Analyzing Consumer Purchase Intent in Beauty Products: Impact of Reviews, Ingredients, and User Context

(COMP3125 Individual Project)

Kledja Caushi
Wentworth Institute of Technology

Abstract— **Consumer decision-making in beauty products is significantly influenced by factors such as product reviews, pricing strategies, and ingredient transparency. This research investigates how these variables affect purchase intention, utilizing a dataset of product reviews and product metadata from Sephora.com. By applying exploratory data analysis, sentiment analysis, and logistic regression modeling, this study evaluates the effects of consumer reviews, controversial ingredients, and user demographics on consumer purchase decisions. Results demonstrate that high product ratings strongly correlate with increased purchase likelihood, whereas controversial ingredients like fragrance and talc exhibit minor negative impacts. Additionally, user-specific factors such as skin type and skin tone notably influence product satisfaction and recommendation rates. The predictive logistic regression model developed in this research achieves high accuracy (ROC AUC of 0.9985), effectively estimating purchase intent based on review and product characteristics. This analysis provides valuable insights into consumer preferences and highlights areas for potential product improvement and targeted marketing strategies in the beauty industry.**

Keywords—Data Analysis, NLP, Logistic Regression, Website Scraping, Predictive Analysis

I. Introduction

Understanding consumer purchase behavior is crucial for product success, particularly in the highly competitive beauty industry. Factors influencing these decisions include user-generated content such as reviews, pricing strategies, ingredient composition, and user demographics. Existing research highlights the significance of online reviews in shaping consumer perceptions, but gaps remain in quantifying the combined impact of reviews, controversial ingredients, and personalized user contexts such as skin type and tone. This study aims to bridge these gaps by investigating the relationship between these factors and consumer purchase intentions on Sephora.com. By leveraging analytical techniques, including sentiment analysis and predictive modeling, this research contributes to a deeper understanding of consumer behavior, aiding businesses in strategic decision-making and enhancing consumer satisfaction. Prior studies emphasize that transparency and consumer trust directly affect sales and brand loyalty; hence, evaluating these elements provides strategic insights for stakeholders in the beauty sector [2].

II. Datasets

A. Source of dataset

The dataset utilized in this study comprises customer reviews and product metadata extracted from Sephora.com, a reputable online beauty retailer. Data collection occurred in April 2024, capturing detailed product reviews and comprehensive product descriptions, including ingredients and pricing information. Notably, the review data was captured directly from Sephora.com by monitoring the website's API requests. This involved identifying and studying the schema provided by the Bazaarvoice Conversations API, as documented at https://developers.bazaarvoice.com/v1.0-ConversationsAPI/reference/get_data-reviews-json.

Subsequently, reviews for all products in our list were systematically captured and saved as raw JSON data (reviews_raw). These raw JSON files were then flattened and cleaned using DuckDB to create the final analysis-ready dataset (reviews_clean). Additionally, product metadata was compiled through direct parsing of Sephora.com product listings and enriched with supplementary data from Kaggle, resulting in the products_raw dataset. This unique data collection approach distinguishes our project and ensures accuracy and relevance to current market trends.

B. Character of the datasets

The analysis involves two primary datasets: reviews_clean and products_raw. The reviews_clean dataset includes approximately 17,000 entries with key features such as user rating (scale 1-5), recommendation status (boolean), helpfulness score, and detailed user context (skin type, skin tone, hair color, eye color, incentivized review status). The products_raw dataset includes product ID, name, brand, pricing (USD), and a comprehensive ingredients list. These datasets were joined on the product ID, merging review-level data with corresponding product information. Table 1 summarizes key dataset characteristics.

Table 1. Data Description

Dataset	**Rows**	**Columns**	**Key Features**
reviews_clean	17,000	14	rating, recommendation, helpfulness, user context attributes (skin type, tone, etc.)
products_raw	2,500	27	product name, brand, price, ingredients

Data preprocessing included converting categorical ingredients into binary flags for controversial ingredients

979-8-3315-0993-4/25/$31.00 ©2025 IEEE

(paraben, fragrance, talc, etc.), normalization of numerical features such as helpfulness scores, and cleansing of inconsistent entries.

III. Methodology

This research aims to answer four primary questions:

1. How do product reviews, ratings, and the number of likes influence the likelihood of a purchase?
2. How do consumer characteristics such as skin type, skin tone, hair color, eye color, and age affect product rating and recommendation likelihood?
3. How do specific ingredients in beauty products, particularly controversial ones, affect consumer purchasing decisions?
4. Can we build a predictive model to estimate purchase probability based on reviews, likes, price, and ingredients?

To address these questions, a combination of exploratory data analysis, sentiment analysis, and logistic regression modeling was used.

A. Exploratory Data Analysis (EDA)

EDA was used to summarize key patterns in consumer reviews and product metadata. This included grouping and visualizing average ratings and recommendation rates based on consumer attributes such as skin type and tone. EDA helped highlight preliminary trends and served as a diagnostic tool before formal modeling. Tools used included DuckDB for extraction, Pandas for grouping and aggregation, and Seaborn/Matplotlib for plotting. Missing values in user contexts were left intact to preserve original response patterns and allow comparison across subgroups.

B. Sentiment and Text Analysis

To understand how language differs between high-rated and low-rated reviews, reviews were grouped by rating category and analyzed using frequency analysis and word clouds. This qualitative step helped uncover the most common themes among satisfied and dissatisfied consumers. Stop words were removed and remaining tokens visualized using the WordCloud library. No stemming or lemmatization was applied to preserve context.

