**Smart Product Pricing Challenge – Team GenWin**

**Methodology:**

We approached the pricing prediction task as a regression problem, utilizing both textual and visual product features. The pipeline consists of three main stages: data preprocessing, feature extraction, and model training. The overall goal is to predict optimal prices by learning complex interactions between product attributes.

**Data Preprocessing:**

Textual Features (catalog\_content):

Cleaned text by removing punctuation, lowercasing, and tokenization.

Extracted product-specific attributes such as brand, item pack quantity (IPQ), and specifications.

Applied TF-IDF vectorization and pre-trained embeddings (e.g., BERT embeddings) to capture semantic meaning.

Image Features (image\_link):

Downloaded product images using download\_images from src/utils.py.

Resized and normalized images for consistency.

Extracted features using a pre-trained CNN model (EfficientNet-B0).

Used extracted embeddings as additional input features for the regression model.

**Feature Engineering:**

Combined text embeddings and image embeddings into a single feature vector.

Created additional numeric features such as IPQ, length of description, and keyword counts.

Handled missing values and normalized continuous features to improve model stability.

**Model Architecture / Algorithms:**

Implemented an ensemble regression approach:

Gradient Boosting Regressor (LightGBM) for tabular and engineered numeric/text features.

Feed-forward Neural Network to handle concatenated text and image embeddings.

Final prediction is a weighted average of outputs from both models, which improved robustness.

Ensured predicted prices are positive floats, aligning with challenge constraints.

**Evaluation:**

Model performance is measured using Symmetric Mean Absolute Percentage Error (SMAPE).

Tuned hyperparameters via cross-validation on the training set to minimize SMAPE.

Final predictions were generated for all 75k test samples, formatted exactly as sample\_test\_out.csv.

**Additional Notes:**

No external price data was used; model trained solely on the provided dataset.

Focused on multi-modal learning (text + image) for improved price prediction accuracy.

The pipeline is reproducible, scalable, and compatible with models up to 8B parameters.

**Conclusion:**

The solution leverages a hybrid approach, combining NLP and computer vision techniques with gradient boosting, ensuring holistic analysis of product details to predict optimal pricing.