Unsupervised Machine Learning

Zomato Restaurant Clustering and Sentiment Analysis

By
Sarthak Arora

Contents

Problem Statement

The Project focuses on analyzing the Zomato restaurant data. You have to analyze the sentiments of the reviews given by the customer in the data and made some useful conclusion in the form of Visualizations. Also, cluster the Zomato restaurants into different segments.

The Analysis also solves some of the business cases that can directly help the customers finding the Best restaurant in their locality and for the company to grow up and work on the fields they are currently lagging in.

This could help in clustering the restaurants into segments. Also the data has valuable information around cuisine and costing which can be used in cost vs. benefit analysis Data could be used for sentiment analysis. Also the metadata of reviewers can be used for identifying the critics in the industry.

Data summary

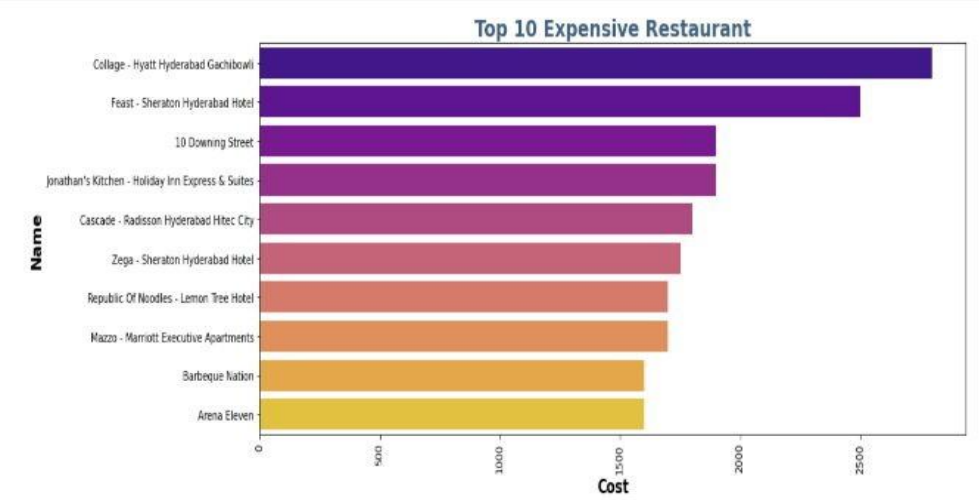
Zomato Restaurant names and Metadata

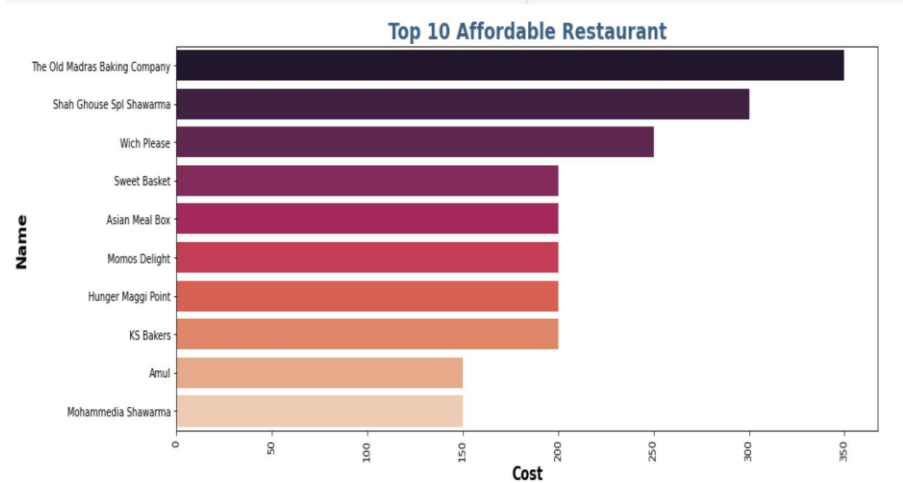
1. **Name** :Name of Restaurants
2. **Links** :URL Links of Restaurants
3. **Cost** :Per person estimated Cost of dining
4. **Collection** :Tagging of Restaurants w.r.t. Zomato categories
5. **Cuisines** :Cuisines served by Restaurants
6. **Timings** :Restaurant Timings

