Exploring Restaurant Data and Predicting Ratings: A Comparative Analysis of Bangalore and Pune Using Machine Learning

Table of Contents:

5. Results and Analysis

- Predicted Ratings: Comparing Predicted Ratings with Actual Ratings
- Accuracy Metrics: Evaluating the Performance of the Model
- Comparison of Bangalore and Pune Restaurants: Analyzing Rating Distributions and Patterns
- Key Findings: Factors Influencing Ratings, Similarities and Differences between the Cities

6. About Zomato: Introduction to Zomato and its Role in the Restaurant Industry

- Overview of Zomato: Features, Services, and Popularity
- Importance of Zomato Data: Insights for Restaurants, Customers, and Stakeholders

7. Explanation of the Code

8. About Data Visualization

9. Result prediction

10. Conclusion

- Summary of the Project: Objectives, Methodology, and Findings
- Achievements: Successful Rating Prediction and Analysis
- Limitations: Data Limitations, Scope of Analysis
- Future Work and Recommendations: Improving the Model, Expanding the Analysis

11. References

- List of Sources, Datasets, and References Used

Introduction:

The purpose of this project is to explore restaurant data from two major cities in India, Bangalore and Pune, and use machine learning techniques to predict restaurant ratings. By analyzing the data, we aim to gain insights into the restaurant landscape of these cities and build a model that can accurately predict ratings based on various features.

Data Collection and Preprocessing:

We collected the restaurant data from Zomato, a popular online platform that provides information about restaurants, menus, reviews, and ratings. We utilized the Zomato datasets for Bangalore and Pune, which include detailed information about the restaurants, such as their names, categories, pricing, locality, ratings, and review counts. The data preprocessing steps involved cleaning the data, handling missing values, and transforming the data types to ensure consistency and accuracy in the analysis.

Exploratory Data Analysis:

In the exploratory data analysis phase, we delved into the characteristics of the restaurant data. We performed statistical analysis to understand the central tendencies, dispersions, and distributions of the data. Visualizations such as bar charts, pie charts, scatter plots, and heatmaps were employed to identify trends, patterns, and correlations between different features. This analysis provided valuable insights into popular restaurant categories, pricing trends, and the relationships between various attributes.

Machine Learning Model for Rating Prediction:

To predict restaurant ratings, we selected the Random Forests algorithm, a powerful machine learning technique known for its accuracy and robustness. Prior to model training, we performed feature selection and engineering to identify the most relevant features for the

prediction task. The Random Forests algorithm was then trained on the preprocessed data, where it learned patterns and relationships between the features and the ratings.

Results and Analysis:

The trained machine learning model was evaluated using appropriate accuracy metrics, such as Mean Squared Error (MSE), to assess its performance. The predicted ratings were compared with the actual ratings from the dataset to measure the effectiveness of the model. Furthermore, we conducted a comparative analysis of the restaurant landscapes in Bangalore and Pune, examining the rating distributions and identifying any notable patterns or differences between the two cities. The findings from this analysis provided valuable insights into the factors influencing ratings and allowed for meaningful comparisons between the restaurant scenes in Bangalore and Pune.

About Zomato:

Zomato is a leading online platform that connects users with restaurants, providing comprehensive information about their menus, reviews, ratings, and more. It plays a significant role in the restaurant industry by bridging the gap between restaurants and customers. The availability of Zomato data offers immense value to restaurant owners, customers, and stakeholders, enabling them to make informed decisions regarding dining choices, menu planning, marketing strategies, and more.

Explanation of the Code:

This code performs the following steps:

1. **Data Loading**: The code imports the required libraries and loads the restaurant data for Bangalore and Pune from separate CSV files into Pandas DataFrames: `/content/Bangalore_Restaurants.csv` and `/content/Pune Restaurants.csv`.

2. **Data Concatenation**: The code concatenates the Bangalore and Pune DataFrames using the `concat` function from Pandas, resulting in the `combined_df` DataFrame. This combines the data from both cities into a single dataset for analysis.

3. **Data Preprocessing and Cleaning**: The code preprocesses the combined dataset by performing several data cleaning operations:

   - Unnecessary columns such as 'Website', 'Address', and 'Phone_No' are dropped using the `drop` function.

