

Skincare Product Analysis

August 4, 2021

A more detailed project documentation

Motivation for the project

With the explosion of skincare brands around the world, the consumer is faced with an ever increasing amount of choices that are not just diverse in terms of product type or category (i.e. cleansers, toners, essences, moisturizers, serums, etc.) but also in ingredient composition and formulation type. Information overload can be frustrating for a consumer and a problem for manufacturers who now have to exert more effort in converting purchases.

Adding to information paralysis, there are costs increasing the barriers to a purchase. It costs time to research the differences between a myriad of products and ingredients. Trial and error is also costly both in terms of time and money. Skincare products are not cheap and it takes time to see results. Hence, a consumer can effectively sample only a handful of products in a year. Besides, there is also a risk that the product may cause adverse reactions which is a major concern especially for consumers with sensitive skin.

With all these hindrances to a product purchase, how can we make the decision just a step easier for the consumer? One of the approaches is to help them understand whether the product is over- or under-priced relative to its competitors, without having to do endless product comparisons. After all, one of the first questions a consumer asks is “Is this worth the money?”

If so, consumers may also want to understand what makes the product cheap or expensive. If buying a cheaper product, what is the trade-off they are making? Is it the lack of the velvety feel or perhaps the use of some ingredients that are not as friendly to sensitive skin? On the other hand, if one has decided to go for a more expensive product, what were they paying for - the brand, a more expensive ingredient, a richer formulation? The project will aim to uncover which product characteristics make a product likely to be cheap, average or expensive. In this way, we can deliver some insights that a consumer can use to make a more informed decision about what they could be paying for.

And if the product is not worth the tag price, can we help a consumer find the next closest product? Is there a 'dupe' for their favorite serum? We can use machine learning to analyse ingredients similarity at scale and deliver recommendations for their next purchase.

Methodology

Data gathering and collection









As the project requires extensive product coverage and detailed information at the product level, e-commerce websites are, by far, the richest data sources and remain unmatched by readily available datasets online. For this project, I have chosen to tap into one of the largest international beauty retailers, Sephora, as the main data source. Sephora sells thousands of products from brands originating from all around the world and across many product types. Its product pages also contain rich product information, including but not limited to number of reviews, likes, ratings, ingredients list, active ingredients, awards received, suitable skin types, skin concerns addressed, and clinical results published. Being a key player in the beauty industry, it also covers the most popular skincare products across geographies.

Data was collected using web scraping with Python, Scrapy and Selenium (to deal with Javascript/dynamic pages and 'lazy load'), creating a product database for over 1,800 products across four main categories (cleansers, eye treatments, moisturizers, treatments).

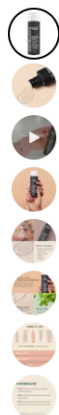
Below is a screenshot of the text fields extracted from the main and product pages.


430 products

Sort by: **Bestselling** ▾

 <p>product name and brand</p> <p>Paula's Choice Skin Perfecting 2% BHA Liquid Exfoliant</p> <p>\$10.00 - \$29.50</p> <p>★★★★★ 433</p> <p>rating number of reviews</p>	 <p>Youth To The People Superfood Antioxidant Cleanser</p> <p>\$12.00 - \$64.00</p> <p>★★★★☆ 4.2K</p>	 <p>only at sephora</p> <p>Farmacy Green Clean Makeup Removing Cleansing Balm</p> <p>\$22.00 - \$34.00</p> <p>★★★★★ 5K</p>	 <p>The Ordinary Glycolic Acid 7% Toning Solution</p> <p>\$8.70</p> <p>★★★★☆ 2.2K</p>
 <p>limited edition</p> <p>fresh Soy Makeup Removing Face Wash</p> <p>\$15.00 - \$69.00</p> <p>★★★★☆ 8.4K</p>	 <p>SK-II Facial Treatment Essence (Pitera Essence)</p> <p>\$99.00 - \$235.00 (\$214.00 value)</p> <p>★★★★☆ 2.7K</p>	 <p>Tatcha The Deep Cleanse Gentle Exfoliating Cleanser</p> <p>\$16.00 - \$62.00</p> <p>★★★★☆ 2.2K</p>	 <p>only at sephora</p> <p>Glow Recipe Watermelon Glow PHA + BHA Pore-Tight Toner</p> <p>\$15.00 - \$34.00</p> <p>★★★★☆ 2.3K</p>