Zomato Restaurant Reviews

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

Exploratory Data Analysis

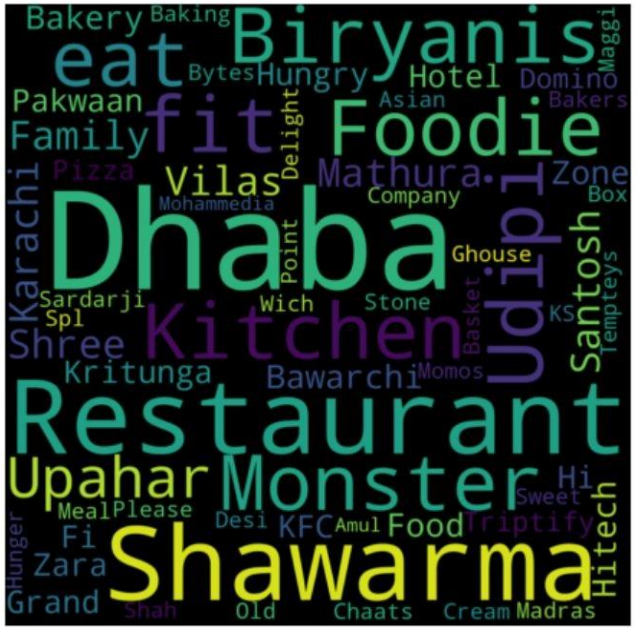




- Finding out the most expensive and most affordable restaurants can help a lot according to different pocket sizes

Exploratory Data Analysis(contd)

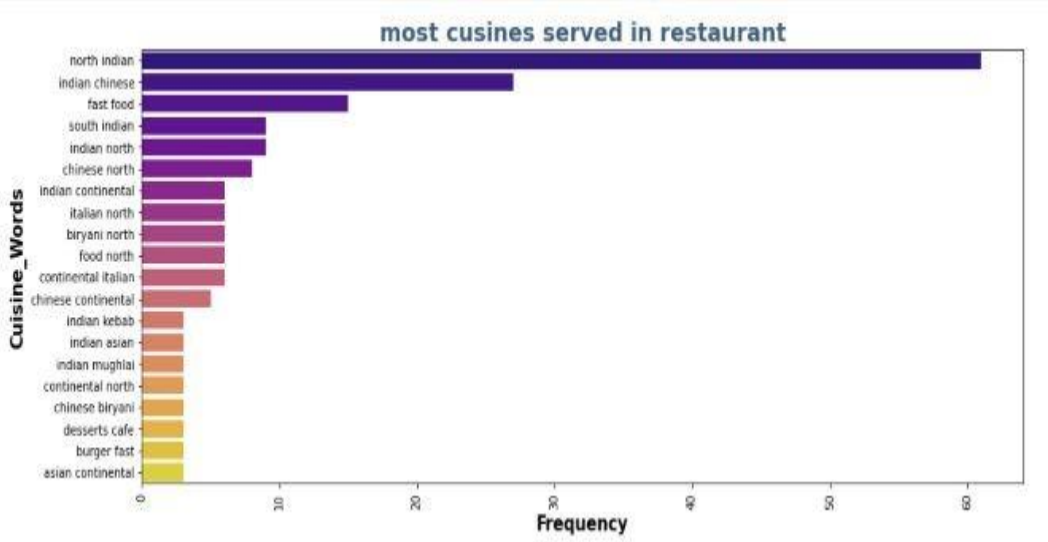
Word cloud for expensive restaurants



Word cloud for affordable restaurants

Exploratory Data Analysis(contd)

