Task : Price Prediction for Airbnb Listings

Goal:

Predict the price of an Airbnb listing based on the features provided in the dataset .

Dataset Info:

- **id:** Unique identifier for each listing
- **name:** Name of the Airbnb listing
- **rating:** Average rating of the listing
- **reviews:** Number of reviews received
- **host_name:** Name of the host
- **host_id:** Unique identifier for the host
- **address:** Location of the listing (city, region, country)
- **features:** Summary of features (number of guests, bedrooms, beds, bathrooms)
- **amenities:** List of amenities provided
- **price:** Price per night in the local currency
- **country:** Country where the listing is located
- **bathrooms:** Number of bathrooms
- **beds:** Number of beds
- **guests:** Number of guests the listing can accommodate
- **toilets:** Number of toilets

- **`bedrooms:`** `Number of bedrooms`
- **`studios:`** `Number of studio units`
- **`checkin:`** `Check-in time`
- **`checkout:`** `Check-out time`

Step 1: Exploratory Data Analysis (EDA)

1. **Understand the Dataset**:
    - Load and inspect the dataset to understand its structure.
    - Identify the data types and check for missing values.
    - Examine the first few records in the dataset.
2. **Check Data Distribution**:
    - Analyze the distribution of the target variable (`price`) and other numerical columns like `rating`, `reviews`, `bedrooms`, etc.
    - Visualize the distribution of the target variable (`price`) using appropriate plots (e.g., histograms, boxplots).
3. **Correlation Analysis**:
    - Investigate correlations between numerical features (e.g., `rating`, `reviews`, `bathrooms`) to identify potential relationships.
    - Use a correlation matrix to visually assess feature interdependencies.
4. **Analyze Categorical Features**:
    - Investigate and visualize the distribution of categorical features like `country`, `host_name`, `bedrooms`, etc.
    - Explore how categorical features relate to the target variable `price`.

Step 2: Data Cleaning

1. **Handle Missing Values**:
    - Identify any missing values in the dataset.
    - Decide how to handle missing values based on the feature's importance (e.g., impute with mean, median, or mode, or drop if necessary).
2. **Outlier Detection and Removal**:
    - Identify potential outliers in the `price` column (e.g., extreme values) using visualization techniques.
    - Remove or adjust outliers where appropriate.
3. **Convert Categorical Features**:
    - Convert categorical variables (e.g., `host_name`, `country`) into numerical representations using methods like one-hot encoding.
4. **Feature Scaling**:
    - Standardize or normalize numerical features where necessary, especially for ANN-based models.

Step 3: Model Creation:

Objective: Create a custom Artificial Neural Network (ANN) model using Keras for the given task.

Evaluation: We expect a high level of Exploratory Data Analysis (EDA) and a good accuracy score on sentiment classification.

Accuracy: Along with the achieved accuracy, the approach you used to attain it will also be considered in the evaluation.

Step 4: Accuracy Improvement

1. **Model Evaluation**:
    - **Metrics**: Evaluate the model using common classification metrics: **accuracy**, **precision**, **recall**, and **F1-score**. These metrics will help assess the overall performance and the balance between true positives and false positives.
    - **Confusion Matrix**: Plot a **Confusion Matrix** to understand the distribution of true positives, true negatives, false positives, and false negatives.
    - **AUC-ROC Curve**: Plot the **AUC-ROC curve** to visualize the tradeoff between the true positive rate and false positive rate. This will help you assess the model's ability to distinguish between duplicate and non-duplicate question pairs.
2. **Hyperparameter Tuning**:
    - Experiment with different hyperparameters (e.g., number of layers, neurons per layer, learning rate, batch size) to improve model performance
3. **Cross-Validation**:
    - Use cross-validation to assess the model's performance across different data splits and ensure its generalizability.
4. **Ensemble Methods**:
    - Consider combining the ANN model with other machine learning models (e.g., Random Forest, Gradient Boosting) to improve the overall accuracy.

Step 5: Final Submission

1. **Submit the Code**:
    - Provide the complete code for the entire workflow, including EDA, data cleaning, feature engineering, model creation, and evaluation.
2. **Provide an Explanation**:
    - Include a detailed explanation of the steps you performed, the reasoning behind each decision, and how the final model was built.