Capstone Project

Zomato Restaurant Clustering and Sentiments Analysis

Team

Rahul Kumar Soni, Lakdawala Ali Asgar

Content

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

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. Zomato is a site where someone can give a review of a restaurant, how the restaurant is, and someone's opinion about the restaurant.

Problem Statement

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

Data Summary

Zomato Restaurant names and Metadata (clustering)

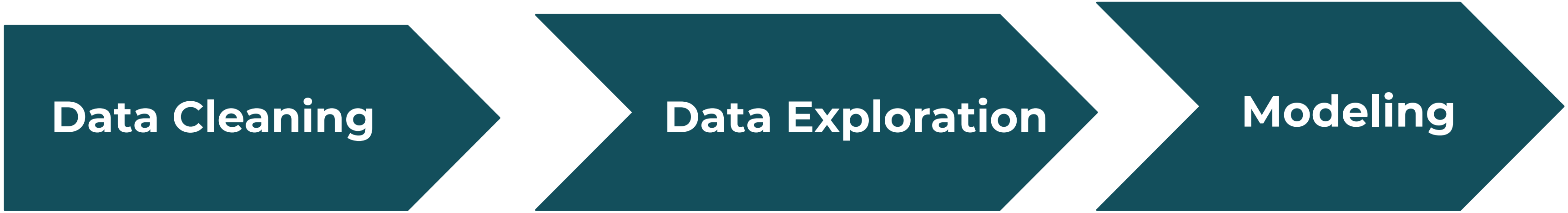
- Name: Name of Restaurants
- Links: URL Links of Restaurants
- Cost: Per person estimated Cost of dining
- Collection: Tagging of Restaurants w.r.t. Zomato categories
- Cuisines: Cuisines served by Restaurants
- Timings: Restaurant Timings Zomato Restaurant reviews

