Zomato Restaurant Clustering and SentimentAnalysis

By Mayank Ghai
Data Science Trainee
AlmaBetter, Bangalore

Abstract:

India is well-known for its unique multi-food cuisine, which is offered in a huge number of restaurants and hotel resorts and symbolizes unity in variety. In India, the restaurant industry is changing rapidly. More People are appealed to the concept of eating restaurant meals, whether they dine outside or have food delivered to their homes. The increasing number of restaurants in every Indian state has encouraged analysis ofthe information to gain some insights, noteworthy facts, and statistics about the Indian food sector. As a result, the purpose of this studyis to analyze Zomato restaurant data in Hyderabad. Zomato is a restaurant aggregator and food delivery service based in India. 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

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 ofrestaurants, and also has food delivery options from partner restaurants in select cities.

The Project focuses on Customers and companies,you have to analyze the sentiments of the reviewsgiven by the customer in the data and make someuseful conclusions in the form of Visualizations. Also, cluster the zomato restaurants into differentsegments. The data is visualized as it becomes easy to analyze data in an instant. The Analysis also solves some of the business cases that can directlyhelp the customers find the Best restaurant in their locality and for the company to grow up and work in the fields they are currently lagging in.

This could help in clustering the restaurants into segments. Also, the data has valuable informationabout cuisine and costs which can be used in cost vs. benefit analysis

Data could be used for sentiment analysis.

Also, the reviewers' metadata can be used to identify the industry's critics.

Introduction:

In today's digitized modern world, the popularity of food apps is increasing due to their functionality to view, book, and order food with a few clicks on the phone for their favorite restaurant or cafes, by surveying the user ratings and reviews of the previously visited customers. Food apps like Zomato provide a secular part where users can rate their experience of the visited restaurant or café. Zomato also provides columns for writing classified user reviews. Sharing on the internet is something we usually do. Giving a review is also a useful activity so that other people on the internet can find out something else and see opinions about things. The usual things are reviewed by someone in the form of experiences, places, objects, and others. When giving a review we usually use text to explain something that we experience with an item, place, or event that we normally experience.

Zomato is a site where someone can give a review of a restaurant, how the restaurant is, and someone's opinion about the restaurant. Restaurant customer satisfaction can be analyzed by their review on Zomato. Sometimes, restaurants see the reviews in Zomato, but they don't get if the reviews are positive or negative to their restaurants. Reviews on Zomato are still in the form of text and can be classified with positive, negative, or neutral ratings. Zomato doesn't have an analysis of how users interact with the reviews and what words will indicate whether they like it or not. We need to extract the words in the review and analyze them so we can know how users interact in Zomato and get customers' satisfaction by their reviews.

In this paper, we propose a method to analyze users' sentiment of Zomato Restaurants. We are using different classifiers to classify the sentiments of users based on their reviews. We also find words that affect the classifier model. Also, we focus on mining customer reviews, authenticating them, and classifying them into positive and negative reviews. We also clustered the restaurants based on their cuisines

Approach:

The approach followed here is to first check the sanctity of the data and then understand the features involved. The events followed were inour approach:

- **Understanding the business problemand the datasets**
  - **Data cleaning and preprocessing-** Both datasets required little cleaning; all that was required was to remove certain null values, convert values to acceptable data types, and selectonly the most significant features. Features like Link, Collections, and Timing, for example, don'thelp distinguish across instances.
  - **Feature Engineering :**

    The process of selecting, modifying, and transforming raw data into meaningful numericalfeatures that machine learning algorithms can exploit is known as feature engineering.
  - **Exploratory data analysis-** of categorical and continuous variables against our target variable.
  - **Restaurant Clustering:** Clustering is done based on the two approaches
    1. K-mean
    2. Principal Component Analysis
  - **Sentiment Analysis:** Sentiment analysis is done using a different machine learning model. The selected model should be able to predict a False positive that is the sentiment is actually negative but the model predicted it as a positive one.

Understanding the Data:

2.1.1 Restaurant Names and Metadata

- Name : Name of Restaurants
- Links : URL Links of Restaurants
- Cost : Per person estimated Cost ofdining
- Collection : Tagging of Restaurants w.r.t. Zomato categories
- Cuisines : Cuisines served by Restaurants
- Timings : Restaurant Timings

2.1.2 Restaurant Reviews

- Restaurant: Name of the Restaurant
- Reviewer: Name of the Reviewer
- Review: Review Text
- Rating: Rating Provided by Reviewer
- MetaData: Reviewer Metadata - No. ofReviews and followers
- Time: Date and Time of Review
- Pictures: No. of pictures posted withreview

Data Cleaning and Preprocessing:

Handling missing values is an important skill in the data analysis process. If there are very few missing values compared to the size of the dataset, we may choose to drop rows that have missing values. Otherwise, it is better to replace them with appropriate values.

It is necessary to check and handle these values before feeding them to the models, to obtain good insights into what the data is trying to say, and to make great characterization and predictions which will in turn help improve the business's growth.

Exploratory Data Analysis:

Exploratory data analysis is a crucial part of data analysis. It is looking through and assessing a dataset to find patterns, trends, and conclusions that may be used to make better data-related decisions. The results are generally summarized using statistical graphics and other data visualization tools. To study the data, pandas is used, while matplotlib and seaborn are used to visualize it.

The following are some essential results from the analysis:

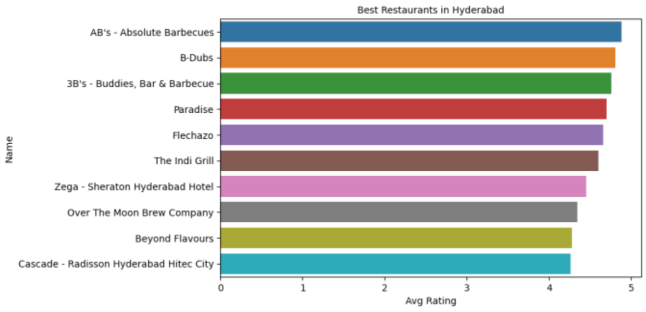
- Best restaurants in the City
- The Most Popular Cuisines in Hyderabad
- Restaurants and their Costs
- Cost-Benefit Analysis

Best restaurants in the City

Food, ambiance, cost, location, ratings, and other considerations all have a role in selecting adecent 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. Reviews are put in place to aid in the above-mentioned judgments. They offer you a sense of what the restaurant is like based on the experiences of people who have visited it multiple times.

To aid in decision-making, the dataset includes the following features: Name, Cost, Total Cuisines, and Average Ratings. The best restaurants in the city would be those with reasonable prices, great ratings, and a large variety of cuisines.

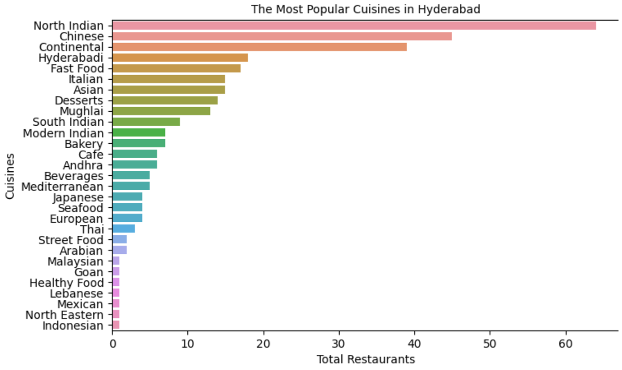
This is a plot of the sorted data, and these are the best restaurants based on the factors indicated above.



The Most Popular Cuisines in Hyderabad:

The most popular cuisines are those that are offered by the majority of restaurants in Hyderabad. Here's a plot of the various cuisines served in Hyderabad, along with the total numberof restaurants that serve them. Despite its locationin South India, North Indian cuisine is the most popular in restaurants, followed by Chinese andContinental cuisines. The variety of cuisines

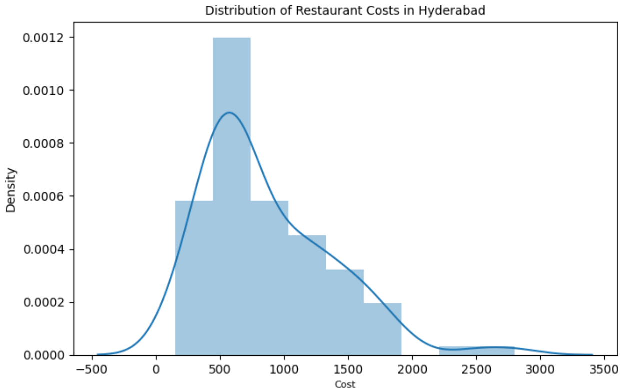
available in Hyderabad demonstrates the city'snumerous dining options.



