Alexandria Petersen
MDA 620 Capstone

Beyond Skin Type: Predicting Skincare Satisfaction with Reviews, Ingredients, and Price

Table of contents

Background & Market Overview

The Evolving Landscape of Skincare: Market Growth and Consumer Trends

The global skincare industry is projected to experience significant expansion, growing from $122.11 billion in 2025 to $194.05 billion by 2032, with a compound annual growth rate (CAGR) of 6.84% (Fortune Business Insights, 2024). This growth reflects increased demand for wellness and self-care solutions across diverse consumer segments.

Consumer behavior is also shifting toward ingredient awareness and transparency. In a 2022 survey**,** 61% of millennial beauty shoppers in the U.S. reported seeking specific skincare ingredients when making purchasing decisions (Statista, 2022). Similarly, 40.2% of consumers prioritize natural ingredients in skincare products, according to a 2021 NielsenIQ report (ESW, 2021).

Despite this growing interest, 69% of consumers admitted to purchasing health or beauty products without fully understanding the label, highlighting a disconnect between awareness and comprehension (Home, 2022). This insight suggests that consumers may struggle to translate product labels into meaningful information about product effectiveness or compatibility with their individual skin needs.

These market trends and behavioral insights present a compelling opportunity for data-driven personalization. Leveraging machine learning and analytics can help bridge the gap between consumer interest and understanding, enabling brands to deliver more tailored and effective skincare recommendations.

The Business Problem: One Size Doesn’t Fit All

Many consumers find it difficult to identify skincare products that work well for their unique skin types. Despite the growing demand for personalized solutions, many brands still rely on generalized product formulations. This approach often leads to inconsistent results, which can frustrate customers and reduce overall satisfaction.

As a result, companies face several key issues. First, high return rates are common when products fail to meet expectations. This not only impacts profits but also signals a deeper problem with product efficacy. Second, customer trust and brand loyalty may decline when users feel a product is not suited to their needs. Third, brands are missing out on opportunities to personalize their offerings by not tailoring products to individual preferences or skin concerns. Finally, rich sources of data**,** like customer reviews and ingredient lists**,** are often underutilized, meaning valuable insights that could inform better product development and recommendations are being overlooked.

This gap between consumer needs and product performance highlights the importance of using data analytics to drive more personalized skincare experiences.

Project Objectives and Goals

The primary goal of this project is to use data to predict which skincare products are most effective for different skin types based on their ingredients. This involves several key objectives.

First, the project aims to predict product success by estimating customer satisfaction through ratings. By analyzing large volumes of review data, we can begin to understand which products are consistently rated highly and what factors contribute to that success.

Next, the project focuses on matching specific ingredients to different skin types. This helps identify which components are most beneficial for skin types such as dry, oily, normal, or combination. The goal is to move beyond generic product recommendations and toward tailored skincare guidance.

To achieve these outcomes, we build and evaluate machine learning (ML) models that can classify and predict product performance. These models help uncover patterns and relationships between ingredients, user feedback, and outcomes.

Ultimately, this approach supports the development of more personalized skincare products by providing insights into what works for whom. It also enables the creation of data-backed ingredient recommendations that can help consumers choose the right products for their skin type.

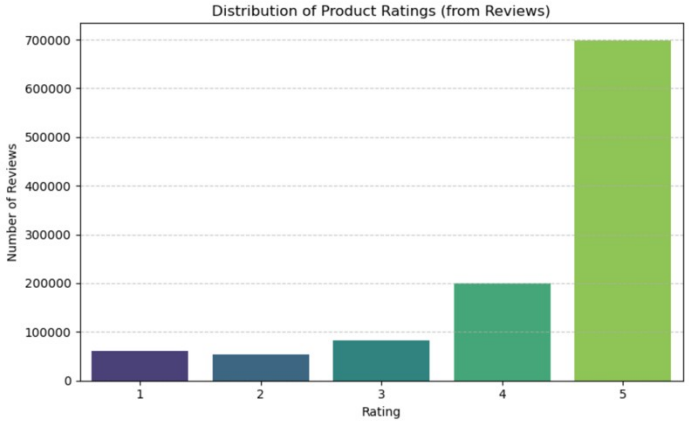
Distribution of Product Ratings

When we look at customer reviews for skincare products, we see that most people give 5-star ratings. This shows that customers are generally happy with the products. However, reviews with lower ratings (1 to 3 stars) are much less common.

