

Zomato Restaurant Rating

Zomato Restaurant Rating

Group Members:

First name	Last Name	Student number
Sneh	Patel	C0869367
Malav	Shah	C0870066
Dhruvraj	Chavda	C0867457
Nandiniben	Patel	C0869773

Table of Contents

Table of Contents

COVER PAGE.....	1
TOPIC NAME AND GROUP DETAILS.....	2
TABLE OF CONTENTS.....	3
ABSTRACT	4
INTRODUCTION	5
ABOUT DATASET	6
UNDERSTANDING PROBLEM STATEMENT	8
IMPORTING NECESSARY LIBRARIES	9
DATA PREPROCESSING	11
DATA VISUALIZATION	12
TRAINING AND TESTING DATA SET	15
MACHINE LEARNING MODELS	16
CONCLUSION.....	17
REFERENCES.....	18

Abstract

- The **Zomato.csv dataset** is a comprehensive collection of information about **restaurants in India**, gathered from the popular restaurant aggregator and food delivery platform Zomato.
- The dataset contains over **50,000 rows of data**, with **17 features** such as **restaurant ID, name, city, address, cuisines, average cost for two, ratings**, and more.
- The goal of this dataset is to provide valuable insights into the restaurant industry in India, including trends in **cuisine preferences, pricing, ratings**, and the **availability of table booking and online delivery services**.
- This information can be used by **researchers, food critics, and consumers** to make **informed decisions about dining choices**.

Introduction

- The Zomato dataset is a **real-time collection** of information about **restaurants all over India**.
- Zomato is a widely used **online food delivery app** that provides users with **ratings and reviews on restaurants**.
- These **ratings and reviews** play a crucial role in determining the **popularity and quality of restaurants**.
- In this **analysis**, we will focus on the city of **Bangalore** and use the **Zomato dataset** to gain **insights** into the factors that influence the establishment of different types of **restaurants** in different **areas** of the city.
- With over **12,000 restaurants** serving dishes from around the world, **Bangalore** is a city with a **diverse and thriving restaurant** scene.
- The dataset provides detailed information on various features of restaurants, such as their **location, cuisine, cost for two people, ratings**, and more.
- By analyzing this data, we can identify **patterns and trends in the restaurant industry in Bangalore**. We can also gain a better understanding of the factors that contribute to the **success** of different types of restaurants in different areas of the city.
- Overall, the Zomato dataset provides us with a **valuable resource** to understand the **restaurant industry in Bangalore**.
- The dataset is a powerful tool for gaining insights into the various factors that influence the establishment and success of restaurants and can be used to inform.

business decisions and consumer choices.

About Dataset:

- The **Zomato Bangalore Restaurants dataset** available on Kaggle contains information on restaurants in Bangalore, India. The dataset consists of **51717 rows and 17 columns**, with each row representing a different **restaurant in Bangalore**.

The columns in the dataset are:

- **url:** The url of the restaurant on the Zomato website.
- **address:** The address of the restaurant.
- **name:** The name of the restaurant.
- **online_order:** Whether online ordering is available or not (Yes/No).
- **book_table:** Whether table booking is available or not (Yes/No).
- **rate:** The overall rating of the restaurant on a scale of 1 to 5.
- **votes:** The total number of votes received by the restaurant.
- **phone:** The phone number of the restaurant.
- **location:** The locality in which the restaurant is situated.
- **rest_type:** The type of restaurant (e.g., Casual Dining, Cafe, Bar, etc.)

- **dish_liked:** The most popular dish at the restaurant.
- **cuisines:** The cuisines served at the restaurant.
- **approx_cost (for two people):** The approximate cost for two people to eat at the restaurant.
- **reviews_list:** A list of tuples containing reviews for the restaurant.
- **menu_item:** The menu items available at the restaurant.
- **listed_in(type):** The type of service offered by the restaurant (Delivery, Dine-out, etc.)
- **listed_in(city):** The city in which the restaurant is listed.
- The dataset provides valuable information about restaurants in Bangalore, including their **location, type of cuisine, cost, and popularity.**
- This information can be used to **analyze the restaurant industry in Bangalore** and to understand the factors that contribute to the **success or failure of restaurants in the city.**

Understanding Problem Statement

- The **primary objective** of this project is to conduct **in-depth Exploratory Data Analysis (EDA)** on the dataset to **gain insights** into the **restaurant industry in Bangalore, India**.
- The analysis will involve examining various factors that affect the **establishment and success** of different types of restaurants in the city, including the types of **cuisine served, the pricing trends**, and the **popularity of various areas**.
- Furthermore, we will aim **to develop** a suitable **machine learning model** that can accurately **predict the rating** of a restaurant based on **specific features**.
- The model will be developed using the insights gained from the EDA, and it will be designed to assist Zomato restaurants in **predicting their ratings accurately**.
- In summary, this project will help us understand the restaurant industry in Bangalore, India, by conducting a detailed EDA and developing a machine learning model that can predict restaurant ratings accurately.
- The project's results will be useful to various **stakeholders, including food critics, researchers, consumers, and restaurant owners and managers**, by providing valuable insights.

