RESTAURANT RATINGS TECHNICAL REPORT

CONTENTS

Executive Summary

FoodieBay, a leading restaurant aggregator, embarked on a data-driven journey to enhance customer experiences, optimize business strategies, and enable efficient decision-making. This project aimed to uncover insights into the factors influencing restaurant ratings and explore machine learning for predicting these ratings.

In the realm of supervised machine learning, two models were employed: Linear Regression and Decision Tree Regressor. The Decision Tree model outperformed, explaining 56.5% of rating variance, signifying its capability to capture intricate relationships between features and ratings.

Unsupervised machine learning, specifically K-Means clustering, was used to group restaurants into clusters. Four clusters emerged as the optimal choice, indicating distinct attributes shared by restaurants within each cluster.

The results reveal a potential for data-driven decisions, enabling FoodieBay to tailor business solutions for partner restaurants and elevate customer satisfaction. The Decision Tree Regressor model offers a robust predictive tool for restaurant ratings.

For ongoing relevance, FoodieBay must periodically update its dataset, retrain models, engage with customers, and collaborate with restaurants. This proactive approach ensures continued success in the dynamic food industry.

This project marks a significant step towards FoodieBay's commitment to data-driven excellence, customer satisfaction, and business growth.

Introduction

FoodieBay, an Indian multinational restaurant aggregator, has been at the forefront of the digital platform revolution, offering customers a convenient way to explore, order, and review restaurants across multiple cities and countries. However, understanding what factors influence restaurant ratings is crucial for both enhancing the dining experience and optimizing business strategies.

Objective
To gain deeper insights into the factors that influence restaurant ratings on the FoodieBay platform. There are two primary analytical tasks in mind:

- To uncover meaningful insights from the dataset provided by FoodieBay.
- To explore the potential for machine learning in predicting restaurant ratings.

Value Proposition
The value proposition of this project is multifaceted:

- Enhanced Customer Experience: By gaining insights into the factors influencing restaurant ratings, FoodieBay can make informed decisions to improve the dining experience for its users.
- Optimized Business Strategies: Understanding the drivers of restaurant ratings enables FoodieBay to offer tailored business solutions to its partner restaurants.
- Efficient Decision-Making: The machine learning model developed as part of this project will empower FoodieBay with the capability to predict restaurant ratings.

Approach

In the context of exploring the FoodieBay dataset and applying both supervised and unsupervised machine learning techniques, we aim to gain insights into the factors influencing restaurant ratings and create meaningful clusters of restaurants based on specific attributes.

Supervised Machine Learning
Initially, we adopt a supervised machine learning approach to predict restaurant ratings based on relevant features. This falls under the category of regression problems as we aim to predict a continuous numerical value (restaurant ratings).

- Linear Regression:
  We utilize a Linear Regression model to establish a linear relationship between the selected features and restaurant ratings. The dataset is split into training and test sets (80% for training and 20% for testing) to evaluate model performance. We evaluate the model using various metrics such as Mean Squared Error (MSE), R-squared (R2), and Root Mean Squared Error (RMSE).
- Decision Tree Regressor:
  We also employ a Decision Tree Regressor, capable of capturing non-linear relationships between features and restaurant ratings. Similar to Linear Regression, we split the dataset, train the model, and evaluate its performance using metrics like Mean Absolute Error (MAE), MSE, RMSE, and R2.

Unsupervised Machine Learning
Subsequently, we employ unsupervised machine learning techniques, specifically K-Means clustering, to group restaurants into clusters based on their attributes. This is an unsupervised clustering problem. The optimal number of clusters is determined through the Elbow Method and Silhouette Score. The cluster assignments are added to the original dataset to facilitate post-analysis.

