Hybrid Skincare Recommendation System

Capstone 3 Final Report

By: Valentina Sanchez

Introduction

Understanding that skincare efficacy is highly individualized. The impact of a product varies significantly across different skin types, with ingredients playing a crucial role in both the potential benefits and risks, such as irritation or allergic reactions. Additionally, this system acknowledges the consumer's interest in finding cost-effective products without compromising on quality. By exploring alternatives and ingredient substitutes that are more readily available, it is often possible to achieve similar results to those of higher-priced items.

The Skincare Product Recommendation System introduces a personalized method for suggesting skincare products, capitalizing on user reviews and detailed product data. Its goal is to align users with the skincare items that most closely align with their individual needs and preferences. This system seeks to simplify the often overwhelming process of choosing from the myriad skincare options on the market, thereby improving the shopping experience with customized recommendations.

Dataset Overview

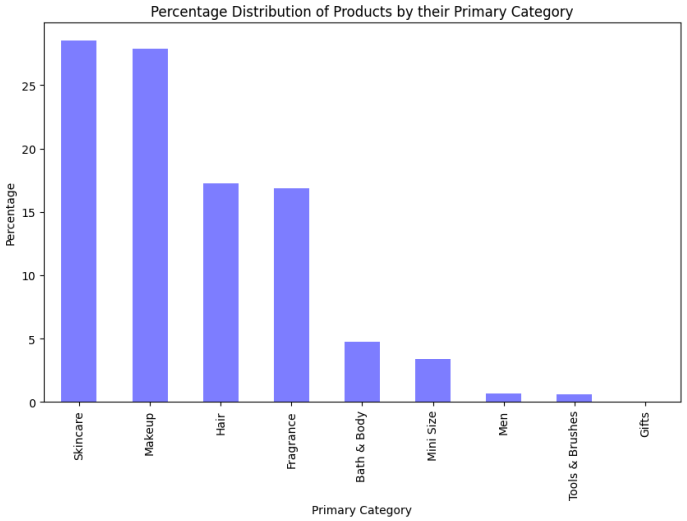
The dataset was sourced from a Kaggle user who performed web scraping on the Sephora website, resulting in the creation of multiple tables. Our dataset is primarily organized into five review tables, each encompassing a distinct range of skincare products along with their reviews, contributed by various authors. For instance, one of the CSV files, named 'reviews0-250.csv,' comprises data on 250 skincare products extracted from the Sephora website.

In addition to the review tables, our dataset includes 'product_info.csv,' a table with a size of (8494, 27). Unlike the review tables, 'product_info.csv' encompasses a broader spectrum of product categories beyond skincare. In our analysis, we will commence by examining this table, defining our criteria for identifying skincare products, and implementing appropriate filters accordingly.

