Unsupervised ML Clustering
Capstone Project

Zomato Restaurant clustering And Sentiment Analysis

by: Madhavi Mali

Data Science Trainee, Alma Better

1.Abstract -

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.

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. Restaurant business in India is always evolving. More Indians are warming up to the idea of eating restaurant food whether by dining outside or getting food delivered. 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 analysing the Zomato restaurant data for each city in India.

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 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 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.Attributes-

Zomato Restaurant names and Metadata

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

Zomato Restaurant reviews (merged with MetaData and Names)

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

I. 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 hypotheses and to check assumptions with the help of summary statistics and graphical representations. It gives us better idea of which feature behaves in which manner compared to target variable.

II. Data Cleaning:

Our Data contains some null values which might tend to disturb our accuracy hence we dropped them at the beginning of our project to get better results.

III. Feature Engineering:
Feature engineering includes scaling and encoding. Scaling is essential for machine learning algorithms that calculate distances between data. If not scale, the feature with a higher value range starts dominating when calculating distances.

IV. Fitting Models:

We used the Multinomial Naïve Bayes , Random Forest Classifier , XGB classifier and Support Vector Classifier. We used k-means clustering algorithm for clustering and then checked the model performance using elbow method to find the number of clusters.

5.Model Building Pre-Requisites

1)Removing Punctuations:

- Punctuations does not carry any meaning clustering.
- So, removing punctuations helps to get rid of unhelpful parts of the data, or noise.

2)Removing Stop words:

- Stop words are basically a set of commonly used words in any language, not just English.
- If we remove the words that are very commonly used in a given language, we can focus on the important words instead.

3)Stemming:

- Stemming is the process of removing a part of a word, or reducing a word to its stem or root.
- Applying stemming to reduce words to their basic form or stem, which may or may not be a legitimate word in the language

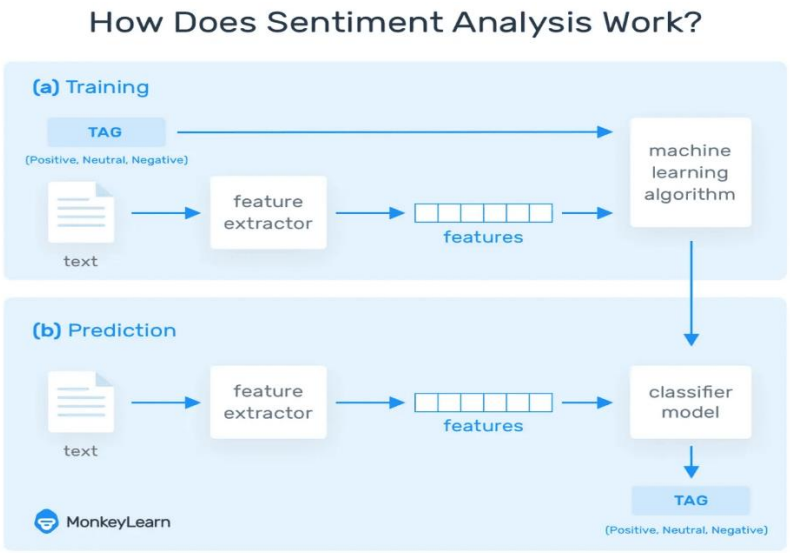
4)Vectorizing the text data using TF_IDF Vectorizer:

- Here we have textual data.
- Clustering algorithms cannot understand textual data.
- So, we use vectorization technique to convert textual data to numerical vectors

6.Sentiment Analysis

Sentiment analysis is a machine learning tool that analyzes texts for polarity, from positive to negative. By training machine learning tools with examples of emotions in text, machines automatically learn how to detect sentiment without human input.

