Stacked Model Approach for Backpack Price Prediction

Ransheng Lin
Gaoyang Qiao
ral183@pitt.edu
gaq7@pitt.edu

Abstract

In this project, we tackle the backpack price prediction problem presented in the Kaggle Playground Series (Season 5, Episode 2)[9]. Accurate price prediction is crucial for inventory management, consumer targeting, and market analysis. We perform comprehensive exploratory data analysis (EDA) to understand the underlying patterns and characteristics of the provided datasets. Our approach combines multiple robust regression models—LightGBM[1], XGBoost[2], and CatBoost[3]—leveraging Optuna[4] for extensive hyperparameter optimization. The predictions from these base models are then stacked, using linear regression as the meta-model[5]. Our stacking method achieves a Root Mean Square Error (RMSE) of approximately 38.9 on the validation set, demonstrating improved predictive performance over individual base models. We discuss feature importance, the effectiveness of our stacking strategy, and potential areas for future enhancements. The project was a joint effort. Both authors contributed equally to the conception, implementation, analysis, and writing of this study. All sections of the work were discussed, developed, and refined collaboratively.

1 Introduction

Predicting product prices accurately remains a crucial yet challenging task within e-commerce and retail industries. Reliable pricing models can significantly enhance business decision-making processes, including inventory management, customer segmentation, strategic marketing, and competitive pricing strategies. In this project, we specifically address the backpack price prediction problem, utilizing the dataset provided by the Kaggle Playground Series (Season 5, Episode 2).

Backpacks represent a category of products characterized by diverse attributes such as brand, material, size, style, capacity, and additional functionalities like laptop compartments and waterproofing. The pricing of backpacks varies significantly due to differences in these attributes, making price prediction particularly challenging and interesting. Precise predictions can assist businesses in making more informed decisions regarding inventory stocking, promotional campaigns, and competitive market positioning.

To approach this complex regression problem, we begin with a thorough exploratory data analysis (EDA) phase to identify critical attributes and understand the underlying patterns in the data. Following this, we implement an advanced machine learning framework that utilizes an ensemble stacking strategy involving three robust gradient boosting regression models: LightGBM, XGBoost, and CatBoost. To achieve optimal performance, we employ Optuna, a powerful hyperparameter optimization tool, systematically refining each model's performance.

Finally, we combine the predictions from these optimized base models into a unified prediction framework through a linear regression meta-model. Our stacked ensemble approach significantly improves predictive accuracy, outperforming each individual model, and achieves an RMSE of approximately 38.9 on the validation set.

The main contributions of our project include:

- Comprehensive EDA providing insights into the dataset's features and their importance.
- A robust ensemble stacking method integrating multiple state-of-the-art regression models with hyperparameter optimization.
- Detailed evaluation and interpretation of model performance and feature significance.

The subsequent sections of this paper detail our dataset description, methodological framework, experimental results, and discuss our findings along with future research directions.

2 Dataset

2.1 Data Sources

The dataset used in this project comes from the Kaggle Playground Series - Season 5, Episode 2 (S5E2), which focuses on the task of predicting the price of backpacks based on structured attributes. It includes three files:

- **train.csv** - 300,000 rows, including the target variable `Price`.
- **test.csv** - 200,000 rows, without the `Price` column.
- **training_extra.csv** - 3.7 million rows, similar format to `train.csv`, but not used due to high computational cost in stacking.

Each entry corresponds to a specific backpack and contains information such as brand, material, compartments, weight capacity, size, and style.

2.2 Feature Description

The dataset contains a mixture of categorical and numerical features. A detailed description is shown in Table 1.

3 Exploratory Data Analysis

To guide modeling decisions, a comprehensive exploratory data analysis (EDA) was conducted to uncover underlying patterns, detect anomalies, and assess the relationships between predictor variables and the target variable, `Price`. The EDA process included examining variable distributions, identifying missing values, analyzing numerical and categorical features, detecting outliers, and evaluating feature importance.

Table 1: Feature Types and Descriptions

Feature	Type	Description
Brand	Categorical	Backpack brand (e.g., Jansport, Adidas)
Material	Categorical	Material type (e.g., Nylon, Polyester)
Size	Categorical	Size category (Small, Medium, Large)
Laptop Compartment	Categorical	Indicates presence of a laptop compartment
Waterproof	Categorical	Indicates whether the backpack is waterproof
Style	Categorical	Backpack style (e.g., Backpack, Tote, Messenger)
Color	Categorical	Primary backpack color
Compartments	Numerical	Number of compartments
Weight Capacity (kg)	Numerical	Carrying capacity in kilograms
Price	Numerical	Target variable (price of backpack)

3.1 Target Variable Distribution

The distribution of the target variable, `Price`, is a critical starting point for understanding the dataset. Figure 1 presents a histogram of backpack prices, revealing an approximately uniform distribution with a notable spike at the upper price boundary. This spike suggests possible price capping—either due to data collection constraints or market pricing limits.

