Capstone Project

Zomato Restaurant Clustering and Sentiment Analysis

Midhun R

Points for Discussion

- Business Task
- Data Summary
- Data Cleaning
- Exploratory Data Analysis
- Feature Engineering
- Text Processing
- Modelling
- Conclusion

Business Task

Two datasets are given: one with metadata of 105 restaurants and the other with reviews given for these restaurants by various reviewers.

The main objective is to understand the existing data and analyze their trends and patterns, so that machine learning models can be built, one for the clustering of restaurants and another for sentiment analysis of reviews.

This is undertaken as an individual project.

Data Summary

	Restaurant Metadata	Review Data
Number of records (rows)	105	10000
Number of features (columns)	6	7
Number of duplicate rows	0	36
Number of columns with missing values	2	5
Number of columns require conversion of data type	2	3

- These irregularities will be handled during data cleaning step.

Data Summary (Contd.)

Restaurant 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 with respect to Zomato categories.
5. Cuisines : Cuisines served by Restaurants.
6. Timings : Restaurant Timings.

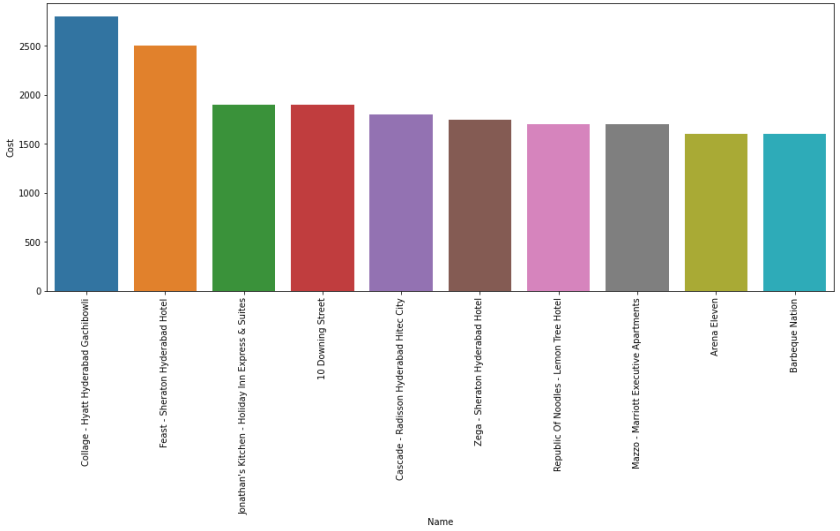
Review Data:

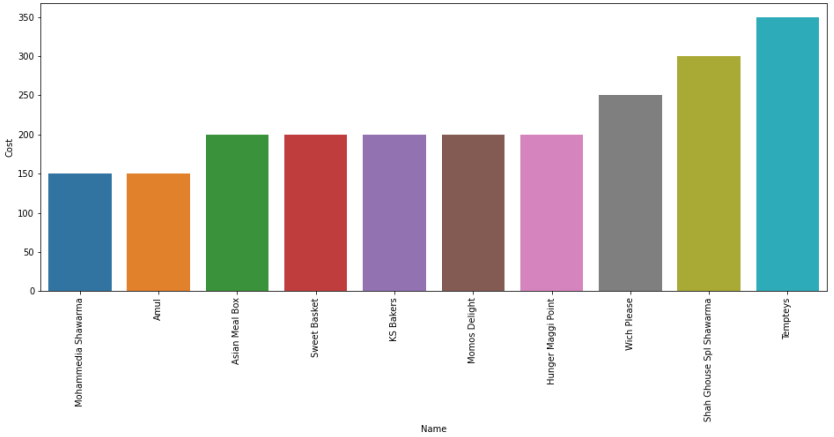
1. Restaurant : Name of the Restaurant.
2. Reviewer : Name of the Reviewer.
3. Review : Review Text.
4. Rating : Rating Provided by Reviewer.
5. Metadata : Reviewer Metadata - No. of Reviews and followers.
6. Time : Date and Time of Review.
7. Pictures : No. of pictures posted with review.