This big difference suggests there may be a positive bias in the reviews—meaning people are more likely to leave good feedback than bad. While that might seem like a good thing, it can make it harder to understand what's not working in a product.

Because of this imbalance, it's important to be careful when using these reviews to train a machine learning model. If most of the data is positive, the model might have trouble learning how to detect the less common negative experiences.

To get a better understanding of what customers really think, we could also use sentiment analysis. This means analyzing the actual text of the reviews to find out how people feel about a product—not just relying on the number of stars they gave it.

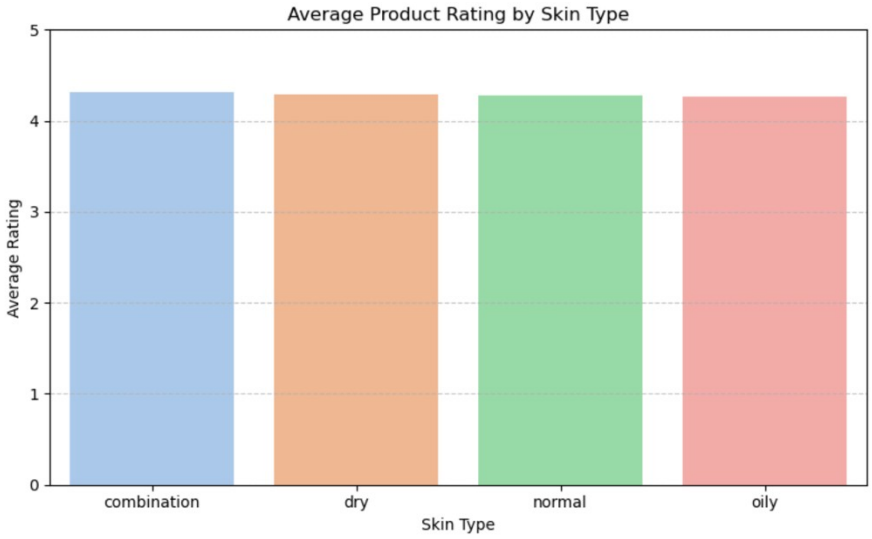


Average Product Rating by Skin Type

When we compare the average ratings of skincare products across different skin types, we find that the scores are very similar. Most products receive an average rating of around 4.3 out of 5, whether the user has dry, oily, normal, or combination skin.

This means that most products perform well across all skin types, and customer satisfaction appears to be consistent no matter what type of skin someone has. There is no big difference in how people with different skin types rate the products.

While the differences are small, they might still be helpful in future analysis—especially when looking more closely at ingredients. Understanding even slight differences could help brands create more personalized and effective products in the future.

