



Predicting Sephora Product Popularity with Linear Regression and XGBoost

Yueran Ling

New York University Shanghai, Shanghai 200124, China
y113167@nyu.edu

Abstract. This project investigates whether product-level information can predict the popularity of Sephora products. The target variable is the loves count, defined as the number of users who have saved or expressed interest in a product. Because the raw loves count is heavily right-skewed, the analysis models log (loves count + 1). The data are split into 80% training and 20% validation. Linear regression and XGBoost are compared using mean squared error (MSE) and R^2 on the log-transformed target. Linear regression achieves a validation MSE of 1.7998 and an R^2 of 0.4503. XGBoost achieves a lower MSE of 1.1018 and a higher R^2 of 0.6635. These results suggest that nonlinear relationships and feature interactions are important for predicting product popularity. The improvement from linear regression to XGBoost indicates that tree-based models capture complex patterns more effectively. Feature importance analysis further reveals that product category, reviews, and variation type are among the most influential predictors. Therefore, product-level attributes alone provide meaningful predictive signals for popularity on e-commerce platforms.

Keywords: Sephora; Product popularity prediction; Machine learning; XGBoost; Feature importance

1 Introduction

Predicting which products will become popular on online beauty retail platforms has clear business value. Merchandising teams need to decide which products to promote, assortment planners need to allocate shelf space, and marketers need to design campaigns around likely bestsellers. Sephora, as a major beauty retailer, generates large amounts of product-level data, including brand, category, price, rating, review count, and flags such as “new” or “online only”. These signals may help explain why some products receive far more customer attention than others. The dataset used in this project is moderately sized: it runs quickly on a personal computer but still contains meaningful patterns, with a mix of continuous variables (price, reviews), binary flags (online only), and high-cardinality categorical variables (brand name). These properties make the problem a natural test case for comparing simple and flexible machine learning models.

Previous studies have shown that machine learning works well for product popularity forecasting. Liu et al. combined LSTM with XGBoost to predict short-term retail popularity and reported high accuracy and reliability [1]. Zhang and Wu applied XGBoost to e-commerce sales forecasting and found that it generated accurate predictions by capturing complex nonlinear relationships [2]. In a retail setting, Dankorpho also obtained a significant enhancement in prediction accuracy using the XGBoost algorithm [3]. These studies confirm that tree-based methods, especially XGBoost, are strong candidates for popularity prediction tasks.

This paper uses an existing Sephora product dataset to compare linear regression with XGBoost, using the loves count (number of users who saved a product) as the target, and interprets feature importance to identify the most predictive signals.

2 Method

2.1 Dataset

The dataset used in this project is a publicly available Sephora product and skincare review dataset from Kaggle [4]. It contains 8,494 product rows and 27 columns. The target variable is the loves count, defined as the number of users who have saved or expressed interest in a product. Because the original loves count distribution is heavily right-skewed, the analysis uses a log transformation: $y = \log(\text{loves count} + 1)$. The addition of 1 ensures that products with zero loves remain in the dataset. This transformation reduces the influence of extreme observations and stabilizes the comparison between models. An 80/20 random split of the data into training and validation sets is performed with a fixed random seed of 42, consistent with standard validation practices [5].

2.2 Feature Engineering

The input features consist of numerical variables (rating, reviews, price in USD, value price in USD, sale price in USD, child count, child max price, child min price), binary variables (limited edition, new, online only, out of stock, sephora exclusive), and categorical variables (brand name, variation type, primary category, secondary category, tertiary category). Missing numerical values are imputed using the median from the training set. Missing categorical values are filled with a "Missing" category and then one-hot encoded. One-hot encoding is a standard method for converting categorical variables into numerical features; the implementation in this project uses scikit-learn [6]. The same preprocessing pipeline is applied to both models to ensure a fair comparison.

2.3 Models

Two models are compared: linear regression and XGBoost.

Linear regression serves as a baseline model, assuming a linear relationship between the features and the log-transformed target and offering an interpretable benchmark. Standard textbooks describe linear regression as a common starting point for regression tasks because it is simple and useful for checking whether a more complex model is necessary. However, linear models may perform poorly when the relationship between features and the target is nonlinear [7].

