Zomato Restaurant Clustering and Sentiment Analysis

Anjali Tidke and Shubham Dukare

Data Science Trainee

Almabetter, Bangalore.

Contents

[1. Introduction 2](#_Toc113803709)

[2. Problem Statement 2](#_Toc113803710)

[3. Attribute Information 3](#_Toc113803711)

[4. Data Inspection and pre-processing 3](#_Toc113803712)

[5. Exploratory Data Analysis 4](#_Toc113803713)

[5.1 Let’s see count of cuisine’s offered by restaurant 4](#_Toc113803714)

[5.2 Let’s see Top 10 Most expensive restaurants 5](#_Toc113803715)

[5.3 Let’s see Top 10 Cheapest Restaurants 5](#_Toc113803716)

[5.4 Let’s see restaurant by best price and rating also. 6](#_Toc113803717)

[5.5 Restaurant rating percentage 6](#_Toc113803718)

[6. Clustering 7](#_Toc113803719)

[6.1 Agglomerative Hierarchial Clustering 7](#_Toc113803720)

[6.2 DBSCAN clustering 8](#_Toc113803721)

[6.3 K Means Clustering 8](#_Toc113803722)

[6.4 mini-batch k-means 9](#_Toc113803723)

[6.5 Validation of all above methods 9](#_Toc113803724)

[7. Sentiment Analysis 9](#_Toc113803725)

[7.1 DecisionTreeClassifier 10](#_Toc113803726)

[7.2 Random Forest Classifier 11](#_Toc113803727)

[7.3 K-Nearest Neighbour Classifier 12](#_Toc113803728)

[7.4 Support Vector Machine 13](#_Toc113803729)

[7.5 Logistic Regression 14](#_Toc113803730)

[7.6 Multinomial Naïve Bayes 15](#_Toc113803731)

[7.7 XG Boost Classifier 16](#_Toc113803732)

[7.8 Light Gradient Boosting Machine Classifier 16](#_Toc113803733)

[8. Hyper parameter tuning 17](#_Toc113803734)

[9. Conclusion 19](#_Toc113803735)

[References 20](#_Toc113803736)

# Introduction

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.

We are provided with the two datasets,

1. Zomato restaurant names and metadata
2. Zomato user reviews

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 Data could be used for sentiment analysis. Also, the metadata of reviewers can be used for identifying the critics in the industry.

# Problem Statement

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.

The Project focuses on Customers and Company, We 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 vizualized 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

Data could be used for sentiment analysis. Also, the metadata of reviewers can be used for identifying the critics in the industry.

# Attribute Information

We are provided with the two datasets, The details about the two datasets given below:

1. Restaurant Name and Metadata

1. Name: Name of the Restaurants

2. Links: URL Links of the Restaurants

3. Cost: Per person estimated Cost of the dining

4. Collection: Tagging of Restaurants with respect to Zomato categories

5. Cuisines: Cuisines served by the Restaurants

6. Timings: Timings of the restaurants

1. Zomato user reviews

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

After understanding the dataset, we imported the required libraries.

# Data Inspection and pre-processing

In data inspection, we understood the data from each column in details.

we changed the data type of cost column to integer and observed null values, treated outliers of numerical data type columns accordingly.

In the next step of data cleaning and preparation, We removed stopwords and punctuations from the collections column. Also, we used the Count vectorizer, Steeming and TF-IDF vectorizer methods and processes to prepare the data.