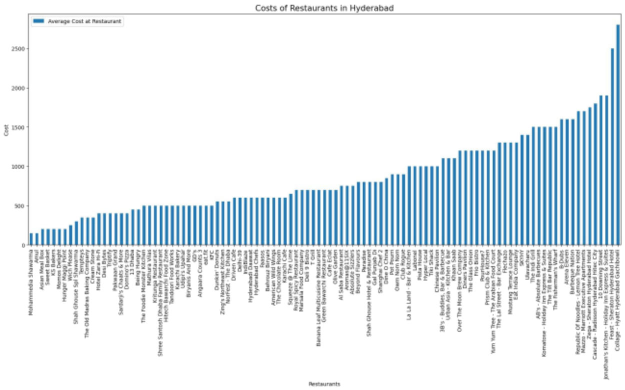
Restaurants and their Costs

The cost per person in Hyderabadi restaurants ranges from 150 INR to 2800 INR. The cheapestrestaurant is Mohammedia Shawarma, while themost expensive is Collage - Hyatt Hyderabad

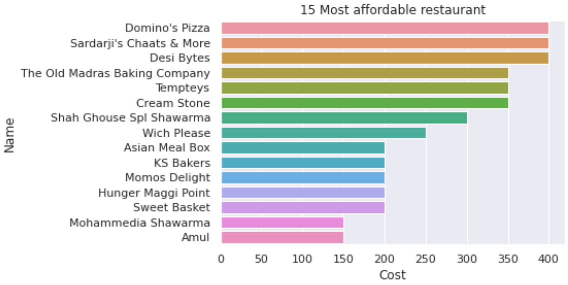
Gachibowli.



Restaurants and their Costs

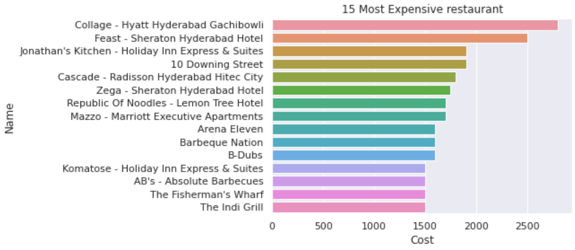


Most affordable restaurant



The cheapest restaurants in the dataset are basically small food joints and bakeries.

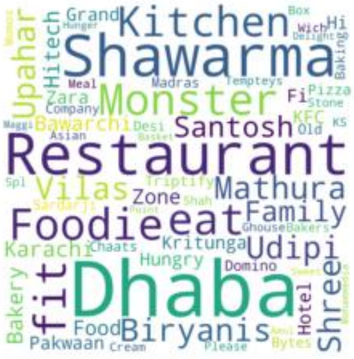
Most Expensive restaurant



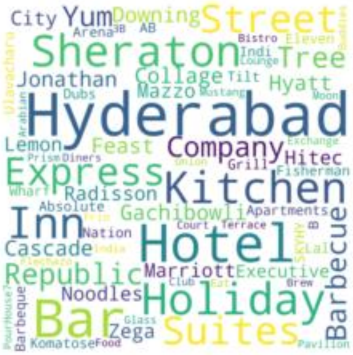
The most expensive restaurants in the dataset are restaurants by 4-star above hotels.

Word Clouds

```
(-0.5, 1399.5, 1399.5, -0.5)
```



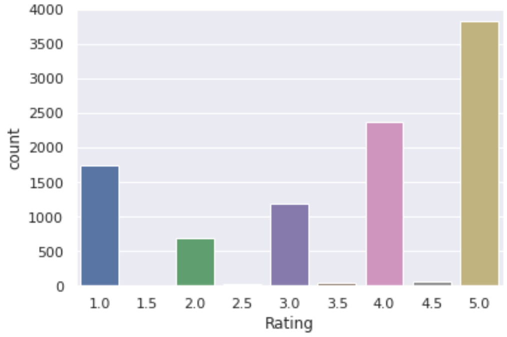
```
(-0.5, 1399.5, 1399.5, -0.5)
```



Reviews

Rating

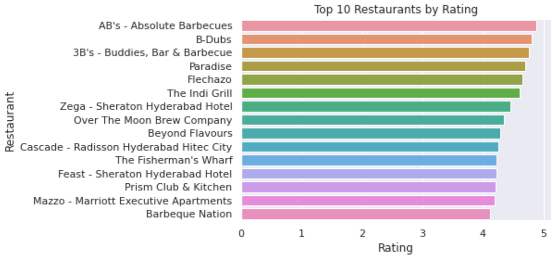
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f045bc1c810>
```



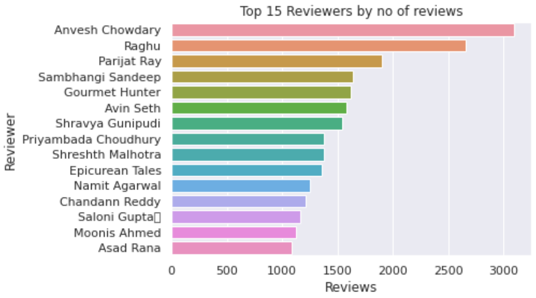
Even if the majority ratings are good, we still have a considerable count of poor ratings

Restaurants by Rating

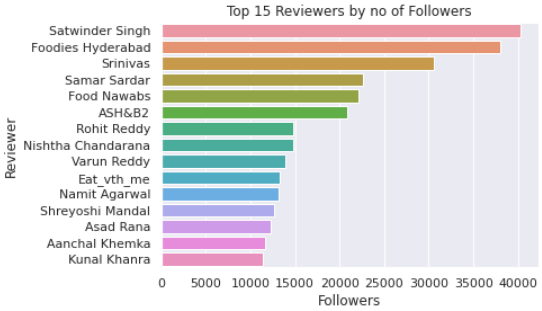
```
Text(0.5, 1.0, 'Top 10 Restaurants by Rating')
```



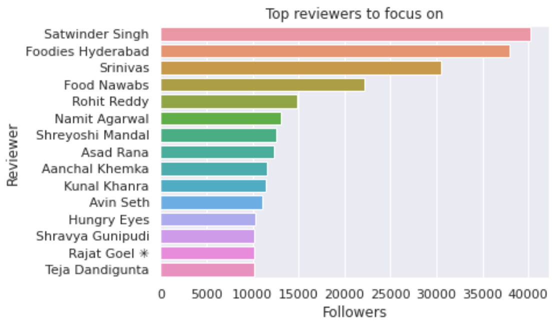
Reviewers by no of reviews



Reviewers by no of Followers

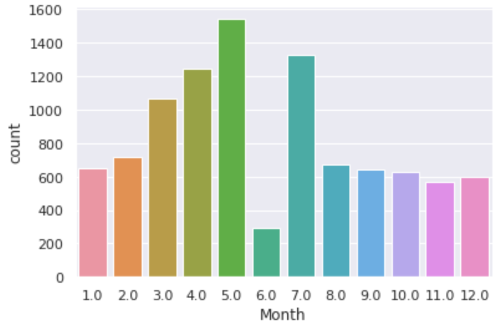


Reviewers to focus on:



These are the reviewer a restaurant should focus on who have reviewed more than 100 restaurants and have followers greater than 10000 with an average rating above 3.5

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f045bf06b10>
```



Most of the reviews are in the month of 5 and 7 months of the year

Cost-Benefit Analysis

Every time you engage in a company endeavor ormake a business choice, you must consider whether the option is worthwhile. A Cost-Benefit Analysis is a method of evaluating the value of a choice by estimating the costs of implementing it and comparing them to the benefits of doing so. Ifthe expected benefits outweigh the costs, you'll profit from the decision; if not, it's time to devise a better strategy.

Zomato is an online food delivery service and a search engine for Indian restaurants.

Zomato is afood delivery service that focuses on internet orders, restaurant reservations, and reward programs. Restaurant chains that want to reach a wider audience, as well as app users who simply want to try out local eateries and cuisines, are the company's target clients. Here is a simple cost-benefit analysis that can be performed based on the limited information available

Costs

When calculating costs, start with direct costs, which are expenses directly tied to the production or development of a product or service (or the implementation of a project or business decision), which in Zomato's case is essentially the mobile app. Maintaining the application, conceptualizing strategies, including restaurants, marketing, food delivery partners, and customer service, necessitates the participation of a large staff. The employees' pay would be a direct cost.

Utilities, rent, partners, advertisements, and other indirect costs are examples.

Other expenses are difficult to quantify, such as negative platform reviews that cause customersto avoid using the app altogether, a poor social network presence, and so on.

