Capstone Project

Zomato Restaurant Clustering and Sentiment Analysis

By

Girish R

Content

- Introduction
- Problem Statement
- Data Summary
- Approach Overview
- Exploratory Data Analysis
- Modelling Overview
- Conclusion

Introduction

In today's competitive restaurant industry, it is crucial for companies to understand the market and target the right audience in order to drive growth and success. This project aims to provide a comprehensive solution to this problem by utilizing advanced data analytics and machine learning techniques to cluster Zomato restaurants into segments based on various factors such as cuisine and cost. By understanding the market segments, the company will be able to target the right audience and make data-driven decisions. Additionally, the project will also analyze customer sentiment and reviews, which will account for 40% of the project, to extract valuable insights and identify areas for improvement that can increase customer satisfaction. Overall, this project aims to empower the restaurant industry with a thorough understanding of the market and actionable recommendations for driving business growth.

Problem Statement

Create hotel clusters based on cuisines and sentiment analysis of the customer reviews

The problem statement of the project is to help the restaurant industry understand the market and target the right audience to drive growth and success by utilizing advanced data analytics and machine learning techniques to cluster Zomato restaurants into segments based on various factors such as cuisine and cost, and analyze customer sentiment and reviews to extract valuable insights and identify areas for improvement that can increase customer satisfaction.

Data Summary

Zomato Restaurant names and Metadata (clustering)

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

Data Summary

Restaurant: Name of the Restaurant (sentiment analysis)

- Reviewer: Name of the Reviewer
- Review: Review Text
- Rating: Rating Provided by Reviewer
- MetaData: Reviewer Metadata - No. of Reviews and followers
- Time: Date and Time of Review
- Pictures: No. of pictures posted with the review

Pipeline

Data Cleaning → **Data Exploration** → **Modeling**

Understanding and Cleaning

- Null value analysis
- Missing value treatment
- Outlier Treatment

Graphical

- Univariate analysis with visualization
- Bivariate Analysis with visualization

Machine Learning

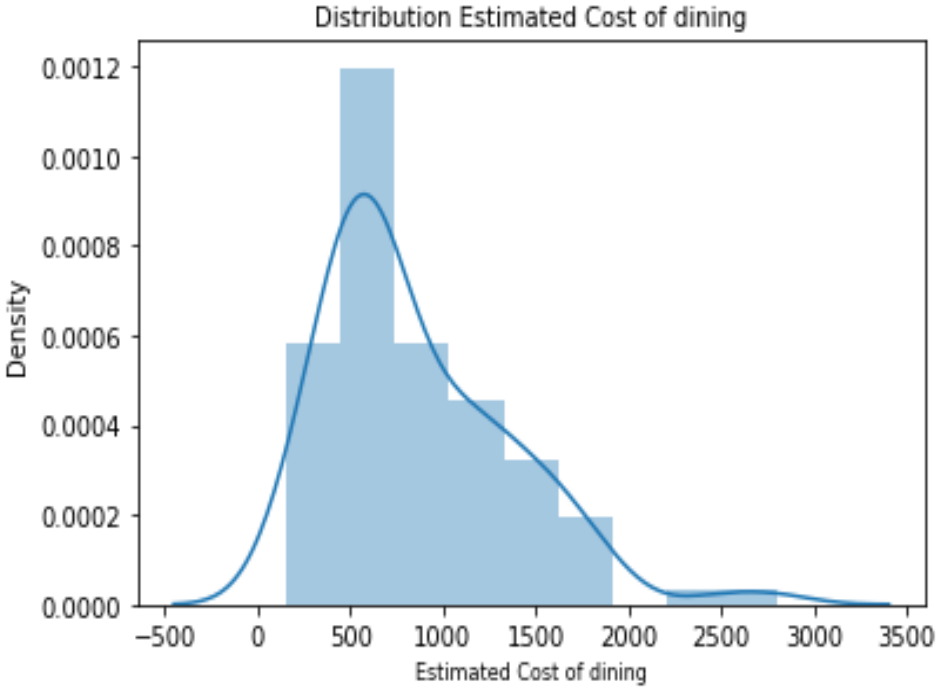
- Clustering
- Sentiment Analysis
- Recommendation System

Basic Exploration

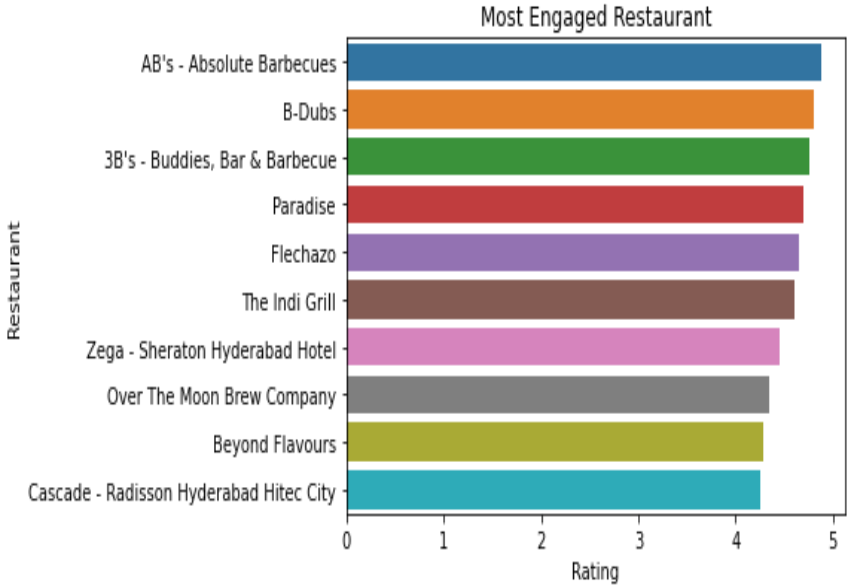
- Data of 105 restaurants.
- Data of 9000 reviews
- 3 years of customer’s reviews
- 0.36 percent null values were present.
- 50 percent of collection data is missing
- Average price of a Restaurant ranges from 150 to 2800