# 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.

## Let’s see count of cuisine’s offered by restaurant

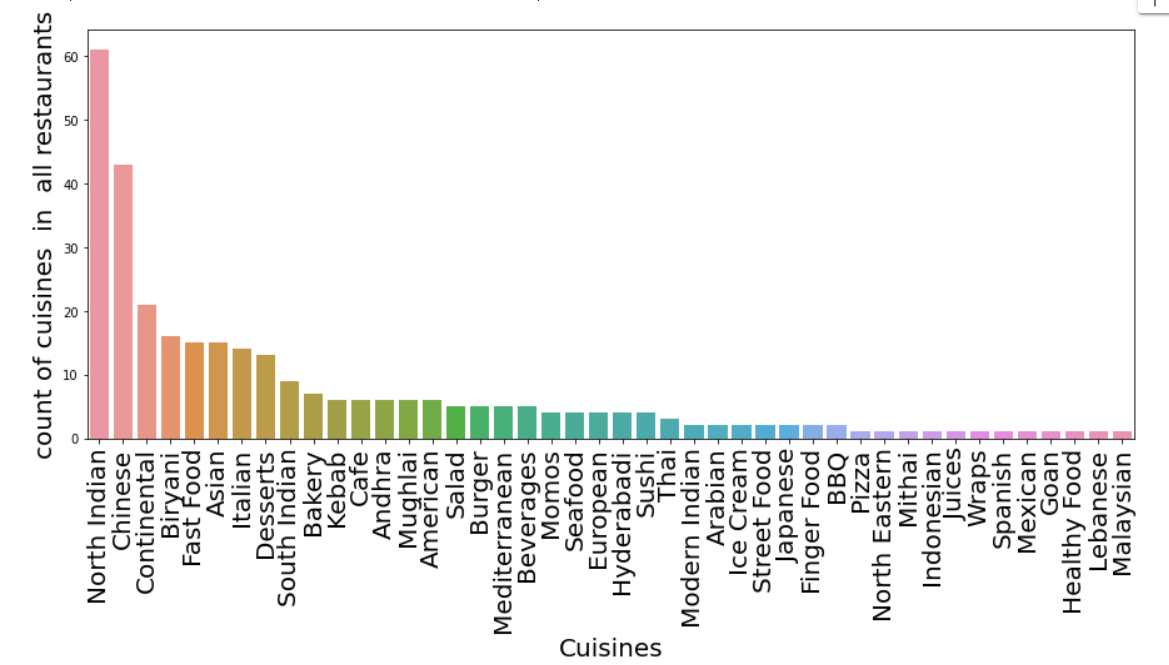


Fig 1. Count of cuisines offered by restaurants

From the above, We can see that north Indian cuisine is available in most of the restaurants followed by Chinese and Continental.

## Let’s see Top 10 Most expensive restaurants

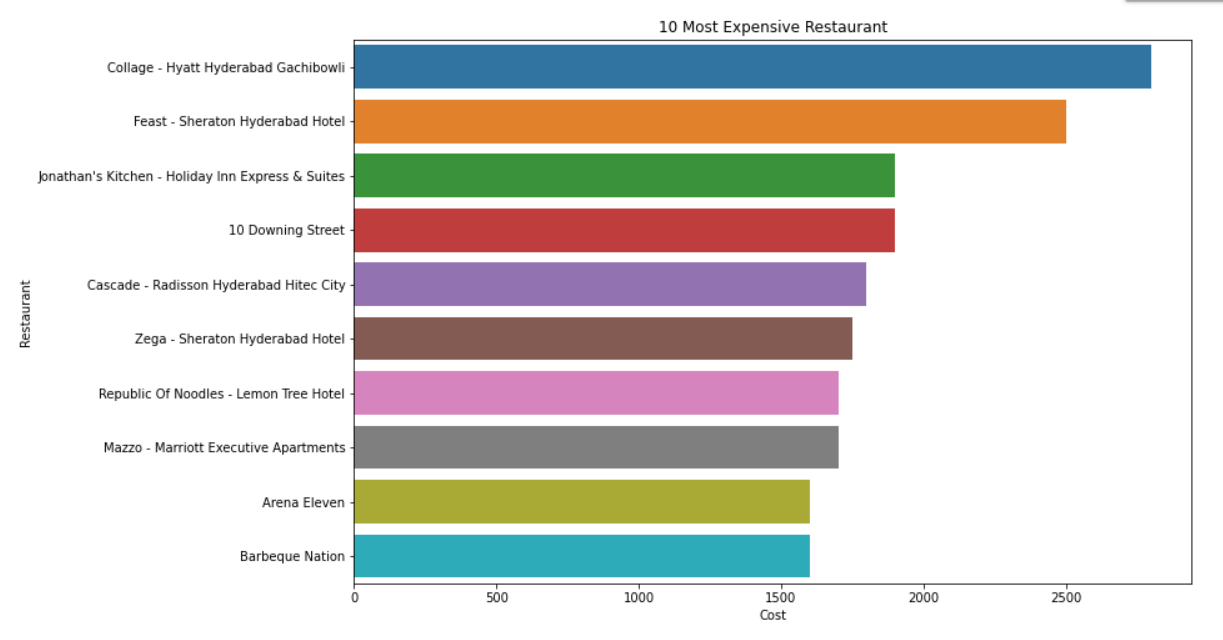


Fig 2: Top 10 most expensive restaurants

The Collage \_Hyatt Hyderabad GachiBowli is the most expensive restaurant in which dishes are more than 2500 cost.