Benefits

Advertising is the primary source of revenue. More restaurants are promoting themselves on Zomato's feed in order to acquire exposure and attention from a huge number of Zomato subscribers and customers.

Zomato charges restaurants a commission basedon the number of orders placed through its food delivery service. The company makes money by charging restaurants a commission for each delivery, which is split between the delivery partners and the company. Due to the high levelof competition and the need to offer large discounts, internet meal delivery represents a

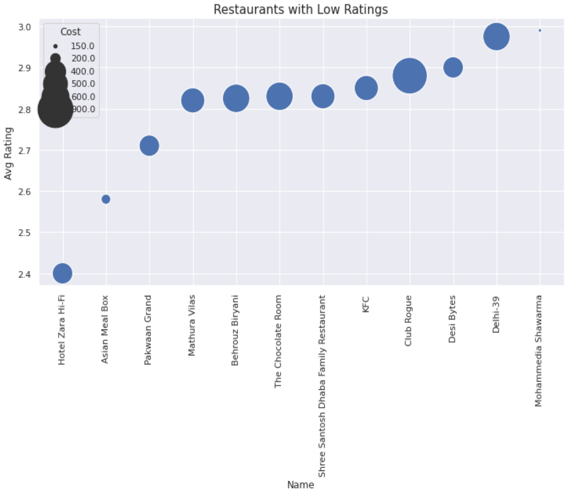
A small fraction of total revenue compared to other revenue streams.

Comparison

The information we have includes the pricing per person, the cuisines available at the restaurant, and the restaurant's average rating. Zomato will have an issue if a restaurant has a poor rating, a high per-person cost, and a limited selection of popular cuisines. Negative reviews are an intangible cost to the business, and as a result, the business will begin to lose everyday application users. The app's users are a valuable asset to the company; because of their enormous viewership, Zomato receives advertising from various restaurants.

Overall, it's critical to identify which restaurants Zomato has to improve in order to improve its overall customer experience, and if improvement tactics fail, they must delist those restaurants.

Here's a scatter plot of the restaurants having thelowest Average Rating according to their per-person Cost.



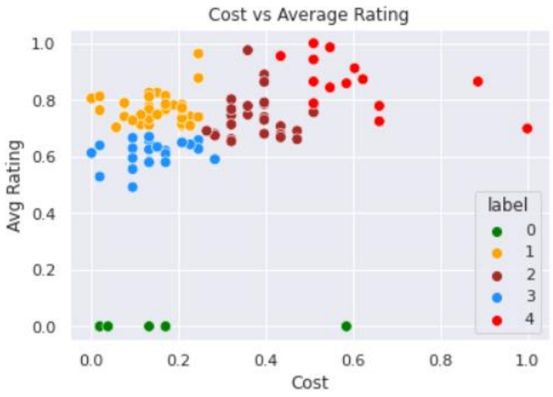
Mohammedia Shawarma has the highest rating among these restaurants and the lowest price, hence it seems profitable enough but some restaurants like Club Rouge have low rating yet high per-person dining cost, this will not generatesignificant revenue and needs improvement.

Restaurant Clustering

Approach 1 Here's a scatter plot of the restaurantclusters formed by K Means Clustering on the basis of just two input variables Cost and AverageRating.

The clusters are fairly distinct from one another. Because there were just two input variables, they were easy to separate and interpret.

- Restaurants with the label 0 were in the names dataset but were not reviewed.
- Restaurants with favorable reviews and inexpensive prices are labeled as label 1.
- Label 2 restaurants are fine dining establishments with good reviews and reasonable prices.
- Restaurants in the Label 3 category are modest eateries with low prices and average reviews.
- Label 4 restaurants are those that are both pricey and have above-average reviews.



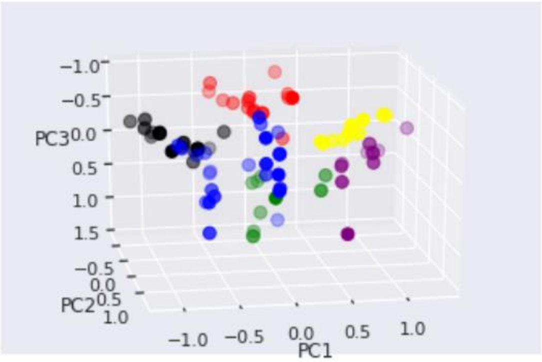
Approach 2 Here's a 3D scatter plot of restaurants clustered on the basis of three principal components.

The data points inside the clusters shared a lot of commonalities.

- Cluster 0 - The eateries in this cluster primarily serve continental and fast meals. The average rating is 3.42, and the average cost is 942 INR, including a2500 INR outlier and a 600 INR mediancost. This means that the eateries in this cluster, with the exception of one, are allrather inexpensive.
- Cluster 1 - The restaurants in Cluster 1 specialize in North Indian cuisine, as

well as other complementing cuisines. The average cost is 823 INR and theaverage rating is 3.63. The prices of these restaurants are slightly higher thanthose in cluster 0.

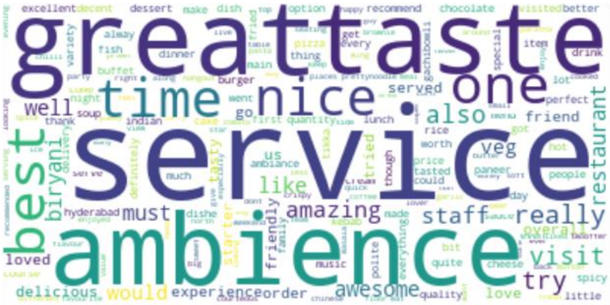
- Cluster 2 - Restaurants in Cluster 2 servea variety of popular cuisines, including North Indian, Chinese, and complimentary. The average rating is 3.77, which is higher than the other two clusters, and the average price is 1331 Indian rupees. These establishments arefine dining restaurants.
- Cluster 3 - The restaurants in this clusterserve a variety of foreign cuisines, including Chinese, Thai, Asian, and seafood, among others. The average ratingis 3.18, owing to the fact that these cuisines aren't particularly popular in Hyderabad, and the average cost is 890INR.
- Cluster 4 - Cluster 4 consists primarily ofsmall eateries, bakeries, and cafes. The average cost is 406 INR and the averagerating is 3.14.
- Cluster 5 - Popular cuisines such as NorthIndian, Chinese, and notably Hyderabadi are accessible at restaurants in cluster 5. The average cost is 674 INR, and the average rating is 3.24. These are casual dining establishments with lower prices and ratings per person than cluster 2.



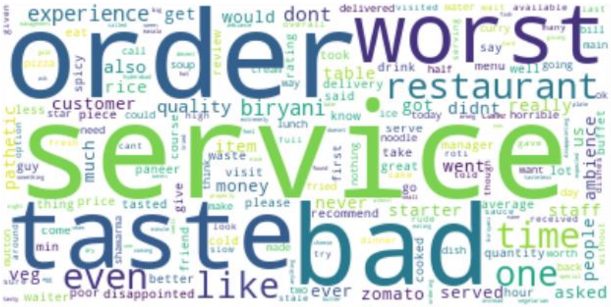
Sentiment Analysis

Sentiment analysis is a machine learning technology that looks for polarity in texts, rangingfrom positive to negative. Machine learning tools learn how to detect sentiment without human input by training them with samples of emotions in text. Sentiment analysis models can be trained to understand things like context, sarcasm, and misapplied words in addition to simple meanings. To command and train machines to perform sentiment analysis, a variety of techniques and complicated algorithms are used.

Positive Word Cloud



Negative Word Cloud

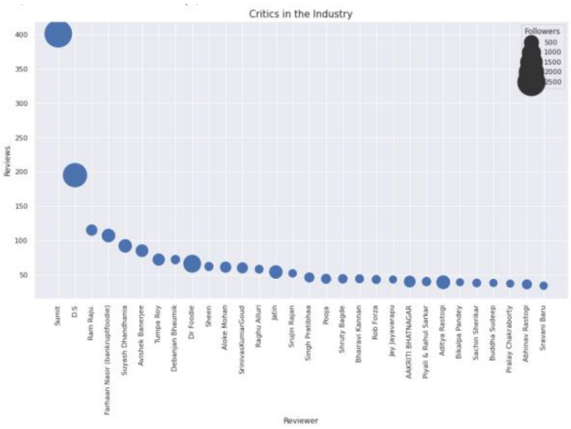


Critics in the Industry

Customers have all the power they need to build or break a firm in today's Internet-driven, social-media world. If clients have a good experience, they tell their friends, family, andacquaintances about it, which leads to new business. All of this word-of-mouth marketing isfree, and when it's posted on a public platform, it's shared with anybody who uses that platform.Customers will complain if you are unable to provide a pleasant client experience for any reason. To decrease the bad marketing impact, Zomato, like any other business, needs to focus on the criticism, particularly with those were genuinely unfavorable but were classifiedas positive, resulting in the loss of a complaint to address.