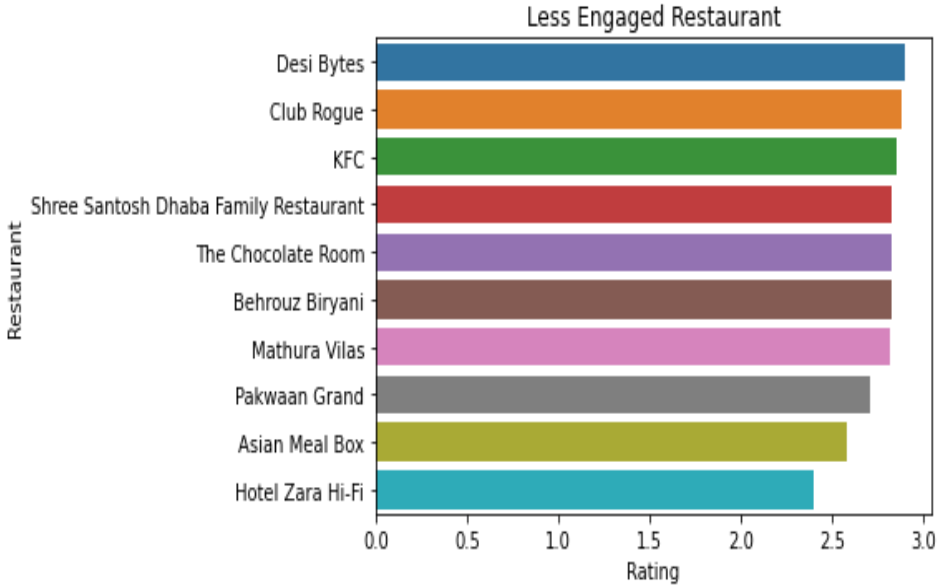
Univariate Analysis

- We utilized a distribution plot to see how the cost of dining varied across all restaurants.
- From the cost distribution curve it can be observed that most of restaurants have cost of dinning in the ranging from Rs.200-Rs.1000 and the median cost is around Rs.700.



Price Point and Maximum Engagement



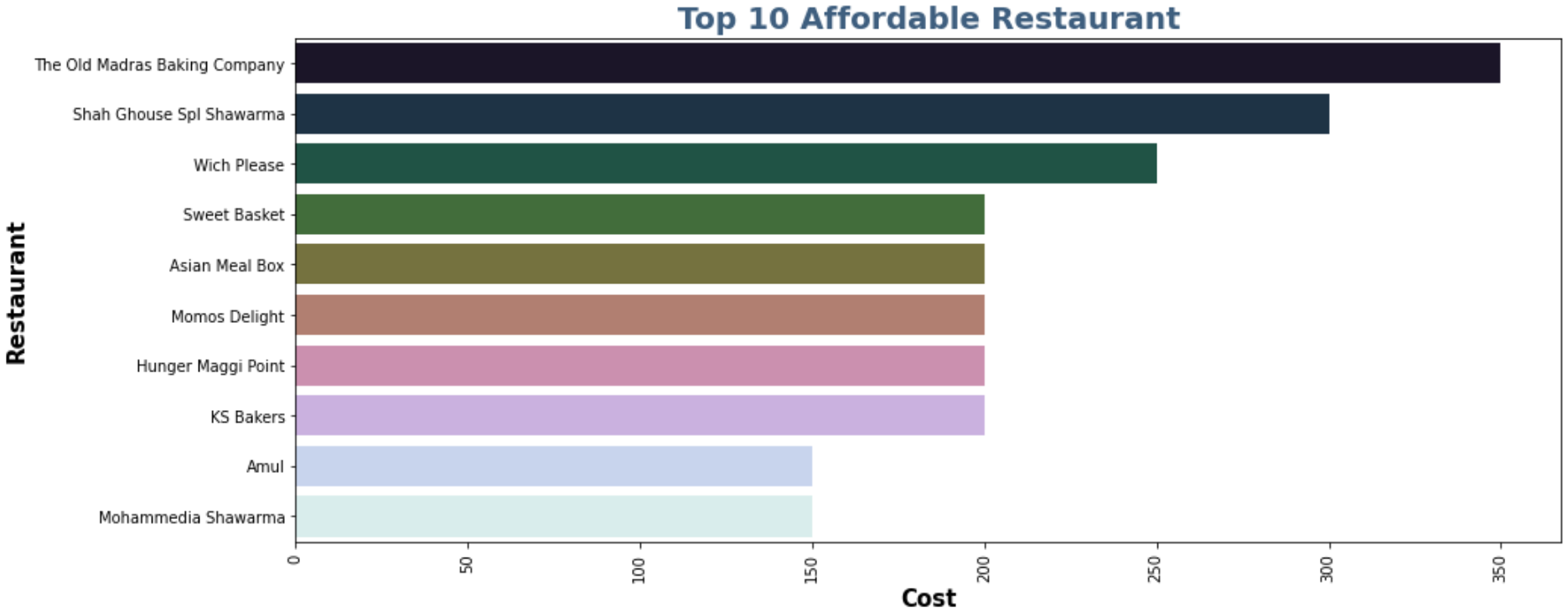


AB's - Absolute Barbecues, show maximum engagement and retention as it has maximum number of rating on average and Hotel Zara Hi-Fi show lowest engagement as has lowest average rating.

10 Most expensive Restaurants

From the above graph we can see the top 10 expensive hotels with Collage-Hyatt Hyderabad Gachibowli being in the top as the most expensive one followed by Feat-Sheraton Hyderabad Hotel

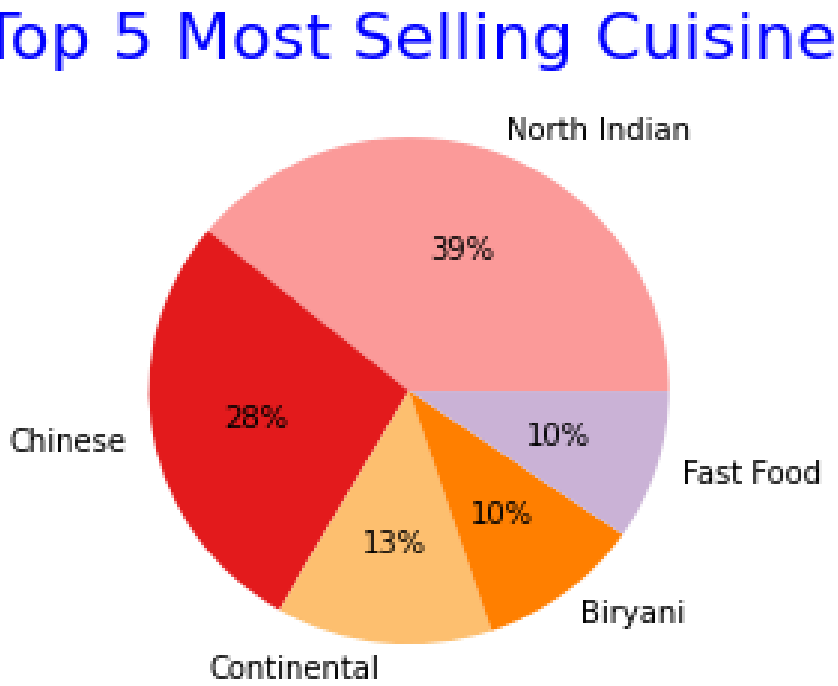
10 most Affordable Restaurants



From the above graph we can see the affordable Restaurants with The old Madras Baking Company at the top which is most affordable according to the given dataset