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

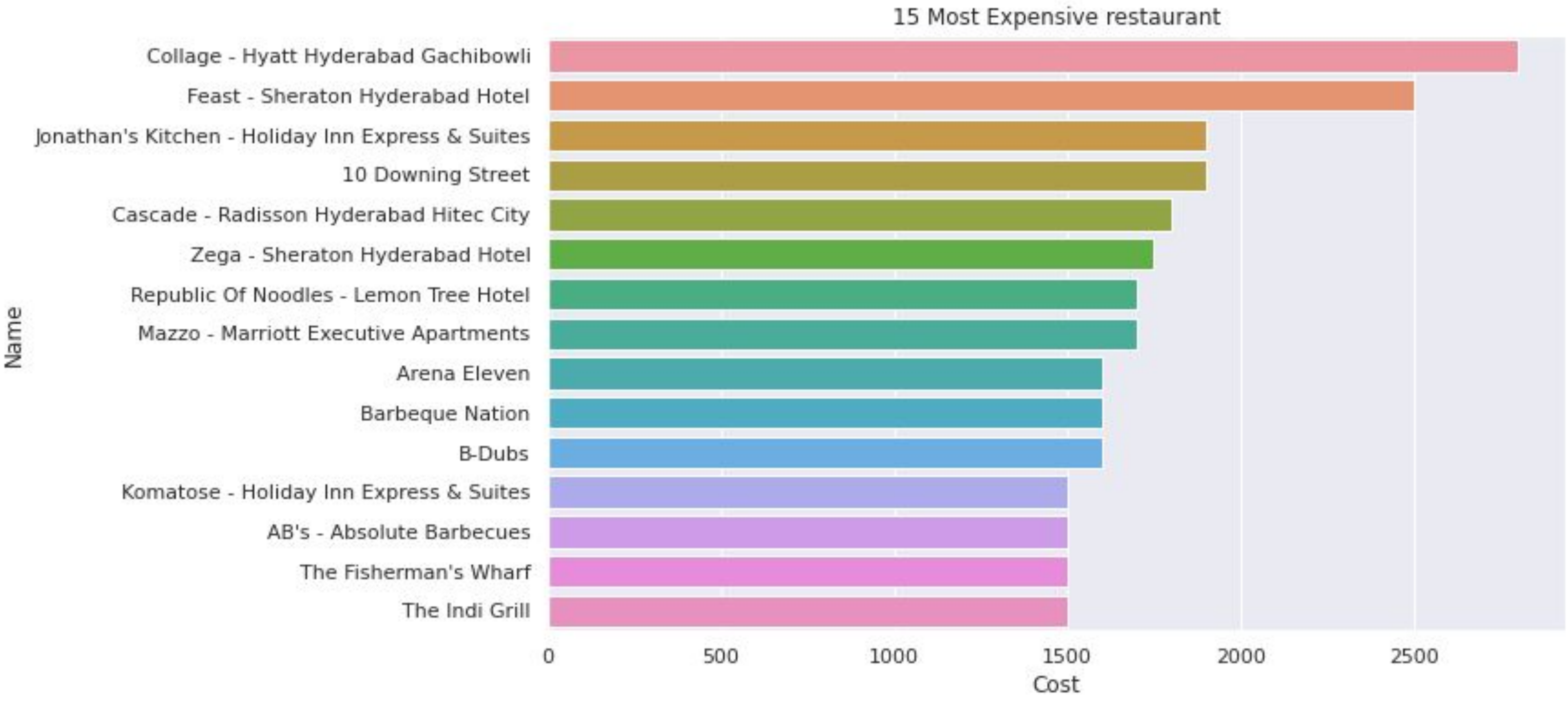
Machine Learning

- Clustering
- Topic Modeling
- Classification

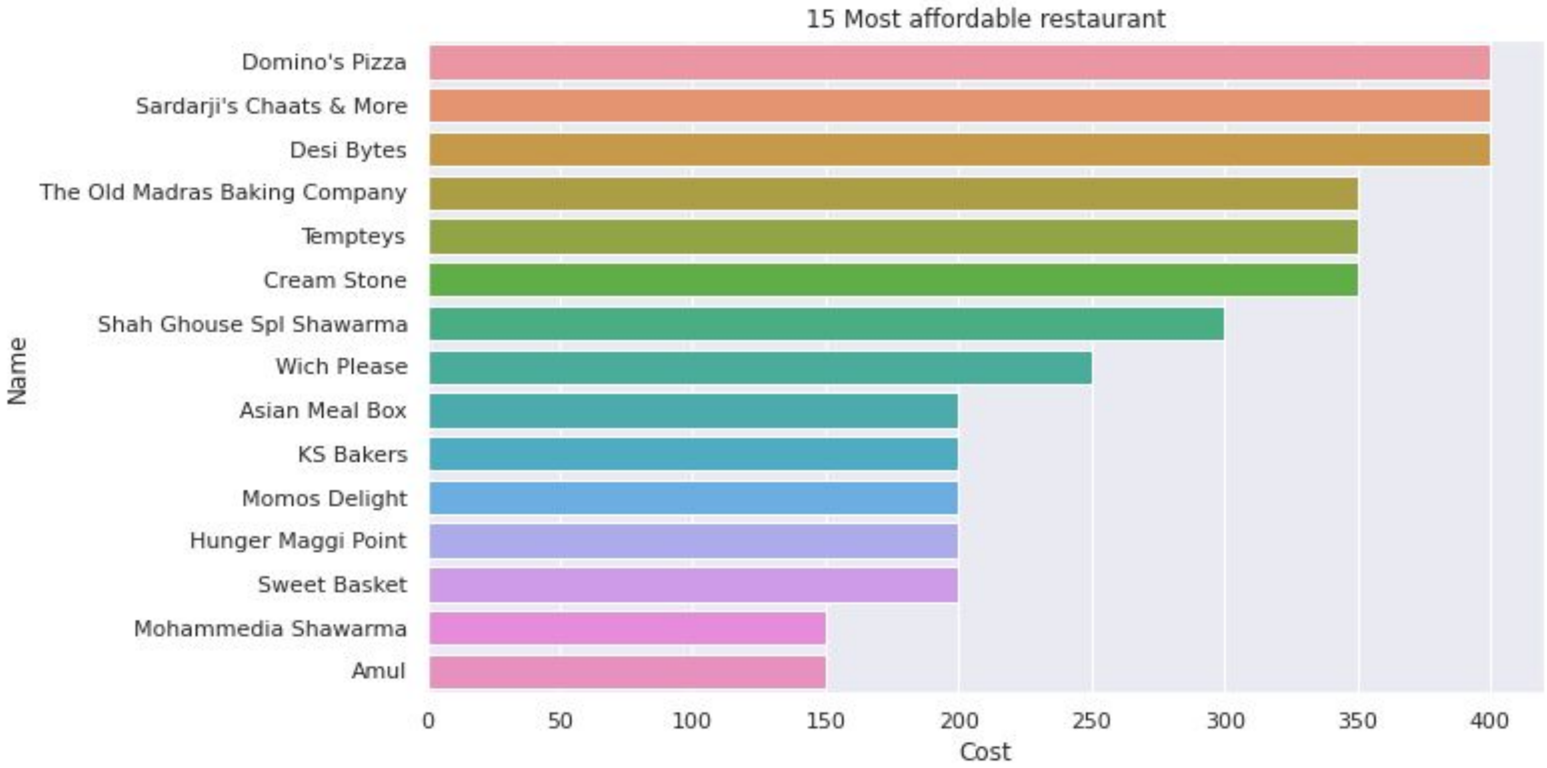
Basic Exploration

- 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 hotel ranges from 200 to 2800