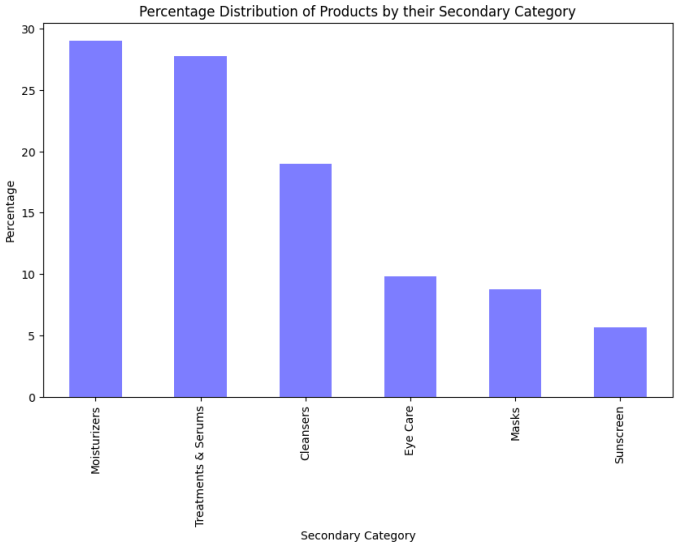
Link to Kaggle dataset

Exploratory Data Analysis

Distribution of Products by the Primary and Secondary Category

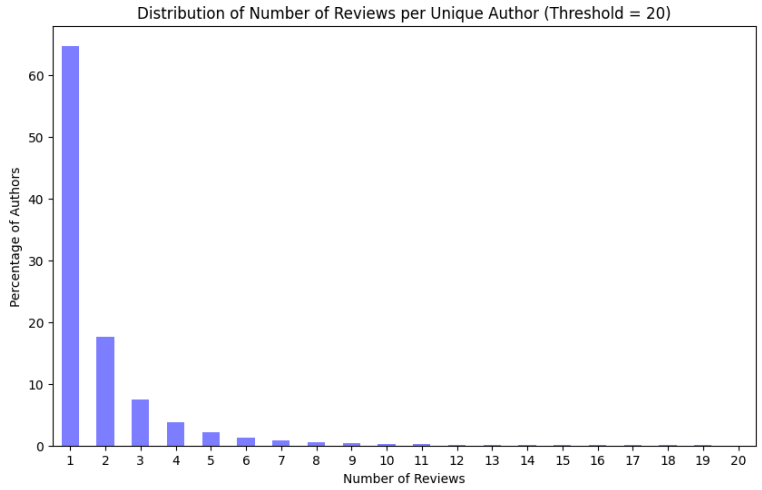


Looking at the product information, the primary category that stands out is skincare as the most populated category among various product types, ensuring a solid data foundation for our skincare-specific recommendation system.



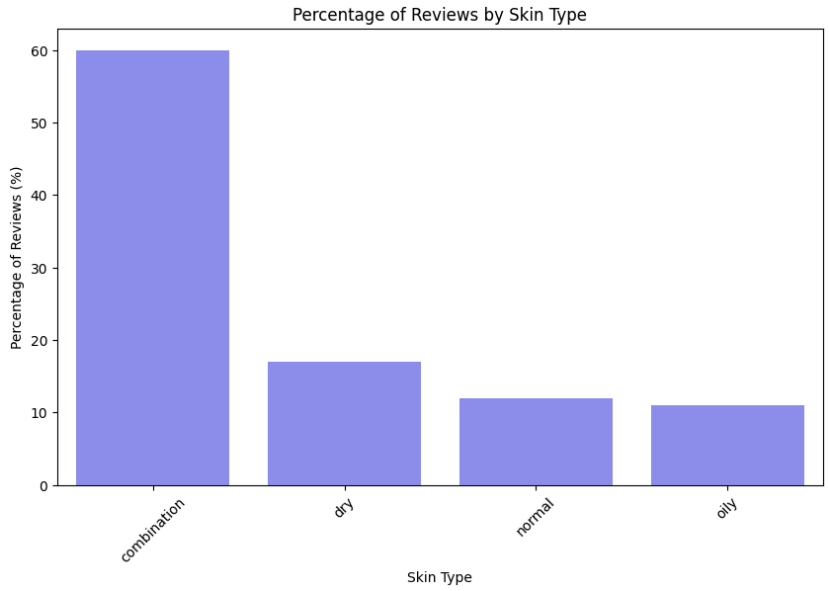
In the secondary product category, we are focusing exclusively on skincare and refining it into detailed subcategories: Moisturizers, Treatments & Serums, Cleansers, Eye Care, Masks, and Sunscreen to improve system relevance. Initially, 'Treatments' and 'Lip Balms & Treatments' were separate categories, but we merged them for simplicity.
For tertiary categories and a more in-depth analysis of products, we removed items that didn't align with our recommendations or those with numerous ingredients aimed at enhancing skin over time, such as Facial Rollers, BB & CC Creams, Face Wipes, etc..

Distribution of Number of Reviews per Unique Author (Threshold = 20)



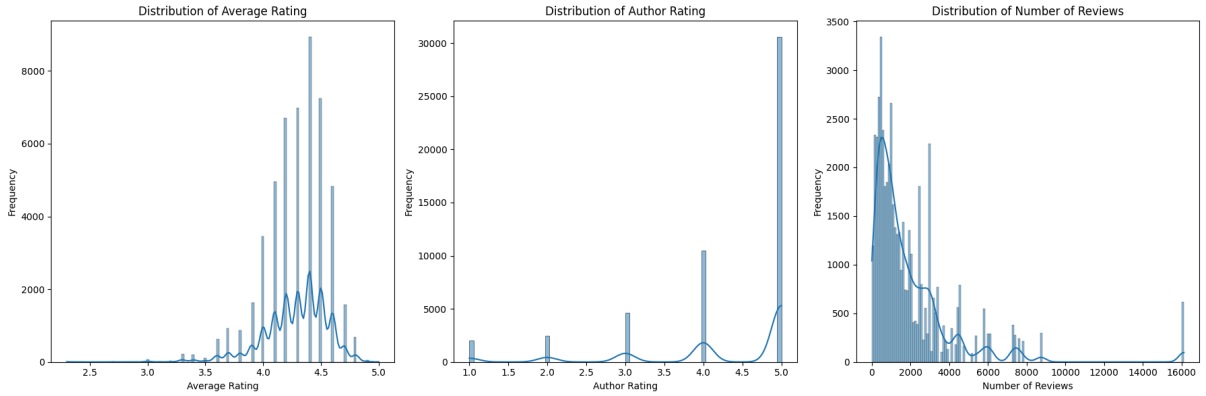
The bar chart illustrates the frequency distribution of reviews penned by distinct authors, setting a maximum threshold at 20 to maintain a clear visual representation of the decline in the number of reviews per author. The number of reviews per unique author, the highest number of reviews from a single author is 257, whereas the mean number of reviews is just 1. Over 60% of authors have contributed a single review, indicating that authors who submit more than a few reviews are quite rare. This pattern suggests that the dataset is predominantly composed of one-time reviewers as opposed to a small group of prolific contributors.