Skincare > Cleansers > Exfoliators





Paula's Choice
Skin Perfecting 2% BHA Liquid Exfoliant

★★★★★ 433 Ask a question | ❤️ 68.4K

price → **\$29.50** or 4 interest-free payments of \$7.38 **Klarna.** ⓘ

size → **Size: 4 oz/ 118 mL**

Standard size
4 oz/ 118 mL

Mini size
1 oz/ 30 mL

☒ **Get It Shipped** You're only \$50.00 away from Free Shipping. [Shipping & Returns](#) 🚚

☐ **Buy Online & Pick Up** ⓘ Select to see availability at [stores near you](#) 🏪


Add to Basket ❤️

Highlights

any mention of special ingredients

 Salicylic Acid

skincare concerns the product is good for

 Good for: Acne/Blemishes

 Good for: Dullness/Uneven Texture

About the Product

[Back to Top](#)

Item 2421360

skin types the
product is suitable for

skincare concerns
addressed

formulation

highlighted ingredients

whether clinical results
are published or not

What it is: A daily leave-on exfoliant with two percent salicylic acid to sweep away dead skin cells, unclog pores, and visibly smooth wrinkles—practically overnight.

Skin Type: Normal, Dry, Combination, and Oily

Skincare Concerns: Pores, Uneven Texture, and Acne and Blemishes

Formulation: Liquid

Highlighted Ingredients:

- Salicylic Acid (BHA) 2%: Penetrates pores to clear blemish-causing buildup and shed dead skin.
- Green Tea: A potent antioxidant that soothes irritated skin while improving visible signs of aging.
- Methylpropanediol: Hydrates while visibly enhancing the skin's glow and the formula's penetration.

Ingredient Callouts: Free of parabens, formaldehydes, formaldehyde-releasing agents, phthalates, mineral oil, retinyl palmitate, oxybenzone, coal tar, hydroquinone, sulfates SLS & SLES, triclocarban, triclosan, and contains less than one percent synthetic fragrance. It is also gluten-free, cruelty-free, and comes in recyclable packaging.

What Else You Need to Know: With a two percent concentration of salicylic acid and a pH range of 3.2 to 3.8, this global bestseller breaks down complexion-dulling buildup on skin's surface and within pores. It leaves skin feeling dramatically smoother and looking clearer and more radiant than before. People notice a natural, visible glow after just one use.

Clinical Results: In an independent consumer panel:

- 91% of users had visibly healthier skin
- 90% saw improved skin texture
- 82% experienced smaller-looking pores

[Show less](#)

Ingredients

active ingredients

full list of ingredients

- Salicylic Acid (BHA) 2%: Penetrates pores to clear blemish-causing buildup and shed dead skin.
- Green Tea: A potent antioxidant that soothes irritated skin while improving visible signs of aging.
- Methylpropanediol: Hydrates while visibly enhancing the skin's glow and the formula's penetration.

Water (Aqua), Methylpropanediol, Butylene Glycol, Salicylic Acid, Polysorbate 20, Camellia Oleifera (Green Tea) Leaf Extract, Sodium Hydroxide, Tetrasodium EDTA.

Wrangling and pre-processing

Data cleaning and preparation is required especially since the information returned through the web-scraping process was unstructured text and not fully standardised. The main steps taken were as follows:

1. Inspected for missing values and deleted rows where appropriate
 - a. Most of the rows with no associated 'size' are related to masks, supplements, devices, peels, make-up removers, gift sets, eye masks, eye gels, massager/rollers, etc. so these can be deleted as they are not the main focus of the study

- b. Deleted rows with no price or no reviews, reducing the number of products in the dataset
 - i. Cleansers: 409 to 363 products
 - ii. Treatments: 578 to 498 products
 - iii. Eye products: 220 to 190 products
 - iv. Moisturizers: 645 to 498 products
2. Text pre-processing to standardize the format of information before converting to the correct data type
 - Removed 'stars' from rating and converted column into numeric type
 - Remove currency sign from price and converted into numeric type
 - Translated both number of likes and number of reviews into digits and converted into numeric data (i.e. 68k to 68,000)
3. From initial inspection, some entries had to be manually corrected.

