💰

Smart Pricing Methodology

Project Overview

This project addresses the **Smart Pricing** task: predicting product prices based on multimodal catalog data that combines textual descriptions, numerical features, and image information.

Pipeline Architecture

The methodology follows a **three-stage pipeline**:

1. Text-only baseline model
2. Multimodal enhancement
3. Stacked ensemble for final inference

Stage 1: Baseline Model

TF-IDF + Ridge Regression

Data Source: `catalog_content` textual field

Preprocessing:

- Missing values filled
- Text vectorization using TF-IDF (max_features = 10,000)
- Captured unigrams and bigrams

Model Training:

- Algorithm: Ridge Regression (α = 1.0)
- Validation: 5-fold cross-validation
- Evaluation metric: Mean Absolute Error (MAE)

Results:

- CV MAE: ≈ **0.1852**
- SMAPE: **15.9%**
- Output: `test_baseline_log.npy`

Stage 2: Multimodal Model

Text + Numeric + Image Features

Feature Engineering

Text Features:

- TF-IDF vectorization
- Dimensionality reduction via TruncatedSVD (128 dimensions)

Image Features:

- Precomputed CNN embeddings
- PCA reduction (64 components)

Numeric Features:

- Pack quantity (ipq)
- Text length
- Word counts
- Currency presence
- Image availability
- Standardized using StandardScaler

Model Training

Algorithm: LightGBM

Validation: 5-fold StratifiedKFold

Results:

- SMAPE: **13.47%** (≈ 13.5%)
- Output: `test_multimodal_log.npy`

Stage 3: Stacked Ensemble

Final Inference

Approach: Stacking predictions from baseline and multimodal models

Process:

1. Load saved predictions (`test_baseline_log.npy` and `test_multimodal_log.npy`)
2. Build 2-feature meta-dataset
3. Train Ridge Regression meta-model to optimally combine both sources

Results:

- SMAPE: **12.8%**
- Output: `test_out_final.csv`

Performance Evaluation

Metrics Summary

All models evaluated using **SMAPE** (Symmetric Mean Absolute Percentage Error):

Model	Features	Algorithm	CV Metric	SMAPE (%)
Baseline	TF-IDF (Text only)	Ridge Regression	MAE = 0.185	**15.9**
Multimodal	TF-IDF + SVD + Image + Numeric	LightGBM	RMSE / SMAPE = 13.47%	**13.5**
Final Stack	Baseline + Multimodal OOF	Ridge (Meta-model)	SMAPE = 12.8%	**12.8**

Performance Improvement

- **Baseline → Multimodal:** 2.4% improvement

- **Multimodal → Stacked:** 0.7% improvement
- **Overall improvement:** 3.1% (15.9% → 12.8%)

Key Findings

> The multimodal and stacked approach substantially outperformed the baseline, highlighting the value of integrating text, image, and numerical signals for price prediction.

Success Factors:

- Multimodal feature integration
- Ensemble learning through stacking
- Proper feature engineering and dimensionality reduction
- Appropriate model selection for each stage