15 Most expensive Restaurants

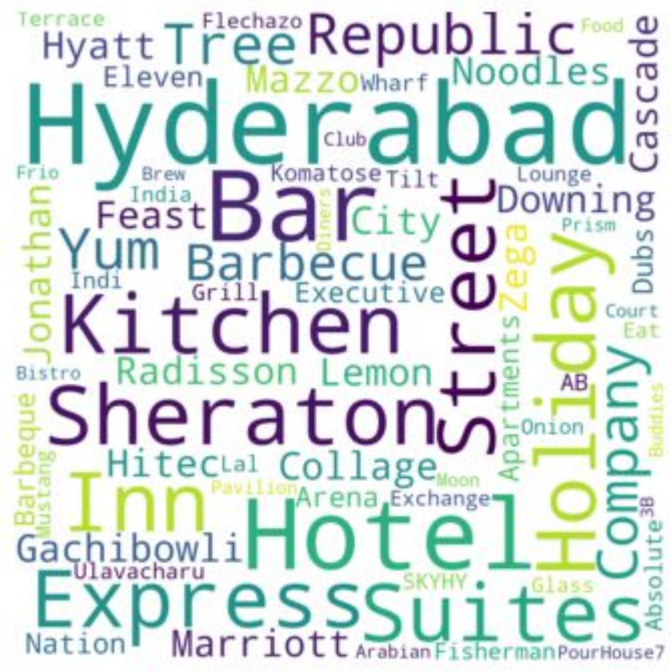


15 most Affordable Restuarents

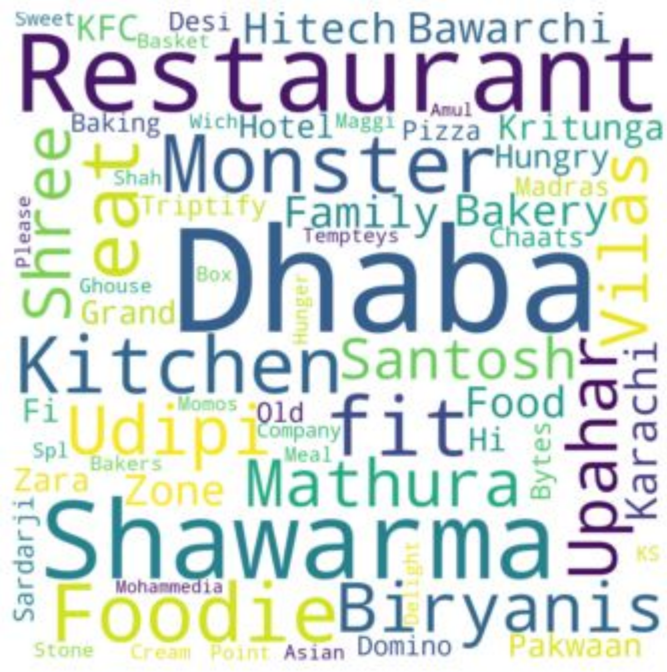


Frequent Keywords Used for Restaurant

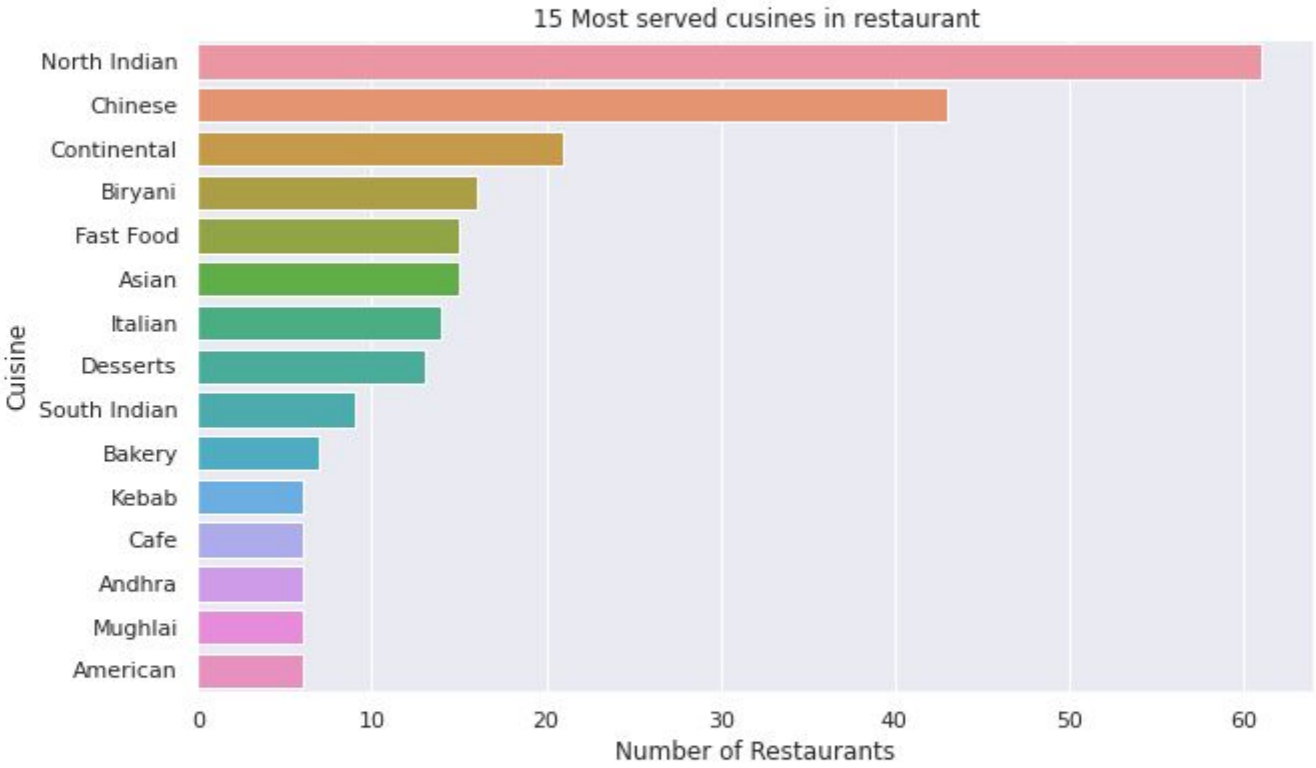
Most Expensive



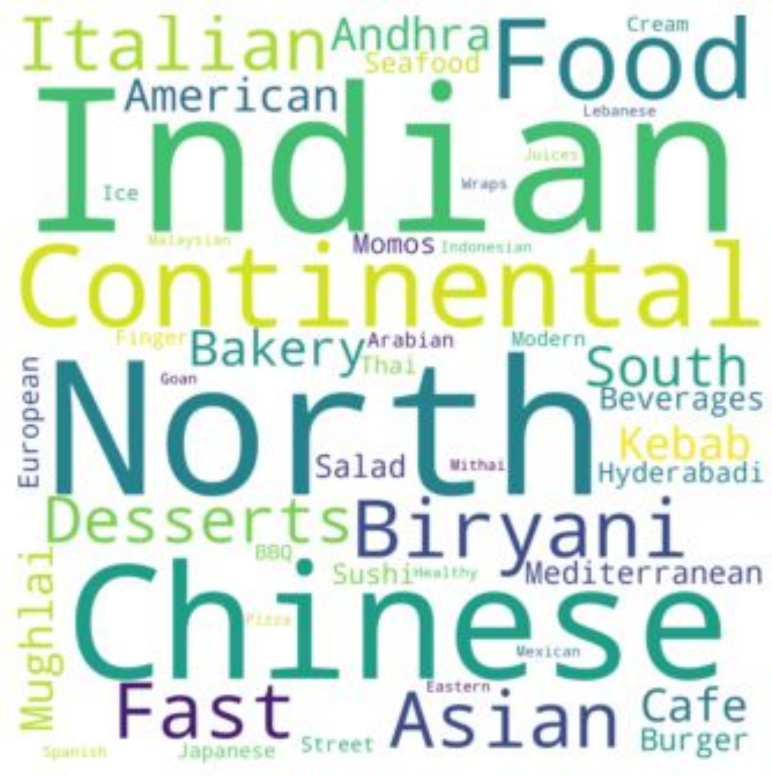
Most Affordable



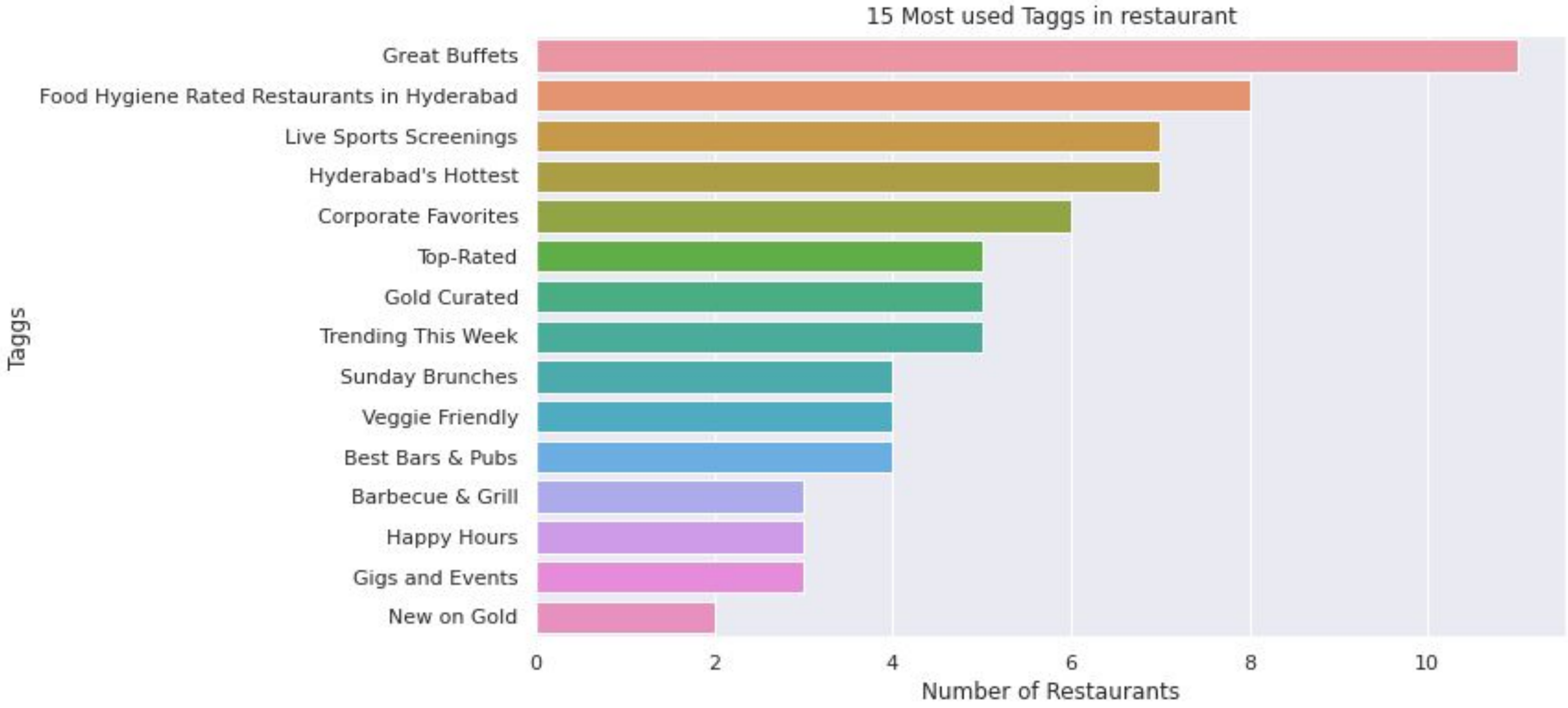
15 Most Served Cuisines



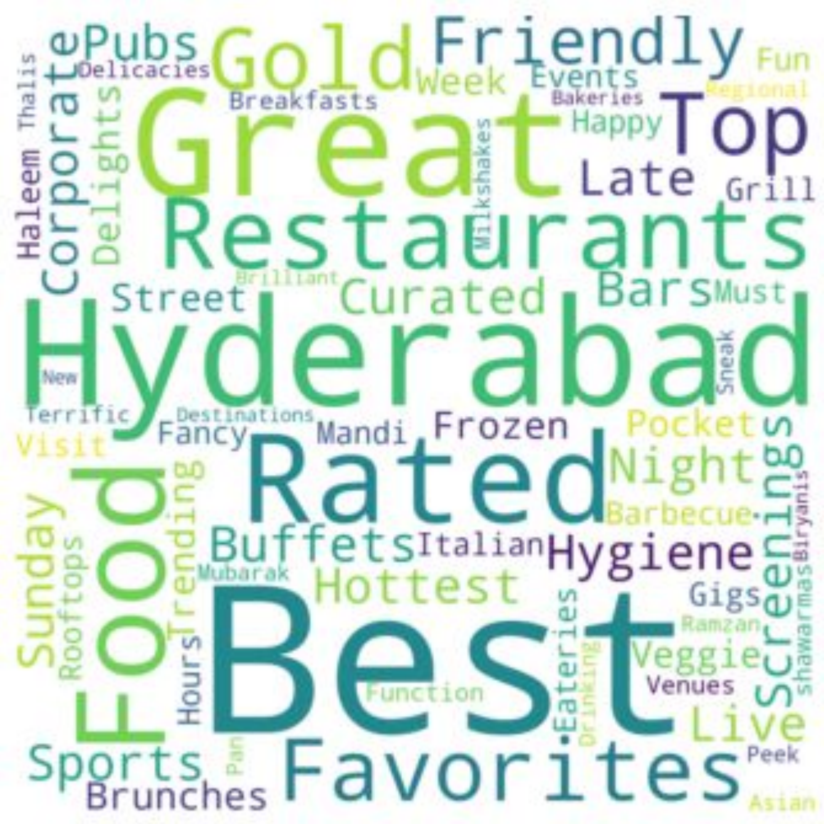
Frequent Keyword Used for cuisine



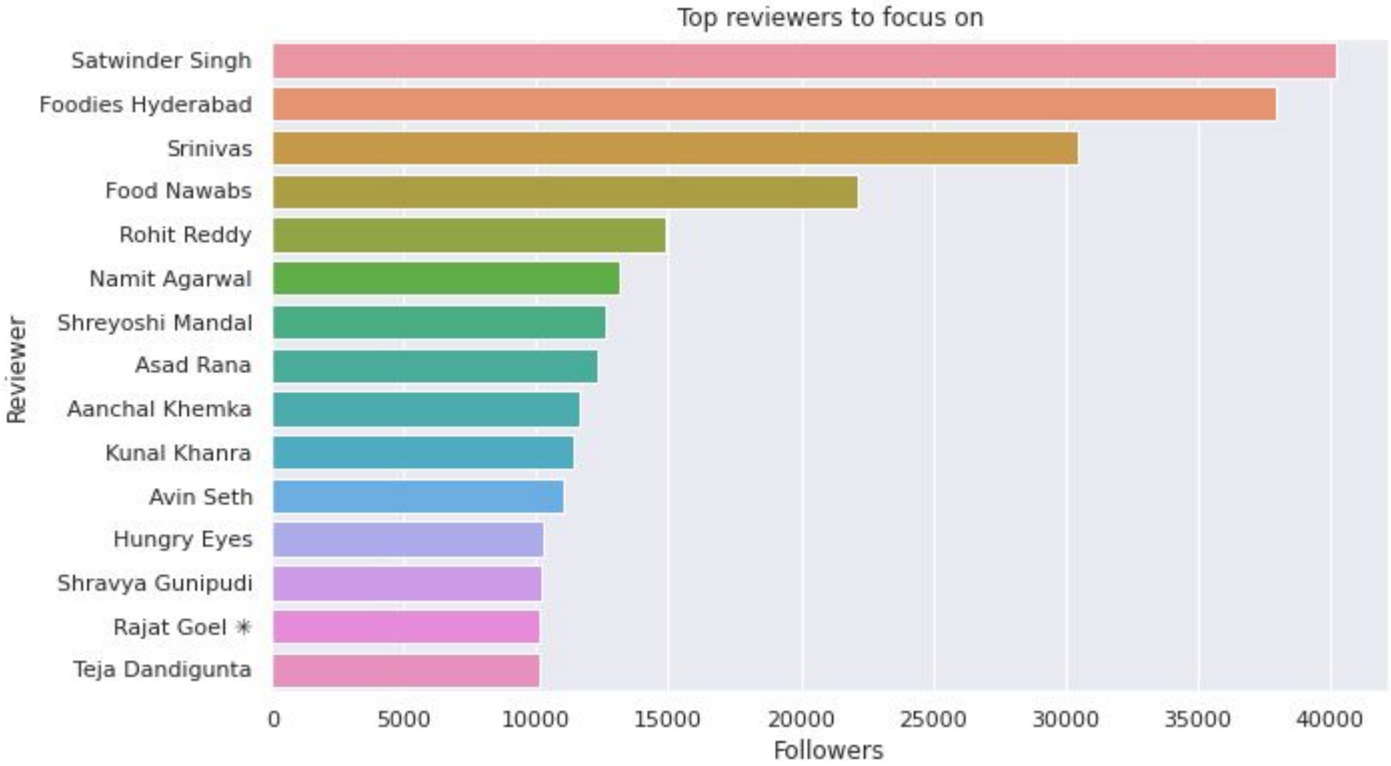
Most used tags for Restaurants



Most used words for Restaurants(Tag)



Food Critics