Importing Necessary Libraries

- **pandas:** Pandas is a powerful data manipulation library that allows data to be easily manipulated and analyzed. It provides data structures for efficiently storing and manipulating large datasets.
- **numpy:** NumPy is a library for numerical computing with Python. It provides fast and efficient numerical operations on arrays and matrices.
- **Label Encoder:** LabelEncoder is a utility class from the sklearn library that allows for converting categorical data into numerical labels, which is required for machine learning models to interpret the data.
- **train_test_split:** train_test_split is a utility function from the sklearn library that allows splitting the dataset into training and testing subsets. It is an essential tool for evaluating the performance of machine learning models.
- **Linear Regression:** Linear Regression is a linear regression model from the sklearn library. It is a simple and widely used model for predicting continuous variables.
- **Decision Tree Regressor:** DecisionTreeRegressor is a decision tree-based regression model from the sklearn library. It is a non-linear model that is suitable for complex datasets.
- **RandomForestRegressor:** RandomForestRegressor is a random forest-based regression model from the sklearn library. It is an ensemble method that combines multiple decision trees to produce a more accurate model.
- **ExtraTreesRegressor:** ExtraTreesRegressor is another ensemble-based regression model from the sklearn library that is like RandomForestRegressor.
- **r2_score:** r2_score is a metric that measures the performance of regression models by calculating the coefficient of determination, which represents the proportion of variance in the target variable that is explained by the model.
- **matplotlib.pyplot:** matplotlib.pyplot is a plotting library that provides a range of

functions for creating different types of visualizations. It is a widely used tool for data visualization and analysis.

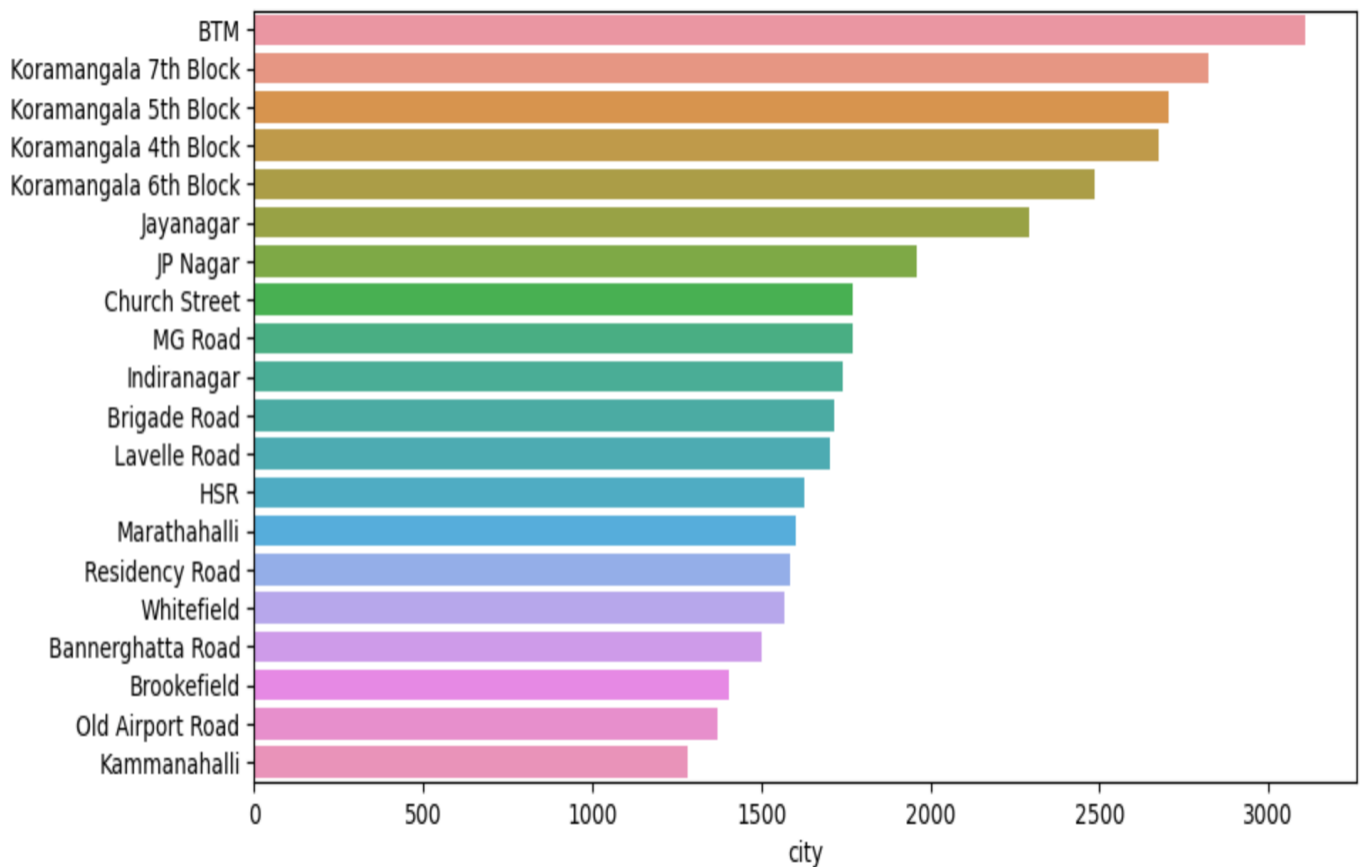
- **seaborn:** seaborn is a visualization library that is built on top of matplotlib. It provides additional visualization functions that can help in creating more complex visualizations and has a more user-friendly interface.

