ML Challenge 2025: Smart Product Pricing Solution

Team Name: Apex
Team Members: Raj Vardhan (Team leader), Sahil Mall, Ayush Jha
Submission Date: [14/10/25]

Executive Summary

We developed an advanced machine learning solution for product price prediction using comprehensive text feature extraction and ensemble modeling. Our approach leverages sophisticated NLP techniques with multiple gradient boosting algorithms, achieving 56.88% SMAPE through robust feature engineering and model optimization.

Methodology

Problem Analysis

Price prediction using textual product descriptions requires extracting semantic patterns, structural characteristics, and implicit quality indicators from unstructured data.

Key Innovations

Multi-layered feature engineering pipeline (148 features)
Three-model ensemble capturing complementary patterns
GPU-accelerated training with early stopping
5-fold cross-validation for robust evaluation

Technical Approach

Feature Engineering Pipeline

Text Processing: TF-IDF vectorization with n-grams (unigrams to trigrams)
Semantic Features: Premium indicators, luxury terminology, quality markers
Structural Features: Text complexity, length metrics, readability scores
Domain Features: Product categories, quantitative attributes, condition status

Ensemble Architecture

XGBoost (GPU): n_estimators=500, max_depth=8, learning_rate=0.05
LightGBM (CPU): n_estimators=500, max_depth=10, learning_rate=0.05
CatBoost (GPU): n_estimators=500, depth=8, learning_rate=0.05
Training Strategy: 5-fold CV, early stopping (50 rounds), log-transformed target

Performance Results

Validation Metrics

Cross-validation SMAPE: 56.88% (±0.54%)
Fold Consistency: 55.96% - 57.53% across folds
Performance Gain: 16.16 percentage points improvement over baseline

Model Interpretation

The ensemble approach demonstrates strong capability in price range estimation using only textual descriptions. Low standard deviation indicates robust generalization across different product segments.

Technical Implementation

Computational Profile

Total Runtime: ~18 minutes
Feature Engineering: ~3 minutes
Model Training: ~15 minutes (GPU-accelerated)
Memory Efficiency: Optimized for large-scale deployment

Key Advantages

Comprehensive Features: 148 engineered features capturing diverse text characteristics
Model Diversity: Complementary algorithms for improved generalization
Optimization: Early stopping prevents overfitting, GPU acceleration reduces training time
Validation: Robust 5-fold cross-validation framework

Conclusion

Our solution demonstrates that sophisticated feature engineering combined with ensemble modeling significantly improves price prediction accuracy using only product descriptions. The 56.88% SMAPE represents a substantial advancement over basic approaches and provides a strong foundation for practical pricing applications.

Enhancement Opportunities

Transformer-based embeddings for semantic understanding
Advanced feature selection and importance analysis
Category-specific hierarchical modeling
Transfer learning from related e-commerce domains

Appendix

Code Repository: [Maintained from original submission]
Production Ready: Efficient batch processing, GPU support, scalable architecture