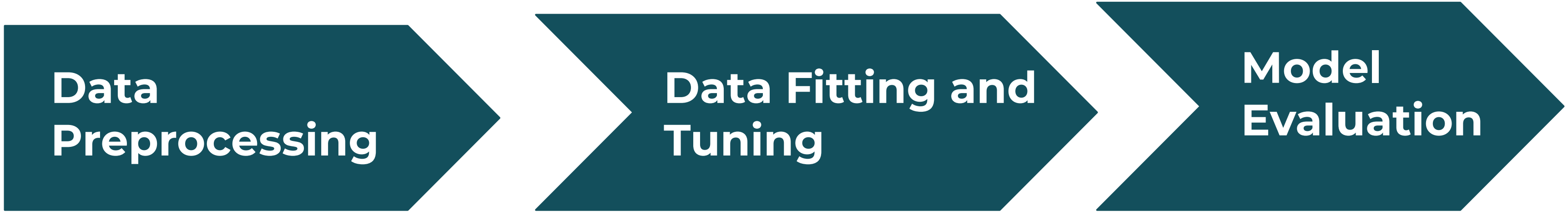
Modeling Overview

Models Used :

- K-means Clustering
- Hierarchical Clustering
- Linear Discriminant Analysis
- Non-negative Matrix Factorization
- Logistic Regression

- Decision Trees
- Random Forest
- Multinomial NB
- XGBoost
- LightGBM

Modeling Steps



Data Preprocessing

- Feature selection
- Feature engineering
- Feature Extraction
- Train test data split(75%-25%)

Data Fitting and Tuning

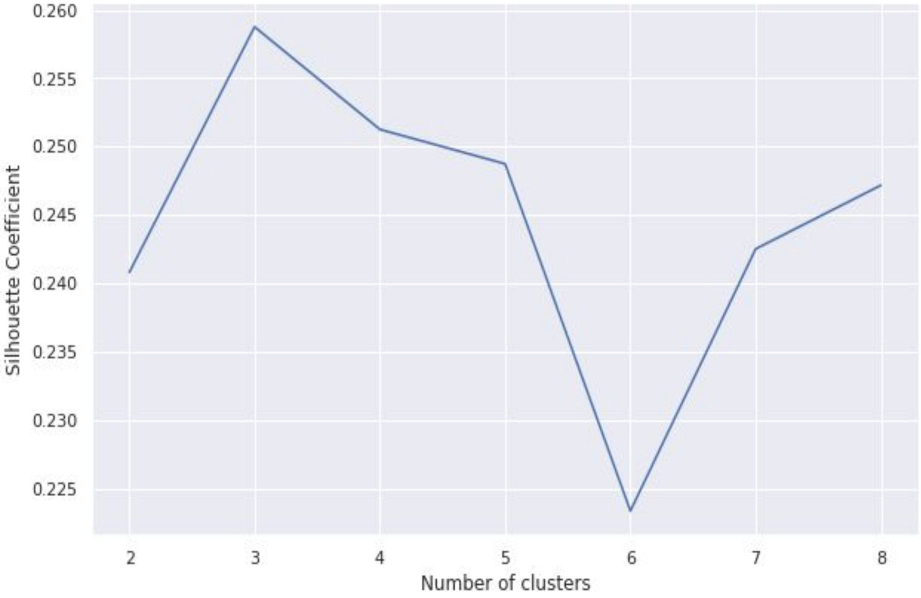
- Start with default model parameters
- Hyperparameter tuning
- Measure scores on training & test data

