Smart Product Pricing Challenge – ML Challenge 2025

Team Name: Nxtgen
Team Leader: Harine K. S
Team Members: Dharshini S, Bhavishya Priyadarshini V, Aboorva J

1. Problem Understanding

In ecommerce, optimal pricing drives profitability and customer trust. Prices depend on factors like brand, specifications, product category, and pack quantity. The goal was to predict product prices using only the provided dataset:

Training data: 75,000 samples with `sample_id`, `catalog_content` (title, description, Item Pack Quantity), `image_link`, and `price`
Test data: 75,000 samples without price
Restriction: No external price lookup

2. Methodology

We focused on textual data (`catalog_content`) as it contains most pricerelevant information.

Pipeline:

1. Data Loading & Cleaning: Handled missing values, standardized text, replaced blanks with "unknown".
2. Text Preprocessing: Removed punctuation/special characters, tokenized text, applied TFIDF vectorization (top 10,000 features).
3. Feature Extraction: Extracted quantity indicators (e.g., "pack of 2", "500 ml") and combined with TFIDF features.
4. Model Selection: Compared Linear Regression and Random Forest; chose LightGBM for highdimensional sparse data and regression performance.

3. Model Architecture & Training

Model: LightGBM Regressor
Hyperparameters: `num_leaves=64`, `learning_rate=0.05`, `n_estimators=2000`, `max_depth=1`, `subsample=0.8`, `colsample_bytree=0.8`, early stopping=50 rounds
Validation: 80/20 train/validation split; internal metric RMSE ≈ 0.68, resulting in low SMAPE

4. Evaluation Metric

Symmetric Mean Absolute Percentage Error (SMAPE):

$$
\text{SMAPE} = \frac{1}{n} \sum \frac{|\text{predicted} - \text{actual}|}{(|\text{predicted}| + |\text{actual}|)/2}
$$

Range: 0% (perfect) to 200%
Lower SMAPE indicates better accuracy
All predicted prices were positive floats

5. Results & Insights

Model captured price trends from descriptive words like “premium”, “organic”, “combo”.
Item Pack Quantity (IPQ) significantly improved accuracy.
TFIDF features alone produced competitive predictions without using images.
Validation SMAPE showed stable and reliable predictions across categories.

6. Future Improvements

Incorporate image features using pretrained CNNs (ResNet, EfficientNet).
Multimodal fusion of text and image embeddings.
Categoryspecific models and hyperparameter tuning via Bayesian optimization.

7. Compliance & Output

No external price data or web scraping used
Model trained solely on provided data
Output file: `test_out.csv` with 75,000 predictions (`sample_id`, `price`)

Conclusion

Using TFIDF + LightGBM, we built a simple, interpretable, and efficient price prediction model based on textual data, producing reliable and fair predictions suitable for largescale ecommerce applications.

Prepared by: Team Nxtgen (Harine K. S, Dharshini S, Bhavishya Priyadarshini V, Aboorva J)