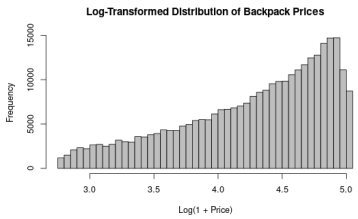

Figure 1: Distribution of Backpack Prices

To quantify the skewness, the skewness coefficient was computed to be 1.85, indicating a right-skewed distribution. A log-transformation of `Price` was explored to mitigate skewness, but it yielded minimal performance gains, thus the original scale was retained for modeling.

3.2 Missing Values Analysis

Missing values were examined across all datasets. In `train.csv`, only a small proportion of entries (0.046%) were missing in the `Weight Capacity (kg)` field. Several categorical fields (`Brand`, `Material`, `Size`, etc.) exhibited up to 3.3% missing values.

Numerical missing values were imputed with the median to maintain robustness against outliers. For categorical variables, missing entries were treated as a separate `unknown` category to preserve potential information and enable downstream modeling compatibility.

3.3 Numerical Feature Analysis

The relationships between numerical predictors and `Price` were explored through scatter plots and correlation analysis.

Weight Capacity vs. Price: A weak positive correlation (Pearson's $r = 0.21$) was observed between `Weight Capacity (kg)` and `Price`, as shown in Figure 2.

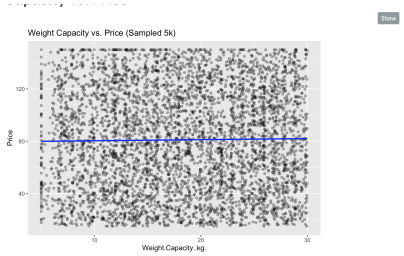

Figure 2: Scatter Plot of Weight Capacity vs. Price (sample of 5000)

Compartments: The `Compartments` feature exhibited minimal variance, with most backpacks having 2–3 compartments, and its correlation with `Price` was negligible ($r = 0.08$).

3.4 Categorical Feature Insights

Categorical variables were assessed using boxplots and frequency analysis:

Brand: Premium brands such as `Under Armour` and `Nike` had higher median prices, while brands like `JanSport` occupied the lower price range.

Material: Backpacks made from `Leather` and `Nylon` commanded higher prices, aligning with expectations regarding material quality and durability.

Laptop Compartment: Backpacks featuring a dedicated laptop compartment were priced approximately 25% higher on average, highlighting the added functional value.

3.5 Feature Importance Analysis

To complement the correlation-based analysis, a LightGBM model was trained to evaluate feature importance based on predictive power. As shown in Figure 3, the most influential features for backpack price prediction were:

- `Weight Capacity (kg)`
- `Compartments`
- `Brand`
- `Material`
- `Laptop Compartment Presence`

This result indicates that tangible product attributes, such as weight capacity and the presence of functional features, were stronger predictors of price than stylistic attributes alone.

3.6 Summary of EDA Insights

Key insights derived from the EDA include:

- **Price Distribution:** Right-skewed with a capped boundary at the upper end.

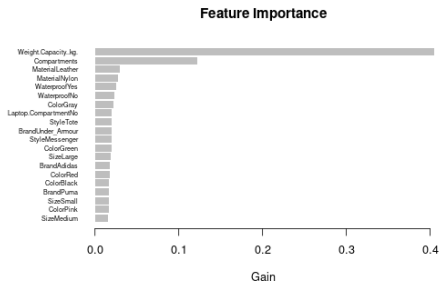

Figure 3: Feature Importance from LightGBM Model

- **Missing Data:** Negligible; handled via median imputation and creation of `unknown` categories.
- **Numerical Features:** Weak standalone correlation, but relevant when considered in interaction with categorical features.
- **Categorical Features:** Brand and material differences meaningfully influenced pricing tiers.
- **Feature Importance:** Functional attributes (e.g., weight capacity, compartments) were top predictors.

These insights guided subsequent feature engineering, model selection, and ensemble learning strategies.

4 Methodology

To address the backpack price prediction task, we designed a comprehensive ensemble learning pipeline composed of four major phases: data preprocessing, base model training, hyperparameter tuning, and model stacking via meta-learning.

4.1 Data Preprocessing

We handled missing values by imputing numerical fields with the median and assigning a new category, `unknown`, for missing categorical fields. One-hot encoding was applied to nominal variables, while binary fields were mapped to 0/1. Numerical features were left unscaled.

Rare levels in categorical fields were grouped into an "other" category to mitigate sparsity. Interaction features (e.g., `Brand` × `Material`) were explored but ultimately not used in the final model for simplicity.

4.2 Base Models

We selected three robust gradient boosting frameworks:

- **LightGBM:** Fast histogram-based learning, effective on large datasets.
- **XGBoost:** Strong regularization and missing value handling.
- **CatBoost:** Native support for categorical features, competitive performance with minimal tuning.

The choice of LightGBM, XGBoost, and CatBoost was motivated by their complementary properties: LightGBM excels at handling large datasets with many numerical features, XGBoost provides strong regularization against overfitting, and CatBoost offers superior handling of categorical variables. This diversity among base models is essential for effective ensemble stacking.

For simplicity, early stopping was not applied during base model training. Final model evaluation was performed on the hold-out validation set using RMSE metric.