Model Evaluation

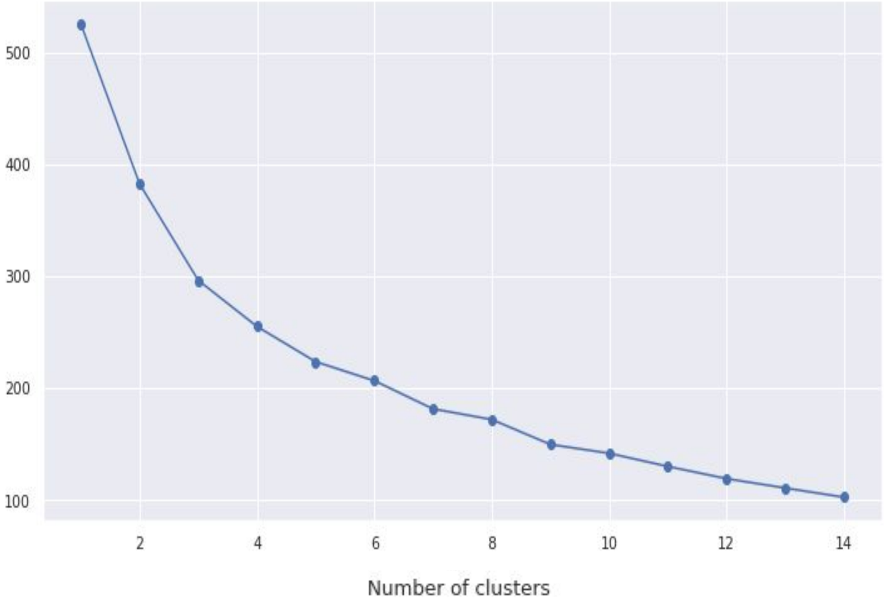
- Model testing
- Compare models

K Means Clustering Plots

Silhouette score



Sum of squares elbow plot



Cuisines in different clusters (K Means)

Cluster 0

'north indian', 'chinese', 'continental', 'mediterranean', 'european', 'seafood', 'biryani', 'hyderabadi', 'american','south indian', 'andhra', 'kebab', 'bbq', 'italian', 'asian','mughlai', 'beverages', 'modern indian', 'desserts', 'spanish', 'japanese', 'salad', 'sushi', 'mexican', 'thai', 'malaysian', 'indonesian', 'goan', 'finger food', 'healthy food'

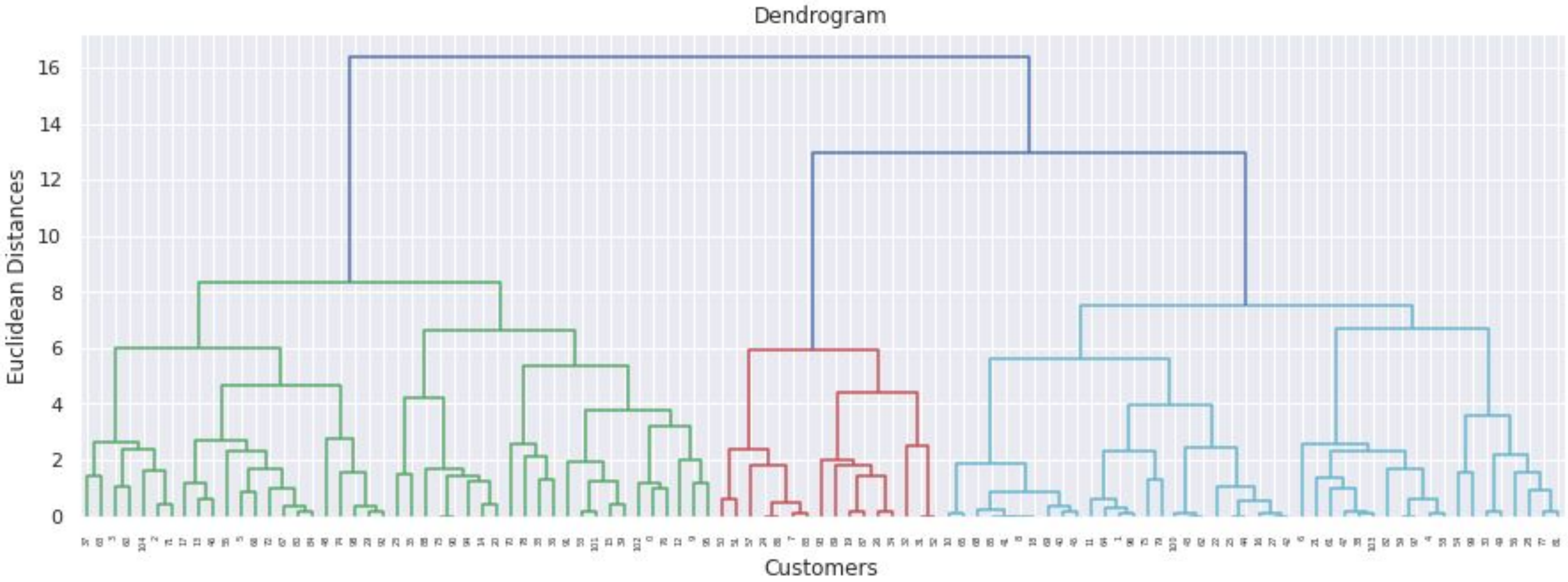
Cluster 1

'ice cream', 'desserts', 'cafe', 'bakery', 'continental','fast food', 'beverages', 'burger', 'biryani', 'north indian','mughlai','juices', 'chinese', 'mithai', 'american', 'wraps'

Cluster 2

'north indian', 'continental', 'american','chinese', 'fast food','salad', 'burger','biryani', 'mughlai','asian','seafood, 'momos',pizza','hyderabadi', 'japanese','sushi', 'finger food','kebab', 'arabian', 'south indian', 'street food', 'lebanese','andhra', 'thai', 'north eastern'

Hierarchical Clustering



Cuisines in different clusters (Hierarchical)

Cluster 0

'north indian', 'chinese', 'continental', 'mediterranean', 'european', 'seafood', 'biryani', 'hyderabadi', 'american', 'south indian', 'andhra', 'kebab', 'bbq', 'mughlai', 'italian', 'asian', 'beverages', 'modern indian', 'desserts', 'spanish', 'japanese', 'salad', 'sushi', 'mexican', 'bakery', 'juices', 'thai', 'malaysian', 'indonesian', 'goan', 'finger food', 'healthy food'