Data Preparation and Exploratory Data Analysis (EDA)

Data Sources and Size:
The dataset provided is a subset of data from FoodieBay, a restaurant aggregator platform operating across various cities in India and other countries. There are 40131 observations and 17 variables,

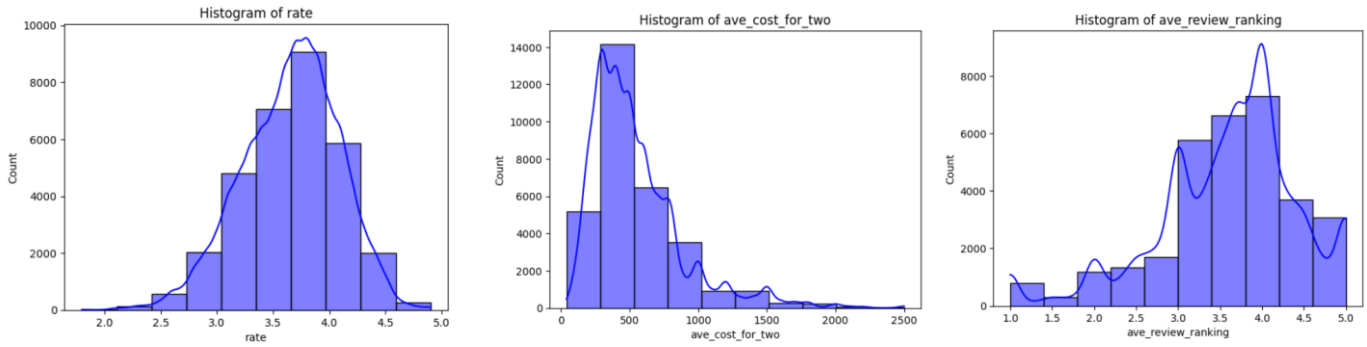
Data Types:
The dataset includes both numerical and categorical features.
Numerical features include 'ave_cost_for_two,' 'ave_review_ranking,' 'votes,' and 'rate.'
Categorical features include 'online_order' and 'book_table.'

Data Cleansing:

- Treating Missing Values
  Rows with missing values in the 'rate' column (target variable) are removed using This is a reasonable approach because missing target values cannot be imputed. Both the review_ranking' and 'ave_cost_for_two' columns are skewed. Due to its resilience to outliers and capacity to produce a representative number, we therefore use the median to fill in for the missing data.
- Converting Categorical Variables to Numerical:
  Categorical variables ('online_order' and 'book_table') are converted to numerical format for use in machine learning models. For 'online_order' and 'book_table,' 'Yes' is encoded as 1, and 'No' is encoded as 0. Additionally, the 'rest_type' and 'listed_in_type' columns, have been mapped specifically with values from 1 to 7.

Exploratory Data Analysis (EDA)