Feature engineering

Majority of the product details were contained in the 'About the Product', 'Ingredients' and 'Highlights' section of the webpage and were retrieved as full text. In order to extract information from this and generate new features, regular expressions were extensively used to detect and list out information such as:

- skin concerns addressed (i.e. dark circles, uneven skin tone, acne, aging, dullness, etc.)
- any excluded ingredients mentioned
- skin types the product is suitable for
- clinical results published
- product formulation
- active or highlighted ingredients
- what it is good for (as noted by Sephora)
- any specific acids being used (i.e. AHA, BHA, glycolic, ascorbic)
- product size

In several cases, it was not a simple binary classification (i.e. whether clinical results were published) or a mutually-exclusive category such that we can proceed directly to one-hot encoding. There could be numerous labels for a feature. For example, a product could target more than one skin concern (i.e. acne and uneven skin tone) and be suitable for a few skin types (i.e. normal, dry and sensitive). In this case, I first checked how many unique labels there were for that feature. If there were too many categories, I narrowed this down to the few main ones by either combining various labels under a larger category or by using 'Other' to represent the minority labels.

1. Created new features based on skin type the product was compatible with. This was derived from the 'about the product section'. (New columns created containing 1 or 0 for each of normal, combination, oily, dry, sensitive)
2. Created new feature based on highlights

As there were numerous unique features based on highlights, several items have been merged under more general categories as the level of the detail is not necessary and this would also prevent creating too many features. The features from this step where:

1) Good for:

- Acne/Blemishes
- Anti-Aging
- Dark Circles or Dark spots
- Dryness
- Dullness/Uneven Texture
- Loss of firmness
- Pores
- Redness
- Hydrating (to be added separately)

2) Ingredients excluded (i.e. Formaldehydes, Mineral Oil, Parabens, Phthalates, Retinyl Palmitate, Silicones, Sulfates SLS & SLES)

- In order to capture the extent to which the product excludes toxic ingredients, I used a count of the number of excluded ingredients instead.

3) Best for:

- Combination
- Dry
- Combo
- Normal
- Oily

4) Clean at Sephora

5) Cruelty-Free

6) Vegan

7) Skincare acids

- Hyaluronic Acid
- Salicylic Acid
- AHA/Glycolic Acid
- Vitamin C

8) Awards

- Community favorite
 - any type of allure award
- 3. Create a new feature capturing whether clinical results are published in the about section
- 4. Extract formulation type for each ingredient and capture this in a separate feature(s)
 - a. Information includes the lightness or richness of the formula as well as the formulation (i.e. oil, serum, lotion, etc.) so two features will be created.
 - i. richness - light, normal, rich/heavy
 - ii. formulation - lotion, cream, serum, liquid, gel, oil, others
- 5. Reduced the subcategory names (i.e. face serums, mists and essences, night creams, eye primer) to a few:
 - a. Exfoliators and peels
 - b. Eye creams and treatments
 - c. Essences, serums and treatments
 - d. Face wash and cleansers
 - e. Toners
 - f. Moisturizers and creams

The ingredients also required some simplification as there were variations for the same ingredient (i.e. Aqua/Water/Eau, Aqua/Eau, Water/Eau, Aqua/Water, Water), leading to an unnecessarily larger dataset comprising c. 6,000 unique ingredients. This was reduced by half through FuzzyMatch similarity.

An important consideration for the project was to use price per volume instead of price alone as price is expected to vary directly with the amount of the product. Referring back to the original problem, the objective is to learn insights about what differentiates a premium from a cheap or an average product. Hence, predicting the exact price of the product is not as important as determining its market positioning so the approach I've taken was to treat this as a classification rather than a price prediction problem. The target variable (price affordability - cheap, average, expensive) was derived by obtaining the dollar price by the size (used oz) and creating the cheap, average and expensive categories by getting the terciles of the price per volume distribution.