C. Predictive Modeling: Logistic Regression

Logistic regression was employed to model the likelihood of a product being recommended, based on key features: rating, price, number of likes, and ingredient flags. Assumptions include independence of predictors and linearity in the log-odds space. The model was chosen for its interpretability and performance in binary classification problems. Python's sklearn.linear_model.LogisticRegression was used for modeling, with stratified train/test splits to address class imbalance. ROC AUC score was used for performance evaluation, and feature importance was assessed through model coefficients [4].

These methods collectively enabled both descriptive insights and predictive inferences to answer the guiding research questions.

IV. Results

This section explicitly addresses each of the four main research questions identified in the Introduction, presenting findings through numerical summaries, visualizations, and interpretive explanations.

A. Influence of Reviews, Ratings, and Likes on Purchase Likelihood

Analysis revealed a strong positive correlation between product ratings and consumer purchase intent. Products with ratings of 4 stars or above exhibited a significantly higher likelihood of being recommended by consumers. Furthermore, the total number of positive feedback counts (likes) correlated positively with product recommendation rates. Visualization through bar charts clearly depicted this relationship, reinforcing the notion that positive consumer engagement and ratings critically influence purchase decisions. [1]

B. Effect of Consumer Characteristics on Ratings and Recommendations

Consumer characteristics such as skin type, skin tone, hair color, eye color, and age were found to significantly influence both product ratings and recommendation likelihood. Exploratory analysis showed that users with 'oily' or 'combination' skin types tended to provide higher ratings compared to those with 'dry' skin. Similarly, users with 'light' or 'medium' skin tones had marginally higher recommendation rates. Hair and eye color had less pronounced but still notable effects, particularly when correlated with product type (e.g., foundations or eye products). Missing or unreported user context attributes were associated with greater variance in ratings. These findings suggest that user-specific attributes play a meaningful role in shaping product experiences and should be considered in personalized product marketing and development strategies. [1][2]

C. Effect of Specific Ingredients on Purchase Decisions

The presence of controversial ingredients, specifically fragrance and talc, had a measurable negative effect on consumer recommendation rates and overall satisfaction. Analysis showed consumers consistently rated and recommended products without these controversial ingredients more highly. This pattern, illustrated through comparative visualizations, highlights consumer preference for ingredient transparency and safety. [3]

D. Predictive Modeling of Purchase Intent

A logistic regression predictive model was developed to estimate the probability of product purchase based on reviews, likes, price, and ingredient data. The model demonstrated exceptional predictive accuracy, achieving a ROC AUC score of 0.9985. Key predictors identified were product rating (strong positive influence), price (moderate negative influence), and presence of controversial ingredients (slight negative influence). This model provides a robust framework for predicting consumer purchase intentions and offers valuable insights for strategic product development and marketing.

V. Discussion

Despite the robust performance of the logistic regression model, certain limitations were identified. The model's reliance on proxy variables for actual purchase data (recommendation status and high ratings) could lead to discrepancies when predicting actual consumer behavior. Additionally, the analysis focused on a limited set of controversial ingredients; future research could explore a broader array of ingredients. Enhancing the model with additional consumer behavior metrics, real purchase data, and expanding sentiment analysis methods could further improve predictive accuracy and actionable insights.

VI. Conclusion

This study offers a comprehensive exploration of the factors influencing consumer purchase intent in the online beauty market, focusing on Sephora.com as a case study. By analyzing user-generated reviews, ingredient lists, product metadata, and individual consumer attributes, we have shown that product ratings, ingredient transparency, and personalized user context significantly drive consumer decisions.

In particular, products with higher ratings and more positive feedback were more likely to be recommended, while the presence of controversial ingredients like fragrance and talc slightly reduced satisfaction metrics. Additionally, consumer features such as skin type and tone were found to meaningfully shape product perception, highlighting the importance of personalization in marketing and product design.

The predictive model developed in this project demonstrates strong performance in estimating purchase likelihood, offering valuable tools for stakeholders in the beauty industry to refine targeting strategies and improve product formulations. These findings underscore the growing value of data-driven approaches in enhancing user experience, trust, and commercial performance.

Future research could expand this framework by integrating real transaction data, exploring causal relationships, or applying natural language models for deeper sentiment extraction. As consumers become increasingly aware of product content and peer feedback, brands that prioritize transparency and personalization will be better positioned to meet evolving expectations in a competitive digital marketplace.

References

[1] Chen, Tao et al. “The Impact of Online Reviews on Consumers' Purchasing Decisions: Evidence From an Eye-Tracking Study.” Frontiers in psychology vol. 13 865702. 8 Jun. 2022, doi:10.3389/fpsyg.2022.865702

[2] T. Macheka, “The effect of online customer reviews and celebrity endorsement on Young Female Consumers’ purchase intentions,” Young Consumers, https://www.emerald.com/insight/content/doi/10.1108/yc-05-2023-1749/full/html

[3] M. K. Reilly, “Controversial beauty ingredients: Chemical ingredients in beauty products are coming under increased scrutiny,” Nutritional Outlook,https://www.nutritionaloutlook.com/view/controversial-beauty-ingredients-chemical-ingredients-in-beauty-products-are-coming-under-increased-scrutiny

[4] Scikit-learn developers. "Logistic Regression." https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html