Reviewers who have been following the most. In order to grow a loyal consumer base, every industry must fight the complaints and critiques itreceives.

Here, an attempt has been made to group consumers with a large number of followers whohave left more reviews with consistently low ratings in order to identify the top critics and the areas that need to be improved.



TF-IDF

Term Frequency Inverse Document Frequency of records is abbreviated as TF-IDF. It's the process of determining how relevant a word in a series or corpus is to a text. The meaning of a word grows in proportion to how many times it appears in the text, but this is offset by the corpus's word frequency (data-set). It's frequently used in information retrieval, text mining, and user modeling searches as a weighting factor.

Model Evaluation

Here are the results for the two models trained by inputting the review text, Logistic Regression, and Random Forest.

In the business challenge, correctly anticipating negative sentiments is critical, but it is even more critical for the models to limit the number of false positives. False positives suggest that the reviews Regression, but having a higher number of false positives. This suggests that Logistic Regressionis penalizing False Positives more aggressively, which is exactly what we want.

Machine Learning Model –

Machine learning is the scientific study of algorithmsand statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference

Building a model by learning the patterns of historicaldata with some relationship between data to make
a data-driven prediction.

Types of Machine Learning

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Unsupervised learning

unsupervised learning is a machine learning technique in which models are not supervised using training dataset. Instead, models itself find the hidden patterns and insights from the given data. It can be compared to learning which takes place in the human brain while learning new things.

Machine Learning models can be understood as a program that has been trained to find patterns within new data and make predictions. These models are represented as a mathematical function that takes requests in the form of input data, makes predictions on input data, and then provides an output in response. First, these models are trained over a set of data, and then they are provided an algorithm to reason over data, extract the pattern from feed data and learn from those data. Once these models get trained, they can be used to predictthe unseen dataset.

Parameters that we need to check to evaluate the machine learning model that we are going to use. These parameters tell us about the model accuracyon training data, but it is also important to get a genuine and approximate result on unseen data otherwise Model is of no use.

Hyperparameter Tunning:

Hyperparameters are sets of information that are used to control the way of learning an algorithm. Their definitions of impact parameters of the models, seen as a way of learning, change from the new hyperparameters. This set of values affects the performance, stability, and interpretation of a model. Each algorithm requires a specific hyperparameters grid that can be adjusted according to the business problem. Hyperparameters alter the way a model learns to trigger this training algorithm after parameters to generate outputs.
We used Grid Search CV for hyperparameter tuning. This also results in cross-validation and in our case we divided the dataset into different folds.
Grid Search CV-Grid:
Search combines a selection of hyperparameters established by the scientist and runs through all of them to evaluate the model's performance. Its advantage is that it is a simple technique that will go through all the programmed combinations. The biggest disadvantage is that it traverses a specific region of the parameter space and cannot understand which movement or which region of the space is important to optimize the model.

Evaluation Metrics:

There are several model evaluation metricsto choose from but since our dataset was highly imbalanced, it is critical to understand which metric should be evaluated to understand the model performance.

- **Accuracy**- Accuracy simply measures how oftenthe classifier correctly predicts. We can define accuracy as the ratio of the number ofcorrect predictions and the total number of predictions. Accuracy is useful when the target class is well balanced but is not a good choice for the unbalanced classes, because if themodel poorly predicts every observation of the majority class, we are going to get pretty high accuracy.
- **Confusion Matrix** - It is a performance measurement criteria forthe machine learning classification problems where we get a table witha combination of predicted and actual values.
- **Precision** - Precision for a label is defined as the number of true positives divided by the number ofpredicted positives.
- **Recall** - Recall for a label is defined as the number of true positives divided by the total number of actualpositives. Recall explains how manyof the actual positive cases we were able to predict correctly with our model.
- **F1 Score** - It's the harmonic mean of Precision and Recall. It is maximum when Precision is equal to Recall.
- **AUC ROC** - The Area Under the Curve (AUC)is the measure of the ability of a classifier to distinguish between classes. When AUC is 0.5,the classifier is not able to distinguish between the classes and when it's closer to 1, the better it becomes at distinguishing them.

Algorithms and Methods

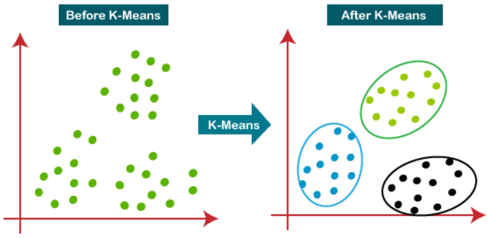
There are two datasets to work with in this problem statement:

- Zomato Restaurant Names and Metadata
- Zomato Restaurant Reviews

The project is divided into two sections, the first one being the clustering of restaurants. Clusteringis the process of separating a population or set of data points into several groups so that data pointsin the same group are more similar than data points in other groups. To put it another way, the goal is to separate groups with similar characteristics and assign them to clusters.

K Means Clustering:

K-Means Clustering is an unsupervised learningalgorithm used in machine learning and data science to handle clustering problems. It's an iterative technique that splits an unlabeled dataset into k clusters, with each dataset belonging to onlyone group with similar qualities. It's a centroid-based approach, which means that each cluster has its own centroid. The main goal of thistechnique is to reduce the sum of distances between data points and the clusters to that they belong to. The technique takes an unlabeled dataset as input, separates it into a k-number of clusters, and continues the procedure until no better clusters are found. In this algorithm, thevalue of k should be predetermined.



The k-means clustering algorithm primarilyaccomplishes two goals:

- Iteratively determines the optimal valuefor K center points or centroids.
- Each data point is assigned to the k-center that is closest to it. A cluster is formed by data points that are close to a specific k-center.

The K-means clustering algorithm's performance is dependent on the very efficient clusters it creates. However, determining the ideal number of clusters is a difficult process. There are several methods for determining the best number ofclusters, but we will focus on the most appropriate approach for determining the numberof clusters or K value. The procedure is as follows:

Elbow Method

One of the most prominent methods for determining the ideal number of clusters is the Elbow approach. This approach makes use of the WCSS value notion. Within Cluster Sum of Squares (WCSS) is a term that describes the total variations within a cluster. The sharp point of bend or a point of the plot looks like an arm, thenthat point is considered as the best value of K.

The Curse of Dimensionality

When we have too many features, it becomes more difficult to cluster observations having too many dimensions causes every observation in thedataset to appear equidistant from every other observation. This is a serious concern since clustering requires a distance measure like Euclidean distance to estimate the similarity between observations. If all of the distances areroughly identical, all of the observations appearto be similarly similar (and equally dissimilar), and no meaningful clusters can be established.

Principal Component Analysis

Principal Component Analysis, or PCA, is a dimensionality-reduction approach for reducing the dimensionality of large data sets by transforming a large collection of variables into asmaller one that retains the majority of the information in the large set.

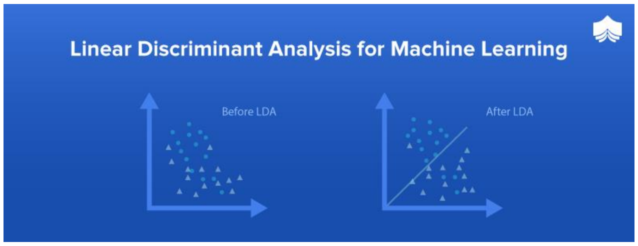
Naturally, reducing the number of variables in a data set reduces accuracy; nevertheless, the idea of dimensionality reduction is to exchange some accuracy for simplicity. Because smaller data sets are easier to study and interpret, and because machine learning techniques can analyze data more easily and quickly without having to deal with unnecessary factors.

PCA's basic concept is to reduce the number of variables in a data collection while retaining asmuch information as feasible. Principal components are new variables that are created by combining or mixing the basic variables in a linear way. The new variables (i.e., principle components) are uncorrelated as a resultof these combinations, and the majority of the information from the initial variables is squeezed or compressed into the first components. For instance, 10-dimensional data gives you ten principal components, but PCA seeks to place as much information as possible in the first component, then as little information as possiblein the second, and so on.

