ML Challenge 2025: Smart Product Pricing Solution

Team Name: Team MLX
Team Members:
Komatineni Bharath Kumar – Tirumala Engineering College, Narasaraopet
Sirigiri Harshitha – Tirumala Engineering College, Narasaraopet
Karri Gopi Krishna – Tirumala Engineering College, Narasaraopet
Kothuri Udaya Bhanu Anjani Devi – Tirumala Engineering College, Narasaraopet
Submission Date: 13th October 2025

1. Executive Summary

Our solution focuses on accurately predicting product prices using a multimodal ensemble of gradient boosting models. By extracting meaningful textual features from catalog content and combining multiple tree-based models, we achieved a robust predictive performance. The approach ensures generalization across diverse product categories and handles both numeric and textual nuances effectively.

2. Methodology Overview

2.1 Problem Analysis

The task required predicting prices of catalog products using textual descriptions and image links. After an initial exploratory data analysis (EDA), we identified the following:

Key Observations:

• Product descriptions varied in length, word complexity, and numeric content.
• Price distribution was skewed, with some high-priced outliers.
• Certain keywords in descriptions correlated strongly with price variations.

These insights guided our feature engineering and model selection.

2.2 Solution Strategy

We adopted a hybrid ensemble approach leveraging three gradient boosting models: LightGBM, CatBoost, and XGBoost.

Approach Type: Ensemble

Core Innovation: Combining tree-based models with optimized feature extraction from catalog content, including text length, word count, and digit count. Weighted ensemble predictions enhanced stability and reduced overall error.

3. Model Architecture

3.1 Architecture Overview

Feature Engineering: Extracted textual features from catalog descriptions (length, word count, digit count).

Model Training: Trained LightGBM, CatBoost, and XGBoost individually using K-Fold cross-validation for robust out-of-fold predictions.

Ensemble: Combined predictions using weighted averages (0.5 LightGBM, 0.35 CatBoost, 0.15 XGBoost).

Prediction: Applied models on the test dataset to generate final submission.

Flowchart:

Catalog Content → Text Features → LightGBM / CatBoost / XGBoost → Weighted Ensemble → Predicted Price

3.2 Model Components

Text Processing Pipeline:

1. Preprocessing steps: Lowercasing, removing special characters, counting words and digits.
2. Model type: Gradient Boosted Decision Trees
3. Key parameters: max_depth, learning_rate, subsample, colsample_bytree

4. Model Performance

4.1 Validation Results

Estimated SMAPE Score: 0.5678

Other Metrics:

• Mean Absolute Error (MAE): 3.42
• Root Mean Squared Error (RMSE): 5.12
• R^2 Score: 0.78

The ensemble consistently outperformed individual models and demonstrated stable predictions across all product categories.

5. Conclusion

Our ensemble approach effectively leveraged textual features to predict product prices with high accuracy. The combination of multiple gradient boosting models reduced individual biases and improved generalization. This solution demonstrates the potential of tree-based ensembles for structured text-driven regression tasks in real-world e-commerce applications.