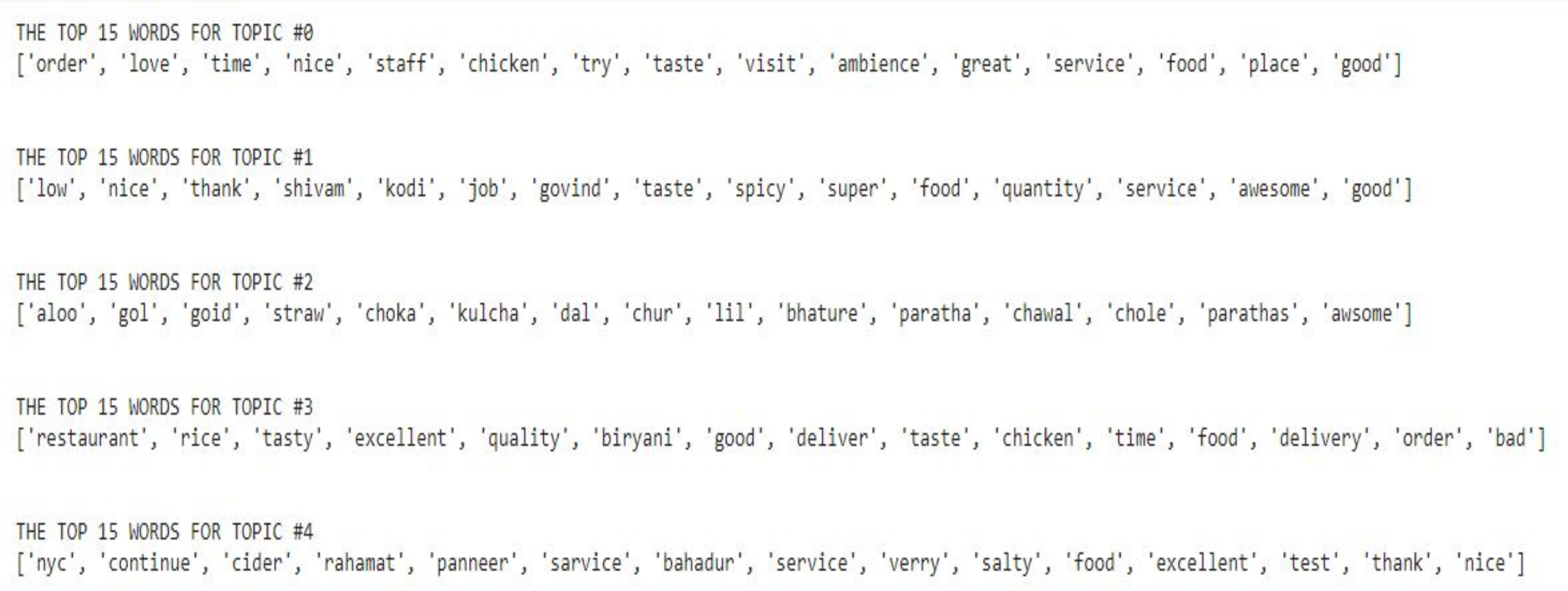
Cluster 2

'ice cream', 'desserts', 'cafe', 'bakery', 'continental', 'fast food', 'beverages', 'burger', 'biryani', 'mithai', 'american', 'wraps'

Cluster 1

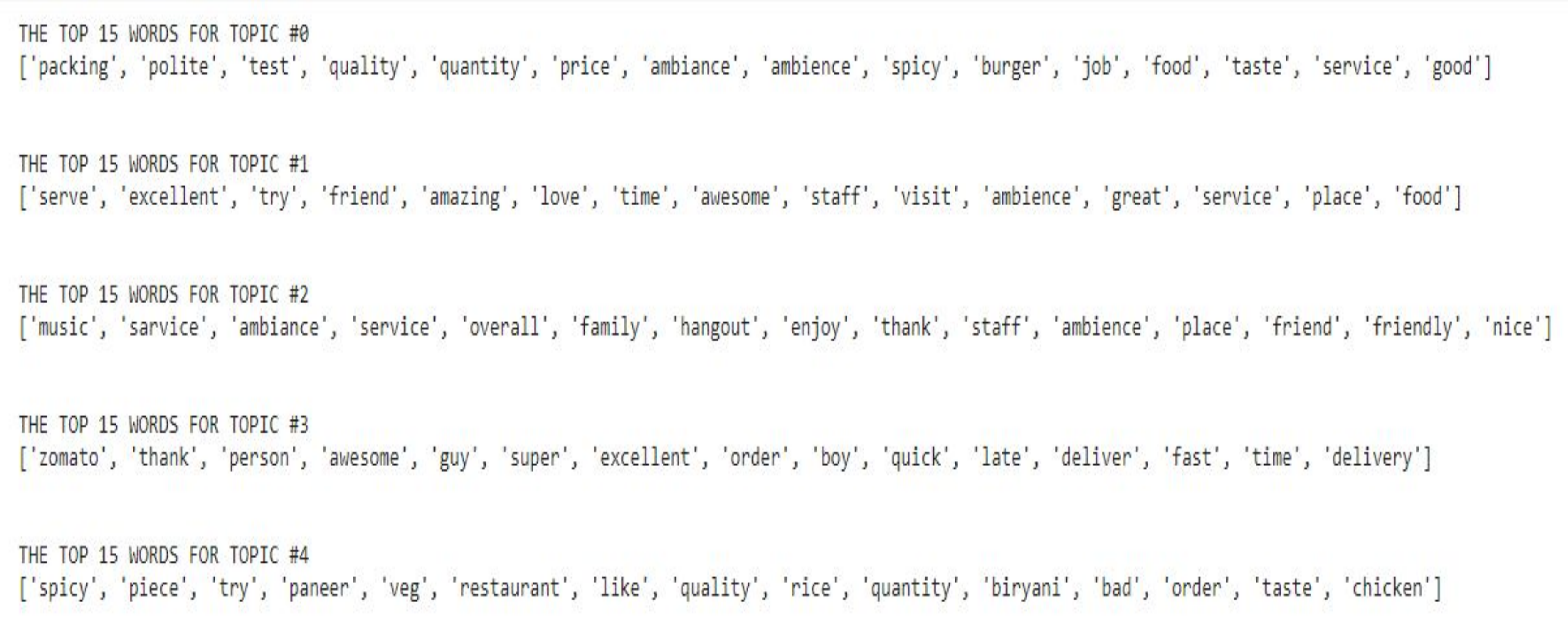
'north indian', 'continental', 'american', 'chinese', 'fast food', 'salad', 'burger', 'biryani', 'mughlai', 'asian', 'seafood', 'momos', 'pizza', 'hyderabadi', 'japanese', 'sushi', 'finger food', 'kebab', 'arabian', 'south indian', 'street food', 'lebanese', 'italian', 'thai', 'north eastern'

LDA top 15 word of each topic



```
THE TOP 15 WORDS FOR TOPIC #0
['order', 'love', 'time', 'nice', 'staff', 'chicken', 'try', 'taste', 'visit', 'ambience', 'great', 'service', 'food', 'place', 'good']


THE TOP 15 WORDS FOR TOPIC #1
['low', 'nice', 'thank', 'shivam', 'kodi', 'job', 'govind', 'taste', 'spicy', 'super', 'food', 'quantity', 'service', 'awesome', 'good']


THE TOP 15 WORDS FOR TOPIC #2
['aloo', 'gol', 'goid', 'straw', 'choka', 'kulcha', 'dal', 'chur', 'lil', 'bhature', 'paratha', 'chawal', 'chole', 'parathas', 'awsome']


THE TOP 15 WORDS FOR TOPIC #3
['restaurant', 'rice', 'tasty', 'excellent', 'quality', 'biryani', 'good', 'deliver', 'taste', 'chicken', 'time', 'food', 'delivery', 'order', 'bad']


THE TOP 15 WORDS FOR TOPIC #4
['nyc', 'continue', 'cider', 'rahamat', 'panneer', 'sarvice', 'bahadur', 'service', 'verry', 'salty', 'food', 'excellent', 'test', 'thank', 'nice']
```

