**ML Challenge 2025: Smart Product Pricing Solution**

**Team Name: Team House of Codesmiths   
Team Members: Satya Venkata Hemanth Challapalli, Girisha Malni N, Syed Ameen, Boodala Akhil  
Submission Date: October 13th, 2025**

**1. Executive Summary**

We developed a regression-based pricing model that predicts product prices using textual descriptions and engineered numeric attributes. Our solution leverages TF-IDF features combined with numeric features such as log-transformed item pack quantities (IPQ), text length, digit counts, and token counts. The final prediction is made using a LightGBM model with 5-fold cross-validation, achieving robust out-of-fold (OOF) performance.

**Key Innovations:**

* Extraction of numeric quantities (IPQ) from textual descriptions and log transformation for better scale handling.
* Hybrid modeling pipeline combining sparse textual TF-IDF features with dense numeric attributes.
* Optimized LightGBM model with carefully selected hyperparameters.

**2. Methodology Overview**

**2.1 Problem Analysis**

The task is to predict product prices using the provided dataset. Exploratory Data Analysis (EDA) highlighted the following patterns:

* Product descriptions frequently contain pack sizes, volume units, or counts (e.g., "500 ml", "pack of 3").
* Price distribution is highly skewed; log-transformation of prices stabilizes variance.
* Sparse textual representation is crucial due to high vocabulary size.

**2.2 Solution Strategy**

We implemented a hybrid feature engineering and modeling approach combining textual and numeric features:

Approach Type: Single-model LightGBM regression  
Core Innovation: Concatenation of sparse TF-IDF vectors with engineered numeric features including log-transformed IPQ, text length, digit counts, and token counts.

**3. Model Architecture**

**3.1 Architecture Overview**

**catalog\_content (text) → clean\_text → TF-IDF → sparse vector**

**│**

**└─ numeric features (ipq\_log, len\_text, num\_digits, num\_tokens)**

**│**

**└─ Concatenate → LightGBM Regression → log(price) prediction**

**3.2 Model Components**

**Text Processing Pipeline:**

* Preprocessing: lowercasing, newline removal, whitespace normalization
* Vectorization: TF-IDF (unigrams + bigrams)
* Key parameters: max\_features=50,000, min\_df=3, ngram\_range=(1,2)

**Numeric Features Pipeline:**

* Features: log-transformed IPQ, text length, number of digits, number of tokens
* Preprocessing: quantity parsing from text, log transform
* Dense numeric input concatenated with TF-IDF sparse features

**LightGBM Model:**

* Objective: regression
* Metric: RMSE
* Hyperparameters: learning\_rate=0.05, num\_leaves=31, min\_data\_in\_leaf=50, feature\_fraction=0.8, bagging\_fraction=0.8, bagging\_freq=5
* Cross-validation: 5-fold

**4. Model Performance**

**4.1 Validation Results**

* OOF SMAPE Score: 52.0602%
* Predictions were log-transformed and exponentiated for submission.

**5. Conclusion**

The hybrid LightGBM model efficiently captures both textual and numeric patterns in product descriptions to predict prices. Key lessons include the importance of:

* Extracting numeric pack/quantity information from text
* Careful text cleaning and normalization
* Combining sparse and dense features for improved regression performance

The solution is fully reproducible and achieves strong out-of-fold performance using only the provided dataset.

**Appendix**

**A. Code Artefacts:**

* **GitHub repository:** [**https://github.com/chsvhemanth/Amazon\_ML\_Challenge**](https://github.com/chsvhemanth/Amazon_ML_Challenge)