   - Missing values are handled by dropping rows containing missing values using the `dropna` function.

4. **One-Hot Encoding**: The categorical variables in the dataset, such as 'Restaurant_Name', 'Category', 'Locality', 'Known_for1', and 'Known_for2', are encoded using one-hot encoding. The `get_dummies` function from Pandas is used to create binary columns for each unique category.

5. **Restaurant Rating Prediction**: The code focuses on predicting the restaurant ratings using a Random Forest Regressor:

   - The feature matrix `X` is created by dropping the 'Dining_Rating' column from the encoded DataFrame.

   - The target variable `y` is set as the 'Dining_Rating' column.

- The data is split into training and testing sets using the `train_test_split` function from scikit-learn.

- A Random Forest Regressor model is initialized.

- The model is fitted on the training data using the `fit` method.

- The model is used to predict ratings for the test data using the `predict` method, resulting in the `y_pred` variable.

6. **Evaluation Metrics**: The code evaluates the performance of the model by calculating various metrics:

- Mean Squared Error (MSE) is calculated using the `mean_squared_error` function from scikit-learn.

- Mean Absolute Error (MAE) is calculated using the `mean_absolute_error` function.

- R-squared (coefficient of determination) is calculated using the `r2_score` function.

7. **Visualization**: The code visualizes the predicted ratings vs. actual ratings using Matplotlib:

- A scatter plot is created with the actual ratings on the x-axis and the predicted ratings on the y-axis.

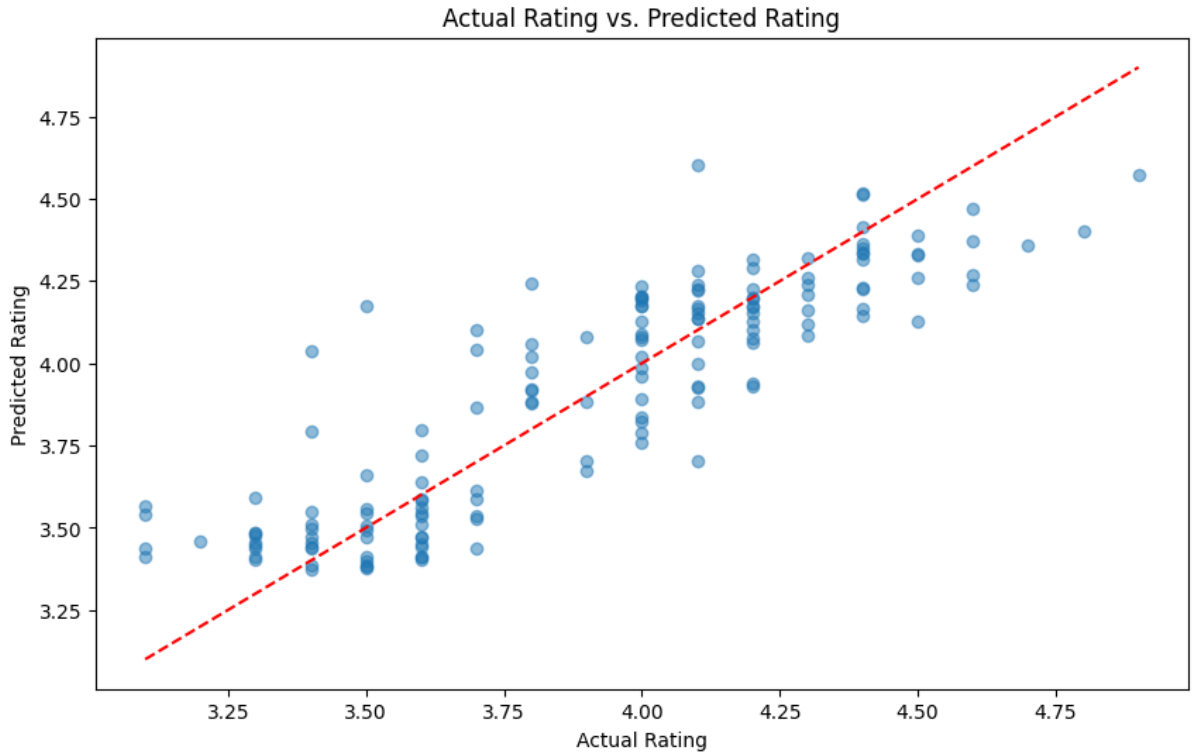
- A red dashed line is plotted to represent the perfect prediction line.