5 Most Served Cuisines

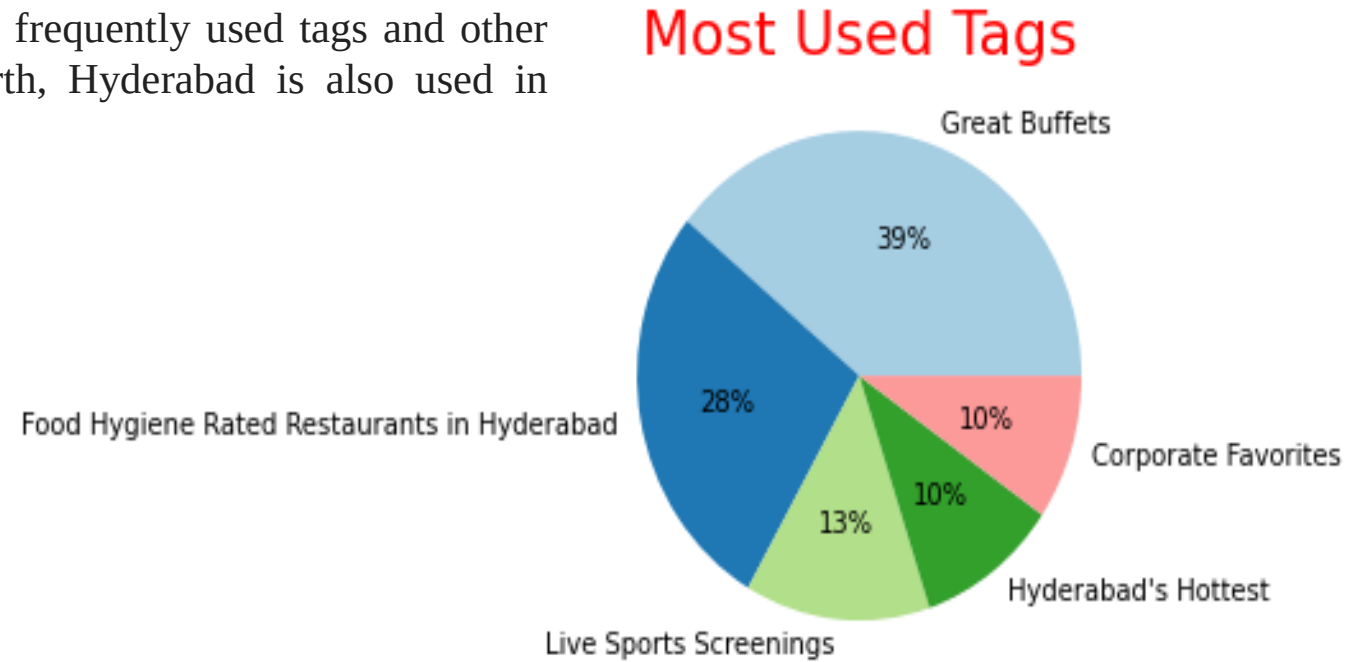
Based on the chart it is clear that most of the hotel sell North Indian food followed by chinese.



Frequent Keyword Used for cuisine

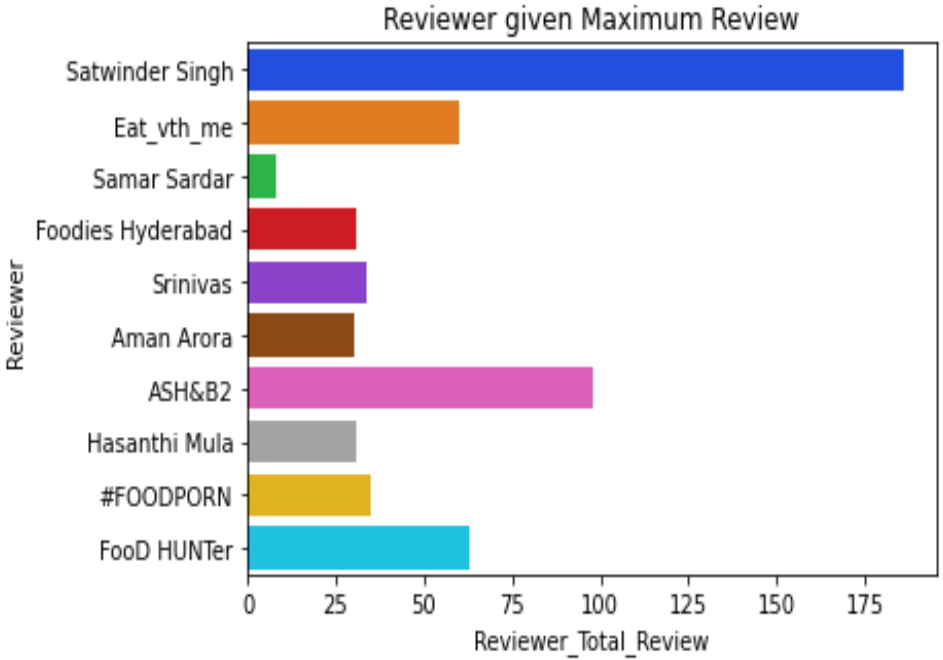
Most used tags for Restaurants

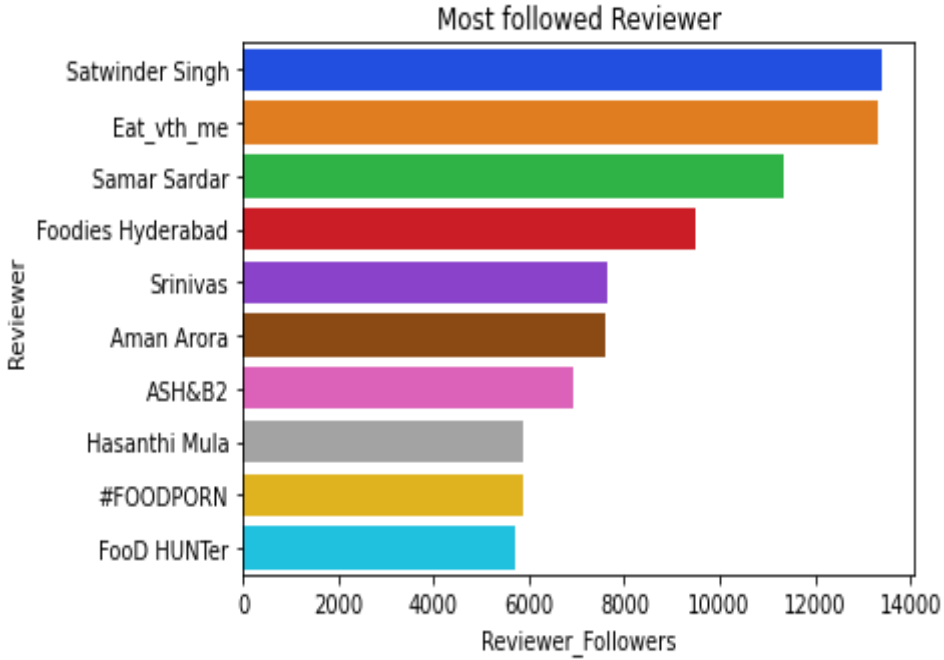
Great Buffets is the most frequently used tags and other tags like great, best, north, Hyderabad is also used in large quantity.