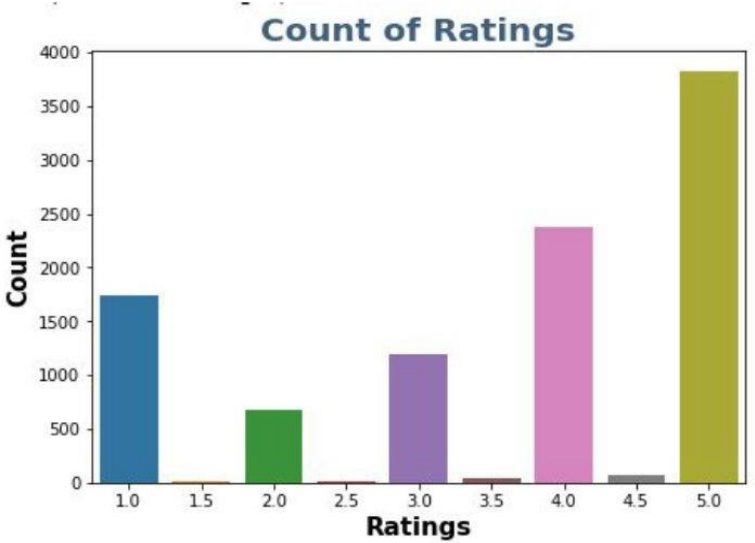
Most cuisines served

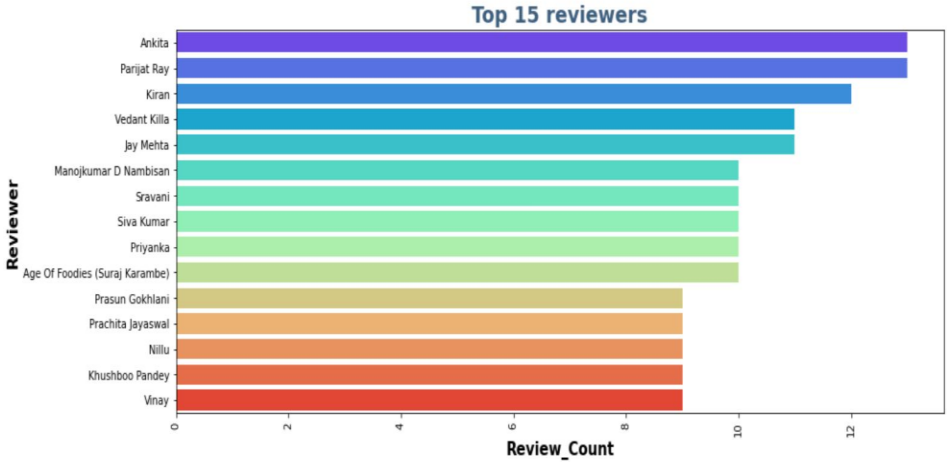


Most cuisines served word cloud

- North-Indian being the most served cuisines followed by the Indian Chinese.

Exploratory Data Analysis(contd)

