# Article on Zomato Data Science project

**1. Problem Definition**

In the competitive world of online food delivery and restaurant discovery, understanding the various factors influencing customer choices is critical. For Zomato, a leading global food discovery platform, optimizing recommendations and improving user experience can significantly impact business performance. This data science project focuses on two key aspects: analyzing the "Average Cost for Two" and "Price Range" of restaurants listed on the platform. These metrics are crucial for both customers and restaurant owners.

**Problem Statement:**

The primary goal is to build predictive models that can accurately forecast the "Average Cost for Two" and "Price Range" of restaurants based on various features such as location, cuisine type, restaurant type, and user ratings. These predictions can help users make informed decisions when choosing a restaurant and assist restaurant owners in positioning their establishments competitively.

In the age of digital dining, platforms like Zomato play a pivotal role in shaping the culinary landscape. For this data science project, we explore two datasets: **Zomato.csv**, containing detailed information about various restaurants, and **country\_code.csv**, providing mappings of country codes to their respective country names. Our primary goal is to analyze and model two critical aspects of the dining experience:

* **Average Cost for Two:** Understanding the typical expenditure for two people in different regions and under varying conditions can provide insights into affordability and dining trends.
* **Price Range:** Classifying restaurants into different price ranges helps customers make informed decisions based on their budget and dining preferences.

**2. Data Analysis**

Zomato.csv contains various features related to restaurants, including but not limited to:

* Restaurant ID
* Restaurant Name
* Country Code
* City
* Address
* Locality
* Cuisines
* Average Cost for Two
* Currency
* Has Table Booking
* Has Online Delivery
* Is Delivering Now
* Price Range
* Aggregate Rating
* Rating Color
* Rating Text
* Votes

country\_code.csv provides a simple mapping:

* Country Code
* Country Name

Dataset Overview:The dataset used for this project comprises restaurant details, including attributes like name, location, cuisines, cost for two, user rating, and price range. This dataset is rich in categorical and numerical features, making it an ideal candidate for a comprehensive data analysis and machine learning modeling.

Data Collection:The data was collected from Zomato's public API and other sources, encompassing a wide range of restaurants across various cities. The dataset includes the following attributes:

Restaurant Name : Name of the restaurant.

Location: City and locality where the restaurant is situated.

Cuisines: Types of cuisines served.

Average Cost for Two: Average expense for two people.

Price Range: Categorization of price (usually on a scale of 1 to 4, where 1 is the cheapest and 4 is the most expensive).

User Rating: Average rating provided by users.

**Initial Observations**

Restaurant Distribution: The dataset includes restaurants from various countries, which can be identified through the Country Code. The country\_code.csv will help decode these numerical values.

Cuisines and Cost: The variety of cuisines offers an opportunity to analyze whether certain types of food are more expensive on average.

Ratings and Reviews: Ratings, votes, and associated textual descriptions provide qualitative insights that can be quantified for analysis.

Service Availability: Features like Has Table Booking and Has Online Delivery offer additional dimensions for understanding the service dynamics.

A preliminary analysis of the dataset revealed the following insights:

Geographical Distribution: A wide range of cities and localities are represented, providing a diverse dataset.

Cuisine Diversity: The dataset contains a variety of cuisines, reflecting the global appeal and diverse offerings of restaurants on Zomato.

Cost and Price Range: There is significant variability in the "Average Cost for Two" and "Price Range," suggesting a wide spectrum of affordability.

**3. EDA Concluding Remarks**

Exploratory Data Analysis (EDA) is crucial for uncovering patterns, anomalies, and relationships in the data. During EDA, the following key insights were identified:

**Key Findings**

1.Correlation Between Variables:

* There is a positive correlation between "User Rating" and "Average Cost for Two," suggesting that higher-rated restaurants tend to be more expensive.
* The "Price Range" is generally aligned with the "Average Cost for Two," confirming that these two features are interrelated.

2.Geographical Trends:

* Certain cities and localities exhibit higher average costs, likely due to the economic status and living costs in those areas.
* Popular tourist destinations and business hubs tend to have a higher concentration of expensive restaurants.

