Predicting Restaurant Ratings Using Machine Learning: A case study on Zomato data

# Introduction

The rise of online food delivery platforms has made user-generated restaurant ratings a crucial factor in consumer decision-making. This project aims to build a predictive model that can estimate a restaurant’s rating on Zomato using features like cost, votes, online ordering, and table booking availability. This study uses machine learning regression models to explore patterns and develop a system capable of predicting ratings based on key restaurant attributes.

# Dataset Overview

The dataset contains information about 51,717 restaurants in Bangalore, sourced from Zomato. It provides a rich collection of features such as restaurant name, location, cuisine type, cost, rating and other service-related attributes. The dataset contains 17 variables including both numerical and categorical. Variables:

1. name
2. url
3. online\_table
4. book\_table
5. rate (This is our target variable)
6. votes
7. phone
8. location
9. rest\_type
10. dish\_liked
11. cuisines
12. reviews\_list
13. approx\_cost (for two people)
14. menu\_item
15. listed\_in(type)
16. listed\_in(city)
17. address

# Methodology

## Data Cleaning

Unnecessary columns like url, address, phone, menu\_item, and dish\_liked were dropped due to high uniqueness or sparsity. Missing values in rate and approx\_cost were cleaned, and rate was converted to float.

## Feature Transformation

Skewed features such as votes and approx\_cost (for two people) were log-transformed. rate was normally distributed, so no transformation was applied.

## Encoding

Binary features (online\_order, book\_table) were label encoded. Multi-category columns like location, rest\_type, cuisines, etc., were one-hot encoded.

## Univariate Analysis

Distributions of key features like rate, votes, and approx\_cost were explored using histograms, KDE plots, and boxplots to understand spread and skewness.

## Bivariate Analysis

Relationships between rate and individual predictors were visualized using boxplots and scatterplots. It helped assess potential influence of features.

## Multivariate Analysis

Correlation and VIF were used to check multicollinearity. All selected features had VIF < 2, indicating low redundancy.

## Feature Selection

Features were selected based on business logic, low multicollinearity, and meaningful relationship with the target variable.

## Model Training and Evaluation

Two models were trained:

Linear Regression (R²: 0.41) — weak baseline.

Random Forest Regressor (R²: 0.68) — improved accuracy and lower error.

Evaluation metrics included MAE, MSE, RMSE, and R².

# Data Preparation

Irrelevant columns like url, address, phone, etc., were removed. Null values in the target column rate were dropped, while missing values in approx\_cost were filled with the median. Highly skewed columns like votes and approx\_cost were log-transformed.

Binary features were label encoded, and multi-class categorical columns were one-hot encoded to make the dataset model-ready.

# Model Building and Evaluation

We first trained a Linear Regression model, but it showed weak performance with an R² score of only 0.41, indicating it could not capture enough variance in ratings.

To improve the results, we trained a Random Forest Regressor, which gave significantly better performance:

MAE: 0.15

RMSE: 0.24

R² Score: 0.68

This shows that the Random Forest Regressor model was better suited for capturing complex, non-linear patterns in the data.

# Conclusion and Key insights

This project aimed to predict restaurant ratings using important features like cost, votes, online order availability, and table booking. While the Linear Regression model showed limited prediction ability, the Random Forest Regressor performed significantly better, capturing more complex relationships.

### Key Insights:

* Restaurants offering online ordering and table booking tend to have higher ratings.
* Higher customer votes are generally associated with better ratings.
* Approximate cost has a moderate influence on rating but not always directly proportional.