Data Cleaning

- 36 duplicate rows in review dataset were removed.
- More than half of the observations of restaurant metadata have missing values in ‘Collections’. Imputing missing values in this feature is possible only by collecting more data. So, this feature was removed since it will lead to inaccurate data analysis.
- There is only 1 missing value in ‘Timings’ and it was imputed with its mode.
- Five out of 7 columns in the review dataset have missing values in them. But only a very small number of data were missing. These observations were removed from the dataset.
- The datatype of ‘Cost’ and ‘Cuisines’ in restaurant metadata was converted to int and list while the datatype of ‘Rating’, ‘Metadata’ and ‘Time’ in review dataset was converted to float, dictionary and datetime respectively.

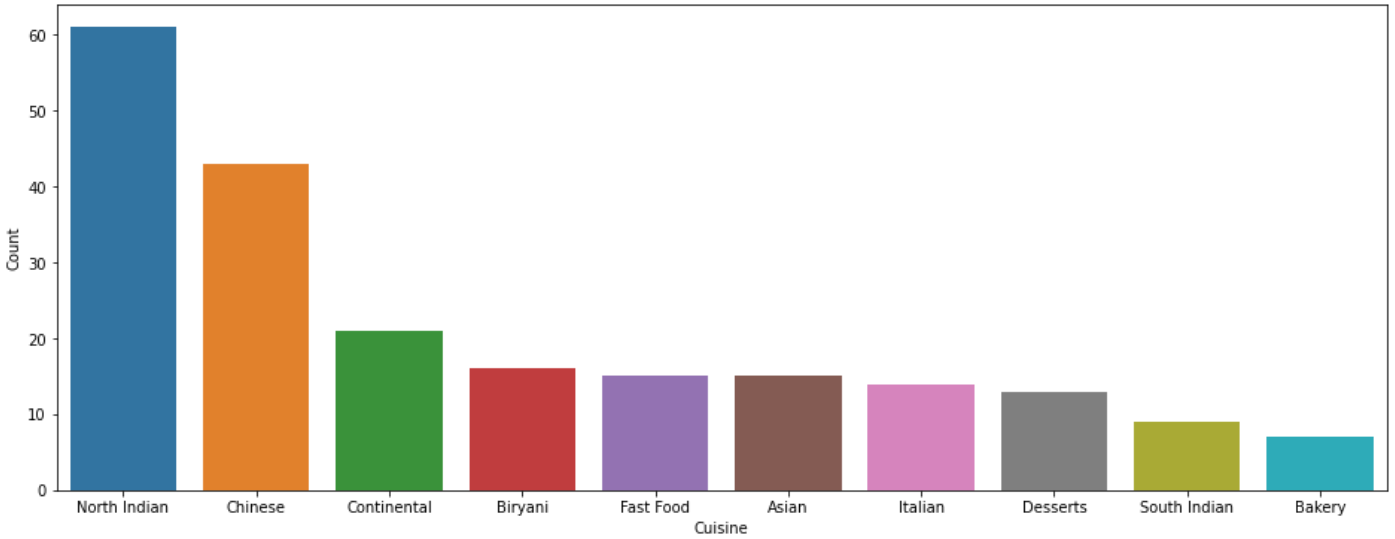
Exploratory Data Analysis





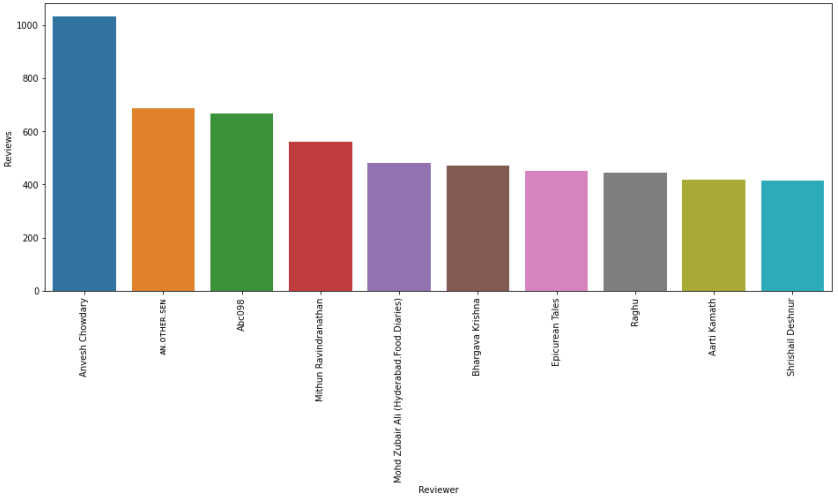
Collage - Hyatt Hyderabad Gachibowli is the most expensive restaurant followed by Feast - Sheraton Hyderabad Hotel. Mohammedia Shawarma and Amul are the most affordable restaurants among the 105 restaurants in the dataset.

