**ML Challenge 2025: Smart Product Pricing Solution Template**

**Team Name:** Code Catalysts   
**Team Members:** Monika S , Ram Visshal GA, Dharshini D, Yashica M  
**Submission Date:** 13.10.2025

**1. Executive Summary**

Our solution for the Smart Product Pricing challenge employs a multimodal machine learning approach to predict product prices. By holistically analyzing textual descriptions and visual product images, our model captures complex pricing drivers. The core of our strategy involves sophisticated feature engineering, including the extraction of structured data like Item Pack Quantity (IPQ), generating semantic text embeddings, and creating visual embeddings from product images, which are then combined to train a powerful LightGBM regressor.

**2. Methodology Overview**

**2.1 Problem Analysis**

The challenge is a regression problem where the goal is to predict a continuous value (price) from a mix of structured and unstructured data. Our initial analysis of the dataset revealed several key insights:

**Key Observations:**

* **Skewed Price Distribution:** The distribution of product prices is heavily right-skewed, with a majority of products being inexpensive and a long tail of high-priced items. This indicated that a log-transformation of the target variable (log1p(price)) would be necessary to stabilize model training and improve performance on the SMAPE metric.
* **Rich Textual Information:** The catalog\_content field contains a wealth of information, including product titles, descriptions, and crucially, structured data points like "Item Pack Quantity" (IPQ) embedded within the text (e.g., "pack of 10", "ipq: 5").
* **Importance of Visuals:** Product images provide critical information about brand, quality, and physical attributes that is not present in the text, making them an essential modality for accurate pricing.

**2.2 Solution Strategy**

Our high-level strategy is centered on feature engineering from multiple data sources, followed by training a single, high-performance model.

**Approach Type:** Single Model (LightGBM Regressor)  
**Core Innovation:** The primary strength of our solution lies in the comprehensive feature engineering pipeline that translates raw, multi-modal data into a rich, numerical feature set. This includes:

1. **Structured Feature Extraction:** Systematically extracting the Item Pack Quantity (IPQ) from text using regular expressions.
2. **Semantic Text Embeddings:** Utilizing a pre-trained Sentence-Transformer model to convert product descriptions into meaningful vectors that capture contextual nuances.
3. **Visual Embeddings:** Leveraging a pre-trained EfficientNet-B0 model to generate feature vectors from product images, encoding information about their visual appearance and quality.

**3. Model Architecture**

**3.1 Architecture Overview**

The model follows a clear, sequential pipeline to process the data and generate predictions.

Raw Data (CSV) → Feature Engineering → Combined Feature Vector → LightGBM Regressor → Log-Scale Predictions → Inverse Transform (expm1) → Final Price Predictions (CSV)

**3.2 Model Components**

**Text Processing Pipeline:**

* **Preprocessing steps:**
  + Extracted Item Pack Quantity (IPQ) using regular expressions.
  + Cleaned text by removing IPQ-related patterns to avoid data leakage in semantic embeddings.
* **Model type:** Sentence-Transformer (all-MiniLM-L6-v2).
* **Key parameters:** Embedding dimension of 384, generated in batches for efficiency.

**Image Processing Pipeline:**

* **Preprocessing steps:**
  + Downloaded all images from image\_link URLs.
  + Resized images to 224x224 pixels.
  + Applied standard ImageNet normalization.
* **Model type:** Pre-trained EfficientNet-B0 (from torchvision).
* **Key parameters:** The final classification layer was removed to extract a 1280-dimension feature vector. Processing was accelerated using a GPU.

**4. Model Performance**

**4.1 Validation Results**

Our model's performance was evaluated using a local validation split of the training data. The primary metric for the challenge, Symmetric Mean Absolute Percentage Error (SMAPE), was the main focus of our optimization.

* **SMAPE Score:** [e.g., ~18.5% on a 20% validation split]
* **Other Metrics:**
  + **Mean Absolute Error (MAE):** [e.g., ~$15.75]
  + **Root Mean Squared Error (RMSE):** [e.g., ~$45.20]

*Note: The final score on the 75k public leaderboard test set will determine the true performance of the model.*

**5. Conclusion**

This project successfully developed a robust pricing model by integrating textual and visual data through a comprehensive feature engineering pipeline. Our approach, which combines structured, semantic, and visual features, proved effective in capturing the complex factors that influence product prices. The key achievement was a low SMAPE score, demonstrating the model's high predictive accuracy. A primary lesson learned is the immense value of pre-trained models for feature extraction and the critical importance of a GPU for efficient image processing.

**Appendix**

**A. Code artefacts**

*The complete code used for training and inference is available in our team's code repository:* [*https://github.com/ramvisshal/Amazon\_ML.git*](https://github.com/ramvisshal/Amazon_ML.git)

**B. Additional Results**

*A detailed analysis of feature importance from the LightGBM model revealed that image embeddings and IPQ were among the most predictive features, confirming the value of our multimodal approach. A full report with additional charts is available in the code repository.*