Capstone Project

Zomato Clustering & Sentiment Analysis

Team Members

Swathi V Hebbar
Charishma Suddala

Contents

- Introduction
- Problem Statement
- Methodology
    - (1) Loading the data
    - (2) Exploratory Data Analysis
    - (3) Treating missing values and outliers
    - (4) Natural Language Processing
    - (5) Clustering
    - (6) Sentimental analysis
    - (7) Building models
    - (8) Hyper parameter tuning
- Conclusion

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 day by day. Zomato a online food delivery website is the perfect example of this.
- The Project focuses on Customers and Company, you have t**o analyze the sentiments of the reviews given by the customer in the data and made some useful conclusion in the form of Visualizations**. Also, **cluster the zomato restaurants into different segments**.
- This could help in clustering the restaurants into segments. Also the data has valuable information around cuisine and costing which can be used in cost vs. benefit analysis.

Data Sets Given

Zomato Restaurant names and Metadata

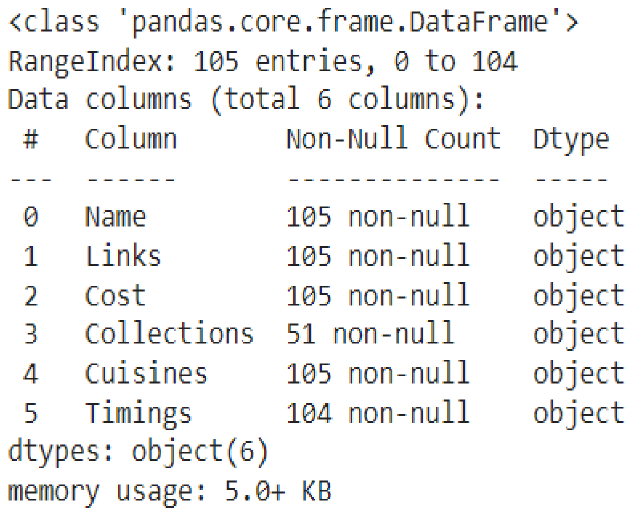
- This dataset gives the information about the restaurants, cuisines, collections available

Zomato Restaurant reviews

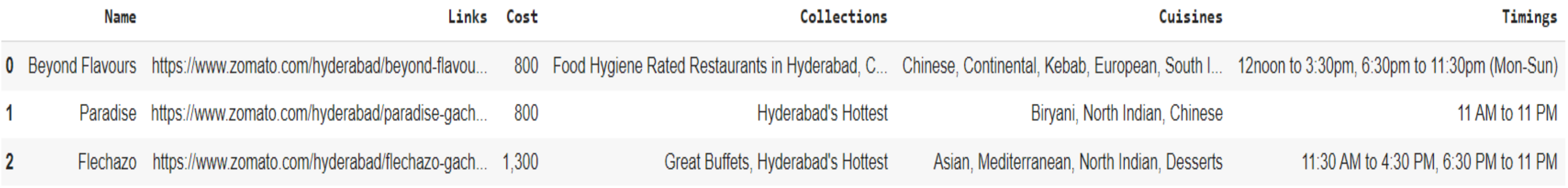
- This Dataset gives the information of the reviewers and followers of the every restaurants which are available

Description of Zomato Restaurant names and Metadata Dataset

1.Name : Name of Restaurants
2.Links : URL Links of Restaurants
3.Cost : Per person estimated Cost of dining
4.Collection : Tagging of Restaurants with respect to Zomato categories
5.Cuisines : Cuisines served by Restaurants
6.Timings : Restaurant Timings



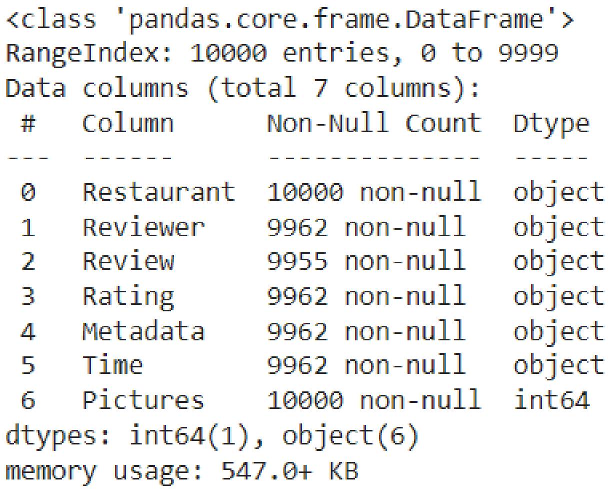
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 105 entries, 0 to 104
Data columns (total 6 columns):
 #   Column       Non-Null Count  Dtype 
---  ------       --------------  ----- 
 0   Name         105 non-null    object
 1   Links        105 non-null    object
 2   Cost         105 non-null    object
 3   Collections  51 non-null     object
 4   Cuisines     105 non-null    object
 5   Timings      104 non-null    object
dtypes: object(6)
memory usage: 5.0+ KB
```



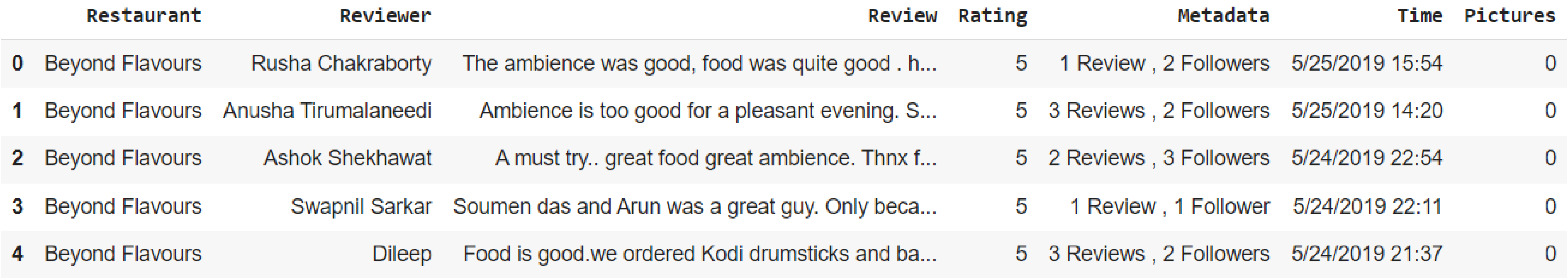
	Name	Links	Cost	Collections	Cuisines	Timings
0	Beyond Flavours	https://www.zomato.com/hyderabad/beyond-flavou...	800	Food Hygiene Rated Restaurants in Hyderabad, C...	Chinese, Continental, Kebab, European, South I...	12noon to 3:30pm, 6:30pm to 11:30pm (Mon-Sun)
1	Paradise	https://www.zomato.com/hyderabad/paradise-gach...	800	Hyderabad's Hottest	Biryani, North Indian, Chinese	11 AM to 11 PM
2	Flechazo	https://www.zomato.com/hyderabad/flechazo-gach...	1,300	Great Buffets, Hyderabad's Hottest	Asian, Mediterranean, North Indian, Desserts	11:30 AM to 4:30 PM, 6:30 PM to 11 PM

Description of Zomato Restaurant 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



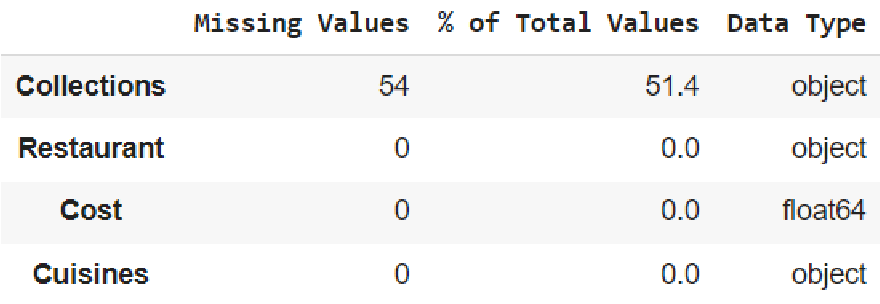
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   Restaurant  10000 non-null  object
 1   Reviewer    9962 non-null   object
 2   Review      9955 non-null   object
 3   Rating      9962 non-null   object
 4   Metadata    9962 non-null   object
 5   Time        9962 non-null   object
 6   Pictures    10000 non-null  int64 
dtypes: int64(1), object(6)
memory usage: 547.0+ KB
```



