# Smart Product Pricing Challenge  Final Report

Date: October 2025

## Executive Summary

This report presents the end-to-end solution developed for the Smart Product Pricing Challenge 2025. The competition's objective was to predict optimal product prices for an e-commerce platform using only product descriptions. Accurate price prediction enables improved customer trust, optimized sales strategies, and balanced profitability. Given the large dataset of product titles and catalog content, this project focuses on building a text-driven machine learning solution that can generalize well to unseen data. A rigorous methodology combining advanced text representation and ensemble learning techniques was applied to achieve stable and interpretable performance.

## Methodology

The modeling pipeline began with comprehensive data preprocessing and feature engineering steps. All textual fields were normalized to lowercase and missing values were filled with the token missing'. A custom regular expression-based extractor was implemented to identify and quantify numerical information embedded within text, such as "Pack of 2" or "Set of 3". Three additional numeric features were derived: the total character count, word count, and inferred pack quantity. These features capture valuable cues related to product specifications and packaging. For textual representation, a Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer was employed to transform the catalog text into numerical vectors. The model used both unigrams and bigrams, limiting the feature space to 2200 dimensions to ensure computational efficiency while retaining semantic richness. This approach effectively represents word importance and relationships within the text. The price target variable was log-transformed using log1p(price) to mitigate skewness and stabilize model training, with inverse transformation applied during inference.

## Modeling Approach

The predictive framework was designed as a three-model Gradient Boosted Decision Tree (GBDT) ensemble. The ensemble comprises CatBoost, LightGBM, and XGBoost regressors, each offering complementary learning characteristics. All three models were trained independently on the same TF-IDF and numeric feature inputs using the log-transformed price as the target variable. Model hyperparameters, such as a learning rate of 0.045 and 2500 boosting iterations, were optimized to balance convergence speed and generalization. Early stopping after 100 non-improving rounds was implemented to prevent overfitting. The ensemble strategy involved averaging the logarithmic predictions from the three base models to form a unified output. This approach reduced individual model variance and improved predictive stability. Predictions were subsequently converted back to their original scale using the exponential inverse transformation. Negative or near-zero predictions were clipped to a minimum threshold of 0.01 to comply with challenge submission constraints.

## Experiments and Evaluation

The model evaluation followed a robust 80/20 train-validation split with a fixed random seed to ensure reproducibility. Performance was measured using the Symmetric Mean Absolute Percentage Error

(SMAPE), the official metric for the competition. SMAPE measures the average relative difference between predicted and actual values, expressed as a percentage. It penalizes both overestimation and underestimation symmetrically, making it ideal for pricing applications. During experimentation, the ensemble achieved validation SMAPE scores ranging between 18% and 22%, demonstrating reliable generalization. Each individual model contributed differently: LightGBM excelled in capturing global patterns, while CatBoost handled categorical interactions effectively, and XGBoost offered robustness to outliers. The ensemble consistently outperformed any single model, validating the benefit of multi-model aggregation.

## Conclusion and Future Scope

The final solution represents a balanced combination of statistical learning rigor and practical design. By fusing TF-IDF-based text representations with handcrafted numerical features and an ensemble of three powerful gradient-boosting frameworks, the system achieved strong validation accuracy and interpretability. The model's results align closely with realistic pricing behavior observed in the training data. Future improvements can focus on integrating transformer-based textual embeddings such as DistilBERT or Sentence-BERT to capture deeper contextual meaning. Additional metadata extraction, like brand or category segmentation, could further improve model explainability. Finally, adopting cross-validation-based ensembling could yield more stable leaderboard performance in large-scale deployment. Overall, this approach establishes a competitive and scalable foundation for data-driven product pricing.