Blog

By: Navin Singh

Title: A machine Learning Analysis

Project Name: Exploring Zomato Dataset

Introduction: The emergence of platforms like Zomato has revolutionized the way people discover and experience food. This analysis will help for foodie person. In this blog we delve into the world of Zomato dataset analysis, where we explore insights that cater to the needs of food enthusiasts, budget-conscious diners, and those seeking values for money restaurants. By leveraging **Machine Learning Techniques**, my aim to predict the **Average cost for two people** and the **Price range** of restaurants, providing valuable insights for the both consumers and businesses.

Project Description: Zomato Data Analysis is a comprehensive exploration of restaurant data from various part of the world. The dataset comprises two main files:

1. Zomato.csv
2. Country_code.csv

Country code --> https://github.com/dsrscientist/dataset4/blob/main/Country-Code.xlsx

Country name --> https://raw.githubusercontent.com/dsrscientist/dataset4/main/zomato.csv

More Details about this Project (find 4th Project) --> https://github.com/ksingh9398/Internship-Project/blob/main/Third-Phase-Evaluation-Projects--1-.pdf

The former contains detailed information about restaurants, including their names, locations, cuisines offered, average cost for two people, and corresponding country names.

All columns of this dataset are:

1. Restaurant Id: Unique id of every restaurant across various cities of the world
2. Restaurant Name: Name of the restaurant
3. Country Code: Country in which restaurant is located

4. City: City in which restaurant is located
5. Address: Address of the restaurant
6. Locality: Location in the city
7. Locality Verbose: Detailed description of the locality
8. Longitude: Longitude coordinate of the restaurant's location
9. Latitude: Latitude coordinate of the restaurant's location
10. Cuisines: Cuisines offered by the restaurant
11. Average Cost for two: Cost for two people in different currencies ��
12. Currency: Currency of the country
13. Has Table booking: yes/no
14. Has Online delivery: yes/ no
15. Is delivering: yes/ no
16. Switch to order menu: yes/no
17. Price range: range of price of food
18. Aggregate Rating: Average rating out of 5
19. Rating colour: depending upon the average rating colour
20. Rating text: text on the basis of rating of rating
21. Votes: Number of ratings casted by people
22. Country Code

Data Exploration and Preparation: My first step is, I import some important library such as pandas, numpy, matplotlib seaborn and warning to avoid warning text, after that I load both dataset and merge in single DataFrame as df. In this dataset there are 22 columns and 9551 rows are present.
Then I observe the all types of details of this dataset using df.info() techniques, then I saw that in this dataset there are three types of dataset (integer, float and object) are present.
Then I check for null values, I got 9 null values in Cuisines columns, I removed this null values by Mode of this columns, then I remove duplicates values and converting data types as needed.

Feature Engineering: with my data cleaned and merged, I proceed to extract relevant features for our Machine Learning Models. I create new features from existing ones, such as deriving the cuisine types offered by

each restaurant or calculation the distance of each restaurant from a central point using its latitude and longitude coordinates.

Exploratory Data Analysis (EDA): EDA plays a crucial role in understanding the underlying patterns and relationships in our data. It's help to visualize various aspects of the data using libraries like **Matplotlib** and **Seaborn**, exploring trends in average cost, ratings distribution, popular cuisines, and more. This analysis helps us gain insight into factors that influence restaurant pricing and customer preferences.

Description of dataset:

This gives the statistical information of the numerical columns. The summary of the dataset looks perfect since there are no negative values

I observe the following statistical points from Description of dataset (df.describe()):

The Counts of all the columns are same which means there are no missing values in the dataset.

the mean value is greater then median(50%) in country code, average cost for two and votes columns which means the data is skewed to right these columns.

the mean values is less then median(50%) in Longitude,Latitude, price range and average rating columns which mean the data is skewed to left these columns.

By summarizing the data we can observe there is a huge difference between 75% and max in Country Code,Longitude, Latitude,Average Cost for two and votes hence there are outliers present in the data.

We can also notice the Standard deviation, min, 25% percentile values from this describe method.

Then I check outliers by Box plot, in this dataset some columns have outliers then I remove this outliers using z-score test. Data loss percentage is 10.92 %.

Then I check skewness and correlation of this dataset;

Model Training and Evaluation: for prediction the average cost for two people and the price range of restaurants, I apply machine learning Algorithms such as Linear Regression and Classification Models (Random Forest Regressor).
I split the data into training and testing dataset, train the models on the training data, and evaluate their performance using metrics like mean squared error (MSE) for regression and accuracy for classification. Additionally, I use techniques like cross validation to ensure the robustness of our models.

Results and Interpretation: After training our models, I analyze the result to understand the factors that contribute most to prediction the average cost and price range of restaurants. We may visualize the model coefficients or feature importance to identify the most influential freatures. Furthermore, I interpret the model predictions to provide actionable insights for consumers and restaurant owners alike.

Conclusion: In this blog post, I have embarked on a fascinating journey through Zomato data analysis using Machine Learning Techniques. By levering the wealth of information available in the dataset, we have gained insights into restaurant pricing and customer preferences, empowering both diners and businesses to make informed decision. As the field of data science continuous to evolve, we look forward to further exploration and discovery in the realm of food analytics.

Thanks

NAVIN SINGH
Data Scientist
Batch No: DS2309