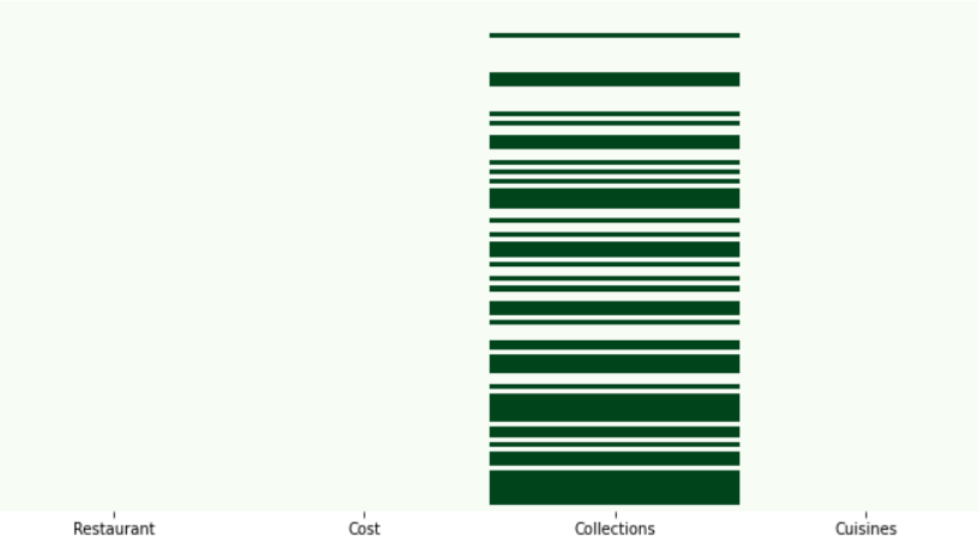
	Restaurant	Reviewer	Review	Rating	Metadata	Time	Pictures
0	Beyond Flavours	Rusha Chakraborty	The ambience was good, food was quite good . h...	5	1 Review , 2 Followers	5/25/2019 15:54	0
1	Beyond Flavours	Anusha Tirumalaneedi	Ambience is too good for a pleasant evening. S...	5	3 Reviews , 2 Followers	5/25/2019 14:20	0
2	Beyond Flavours	Ashok Shekhawat	A must try.. great food great ambience. Thnx f...	5	2 Reviews , 3 Followers	5/24/2019 22:54	0
3	Beyond Flavours	Swapnil Sarkar	Soumen das and Arun was a great guy. Only beca...	5	1 Review , 1 Follower	5/24/2019 22:11	0
4	Beyond Flavours	Dileep	Food is good.we ordered Kodi drumsticks and ba...	5	3 Reviews , 2 Followers	5/24/2019 21:37	0

Exploratory Data Analysis

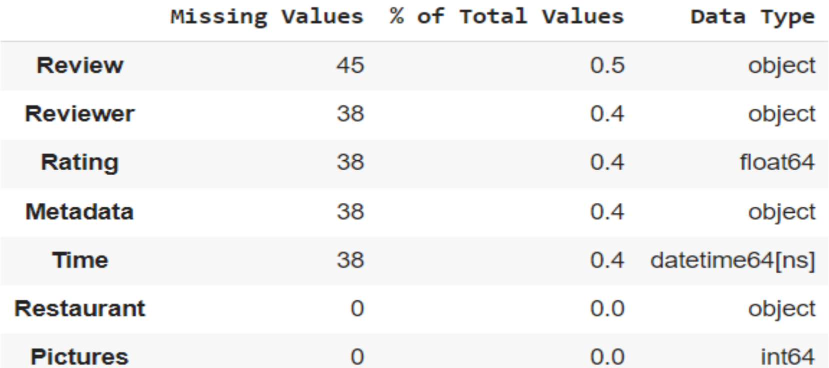
Zomato Restaurant names and Metadata



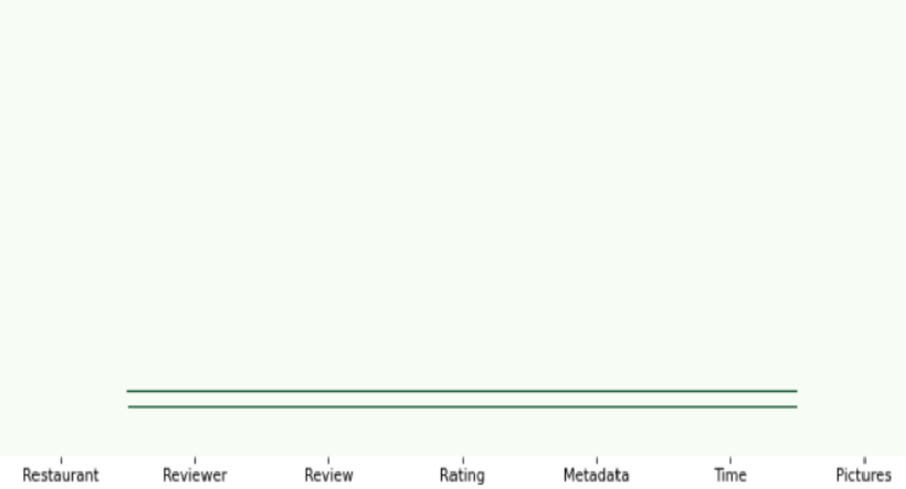
	Missing Values	% of Total Values	Data Type
Collections	54	51.4	object
Restaurant	0	0.0	object
Cost	0	0.0	float64
Cuisines	0	0.0	object



Zomato Restaurant reviews

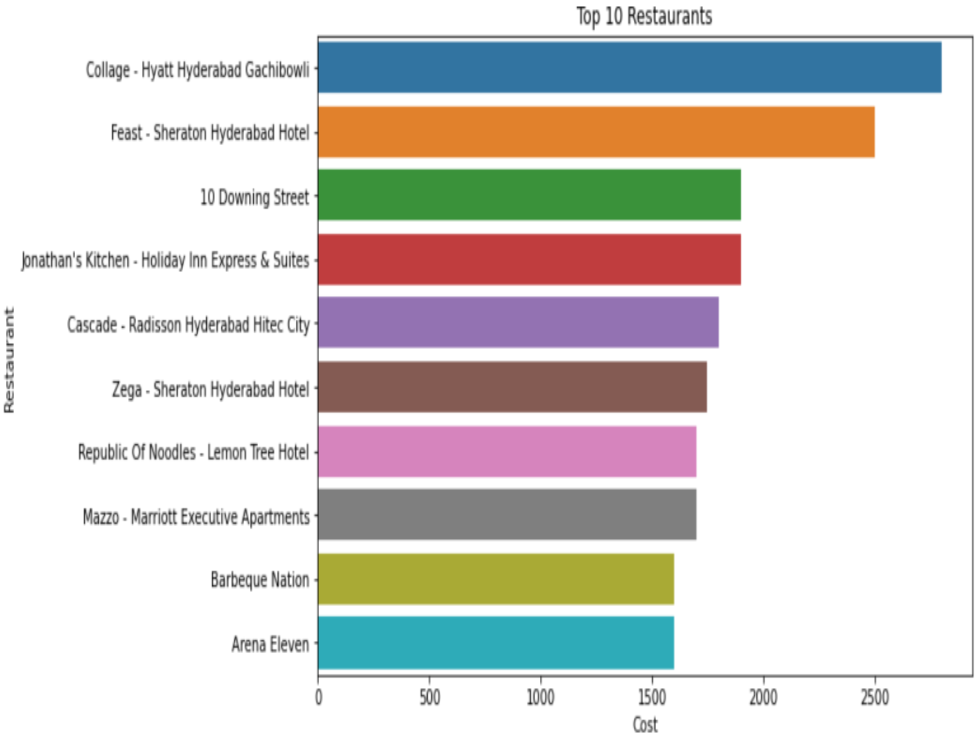


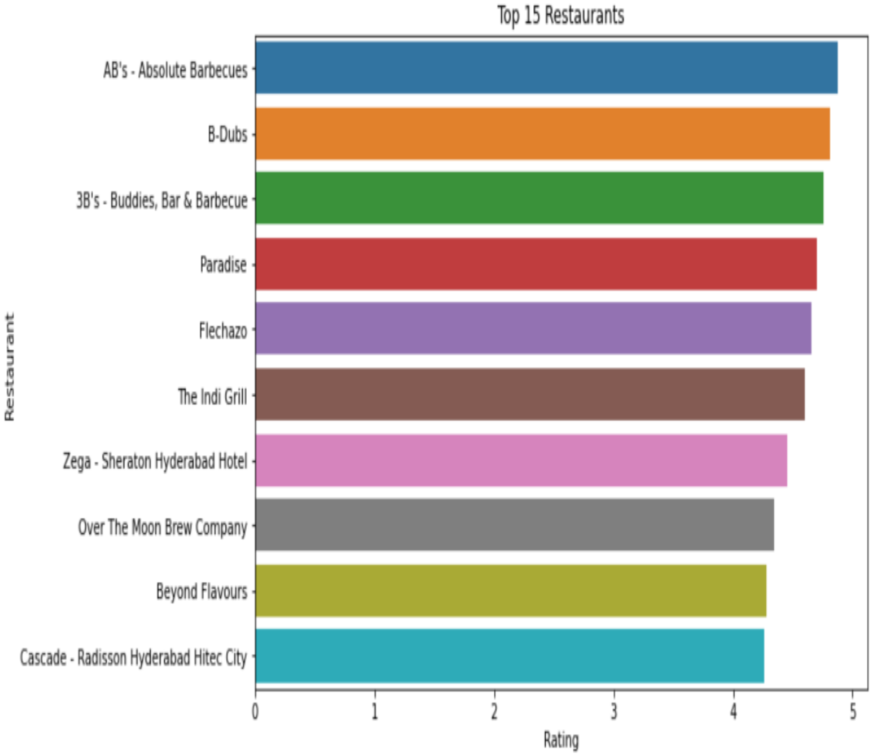
	Missing Values	% of Total Values	Data Type
Review	45	0.5	object
Reviewer	38	0.4	object
Rating	38	0.4	float64
Metadata	38	0.4	object
Time	38	0.4	datetime64[ns]
Restaurant	0	0.0	object
Pictures	0	0.0	int64