Most Popular Critics

Satwinder singh is the most popular critic who has maximum number of follower and on an average he give 3.5 rating





Modelling Overview

Models Used :

- K-means Clustering
- Hierarchical Clustering
- Sentimental Analysis (Unsupervised)
  - LDA(Latent Dirichlet Allocation)
- Sentimental Analysis (Supervised)
  - Logistic Regression
  - XGBooster
- Recommendation System

Modeling Steps

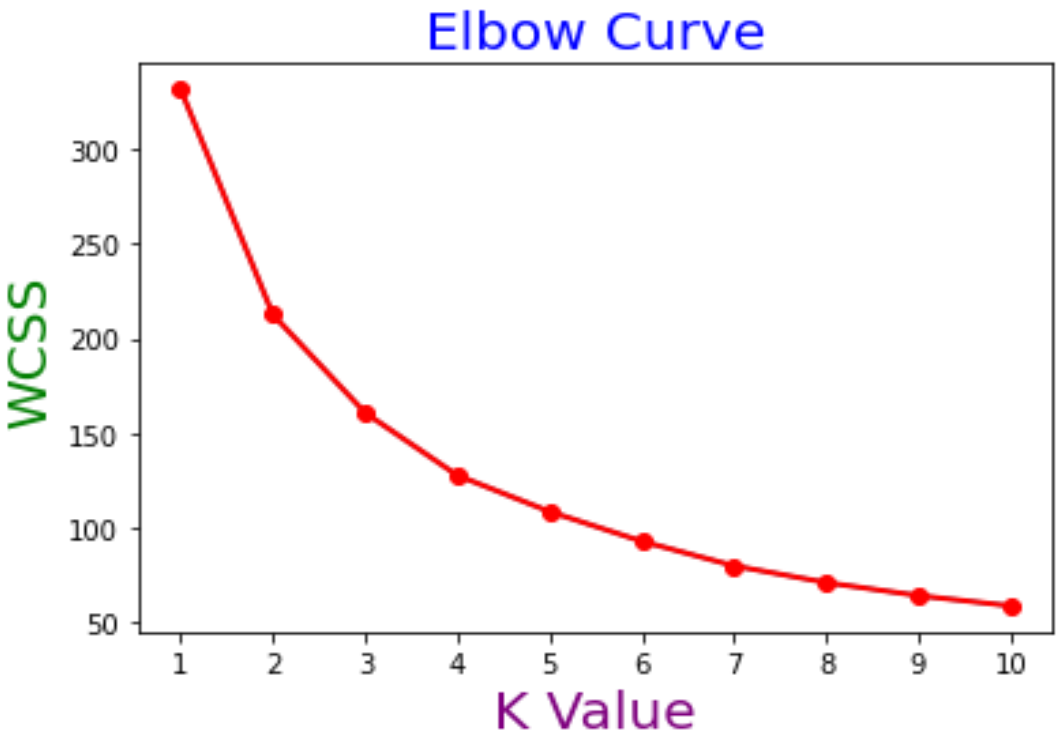
- Feature selection
- Feature engineering
- Feature Extraction
- Train test data split(80%-20%)

- Start with default model parameters
- Hyperparameter tuning
- Measure scores on training & test data

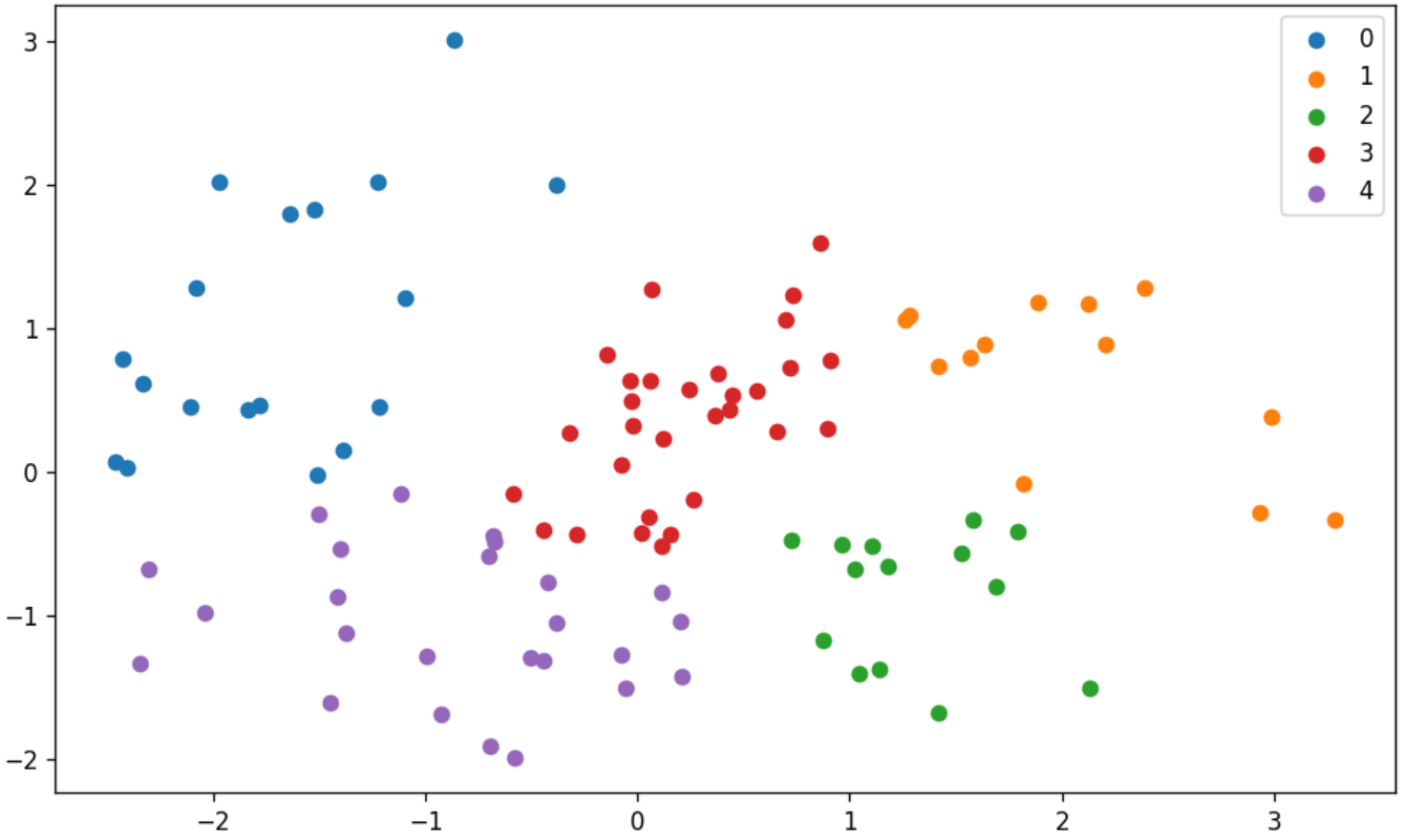
- Model testing
- Compare models

K Means Clustering Plots

ELBOW METHOD



Visualizing the clusters and the datapoints in each clusters



Cuisines in different clusters (K Means)