Data Preprocessing

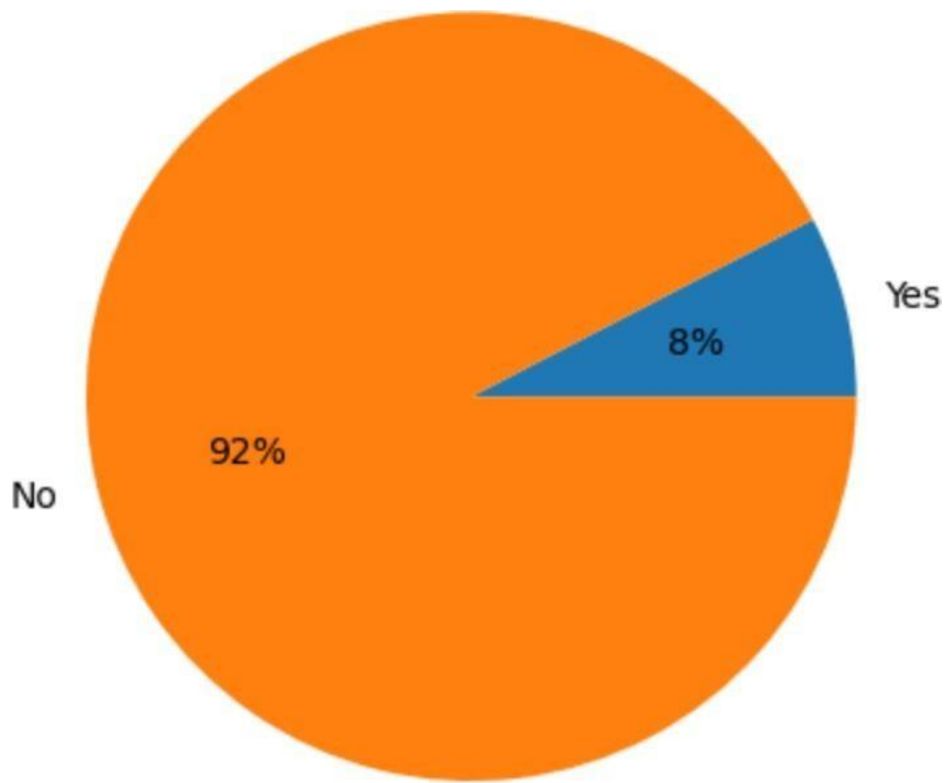
- **Loading the dataset:** The dataset needed to be loaded into a panda Data Frame as the first step. The dataset was read from the CSV file using the `read_csv ()` function. Then, we looked at the data to see if there were any anomalies or missing values that might have an impact on our analysis.
- **Handling missing values:** We noticed that some of the dataset's columns had blank values. We dealt with missing values by removing columns that had more than 50% of their values missing and filling in the gaps with the column's mode.
- **Cleaning the data:** To clean the data, we eliminated any duplicate records and dropped any unnecessary columns. Additionally, we made a few minor changes to the data, such as changing the strings to lowercase and removing the currency symbol from the cost columns.
- **Encoding categorical variables:** We used the `LabelEncoder` function from the `sklearn` library to encode categorical variables as numerical labels. Given that machine learning models typically require numerical data, this was necessary.
- After preprocessing the data, we split the dataset into **training and test sets** and applied various **machine learning algorithms** to **predict the rating of restaurants**.
- Overall, the preprocessing steps performed on the Zomato dataset were necessary to ensure that the data was in a suitable format for analysis and helped to avoid any potential errors or biases in our analysis.