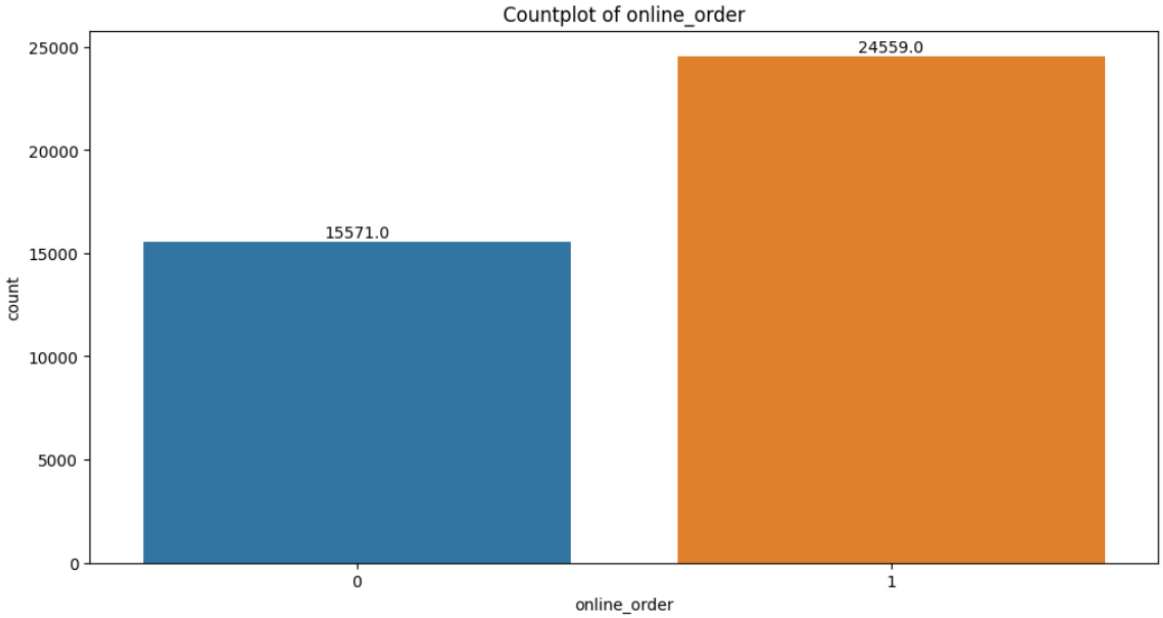
Univariate



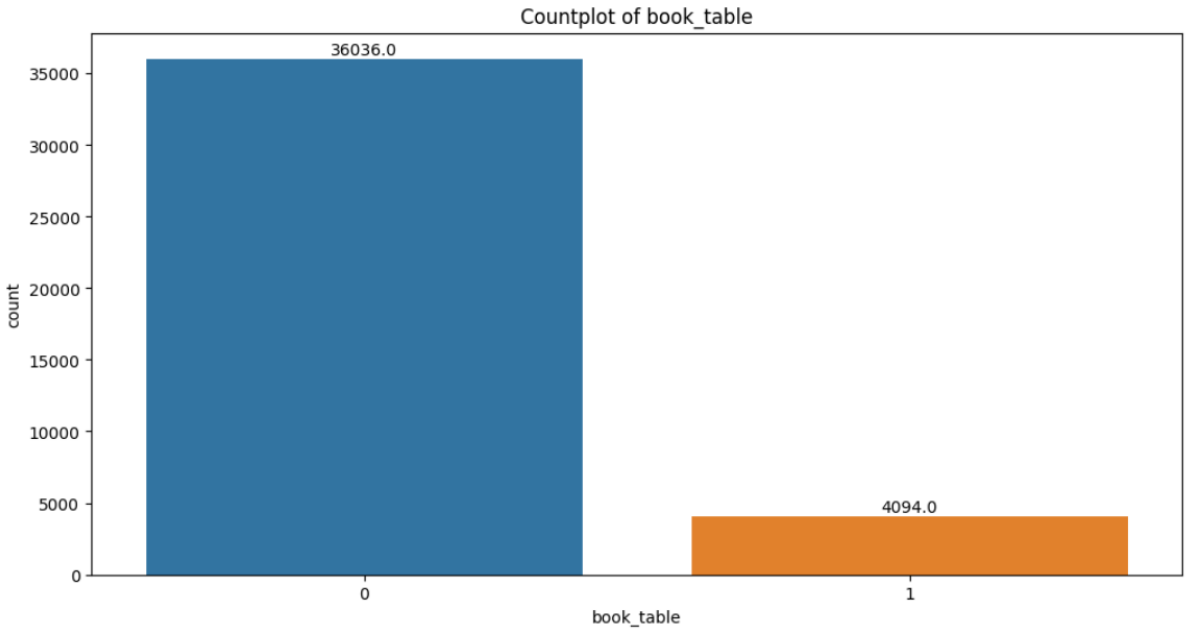
The distribution of user ratings is right-skewed. The central tendency, or average rating, is approximately 4.0. The spread of the data is relatively small, meaning that most restaurants have ratings that are close to the average.

The most popular price point for restaurants in India is ₹250-₹500 for two. There are a significant number of affordable restaurants with an average cost for two below ₹500. The spread of the data suggests that there is a wide range of options available to consumers, regardless of their budget.

The histogram for average review ranking is right-skewed. The central tendency of the data is around 4.5.The spread of the data is relatively small.