Costly Restaurants

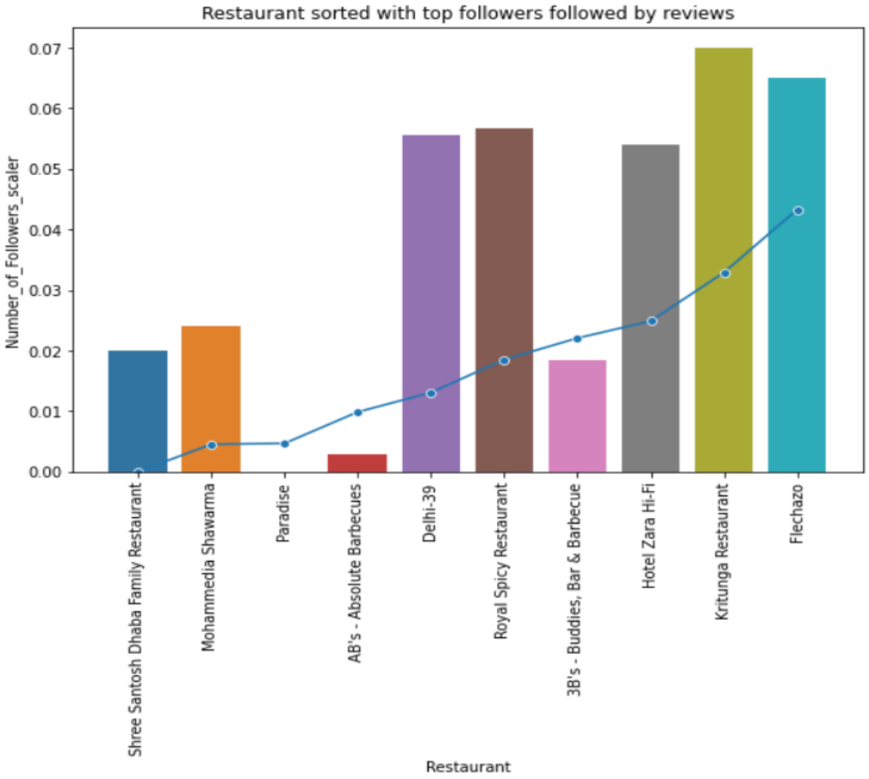
Top Restaurants based on Rating

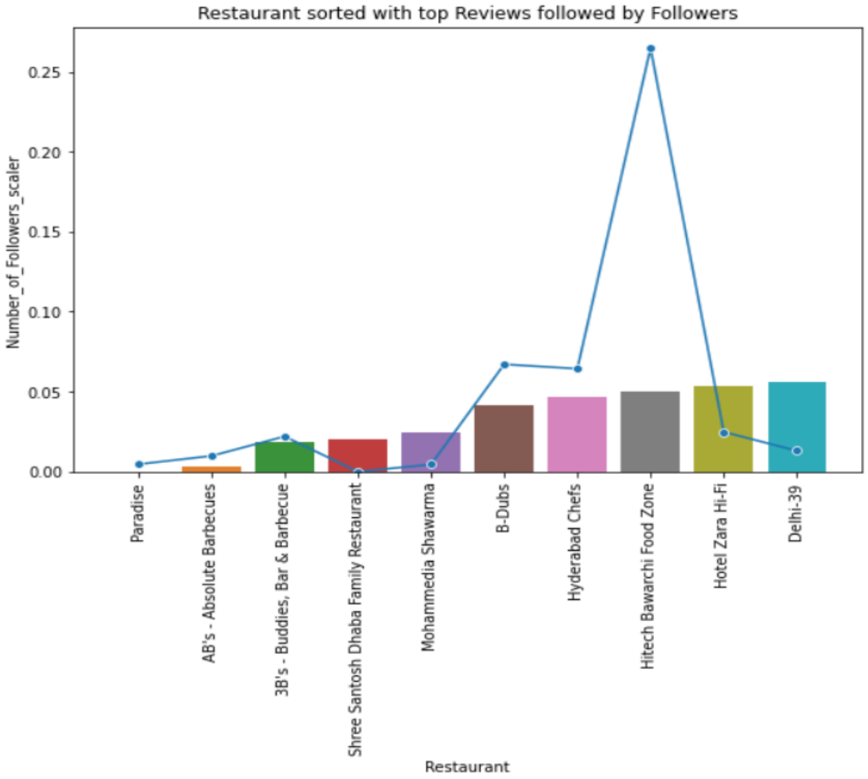




These graphs represents the top 10 restaurants with respect to cost and Rating

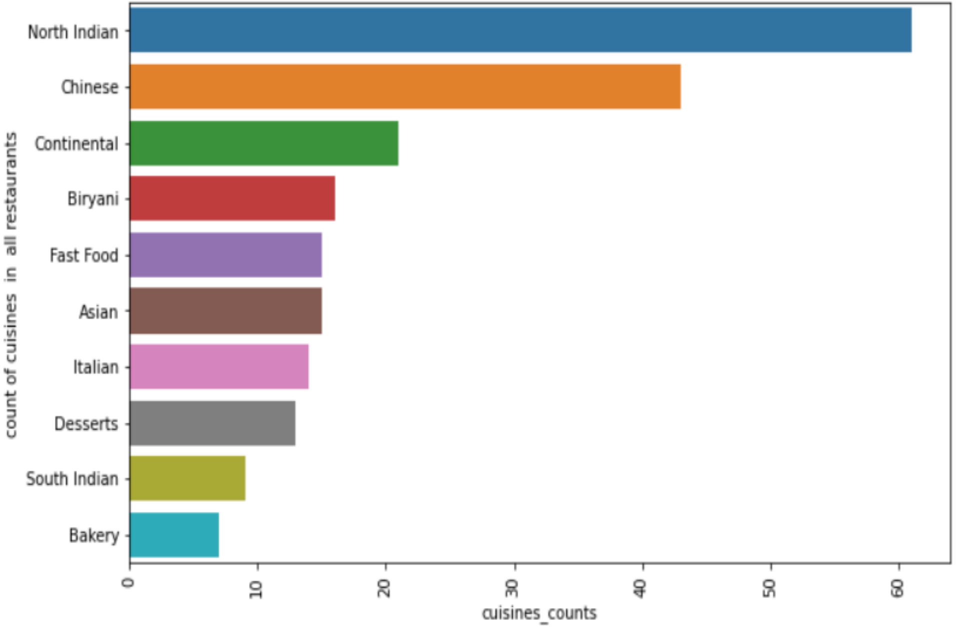
Graph which represents the number of followers and number of reviews for each restaurants





Restaurants with maximum followers

Restaurants with more reviews



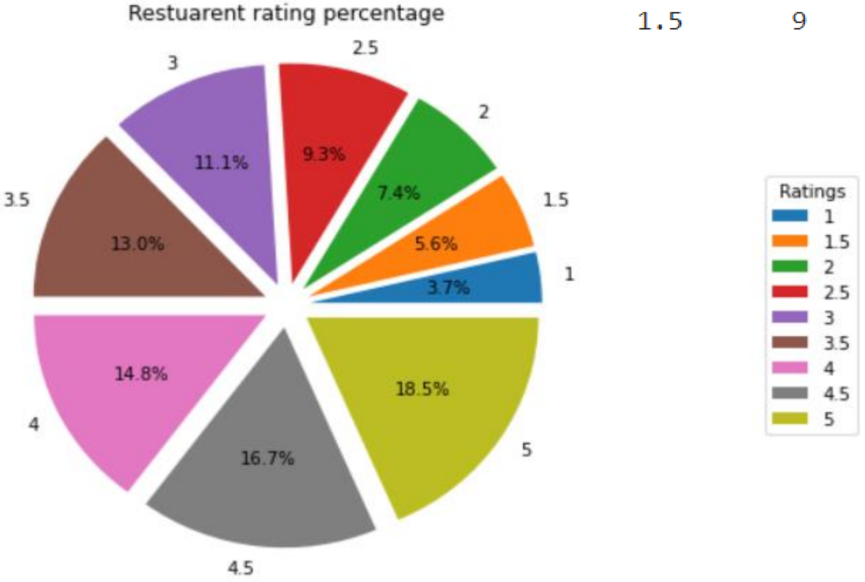
Famous Cuisines offered by Restaurants

'North Indian' is the **Popular Cuisine** which is offered by almost many restaurants. And **'Malaysian'** is the **rare cuisine**.

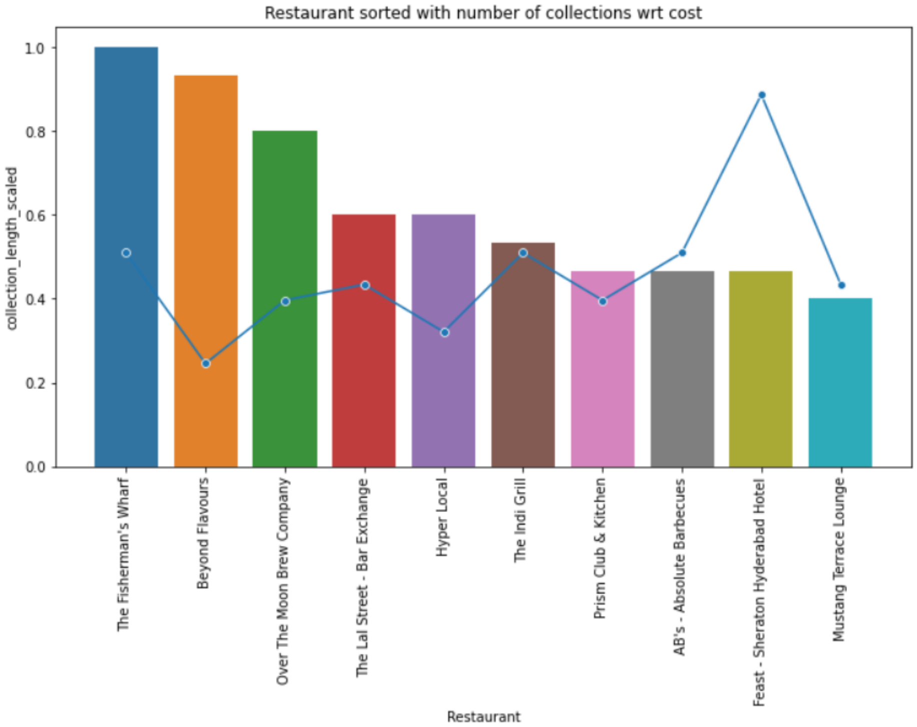
Restaurants available wrt different Ratings