## Let’s see Top 10 Cheapest Restaurants

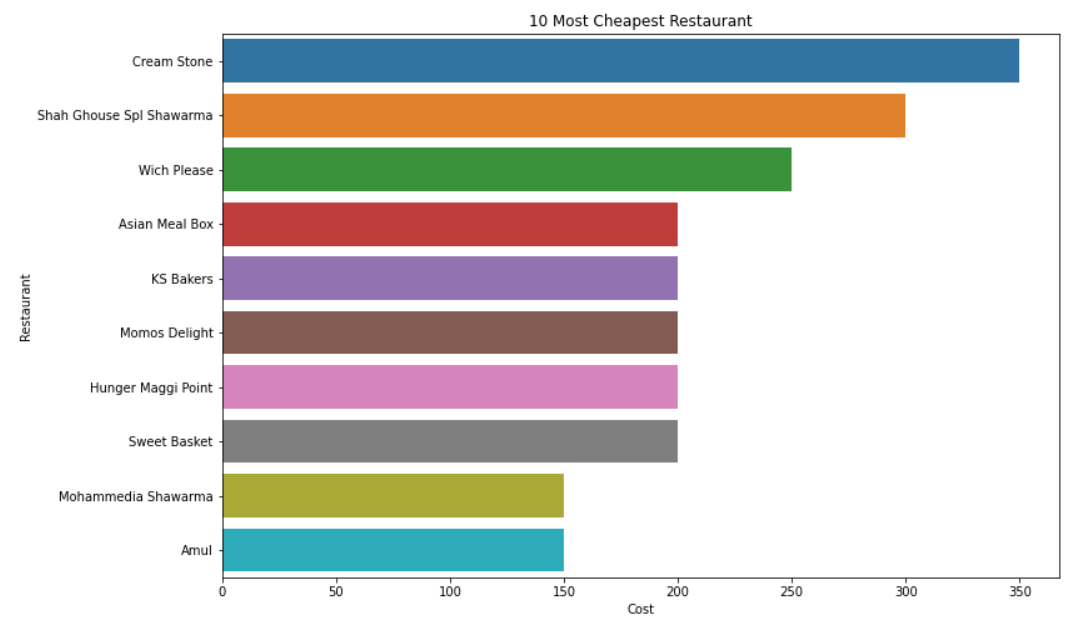


Fig 3: Top 10 cheapest restaurants

We can see Amul is the cheapest restaurant.

1. Now, which restaurant to go?

## Let’s see restaurant by best price and rating also.

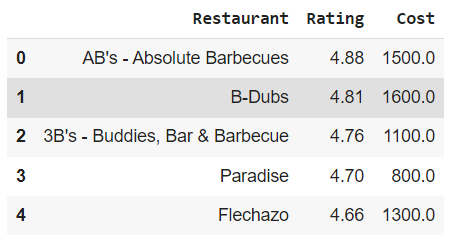


Fig 4: Best restaurants

Observation is that AB's - Absolute Barbecues is the best restaurant with rating and price.

## Restaurant rating percentage

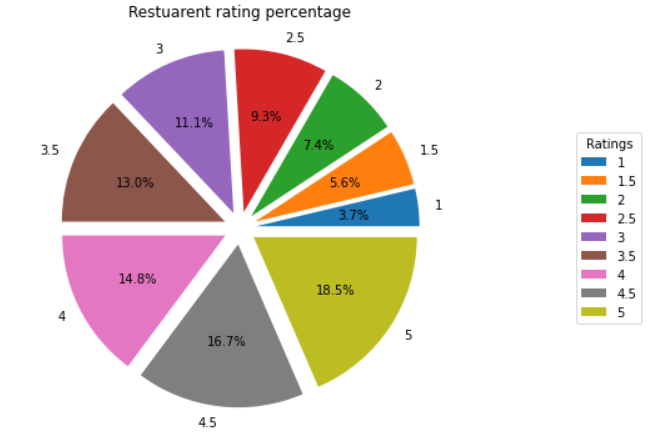


Fig 6: Restaurant rating percentage

We can see that majority of the restuarants has >3 rating.

The next process is to prepare the data to train the models. It includes separating columsn if needed, hot encoding, scaling.

Here, we extracted rating, no. of reviews, no. of followers, cuisines for each restaurant.

# Clustering

Clustering or cluster analysis is a machine learning technique, which groups the unlabelled dataset. It can be defined as "A way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group."

It does it by finding some similar patterns in the unlabelled dataset such as shape, size, color, behavior, etc., and divides them as per the presence and absence of those similar patterns.

It is an unsupervised learning method, hence no supervision is provided to the algorithm, and it deals with the unlabeled dataset.

