ML Challenge 2025: Smart Product Pricing Solution

Team Name: Team EPOCH Engineers

Team Members: Shiv Swagat Subudhi, Trilokesh Das, Sahil Mohanty

Submission Date: 13 October 2025

1. Executive Summary

Our solution implements a hybrid ensemble model that combines semantic text embeddings from SentenceTransformer (all-mpnet-base-v2),

TF-IDF features, and engineered numerical attributes to understand complex relationships between product descriptions and prices.

The dual LightGBM architecture, integrated through Ridge regression stacking, effectively captures both linguistic context and numerical patterns,

allowing for more accurate and data-driven product price predictions.

2. Methodology Overview

2.1 Problem Analysis

The challenge focused on predicting optimal product prices based on product descriptions, packaging, and textual data.

Through EDA, we observed that textual details such as product name, unit quantity, and pack size had a strong correlation with price levels.
Handling outliers, skewed price distributions, and sparse text patterns was critical for achieving stable predictions.

Key Observations:

- Strong correlation between packaging terms (e.g., 'Pack of', 'ml', 'kg') and price.
- Text skew handled using log1ptransformation and 99th percentile clipping.
- Image embeddings optional but excluded for faster and consistent performance.

2.2 Solution Strategy

We designed a multimodal hybrid ensemble pipeline leveraging semantic, lexical, and numeric signals.

MPNet embeddings capture contextual meaning, TF-IDF models extract keyword importance, and engineered numeric features encode quantitative product attributes.

Dual LightGBM regressors model both dense and sparse representations, and a Ridge regression layer fuses their predictions for optimal generalization.

Approach Type: Hybrid Ensemble (Text + Numeric)
Core Innovation: Multilevel fusion of contextual embeddings and TF-IDF representations through Ridge stacking.

3. Model Architecture

3.1 Architecture Overview

The architecture integrates multiple parallel components:

1. Text Embedding Stream: SentenceTransformer (all-mpnet-base-v2) with TruncatedSVD dimensionality reduction.

2. Sparse Text Stream: TF-IDF with LightGBM regression.

3. Numeric Feature Stream: Engineered length, count, and digit features added to dense embeddings.

Predictions from these models are aggregated using Ridge regression to enhance robustness.

3.2 Model Components

Text Processing Pipeline:

- Preprocessing: Lowercasing, stopword removal, NaN fill, unit extraction (pack/ml/kg).
- Model Type: SentenceTransformer (all-mpnet-base-v2) + LightGBM.
- Key Parameters: embedding_dim=768, svd_dim=256, learning_rate=0.035, num_leaves=256.

Image Processing Pipeline:

- Preprocessing: Optional CLIP-based embeddings.
- Model Type: ViT-B/32 (disabled for runtime optimization).
- Key Parameters: embedding_dim=512, normalization=True.

4. Model Performance

Validation Results:

- SMAPE Score: Achieved strong generalization across validation folds.
- Other Metrics: Consistent RMSE across TF-IDF and dense branches with stable Ridge fusion.
- Runtime: Efficient execution on A100 GPU without overfitting indicators.

5. Conclusion

Our hybrid ensemble effectively integrates contextual, lexical, and numeric representations to model complex pricing relationships.

The modular design ensures scalability, adaptability, and interpretability for diverse product datasets. Future work includes leveraging transformer-based multimodal fusion and advanced hyperparameter tuning for further accuracy gains.