Zomato Restaurant Clustering And Sentiment Analysis

Sharad Tawade, Sagar Malik
Vinay Kumar
Data science trainee,
AlmaBetter

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 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.

Sentiment Analysis or opinion mining indicates the utilization of computational linguistics, textual analysis, biostatistics, and NLP (Natural Language Processing) to consistently recognize, skin, compute and grasp effective levels and abstract information.

In this Project we focus on Customers and Company, we 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 using clustering algorithms.

Keywords:eda, Food Reviews, NLP, Sentiment Analysis, Text Visualization, Clustering.

1.Problem Statement

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 analyzing the Zomato restaurant data for each city in India.

The main objective is to cluster the restaurants into segments and sentiment analysis of reviews. Also the data has valuable information around cuisine and costing which can be used in cost vs. benefit analysis.

Attribute Information:

Zomato Restaurant names and Metadata:

- **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 Restaurants
- **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.

2. Introduction

In a globally networked world, where the internet is a boon to the immediate feedback of sentiments that are based on emotions. People review the services and products being offered to them on various websites. These websites not only help customers to share their reviews but also allows other customers and business owners to improve the range of their services. Users with different backgrounds furnish their reviews in different languages and scripts which are gold for the opinion miners. Online visualization of the user's feedback is necessary, which is indeed responsible for the overall insight of a user and is increasing the indulgence of individuals in the field of opinion mining and analysis. Sentiment analysis is responsible for the wrong beliefs prevailing in the user's mind regarding any service or product. It chooses the best of comments and analyzes the same for a specific set of interests.

We here use many classifiers for sentiment analysis like MultinomialNB , Random Forest, XGB and SVM. We use clustering techniques K-Means and PCA (Principal Components Analysis) for cuisines.

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:

In this process we drop unnecessary columns and convert some columns in useful format.

III. Features Selection:

The feature selection process is based on a specific machine learning algorithm that we are trying to fit on a given dataset. It follows a greedy search approach by evaluating all the possible combinations of features against the evaluation criterion.

For sentiment analysis, we have used rating and reviews features. - For clustering we got cost, cuisine and timing of the restaurant as the features to build the model.

IV. Text Pre-processing:

Text processing is essential for NLP algorithms. It contains Removing punctuation, removing stop words and lemmatization.

V. Fitting Models for sentiment and clustering:

At first we tried the MultinomialNB model for sentiment analysis and then used a Random Forest Classifier, XGB Model, Support Vector Machine and compared the results. For clustering of cuisines we used K-Means clustering and PCA(Principal Component Analysis).

5. Algorithms

I. MultinomialNB:

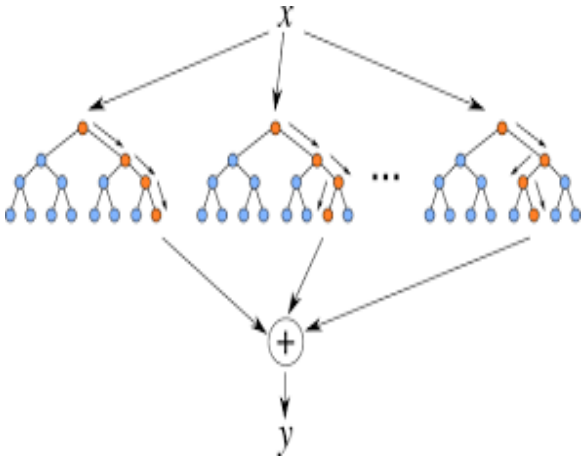
Naive Bayes classifier for multinomial models.The multinomial Naive Bayes classifier is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts. However, in practice, fractional counts such as tf-idf may also work.

MultinomialNB implements the naive Bayes algorithm for multinomial distributed data, and is one of the two classic naive Bayes variants used in text classification (where the data are typically represented as word vector counts, although tf-idf vectors are also known to work well in practice). The distribution is parametrized by vectors **θy=(θy1,…,θyn)** for each class **y**, where **n** is the number of features (in text classification, the size of the vocabulary) and **θyi** is the probability P(xi∣y) of feature **i** appearing in a sample belonging to class **y**.