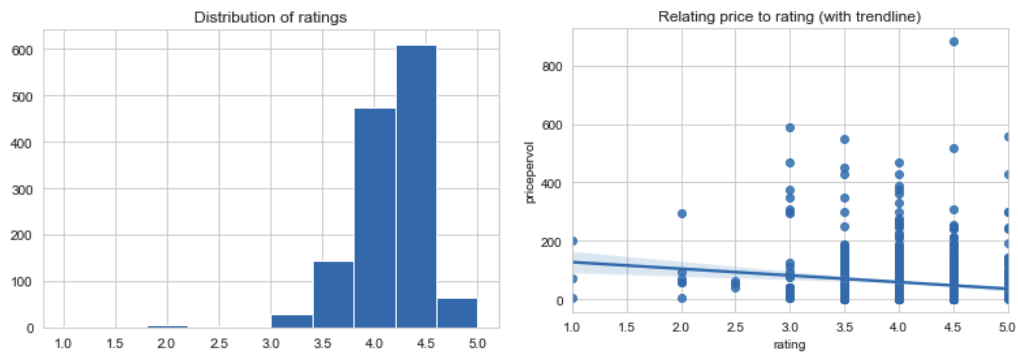
1. Create a new (uniform) feature for size or volume
 - Product information was typically in the ff. format: '## oz/ ## mL', '## g/ ## mL'
 - Using regex, I extracted the available information whether this was in oz, mL and g. Since most had the sizes in oz, I used this as the unit for size and converted mL or g

- brand
- product_type: category of product (i.e. essences, serums and treatments, moisturizers and creams, face wash and cleansers, eye creams and treatments, toners, exfoliators and peels)
- num_likes: number of likes
- rating: rating from 1-5 stars
- num_reviews: number of reviews received
- sensitive_type: whether product is suitable for sensitive skin
- combination_type: whether product is suitable for combination skin
- oily_type: whether product is suitable for oily skin
- normal_type: whether product is suitable for normal skin
- dry_type: whether product is suitable for dry skin
- clean_sephora: clean if the product does not contain the sulfates SLS and SLES, parabens, formaldehydes and formaldehyde-releasing agents, phthalates, mineral oil, retinylpalmitate, oxybenzone, coal tar, hydroquinone, triclosan, and triclocarban
- cruelty_free: whether or not it was not tested on animals
- vegan
- Acne/Blemishes, Anti-Aging, Dark Circles, Dark spots, Dryness, Dullness/Uneven Texture, Hair Dryness, Hydrating, Loss of firmness, Pores, Redness: whether or not the product targets these skin concerns
- num_excl_ingr: number of specific ingredients not used in the formulation of the product
- best_for_skin_type: whether it is highlighted to be best for a specific skin type
- Hyaluronic Acid, Salicylic Acid, AHA/Glycolic Acid, Vitamin C: whether or not the product contains these skincare acids
- award: number of awards won
- affordability: cheap, average or expensive according to \$ price per oz.
- clinical_results: whether clinical results have been noted in the 'about the product' section in Sephora
- formulation_type: whether it is a cream, serum, liquid, gel, oil, lotion or other
- richness: indicates whether the product is lightweight, normal or heavy
- various ingredients used: whether the product uses the ingredient or not

