Pankaj Chaddah, two Delhi IIT graduates, in 2008. Till November 2010, Zomato was known as “Foodiebay.” Once they saw their colleagues who were seeking menus of different restaurants to order food. That’s when the idea took birth, and they thought of converting these manual menus into a digital format. In the year 2012, Zomato had spread its wings across the globe and started to list out the number of restaurants in the market.

Zomato’s Business Model is aimed at providing quality food services, information related to restaurants, their menus and user reviews. The Business Model of Zomato consists of providing food delivery services, information, user reviews and menu’s of partner restaurants. It has created a revolution in industries doing food business by including different restaurants and facilitating people to look for restaurants more conveniently.

The main work of Zomato is to suggest local and nearby restaurants to users and receive orders from them. Users can place orders from their favorite restaurant based on ratings and reviews shared by previous customers.

Step 1: From the desiccated app solution or website, users can explore various restaurants and order meals.
Step 2: Particular restaurant owners receive an order request and start preparing a meal.
Step 3: Once the food is ready to dispatch, it will be handed over to delivery providers.
Step 4: Delivery providers deliver the meal to the customer’s preferred location.
Step 5: From the given payment options, customers can make payments and share reviews based on their experience.

Steps involved:

Treating Null Values

Our dataset has missing values with many features which could lead us to loss of information and can mislead to wrong results so addressing Null values are important Null values are imputed with closest mean value as possible.

Exploratory Data Analysis

After loading the dataset, we performed EDA to get insights and also better understanding of the dataset. This process helped us figure out various aspects and relationships between the features. It gave us a better idea of which feature behaves in which manner compared to the other features.

Building structured multi-plot grids and graphs

When exploring multidimensional data, a useful approach is to draw multiple instances of the same plot on different subsets of your dataset. It allows a viewer to quickly extract a large amount of information about a complex dataset. Matplotlib offers good support for making figures with multiple axes; seaborn builds on top of this to directly link the structure of the plot to the structure of your dataset.

We have used count plot, box plot, distplot and pie chart in multi-plots for various features and the graphs provide a high level of convenience for comparison

Multicollinearity for all the variables has been plotted and some highly correlated variables have been detected.

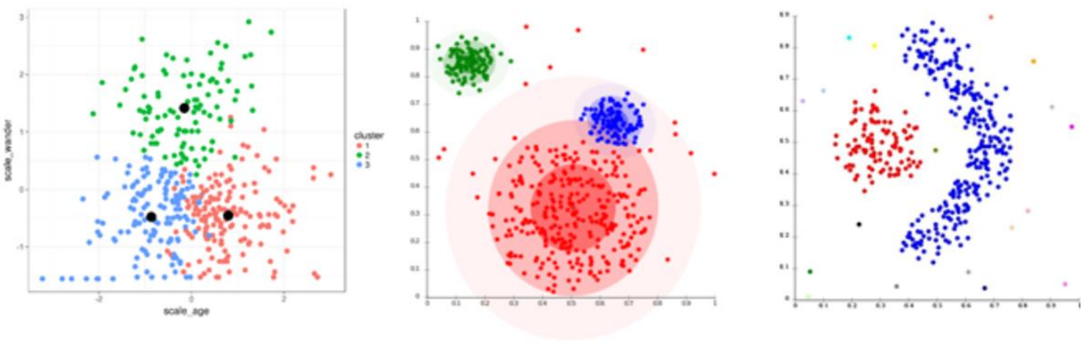
Text Pre-processing

Text preprocessing is an approach for cleaning and preparing text data for use in a specific context. Developers use it in almost all-natural language processing (NLP) pipelines, including voice recognition software, search engine lookup, and machine learning model training. It is an essential step because text data can vary. From its format (website, text message, voice recognition) to the people who create the text (language, dialect), there are plenty of things that can introduce noise into your data.

The ultimate goal of cleaning and preparing text data is to reduce the text to only the words that you need for your NLP goals.

Clustering

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 than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.



KMeans Clustering

K-means clustering algorithm computes the centroids and iterates until we it finds optimal centroid. It assumes that the number of clusters are already known. It is also called flat clustering algorithm. The number of clusters identified from data by algorithm is represented by ‘K’ in K-means.

In this algorithm, the data points are assigned to a cluster in such a manner that the sum of the squared distance between the data points and centroid would be minimum. It is to be understood that less variation within the clusters will lead to more similar data points within same cluster.

Working of K-Means Algorithm

We can understand the working of K-Means clustering algorithm with the help of following steps −

Step 1 − First, we need to specify the number of clusters, K, need to be generated by this algorithm.

Step 2 − Next, randomly select K data points and assign each data point to a cluster. In simple words, classify the data based on the number of data points.

Step 3 − Now it will compute the cluster centroids.

Step 4 − Next, keep iterating the following until we find optimal centroid which is the assignment of data points to the clusters that are not changing any more.

DBSCAN Clustering

DBSCAN is a popular density-based data clustering algorithm. To cluster data points, this algorithm separates the high-density regions of the data from the low-density areas. Unlike the K-Means algorithm, the best thing with this algorithm is that we don’t need to provide the number of clusters required prior.

