Smart Product Pricing Challenge

Unstop Amazon ML Challenge 2025

Methodology Documentation

Sumit Kumar (Leader) Ajitabh Ranjan
Harsh Raj Disha Tribedy

October 24, 2025 — Submission: `final_safe_submission.csv`

We developed a multimodal ensemble solution combining CLIP text embeddings, URL-based image features, and multiple ML models. Key innovations include memory-optimized processing, hyperparameter tuning, log-space ensemble blending, and distribution-aware post-processing. Final submission meets all formatting and validation requirements.

Problem Overview

Hackathon: Unstop Amazon ML Challenge 2025
Problem: Smart Product Pricing Challenge
Goal: Predict product prices using catalog content and image URLs
Metric: SMAPE (Symmetric Mean Absolute Percentage Error)
Data: 75K train + 75K test samples

Technical Approach

Feature Engineering

- **Text Features:** CLIP embeddings (384D → 32D PCA), TF-IDF (15D)
- **Image Features:** URL-based (protocol, domain, file type)
- **Numeric Features:** item_pack_qty, catalog_len + transformations
- **Feature Selection:** 61 features → 54 after selection

Model Architecture

- **XGBoost:** 800 trees, early stopping, tuned parameters
- **LightGBM:** 800 trees, early stopping, tuned parameters
- **Ridge:** L2 regularization, grid search
- **ElasticNet:** L1/L2 regularization, grid search

Optimization

- **Hyperparameter Tuning:** Manual grid search for learning rates, tree depth, and regularization
- **Memory Optimization:** Batch processing and dtype optimization
- **Feature Selection:** Variance thresholding + SelectKBest

Pipeline Overview

Figure 1: CLIP embeddings → PCA → Ensemble → Post-processing

Ensemble Strategy

Weighted Blending

$$w_i = \frac{1/\mathrm{RMSE}_i}{\sum_j 1/\mathrm{RMSE}_j}$$

- Log-space predictions for stability
- Inverse RMSE weighting
- Final exponentiation to original scale

Validation Performance

Model	RMSE	Weight
XGBoost	36.75	35%
LightGBM	36.93	35%
Ridge	38.86	15%
ElasticNet	38.86	15%

Post-processing

- **Clipping:** 0.1th to 99.5th percentiles of training prices
- **Alignment:** Mean correction to match training distribution
- **Rounding:** All prices to 2 decimal places
- **Validation:** No negative/zero/missing values

Final Results

Final SMAPE Score: 52.1

(As per Competition Management — Official Evaluation)

- **Training Stats:** Mean=$23.65, Median=$14.00
- **Final Submission:** Mean=$23.65, Median=$17.20
- **Price Range:** $2.48 - $72.05
- **Samples:** 75,000 complete predictions
- **Leaderboard: Final SMAPE = 52.1** (Official Result)

Submission Files

- **Main:** `final_safe_submission.csv`
- **Variants:** +1%, +2% bias versions for LB testing
- **Format:** Exact `sample_id,price` columns
- **Validation:** All sanity checks passed

Future Improvements

- **Advanced Multimodal:** Incorporate BLIP, CLIP ViT-L for better embeddings
- **Feature Analysis:** Expand feature selection using SHAP importance
- **Robust Ensembling:** Cross-validation folds for better ensemble weights
- **Architecture:** Transformer-based multimodal encoders

Team zyro — Code and implementation available in submission materials
Methodology: CLIP Multimodal + Ensemble Models + Distribution Alignment