3.Cuisine Impact:

* Some cuisines, such as fine dining and international cuisines, are associated with higher costs compared to local and fast-food options.

4.Rating Distribution:

* The distribution of user ratings suggests a skew towards higher ratings, indicating a general satisfaction with the restaurants listed on the platform.

5.Challenges Identified

* Missing Values: Some records had missing values, particularly in the "Average Cost for Two" and "Price Range" columns.
* Outliers: A few data points had extreme values, which could potentially skew the analysis and model predictions.
* Categorical Data: Handling categorical features like "Cuisines" and "Location" requires careful preprocessing to ensure meaningful model input.

**4. Pre-processing Pipeline**

To ensure the data is suitable for model training, a robust preprocessing pipeline was implemented. This pipeline includes several crucial steps:

**Data Cleaning**

1.Handling Missing Values:

* Missing values in numerical columns were imputed using the median value of the respective columns.
* Missing values in categorical columns were filled with a placeholder indicating missing information.

2.Outlier Treatment:

* Outliers in the “Average Cost for Two” were identified using the IQR method and capped at a reasonable limit.

**Feature Engineering**

1.Encoding Categorical Variables:

* Location and Cuisines: One-hot encoding was used for these features to convert them into numerical format.
* Price Range: As a categorical variable with an ordinal nature, label encoding was applied.

2.Feature Creation:

* Created new features like "Total Number of Cuisines" offered by a restaurant.

3.Scaling and Normalization:

* Numerical features like "Average Cost for Two" were normalized using standard scaling to ensure a uniform range across features.

**5. Building Machine Learning Models**

With a clean and well-preprocessed dataset, the next step is to build machine learning models to predict the "Average Cost for Two" and "Price Range."

**Model Selection**

Given the nature of the problem, the project involved two types of predictive tasks:

1. **Regression Task:** Predicting the "Average Cost for Two" as a continuous variable.
2. **Classification Task:** Predicting the "Price Range" as a categorical variable.

**Regression Models**

Several regression models were tested to predict the "Average Cost for Two":

* Linear Regression: A basic model to establish a baseline.
* Decision Tree Regressor: To capture non-linear relationships.
* Random Forest Regressor: An ensemble method for better performance.
* Gradient Boosting Regressor: Another ensemble technique focusing on improving the model's accuracy.

**Classification Models**

For predicting the "Price Range," the following classification models were explored:

* Logistic Regression: A simple model to understand baseline performance.
* Decision Tree Classifier: To capture the decision-making process.
* Random Forest Classifier: An ensemble method for higher accuracy.
* Gradient Boosting Classifier: For improved prediction capabilities.

**Model Evaluation**

To evaluate the models, the dataset was split into training and testing sets using an 80-20 split. Cross-validation was also used to assess the model's performance across different subsets of the data.

Metrics Used:

* Regression Metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R²) were used to evaluate regression models.
* Classification Metrics: Accuracy, Precision, Recall, and F1-Score were used to assess

**Model Performance**

* Regression Models: The Gradient Boosting Regressor achieved the best performance, with the lowest MAE and MSE, and the highest R² value.
* Classification Models: The Gradient Boosting Classifier also outperformed other models, achieving the highest accuracy and F1-Score.

**6. Concluding Remarks**

This data science project on Zomato's restaurant data provided valuable insights into predicting key metrics like "Average Cost for Two" and "Price Range. " The project's success can be attributed to a thorough exploratory data analysis, a robust preprocessing pipeline, and the application of various machine learning models.

Key Takeaways

* Modeling Capabilities: The Gradient Boosting models proved to be the most effective for both regression and classification tasks, highlighting the importance of ensemble methods in capturing complex patterns in data.
* Feature Importance: Features such as "Location," "Cuisines," and "User Rating" were critical in predicting both target variables, underscoring their significance in the food and restaurant industry.
* Business Implications: The models developed can be leveraged by Zomato to enhance user experience by providing more accurate cost and price range estimates. Restaurant owners can also use these insights to strategically position their offerings.