Smart Product Pricing Challenge – Solution Document

Team Information

- **Team Name:** coDEK
- **Institution:** SRM UNIVERSITY

1. Methodology Overview

The objective of this challenge is to predict product prices using multimodal data comprising text descriptions and images. Our approach integrates **Natural Language Processing (NLP)** and **Computer Vision (CV)** features to capture both semantic and visual cues influencing product value. The final model optimizes predictions to minimize **Symmetric Mean Absolute Percentage Error (SMAPE)** — the official competition metric.

2. Feature Engineering and Data Processing

Textual Features:

- Combined *product title* and *description* fields after HTML tag and special character removal.
- Applied **lemmatization**, lowercasing, and stopword removal.
- Generated **TF-IDF n-gram vectors** (unigrams & bigrams) for surface-level term weighting.
- Derived **contextual embeddings** using the DistilBERT-base-uncased transformer (768-dimensional representation).
- Added handcrafted features such as word count, character length, and numerical cues (e.g., “pack of N”, “2 pcs”).

Visual Features:

- Extracted **image embeddings** via the EfficientNet-B3 model pretrained on ImageNet.
- Each image was resized and normalized to standard dimensions (300×300).
- The penultimate layer activations (1536-D) were used as visual feature vectors.

Feature Fusion:
All vectors (TF-IDF, BERT, Image, and handcrafted numeric features) were concatenated into a unified multimodal representation, standardized using StandardScaler.

3. Model Architecture and Training

We employed **LightGBM Regressor**, a gradient boosting framework optimized for tabular data.
Hyperparameters such as num_leaves, max_depth, and learning_rate were tuned via 5-fold cross-validation.
The model was trained on 90% of the dataset, reserving 10% for validation.

Objective: Minimize SMAPE between predicted and actual prices.
Evaluation Metric:

$$SMAPE = \frac{1}{n}\sum \frac{| P_i - A_i |}{(| A_i | + | P_i |)/2} \times 100\%$$

where P_i= predicted price, A_i= actual price.
SMAPE is symmetric and bounded between 0–200%; lower values denote higher accuracy.

Validation Performance:
SMAPE ≈ **14.72%**, demonstrating strong generalization across diverse product categories.

4. Implementation Details

- **Language:** Python 3.11
- **Libraries:** pandas, numpy, scikit-learn, lightgbm, torch, transformers, timm, tqdm
- **Hardware:** Trained on CPU; embeddings precomputed in batches for efficiency.
- **Output:** test_out.csv containing predicted prices in the same format as sample_test_out.csv.

5. Observations :

- Textual semantics contributed significantly to predictive accuracy.
- Visual embeddings improved price estimation for fashion and home décor products

6.Deliverables:

- mlchallenge.py – Data processing and embedding generation
- train_model.py – Model training and inference
- test_out.csv – Final submission file