Exploratory Data Analysis

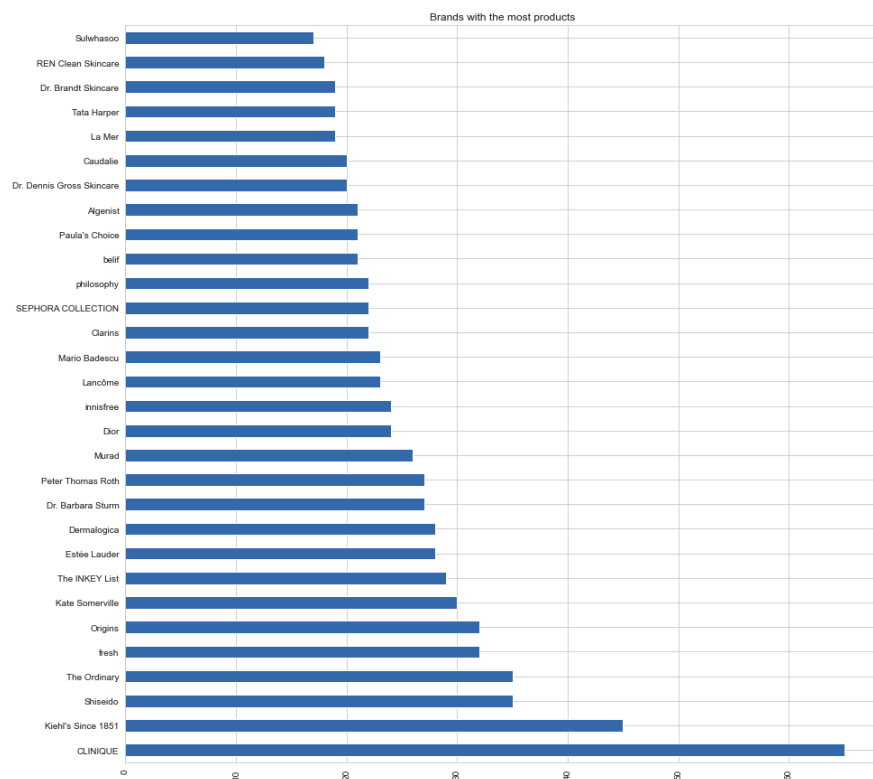
Selected Insights

- 1) **Rating:** c. 80% of the products have a rating of 4-4.5; slight negative relationship between price and rating.