To capture nonlinear patterns, tree-based methods are a natural choice. Classification and regression trees (CART) introduced recursive partitioning [8]. Random forests improved predictive accuracy by averaging many randomized decision trees [9]. Gradient boosting introduced a different ensemble strategy: instead of training independent trees, it fits new trees to correct previous errors [10]. XGBoost extends gradient boosting with regularization and scalable engineering, which is one reason it is widely used for tabular machine learning [11]. In this project, XGBoost is compared with linear regression to test whether a nonlinear model improves the prediction of Sephora product popularity.

The models are evaluated using mean squared error (MSE) and R^2 on the log-transformed target. A lower MSE corresponds to a smaller prediction error, whereas a higher R^2 implies that a larger proportion of the variance in the target is explained by the model. Evaluation is performed on the validation set rather than the training set to measure generalization performance [5].

3 Results

The dataset contains 8,494 product rows and 27 columns, which satisfies the project guideline of using a dataset with more than 300 rows and fewer than 10,000 rows. The validation results are shown in Table 1. XGBoost outperforms Linear Regression, reducing validation MSE from 1.7998 to 1.1018 and increasing R^2 from 0.4503 to 0.6635. These results suggest that product popularity is not explained only by simple linear effects. XGBoost performs better because it can model nonlinear relationships and interactions.

As shown in Fig. 1, the most important features include primary category (Makeup), reviews, variation type (Color), online only, tertiary category (Leave-In Conditioner), variation type (Type), and several brand or category indicators. This indicates that product category, engagement, variation type, online availability, and brand all help predict popularity.

Three of these features deserve further discussion from a business perspective. Primary category (Makeup) ranks highest. A plausible reason is that makeup products constitute a large share of Sephora's assortment and tend to generate more user engagement (e.g., swatches, tutorials, seasonal trends) than skincare or hair care. Reviews are another top predictor. This aligns with the social proof effect: consumers often rely on peer ratings and written feedback when deciding which products to save or buy, so a high review count naturally correlates with higher popularity. Online only also appears important. Products sold exclusively online may benefit from targeted digital marketing campaigns and scarcity cues (limited availability), which drive user

savings and interest. Feature importance should not be interpreted as causality; reviews, for example, may be correlated with love count rather than causing it.

Linear regression explains about 45.0% of the validation variation in log-transformed love count, while XGBoost explains about 66.3%. The log transformation makes the evaluation more stable and makes the prediction task closer to comparing relative popularity levels.

Table 1. Validation performance on log-transformed loves_count.

Model	Validation MSE	Validation R ²	Notes
Linear Regression	1.7998	0.4503	Baseline model
XGBoost	1.1018	0.6635	Best validation performance

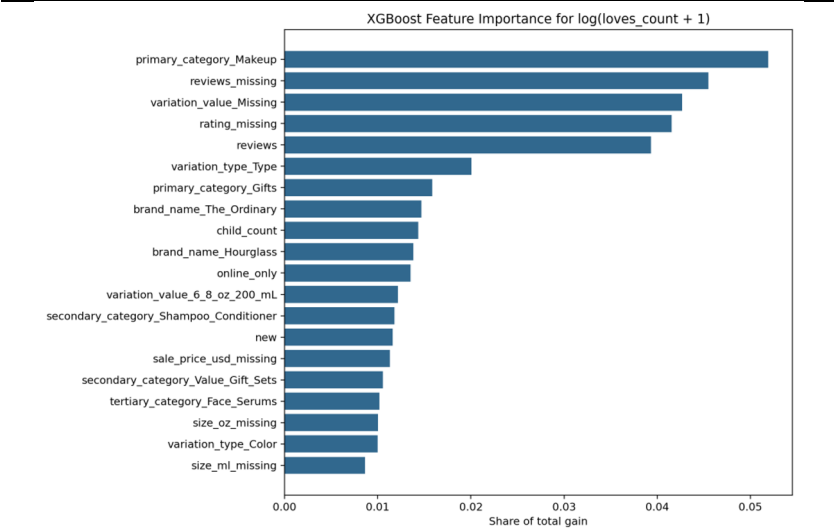


Fig. 1. Top 20 XGBoost feature importances (Photo/Picture credit: Original).

4 Discussion

The results show that nonlinear modeling is valuable for predicting Sephora product popularity. XGBoost performs better because it captures interactions among product attributes, not just linear effects. For example, a product category may matter differently when combined with brand, review count, online availability, or variation type. This is consistent with the known strength of boosted trees on tabular data [11]. The advantage of XGBoost over other boosting methods is also seen in a 2025 e-commerce study by Zhang, who compared GBDT, LightGBM, and XGBoost for online sales prediction and found XGBoost consistently gave the highest accuracy [12].