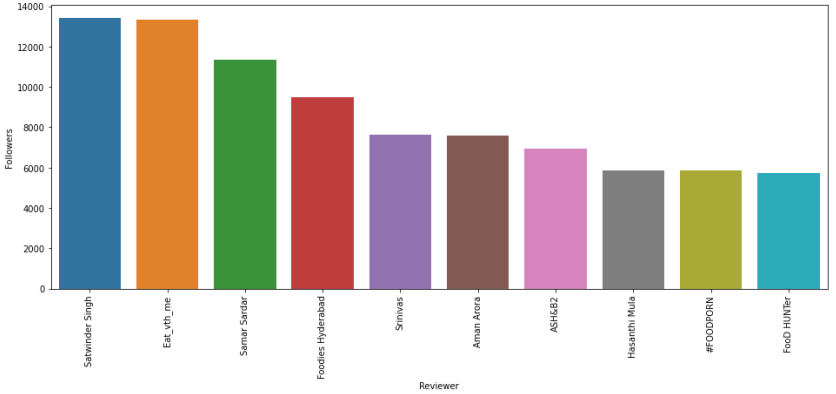
Exploratory Data Analysis (Contd.)



North Indian is the most popular cuisine by availability, followed by Chinese.

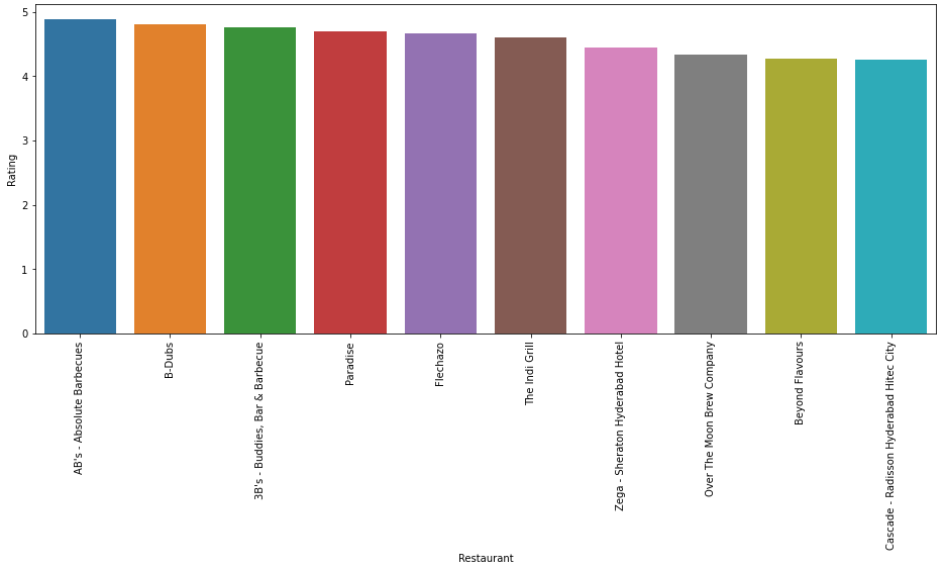
Exploratory Data Analysis





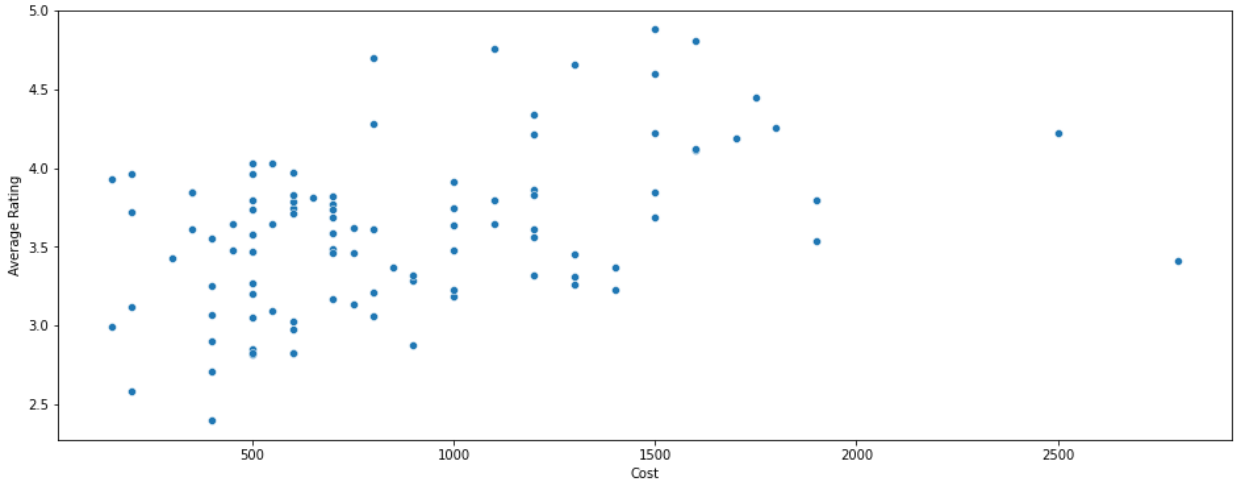
Anvesh Chowdary has written the most number of reviews with over 1000 reviews. Satwinder Singh is the most popular reviewer based on the number of followers.

Exploratory Data Analysis (Contd.)



AB's - Absolute Barbecues has the highest average rating followed by B-Dubs and 3B's - Buddies, Bar & Barbecue.

Exploratory Data Analysis (Contd.)



Some linear relationship exists between the average rating of restaurants and the cost of food.

Feature Engineering

- ‘Links’ and ‘Timings’ are removed from the restaurant metadata as are not useful in the clustering of restaurants
- The average of ‘Rating’ was calculated and then merged it with the clustering dataframe.
- ‘Cuisines’, stored as lists, was encoded using multilabel encoder since list is iterable and multiple cuisines were combined to reduce the number of dimensions.
- ‘Tokenized_Review’ and ‘Rating’ are retained from review dataset since only input text data and target sentiment data are necessary for sentiment analysis.