Cuisine List for Cluster : 4

['Chinese' 'Continental' 'Kebab' 'European' 'South Indian' 'North Indian' 'Biryani' 'Seafood' 'Beverages' 'Healthy Food' 'American' 'Japanese' 'Italian' 'Salad' 'Sushi' 'Mexican' 'Bakery' 'Mughlai' 'Juices' 'Andhra' 'Hyderabadi' 'Spanish' 'Finger Food' 'Thai' 'Indonesian' 'Asian' 'Momos' 'Desserts' 'Cafe' 'Burger' 'Fast Food']

Cuisine List for Cluster : 0

['Biryani' 'North Indian' 'Chinese' 'Asian' 'Mediterranean' 'Desserts' 'Continental' 'Seafood' 'Goan' 'Kebab' 'BBQ' 'European' 'American' 'Italian' 'South Indian' 'Modern Indian' 'Sushi']

Cuisine List for Cluster : 1

['Lebanese' 'Ice Cream' 'Desserts' 'North Indian' 'Fast Food' 'Asian' 'Beverages' 'Bakery' 'Momos' 'Pizza' 'Street Food' 'Arabian']

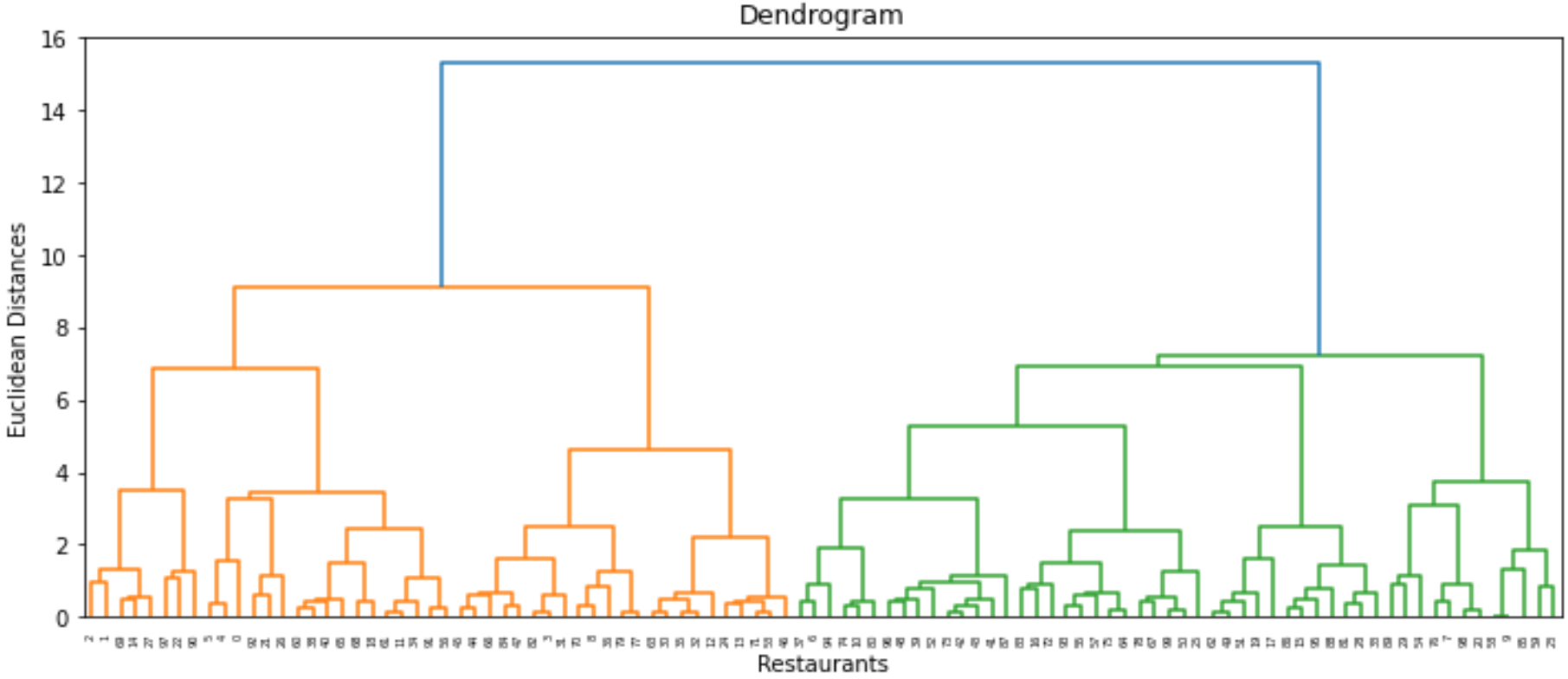
Cuisine List for Cluster : 2

['Street Food' 'North Indian' 'Fast Food' 'Burger' 'Chinese' 'Biryani' 'Mughlai' 'South Indian' 'Desserts' 'Kebab' 'Cafe']