II. Random Forest:

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

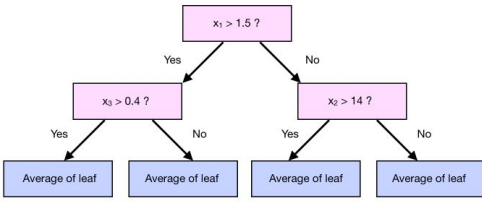
One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification.



III. XGBoost:

To understand XGBoost we have to know gradient boosting beforehand.

- Gradient Boosting:



Gradient boosting is a type of machine learning boosting. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set the target outcomes for this next model in order to minimize the error.

How are the targets calculated? The target outcome for each case in the data depends on how much changing that case's prediction impacts the overall prediction error.

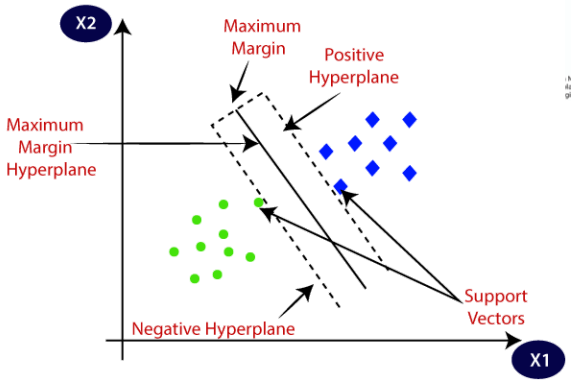
XGBoost is one of the fastest implementations of gradient boosting trees. It does this by tackling one of the major inefficiencies of gradient boosted trees: considering the potential loss for all possible splits to create a new branch (especially if you consider the case where there are thousands of features, and therefore thousands of possible splits). XGBoost tackles this inefficiency by looking at the distribution of features across all data points in a leaf and using this information to reduce the search space of possible feature splits.

IV. Support Vector Machine:

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

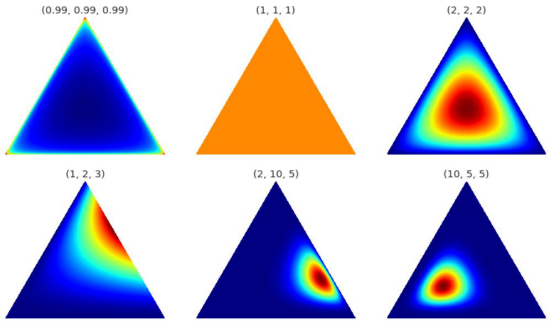
SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence the algorithm is termed as Support Vector Machine.



V. Latent Dirichlet Allocation(LDA):

It is one of the most popular topic modeling methods. Each document is made up of various words, and each topic also has various words belonging to it. The aim of LDA is to find topics a document belongs to, based on the words in it. It assumes that documents with similar topics will use a similar group of words. This enables the documents to map the probability

distribution over latent topics and topics are probability distribution.



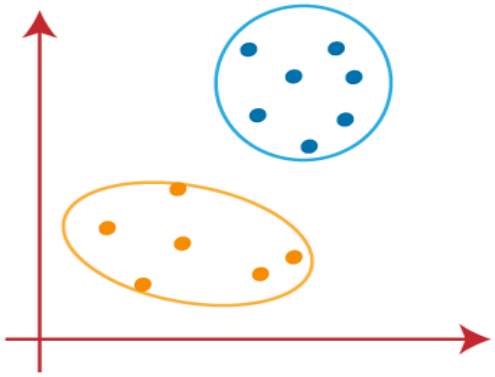
The Dirichlet distribution is parameterized by the vector α, which has the same number of elements K as the multinomial parameter θ. Above is the visualization of the Dirichlet distribution, for our purpose, we can assume that corners/vertices represent the topics with words inside the triangle (the word is closer to the topic if it frequently relates with it.) or vice-versa.