- Since Rating is a continuous numerical data, it was converted to categorical data to use as the target feature for classification. The value of 1 represents positive review and 0 represents negative review.

Text Processing

- Stemming is used for text normalization since getting base words is more crucial than the meaning of words to determine which class the text data belongs to.
- TF-IDF is used for feature extraction from text since just the importance of words also needs to be considered.

Wordcloud for positive reviews

Wordcloud for negative reviews

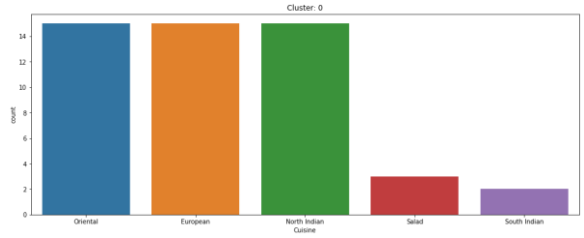
Modelling

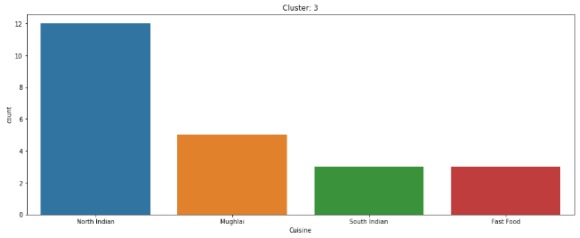
Restaurant clustering based on cuisines

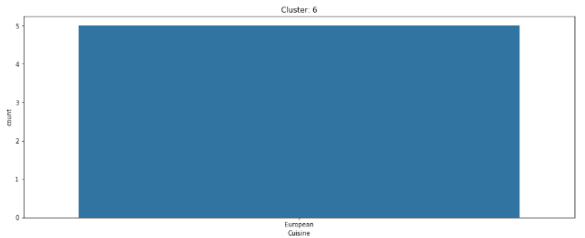
- Feature scaling was performed on the data by normalization.
- Since the clustering dataframe has a large number of features, dimensionality reduction was also performed using principal component analysis (PCA) with 3 principal components.
- Clustering was done using 2 different algorithms:
  1. K means
  2. DBSCAN
- The model built by K means algorithm created 8 clusters while the one built by DBSCAN algorithm created 7 clusters.