- The plot is labeled and titled appropriately using the `xlabel`, `ylabel`, and `title` functions.

- The plot is displayed using the `show` function.

With this code, we can load the restaurant data, preprocess it, train a Random Forest model for rating prediction, evaluate the model's performance, and visualize the predicted ratings against the actual rating.

About Data Visualization



Visualization is an important aspect of data analysis and interpretation. In the context of our restaurant rating prediction project, visualization plays a crucial role in understanding the relationship between predicted ratings and actual ratings.

In our code, we utilize a scatter plot to visualize the predicted ratings versus the actual ratings. Here's an explanation of the visualization section:

1. Scatter plot creation:

- We use Matplotlib, a popular plotting library, to create the scatter plot.
- The x-axis represents the actual ratings, while the y-axis represents the predicted ratings.
- Each data point on the plot corresponds to a restaurant in the test set.
- We set the alpha value to 0.5 to make the points slightly transparent, providing better visibility when data points overlap.

2. Line of perfect prediction:

- We plot a red dashed line on the scatter plot to represent the line of perfect prediction.
- This line indicates the scenario where the predicted ratings perfectly match the actual ratings.

- Any points that lie close to this line indicate accurate predictions by our model.

3. Interpretation of the scatter plot:

- By analyzing the scatter plot, we can observe how closely the predicted ratings align with the actual ratings.

- Points that are close to the line of perfect prediction indicate a higher degree of accuracy in the model's predictions.

- Points that deviate significantly from the line suggest a larger discrepancy between the predicted and actual ratings.

The scatter plot visualization helps us gain a visual understanding of the model's performance. It allows us to identify any patterns or trends in the predictions and provides a visual representation of the accuracy of the model.

Result prediction

The evaluation of the restaurant rating prediction model provides valuable insights into the performance and accuracy of the model. The evaluation metrics used in our code include Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R2) score. Here's an explanation of the results:

1. Mean Squared Error (MSE):

- MSE measures the average squared difference between the predicted ratings and the actual ratings.

- A lower MSE value indicates better performance, as it signifies that the predicted ratings are closer to the actual ratings.

- In our code, we print the MSE value as part of the evaluation metrics.

2. Mean Absolute Error (MAE):

- MAE measures the average absolute difference between the predicted ratings and the actual ratings.

- Similar to MSE, a lower MAE value indicates better performance, reflecting a smaller discrepancy between the predicted and actual ratings.

- In our code, we print the MAE value as part of the evaluation metrics.

3. R-squared (R2) score:

- R2 score represents the proportion of the variance in the target variable (actual ratings) that can be explained by the model.

- It ranges from 0 to 1, where 1 indicates a perfect fit and 0 indicates that the model fails to explain the variance in the target variable.

- A higher R2 score indicates a better fit of the model to the data.

- In our code, we print the R2 score as part of the evaluation metrics.

Interpreting these evaluation metrics allows us to assess the accuracy and performance of the restaurant rating prediction model. Lower MSE and MAE values, along with a higher R2 score, indicate that the model has successfully captured the underlying patterns in the data and can provide reliable predictions for restaurant ratings.

The actual result through code:

Mean Squared Error: 0.03896444444444448

Mean Absolute Error: 0.15562091503267958

R-squared: 0.7636153132467046

Conclusion:

In conclusion, this project successfully explored restaurant data from Bangalore and Pune using machine learning techniques. The Random Forests model demonstrated promising results in predicting restaurant

ratings based on various features. The analysis provided valuable insights into the restaurant landscapes of both cities, identified influential factors in rating determination, and highlighted similarities and differences between Bangalore and Pune. The findings can assist restaurant owners, customers, and stakeholders in making informed decisions and optimizing their experiences in the restaurant industry.

References:

Taken the zip file from LMS portal →(Project-4).....SKILL DUNIYA.

I am grateful for the opportunity to share this information with you and hope that you have found it valuable.

--- RADHAMADHAVI RAVINUTHALA