Amazon ML Challenge 2025 - Approach by Deep Morph

Team Members: Pranav A, Devanarayanan V S, Devesh, Siddhant Erande
Submitted on: 13th October 2025

Our solution is a SMAPE-optimized weighted ensemble of five diverse models, encapsulating a feature engineering pipeline that extracts and canonicalizes quantitative data directly from unstructured text. This approach combines a deep semantic understanding of product descriptions with explicit, structured attributes to deliver highly accurate and generalizable price predictions.

Methodology

The goal was to predict product prices from catalog content (item name and price) using analysis of text, numeric and categorical features. The approach involved feature engineering to extract structured signals from unstructured text, training **multiple regressors** with diversity (linear, tree-based and neural) and blending them into an **ensemble** for more accurate predictions. A **90/10 train-validation** split is used for hyperparameter tuning and weight optimization via **Nelder-Mead minimization** of Symmetric Mean Absolute Percentage Error (SMAPE). Models are retrained on full data post-validation for better results. No external data or pre-trained price-specific embeddings are used to avoid leakage.

Feature Engineering

Text Cleaning & Extraction: Catalog content is cleaned by removing redundant sections (like Value/Unit patterns via regex). Key entities like brand, category (eg: soup, sauce using rule-based regex), sub-category, flavor profile, and nutritional info (eg: protein grams, calories) are extracted.
Quantity Normalization: Units were canonicalized (eg: lb → g, fl oz → ml) with scaling factors. Pack size is detected via patterns (eg: "Pack of 12"). Derived features include log-transformed values (eg: log_total_g), inverses (eg: inv_pack_size) and flags (eg: is_organic, is_high_protein).
Text Features: TF-IDF (word/char n-grams) + TruncatedSVD (128 components) for tree-based models - raw text tokenized for Transformer.
Numeric/Categorical: Scalers (StandardScaler) normalize numerics - LabelEncoders handle categories (eg: unit, brand). Additional metadata like text length, premium/bulk keyword counts, and quality flags (eg: gluten-free) enhance signals.
Handling Edge Cases: Defaults for missing values (eg: pack_size=1, unit='<unk>'); clipping (eg: prices ≥0.99) prevents invalid outputs.

Model Architectures & Algorithms

Ridge Regression: Linear baseline with TF-IDF (word 1-2grams, char 3-4grams, max_features=100k) + sparse numeric features. Alpha=3.0, trained on log(price) for scale invariance.
XGBoost, CatBoost & LightGBM: Regressors on TF-IDF+SVD + engineered features. Params: depth=6, lr=0.05, MAE loss, L1 & L2 reg, GPU-accelerated with early stopping. Log(price) being target, categorical handling was done via encoders.
Transformer Model: Used pretrained all-mpnet-base-v2 Sentence Transformer as base + embeddings for unit/category/flavor + MLP head (768→384→192→1). Mean-pooled hidden states concatenated with scaled numerics. Trained with mixed MAE + SMAPE loss (alpha=0.5), AMP, CosineWarmup scheduler, DDP for multi-GPU. Batch=32, epochs=10, patience=3. First training only the MLP head, then unfreezing and fine-tuning the transformer encoder with a lower learning rate.

Ensemble Strategy

Predictions from all models are Combined using optimized weights (Ridge Regression: 0.0479,XGBoost: -0.1483, CatBoost: 0.3207, LightGBM: 0.1004,Transformer: 0.6793,) minimizing validation SMAPE. Final model retrained on full train.
Final SMAPE Score: 43.56318559 (by eval in unstop)

Note on Images: Although image links were provided in the dataset, they were not incorporated into the final model. An experimental multimodal approach combining text features with image embeddings resulted in a significantly higher validation SMAPE of over 70%, hence images were not used.
Link to access code: https://drive.google.com/drive/folders/1KLu5k95qtXQaCg2veGns8yNu7M0HLlTS?usp=sharing