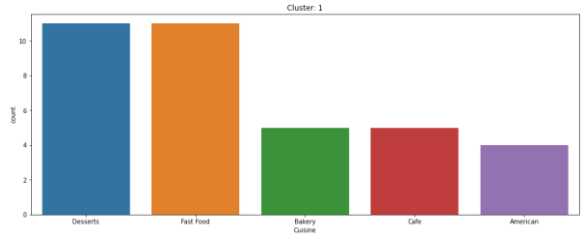
Modelling (Contd.)

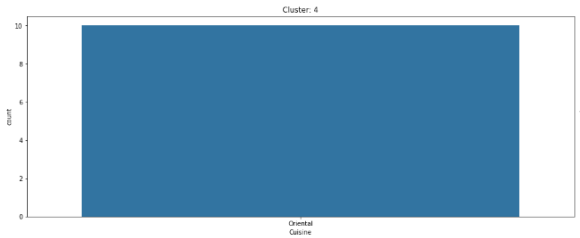
Clusters created by K means algorithm

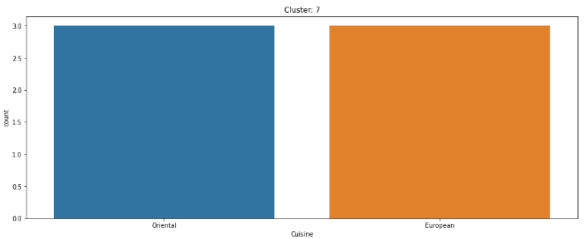


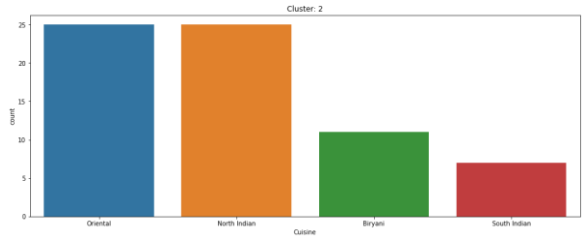


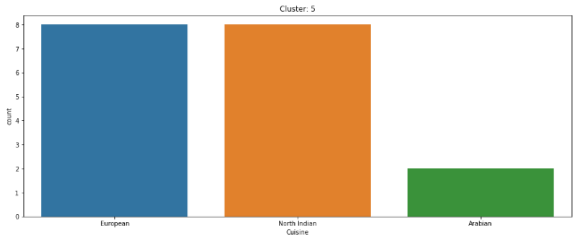






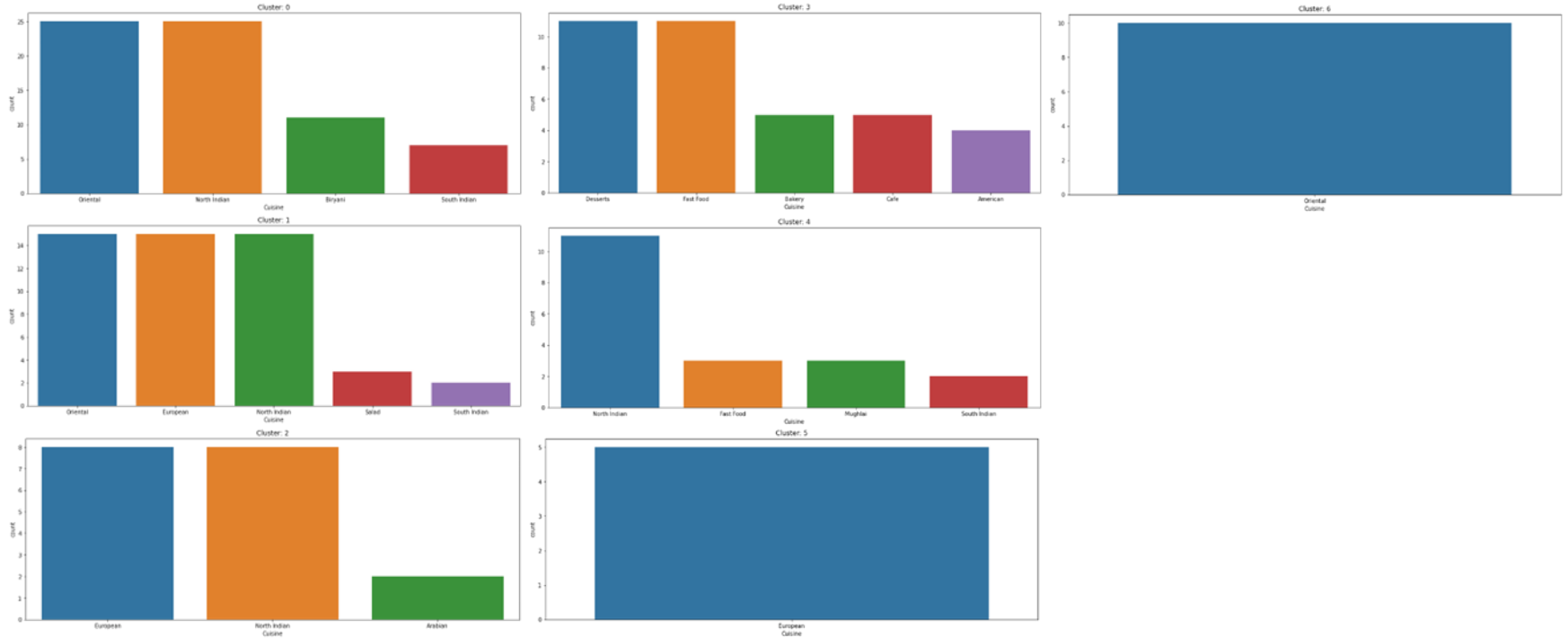






Modelling (Contd.)

Clusters created by DBSCAN algorithm



Modelling (Contd.)

