Report: Multimodal Regression Model

Introduction

This report outlines the implementation for predicting a target variable using multimodal data, including tabular, textual, and image data. The task is divided into two steps:

1. Optimization using tabular data only
2. Multimodal learning with text and image integration

Step 1: Optimization with Tabular Data

Objective

The goal for this step was to develop a regression model that predicts the target variable using only tabular data. Model performance is evaluated using **Mean Absolute Error (MAE)** and the **R2 score**.

Implementation

1. **Tabular Data Preprocessing**:
    - The `TabularPreprocessor` class performs essential preprocessing operations:
        - Missing Value Imputation: median for numerical data, constant for categorical.
        - Categorical Encoding: Applies `LabelEncoder` for categorical features.
        - Feature Scaling: Standardizes numerical features using `StandardScaler`.
        - Target Transformation: Log transformation or normalization is optionally applied to the target.
2. **Model and Pipeline**:
    - A `GradientBoostingRegressor` model was chosen with standard parameters (e.g., `n_estimators=100`, `max_depth=5`) for optimizing target prediction.

- The `Step1Pipeline` class integrates:
  - Tabular data preprocessing.
  - Dataset splitting for train/test sets.
  - Model training and evaluation.

Results

Following training, the tabular model achieved the following performance metrics:

- **Train MAE**: 6130.85
- **Test MAE**: 6525.80
- **Train R2 Score**: 0.8038
- **Test R2 Score**: 0.6971

Step 2: Multimodal Learning with Text and Images

Objective

This step extended the model to incorporate textual and visual data alongside tabular data. The `description` column and corresponding images were used to enhance target prediction through multimodal embeddings.

Implementation

1. **Multimodal Preprocessing**:
   - The `MultimodalPreprocessor` class combines `TabularPreprocessor` functionality with text and image preprocessing:
     - **Text Data**: Tokenization and embedding generation for textual data using `CLIP`.
     - **Image Data**: Each image is resized to 224×224, converted to a tensor, and normalized.
2. **Multimodal Dataset and Model**:
   - `MultimodalDataset`: This dataset class prepares each sample with tabular, text, and image data for model input.
   - `MultimodalModel`: The model structure includes:
     - A fully connected network to process tabular data.

    - CLIP to generate embeddings for both text and image data, with projection layers to align the embeddings in a common latent space.
    - A fusion network that concatenates tabular, textual, and image embeddings to predict the target.

3. **Pipeline and Training**:
    - The Step2Pipeline defines the multimodal training process:
        - Tabular, text, and image data are combined and split into train/test sets.
        - The CLIP model (using the ViT-B-32 architecture) generates embeddings for text and image data.
        - Training is conducted using MSE loss and the Adam optimizer, with the best model saved during validation.

Results

During multimodal training, the following performance metrics were recorded:

- **Training Loss**: 0.4017
- **Validation Loss**: 0.1155
- **Validation MAE**: 9411.12
- **Validation RMSE**: 44776.07

Conclusion

This implementation effectively addresses a multimodal regression challenge:

- **Step 1** optimized the model with tabular data alone, achieving reasonable MAE and R2 scores.
- **Step 2** incorporated text and image data, utilizing CLIP for multimodal embeddings and showing improved feature integration, though the MAE and RMSE scores suggest room for fine-tuning and additional adjustments.

The project is completed with a Dockerfile, enabling the end-to-end pipeline to run in a reproducible environment.