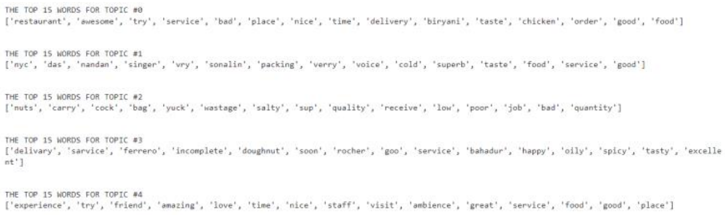
Sentiment Analysis, the second half of the project, is carried out using supervised machinelearning methods like Logistic Regression, Multinomial Naïve Bayes, Decision Tree, Random Forest, XG Boost, and Light GBM Classification.

LDA:

Linear Discriminant Analysis or LDA is a dimensionality reduction technique. It is used as a pre-processing step in Machine Learning and applications of pattern classification. The goal of LDA is to project the features in higher dimensional space onto a lower-dimensional space in order to avoid the curse of dimensionality and also reduce resources and dimensional costs. The original technique was developed in the year 1936 by Ronald A. Fisher and was named Linear Discriminant or Fisher's Discriminant Analysis. The original Linear Discriminant was described as a two-class technique. The multi-class version was later generalized by C.R Rao as a Multiple Discriminant Analysis. They are all simply referred to as Linear Discriminant Analysis. LDA is a supervised classification technique that is considered a part of crafting competitive machine learning models. This category of dimensionality reduction is used in areas like image recognition and predictive analysis in marketing.



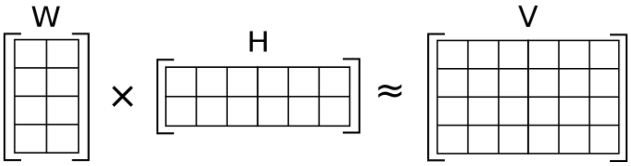
LDA top 15 words of each topic



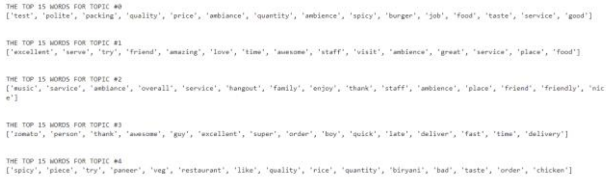
```
THE TOP 15 WORDS FOR TOPIC #0
['restaurant', 'awesome', 'try', 'service', 'bad', 'place', 'nice', 'time', 'delivery', 'biryani', 'taste', 'chicken', 'order', 'good', 'food']

THE TOP 15 WORDS FOR TOPIC #1
['nyc', 'das', 'nandan', 'singer', 'vry', 'sonalin', 'packing', 'verry', 'voice', 'cold', 'superb', 'taste', 'food', 'service', 'good']

THE TOP 15 WORDS FOR TOPIC #2
['nuts', 'carry', 'cock', 'bag', 'yuck', 'wastage', 'salty', 'sup', 'quality', 'receive', 'low', 'poor', 'job', 'bad', 'quantity']

THE TOP 15 WORDS FOR TOPIC #3
['delivary', 'sarvice', 'ferrero', 'incomplete', 'doughnut', 'soon', 'rocher', 'goo', 'service', 'bahadur', 'happy', 'oily', 'spicy', 'tasty', 'excelle
nt']

THE TOP 15 WORDS FOR TOPIC #4
['experience', 'try', 'friend', 'amazing', 'love', 'time', 'nice', 'staff', 'visit', 'ambience', 'great', 'service', 'food', 'good', 'place']
```

Non-negative matrix Factorization:

NMF stands for non-negative matrix factorization, a technique for obtaining a low-rank representation of matrices with non-negative or positive elements. Such matrices are common in a variety of applications of interest. For example, images are nothing but matrices of positive integer numbers representing pixel intensities. In information retrieval and text mining, we rely on term-document matrices for representing document collections. In recommendation systems, we have utility matrices showing customers’ preferences for items.



NMF Top 15 words of each Topic



```
THE TOP 15 WORDS FOR TOPIC #0
['test', 'polite', 'packing', 'quality', 'price', 'ambiance', 'quantity', 'ambience', 'spicy', 'burger', 'job', 'food', 'taste', 'service', 'good']

THE TOP 15 WORDS FOR TOPIC #1
['excellent', 'serve', 'try', 'friend', 'amazing', 'love', 'time', 'awesome', 'staff', 'visit', 'ambience', 'great', 'service', 'place', 'food']

THE TOP 15 WORDS FOR TOPIC #2
['music', 'sarvice', 'ambiance', 'overall', 'service', 'hangout', 'family', 'enjoy', 'thank', 'staff', 'ambience', 'place', 'friend', 'friendly', 'nic
e']

THE TOP 15 WORDS FOR TOPIC #3
['zomato', 'person', 'thank', 'awesome', 'guy', 'excellent', 'super', 'order', 'boy', 'quick', 'late', 'deliver', 'fast', 'time', 'delivery']

THE TOP 15 WORDS FOR TOPIC #4
['spicy', 'piece', 'try', 'paneer', 'veg', 'restaurant', 'like', 'quality', 'rice', 'quantity', 'biryani', 'bad', 'taste', 'order', 'chicken']
```

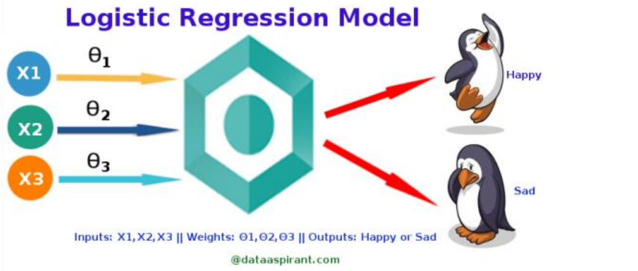
Modeling

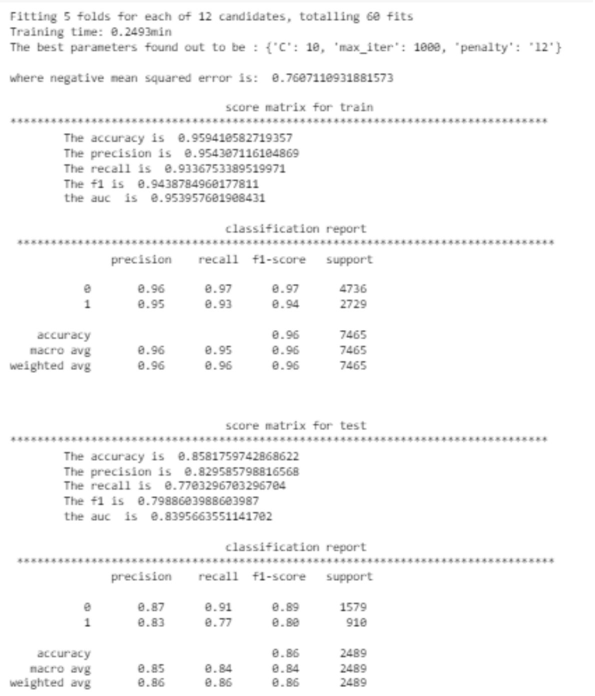
Logistic Regression

Logistic regression is a statistical analytic approach for predicting a binary outcome, such asyes or no. A logistic regression model analyses the relationship between one or more existing independent variables to predict a dependent datavariable. Except for how they are employed, Logistic Regression is very similar to Linear Regression.

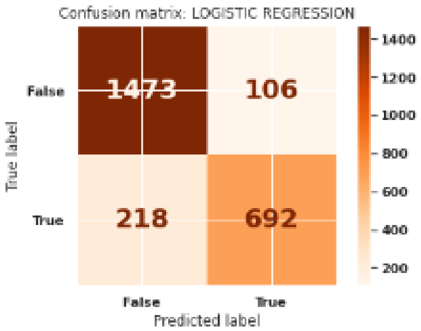
Instead of fitting a regression line, we fit a "S" shaped logistic function in logistic regression, which predicts two maximum values (0 or 1). Because of its capacity to generate probabilitiesand classify fresh data, Logistic Regression is akey machine learning technique.

The sigmoid function is a mathematical functionfor converting anticipated values into probabilities.
It maps any real value into another value within arange of 0 and 1.
The logistic regression's value must be between 0and 1, and it cannot exceed this limit, resulting ina "S" curve. The Sigmoid function, often known as the logistic function, is the S-form curve.
The concept of the threshold value is used in logistic regression to describe the probability of
either 0 or 1. Values over the threshold value tendto be 1, while those below the threshold value tend to be 0.