- Silhouette coefficient was used to compare the two clustering models to find the out the best among them. Higher the silhouette score, better the model.

  ```
  K Means: 0.676246
  DBSCAN: 0.646559
  ```

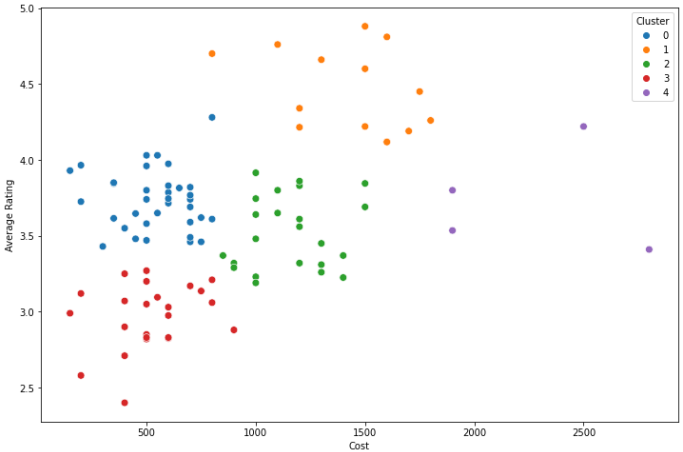
- Both the models clustered the restaurants based on cuisines similarly, except for 1 missing repeated cluster in the K means model.
- The model built using K means algorithm has higher silhouette coefficient than that built using DBSCAN algorithm by only 4.6%