After applying this clustering technique, each cluster or group is provided with a cluster-ID. ML system can use this id to simplify the processing of large and complex datasets.

## Agglomerative Hierarchial Clustering

Also known as bottom-up approach or hierarchical agglomerative clustering (HAC).

A structure that is more informative than the unstructured set of clusters returned by flat clustering. This clustering algorithm does not require us to prespecify the number of clusters.

Bottom-up algorithms treat each data as a singleton cluster at the outset and then successively agglomerates pairs of clusters until all clusters have been merged into a single cluster that contains all data.

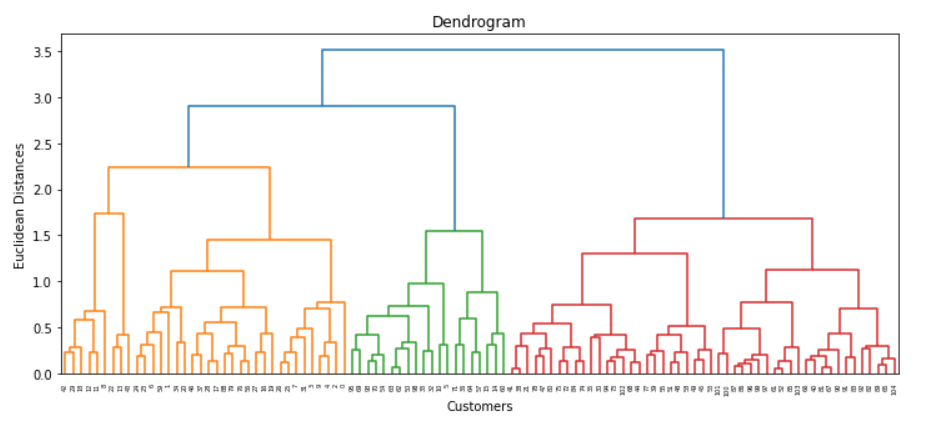


Fig 7: Dendogram

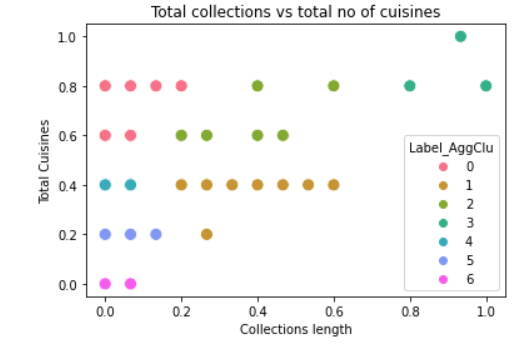
**

Fig 8: Total collections vs total no of cuisines

## DBSCAN clustering

K-Means clustering may cluster loosely related observations together. Every observation becomes a part of some cluster eventually, even if the observations are scattered far away in the vector space. Since clusters depend on the mean value of cluster elements, each data point plays a role in forming the clusters. A slight change in data points might affect the clustering outcome. This problem is greatly reduced in DBSCAN due to the way clusters are formed. This is usually not a big problem unless we come across some odd shape data.

Another challenge with k-means is that you need to specify the number of clusters (“k”) in order to use it. Much of the time, we won’t know what a reasonable k value is a priori.

What’s nice about DBSCAN is that you don’t have to specify the number of clusters to use it. All you need is a function to calculate the distance between values and some guidance for what amount of distance is considered “close”. DBSCAN also produces more reasonable results than k-means across a variety of different distributions.

## 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 is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.

The elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned center.

## mini-batch k-means

Mini Batch K-means algorithm‘s main idea is to use small random batches of data of a fixed size, so they can be stored in memory. Each iteration a new random sample from the dataset is obtained and used to update the clusters and this is repeated until convergence.

## Validation of all above methods

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 and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them. A variety of measures have been proposed in the literature for evaluating clustering results. The term clustering validation is used to design the procedure of evaluating the results of a clustering algorithm.

Silhouette score –

Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.

1: Means clusters are well apart from each other and clearly distinguished.

0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.

-1: Means clusters are assigned in the wrong way.

Average silhouette method computes the average silhouette of observations for different values of k. The optimal number of clusters k is the one that maximize the average silhouette over a range of possible values for k.