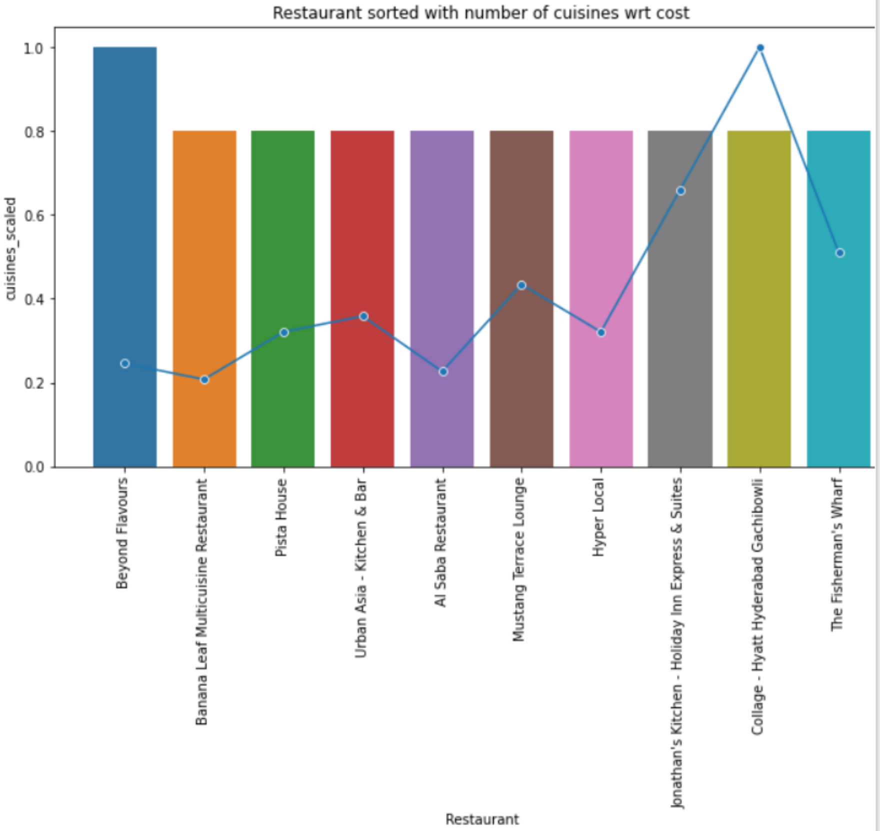
- 3826 restaurants available with rating 5

```
5.0    3826
4.0    2373
1.0    1735
3.0    1192
2.0     684
4.5      69
3.5      47
2.5      19
1.5       9
```

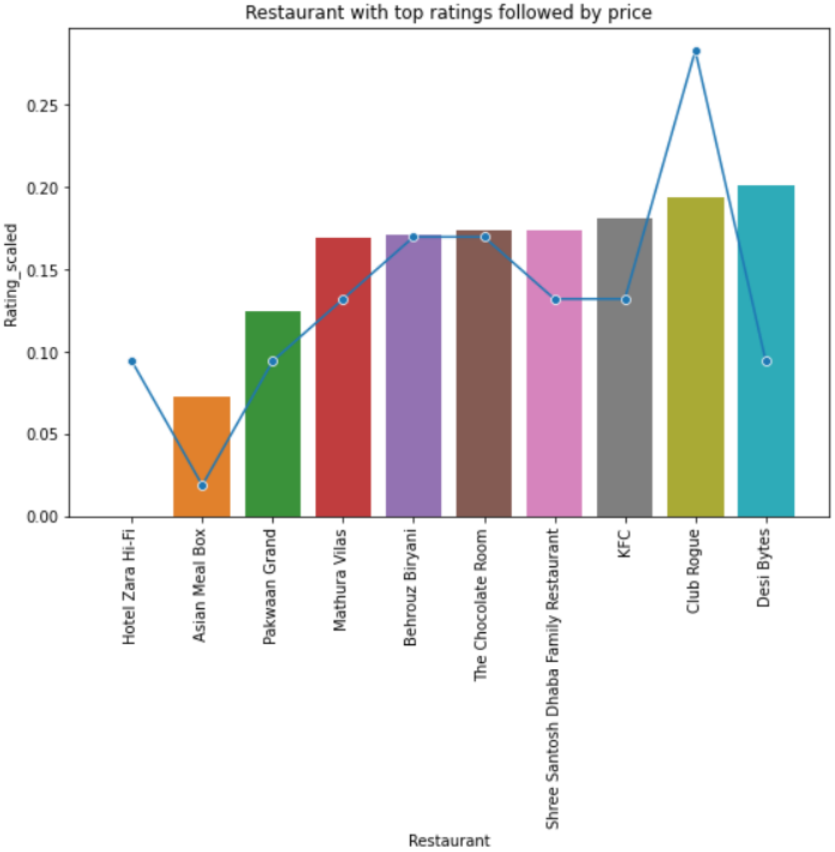


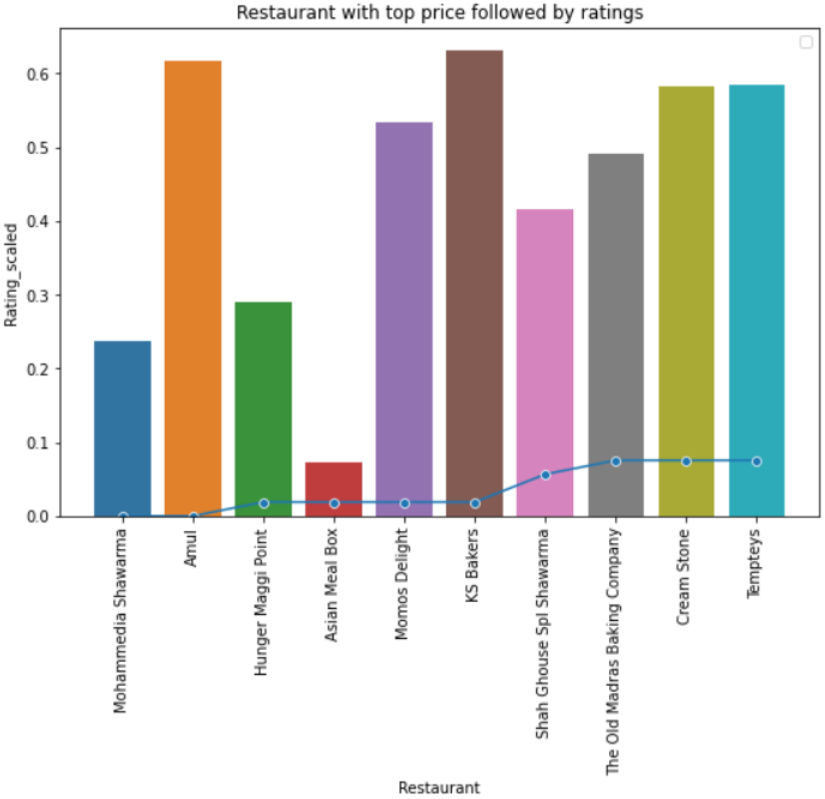
Total number of collections and cuisines wrt cost offered by each restaurant


Restaurant sorted with number of collections wrt cost

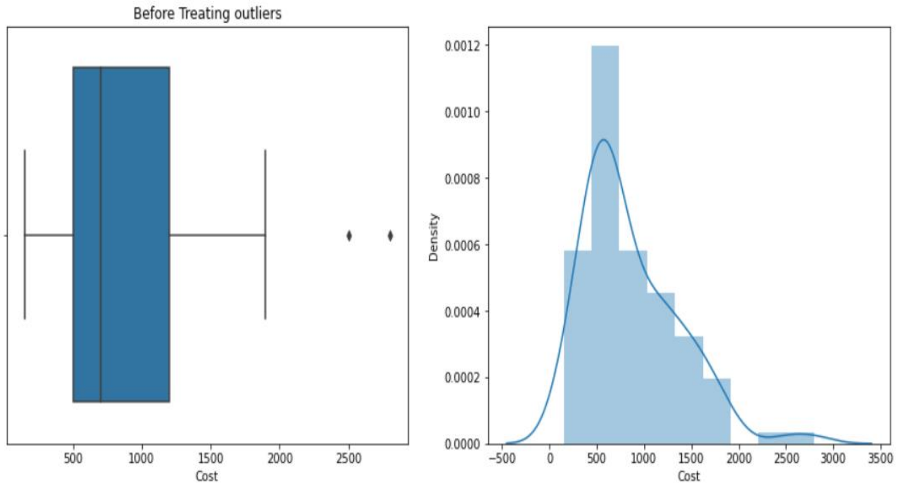

Restaurant sorted with number of cuisines wrt cost

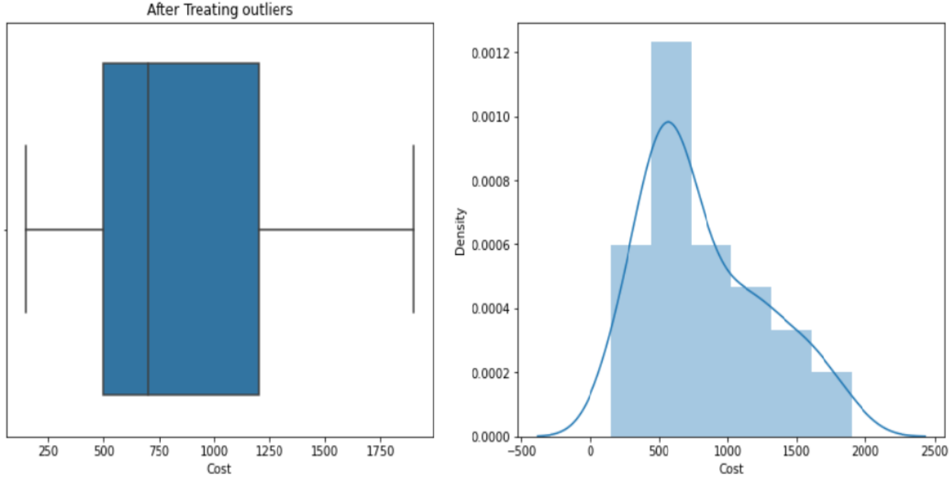
Best Restaurant with respect to ratings and price