To put it simply, machine learning allows computers to learn new tasks without being expressly programmed to perform them. Sentiment analysis models can be trained to read beyond mere definitions, to understand things like, context, sarcasm, and misapplied words



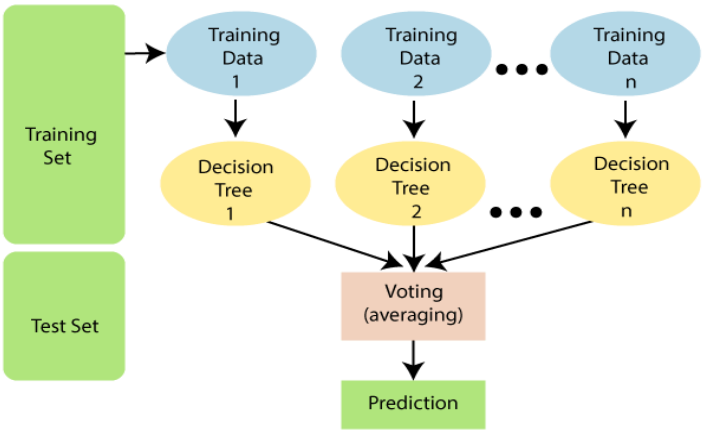
Random Forest

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "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.

The below diagram explains the working of the Random Forest algorithm:



Naive Bayes

Naive Bayes is a fairly simple group of probabilistic algorithms that, for sentiment analysis classification, assigns a probability that a given word or phrase should be considered positive or negative.

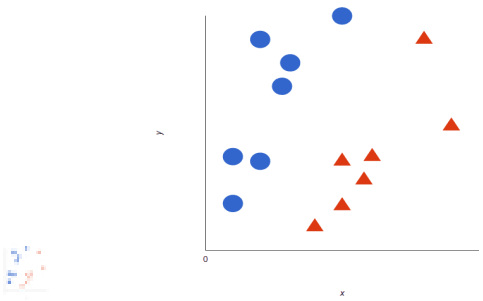
Essentially, this is how Bayes' theorem works. *The probability of A, if B is true, is equal to the probability of B, if A is true, times the probability of A being true, divided by the probability of B being true*

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Support Vector Machines (SVM)

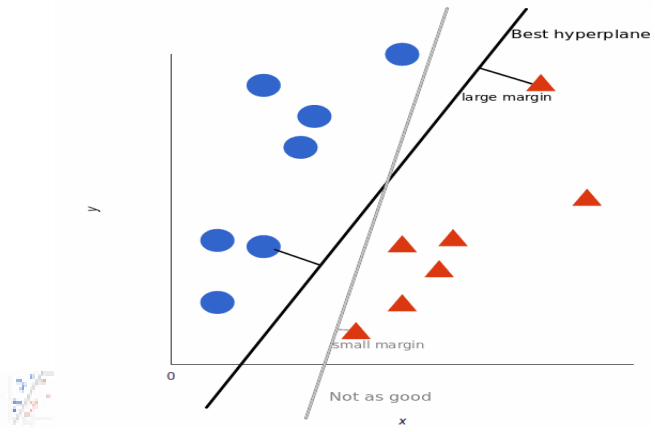
A support vector machine is another supervised machine learning model, similar to linear regression but more advanced. SVM uses algorithms to train and classify text within our sentiment polarity model, taking it a step beyond *X/Y* prediction.

For a simple visual explanation, we'll use two tags: *red* and *blue*, with two data features: *X* and *Y*. We'll train our classifier to output an *X/Y* coordinate as either *red* or *blue*.

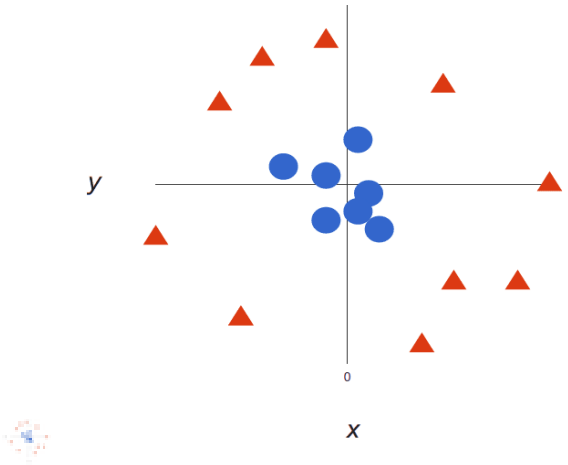


The SVM then assigns a hyperplane that best separates the tags. In two dimensions this is simply a line (like in linear regression). Anything on one side of the line is *red* and anything on the other side is *blue*. For sentiment analysis this would be *positive* and *negative*.