NMF Top 15 word of each Topic

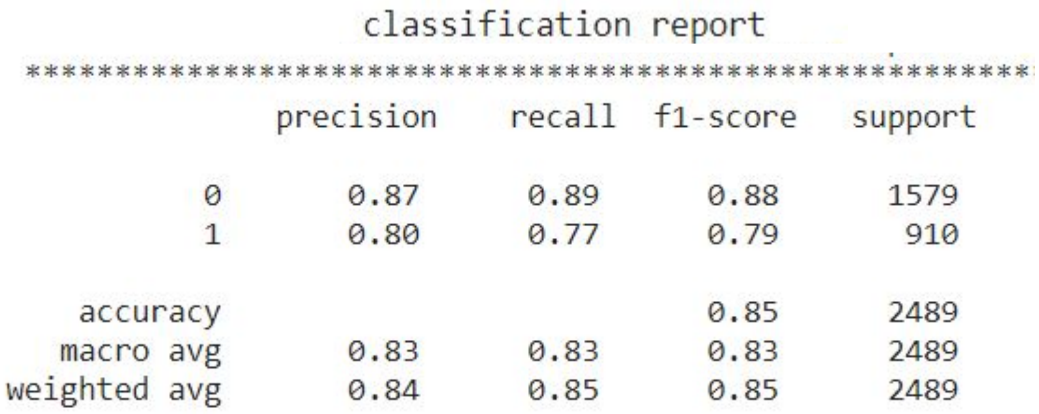


```
THE TOP 15 WORDS FOR TOPIC #0
['packing', 'polite', 'test', 'quality', 'quantity', 'price', 'ambiance', 'ambience', 'spicy', 'burger', 'job', 'food', 'taste', 'service', 'good']


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


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


THE TOP 15 WORDS FOR TOPIC #3
['zomato', 'thank', 'person', 'awesome', 'guy', 'super', 'excellent', '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', 'order', 'taste', 'chicken']
```

Logistic Regression

Parameters :

- C = 10
- Max_iter = 1000
- Penalty = L2

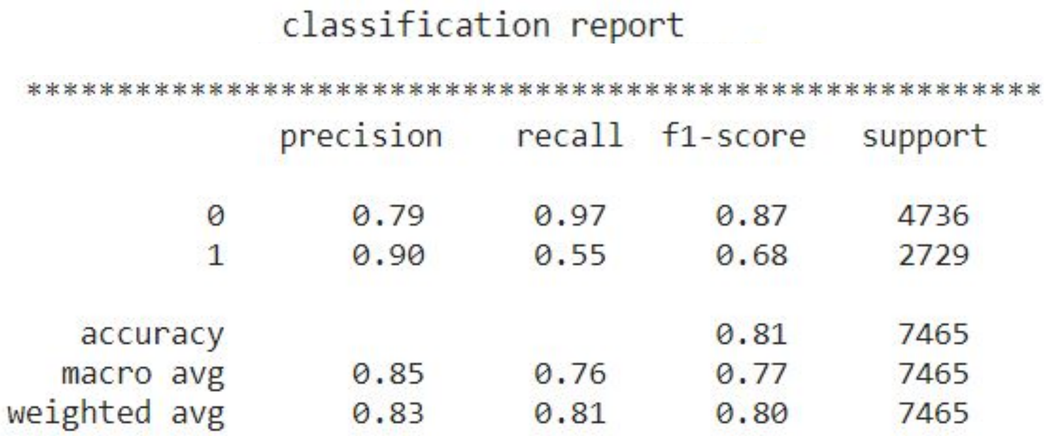


```
                     classification report
*********************************************************
              precision    recall  f1-score   support

           0       0.87      0.89      0.88      1579
           1       0.80      0.77      0.79       910

    accuracy                           0.85      2489
   macro avg       0.83      0.83      0.83      2489
weighted avg       0.84      0.85      0.85      2489
```

Random Forest Metrics

Parameters :

- **max_depth=15**
- **n_estimators=125**
- **criterion: entropy**

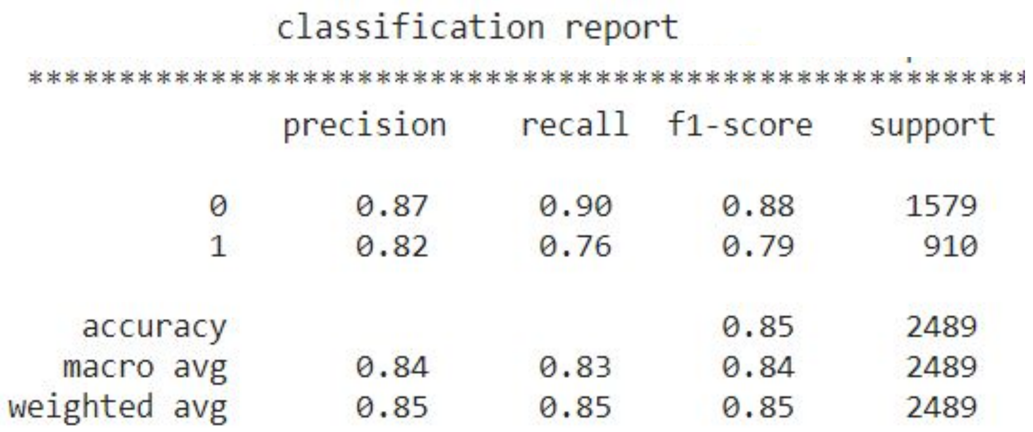


```
               classification report
 ********************************************************
              precision    recall  f1-score   support

           0       0.79      0.97      0.87      4736
           1       0.90      0.55      0.68      2729

    accuracy                           0.81      7465
   macro avg       0.85      0.76      0.77      7465
weighted avg       0.83      0.81      0.80      7465
```

XGBoost Modelling

Parameters :

- max_depth= 15
- n_estimators=125
- criterion: entropy

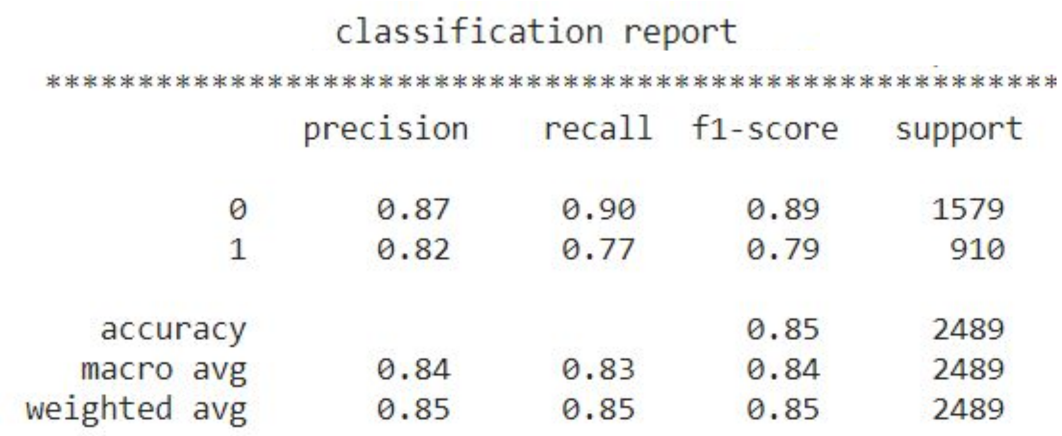


```
               classification report
 *******************************************************
              precision    recall  f1-score   support

           0       0.87      0.90      0.88      1579
           1       0.82      0.76      0.79       910

    accuracy                           0.85      2489
   macro avg       0.84      0.83      0.84      2489
weighted avg       0.85      0.85      0.85      2489
```