Cuisine List for Cluster : 3

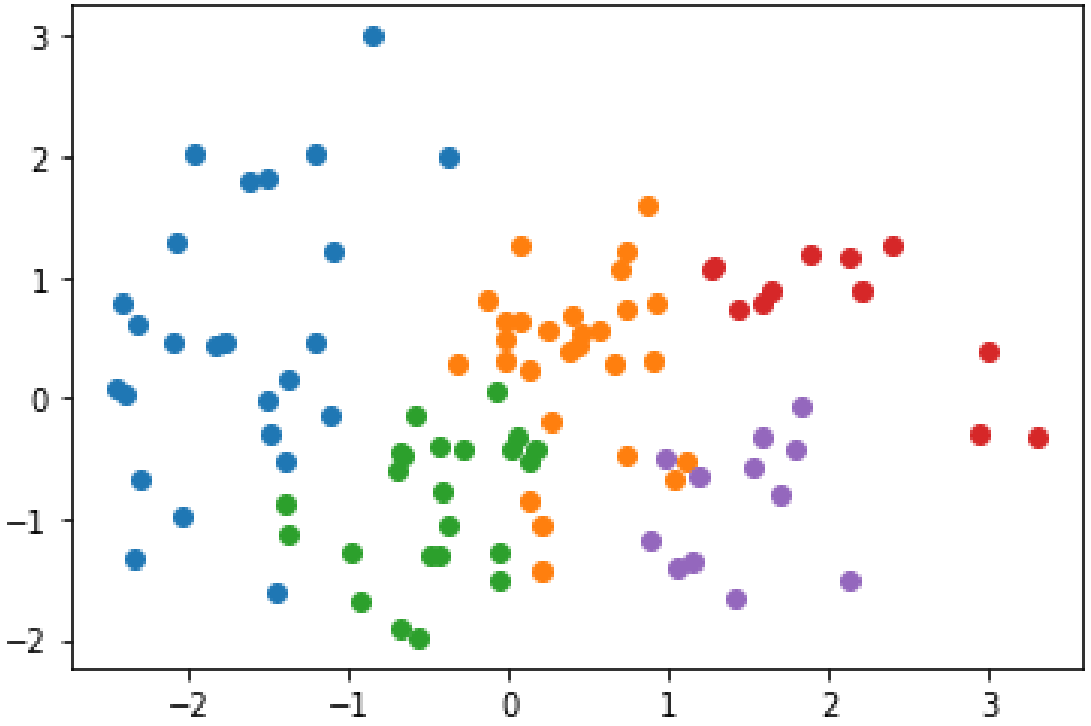
['Continental' 'American' 'Chinese' 'North Indian' 'Italian' 'Finger Food' 'Andhra' 'South Indian' 'Arabian' 'Biryani' 'Cafe' 'Desserts' 'Bakery' 'Fast Food' 'Wraps' 'Asian' 'Momos' 'Hyderabadi' 'Mughlai' 'Beverages' 'Burger' 'Salad' 'North Eastern' 'Seafood']

Hierarchical Clustering



Agglomerative Hierarchical Clustering

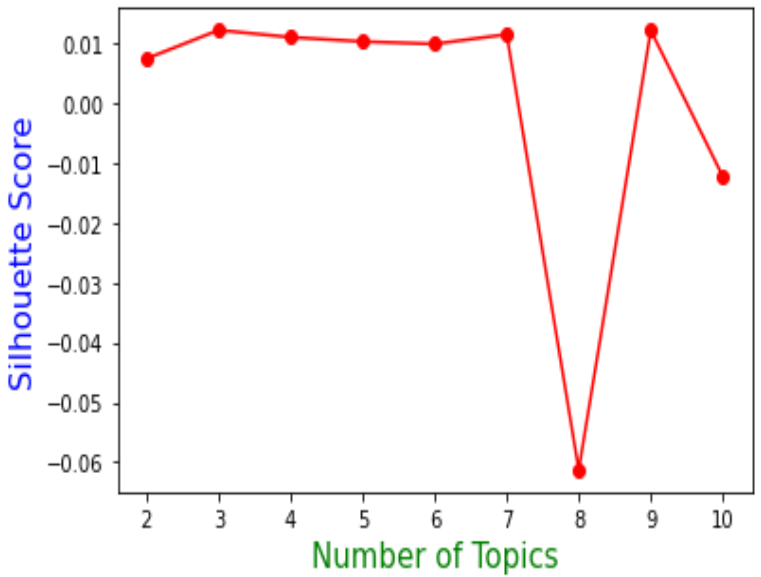
I have used Hierarchical Clustering - Agglomerative Model to cluster the restaurants based on different features. This model uses a down-top approach to cluster the data. I have used Silhouette Coefficient Score and used clusters = 6 and then visualized the clusters and the datapoints within it.



Silhouette Coefficient: 0.247
davies_bouldin_score 1.151

Sentimental Analysis (Unsupervised)

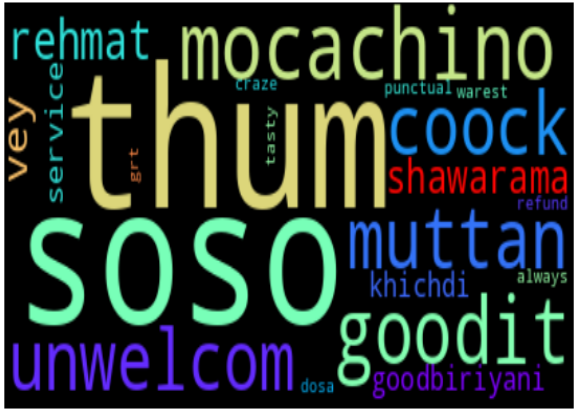
LDA



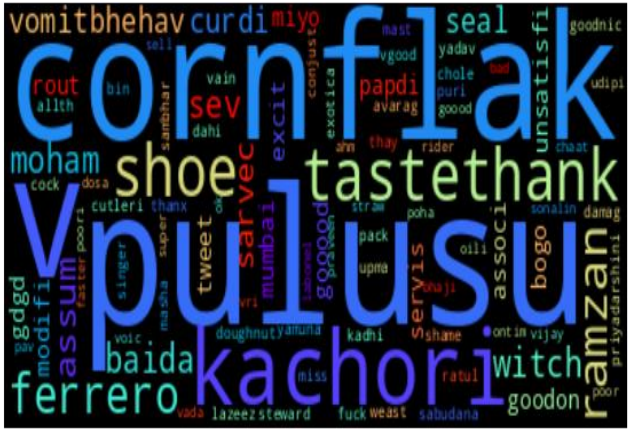

```
Prediction =  1

1.00    406
5.00    146
4.00     60
2.00     54
3.00     48
4.50      1
2.50      1
Name: Rating, dtype: int64
============================
Prediction =  0

5.00    79
1.00    20
4.00    17
3.00    16
2.00     4
Name: Rating, dtype: int64
----------------------------
```

```
Prediction =  2

5.00    2409
4.00    1584
3.00     711
1.00     470
2.00     307
4.50      30
3.50      22
2.50      12
1.50       4
Name: Rating, dtype: int64
============================
Prediction =  3

5.00    1198
1.00     839
4.00     712
3.00     418
2.00     319
4.50      38
3.50      25
2.50       6
1.50       5
Name: Rating, dtype: int64
```