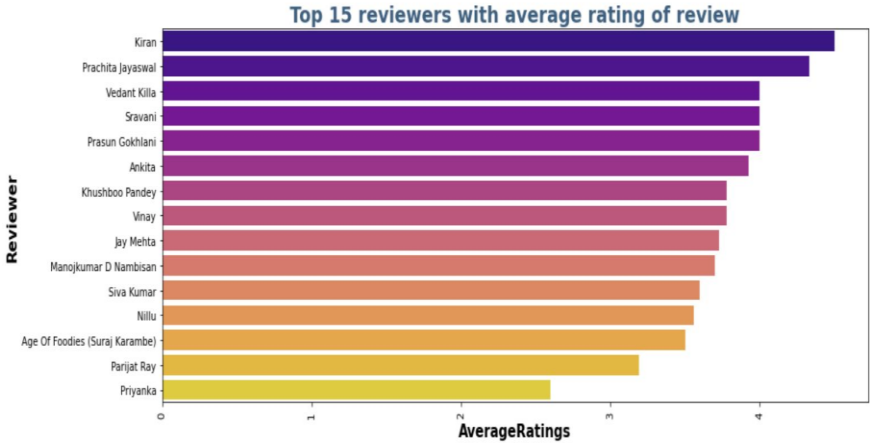




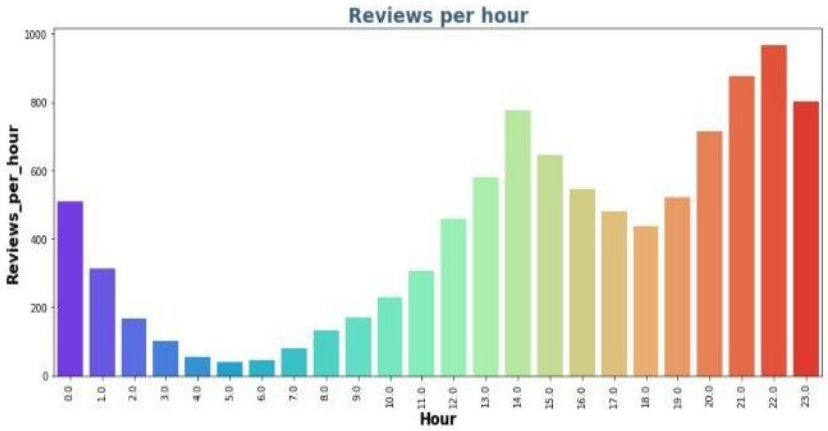
- Ratings with 5 have more in count
- Ankita has reviewed the most when compared to the others

Exploratory Data Analysis(contd)

TOP average rating by the reviewers



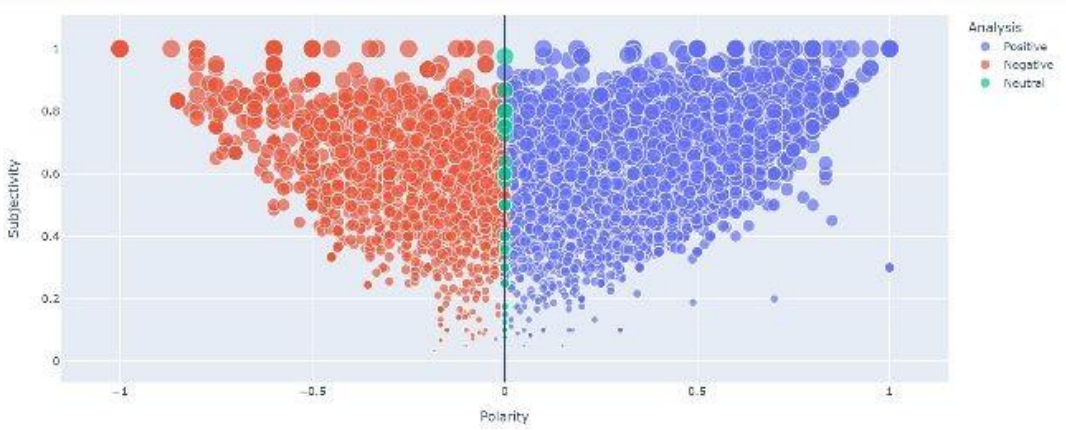
Reviews per hour



- Kiran is the most satisfied customer it seems as he has nearly 5 star rating average
- Reviews are high at the time of 22.00 hrs