Treatment of Missing Values and Outliers





Treatment of missing values

- Zomato Restaurant names and Metadata Dataset contains 54% missing values in collections feature. Since it is a string, it is treated by replacing null values with Unknown.
- As we have seen from missing values graph of Zomato Restaurant reviews, it has null values at the same position of every feature. So these null values are dropped.

Treatment of Outliers

Outliers treatment in this data set is treated by z score.

Natural Language Processing

- Natural Language Processing (NLP) refers to AI method of communicating with an intelligent systems using a natural language such as English.
- Since we have sentences in or dataset we have used, NLP to process them.

Steps Involved

1. Removing of stop words punctuations, emojis etc from the text.
2. Count Vectorizer : **CountVectorizer** is used to transform a corpora of text to a vector of term / token counts.
3. Stemming and lemmatization: these two are the text normalization techniques. These methods are used to process the text accordingly.
4. TFIDF vectorizer: TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is very common algorithm to transform text into a meaningful representation of numbers which is used to fit machine algorithm for prediction

Feature Engineering

- Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning.
- Feature Engineering consists of various process :

  (1) Feature Creation (2) Transformation (3) Feature Selection

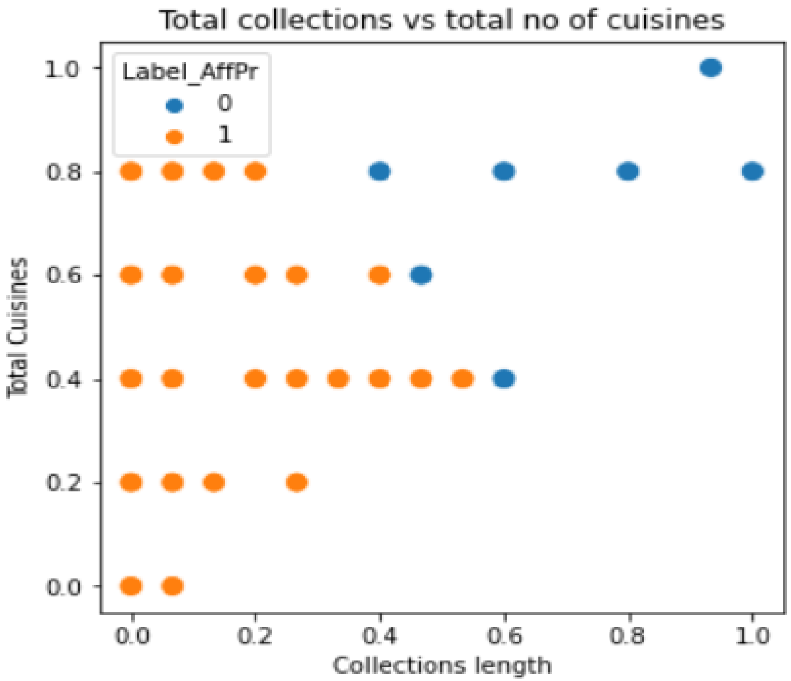
Clustering

- Cluster analysis, or clustering, is an unsupervised machine learning task. Similarity between observations is defined using some inter-observation distance measures or correlation-based distance measures.
- It involves automatically discovering natural grouping in data. Unlike supervised learning (like predictive modeling), clustering algorithms only interpret the input data and find natural groups or clusters in feature space.
- Cluster analysis is an iterative process where subjective evaluation of the identified clusters is fed back into changes to algorithm configuration until a desired or appropriate result is achieved.

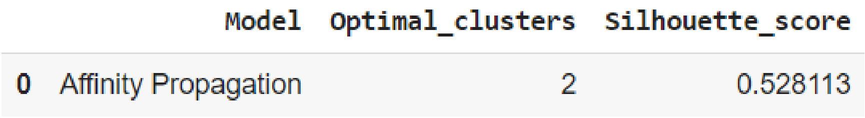
Clustering techniques we have used are

- ➢ **Affinity Propagation**
- ➢ **Hierarchial Clustering**
- ➢ **dbscan clustering**
- ➢ **K Means Clustering**
- ➢ **mini-batch k-means**

Affinity Propagation

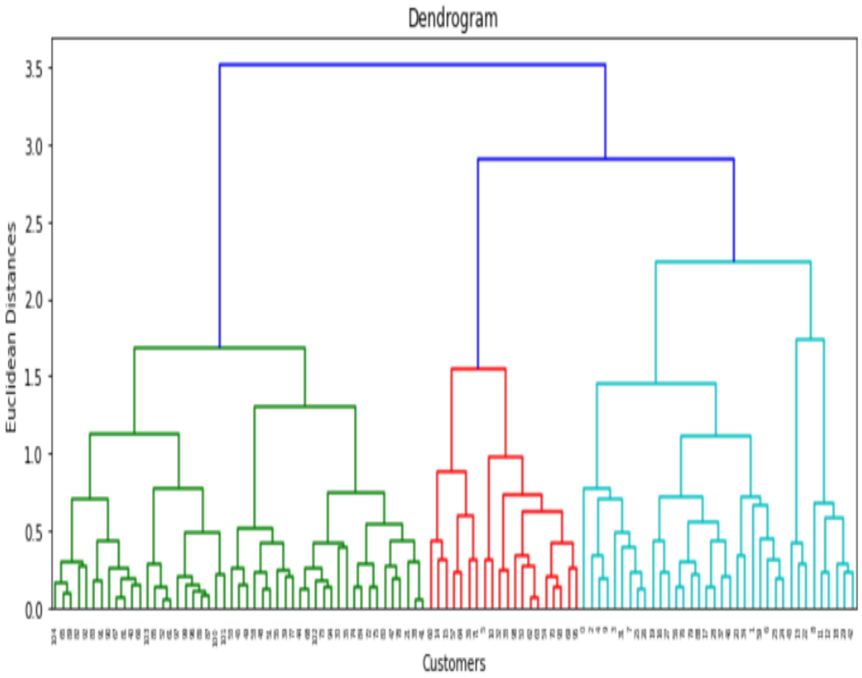


- Affinity Propagation involves finding a set of exemplars that best summarize the data.
- It is implemented via the Affinity Propagation class and the main configuration to tune is the “damping” set between 0.5 and 1, and perhaps “preference.”
- Clustering has done on Total Cuisines and collections length



	Model	Optimal_clusters	Silhouette_score
0	Affinity Propagation	2	0.528113

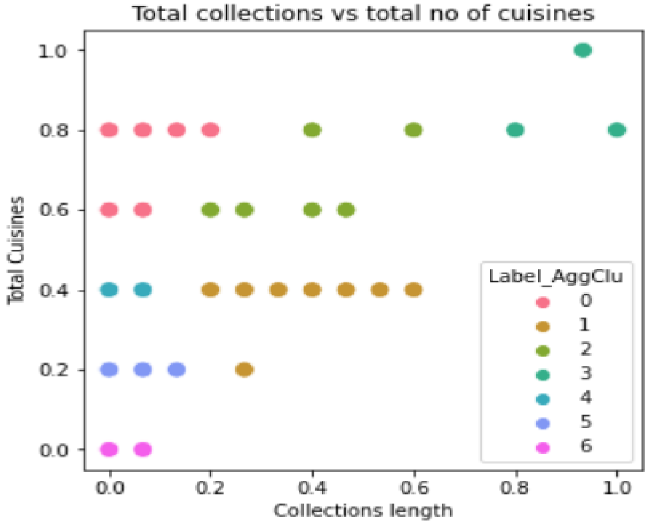
Hierarchial Clustering



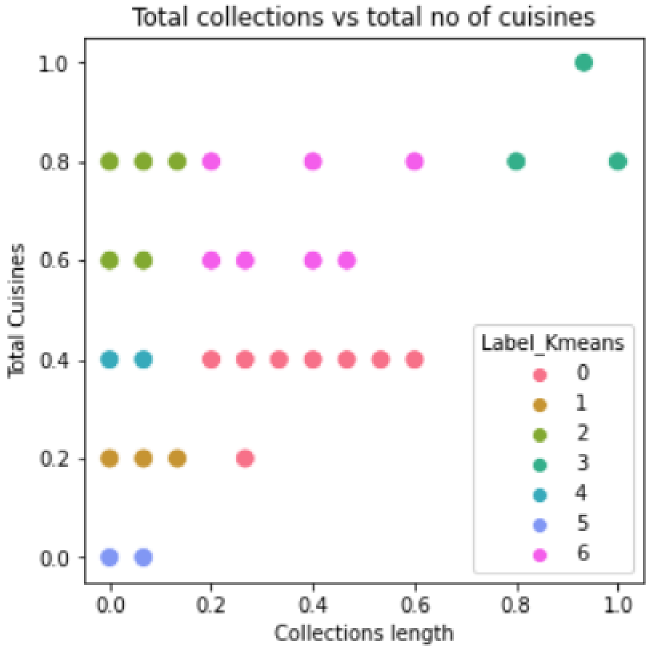
Hierarchical clustering is used to group together the unlabeled data points having similar characteristics.

