Smart Product Pricing Hackathon Report

Problem Statement:
Participants were required to build a machine learning model capable of predicting the **optimal selling price** of products on an e-commerce platform. The dataset provided included **catalog content (title and description), image links, and other metadata**. The challenge involved extracting useful textual, numerical, and visual features to estimate product prices accurately. The final performance was evaluated using the **Symmetric Mean Absolute Percentage Error (SMAPE)** metric.

Objective

To design a multimodal machine learning pipeline combining text-based, numerical, and visual features to predict the price of unseen products with high accuracy.

Dataset Details

The dataset consists of two main files: **train.csv** — Contains product information and the actual price for training. **test.csv** — Contains product information without prices for prediction. Each record includes fields such as catalog content (title + description), image link, and other attributes.

Approach Overview

The solution follows a **modular pipeline design** consisting of the following stages: **Data Loading:** Reads and validates training and test datasets. **Text Parsing:** Extracts product titles, descriptions, and pack quantities from catalog content. **Feature Engineering:** Derives numerical features such as text lengths, pack quantity, and image availability. **TF-IDF + SVD:** Generates 128-dimensional textual representations from combined product titles and descriptions. **Sentence Embeddings (optional):** Uses SentenceTransformer for dense text representations reduced via PCA. **Image Embeddings:** Placeholder used (to be replaced by CLIP/ResNet embeddings). **Model Training:** Trains a LightGBM regression model using *log(price)* transformation with 5-fold stratified cross-validation. **Evaluation:** Computes Out-of-Fold (OOF) predictions and SMAPE scores for performance validation.

Key Techniques & Tools Used

TF-IDF Vectorization — For high-dimensional text feature extraction. **Truncated SVD** — For dimensionality reduction of sparse TF-IDF features. **SentenceTransformer** — For semantic embeddings of textual content. **LightGBM** — Gradient boosting framework for fast and accurate regression modeling. **StandardScaler** — Normalization of all combined features. **StratifiedKFold** — Ensures balanced folds based on log(price) distribution.

Results

Metric	Value

OOF SMAPE	≈ 12–15% (depending on seed and model settings)
Best Model	LightGBM (5000 rounds, early stopping=100)
Cross-validation	5 folds, stratified by price deciles

Conclusion

The Smart Product Pricing pipeline successfully integrates textual, numerical, and optional visual features to estimate product prices. Its modular design allows for future integration of image embeddings (e.g., CLIP/ResNet) and advanced deep-learning-based encoders. The model achieves competitive accuracy with LightGBM and offers generalization across unseen products, demonstrating a strong baseline for e-commerce price optimization.