In order to maximize machine learning, the best hyperplane is the one with the largest distance between each tag:

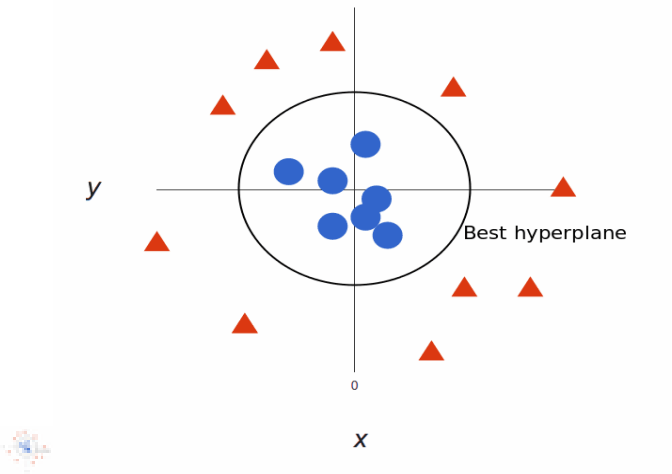


However, as data sets become more complex, it may not be possible to draw a single line to classify the data into two camps:



Using SVM, the more complex the data, the more accurate the predictor will become. Imagine the above in three dimensions, with a *Z* axis added, so it becomes a circle.

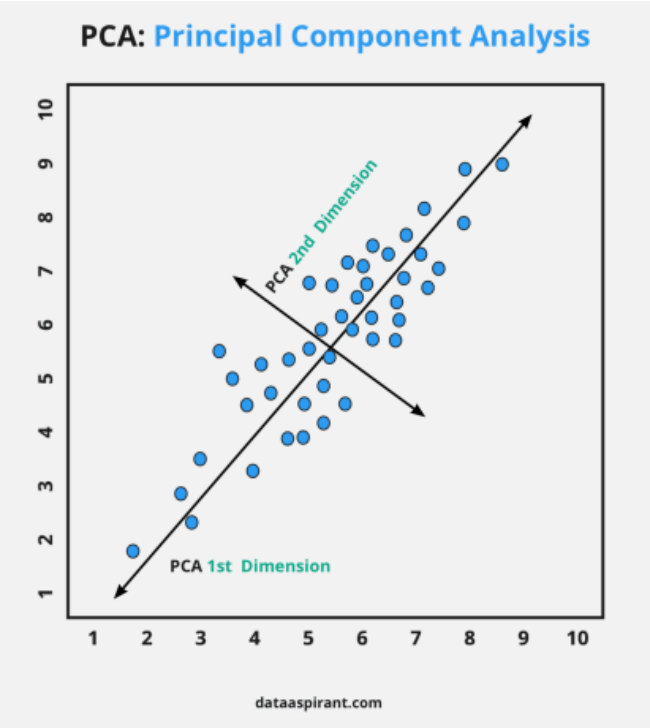
Mapped back to two dimensions with the best hyperplane, it looks like this:



Very simply put, SVM allows for more accurate machine learning because it's multidimensional.

7.Model Building

1) Principal Component Analysis:



A dimensionality-reduction technique in which transformation of high dimensional correlated data is performed into a lower-dimensional set of uncorrelated components also referred to as principal components.

Its Main Objectives are:

- To visualize the high dimensionality data.
- To introduce improvements in classification.
- To obtain a compact description.
- To capture as much variance in the data as possible.
- To decrease the number of dimensions in the dataset.
- To search for patterns in the dataset of high dimensionality.
- To discard noise.

2)K-Means Clustering Algorithm

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. In this topic, we will learn what is K-means clustering algorithm, how the algorithm works, along with the Python implementation of k-means clustering.

What is K-Means Algorithm?

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabelled 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.