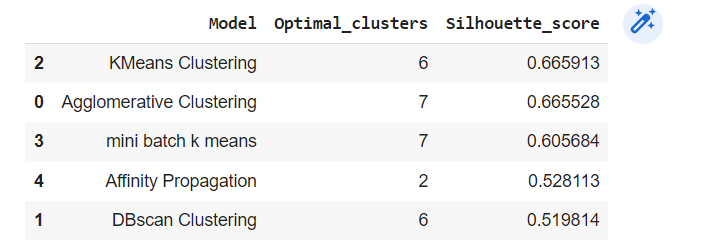


Fig 9: Silhouette score

# Sentiment Analysis

The next part is of sentiment analysis. For the sentiment analysis, The first step is text pre-processing.

In text pre-processing, we have to remove the following from the dataset.

(1) Punctuations

(2) Emojis

(3) multiple spaces

(4) stop words and special characters

(5) \n present in review

(6) Numbers and ratings specified within the review

(7) need to use lemmatization

After performing above operations, we have to clean the dataset and verify once.

Next step is to build the models.

We used following models for the sentiment analysis.

## DecisionTreeClassifier

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

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.

The decisions or the test are performed on the basis of features of the given dataset.

It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.

A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

Below diagram explains the general structure of a decision tree:

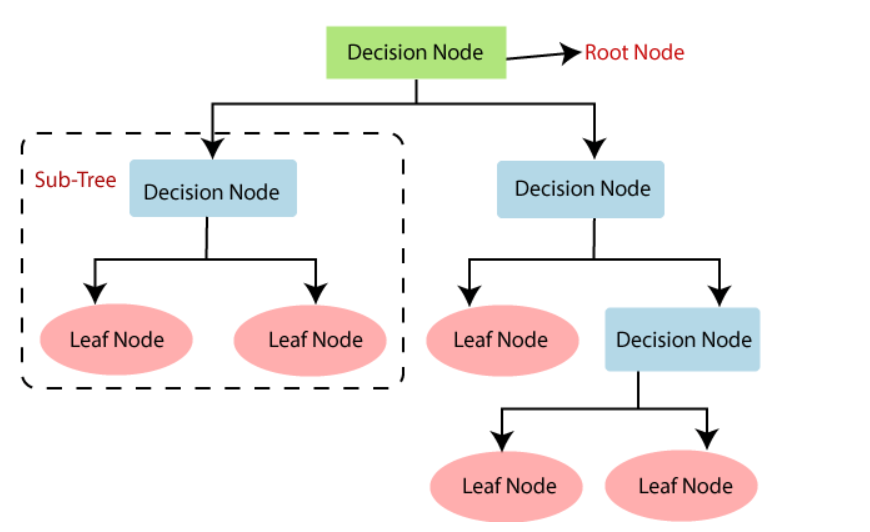


Fig 10: Decision Tree classifier

There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model. Below are the two reasons for using the Decision tree:

Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.

The logic behind the decision tree can be easily understood because it shows a tree-like structure.

We are getting an accuracy of 0.75 with this model.

## Random Forest Classifier

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:

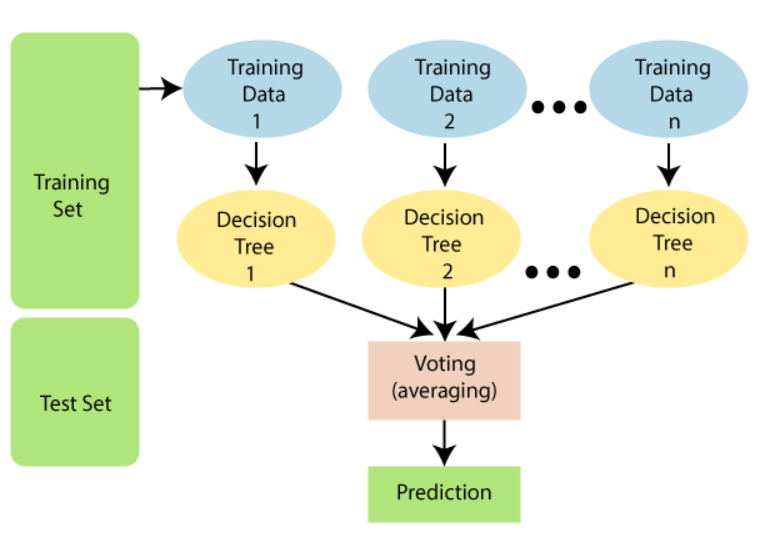


Fig 11: Random Forest Classifier

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier:

There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.

The predictions from each tree must have very low correlations.

Below are some points that explain why we should use the Random Forest algorithm:

It takes less training time as compared to other algorithms.

It predicts output with high accuracy, even for the large dataset it runs efficiently.

It can also maintain accuracy when a large proportion of data is missing.