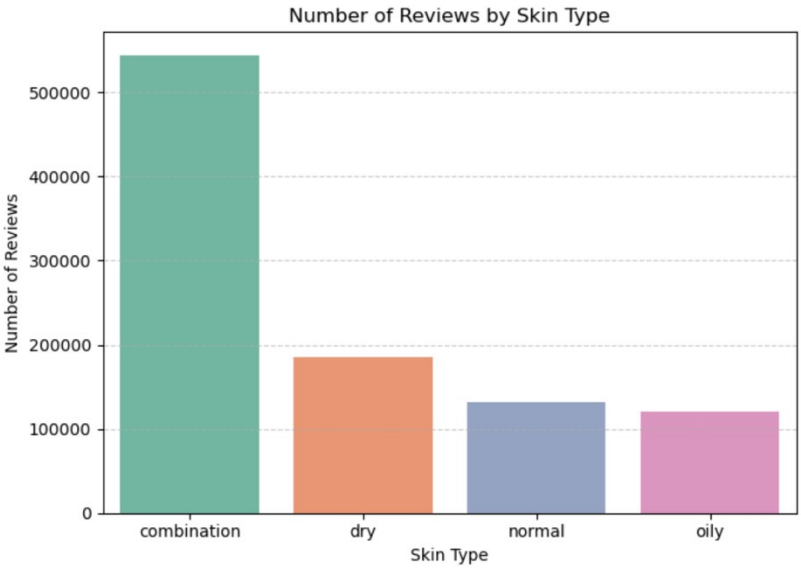


Number of Reviews by Skin Type

When we look at how many reviews were written by people with different skin types, we see a big difference. More than 500,000 reviews came from users with combination skin. In comparison, other skin types—like dry, normal, and oily—had far fewer reviews.

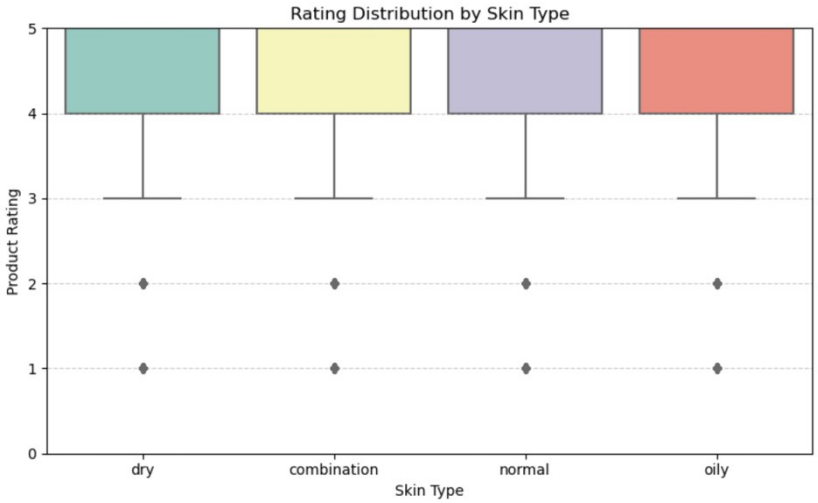
This means that combination skin is overrepresented in the data, while other skin types are underrepresented. This kind of imbalance can affect the results of our models. If most of the training data comes from one group, the model might learn patterns that don't work well for others.

To fix this, we could rebalance the data by adding more reviews for underrepresented skin types or by adjusting the way the model learns from each group. That way, we can make more fair and accurate predictions for all skin types. A more balanced dataset would also help us make better comparisons and learn what truly works best for each type of skin.



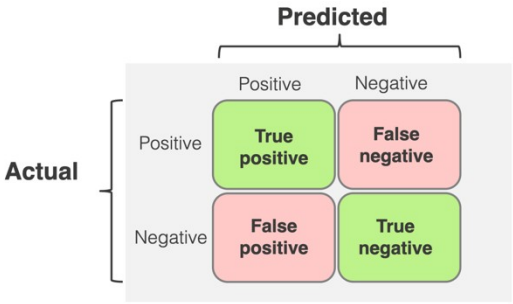
Rating Distribution by Skin Type

When we look at the distribution of product ratings for each skin type, we see that skincare products are consistently rated highly across all categories. The median rating is around 5 stars, and there are very few low-rated outliers. This means that no matter the user's skin type—whether it's dry, oily, combination, or normal—most people are giving similar high ratings. This limited variation suggests that skin type by itself may not strongly influence customer satisfaction. Instead, it points to the importance of exploring other features, such as specific ingredients, that could have a greater impact. These results provide a strong starting point for deeper modeling efforts, especially those that look at how ingredients interact with different skin types.



Logistic Regression Model: Predicting Positive Ratings

The logistic regression model achieved an overall accuracy of 81.99%, which might seem impressive at first. It means that over 81% of the predictions matched the actual outcomes. However, a closer look at the confusion matrix revealed a major issue—the model only predicted positive ratings. It failed to recognize any negative reviews, meaning it was only learning from the dominant class in the dataset. This highlights a class imbalance problem, where there were far more positive reviews than negative ones. As a result, the model learned to assume that nearly all reviews would be positive. This case shows that high accuracy can be misleading, especially when the dataset is heavily skewed toward one outcome. A more balanced dataset or different evaluation metrics (like precision, recall, or F1 score) would give a clearer picture of true model performance.