TOP 100 WORDS FOR TOPIC #0

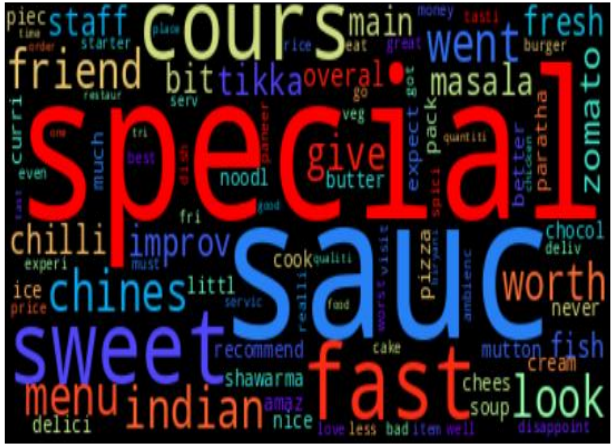


TOP 100 WORDS FOR TOPIC #1



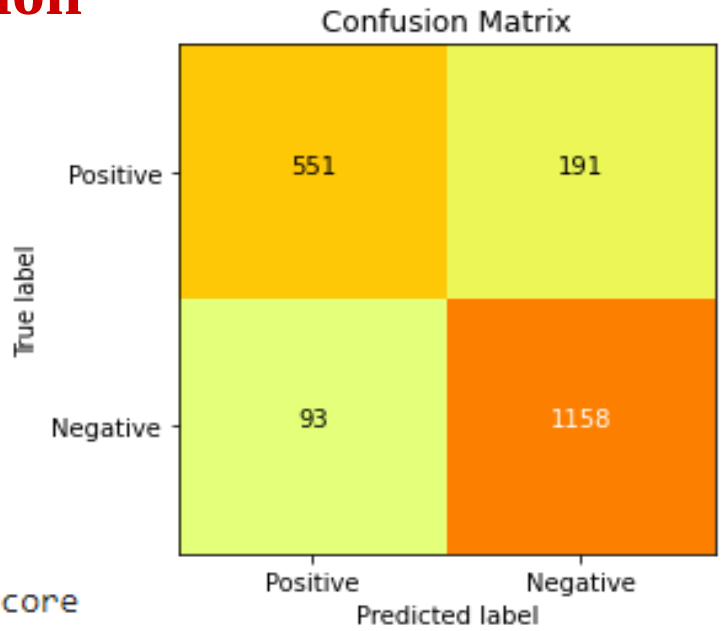
TOP 100 WORDS FOR TOPIC #2

TOP 100 WORDS FOR TOPIC #3



Sentimental Analysis (Supervised)
Logistic Regression

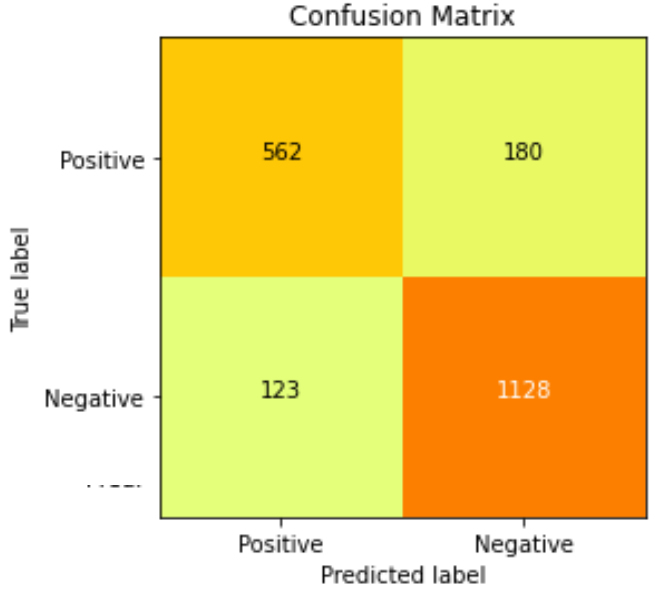
- ✓ 580 instances are labeled as True Positive (correctly predicted as positive)
- ✓ 183 instances are labeled as False Positive (incorrectly predicted as positive)
- ✓ 1148 instances are labeled as True Negative (correctly predicted as negative)
- ✓ 82 instances are labeled as False Negative (incorrectly predicted as negative)




```
[[ 551  191]
 [  93 1158]]
Metric        Score
--------  --------
ROC AUC    0.834124
Precision  0.858414
Recall     0.925659
F1         0.890769
Accuracy   0.857501
```

XGBoost Modelling

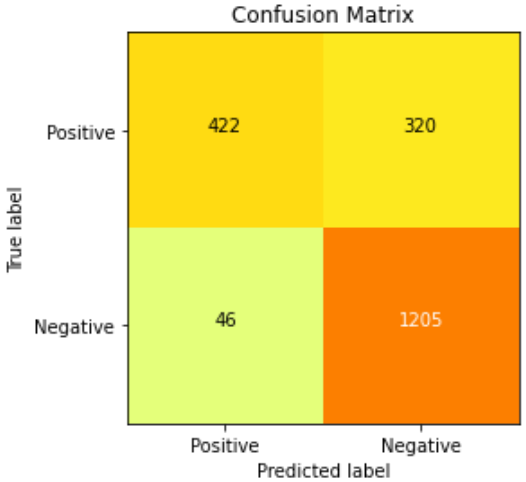
- ✓ 445 instances are labeled as True Positive (correctly predicted as positive)
- ✓ 318 instances are labeled as False Positive (incorrectly predicted as positive)
- ✓ 1153 instances are labeled as True Negative (correctly predicted as negative)
- ✓ 77 instances are labeled as False Negative (incorrectly predicted as negative)




```
[[ 562  180]
 [ 123 1128]]
Metric        Score
---------  --------
ROC AUC    0.829546
Precision  0.862385
Recall     0.901679
F1         0.881594
Accuracy   0.847968
```

Cross- Validation & Hyperparameter Tuning

Logistic Regression

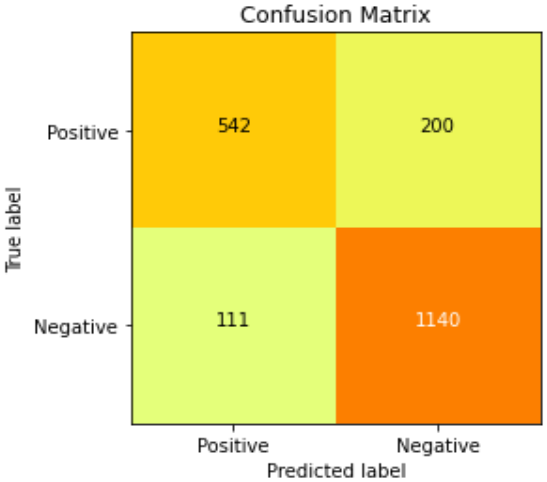



```
[[ 422  320]
 [  46 1205]]
Metric        Score
---------  --------
ROC AUC    0.765981
Precision  0.790164
Recall     0.963229
F1         0.868156
Accuracy   0.816357
```