Data Visualization

- **Data visualization** is an essential aspect of data analysis as it helps to **gain insights into the data** by presenting it in a graphical format.
- In this project, we used various visualization techniques to explore the Zomato dataset and understand the relationship between the different variables.
- We used **matplotlib and seaborn libraries** to create visualizations such as **bar plots, plots, pie chart**.
- These visualizations helped us to understand the distribution of different variables and their relationships with each other.
- We also used visualizations to identify **trends and patterns in the data**, such as the distribution of the ratings of different restaurants and the types of cuisines that are most popular in Bangalore.
- Overall, data visualization played a critical role in helping us gain insights into the Zomato dataset and provided a clear and concise way to present **our findings to stakeholders**.
- [Following image describe Finding top 20 city with having more restaurants.](#)
- The **top 20 cities** in the Zomato dataset are displayed in a horizontal bar plot using the seaborn library, based on the number of restaurants in each city.
- This visualization aids in understanding how restaurants are distributed among various cities and can reveal which **cities have the highest concentration of restaurants in the dataset**.



- [Following image describe Restaurants are having both online and table booking.](#)
- The proportion of restaurants in the Zomato dataset that provide both online ordering and table reservations is depicted in this pie chart.
- This visualization explains how many **restaurants** in the dataset **offer both online order and table booking services** as well as what proportion of restaurants do so.



Training and Testing Data Set

- In machine learning, **training and testing** are crucial steps.
- The training set and the testing set are created after the data has **undergone preprocessing**.
- The model is trained on the training set, and the testing set is used to assess the **model's performance**.
- The objective is to develop a model that accurately predicts the target variable and generalizes well to new data.
- Predicting the restaurant rating using the Zomato dataset's various features is the objective. We must divide the dataset into a training set and a testing set in order to accomplish this.
- The machine learning model will be trained using the training set, and its performance will be assessed using the testing set.
- A typical split is **25%** for testing and **75%** for training. In other words, 75% of the data are used to train the model, and the remaining 25% are used to assess how well the trained model performs.

Machine Learning Models

- **Linear Regression:** Linear Regression is a simple yet powerful algorithm that works by finding the best line that fits the data.
- **Decision Tree:** A Decision Tree is a flowchart-like structure in which each internal node represents a "test" on a feature, each branch represents the outcome of the test, and each leaf node represents a class label (or regression value) assigned to the instance.
- **Extra Trees:** Extra Trees is another ensemble learning algorithm that builds multiple decision trees in a random fashion and combines their predictions to obtain a more accurate and stable prediction.
- **Random Forest:** Random Forest is an ensemble learning algorithm that builds multiple decision trees and combines their predictions to obtain a more accurate and stable prediction.
- All of these algorithms can be applied to the Zomato dataset to create regression models that forecast a restaurant's rating based on specific features.

Conclusion

- Based on the accuracy scores, it appears that the **Extra Trees model** performed the best with an **accuracy score** of 0.886. T
- The **Random Forest model** performed the second-best with an **accuracy score of 0.821**, followed by the **Decision Tree model** with an accuracy score of **0.783**.
- The **Linear Regression model** had the lowest accuracy score of **0.268**.
- This indicates that the **Extra Trees model** is the **best fit** for the given data and can be used **to predict restaurant ratings based on various features**.
- Our analysis demonstrated that characteristics like **location, restaurant type, and cuisine** were **significant predictors of restaurant rating**.
- Restaurant owners and investors may find this information helpful in making decisions about where to locate their businesses and what kinds of foods to serve.

References

Zomato Bangalore Restaurants. (2019, March 31). Kaggle.

<https://www.kaggle.com/datasets/himanshupoddar/zomato-bangalore-restaurants?resource=download>

MaharanaSaroj. (2022, December 5).

ZomatoResturantRating_End2End/Zomato_Resturant_EDA_RatingPrediction.ipynb at main · MaharanaSaroj/ZomatoResturantRating_End2End. GitHub.

https://github.com/MaharanaSaroj/ZomatoResturantRating_End2End/blob/main/Zomato_Resturant_EDA_RatingPrediction.ipynb