DBSCAN algorithm group points based on distance measurement, usually the Euclidean distance and the minimum number of points. An essential property of this algorithm is that it helps us track down the outliers as the points in low-density regions; hence it is not sensitive to outliers as is the case of K-Means clustering.

The following are the DBSCAN clustering algorithmic steps:

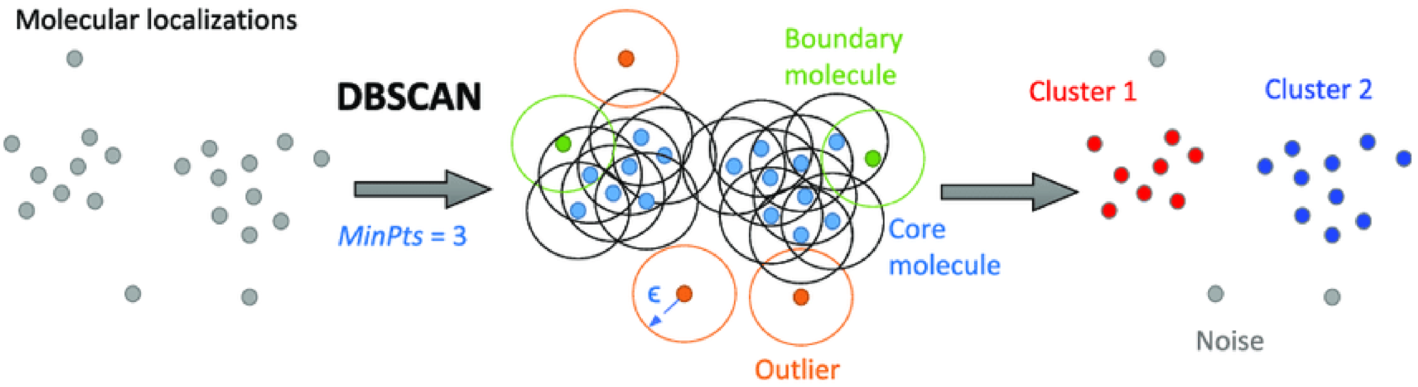
Step 1: Initially, the algorithms start by selecting a point (x) randomly from the data set and finding all the neighbor points within Eps from it. If the number of Eps-neighbours is greater than or equal to MinPoints, we consider x a core point. Then, with its Eps-neighbours, x forms the first cluster.

After creating the first cluster, we examine all its member points and find their respective Eps -neighbors. If a member has at least MinPoints Eps-neighbors, we expand the initial cluster by adding those Eps-neighbours to the cluster. This continues until there are no more points to add to this cluster.

Step 2: For any other core point not assigned to cluster, create a new cluster.

Step 3: To the core point cluster, find and assign all points that are recursively connected to it.

Step 4: Iterate through all unattended points in the dataset and assign them to the nearest cluster at Eps distance from themselves. If a point does not fit any available clusters, locate it as a noise point.



Vader Model

VADER (Valence Aware Dictionary and Entiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media.

It is used for sentiment analysis of text which has both the polarities i.e., positive/negative. VADER is used to quantify how much of positive or negative emotion the text has and also the intensity of emotion.

Advantages

- It does not require any training data.
- It can very well understand the sentiment of a text containing emoticons, slangs, conjunctions, capital words, punctuations and much more.
- It works excellent on social media text.
- VADER can work with multiple domains.

TextBlob

TextBlob is a Python library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

Libraries Used

Pandas

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals.

Pandas is mainly used for data analysis and associated manipulation of tabular data in Data Frames. Pandas allows importing data from various file formats such as comma-separated values, JSON, Parquet, SQL database tables or queries, and Microsoft Excel. Pandas allows various data manipulation operations such as merging, reshaping, selecting, as well as data cleaning, and data wrangling features. The development of pandas introduced into Python many comparable features of working with Data frames that were established in the R programming language. The pandas library is built upon another library NumPy, which is oriented to efficiently working with arrays instead of the features of working on Data frames.

NumPy

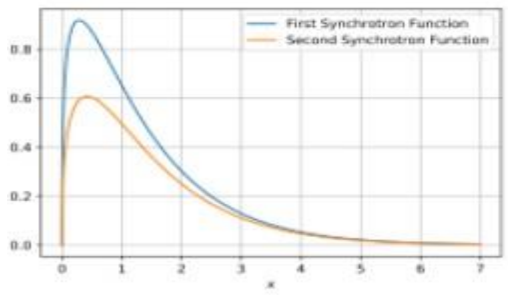
NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. NumPy targets the CPython reference implementation of Python, which is a non-optimizing bytecode interpreter. Mathematical algorithms written for this version of Python often run much slower than compiled equivalents due to the absence of compiler optimization. NumPy addresses the slowness problem partly by providing multidimensional arrays and functions and operators that operate efficiently on arrays; using these requires rewriting some code, mostly inner loops, using NumPy.