6. Clustering Algorithms

I. K-Means Clustering:

The K-means clustering algorithm computes centroids and repeats until the optimal centroid is found. It is presumptively known how many clusters there are. It is also known as the flat clustering algorithm. The number of clusters found from data by the method is denoted by the letter 'K' in K-means.

In this method, data points are assigned to clusters in such a way that the sum of the squared distances between the data points and the centroid is as small as possible. It is essential to note that reduced diversity within clusters leads to more identical data points within the same cluster.



K-means implements the Expectation-Maximization strategy to solve the problem. The Expectation-step is used to assign data points to the nearest cluster, and the Maximization-step is used to compute the centroid of each cluster.

II. K-Means with Principal Component Analysis(PCA):

PCA is fundamentally a dimensionality reduction algorithm, but it can also be useful as a tool for visualization, for noise filtering, for feature extraction and engineering, and much more.

principal components are the continuous solutions to the discrete cluster membership indicators for K-means clustering. New lower bounds for K-means objective function are derived, which is the total variance minus the eigenvalues of the data covariance matrix. These results indicate that unsupervised dimension reduction is closely related to unsupervised learning.

It is a common practice to apply PCA (principal component analysis) before a clustering algorithm (such as k-means). It is believed that it improves the clustering results in practice (noise reduction).

PCA's main weakness is that it tends to be highly affected by outliers in the data. For this reason, many robust variants of PCA have been developed, many of which act to iteratively discard data points that are poorly described by the initial components.

6. Model Performance

I. Confusion Matrix:

The confusion matrix is a table that summarizes how successful the classification modelis at predicting examples belonging to various classes. One axis of the confusion matrix is the label that the model predicted, and the other axis is the actual label.

II. Precision/Recall:

Precision is the ratio of correct positive predictions to the overall number of positive predictions : TP/TP+FP

Recall is the ratio of correct positive predictions to the overall number of positive examples in the set: TP/FN+TP.

III. Accuracy:

Accuracy is given by the number of correctly classified examples divided by the total number of classified examples. In terms of the confusion matrix, it is given by: TP+TN/TP+TN+FP+FN

IV. Area under ROC Curve(AUC):

ROC curves use a combination of the true positive rate (the proportion of positive examples predicted correctly, defined exactly as recall) and false positive rate (the proportion of negative examples predicted incorrectly) to build up a summary picture of the classification performance.

7. HyperParameter Tuning

Hyper-parameters are those sets of information that are used to control our parameters in order to get good results. We used Grid Search CV for hyper parameter tuning.

Grid Search CV :

It is the process of performing hyperparameter tuning in order
to determine the optimal values for a given model. As mentioned above, the performance of a model significantly depends on the value of hyperparameters. Note that there is no way to know in advance the best values for hyperparameters so ideally, we need to try all possible values to know the optimal values. Doing this manually could take a considerable amount of time and resources and thus we use GridSearchCV to automate the tuning of hyperparameters.

GridSearchCV is a function that comes in Scikit-learn's(or SK-learn) model_selection package.So an important point here to note is that we need to have the Scikit-learn library installed on the computer. This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set. So, in the end, we can select the best parameters from the listed hyperparameters.

8. Conclusion:

That's it! We reached the end of our exercise.Starting with loading the data so far we have done EDA, null values treatment, text preprocessing, and then model building.

We got the best cluster as 5 in K-Means and Principal Component Analysis(PCA). We plot polarity and subjectivity plot for sentiment analysis, Polarity tells how positive or negative the text is. The subjectivity tells how subjective or opinionated the text is, from this plot positive feedbacks are more. For sentiment analysis we used supervised techniques. We got the best model as an SVM (Support Vector Machine) classifier.