Team Name: EpStrikers

Team Members: Mohit Patel, Raj Desai, Utsav Bhalani

1. Project Overview and Target

The primary objective of this project is to accurately predict product prices, aiming for a SMAPE (Symmetric Mean Absolute Percentage Error) below 45%. The entire methodology centers on a robust Stacked Ensemble architecture. This ensemble uses three diverse base models—LightGBM, XGBoost, and Ridge Regression—and a final Ridge Regression model as the meta-learner to make the ultimate price prediction. Model validation relies on a **7-Fold Stratified Cross-Validation**, which stratifies on price quantiles (or buckets) to ensure each fold has a representative distribution of the target prices.

2. Data Preparation and Feature Engineering

The approach maximizes information extraction from both the product's catalog text (`catalog_content`) and its associated image embeddings, resulting in a multi-modal feature set.

A. Data Preprocessing

Data preparation begins with aggressive **outlier removal** on the `price` column using a $1\%/99\%$ quantile-based Interquartile Range (IQR) method to stabilize training. Critically, the target variable, `price`, is transformed using the natural logarithm ($\log(1+\text{price})$) to mitigate the effects of its highly skewed distribution.

B. Comprehensive Feature Streams

The final combined feature matrix includes 521 features per sample. These features are grouped into six distinct streams:

1. **Comprehensive Numerical Features:** A custom text parser extracts 42 numerical features from the `catalog_content`. These include detailed metrics like (pack quantity ,total volume}), Item-Per-Quantity (IPQ) values, text statistics (e.g., word count, digit ratio), numeric patterns (e.g., maximum and

skew of numbers found), and binary category indicators (e.g., `is_food`, `is_organic`). These raw features are then scaled using QuantileTransformer to achieve a near-normal distribution, improving their performance in the models.

2. **TF-IDF and SVD Features:** A TfidfVectorizer processes the cleaned product text, using 1 to 3 word phrases (n-grams) and retaining $20,000$ features. This sparse matrix is then reduced to 200 components using TruncatedSVD to capture the most important semantic variance while controlling model complexity.

3. **Image and Text Embeddings (PCA):** High-dimensional, pre-trained image (512) and text (384) embeddings are loaded. To make them efficient for the tree-based models, Principal Component Analysis (PCA) is applied to each set, reducing both the image and text embeddings to 100 components each.

4. **Categorical Clusters:** To create new, discrete categorical features from the continuous embeddings, K-Means Clustering is applied. This generates 30 image clusters and 50 text clusters, which are then one-hot encoded and added to the feature matrix.

3. Stacked Ensemble Modelling

3.1 Base Model Training

The three base models—LightGBM, XGBoost, and Ridge—are trained using the full 521-feature set within the 7-Fold Stratified Cross-Validation framework. Key configuration changes in this enhanced V2 approach include increasing the number of estimators (5000 for LGBM, 4000 for XGBoost) and decreasing the learning rate (0.02) for both boosting models to improve convergence and generalization. LightGBM also features increased depth and leaf count, and includes `extra_trees` regularization.

Average Cross-Validation SMAPE:

LightGBM: 48.01%

XGBoost: 48.89%

Ridge: 56.41%

3.2 Stacking Layer and Final Result

The Out-of-Fold (OOF) predictions from the three base models are concatenated to form the training set for the second-level meta-learner.

Meta-Learner: A Ridge Regression model (alpha=1.0) is trained on the OOF predictions.

Final Stacked OOF SMAPE: $47.5357%

The meta-learner's weights reveal its dependency: LightGBM contributes most significantly (0.887), followed by XGBoost (0.288), while the base Ridge model has a minor correctional role (0.132). Although the final result of 47.5357% did not quite meet the initial target of <45%, the stacked ensemble demonstrates strong performance and robust feature utilization across multiple data modalities.