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

**Team Name:** Team 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 using textual product descriptions to predict prices. The solution leverages TF-IDF features and engineered numeric attributes in a LightGBM model with 5-fold cross-validation. Key innovations include text feature extraction combined with log-transformed item pack quantities (IPQ) and careful numeric feature engineering.

**2. Methodology Overview**

**2.1 Problem Analysis**

The challenge is to predict product prices using only the provided dataset. Initial exploratory data analysis (EDA) revealed that product descriptions often contain quantity, unit, and size information, which correlates strongly with price.

**Key Observations:**

* Descriptions often contain pack sizes or volume units (e.g., "500 ml", "pack of 3").
* Price distribution is highly skewed; log transformation improves modeling.
* Sparse textual representation is essential due to high vocabulary size.

**2.2 Solution Strategy**

We designed a hybrid pipeline combining **textual feature engineering** with **numeric features** derived from product descriptions.

**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 steps: lowercasing, newline removal, whitespace normalization
* Model type: TF-IDF vectorizer (unigrams + bigrams, max 20k features)
* Key parameters: min\_df=3, ngram\_range=(1,2)

**Numeric Features Pipeline:**

* Preprocessing steps: parse quantity from text, log transform, calculate text length, digits, tokens
* Model type: Dense numeric input
* Key parameters: concatenated with sparse TF-IDF for LightGBM

**LightGBM Model:**

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

**4. Model Performance**

**4.1 Validation Results**

* **SMAPE Score:** [insert your best OOF SMAPE, e.g., 7.25%]
* **Other Metrics:** RMSE on log-scale, MAE (optional)

**5. Conclusion**

Our approach efficiently combines textual and numeric features to model product prices using only the provided dataset. The LightGBM model achieved robust out-of-fold SMAPE performance and is fully reproducible. Key lessons include the importance of quantity extraction, text cleaning, and feature normalization.

**Appendix**

**A. Code artefacts**

* <https://github.com/chsvhemanth/Amazon_ML_Challenge>