Sentiment Analysis

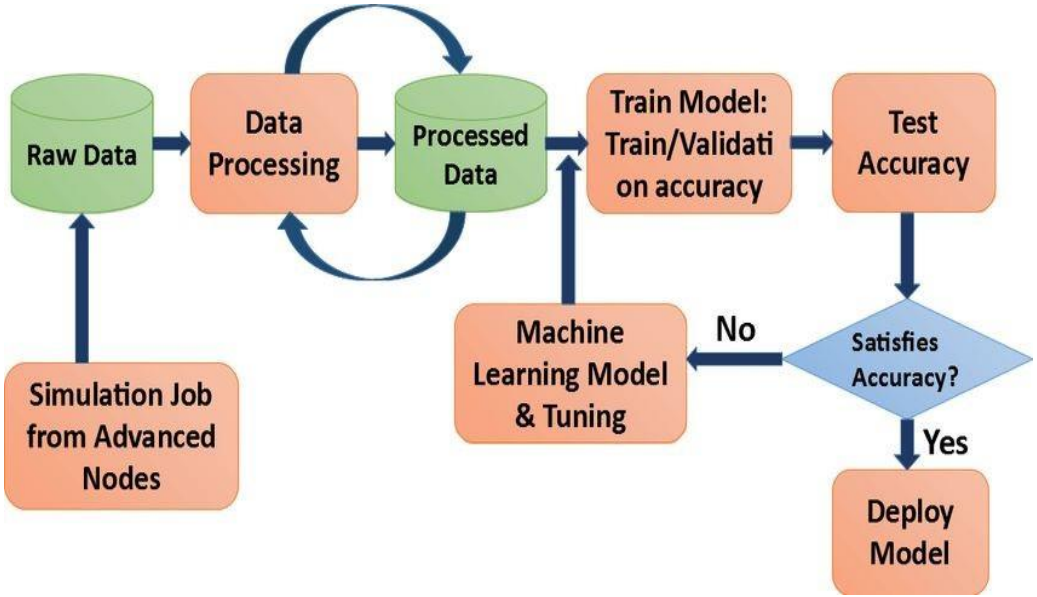
- After completing the necessary text processing part, which contained removing punctuation, Removing stop words & Lemmatization, we move towards Sentiment Analysis



- The subjectivity column that showcases the sentiment is visualized above, where lite purple being ***Positive***, red being ***Negative*** and green being ***neautral***

Models performed

- Multinomial Naive Bayes
- Random Forest Classifier
- XGB Classifier
- Support Vector Classifier



Models performance

Multinomial Naive Bayes

The classification report on the train data is :

	precision	recall	f1-score	support
0	1.00	0.82	0.90	2461
1	0.06	1.00	0.11	28
accuracy			0.82	2489
macro avg	0.53	0.91	0.51	2489
weighted avg	0.99	0.82	0.89	2489

Train accuracy is: 0.8365706630944407
Test accuracy is: 0.823222177581358

Random Forest Classifier

The classification report on the train data is :

	precision	recall	f1-score	support
0	1.00	0.81	0.90	2487
1	0.00	1.00	0.01	2
accuracy			0.81	2489
macro avg	0.50	0.91	0.45	2489
weighted avg	1.00	0.81	0.90	2489

Train accuracy is: 0.8171466845277964
Test accuracy is: 0.8127762153475291

Models performance

XGB Classifier

The classification report on the train data is :

	precision	recall	f1-score	support
0	0.97	0.95	0.96	2071
1	0.76	0.86	0.81	418
accuracy			0.93	2489
macro avg	0.87	0.90	0.88	2489
weighted avg	0.94	0.93	0.93	2489