However, prediction does not imply causation. The model predicts love count from product information, but some variables—like reviews—may be correlated with popularity rather than causing it. Thus, the model is useful for identifying predictive patterns, not for claiming that any single feature directly drives popularity.

Interpretability also matters. Linear regression is easier to explain because each coefficient has a clear meaning. XGBoost performs better but is less transparent. Feature importance gives a rough summary, but it does not show interactions or specific predictions. More detailed methods like SHAP could help; Liu et al. used SHAP in an e-commerce setting to explain how product description phrases affect sales forecasts, improving transparency [13].

A limitation is that the dataset is cross-sectional—it captures product information at one point in time, not how popularity evolves. Therefore, the project cannot tell whether a product becomes popular before or after reviews increase. Future work could use time-stamped data to separate early attributes from later feedback, which would be better for predicting pre-launch popularity.

Other future improvements include adding text features from product names, descriptions, or customer reviews. Beauty products often depend on qualitative signals (skin concerns, formulation, shade range, brand identity) that structured variables miss. Also, hyperparameter tuning and cross-validation could test whether the XGBoost results are stable across different data splits.

5 Conclusion

This project predicts Sephora product popularity using love count as the target variable. Since the target is highly skewed, models are trained on $\log(\text{loves count} + 1)$. Linear regression is compared with XGBoost using an 80/20 train-validation split, and performance is evaluated using MSE and R^2 . The experimental results show that XGBoost performs better than linear regression. XGBoost achieves a validation MSE of 1.1018 and R^2 of 0.6635, compared with linear regression's MSE of 1.7998 and R^2 of 0.4503. This supports the conclusion that nonlinear machine learning methods can be useful for predicting consumer product popularity from product-level features.

There are several limitations. The dataset is observational, so the model cannot establish causal relationships. Some predictive features, such as reviews, are closely related to popularity itself. Future work could remove engagement variables to predict pre-launch popularity, use richer text features from product names or ingredients, tune model hyperparameters more carefully, or test whether the model generalizes to other retailers.

Overall, the project demonstrates a complete machine learning workflow: identifying a target variable, cleaning and transforming data, building a fair model comparison, evaluating with quantitative metrics, and interpreting model behavior. The strongest conclusion is that XGBoost is a better predictor than linear regression for this Sephora dataset, but the result should be understood as predictive rather than causal.

References

1. Liu, X., Deng, C., Lu, B., Mo, Y., Li, P., Chen, K.: Short-term Popularity Forecasting Method of Retail Products Based on Combined LSTM-XGBoost Model. In: 2024 International Conference on Intelligent Computing and Data Mining (ICDM), pp. 129–133 (2024)
2. Zhang, M., Wu, L.: Sales Forecasting and Data-Driven Marketing Strategies for E-Commerce Platforms Using XGBoost. *International Journal of Intelligent Information Technologies* (2025)
3. Dankorpo, P.: Sales Forecasting for Retail Business using XGBoost Algorithm. *Journal of Computer Science and Technology Studies* (2024)
4. Kaggle: Sephora Products and Skincare Reviews, <https://www.kaggle.com/datasets/nadyinky/sephora-products-and-skincare-reviews>, last accessed 2026/06/16
5. James, G., Witten, D., Hastie, T., Tibshirani, R., Taylor, J.: *An Introduction to Statistical Learning: With Applications in Python*. Springer (2023)
6. Breiman, L.: Random Forests. *Machine Learning* 45, 5–32 (2001)
7. Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer (2009)
8. Breiman, L., Friedman, J., Olshen, R., Stone, C.: *Classification and Regression Trees*. Wadsworth (1984)
9. Friedman, J.: Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics* 29(5), 1189–1232 (2001)
10. Lundberg, S.M., Lee, S.I.: A Unified Approach to Interpreting Model Predictions. In: *Advances in Neural Information Processing Systems* (2017)
11. Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794 (2016)
12. Zhang, J.: Enhancing Predictive Models in E-Commerce: A Comparative Study Using XGBoost Across Diverse Scenarios. *ITM Web of Conferences* 70, 02014 (2025)
13. Liu, R., et al.: Which product description phrases affect sales forecasting? An explainable AI framework by integrating WaveNet neural network models with multiple regression. *Decision Support Systems* 176, 114065 (2024)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

