Zomato Restaurant Clustering and Sentiment Analysis

Pavan Potnuru, Mohammed Haseebuddin
Data science trainees,
Alma Better, Bangalore

Abstract:

India is quite famous for its diverse multi cuisine available in a large number of restaurants and hotel resorts, which is reminiscent of unity in diversity. 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 Project focuses on Customers and Company, you have to analyze the sentiments of the reviews given by the customer in the data and made some useful conclusion in the form of Visualizations. Also, cluster the zomato restaurants into different segments. This could help in clustering the restaurants into segments.

1.Introduction

Online food-delivery platforms are expanding choice and convenience, allowing customers to order from a wide array of restaurants with a single tap of their mobile phone. The online food delivery market is no longer the underdog but has evolved into a champion. Having a food ordering marketplace platform was considered a state-of-the-art innovation in the early 2000s but today the segment has expanded to different demographics across the globe. Thanks to the increasing number of online users and vendors, the delivery providers have a sustainable business model. These businesses have immense opportunities to expand their services to a new geographical area and cater to new demographics.

2.Problem Statement

The Project focuses on Customers and Company, you have to analyse the sentiments of the reviews given by the customer in the data and made some useful conclusion in the form of Visualizations. Also, cluster the zomato restaurants into different segments. The data is visualized as it becomes easy to analyse data at instant. The Analysis also solve 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

3. Feature description

For this project we were given two data sets namely, “Zomato Restaurant names and meta data” and “Zomato Restaurant reviews”. Let us go through the features present in both the data sets.

Zomato Restaurant names and meta data

- 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

Zomato 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

5. Exploratory Data Analysis

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

The following are the various steps performed as a part of Exploratory Data Analysis:

- Data Preparation
- Data Cleaning
- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis

5.1 Data Preparation

Firstly, we imported libraries and dataset, some of the libraries used are NumPy, pandas, matplotlib, seaborn, warnings. Once the data is collected, process of analysis begins. But data has to be translated in an appropriate form. This process is known as Data Preparation.

5.2 Data Cleaning

The raw data received in the data set might not be directly suitable for analysis due to presence of unwanted data like, duplicate values, null values, outliers etc. We need to handle them first before we proceed with further analysis.

Removing Duplicates: The “Zomato Restaurant names and Meta data” dataset provided for this analysis consists of almost 105 records. None of them are duplicated and all of them are unique. Similarly, the “Zomato Restaurant Reviews” dataset provided consists of 10000 records and none of the are duplicated and all are unique. It is better to check for duplicate values before modelling.

Handling null/missing values: It is also possible that the given data set can contain missing information for some or all features in some records, we need to either remove them or find alternatives to fill up the null values. In “Zomato Restaurant names and meta data” dataset, more than half of the collections column are null values. So, we are not using that column for the analysis. Also, in “Zomato Restaurant Reviews” dataset, there are few null values in few columns. As the number of null values are low, we have dropped them. After splitting the meta data into two different columns namely, “Reviews” and “Followers” we found some more null values. As now the number is more, we have replaced null values with 0.

Feature Handling: We can manipulate some of the features according to our requirements to draw required information

from the data. For example, we have divided the “Time” column from “Zomato Restaurant Reviews” dataset to extract three more new features called, “year”, “month” and “hour” columns. Also, we have split the “Meta Data” column into “Reviews” and “Followers”.

5.3 Univariate Analysis: In Univariate Analysis, we choose a single feature from the data and try to determine what the output or the target value is, i.e., one feature/variable at a time.

• Understand the trends and patterns of data
• Analyze the frequency and other such characteristics of data
• Know the distribution of the variables in the data.
• Visualize the relationship that may exist between different variables.

5.4 Bivariate Analysis: In a Bivariate Analysis, we try to analyze two features instead of one, and finally determine the classification of output we are looking for. It is a methodical statistical technique applied to a pair of variables (features/ attributes) of data to determine the empirical relationship between them. In order words, it is meant to determine any concurrent relations.

5.5 Multivariate Analysis: In Multivariate analysis we analyze three or more different features at a time, to understand the relationship between all the features involved.

6. Model Building

6.1 Pre-requisites

Feature Engineering: Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning. In order to make machine learning work well on new tasks, it might be necessary to design and train better features.

Text Processing

Data Preprocessing is the most essential step for any Machine Learning model. How well the raw data has been cleaned and preprocessed plays a major role in the performance of the model. Likewise in the case of NLP, the very first step is Text Processing.

The various preprocessing steps that are involved are:

- Lower Casing
- Tokenization
- Punctuation Mark Removal
- Stop Word Removal
- Stemming
- Lemmatization

● Lower Casing: Converting a word to lower case (NLP -> nlp). Words like Book and book mean the same but when not converted to the lower case those two are represented as two different words in the vector space model (resulting in more dimensions).