Train accuracy is: 0.9880776959142665
Test accuracy is: 0.9369224588188028

Support Vector Classifier

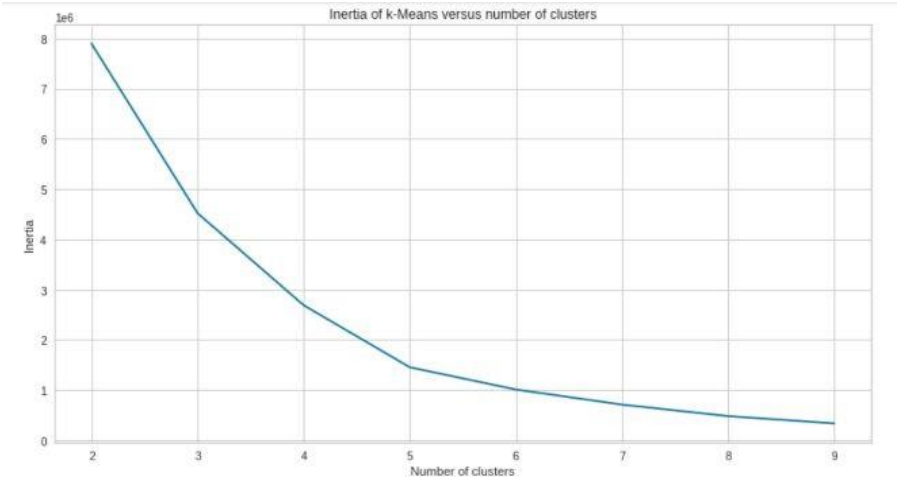
The classification report on the train data is :

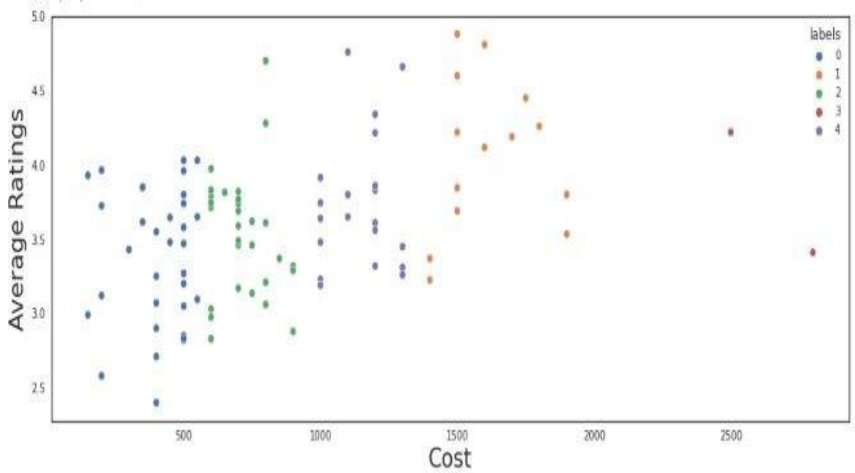
	precision	recall	f1-score	support
0	0.99	0.93	0.96	2145
1	0.69	0.93	0.79	344
accuracy			0.93	2489
macro avg	0.84	0.93	0.87	2489
weighted avg	0.95	0.93	0.94	2489