Logistic Regression Model
X1
X2
X3
θ1
θ2
θ3
Happy
Sad
Inputs: X1,X2,X3 || Weights: Θ1,Θ2,Θ3 || Outputs: Happy or Sad
@dataaspirant.com



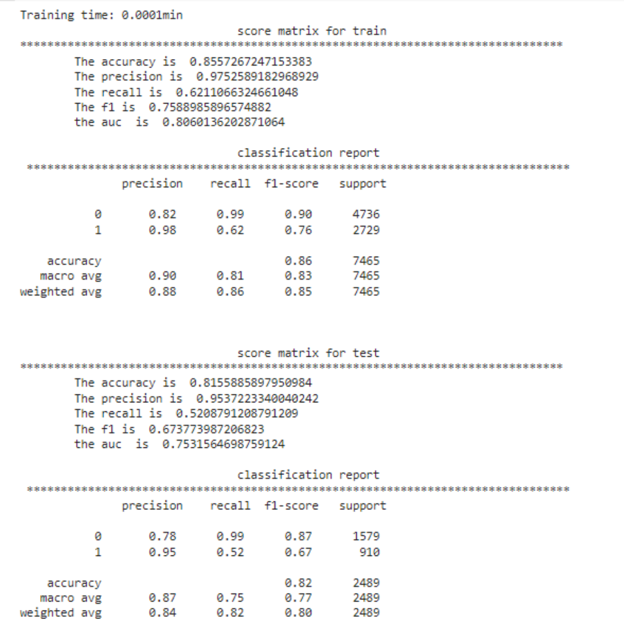
```
Fitting 5 folds for each of 12 candidates, totalling 60 fits
Training time: 0.2493min
The best parameters found out to be : {'C': 10, 'max_iter': 1000, 'penalty': 'l2'}

where negative mean squared error is:  0.7607110931881573

                            score matrix for train
********************************************************************************
        The accuracy is  0.959410582719357
        The precision is  0.954307116104869
        The recall is  0.9336753389519971
        The f1 is  0.9438784960177811
        the auc  is  0.953957601908431

                            classification report
 ********************************************************************************
              precision    recall  f1-score   support

           0       0.96      0.97      0.97      4736
           1       0.95      0.93      0.94      2729

    accuracy                           0.96      7465
   macro avg       0.96      0.95      0.96      7465
weighted avg       0.96      0.96      0.96      7465


                            score matrix for test
********************************************************************************
        The accuracy is  0.8581759742868622
        The precision is  0.829585798816568
        The recall is  0.7703296703296704
        The f1 is  0.7988603988603987
        the auc  is  0.8395663551141702

                            classification report
 ********************************************************************************
              precision    recall  f1-score   support

           0       0.87      0.91      0.89      1579
           1       0.83      0.77      0.80       910

    accuracy                           0.86      2489
   macro avg       0.85      0.84      0.84      2489
weighted avg       0.86      0.86      0.86      2489
```



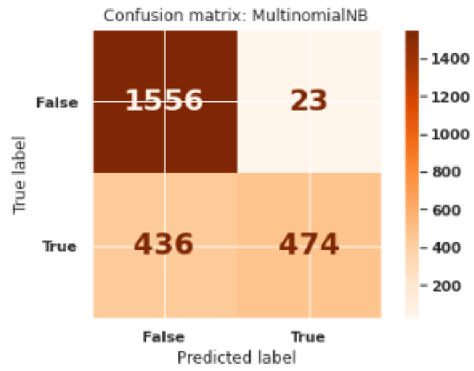
Multinomial NB:

The multinomial Naive Bayes algorithm is a probabilistic learning method that is mostly used in Natural Language Processing (NLP). The algorithm is based on the Bayes theorem and predicts the tag of a text such as a piece of email or newspaper article. It calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output.

A naive Bayes classifier is a collection of many algorithms where all the algorithms share one common principle, and that is each feature being classified is not related to any other feature. The presence or absence of a feature does not affect the presence or absence of the other feature.



```
Training time: 0.0001min
                                score matrix for train
********************************************************************************
        The accuracy is  0.8557267247153383
        The precision is  0.9752589182968929
        The recall is  0.6211066324661048
        The f1 is  0.7588985896574882
        the auc  is  0.8060136202871064

                                classification report
 ********************************************************************************
              precision    recall  f1-score   support

           0       0.82      0.99      0.90      4736
           1       0.98      0.62      0.76      2729

    accuracy                           0.86      7465
   macro avg       0.90      0.81      0.83      7465
weighted avg       0.88      0.86      0.85      7465



                                score matrix for test
********************************************************************************
        The accuracy is  0.8155885897950984
        The precision is  0.9537223340040242
        The recall is  0.5208791208791209
        The f1 is  0.673773987206823
        the auc  is  0.7531564698759124

                                classification report
 ********************************************************************************
              precision    recall  f1-score   support

           0       0.78      0.99      0.87      1579
           1       0.95      0.52      0.67       910

    accuracy                           0.82      2489
   macro avg       0.87      0.75      0.77      2489
weighted avg       0.84      0.82      0.80      2489
```

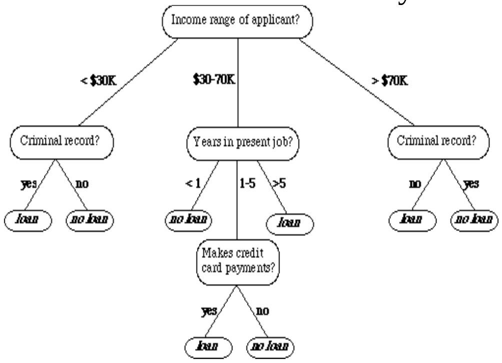


Decision Tree:

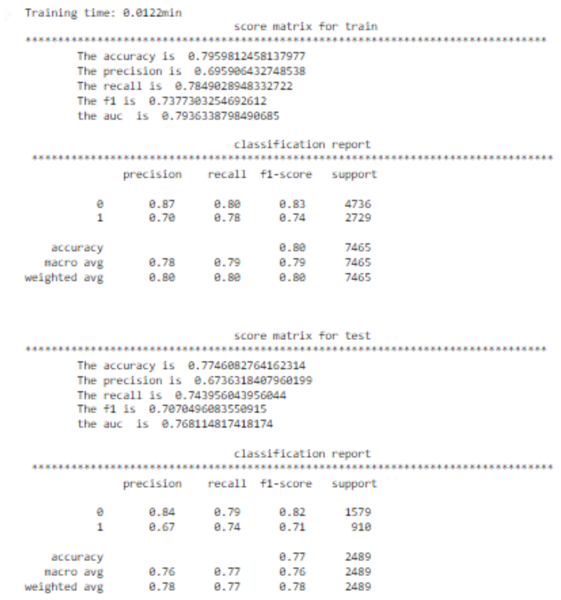
A Decision tree is a type of supervised learning algorithm that can be used in classification as well as regressor problems. The input to a decision tree can be both continuous as well as categorical. The decision tree works on an if-then statement. A decision tree tries to solve a problem by using tree representation (Node and Leaf)

- Assumptions while creating a decision tree: Initially all the training set is considered as a root
- Feature values are preferred to be categorical, if continuous then they are discretized
- Records are distributed recursively based on attribute values
- Which attributes are considered to be in the root node or internal node is done by using a statistical approach.

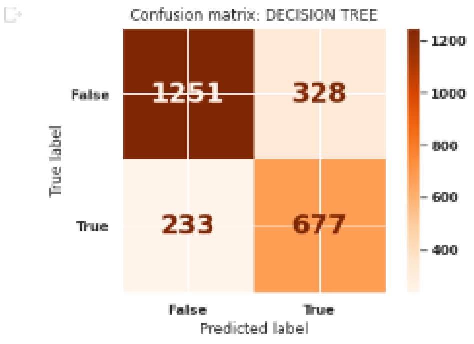
In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches



It's better to have a much more generalized model for future data points. Businesses prefer the model to be interpretable to understand the patterns and strategize accordingly unlike any scientific. The facility where the results matter much more than interpretability. If interpretability is important then sticking with tree-based algorithms when most of the features are categorical; is beneficial and using tuned Hyperparameters to grow the tree deep enough without overfitting.