LightGBM

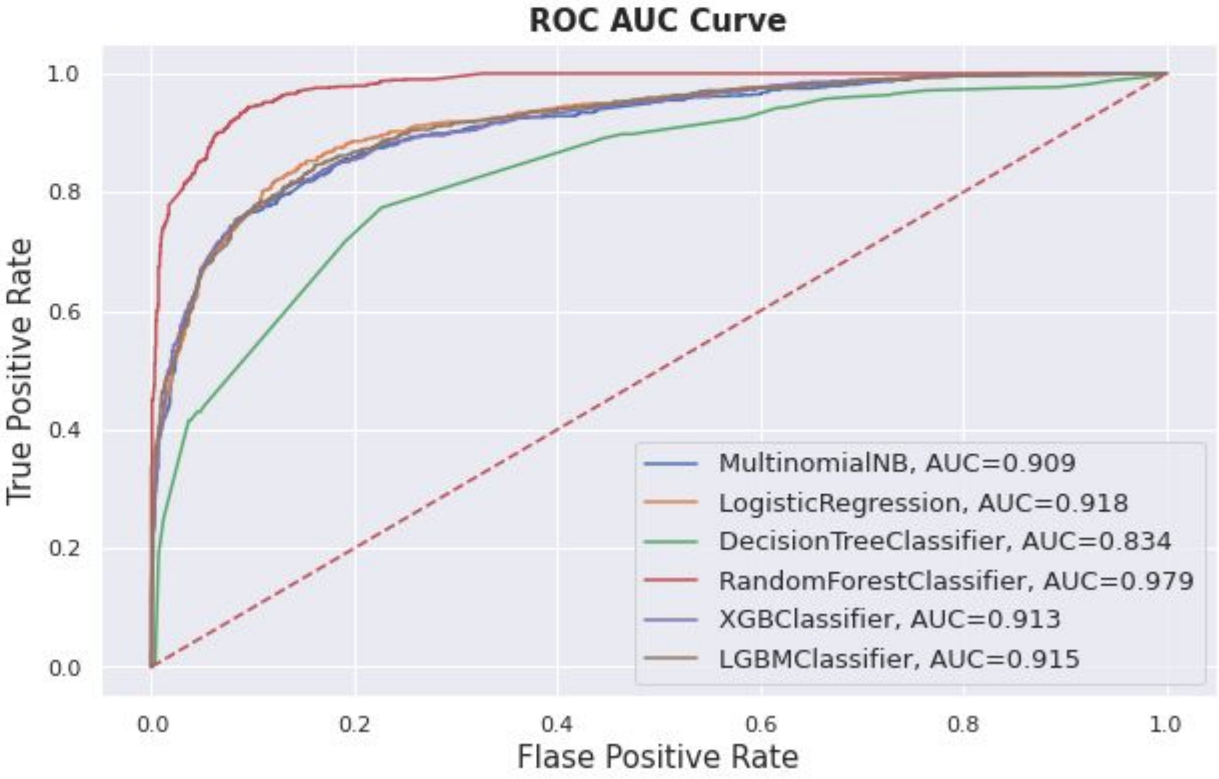
Parameters :

- max_depth=25
- n_estimators: 125

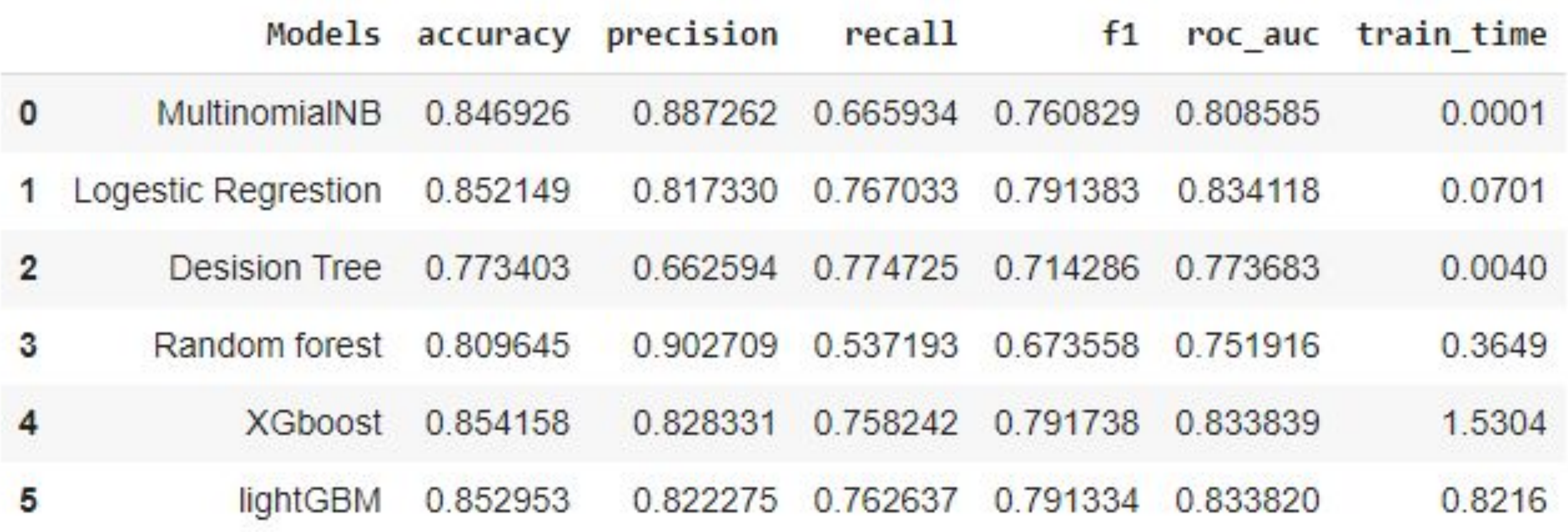


```
                 classification report
*******************************************************
              precision    recall  f1-score   support

           0       0.87      0.90      0.89      1579
           1       0.82      0.77      0.79       910

    accuracy                           0.85      2489
   macro avg       0.84      0.83      0.84      2489
weighted avg       0.85      0.85      0.85      2489
```

AUC-ROC curve comparison



Score Matrix



	Models	accuracy	precision	recall	f1	roc_auc	train_time
0	MultinomialNB	0.846926	0.887262	0.665934	0.760829	0.808585	0.0001
1	Logestic Regrestion	0.852149	0.817330	0.767033	0.791383	0.834118	0.0701
2	Desision Tree	0.773403	0.662594	0.774725	0.714286	0.773683	0.0040
3	Random forest	0.809645	0.902709	0.537193	0.673558	0.751916	0.3649
4	XGboost	0.854158	0.828331	0.758242	0.791738	0.833839	1.5304
5	lightGBM	0.852953	0.822275	0.762637	0.791334	0.833820	0.8216

Challenges

- Feature engineering.
- Finding optimum number of Cluster
- Text preprocessing

Conclusion

- We got best cluster as 3 in k means and in hierarchical

- Best no of cluster for sentiment analysis (unsupervised) is 2 i.e. for positive and negative reviews

- Best model we found for sentiment analysis(Supervised) are Lightgbm and logistic regression

Thank You