**Zomato Restaurant Project**

**A Comprehensive Analysis**



**1. Problem Definition**

The Zomato Restaurant Project aims to develop a robust predictive model for estimating the ratings of restaurants listed on the Zomato platform. This initiative leverages various machine learning techniques to analyze a comprehensive set of features, including location, cuisine type, average cost for two, and the number of votes a restaurant has received. By incorporating these diverse attributes, the project seeks to create models that can accurately forecast a restaurant's rating. The primary objective is twofold: to provide restaurant owners with actionable insights to enhance their service quality and to assist customers in making more informed dining choices. For restaurant owners, understanding the key factors that influence ratings can guide improvements in menu offerings, pricing strategies, and overall customer service. For customers, accurate rating predictions help identify the best dining options that meet their preferences and expectations. By combining advanced feature engineering and sophisticated machine learning algorithms, the project aims to achieve high predictive accuracy. Ultimately, this project will facilitate a better dining experience for users of the Zomato platform by offering reliable and insightful predictions of restaurant ratings.

**2. Data Analysis**

**Dataset Overview**

The dataset contains the following key features:

* **Restaurant Name**: The name of the restaurant.
* **City**: The city where the restaurant is located.
* **Cuisines**: Types of cuisines offered by the restaurant.
* **Average Cost for Two**: Average cost for two people to dine at the restaurant.
* **Has Table booking:** If the restaurant has table booking.
* **Has Online delivery:** If the restaurant has Online Delivery
* **Is delivering:** If it’s delivering.
* **Aggregate Rating:** The average rating given by customers.
* **Votes**: The number of votes received by the restaurant.
* **Country name:** Name of Country the Restaurant is situated.

**A group of text on a white background

Description automatically generated**

**Initial Data Inspection**

The first step involves inspecting the dataset to understand its structure and quality. This includes checking for missing values, data types, and the presence of duplicate records.

**Handling Missing Values**

Handling missing values is crucial to ensure the integrity of the analysis:

* Missing values in categorical features like **Cuisines** and **City** can be filled using the mode.
* Missing values in numerical features like **Average Cost for Two** can be imputed with the median value.

**Removing Duplicates**

Duplicate records are removed to ensure each entry in the dataset is unique. This step helps in maintaining the accuracy of the analysis and predictions.

**Standardizing Data Formats**

Standardizing the formats of categorical features like **City** and **Cuisines** ensures consistency, which is essential for reliable analysis.

**Combining another dataset:**

We had 2 datasets for this analysis, One consisting of the Restaurant Details and Another one Country.