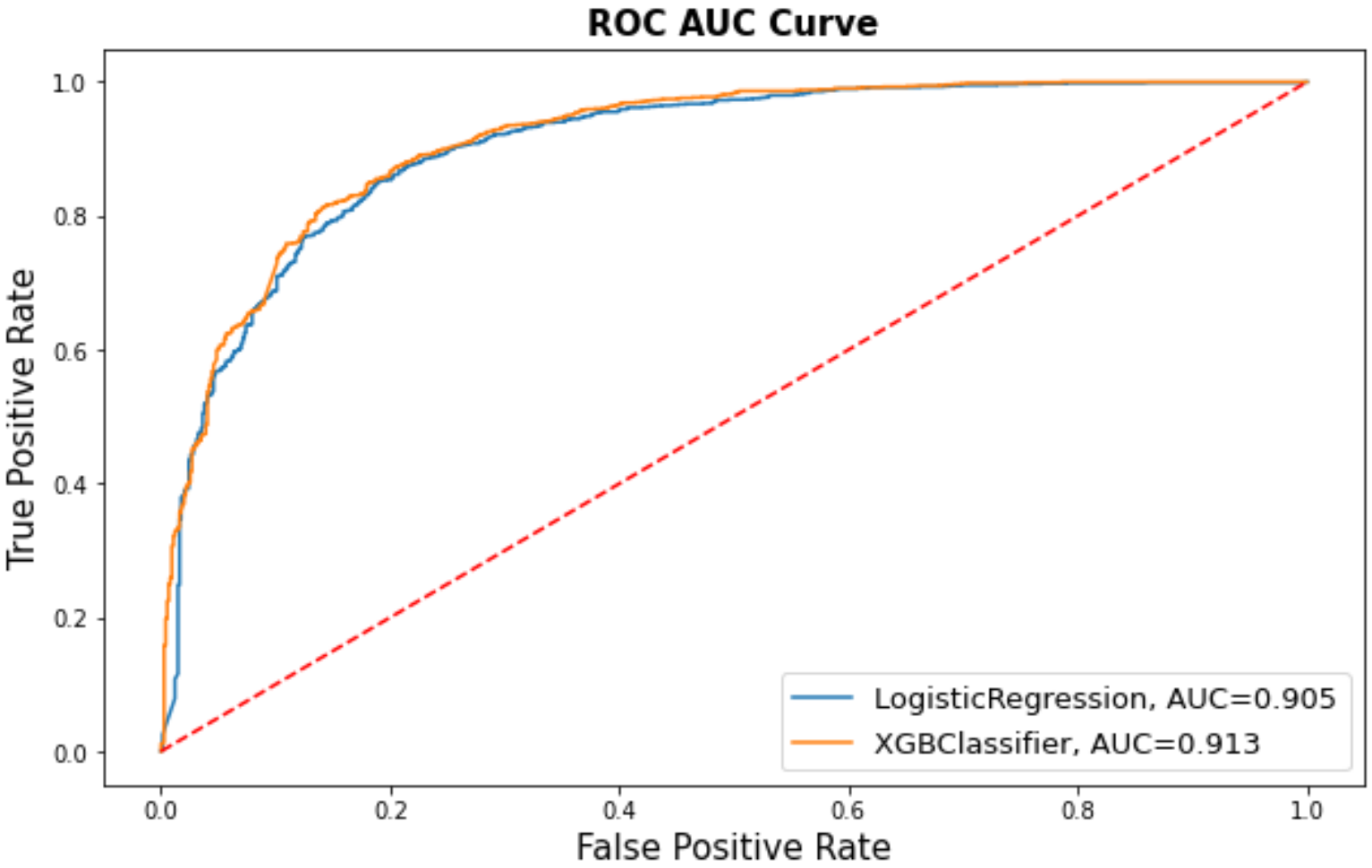
After tunning

- ✓ True Positive is assigned to 566 instances (correctly predicted as positive).
- ✓ False Positive is assigned to 197 instances (incorrectly predicted as positive).
- ✓ 1100 cases are classified as True Negative (correctly predicted as negative).
- ✓ False Negative is assigned to 130 instances (incorrectly predicted as negative).

XGBoost




```
[[ 542  200]
 [ 111 1140]]
Metric        Score
---------  --------
ROC AUC    0.820865
Precision  0.850746
Recall     0.911271
F1         0.879969
Accuracy   0.843954
```


ROC AUC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
LogisticRegression, AUC=0.905
XGBClassifier, AUC=0.913

Recommendation System

Content-based filtering is a recommendation system technique that recommends items to users based on their previous preferences or interactions with items. It works by analysing the attributes of the items and user preferences, and recommending items that have similar attributes.

In a restaurant recommendation system, for example, a user's profile may include information about the restaurant genres they prefer. If a user enjoys a particular Chinese, Italian, or Indian restaurant, the system will suggest other Chinese, Italian, or Indian restaurants to them.

	USER	RESTAURANT	SCORE
73472	Shree	Hitech Bawarchi Food Zone	0.81
13143	Santosh	Karachi Cafe	0.77
23447	Kiran Thota	Al Saba Restaurant	0.87
41470	Naveen Reddy	Tiki Shack	0.73
17390	Amar	Tandoori Food Works	0.86
7241	Khaane_mey_kya_hey By Rony Samuel	Sardarji's Chaats & More	0.64
55792	Paridhi Mehra	American Wild Wings	0.63
19747	Mahesh	eat.fit	0.77
19255	Sriram Reddy	Karachi Cafe	0.80
51981	Poojitha Challagali	Shah Ghouse Spl Shawarma	0.70

Content-based filtering can also be used to suggest items to new users who have yet to interact with the system. The system will recommend items based on their attributes rather than the user's previous preferences in this case.

Conclusion

Using a dataset of customer evaluations for the meal delivery service Zomato, clustering and sentiment analysis were conducted. To comprehend the customer's experience and learn more about their input, this analysis was conducted.

Our analysis of customer evaluations for Zomato utilized clustering and sentiment analysis to gain insights into customer satisfaction levels and identify areas for improvement. By clustering customers into positive and negative groups and using sentiment analysis to classify reviews as positive or negative, we were able to gain a comprehensive understanding of customer feedback. Our findings can be used to inform business decisions and improve the Zomato service.

From the EDA we got to know about,

- AB's - Absolute Barbecues, show maximum engagement and retention as it has maximum number of rating on average and Hotel Zara Hi-Fi show lowest engagement as has lowest average rating.
- Most Expensive Restaurants
- Most Affordable Restaurants
- Great Buffets is the most frequently used tags.

- I have chosen XGBoost model which is hyperparameter optimized for my final prediction.
- As a result of its high regularization, XGBoost is more resistant to overfitting and more adaptable to new data. With XGBoost, a supervised learning system, sentiment labels can be predicted by training on labelled data.
- The ensemble aspect of XGBoost can aid in enhancing sentiment analysis performance by pooling the predictions of various models.
- In sentiment analysis, when the model needs to generalize to new data, XGBoost's regularization can help to reduce overfitting and make the model more robust to unseen data.

Thank You