We are getting an accuracy of 0.84 with this model.

## K-Nearest Neighbour Classifier

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.

K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.

It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset. Consider the below diagram:

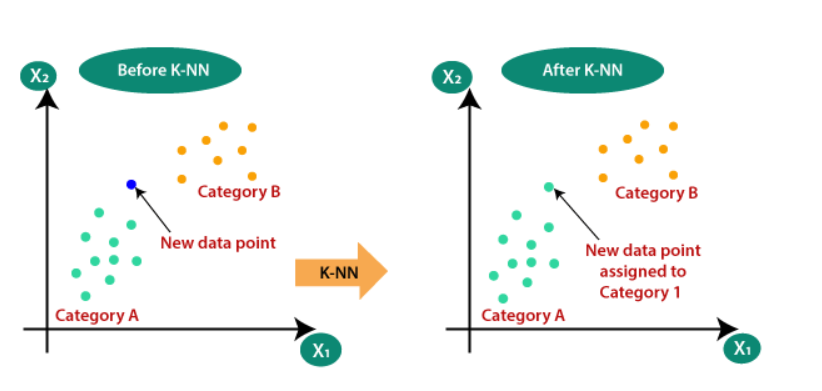


Fig 12: KNN Classifier

We are getting an accuracy of 0.65 with this model.

## 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. However, primarily, it is used for Classification problems in Machine Learning.

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 as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:

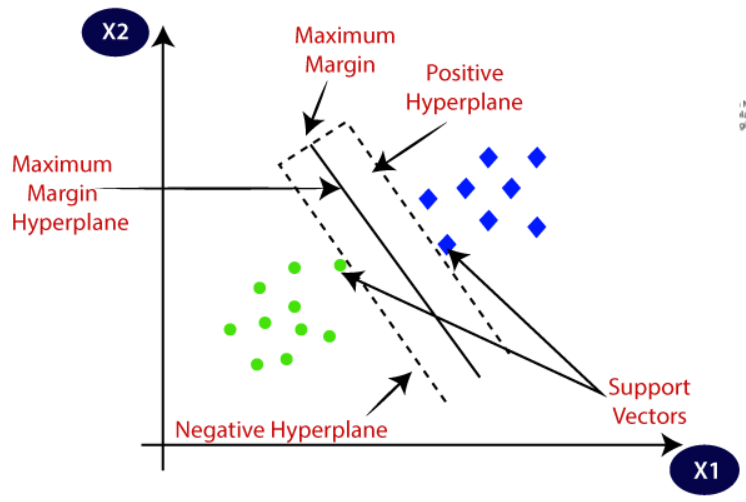


Fig 11: Support Vector Machine

We are getting an accuracy of 0.85 with this model.

## Logistic Regression

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.

Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:

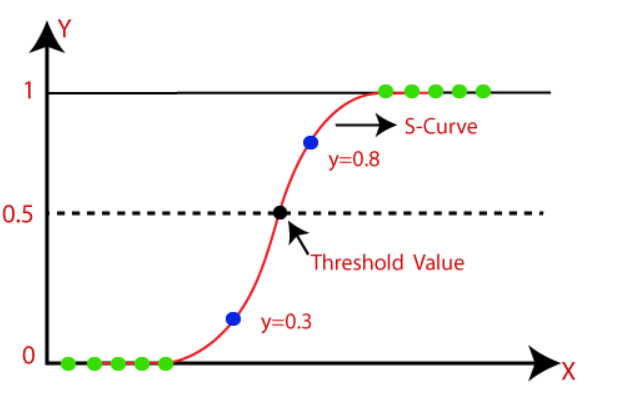


Fig12: Logistic Regression

The sigmoid function is a mathematical function used to map the predicted values to probabilities.

It maps any real value into another value within a range of 0 and 1.

The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.

In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

We are getting an accuracy of 0.85 with this model.

## Multinomial Naïve Bayes

Multinomial Naive Bayes is one of the variations of the Naive Bayes algorithm in machine learning which is very useful to use on a dataset that is distributed multinomially. When there are multiple classes to classify, this algorithm can be used because to predict the label of the text it calculates the probability of each label for the input text and then generates the label with the highest probability as output.

Some of the advantages of using this algorithm for multinomial classification are:

It is easy to use on continuous and discrete data

It can handle large data sets

It can classify data with multiple labels.

We are getting an accuracy of 0.79 with this model.

## XG Boost Classifier

XGBoost is an implementation of Gradient Boosted decision trees.

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. It can work on regression, classification, ranking, and user-defined prediction problems.