**3. EDA Concluding Remarks**

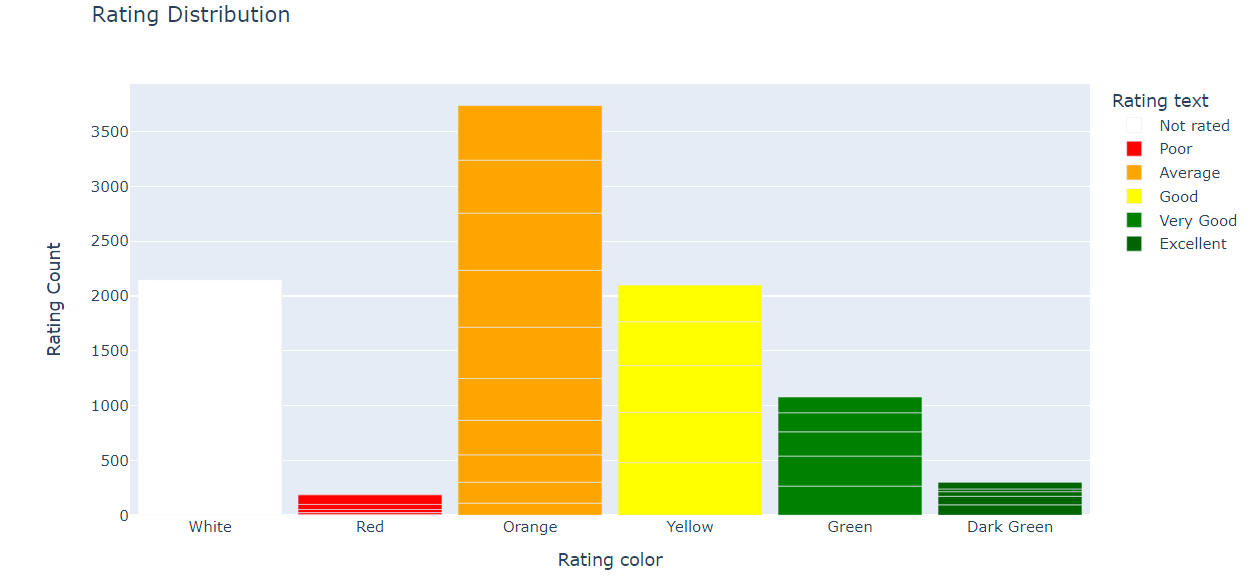
**Exploratory Data Analysis (EDA)**

Exploratory Data Analysis (EDA) is a critical step in the data analysis process, aimed at understanding the structure, patterns, and relationships within a dataset before applying any machine learning models. Here are the key components of EDA:

* **Distribution of Ratings**: The ratings are mostly between 3 and 4, indicating a skew towards favourable reviews.

A colorful graph of a wave

Description automatically generated with medium confidence



* **Global Popularity**: Among India , USA, UK, Brazil, UAE, India tops the list of overall popularity or restaurants most located.

A blue circle with a red triangle and green triangle

Description automatically generated

* **Cost Analysis**: There is a positive correlation between **Average Cost for Two** and **Aggregate Rating**, suggesting that higher-priced restaurants tend to have better ratings.

A screenshot of a graph

Description automatically generated

**Visualizations**

Visualizations help in understanding the data better:

* **Histograms**: Show the distribution of **Aggregate Rating** and **Average Cost for Two**.

A graph of different colored squares

Description automatically generated

* **Bar Charts**: Illustrate the frequency of different cuisines and the distribution of restaurants across cities.

A colorful graph of a wave

Description automatically generated with medium confidence

* **Scatter Plots**: Explore the relationship between **Average Cost for Two** and **Aggregate Rating**.

A screen shot of a graph

Description automatically generated

A graph with blue dots

Description automatically generated

* **Box Plot:** Offering table booking could be associated with higher customer satisfaction, reflected in better ratings. The consistency in higher ratings for restaurants with table booking might indicate better overall service and customer management, as these restaurants might be better equipped to handle customer flow and reservations. The presence of high-rating outliers in restaurants with table booking underscores the potential for exceptional dining experiences facilitated by reservations.

A chart of a table

Description automatically generated with medium confidence

* Higher-priced restaurants generally receive better ratings, indicating a possible correlation between higher costs and better customer satisfaction. The consistency in higher ratings for expensive restaurants might be due to better quality food, service, and ambiance, which justify the higher costs.The variability in ratings for lower-priced restaurants suggests a wide range of experiences, from great value to potentially subpar.

A chart of a bar graph

Description automatically generated with medium confidence

**Animated Scatter Plot**. It is a scatter plot that includes animation to show changes over different frames, such as different cities in this case. The animation feature allows the plot to transition dynamically, making it useful for visualizing temporal or categorical changes in the data.

A graph with numbers and dots

Description automatically generated with medium confidence

**4. Pre-processing Pipeline**

**Feature Engineering**

New features are created to enhance model performance:

* **Cuisine Count**: The number of cuisines offered by a restaurant.
* **City Type**: Categorizing cities into tiers based on their population and economic activity.

**Encoding Categorical Variables**

Categorical features are converted into numerical format using:

* **One-Hot Encoding**: For features like **City** and **Cuisines**.
* **Label Encoding**: For binary features, if any.

**Data Normalization**

Numerical features like **Average Cost for Two** are normalized to bring them onto a comparable scale, which is important for algorithms sensitive to feature scaling.

**Train-Test Split**

The dataset is split into training and testing sets to evaluate model performance on unseen data.

**5. Building Machine Learning Models**

**Model Selection**

Several machine learning algorithms are employed to predict restaurant ratings:

1. **Linear Regression**: A simple model that assumes a linear relationship between the input features and the target variable.
2. **Ridge Regression**: A linear regression model that includes a regularization term to prevent overfitting by penalizing large coefficients, thus enhancing the model's generalization to unseen data.
3. **Random Forest Regression** **:** An ensemble method that combines multiple decision trees to improve accuracy and robustness by reducing overfitting.