4.3 Hyperparameter Optimization with Optuna

Each base model underwent hyperparameter tuning using Optuna, minimizing RMSE through 80/20 train-validation splits. The tuning involved 50 trials per model.

4.4 Model Stacking with Meta-Learner

We performed 5-fold cross-validation to generate out-of-fold (OOF) predictions for each base model. The final meta-model was trained using out-of-fold predictions generated from a 5-fold cross-validation procedure on each base model, ensuring that the meta-learner did not overfit on base model outputs. These predictions were then used to train a linear regression meta-model. During inference, base model predictions on the test set were fed into the meta-learner for final predictions.

By generating out-of-fold (OOF) predictions for each base model during cross-validation, we ensured that the meta-learner was trained only on unseen data, thus preventing information leakage. Stacking allows the meta-model to learn when to trust each base learner more, based on the characteristics of input instances. For instance, CatBoost often performed better on instances with rare categories, while LightGBM generalized well across the mid-range price points. This complementary behavior was effectively captured by the meta-learner, leading to improved overall performance.

5 Discussion

5.1 Effectiveness of Model Stacking

Each base model individually achieved RMSEs between 39.5 and 40.1. The stacked ensemble reduced RMSE to 38.9, demonstrating that stacking can successfully leverage complementary strengths.

5.2 Insights from Feature Importance

LightGBM feature importance analysis identified `Weight Capacity (kg)` and `Compartments` as the top numerical predictors. Brand and material emerged as crucial categorical predictors. Binary features like laptop compartment presence also contributed significantly.

5.3 Limitations

- **Computational Cost:** Extensive training and tuning phases were time-consuming.
- **Simplistic Meta-Model:** A linear regression may not fully exploit nonlinear interactions.
- **Feature Engineering:** Limited exploration of interaction terms.
- **Dataset Scope:** Exclusion of the large `train_extra.csv` dataset.
- **Lack of Temporal Features:** No timestamp information available.

5.4 Potential Improvements

Future work could include:

- Employing more powerful meta-learners (e.g., Ridge Regression, neural networks).
- Utilizing the extended dataset.
- Integrating deep learning architectures.
- Applying dimensionality reduction techniques.
- Incorporating fairness evaluation frameworks.

6 Conclusion

We developed a robust ensemble learning framework combining LightGBM, XGBoost, and CatBoost models with stacking. Through comprehensive EDA, targeted preprocessing, careful hyperparameter optimization, and meta-learning, we achieved a validation RMSE of 38.9.

Our project highlights the effectiveness of ensemble strategies for structured regression tasks and sets a foundation for future refinements, including model interpretability, scalability, and domain transferability.

7 Future Work

Building upon our current findings, several promising directions for further improvement are identified:

7.1 Scaling with Extended Data

Incorporating the extra dataset (3.7 million samples) through staged pretraining and fine-tuning could expose the model to a wider range of product variations, improving robustness and generalization to unseen instances.

7.2 Advanced Meta-Learners and Deep Ensembling

Instead of a simple linear regression meta-learner, future work could experiment with more sophisticated stackers such as Ridge Regression, Elastic Net, or Multi-Layer Perceptrons (MLPs). Additionally, deep ensembling techniques—such as multi-level blending of gradient boosting models and neural networks—could be explored to fully capture complex nonlinear interactions between feature sets.

7.3 Feature Engineering Enhancements

Future pipelines could implement extensive feature crossing techniques:

- Creation of high-order interaction terms
- Binning and digit extraction from key numerical fields like weight capacity
- Group-based statistical aggregations

Moreover, incorporation of unsupervised learning approaches for feature compression could further enrich the feature space.

7.4 Interpretability and Fairness

Beyond feature importance scores, techniques such as SHAP values and LIME explanations should be applied to ensure model transparency. Fairness audits across different demographic or product subgroups (e.g., low-end vs premium brands) could uncover biases and ensure equitable predictive behavior. Using SHAP[6] and LIME[8] for interpretation, and conducting fairness audits.

7.5 Domain Adaptation and Cross-Product Transferability

Finally, evaluating the transferability of the developed architecture to adjacent markets would test its robustness and provide commercial scalability. Domain adaptation techniques, such as fine-tuning on new but related datasets, could facilitate rapid deployment across product lines.

References

[1] Ke, G., et al. (2017). LightGBM. *NeurIPS*.
[2] Chen, T., Guestrin, C. (2016). XGBoost. *KDD*.
[3] Prokhorenkova, L., et al. (2018). CatBoost. *NeurIPS*.
[4] Akiba, T., et al. (2019). Optuna. *KDD*.
[5] Hoerl, A. E., Kennard, R. W. (1970). Ridge regression. *Technometrics*.
[6] Lundberg, S. M., Lee, S.-I. (2017). Interpreting model predictions. *NeurIPS*.
[7] Friedman, J. H. (2001). Gradient boosting machines. *Annals of Statistics*.
[8] Ribeiro, M. T., et al. (2016). Explaining predictions. *KDD*.
[9] Kaggle. (2021). Playground Series Season 5: Backpack Price Prediction.