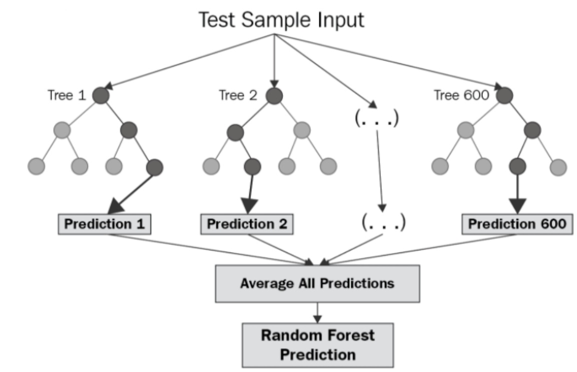
```
Training time: 0.0122min
                                    score matrix for train
********************************************************************************
        The accuracy is  0.7959812458137977
        The precision is  0.695906432748538
        The recall is  0.7849028948332722
        The f1 is  0.7377303254692612
        the auc  is  0.7936338798490685

                                    classification report
 ********************************************************************************
              precision    recall  f1-score   support

           0       0.87      0.80      0.83      4736
           1       0.70      0.78      0.74      2729

    accuracy                           0.80      7465
   macro avg       0.78      0.79      0.79      7465
weighted avg       0.80      0.80      0.80      7465


                                    score matrix for test
********************************************************************************
        The accuracy is  0.7746082764162314
        The precision is  0.6736318407960199
        The recall is  0.743956043956044
        The f1 is  0.7070496083550915
        the auc  is  0.768114817418174

                                    classification report
 ********************************************************************************
              precision    recall  f1-score   support

           0       0.84      0.79      0.82      1579
           1       0.67      0.74      0.71       910

    accuracy                           0.77      2489
   macro avg       0.76      0.77      0.76      2489
weighted avg       0.78      0.77      0.78      2489
```

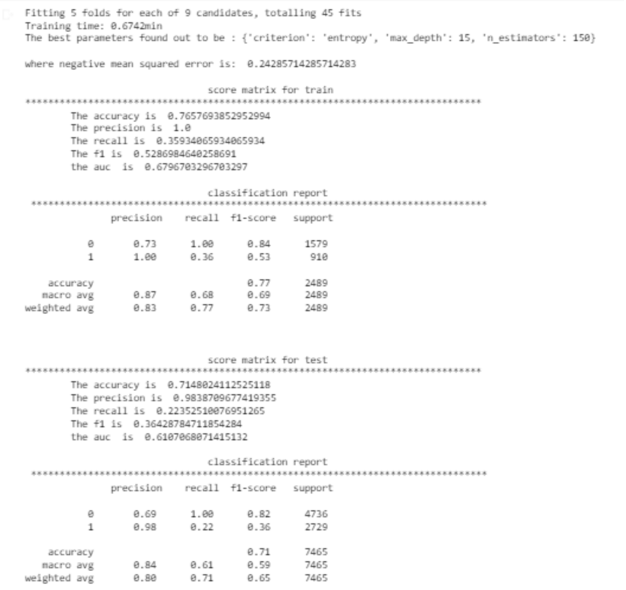


Random Forest

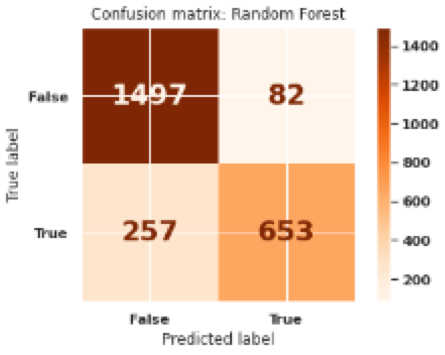
Random forest is a supervised machine learningalgorithm that is commonly used to solve classification and regression problems. It creates decision trees from various samples, and uses themajority vote for classification and the average for regression.

One of the most essential characteristics of the Random Forest Algorithm is that it can handle data sets with both continuous and categorical variables, as in regression and classification. Forclassification challenges, it produces better results.





```
Fitting 5 folds for each of 9 candidates, totalling 45 fits
Training time: 0.6742min
The best parameters found out to be : {'criterion': 'entropy', 'max_depth': 15, 'n_estimators': 150}

where negative mean squared error is:  0.24285714285714283

                                    score matrix for train
********************************************************************************
        The accuracy is  0.7657693852952994
        The precision is  1.0
        The recall is  0.35934065934065934
        The f1 is  0.5286984640258691
        the auc  is  0.6796703296703297

                                    classification report
 ********************************************************************************
              precision    recall  f1-score   support

           0       0.73      1.00      0.84      1579
           1       1.00      0.36      0.53       910

    accuracy                           0.77      2489
   macro avg       0.87      0.68      0.69      2489
weighted avg       0.83      0.77      0.73      2489


                                    score matrix for test
********************************************************************************
        The accuracy is  0.7148024112525118
        The precision is  0.9838709677419355
        The recall is  0.22352510076951265
        The f1 is  0.36428784711854284
        the auc  is  0.6107068071415132

                                    classification report
 ********************************************************************************
              precision    recall  f1-score   support

           0       0.69      1.00      0.82      4736
           1       0.98      0.22      0.36      2729

    accuracy                           0.71      7465
   macro avg       0.84      0.61      0.59      7465
weighted avg       0.80      0.71      0.65      7465
```



XGBoost Classification:

XGBoost is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data.XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

The implementation of the algorithm was engineered for the efficiency of computing time and memory resources. A design goal was to make the best use of available resources to train the model. Some key algorithm implementation features include:

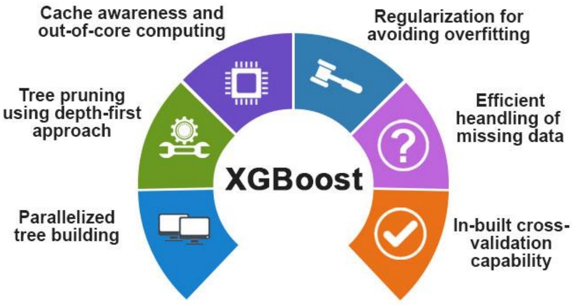
- **Sparse Aware** implementation with automatic handling of missing data values.
- **Block Structure** to support the parallelization of tree construction.
- **Continued Training** so that you can further boost an already fitted model on new data.

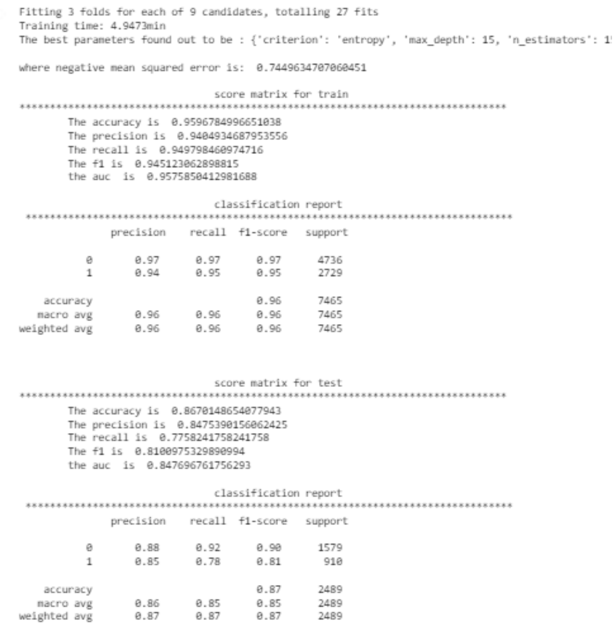
XGBoost is free open source software available for use under the permissive Apache-2 license.

Why Use XGBoost?

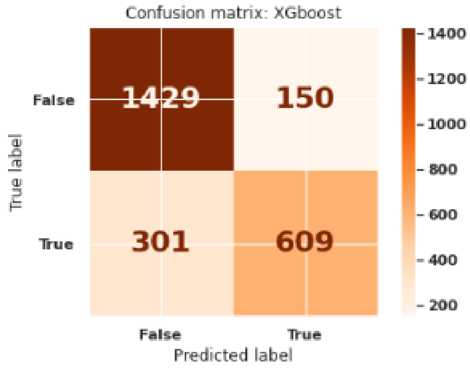
The two reasons to use XGBoost are also the two goals of the project:

1. Execution Speed.
2. Model Performance.


Cache awareness and out-of-core computing
Regularization for avoiding overfitting
Tree pruning using depth-first approach
Efficient heandling of missing data
XGBoost
Parallelized tree building
In-built cross-validation capability