Logistic Regression Model: How Can We Improve It?

After balancing the dataset to include an equal number of positive and negative reviews, the model's accuracy dropped to 51.1%. While this is lower than the previous result, it is more realistic and meaningful. In a balanced dataset with a 50/50 class split, an accuracy near 50% is expected if the model is still learning and hasn’t yet picked up strong predictive patterns.

According to the confusion matrix, the model correctly identified around 20,000 true positives (positive reviews predicted correctly as positive) and about 20,000 true negatives (negative reviews predicted correctly as negative). However, it also made a lot of mistakes—misclassifying nearly 39,000 reviews in total. These errors include both false positives (predicting a review is good when it's not) and false negatives (missing a good review).

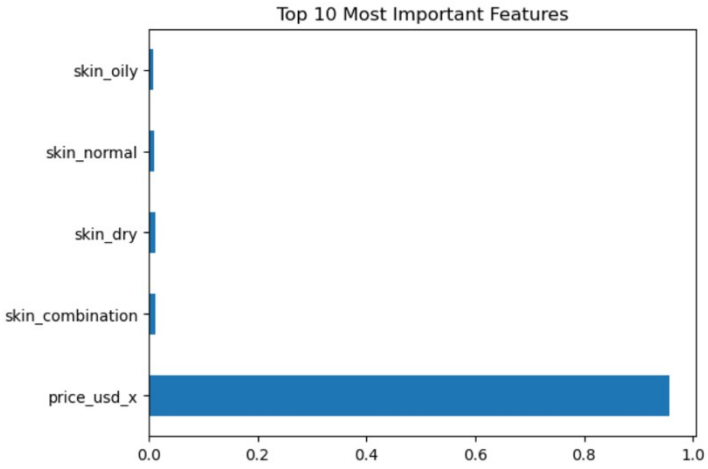
This result suggests that while the model is starting to learn from both classes, it needs improvement. Possible next steps include trying more advanced models (like Random Forest or XGBoost), fine-tuning features, or incorporating text-based sentiment from the reviews.

```
Accuracy: 0.5113669688508078
Confusion Matrix:
 [[19992 19347]
 [18976 20114]]
```

What Drives Positive Skincare Ratings?

One of the most important findings from this analysis is that product price (price_usd_x) is the strongest predictor of whether a skincare product receives a high customer rating. According to the feature importance chart from the model, price had a significantly larger impact than any other feature, including skin type (oily, dry, normal, combination). This suggests that consumers may associate higher prices with better quality, regardless of whether the product is tailored to their skin type.

This insight has a few important implications. First, brands may benefit from emphasizing value and pricing in their marketing strategies. Second, it shows that personalization strategies need to go deeper than just matching skin types. To truly improve satisfaction and product effectiveness, companies should consider analyzing ingredients, product categories, and individual review feedback to offer smarter recommendations.



Recommended Products for Dry Skin

To create a data-driven recommendation list, products were filtered to include only those with 50 or more reviews and listed for users with dry skin. This ensured that only reliable and well-supported products were considered. From this filtering, several top-rated products emerged.

Examples of highly rated skincare items include Barrier+ Triple Lipid + Collagen and Super Rich Repair Moisturizer, both with an average rating of 4.89 stars. Another top performer is the Evercalm Barrier Support Face Oil, which has a rating of 4.88. These products not only received strong reviews but also fell into a reasonable price range—most cost between $40 and $70, showing that high value does not always mean high cost.