Agglomerative Hierarchial Clustering

- In agglomerative hierarchical algorithms, each data point is treated as a single cluster and then successively merge or agglomerate (bottom-up approach) the pairs of clusters.

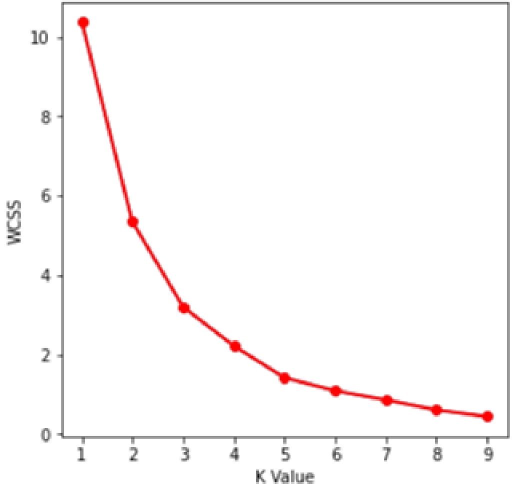


K means Clustering



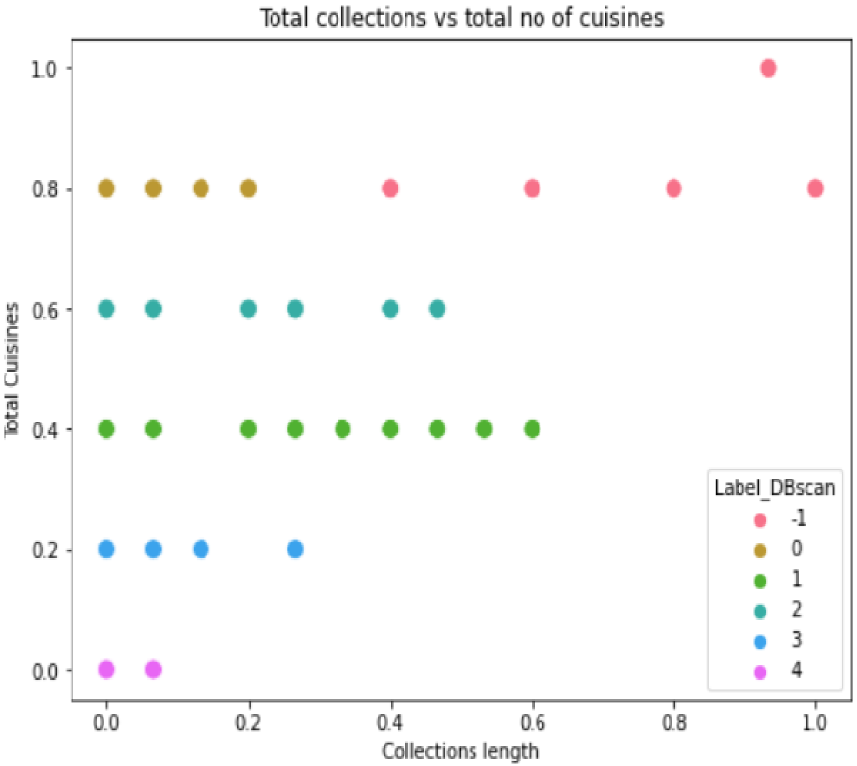
- It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.
- K defines the number of pre-defined clusters that need to be created in the process.

Elbow Method



K value is determined by elbow method

dbScan Clustering

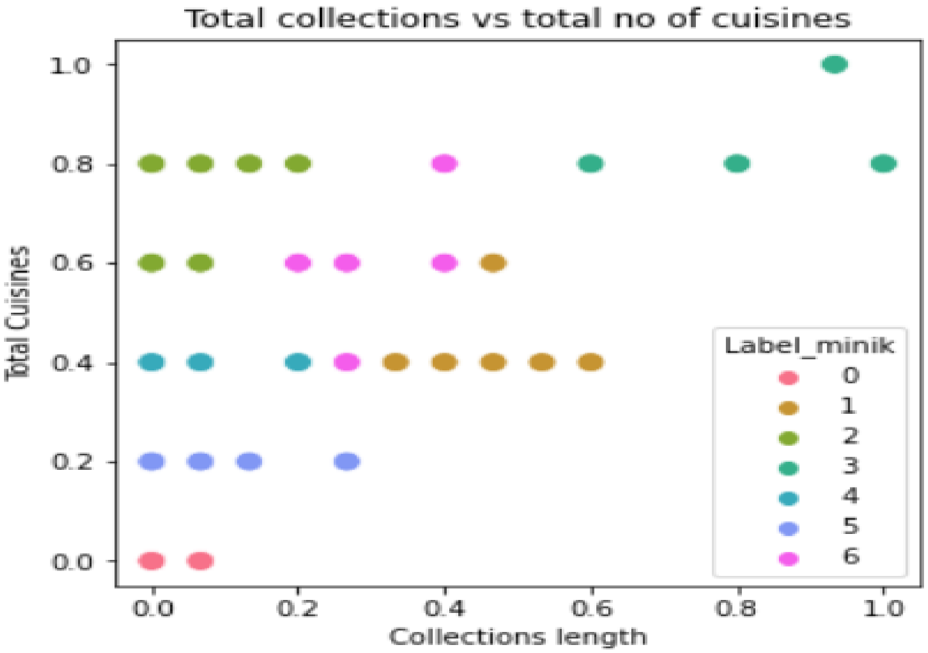


- DBSCAN stands for **d**ensity-**b**ased **s**patial **c**lustering of **a**pplications with **n**oise. It is able to find arbitrary shaped clusters and clusters with noise (i.e. outliers).
- The main idea behind DBSCAN is that a point belongs to a cluster if it is close to many points from that cluster.
- There are two key parameters of DBSCAN:

eps: The distance that specifies the neighborhoods. Two points are considered to be neighbors if the distance between them are less than or equal to eps.

minPts: Minimum number of data points to define a cluster.

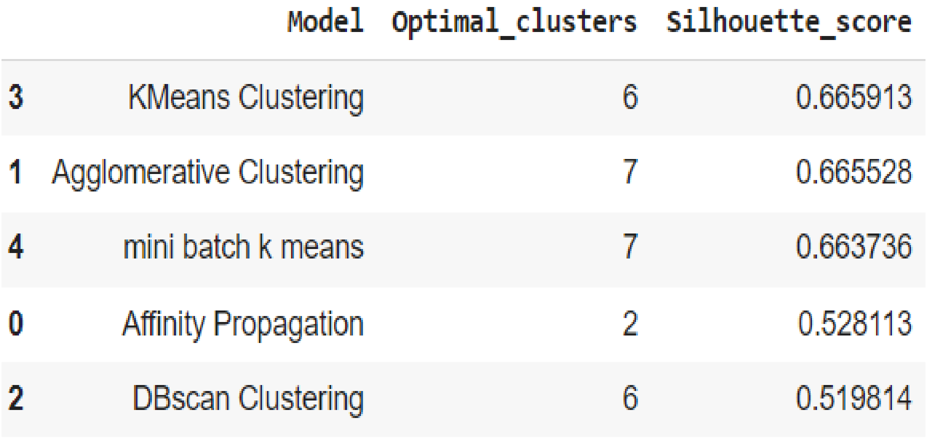
Mini-batch k-means



- Mini-Batch K-Means is a modified version of k-means that makes updates to the cluster centroids using mini-batches of samples rather than the entire dataset, which can make it faster for large datasets, and perhaps more robust to statistical noise.

- It is implemented via the Mini Batch KMeans class and the main configuration to tune is the “n_clusters” hyperparameter set to the estimated number of clusters in the data.

Validation



	Model	Optimal_clusters	Silhouette_score
3	KMeans Clustering	6	0.665913
1	Agglomerative Clustering	7	0.665528
4	mini batch k means	7	0.663736
0	Affinity Propagation	2	0.528113
2	DBscan Clustering	6	0.519814

- The term clustering validation is used to design the procedure of evaluating the results of a clustering algorithm.

Silhouette score -

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.

Through this data frame, we can conclude that the optimal number of clusters are 7 (or 6). Except the affinity propagation, all other models are giving the optimal clusters as 7(or 6).

Sentimental Analysis

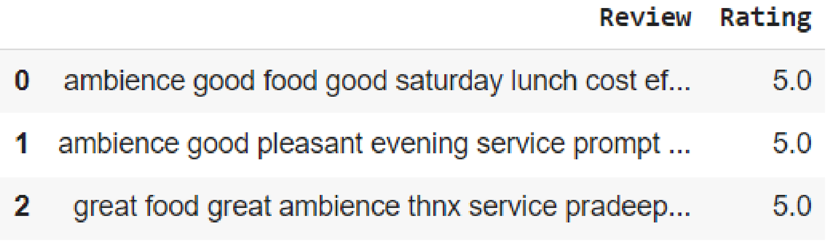
- Sentimental Analysis is the process of classifying whether a block of text is positive, negative, or, neutral.
- The goal which Sentiment analysis tries to gain is to analyze people's opinion in a way that it can help the businesses expand.
- It focuses not only on polarity (positive, negative & neutral) but also on emotions (happy, sad, angry, etc.).