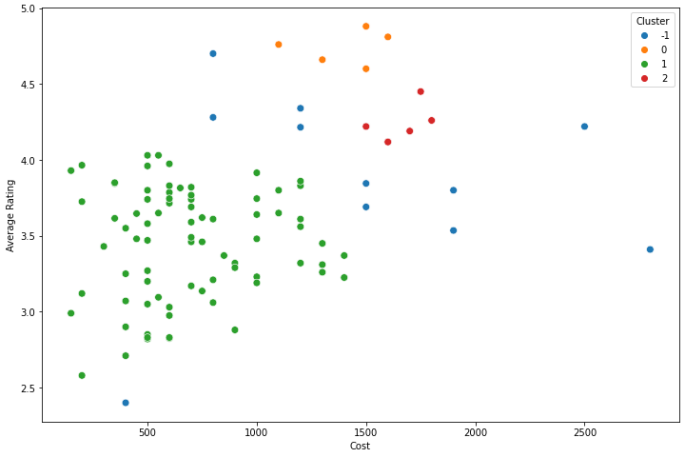
Modelling (Contd.)

Restaurant clustering based on cost and rating

- Feature scaling was performed on the data by normalization.
- Clustering was done using 2 different algorithms:

  1. K means
  2. DBSCAN





Modelling (Contd.)

- Silhouette coefficient was used to compare the two clustering models to find the out the best among them. Higher the silhouette score, better the model.

```
K Means: 0.461043
DBSCAN: 0.392883
```

- The K means algorithm was more successful than DBSCAN algorithm in creating accurate and distinguishable clusters.
- The model built using K means algorithm has higher silhouette coefficient than that built using DBSCAN algorithm by 18%.

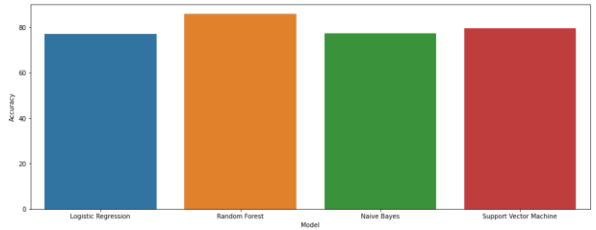
Modelling (Contd.)

Sentiment analysis

- Input and target data, which in this case were vectorized corpus and sentiment respectively, were split into training and test data with 25% test data.
- Training and test data of independent features were scaled by standardization.
- An over sampling method called SMOTE was used to balance the dataset. Undersampling was not employed as the dataset is already small and undersampling will lead to loss of information.
- Model training was done using 4 different algorithms:
    1. Logistic regression
    2. Random forest
    3. Naïve Bayes
    4. Support Vector Machine

Modelling (Contd.)

	Model	Accuracy	Precision	Recall	F1Score
0	Logistic Regression	77.059060	77.225251	77.059060	77.130006
1	Random Forest	85.777421	85.682893	85.777421	85.669634
2	Naive Bayes	77.380474	77.435115	77.380474	77.406250
3	Support Vector Machine	79.509843	79.564589	79.509843	78.808569




```
The best parameters:
min_samples_split=10
min_samples_leaf=2
max_leaf_nodes=100
max_features=auto
max_depth=None

Best F1 Score: 86.4619%
```

- Evaluation metrics like Accuracy, Precision, Recall and F1 Score were calculated for each model.
- F1 Score was used to compare different models, since the dataset is imbalanced, and find out which one is better. Higher the F1 Score, better the model.
- The model built using the random forest algorithm has the highest F1 Score. So, it can be used for sentiment analysis after performing hyperparameter tuning.
- Best parameters and accuracy were found after hyperparameter tuning.

Conclusion

EDA Conclusions

- Collage - Hyatt Hyderabad Gachibowli is the most expensive restaurant and Mohammedia Shawarma, and Amul are the most affordable ones.
- North Indian cuisine is the most popular cuisine.
- Anvesh Chowdary is the most experienced reviewer while Satwinder Singh is the most popular one.
- AB's - Absolute Barbecues is the highest rated restaurant.
- Some linear relationship exists between the average rating of restaurants and the cost of food.

Modelling Conclusions

- Either of the two models, trained using K means algorithm or DBSCAN algorithm, can be chosen for clustering the restaurant dataset based on cuisines, depending on the number of clusters preferred and whether outliers be included.
- The model built using K means algorithm is selected for clustering based on cost and ratings.
- For sentiment analysis, the model built using random forest algorithm was chosen over others.
- If model interpretability is more important than accuracy, model built using logistic regression should be chosen.