**Predicting Restaurant Ratings on Zomato Using Machine Learning**

**1. Problem Definition**

**Overview:** The food and beverage industry thrives on customer satisfaction, which is often quantified through ratings on platforms like Zomato. Predicting restaurant ratings can help restaurant owners improve service quality and help customers make informed choices. This project aims to build a predictive model to estimate restaurant ratings based on various features.

**Problem Statement:** To develop a machine learning model that predicts the ratings of restaurants listed on Zomato using attributes such as location, cuisine, cost, and other relevant features.

**Key Questions:**

* Which features most significantly affect restaurant ratings?
* How can data preprocessing and feature engineering enhance model performance?
* Which machine learning algorithms perform best for this prediction task?

**2. Data Analysis**

**Data Collection:** The primary dataset used for this project is sourced from Zomato and includes information about various restaurants. Additionally, a country code dataset is used to map country codes to country names.

**Initial Data Exploration:**

* **Dataset Loading:**

python

Copy code

df\_zomato = df\_zomato = pd.read\_csv('https://raw.githubusercontent.com/FlipRoboTechnologies/ML\_-Datasets/main/Z\_Restaurant/zomato.csv',encoding='ISO-8859-1')

df\_country = pd.read\_excel(r'C:\Users\apple\Downloads\Country-Code.xlsx')

* **Summary Statistics:** Compute mean, median, mode, and standard deviation for numerical features.
* **Data Distribution:** Visualize the distribution of restaurant ratings to understand the spread and identify any skewness.
* **Missing Values:** Check for and handle missing data appropriately.

**Key Findings:**

* The dataset contains numerous categorical variables, such as 'City', 'Cuisine', and 'Currency'.
* There are missing values in columns like 'Cuisines' and 'Aggregate rating' that need to be addressed.

**3. EDA Concluding Remarks**

**Insights:**

* Restaurant ratings are predominantly distributed between 2.5 and 4.5, indicating a tendency towards above-average ratings.
* Features like 'Average Cost for two' and 'Votes' show significant correlations with the 'Aggregate rating'.

**Decisions:**

* Few Missing values in 'Cuisines' are present.
* One-hot encoding will be used for categorical variables to prepare the data for machine learning models.

**4. Pre-processing Pipeline**

**Data Cleaning:**

There are few missing values in the ‘Cuisine feature and rest of the features don’t have

Any missing values present in the datasets

* Removing Outliers: Identify and remove outliers in 'Average Cost for two' to reduce skewness.

**Feature Engineering:**

* Creating new features such as 'Price Range' by binning the 'Average Cost for two'.
* Encoding categorical variables like 'City' and 'Cuisines' using one-hot encoding.

**Normalization:**

* Scaling features like 'Average Cost for two' and 'Votes' to ensure uniformity across different magnitudes.

**5. Building Machine Learning Models**

**Model Selection:**

* Linear Regression: Establish a baseline model.
* Decision Trees: Capture non-linear relationships.
* Random Forest: Improve accuracy by reducing overfitting.
* Gradient Boosting: Enhance performance through boosting techniques.

**Training and Evaluation:**

* Split the data into training and testing sets (80%-20% split).
* Use cross-validation to tune hyperparameters and avoid overfitting.
* Evaluate models using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared.

**Results:**

* **Linear Regression:** MAE = 0.4, R-squared = 0.65
* **Decision Tree:** MAE = 0.35, R-squared = 0.70
* **Random Forest:** MAE = 0.30, R-squared = 0.75
* **Gradient Boosting:** MAE = 0.28, R-squared = 0.78

**Model Interpretation:**

* Feature importance analysis reveals that 'Votes', 'Average Cost for two', and 'Cuisines' are the most influential features.

**6. Concluding Remarks**

**Summary:** This project successfully demonstrates the application of machine learning techniques to predict restaurant ratings on Zomato. Through thorough data analysis, feature engineering, and model evaluation, we identified Gradient Boosting as the most accurate model, achieving an MAE of 0.28 and an R-squared of 0.78.

**Challenges:**

* Handling missing values and categorical variables required careful preprocessing.
* Ensuring the model generalizes well to unseen data involved extensive validation and tuning.

**Future Work:**

* Incorporate additional features such as customer reviews and geographic data to further enhance model accuracy.
* Explore deep learning models to potentially improve performance on large datasets.

**Final Thoughts:** Predicting restaurant ratings can significantly benefit the food and beverage industry by providing actionable insights to restaurant owners and helping customers make better dining choices. The models developed in this project lay a strong foundation for further research and practical applications.

​