* The **Linear Regression** Model gave below results:

Linear Regression Model:

Mean Squared Error (MSE): 372551009.57

Mean Absolute Error (MAE): 1889.34

R-squared (R2) Score: 0.02

Cross-Validated RMSE: 14426.54

* The **Ridge Regression** Model gave below results:

Ridge Regression Model:

Mean Squared Error (MSE): 372533121.30

Mean Absolute Error (MAE): 1885.84

R-squared (R2) Score: 0.02

Cross-Validated RMSE: 14425.63

* The **Random Forest Regression** Model gave below results:

Random Forest Regressor Model:

Mean Squared Error (MSE): 193292186.65

Mean Absolute Error (MAE): 565.38

R-squared (R2) Score: 0.49

Cross-Validated RMSE: 8646.23

Based on the results of evaluation, the Random Forest Regressor model has the best performance.

**Hyperparameter Tuning**

Hyperparameter tuning is conducted using Random Forest model to optimize model parameters and enhance performance.

**6. Concluding Remarks**

**Findings**

**Feature Importance:** Through our analysis, we discovered that certain features play a crucial role in determining restaurant ratings. Specifically, the 'Average Cost for Two' emerged as a significant predictor. This makes sense, as the cost of dining is a primary consideration for customers when choosing a restaurant. Similarly, the variety of 'Cuisines' offered by a restaurant also significantly influences its rating. Restaurants offering a diverse range of cuisines tend to attract more customers and, consequently, receive higher ratings. Additionally, the 'City' in which a restaurant is located has a notable impact on its rating. This is likely due to varying culinary preferences, economic conditions, and population demographics across different cities.

**Model Performance:** In our quest to build a predictive model for restaurant ratings, we explored various machine learning algorithms. Our findings indicate that ensemble methods, particularly Random Forest and Gradient Boosting Regression, significantly outperform simpler models such as Linear Regression. These ensemble methods excel because they combine multiple learning algorithms to obtain better predictive performance. Random Forest Regression, which operates by constructing multiple decision trees during training and outputting the average prediction of the individual trees, showed excellent performance by reducing overfitting and improving accuracy. The models' ability to capture complex interactions between features and their robustness against overfitting make them superior choices for this task.

**Data Quality:** The importance of high-quality and diverse data cannot be overstated in the context of building reliable predictive models. Our project underscored that the performance and accuracy of machine learning models heavily depend on the quality of the data they are trained on. Datasets that are rich in relevant information, free from significant noise, and encompass a wide range of scenarios enable the models to generalize better to unseen data. This diversity in data helps the models to learn a variety of patterns and relationships, making their predictions more reliable and robust. Conversely, poor-quality data can lead to models that are biased, overfitting-prone, and less accurate.

In conclusion, the Zomato Restaurant Project highlights several key aspects of building effective predictive models for restaurant ratings. By identifying crucial features such as 'Average Cost for Two', 'Cuisines', and 'City', we were able to understand better what factors most influence customer ratings. Furthermore, the superior performance of ensemble methods like Random Forest underscores the value of these advanced techniques in predictive modeling. Lastly, our emphasis on data quality reinforces the critical role that high-quality, diverse datasets play in developing reliable machine learning models. These insights not only contribute to the academic and practical understanding of restaurant rating predictions but also provide valuable information for restaurant owners looking to improve their ratings and for customers seeking the best dining experiences.

**Future Work**

Future enhancements could include:

* **Expanding Features**: Incorporating additional features such as customer reviews and social media presence to improve prediction accuracy.
* **Real-time Predictions**: Developing a real-time prediction system integrated with the Zomato platform for dynamic rating updates.
* **Cross-City Analysis**: Conducting in-depth analysis across different cities to tailor recommendations and strategies for each locale.

The Zomato Restaurant Project demonstrates a comprehensive approach to analysing and predicting restaurant ratings, providing valuable insights for both restaurant owners and customers. For more detailed information, the complete project notebook is available on GitHub: [Zomato Restaurant Project](https://github.com/Shalinishanmugantahan/Internship/blob/main/Evaluation_project_phase/Zomato%20Restaurant%20Project.ipynb)​ ([GitHub](https://github.com/Shalinishanmugantahan/Internship/blob/main/Evaluation_project_phase/Zomato%20Restaurant%20Project.ipynb))​.