Train accuracy is: 0.9961152042866711
Test accuracy is: 0.9188429087987143

Clustering

K-Means Clustering



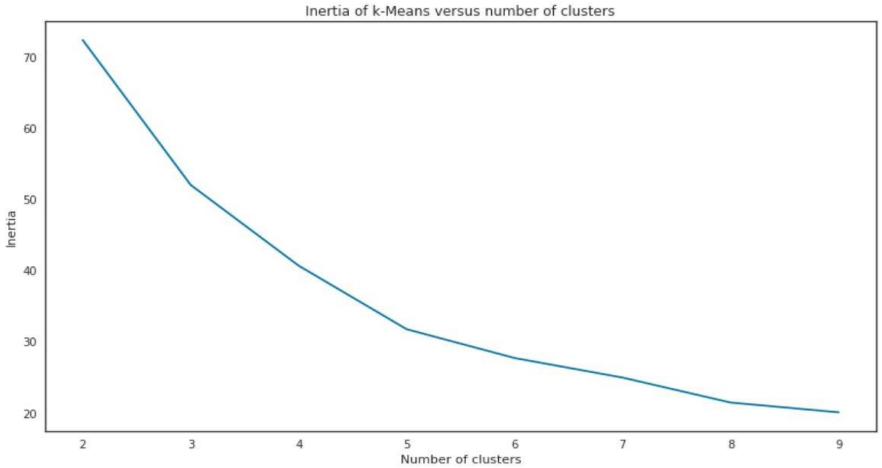


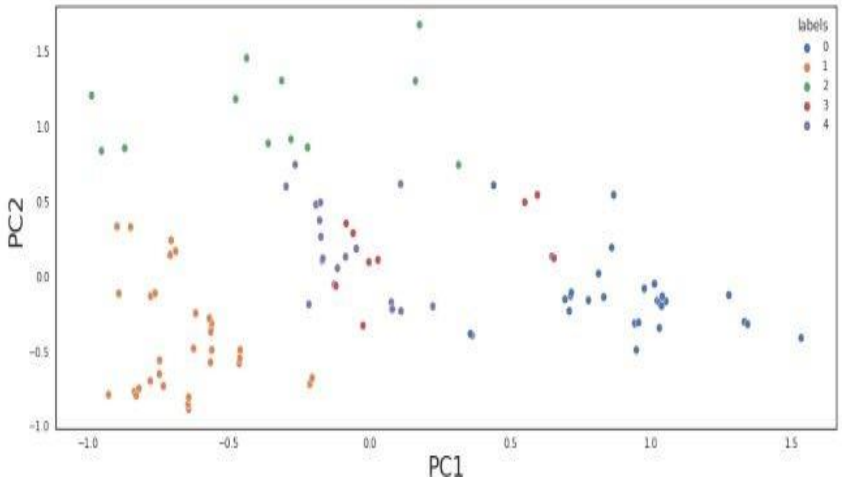
- According to the elbow curve we should have 5 clusters for the best results

- 5 clusters on the average rating and the cost

Clustering(contd)

PCA - Principal Component Analysis





- According to the elbow curve we should have 5 clusters for the best results using PCA.
- 5 clusters on the average rating and the cost using PCA

Top 3 Cuisines in 5 clusters K-Means

```
Top cuisines in cluster 0
 northindian    16
chinese         9
fastfood        8
dtype: int64

Top cuisines in cluster 1
 northindian    11
continental     6
asian           5
dtype: int64

Top cuisines in cluster 2
 northindian    18
chinese        18
biryani        11
dtype: int64

Top cuisines in cluster 3
 asian          2
italian        2
continental    2
dtype: int64

Top cuisines in cluster 4
 northindian    14
chinese         9
italian         7
dtype: int64
```

Model Validation

- As it is clear form the validation table that both XGB and SVM (Classifier) are working exceptionally well than other models.
- So we can choose between any one of them for the production

	Model_Name	Training_accuracy	Test_accuracy
0	MultinomialNB	0.8371	0.8232
1	Random Forest	0.8140	0.8107
2	XGB	0.9880	0.9369
3	Support Vector Machine	0.9961	0.9188

Conclusion

- The most popular cuisines are the cuisines which most of the restaurants are willing to provide. The most popular cuisines in Hyderabad are North Indian, Chinese, Continental, and Hyderabadi.
- The cheapest is the food joint called Mohammedia Shawarma and the costliest restaurant is Collage – Hyatt Hyderabad Gachibowli.
- Sentiment Analysis was done on the reviews and a model was trained in order to identify negative and positive sentiments.
- SVM and XGB both performed well and we can choose any one them.
- SVM and XGB are having 0.9188 and 0.9369 of testing accuracy respectively.
- We got best cluster as 5 in K-Means and Principal Component Analysis(PCA).

References

- 1. Machine Learning Mastery
- 2. GeeksforGeeks
- 3. Analytics Vidhya Blogs
- 4. Towards Data Science Blogs
- 5. Built in Data Science Blogs
- 6. Scikit-Learn Org
- 7. Jovian.ai
- 8. Youtube


Thank You!