Smart Product Pricing Challenge - Solution Documentation

ML Challenge 2025 — Team: MannusAi

Problem: Predict product prices from catalog_content (text) and product images using machine learning. Evaluation metric: SMAPE.

1. Methodology Used

Our approach implements a multi-modal ensemble combining textual, visual, and engineered tabular features. The pipeline consists of: (1) data preprocessing and cleaning, (2) multi-modal feature extraction, (3) ensemble model training using gradient-boosted trees, and (4) prediction aggregation. We apply log-transformation to prices for training stability. All features are derived exclusively from the provided dataset without external price lookup.

2. Model Architecture/Algorithms Selected

- **Text Processing:** DistilBERT for semantic embeddings (768-dim) + TF-IDF vectorization (1000 features) from cleaned catalog_content
- **Visual Processing:** ResNet50 feature extractor producing 2048-dimensional image embeddings from downloaded product images
- **Ensemble Method:** Weighted combination of LightGBM and XGBoost regressors (50:50 ratio) trained on concatenated multi-modal features
- **Training:** Models trained on log(price+1) with MAE objective, early stopping on validation SMAPE, final predictions via expm1 transformation

3. Feature Engineering Techniques Applied

- **Text Features:** Regex extraction of Item Pack Quantity (IPQ), product descriptions, bullet points; text cleaning and normalization; BERT embeddings and TF-IDF vectors
- **Numeric Features:** Pack size, ounce/fluid ounce quantities, total volume calculations, product counts; binary flags for organic/gluten-free products
- **Image Features:** ResNet50 embeddings from preprocessed images (resize to 256×256, center crop to 224×224, normalization); random feature vectors for missing images
- **Data Processing:** Log transformation of target variable, StandardScaler normalization for numeric features, missing value imputation with median/mode

4. Additional Relevant Information

Validation Strategy: 85:15 train-validation split with stratification; monitored both MAE and SMAPE during training; early stopping based on validation performance.

Missing Data Handling: Graceful degradation for missing images using fixed random vectors; median imputation for numeric features.

Compliance: All model development used only the provided training dataset. No external price databases, web scraping, or third-party pricing sources were utilized.

Output Format: Final predictions saved as CSV with sample_id and price columns matching the required submission format.