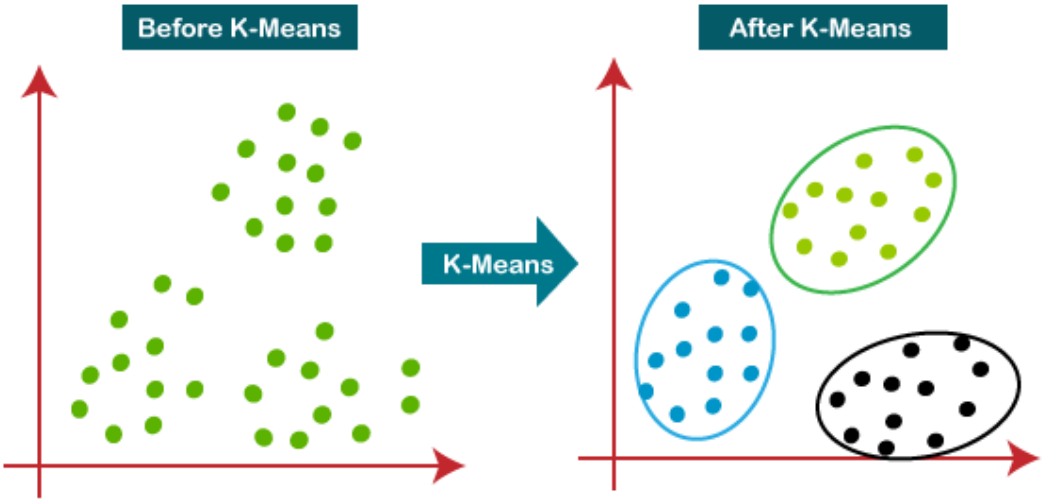
The algorithm takes the unlabelled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks:

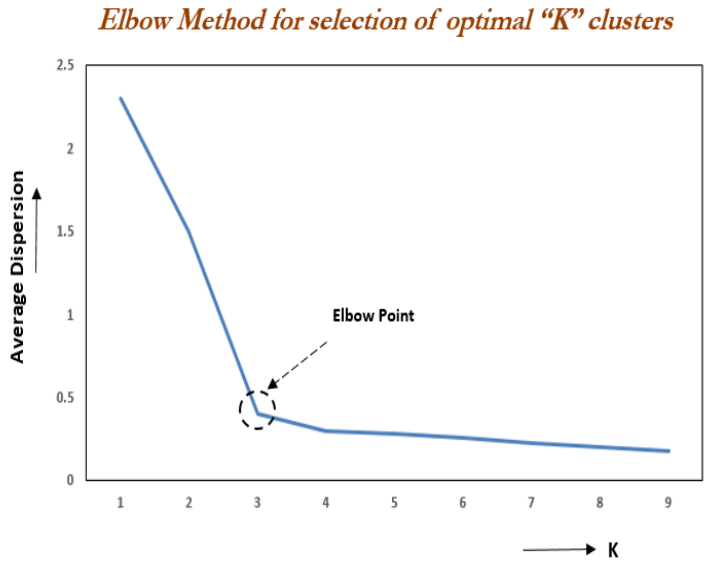
- Determines the best value for K centre points or centroids by an iterative process.
- Assigns each data point to its closest k-centre. Those data points which are near to the particular k-center, create a cluster.

Hence each cluster has data points with some commonalities, and it is away from other clusters.

The below diagram explains the working of the K-means Clustering Algorithm:



Elbow method to find optimum k value:



Elbow Method is an empirical method to find the optimal number of clusters for a dataset. In this method, we pick a range of candidate values of k, then apply K-Means clustering using each of the values of k. Find the average distance of each point in a cluster to its centroid, and represent it in a plot. Pick the value of k, where the average distance falls suddenly.

8. Conclusion:

The analysis is done with Zomato data. We used the Random forest classifier, XGB classifier, Naïve bayes classifier and Support Vector Machine to find the accuracy . We used k-means clustering algorithm and then checked the model performance elbow method to find the optimal number of clusters and finally plotted the Word cloud to visualize various clusters.

The experimental results showed that:

• 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 0.9188 and 0.9369 of testing accuracy respectively.

• We got best cluster as 5 in K-Means and Principal Component Analysis(PCA)