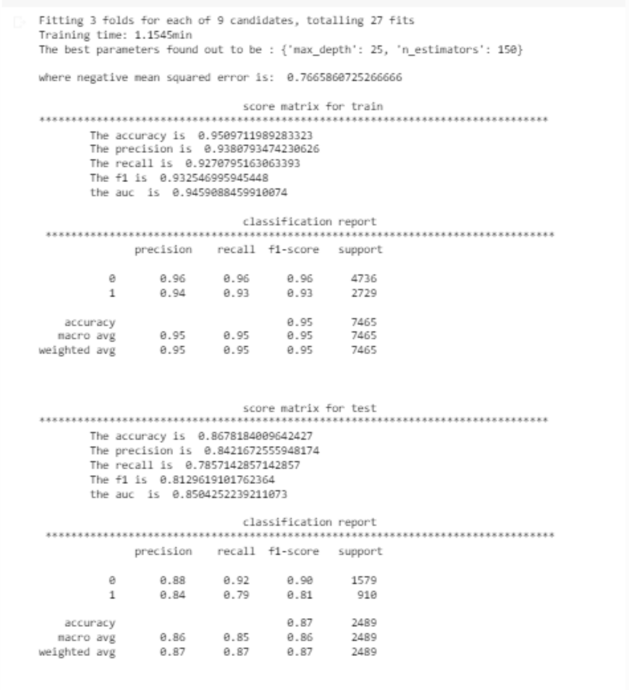
```
Fitting 3 folds for each of 9 candidates, totalling 27 fits
Training time: 4.9473min
The best parameters found out to be : {'criterion': 'entropy', 'max_depth': 15, 'n_estimators': 1

where negative mean squared error is:  0.7449634707060451

                                score matrix for train
********************************************************************************
        The accuracy is  0.9596784996651038
        The precision is  0.9404934687953556
        The recall is  0.949798460974716
        The f1 is  0.945123062898815
        the auc  is  0.9575850412981688

                                classification report
 ********************************************************************************
               precision    recall  f1-score   support

           0       0.97      0.97      0.97      4736
           1       0.94      0.95      0.95      2729

    accuracy                           0.96      7465
   macro avg       0.96      0.96      0.96      7465
weighted avg       0.96      0.96      0.96      7465


                                score matrix for test
********************************************************************************
        The accuracy is  0.8670148654077943
        The precision is  0.8475390156062425
        The recall is  0.7758241758241758
        The f1 is  0.8100975329890994
        the auc  is  0.847696761756293

                                classification report
 ********************************************************************************
               precision    recall  f1-score   support

           0       0.88      0.92      0.90      1579
           1       0.85      0.78      0.81       910

    accuracy                           0.87      2489
   macro avg       0.86      0.85      0.85      2489
weighted avg       0.87      0.87      0.87      2489
```



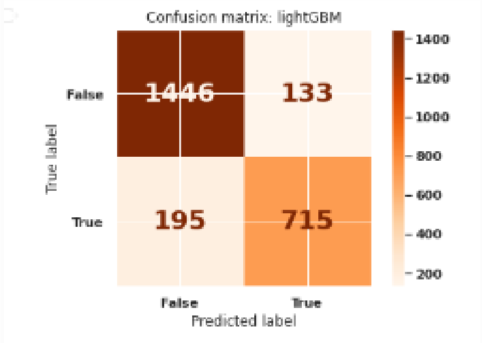
LightGBM Classification:

LightGBM is a gradient boosting framework that uses tree-based learning algorithms. It is designed to be distributed and efficient with the following advantages:

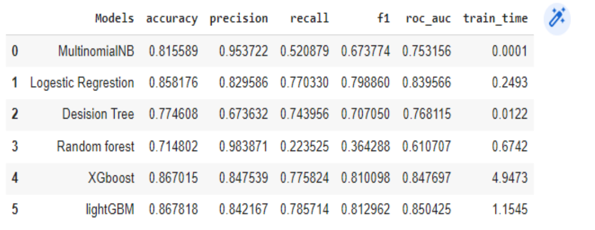
- Faster training speed and higher efficiency.
- Lower memory usage.
- Better accuracy.
- Support of parallel, distributed, and GPU learning.
- Capable of handling large-scale data.



```
Fitting 3 folds for each of 9 candidates, totalling 27 fits
Training time: 1.1545min
The best parameters found out to be : {'max_depth': 25, 'n_estimators': 150}

where negative mean squared error is:  0.7665860725266666

                              score matrix for train
********************************************************************************
       The accuracy is  0.9509711989283323
       The precision is  0.9380793474230626
       The recall is  0.9270795163063393
       The f1 is  0.932546995945448
       the auc  is  0.9459088459910074

                              classification report
 ********************************************************************************
              precision    recall  f1-score   support

           0       0.96      0.96      0.96      4736
           1       0.94      0.93      0.93      2729

    accuracy                           0.95      7465
   macro avg       0.95      0.95      0.95      7465
weighted avg       0.95      0.95      0.95      7465


                              score matrix for test
********************************************************************************
       The accuracy is  0.8678184009642427
       The precision is  0.8421672555948174
       The recall is  0.7857142857142857
       The f1 is  0.8129619101762364
       the auc  is  0.8504252239211073

                              classification report
 ********************************************************************************
              precision    recall  f1-score   support

           0       0.88      0.92      0.90      1579
           1       0.84      0.79      0.81       910

    accuracy                           0.87      2489
   macro avg       0.86      0.85      0.86      2489
weighted avg       0.87      0.87      0.87      2489
```



Evaluation Matrix:



	Models	accuracy	precision	recall	f1	roc_auc	train_time
0	MultinomialNB	0.815589	0.953722	0.520879	0.673774	0.753156	0.0001
1	Logestic Regrestion	0.858176	0.829586	0.770330	0.798860	0.839566	0.2493
2	Desision Tree	0.774608	0.673632	0.743956	0.707050	0.768115	0.0122
3	Random forest	0.714802	0.983871	0.223525	0.364288	0.610707	0.6742
4	XGboost	0.867015	0.847539	0.775824	0.810098	0.847697	4.9473
5	lightGBM	0.867818	0.842167	0.785714	0.812962	0.850425	1.1545

In a business problem, predicting the negative sentiments correctly is really important but is more important for the models to reduce the number of false positives.

False positives indicate that the reviews were actually negative but they were categorized as positive and this will lead to missing a complaint to work on.

Even though the number of false negatives is higher in the case of Logistic Regression than Random Forest, it is performing better in terms of reducing False positives. This indicates that Logistic Regression is penalizing False positives more just as we want.

Conclusion and Recommendations:

Conclusion:

Clustering is the process of identifying unique groupings or "clusters" within a data set. The program constructs groups using a machine language algorithm, and items in a comparable group will have similar features in general.
One of the challenges that organizations have is figuring out how to arrange the massive volumes of data accessible into usable structures. Alternatively, divide a large heterogeneous group into smaller homogenous groupings. Cluster analysis is an exploratory data analysis tool that seeks to group things so that the degree of relationship between two objects is greatest if they belong to the same group and minimal if they don't.
This enables businesses to assist their clients in quickly locating the information they require. This analysis included all of the essential subjectsin both the business and technological domains.
Some important insights to draw from the analysis include:

- The best restaurants in Hyderabad are AB's - Absolute Barbecues, B-Dubs, and 3B's - Buddies, Bar & Barbecue.
- The most popular cuisines are the cuisines that most of the restaurants are willing to provide. The most popular cuisines in Hyderabad are North Indian, Chinese, Continental, and Hyderabadi.
- The restaurants in Hyderabadi have a

flexible per person cost of 150 INR to 2800 INR. The cheapest is the food jointcalled Mohammedia Shawarma and the costliest restaurant is Collage - Hyatt Hyderabad Gachibowli.

- Upon conducting a basic cost-benefitanalysis on Zomato with a few assumptions one basis of the little business understanding that could be gathered, it can be concluded that it is important to separate out the restaurants with the lowest rating in order to improveits overall customer experience. These restaurants were small food joints or restaurants with high prices according to the food they were serving. Efforts should be made to advertise more and analyze thereviews, especially for these restaurants, and work on them. MohammediaShawarma seems to be profitable.
- Restaurant Clustering was done in two approaches. First with just two features and then with all of them. K means Clustering worked well in the first approach but as we increase the dimensions, it isn't able to distinguish the clusters hence principal component analysis was done and then clustered into 6 clusters. The similarities in the data points within the clusters were pretty great.
- Critics in the Industry were identified bygrouping the customers with a good number of followers who have given more reviews with constantly low ratings. Sumit, D.S, and Ram Raju are the top three critics.
- Sentiment Analysis was done on the reviews and a model was trained in orderto identify negative and positive sentiments. Even though the number of false negatives is lower in the case of Multinomial NB and Logistic Regression than in Light GBM, it is performing better in terms ofreducing False positives. This indicates that Multinomial NB and Logistic Regression is penalizing False positives more just as we want.

Challenges:

- Because the data was provided in a raw format in string format, the project's main problem was extracting key information from the dataset in numerical form.

Recommendations:

- Negative reviews should be approachedto reach a win-win solution.
- Ratings should be gathered according to categories, such as packing, delivery, taste, quality, amount, and service. This would aid in identifying and addressing lagging fields.

References:

- Machine Learning Mastery
- GeeksforGeeks
- Analytics Vidhya Blogs
- Towards Data Science Blogs
- Built-inin Data Science Blogs
- Scikit- Learn