zomato

Capstone Project 4

Zomato Restaurant Clustering And Sentiment Analysis

MANOHAR JHA

CONTENTS:

zomato

- Project Summary
- Problem Statement
- Understanding the Dataset
- EDA
- Data Wrangling
- All Manipulations
- Selecting algorithm
- Modeling
- Conclusion

MANOHAR JHA

Project Summary :

zomato

Zomato is an Indian restaurant aggregator and food delivery start-up founded in 2008 by Deepinder Goyal and Pankaj Chaddha. Zomato provides restaurant information, menus and user-reviews, and also offers food delivery options from partner restaurants in select cities. India is quite famous for the diverse multi-cuisine available in a large number of restaurants and hotel resorts, which reminds of unity in diversity. The restaurant business is growing significantly in India. Most Indians are loving the idea of eating restaurant food, whether it's eating out or ordering food. An increasing number of restaurants in every state of India are going to inspect the data to get some insights, interesting facts and figures about the Indian food industry in each city. Therefore, this project focuses on analyzing Zomato restaurant data from each city in India.

This project focuses on customers and company, you have to analyze the sentiments of customer reviews in the data and draw some useful conclusions in the form of visualization. Apart from this, Zomato restaurants will also be grouped into different segments. Data is visualized because it becomes easier to analyze the data quickly. The analysis also solves some business cases which can directly help the customers to find the best restaurants in their locality and can help the company to move ahead and work on the areas in which they are currently lagging behind. It may help to divide the restaurant into sections. Additionally the data contains valuable information about dishes and costs that can be used in cost vs benefit analysis The data can be used for sentiment analysis. Additionally reviewers' metadata can be used to identify critics in the industry.

MANOHAR JHA

Problem Statement :

zomato

Zomato which is an Indian restaurant and food delivery start-up takes orders through online mode and delivers food items to people's doorsteps. We have some data of that which is from Hyderabad. In that data, we have to analyze it by applying the algorithm of Unsupervised ML, in which we have to understand the sentiments of the customers and based on the information about cuisine, cost and customer reviews, we have to find out whether the customers are trying to find the best restaurant in their area. That no. So that the company can be helped in development and improvement in the food industry.

MANOHAR JHA

Understanding the Dataset :

zomato

This is data from a Zomato restaurant, which has 10105 rows and 13 columns. In this, all the data comes in type object, only one is in int. 36 Duplicate Values in review dataset and both has missing values.

MANOHAR JHA

EDA :

zomato

Exploratory Data Analysis (EDA) is a critical phase in the data analysis process that involves examining and visualizing data to gain insights, identify patterns, and uncover potential relationships. This report outlines the key findings and observations from the EDA conducted on the Zomato Restaurant Clustering And Sentiment Analysis dataset.

1. Understand the Data Structure
2. Find mean, median, standard deviation, etc
3. Explore the distribution of key variables
4. Visualize distributions using histograms, box plots
5. Handling missing values

MANOHAR JHA

Data Wrangling:

zomato

First of all I copied both the datasets with different names. Then I checked the price values in a dataset and converted them to integers. Then I checked the value count of Rating. Dropped some unnecessary columns. After that I replaced the value. After that reviews and followers were extracted from a dataset. After that the datetime was extracted. Then checked the value count of some columns. Then split the text data into a dataset. Then added its value in a separate column. After add year and Rating column in meta_df dataset and find missing value and fit median value.

MANOHAR JHA

All Manipulations:

zomato

First of all I copied both the datasets with different names. Then I checked the price values in a dataset and converted them to integers. Then I checked the value count of Rating. Dropped some unnecessary columns. After that I replaced the value. After that reviews and followers were extracted from a dataset. After that the datetime was extracted. Then checked the value count of some columns. Then split the text data into a dataset. Then added its value in a separate column. After add year and Rating column in meta_df dataset and find missing vale and fit median value and convert 'Cost' column to numeric.

MANOHAR JHA

Selecting algorithm :

I have used three machine-learning models:

1. Logistic Regression:
2. K Nearest Neighbors (KNN):
3. Random Forest Classifier:

These models are trained and evaluated using metrics such as accuracy, precision, recall, and AUC-ROC scores.

Modeling :

After using all these models we find that they are having the highest testing accuracy and precision in Logistic Regression. In Logistic Regression, Test accuracy- 0.875063, Precision- 0.874544, Recall Auc- 0.939655, Roc Score- 0.849883 are coming. which is the most of all.

Logistic Regression is a binary classification algorithm used to predict the probability of an instance belonging to a particular class. It's particularly suitable when the dependent variable is binary, meaning it has only two possible outcomes.

In your case, it seems like you've used Logistic Regression for a binary classification task, and the model has demonstrated strong performance based on the provided metrics (Test accuracy: 0.875063, Precision: 0.874544, Recall AUC: 0.939655, ROC Score: 0.849883).

Conclusion :

zomato

After a comprehensive analysis of various machine learning models, it is evident that Logistic Regression outperforms others in terms of testing accuracy, precision, recall AUC, and ROC score. The obtained results showcase the robustness and effectiveness of Logistic Regression in the given context. With a testing accuracy of 87.51%, precision of 87.45%, recall AUC of 93.97%, and a ROC score of 84.99%, Logistic Regression demonstrates a superior ability to accurately classify and predict outcomes.

These findings suggest that, for the specific task at hand, Logistic Regression is a reliable choice for modeling and prediction. The model not only achieves high overall accuracy but also excels in correctly identifying positive instances, as reflected in the precision and recall AUC metrics. The ROC score further supports the model's discriminative power and its ability to balance sensitivity and specificity.

MANOHAR JHA