Using NumPy in Python gives functionality comparable to MATLAB since they are both interpreted, and they both allow the user to write fast programs as long as most operations work on arrays or matrices instead of scalars. In comparison, MATLAB boasts a large number of additional toolboxes, notably Simulink, whereas NumPy is intrinsically integrated with Python, a more modern and complete programming language. Moreover, complementary Python packages are available; SciPy is a library that adds more MATLAB-like functionality and Matplotlib is a plotting package that provides MATLAB-like plotting functionality. Internally, both MATLAB and NumPy rely on BLAS and LAPACK for efficient linear algebra computations.

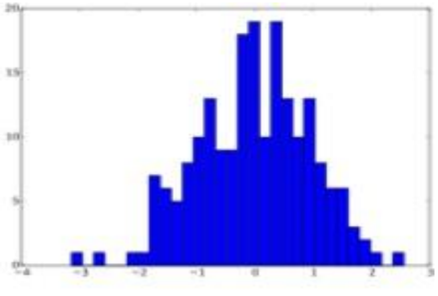
Python bindings of the widely used computer vision library OpenCV utilize NumPy arrays to store and operate on data. Since images with multiple channels are simply represented as three-dimensional arrays, indexing, slicing or masking with other arrays are very efficient ways to access specific pixels of an image. The NumPy array as a universal data structure in OpenCV for images, extracted feature points, filter kernels and many more vastly simplifies the programming workflow and debugging.

Matplotlib

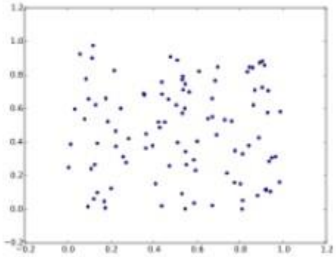
Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged. SciPy makes use of Matplotlib.



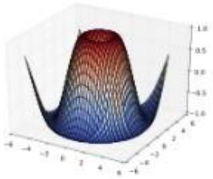
Line plot



Histogram



Scatter plot



3D plot

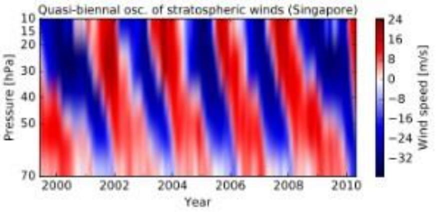
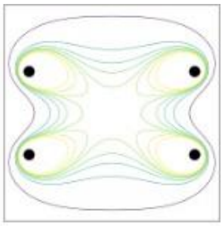
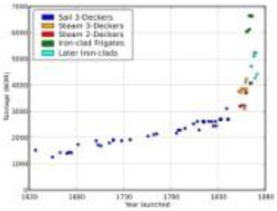


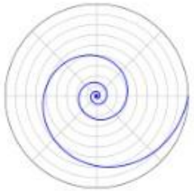
Image plot



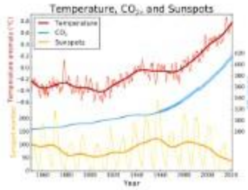
Contour plot



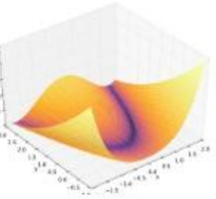
Scatter plot



Polar plot



Line plot



3-D plot

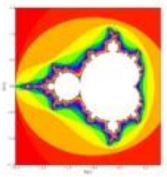


Image plot

Seaborn

Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures.

Seaborn helps you explore and understand your data. Its plotting functions operate on data frames and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them. Behind the scenes, seaborn uses matplotlib to draw its plots

Conclusion:

- **North Indian** cuisine is most common cuisine found in the restaurants.
- **Collage - Hyatt Hyderabad Gachibowli** is most expensive restaurant.
- **Amul** and **Mohammedia Shawarma** are the most affordable restaurants.
- The Restaurants are clustered on cuisines into **15** clusters by using KMeans clustering algorithm with the Silhouette score of **0.195**.
- **DBSCAN algorithm** is also used to cluster the restaurants into **15** clusters and also helps us to detect the outliers with the Silhouette score of 0.107.
- **Anvesh Chowdary** has given the greatest number of reviews.
- **AB's – Absolute Barbecues** is the top-rated restaurant.
- Almost 79 percent of the observations have Positive sentiment and 14 and 7 percent of the observations have Neutral and Negative sentiments respectively
- **Good** is the most common word in the Highly positive sentiment.
- **Worst** is the most common word in the Highly negative sentiment.
- **AB's – Absolute Barbecues, The Indi grill** and **B-Dubs** are the restaurants with the greatest number of positive reviews.
- **Arena Eleven** and **Banana leaf Multicuisine** restaurant are the restaurants with the greatest number of negative reviews.
- **Udipi's Upahar** is the most **affordable** restaurant with the **best** rating.
- **Feast - Sheraton Hyderabad Hotel** is the most **expensive** restaurant with the **best** rating.
- **Asian Meal Box** is the most **affordable** restaurant with the **worst** rating.
- **Club Rogue** is the **expensive** restaurant with **worst** rating.