Capstone Project Summary

Team Member's Name, Email and Contribution:

Tito Varghese

Email - tito.varghese1992@gmail.com

Individual Contributions

- Data Loading
- Data Inspection and Cleaning
- Exploratory Data Analysis
- Text Preprocessing Reviews
- Topic Modelling(LDA, NNMF)-Sentimental Analysis
- Supervised Classification Model Formulation- Sentimental Analysis
- Logistic Regression
- Decision Tree
- Random Forest
- XGboost
- LightGBM
- Clustering the restaurants- K- Means,PCA,DBSCAN and Hierarchical
- Clustering Analysis
- Cost Benifit Analysis
- Technical Docomentation
- Presentation Slide

Please paste the GitHub Repo link.

GitHub Profile Link: - https://github.com/7692TITO

GitHub Repository Link: - https://github.com/7692TITO/ZOMATO_RESTAURANT_CLUSTERING_AND_SENTIMENT_ANALYSIS

Zomato Restaurant Clustering and Sentiment Analysis
Summary

The Project focuses on Customers and Company, 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 data is vizualized as it becomes easy to analyse data at instant. The Analysis also solve 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.

In this Unsupervised Capstone Project, we will be looking into multiple unsupervised models and supervised models for our sentiment analysis ,at the end try to come up with a best model. We are only focussing on all that algorithm which has been taught to us till now in our class. LDA,NNMF for topic modelling and a few more clustering algorithm like K-Means,Hierarchical,DBSCAN and PCA, we have implemented in this capstone project.Apart from that ,we will be using supervised ml classification model like Logistic,XGboost,Random Forest,Decision Tree,LightGBM, for the sentiment analysis.

Pipeline to be followed:-

- Dataset Inspection & Cleaning
- Feature Engineering
- Exploratory Data Analysis
- Text Preprocessing
- Topic Modelling
- Supervised ML Classification Models
- Model Summary
- ROC-AUC Plot
- Clustering Models
- Cluster Analysis
- Cost Benfit Analysis

Some of the important conclusions to be drawn are:

- The most popular cuisines available in the most of the restaurant is the North Indian and Chinese.
- The collage-Hyatt Hyderbad Gachibowli is the most expensive restaurant available and the most affordable restaurant for the customers are the Amul and Mohammedia Shawarma.
- The most expensive Cuisine is the Modern India cuisine which cost around 2000 rupees and the least expensive item available at a cost of 200 rupees is the Mithai
- Great Buffet is one of the most common Tags given to the zomato restaurants with nearly more than 40 restaurants.
- Namit Aggarwal is one of the active reviewer based on number of followers and ratings provided
- Sentiment Analysis was done on the reviews and a model was trained in order to identify negative and positive sentiments. The best model for sentiment analysis we found out to be LightGBM and Logistic Regression model.
- Restaurant Clustering was done in with just two features Cost and Rating. Kmeans Clustering gave us optimal cluster value as 5 but we have done our clustering analysis based upon the principal

component analysis because the similarities in the data points within the clusters were pretty great. We have got optimal clusters as 4 clusters in PCA.

Cluster 0 - The mostly available cuisine in the restaurants in cluster 0 is the South east asian cuisines followed by Indian cuisine and the restaurants in cluster 0 does not having Beverges and Deserts available. The average rating is 3.46 and the average cost is 923 INR which includes an outlier of cost 1150 INR and median cost of 525 INR. This means the restaurants are basically in general cheaper in nature in this cluster.

Cluster 1 - The mostly available cuisine in the restaurants of cluster 1 is the Beverges and Desserts cuisines followed by Fast Food cuisine.The average rating is 3.61 and the average cost is 736 INR which includes an outlier of cost 2500 INR and median cost of 600 INR.. These restaurants are slightly higher in prices than cluster 0.

Cluster 2 - The mostly in demand and available cuisine in the restaurants of cluster 2 is the Continental cuisines followed by Indian cuisine.The average rating is 3.82 and the average cost is 1052 INR which includes an outlier of cost 2800 INR and median cost of 1100 INR.. These restaurants are fine dining restaurants and expensive as well compared to other clusters like 0 and 1.

Cluster 3 - The mostly in demand and available cuisine in the restaurants of cluster 3 is the Indian cuisines.The restaurants in cluster 2 have a very few demand of fast food and beverages.The average rating is 3.48 and the average cost is 769 INR which includes an outlier of cost 1700 INR and median cost of 700 INR.. These restaurants are afforadable and mostly preferred cuisine in high numbers in these restaurants.

Recommendations

- Ratings should be collected on a category basis such as rating for packaging, delivery, taste, quality, quantity, service, etc. This would help in targetting specific fields that are lagging.