These results show how brands can use simple filtering techniques to offer smarter, more targeted recommendations for consumers based on their skin type. By combining review count, rating, and price, companies can guide customers toward effective, trusted products.

	product_name_x	avg_rating	review_count	avg_price
0	Barrier+ Triple Lipid + Collagen + Niacinamide...	4.888889	117	69.0
1	Super Rich Repair Moisturizer	4.888889	54	94.0
2	Evercalm Barrier Support Face Oil	4.884615	52	60.0
3	Luxury Sun Ritual Pore Smoothing Sunscreen SPF 30	4.881356	59	38.0
4	Ultralight Moisture-Boosting Botanical Oil	4.873239	71	44.0
5	Truth Barrier Booster Orange Ferment Vitamin C...	4.866071	112	48.0
6	Silk Rice Makeup-Removing Cleansing Oil	4.858974	78	46.0
7	Rénergie Lift Multi-Action Ultra Dark Spot Cor...	4.850000	60	135.0
8	Daily Milkfoliant Exfoliator	4.847328	131	65.0
9	Juneberry & Collagen Hydrating Cold Cream Clea...	4.843137	51	39.0

Top-Rated Products for Dry Skin: Data-Driven Picks

Based on the analysis, the best skincare products for dry skin tend to include hydrating ingredients such as hyaluronic acid, niacinamide, glycerin, and squalane. These ingredients are known to support moisture retention and skin barrier repair—key needs for people with dry skin.

Most top-rated products for this skin type fall within a $38 to $69 price range, showing that effective skincare doesn't have to be extremely expensive. The product selection was guided by clear criteria: a high average rating, a minimum of 50 reviews, and the presence of key ingredients.

Across the top recommendations, the most common ingredients included Hyaluronic Acid, Niacinamide, Squalane, Oat Extract, Jojoba Oil, and Rice Bran Oil. These results empower users with dry skin to make confident, affordable, and evidence-backed skincare choices, using data instead of guesswork.

Model Building & Methodology

To predict whether a skincare product would receive a positive rating (defined as 4 stars or higher), a Random Forest model was trained using features such as product price and skin type. The model's feature importance chart showed which inputs had the greatest influence on prediction results.

The most important feature by far was product price (price_usd_x), which dominated all other factors. In contrast, skin type features (e.g., dry, oily, normal, combination) had very little impact on the model's decisions. This suggests that consumers may be more influenced by a product’s cost—possibly associating a higher price with better quality—than by how well the product is tailored to their skin type.

While the model performed reasonably well, these findings highlight opportunities for improvement. Including more detailed information such as ingredient content or

performing sentiment analysis on customer reviews could make future models more accurate and personalized.

Key Findings

This analysis uncovered several important insights about skincare product reviews. First, price was the most influential factor in predicting whether a product received a positive rating. Consumers appeared to associate higher price with higher quality, even more than with personalized product features.

Interestingly, skin type had very little impact on prediction accuracy. Although features like oily, dry, or normal skin were included in the model, they didn't significantly affect the outcome. This shows that skin type alone does not determine user satisfaction and highlights the need to explore additional factors.

The study found that ingredients may be the missing link in personalization. When products were filtered for effectiveness by skin type (like dry skin), certain ingredients such as hyaluronic acid, squalane, niacinamide, and oat extract repeatedly showed up in top-rated products. This suggests that ingredients play a stronger role in satisfaction than skin type labels.

Lastly, the dataset showed a clear imbalance, with most reviews being 4 or 5 stars. This skew toward positive ratings impacted model fairness and performance. Addressing this required undersampling and using better evaluation methods like the confusion matrix to get a more realistic understanding of model behavior.

References:

- Nadyinsky, N. (2022). *Sephora products and skincare reviews* [Data set]. Kaggle. https://www.kaggle.com/datasets/nadyinky/sephora-products-and-skincare-reviews/data
- ESW. (2021). *NielsenIQ reveals 40.2% of consumers prioritize natural ingredients*. Retrieved from https://esw.com
- Fortune Business Insights. (2024). *Skincare market size, share & COVID-19 impact analysis*. Retrieved from https://www.fortunebusinessinsights.com
- Home. (2022). *Health and beauty product label confusion survey*. Retrieved from https://home.com
- Statista. (2022). *Millennial shopping behavior in beauty*. Retrieved from https://www.statista.com