Percentage of Reviews by Skin Type



A vast majority of the reviews come from individuals with combination skin, which accounts for over 60% of the reviews. This could either mean that most combination skin users either hate or love products enough to write a review about them, they tend to have the most problematic skin. The least number of reviews are from those with oily skin over 10%.

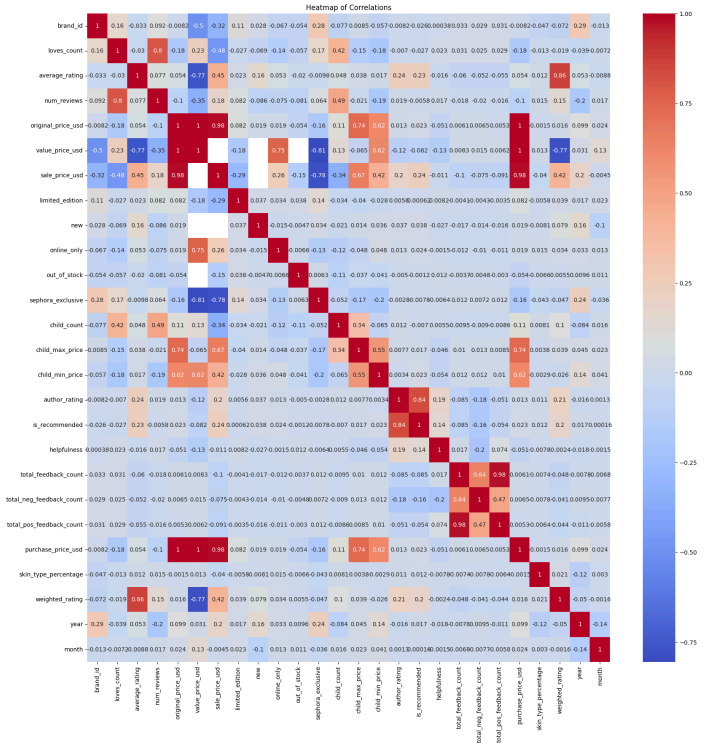
Distribution of Average Rating, Author Rating and Number of Reviews



In all three distributions, the presence of long tails suggests that while there are common values that occur frequently (such as a large number of products with average ratings around 4 to 5), there are also outliers with very high or very low values. This kind of information is valuable for understanding consumer behavior and product reception in the context of the skincare recommendation system. For the recommendation engine to take into account the skewness and high variability and treat it with normalization.

Heatmap of Correlations

This heatmap displays the correlation coefficients between various variables in a dataset. High positive correlations (dark blue squares) can be seen between variables like 'num_reviews' and 'total_feedback_count', which makes intuitive sense since more reviews likely lead to more feedback. There are also strong negative correlations (dark red squares) observable, such as between 'out_of_stock' and 'num_reviews', indicating that items that are out of stock may have fewer reviews. Variables like 'brand_id', 'original_price_usd', and 'sale_price_usd' show little to no correlation (white squares) with many other variables, suggesting they don't have a linear relationship with them or are independent.



Data Preprocessing

The process begins with a comprehensive cleaning and preparation phase. Essential columns such as author ID, product name, product ID, and various ratings and categories are selected to be included. The dataset is then organized by author ID, ensuring a structured foundation for further analysis.

A function is developed to clean the 'ingredients' column. This involves the elimination of parenthesis, numerical data, and extraneous symbols. The function further addresses the uniformity of ingredient terminology, specifically targeting various representations of water, to standardize and simplify ingredient lists for improved analytical clarity.