We are getting an accuracy of 0.82 with this model.

## Light Gradient Boosting Machine Classifier

LightGBM is a gradient boosting framework based on decision trees to increases the efficiency of the model and reduces memory usage.

It uses two novel techniques: Gradient-based One Side Sampling and Exclusive Feature Bundling (EFB) which fulfills the limitations of histogram-based algorithm that is primarily used in all GBDT (Gradient Boosting Decision Tree) frameworks. The two techniques of GOSS and EFB described below form the characteristics of LightGBM Algorithm. They comprise together to make the model work efficiently and provide it a cutting edge over other GBDT frameworks.

We are getting an accuracy of 0.85 with this model

From the above, we can see that SVM, Logistic Regression, Random Forest and LGBM are giving us good score.

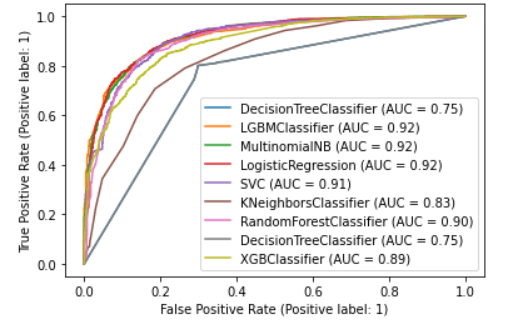


Fig 13: ROC Curve

According to the above ROC curve, LGBM, Multinomial NB and Logistic regression are performing good.

# Hyper parameter tuning

Our next step is to perform hyper parameter tuning for the models. Let’s choose the models for hyper parameter tuning.

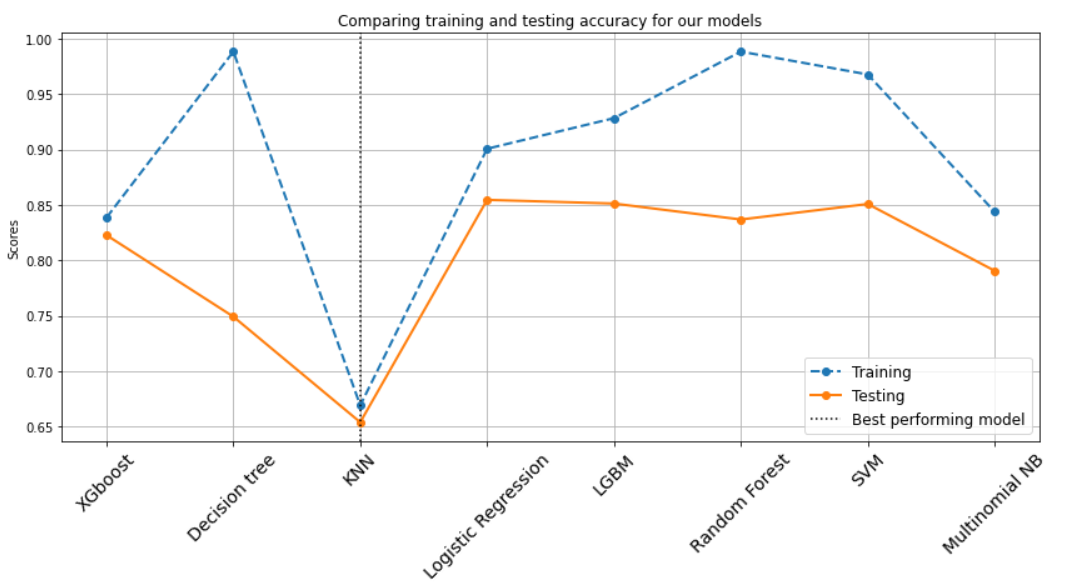


Fig 14: Training and Testing accuracy of models

We can see that our models are quite overfitting. But if we consider those models with less overfitting and more than 80% accuracy then it would be

(1)XGboost

(2) Logistic regression

(3) LGBM and

(4) SVM

Now we do the hyperparameter tuning for these models and we choose the best amongst these.

The following metrics were used to evaluate the performance of our pipelines.

**Accuracy**: Accuracy is the quintessential classification metric. Accuracy is the proportion of true results among the total number of cases examined. It is easily suited for binary as well as a multiclass classification problem. Accuracy is a valid choice of evaluation for classification problems which are well balanced and not skewed or no/less class imbalance.

Accuracy = (TP + TN) / (TP + TN + FP + FN)

**Precision (Macro):** Precision answers what proportion of predicted positives is truly positives. Precision is a valid choice of evaluation metric when we want to be very sure of our prediction.