● Tokenization: It is the process of tokenizing or splitting a string, text into a list of tokens. One can think of token as parts like a word is a token in a sentence, and a sentence is a token in a paragraph.

● Punctuation Mark Removal: The punctuation removal process will help to treat each text equally. For example, the word data and data! are treated equally after the process of removal of punctuations.

● Stop Word Removal: The idea is simply removing the words that occur commonly across all the documents in the corpus. Typically, articles and pronouns are generally classified as stop words.

● Stemming: This is the process of reducing a word to its word stem that affixes to suffixes and prefixes.
● Lemmatization: This is process of the grouping together of different forms of the same word and converting words into base or root form.

6.2 Sentiment Analysis

Sentiment analysis (or opinion mining) is a natural language processing (NLP) technique used to determine whether data is positive, negative or neutral. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs. In this project we will try to see the sentiment of the review given by the customers.

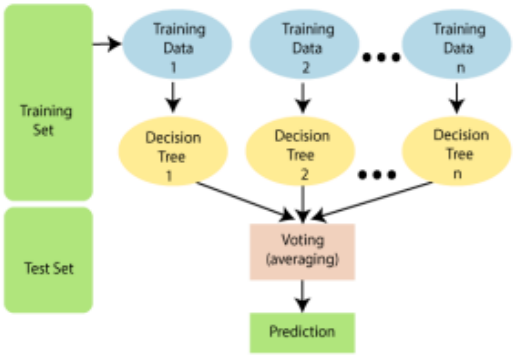
Subjectivity: Subjectivity quantifies the amount of personal opinion and factual information contained in the text. Subjectivity lies between [0,1]. The higher subjectivity means that the text contains personal opinion rather than factual information.

Polarity: Polarity score tells us how positive or negative the text is, it ranges between (- 1 to +1) if score in negative it means sentiment is negative, if score is positive, it means sentiment is positive. If the Polarity score is 0, it means the text is neither positive nor negative but neutral.

6.3 Model building

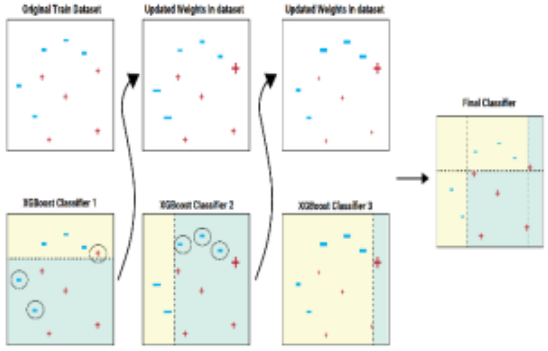
Naïve Bayes Classifier: Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

Random Forest: Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

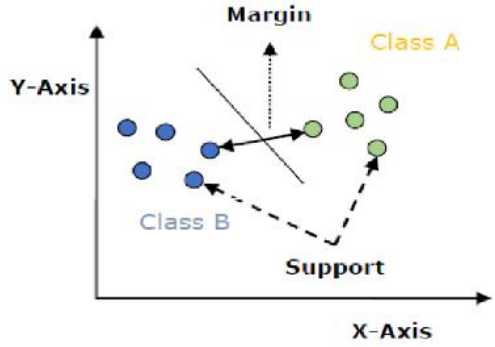


XGBoost Algorithm: In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. XGBoost comes under the boosting ensemble

techniques which combines the weakness of primary learners to the next strong and compatible learners.



Support Vector Machines (SVM):

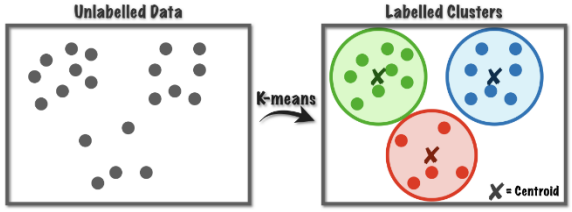


An SVM model is basically a representation of different classes in a hyperplane in multidimensional space. The hyperplane will be generated in an iterative manner by SVM so that the error can be minimized. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH).

6.4 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.

K Means Clustering: K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on. It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabelled dataset on its own without the need for any training. It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.



7. Conclusion

- 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 cheapest is the food joint called Mohammedia Shawarma and the costliest restaurant is Collage – Hyatt Hyderabad Gachibowli.
- Sentiment Analysis was done on the reviews and a model was trained in order to identify negative and positive sentiments.
- SVM and XGB both performed well, and we can choose any one them.

- SVM and XGB are having 0.921 and 0.981 of testing accuracy, respectively.
- We got best cluster as 5 in K-Means and Principal Component Analysis (PCA).