The cleaned ingredients undergo TF-IDF (Term Frequency-Inverse Document Frequency) vectorization. This transformation converts the textual ingredient data into a structured numerical format, making it compatible with machine learning algorithms and enhancing the model's ability to discern patterns and similarities among products.

Categorical attributes, specifically 'skin_type' and 'secondary_category', are encoded using Label Encoding. This step transforms these categories into a numerical format, allowing the model to incorporate these variables effectively into its predictive framework.

Numerical features such as author rating, average rating, the number of reviews, and original price are normalized. This normalization process adjusts these variables to a common scale, reducing potential biases and variability that could distort the model's performance.

An interaction weight is calculated, integrating average ratings, the suitability of products for specific skin types, and a novel approach to highlight lesser-known products by inversely relating their interaction weight to the number of reviews they have received. This metric aims to balance various aspects of product relevance and user preference, enhancing the recommendation system's accuracy and user satisfaction.

Model Development & Evaluation

In our quest to enhance recommendation accuracy and relevance, we employed the LightFM model, a robust hybrid framework that adeptly merges the strengths of collaborative filtering and content-based methods. The hybrid approach is particularly beneficial as it overcomes typical limitations associated with singular recommendation strategies by not merely suggesting items popular among similar users, and it excels beyond content-based filtering in personalization by incorporating user preferences and behaviors into its recommendations. Throughout the model selection process, we conducted an exhaustive grid search to identify the optimal configuration for LightFM, as well as for two other models: k-Nearest Neighbors (k-NN) and Singular Value Decomposition (SVD).

Models & best params.	AUC	Average RMSE
LightFM `{no_components=20, epochs=10,num_threads= 4}`	0.820	-
k-NN `{'k': 20, 'sim_options': {'name': 'msd', 'user_based': False}`	-	0.282
SVD `{'lr_all': 0.01, 'n_epochs': 30, 'n_factors': 70, 'reg_all': 0.1}`	-	0.2560

These results highlight the distinct strengths and limitations of each model within the recommendation system framework. LightFM's impressive AUC score reflects its strong capacity to distinguish between items that a user will like or dislike, affirming its proficiency in delivering tailored recommendations. Conversely, while the k-NN and SVD models excel in predictive accuracy, as evidenced by their efforts to minimize RMSE, they fall short in integrating item features into their recommendation process. This approach confines them to relying solely on product ratings, potentially skewing recommendations towards popular items without considering the richness of item attributes. This could inadvertently overlook less popular but highly relevant items, underscoring the importance of feature incorporation for a more balanced and content-aware recommendation strategy.

Conclusion

Overall, the Hybrid Skincare Recommendation System leverages the strengths of collaborative filtering and content-based methods through the LightFM model. This approach ensures a personalized shopping experience by matching users with products that closely align with their unique skin needs and preferences.
Our evaluation reveals that the LightFM model, with its superior AUC score, excels in differentiating between relevant and irrelevant products for users, showcasing its effectiveness in personalizing recommendations. On the other hand, the k-NN and SVD models, while demonstrating commendable predictive accuracy through their focus on minimizing RMSE, are constrained by their reliance on product ratings alone. This limitation potentially biases recommendations towards popular items, neglecting the comprehensive benefits of considering product attributes.

Further Work

For future enhancements of the Hybrid Skincare Recommendation System, several strategic improvements and explorations could be made. Firstly, a more comprehensive incorporation of available dataset attributes, such as submission time and the helpfulness of reviews, could provide deeper insights into user preferences and product effectiveness over time.

Additionally, experimenting with an ensemble approach that combines the strengths of SVD and k-NN models may offer superior recommendation accuracy by leveraging the unique advantages of both algorithms. Further augmenting the system by integrating a broader array of features beyond ratings, including textual review analysis and product attributes, could transition the models towards a more hybrid nature. This would not only enhance personalization but also the system's ability to cater to nuanced user needs.

Developing an intuitive and user-friendly interface for the recommendation system is crucial for fostering user engagement and soliciting valuable feedback, which in turn fuels the system's learning and adaptability.

Moreover, prioritizing diversity and novelty in the recommendation algorithm can mitigate the bias towards popular products, thereby offering users a richer discovery experience and potentially uncovering hidden gems within the skincare domain.

Collectively, these initiatives aim to refine and expand the capabilities of the recommendation system, ensuring it remains at the forefront of personalization and user satisfaction in the evolving landscape of skincare retail.