Steps Involved

1. Text Processing 2. Feature Engineering 3. Train-Test Split 4. Building Models

Test-Processing

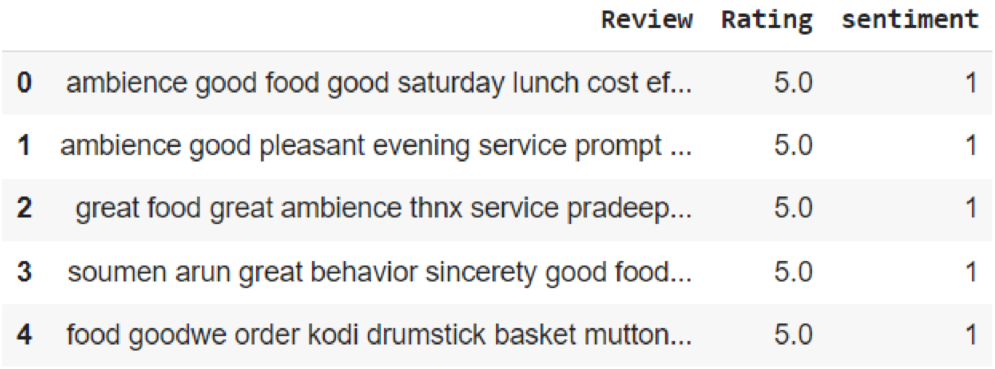
In this step all the stop words, punctuations, emojis, special characters etc are removed from the text. Then lemmatization has been applied



	Review	Rating
0	ambience good food good saturday lunch cost ef...	5.0
1	ambience good pleasant evening service prompt ...	5.0
2	great food great ambience thnx service pradeep...	5.0

Feature Engineering

A new feature sentiment is created according to the rating. If the rating > 3.5 then the sentiment is positive(1). If the rating < 3.5 then the sentiment is considered as negative(0).



	Review	Rating	sentiment
0	ambience good food good saturday lunch cost ef...	5.0	1
1	ambience good pleasant evening service prompt ...	5.0	1
2	great food great ambience thnx service pradeep...	5.0	1
3	soumen arun great behavior sincerety good food...	5.0	1
4	food goodwe order kodi drumstick basket mutton...	5.0	1

Train-Test Split

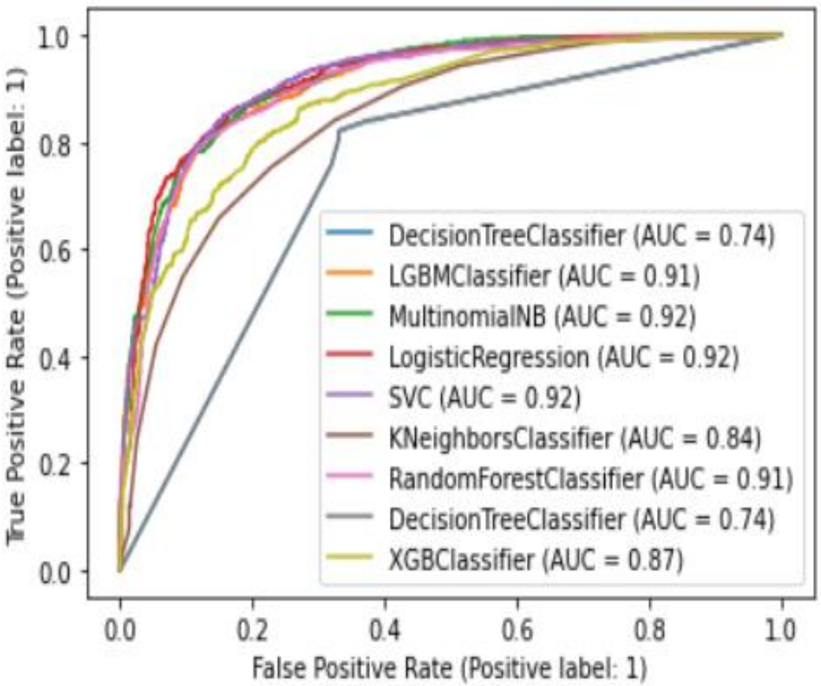
The dataset is split into Train – Test datasets. This is done to ensure that our test dataset is completely isolated and there is no information leakage during the training process of machine learning models. Since we have considered Review column Tfidf vectorizer id applied on it.

Building Models

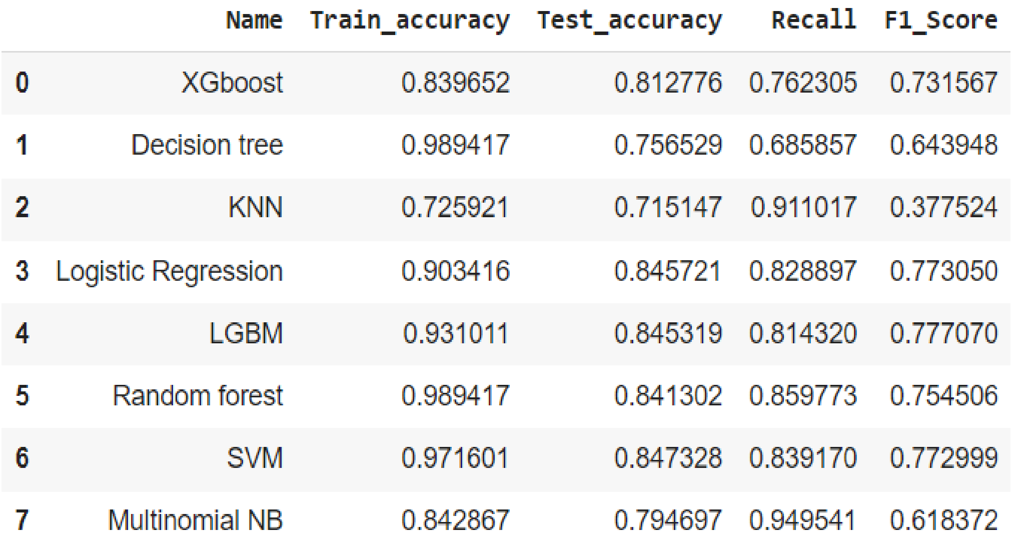
- Since we have sentiment as a class(1 and 0) we have build classification models on this data.

- There are many classification models available in supervised machine learning. The models which we have used are,
  1. Logistic regression
  2. Decision Tree
  3. Random Forest
  4. K – nearest neighbor
  5. XGBoost
  6. LGBM
  7. Support Vector Machine (SVM)
  8. Multinomial Naïve Baye’s