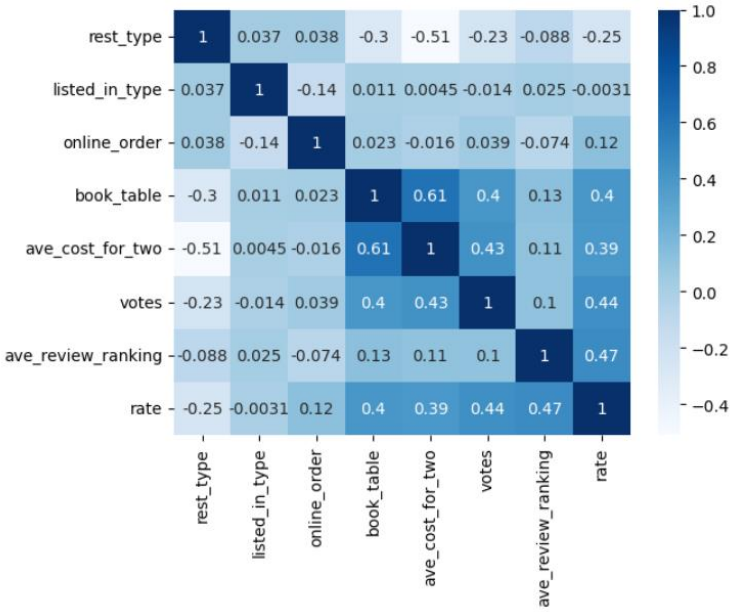


It can be observed most restaurants have a relatively low number of online orders. About 61% of the total orders are online orders



It is evident that the booking of table does not take place very often accounting for 89.7% of observations in the entire dataset.

Multivariate Analysis



During the process of feature selection, we have taken into account the correlation between certain variables and the target variable 'rate,' as well as their potential impact on predictive modelling and analysis. Here is a summary of our key findings:

- Average review ranking (Correlation: 0.47):
- Votes (Correlation: 0.44):
- Average cost for two (Correlation: 0.38):
- Book table (Correlation: 0.4):
- Online order (Correlation: 0.12):

On the other hand, variables related to 'Listed in type' and 'restaurant type' represent the locations and types of restaurant outlets, but they do not have a substantial impact on ratings.

Model Development and Evaluation

Supervised Machine Learning:

We implemented two supervised machine learning models. These models were used to predict restaurant ratings based on selected features, including 'online_order,' 'book_table,' 'ave_cost_for_two,' 'ave_review_ranking,' and 'votes.' The dataset was further split into 80:20 ratio into training and testing sets.

Linear Regression Model:

The Linear Regression model aimed to establish a linear relationship between the selected features and restaurant ratings.

Performance Metrics:

- Mean Squared Error
  The MSE is approximately 0.1048, which means that, on average, the model's predictions deviate from the actual ratings by this amount squared.
- R-squared
  An R2 score of 0.4381 indicates that the model explains approximately 43.81% of the variance in restaurant ratings. This suggests that the model captures a moderate amount of the variation in ratings but may not account for all influencing factors.
- Root Mean Square Error
  RMSE is the square root of the MSE and is expressed in the same units as the target variable (restaurant ratings in this case).

Decision Tree Regressor Model:

The Decision Tree Regressor captured complex non-linear relationships between features and restaurant ratings.

Performance Metrics:

- Mean Absolute Error
  The MAE is approximately 0.204, which means that, on average, the model's predictions deviate from the actual ratings by this amount.
- Mean Squared Error
  An MSE of approximately 0.081 suggests that, on average, the squared prediction errors are relatively small.
- Root Mean Square Error
  An RMSE of approximately 0.285 indicates that, on average, the model's predictions deviate from the actual ratings by this amount.
- R-squared
  An R2 score of 0.565 indicates that the model explains approximately 56.5% of the variance in restaurant ratings. This suggests that the model captures a relatively good amount of the variation in ratings, indicating a reasonable level of predictive power.

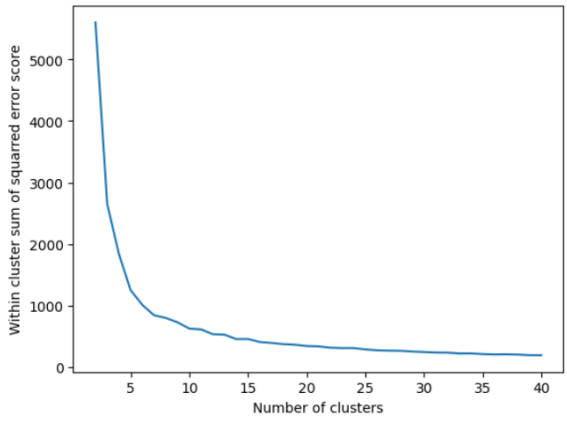
Unsupervised Machine Learning:

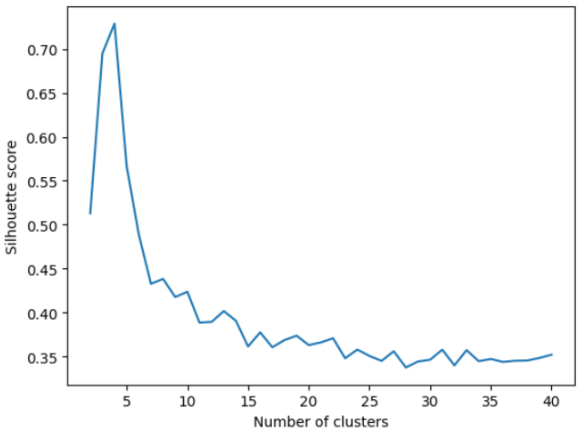
Selected features have undergone Min-Max scaling to standardize these features, ensuring that each feature has the same scale, which is important for K-Means clustering.

Cluster Evaluation Metrics

- Within-Cluster Sum of Squares
  The WCSS value of 1249.999 suggests that, on average, the data points within each cluster are relatively close to their cluster centroids.
- Davies-Bouldin Index
  The computed DBI value is 0.561, which indicates that the clusters are reasonably well-separated and compact. A DBI value below 1 is generally considered good.
- Silhouette Score
  A Silhouette Score of 0.566 indicates that the clusters are reasonably well-separated and data points within each cluster are relatively similar to each other compared to data points in other clusters.

Determining the Number of Clusters





When the number of clusters is raised from 3 to 4, the WCSS drops significantly. A Silhouette Score of 0.729 suggests that the data points within each cluster are similar to each other and distinct from data points in other clusters at k=4 and hence it is the most suitable choice.

Solution and Recommendation

Interpretation and Discussion of Results:

Supervised Machine Learning:

- The Linear Regression model explains approximately 43.81% of the variance in restaurant ratings. It captures a moderate amount of variation but may not account for all influencing factors.
- The Decision Tree Regressor model, with an R2 score of approximately 56.5%, performs relatively well in explaining the variation in restaurant ratings.
- Both models provide insights into the relationships between selected features (e.g., online orders, cost for two, review ranking, votes) and restaurant ratings.

Unsupervised Machine Learning:

- The optimal number of clusters (k) was determined to be 4, based on the Elbow Method and Silhouette Score.
- The clusters were found to be reasonably well-separated and data points within each cluster were relatively similar to each other compared to data points in other clusters.

Solution recommendation

The Decision Tree Regressor model appears to perform better than the Linear Regression model, based on the evaluation metrics. The relatively low MAE (0.204) and RMSE (0.285) values indicate that the model's predictions are reasonably close to the actual ratings on average. The R-squared value of 0.565 indicates that the Decision Tree model explains a significant portion of the variance in restaurant ratings. This suggests that the model is more effective at capturing the relationships between the input features and ratings.

Future engagements with the client

- Ongoing monitoring and analysis of restaurant ratings to identify trends and opportunities.
- Further refinement of the machine learning model to improve predictive accuracy.
- Exploration of additional data sources or features that may enhance the understanding of restaurant ratings.
- Collaboration with partner restaurants to implement recommendations and measure the impact on ratings and customer satisfaction.

Technical recommendations

Suggestions for maintenance of accuracy and relevance over time

- Regular Data Updates: Periodically update the dataset with new restaurant data and customer reviews. This will help the models adapt to changing trends and preferences.
- Retraining Models: Revisit and retrain the machine learning models when significant changes occur in the restaurant industry or customer behaviour.
- Customer Engagement: Engage with customers to gather feedback and preferences. Use customer surveys and reviews to fine-tune restaurant recommendations and enhance the user experience.
- Regular Model Evaluation: Periodically assess model performance using updated evaluation metrics and consider retraining or fine-tuning models accordingly.
- Collaboration with Restaurants: Collaborate with partner restaurants to gather insights and feedback on the impact of recommendations and optimizations. This can lead to mutually beneficial improvements in service quality.