Precision Macro = (Sum of Precision for each individual class) / (No. of Classes)

**Recall (Macro):** Recall answers what proportion of actual Positives is correctly classified. Recall is a valid choice of evaluation metric when we want to capture as many positives as possible.

Recall Macro = (Sum of Recall for each individual class) / (No. of Classes)

**F1 Score**: The F1 score is a number between 0 and 1 and is the harmonic mean of precision and recall. F1 score sort of maintains a balance between the precision and recall for classifier. If precision is low, the F1 is low and if the recall is low again F1 score is low. The F1 score manages the tradeoff.

F1 Score = 2 \* (Precision \* Recall) / (Precision + Recall)

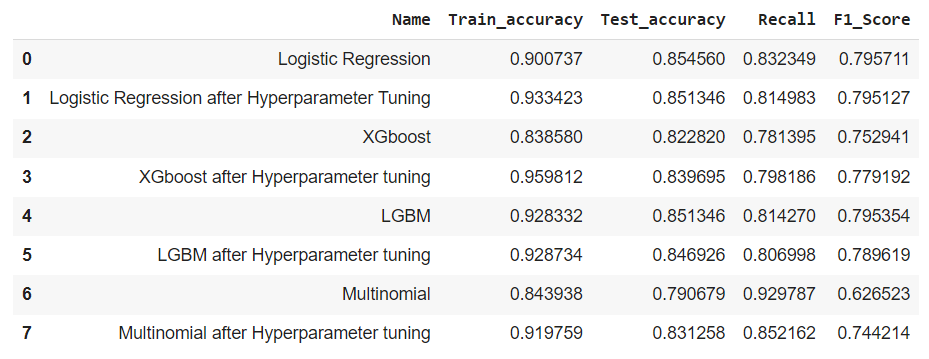


Fig 15: Accuracies after hyper parameter tuning

We are not seeing that much of change in recall and accuracy.

After hyper-parameter tuning, we can see that XGBoost model is overfitting.

We can say it’s been tuned to its perfect fit.

# Conclusion

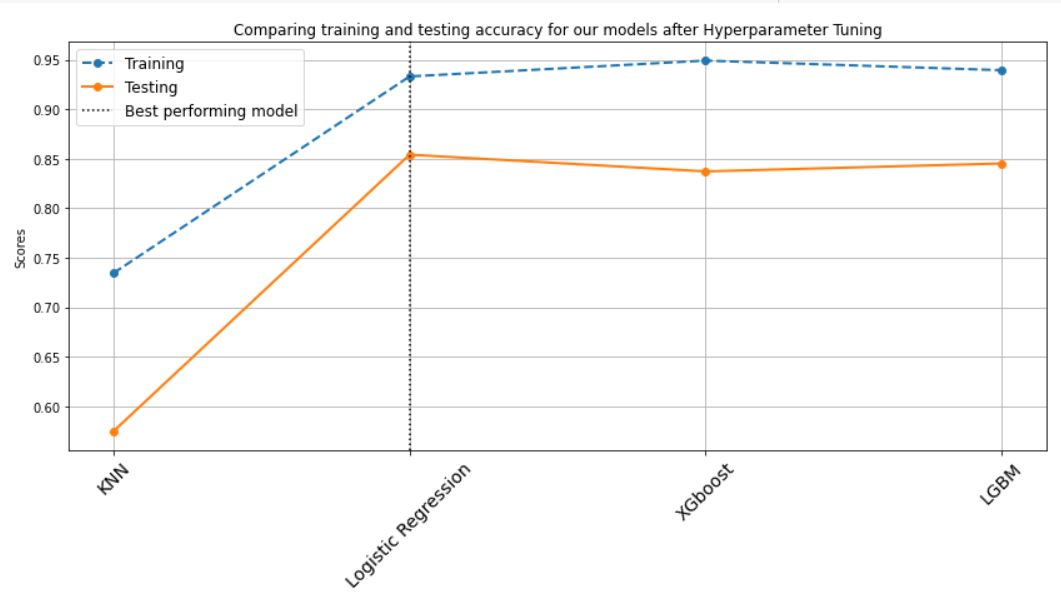


Fig 16: Comparing train and test accuracies after hyper parameter tuning

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

From the above, we can see that Logistic regression model is working fine as compared to other models. Its accuracy and recall are maximum. Finally, our conclusion is that the Logistic regression model is the best suitable model for this sentimental analysis of the project.

References:

Almabetter

Geeksforgeeks.org

W3schools.com

Tutorialspoint.com