ROC Curve for different models

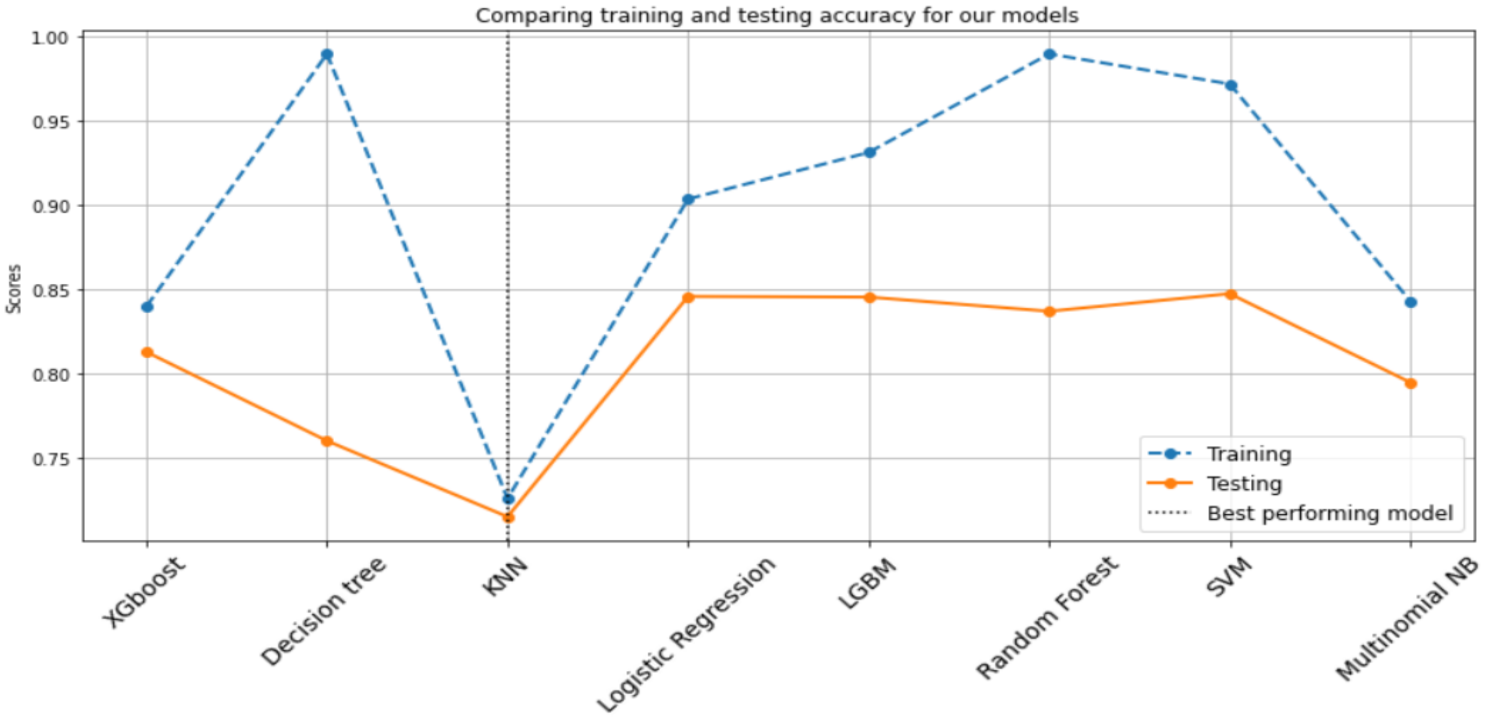


Evaluation metric for all the models

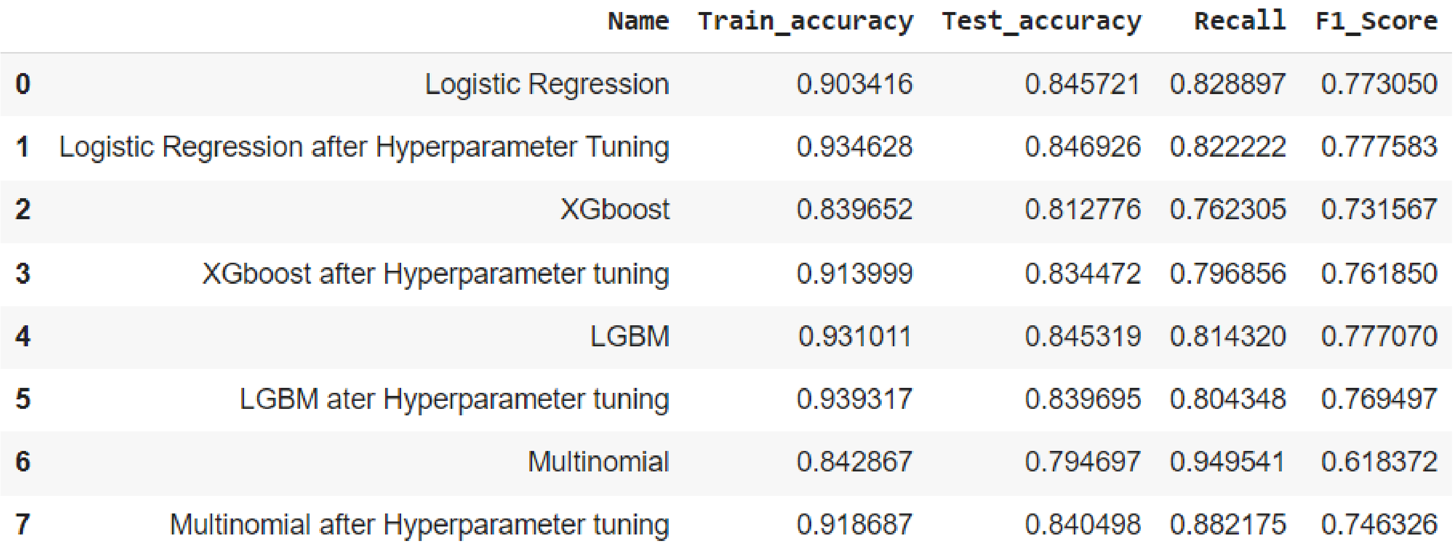


	Name	Train_accuracy	Test_accuracy	Recall	F1_Score
0	XGboost	0.839652	0.812776	0.762305	0.731567
1	Decision tree	0.989417	0.756529	0.685857	0.643948
2	KNN	0.725921	0.715147	0.911017	0.377524
3	Logistic Regression	0.903416	0.845721	0.828897	0.773050
4	LGBM	0.931011	0.845319	0.814320	0.777070
5	Random forest	0.989417	0.841302	0.859773	0.754506
6	SVM	0.971601	0.847328	0.839170	0.772999
7	Multinomial NB	0.842867	0.794697	0.949541	0.618372

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

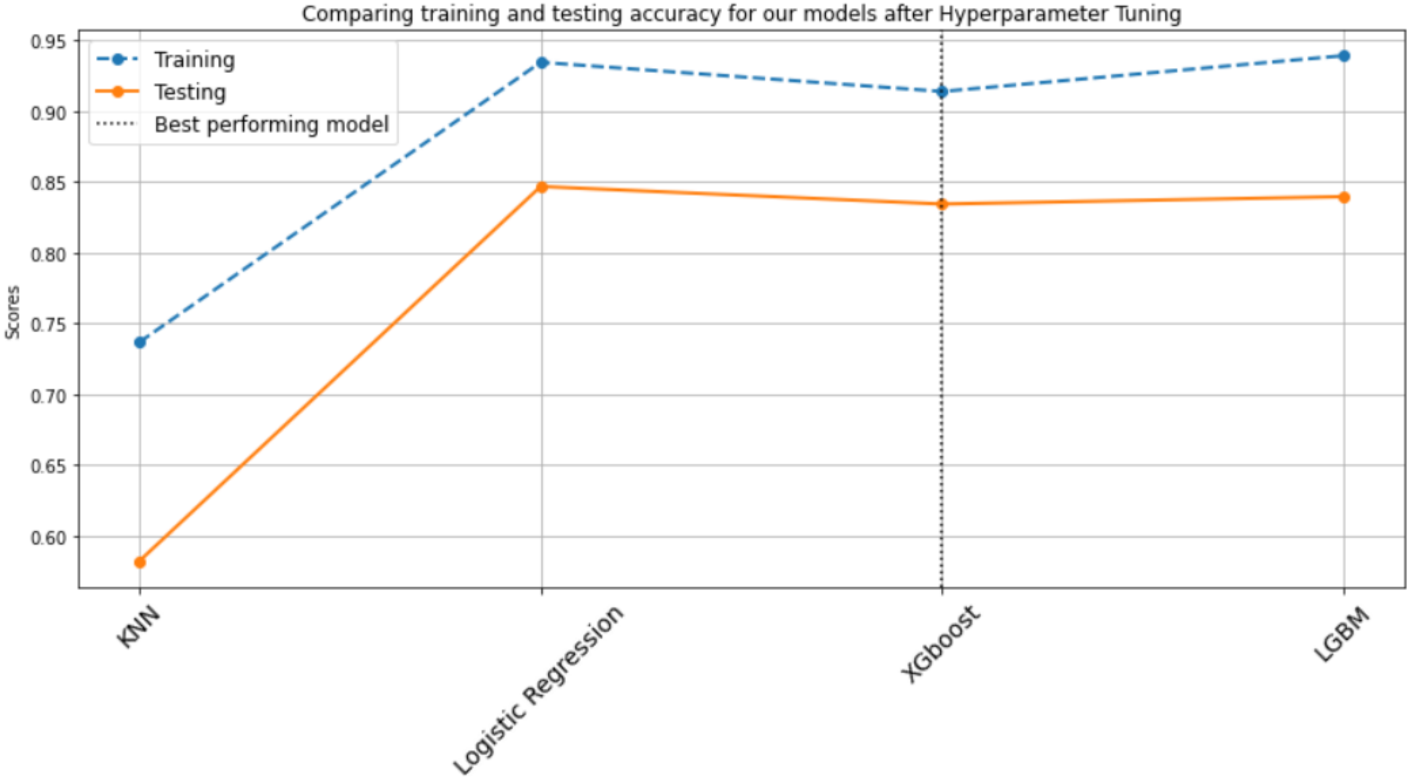

Comparing training and testing accuracy for our models
Scores
1.00
0.95
0.90
0.85
0.80
0.75
XGboost
Decision tree
KNN
Logistic Regression
LGBM
Random Forest
SVM
Multinomial NB
Training
Testing
Best performing model

Hyperparameter Tuning



	Name	Train_accuracy	Test_accuracy	Recall	F1_Score
0	Logistic Regression	0.903416	0.845721	0.828897	0.773050
1	Logistic Regression after Hyperparameter Tuning	0.934628	0.846926	0.822222	0.777583
2	XGboost	0.839652	0.812776	0.762305	0.731567
3	XGboost after Hyperparameter tuning	0.913999	0.834472	0.796856	0.761850
4	LGBM	0.931011	0.845319	0.814320	0.777070
5	LGBM ater Hyperparameter tuning	0.939317	0.839695	0.804348	0.769497
6	Multinomial	0.842867	0.794697	0.949541	0.618372
7	Multinomial after Hyperparameter tuning	0.918687	0.840498	0.882175	0.746326

Conclusion



We can observe that Logistic regression is working good. Its accuracy and recall is more when compared to other models. So we conclude that that Logistic regression is the best model in this sentimental analysis.

Challenges faced

- In the metadata(for clustering), we had only 100 rows and 4 variables to learn. After building the models, we found the silhouette score different number of clusters. We mainly focused on the silhouette score to evaluate the models. We were able to secure only around 0.6 silhouette score from all different models whose optimal number of clusters were to be 6 or 7.

- In the reviews data(for sentiment analysis), to find the feature for the analysis was a tedious task. We made the split for rating and created another feature which is further used for sentimental analysis. But while creating this another feature, we first took 3 partitions for rating (average, good and best). But we did not get good result from the feature with 3 splits. Later another feature was created with 2 splits from rating(1 and 0)

- Overall, we succeeded with good Silhouette score of 0.6 with 6 optimal clusters in clustering.
  And in Sentiment Analysis, we got the train and test accuracy as good as 0.9 and 0.84.

References

- Analytics Vidhya
- GeeksforGeeks
- Medium

Thank You