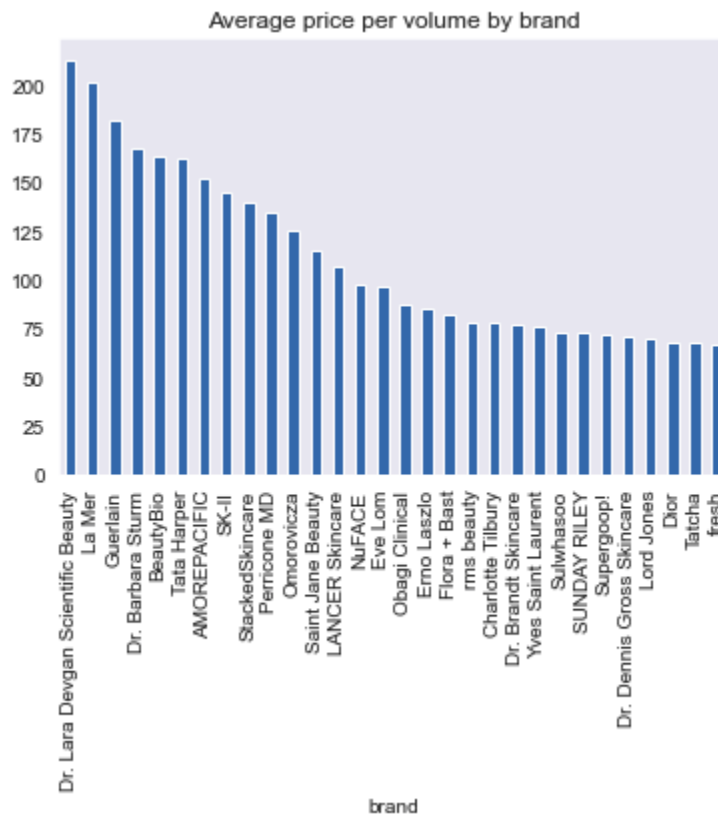


2) Brands

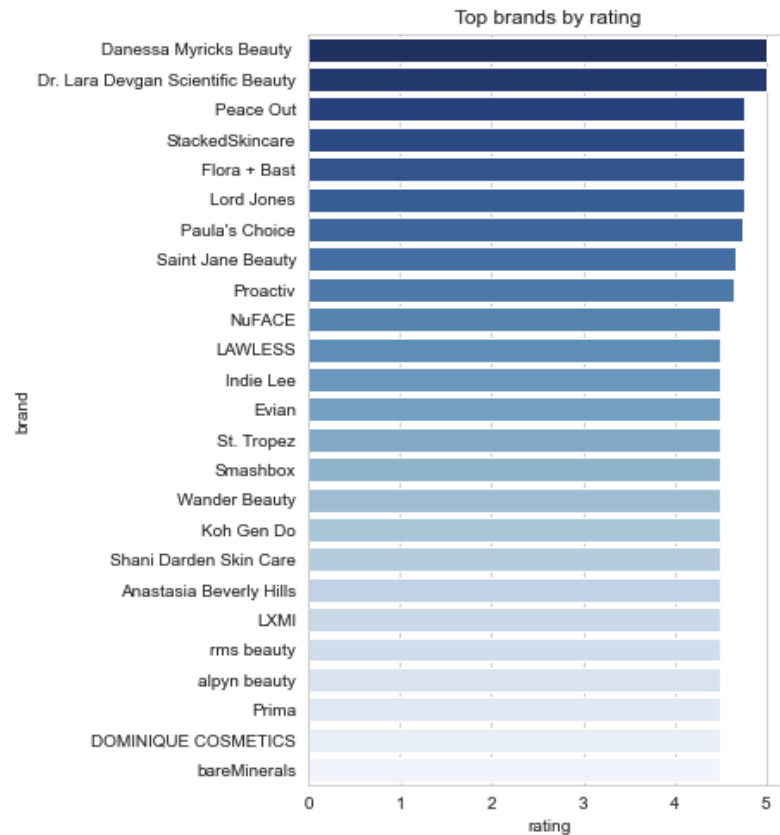
a) Which brands offer the most products? Clinique, Kiehls, The Ordinary, Shiseido



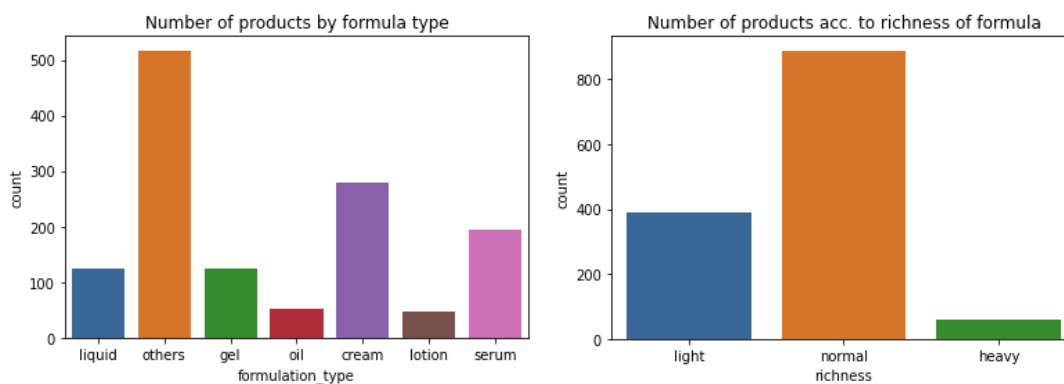
b) What brands are the most expensive? Dr. Lara, La Mer, Guerlain, Dr. Barbara Sturm



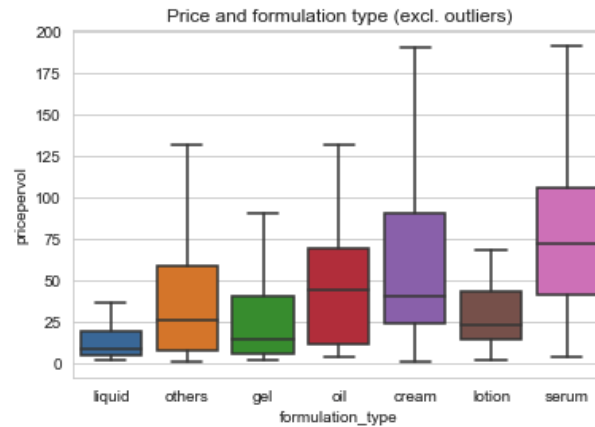
- c) Which brands have the highest rating? Danessa, Dr. Lara, PeaceOut, Stacked Skincare



- 3) **Awards:** Only 9% of the products have been awarded, indicating a selective handpicking of top products
- 4) **Clinical results:** A fair proportion have published clinical results (39%), possibly to provide more credibility to the product or brand
- 5) **Formulation type:**
 - Majority of products are formulated as creams and serums. There are also less products with a heavy formula.

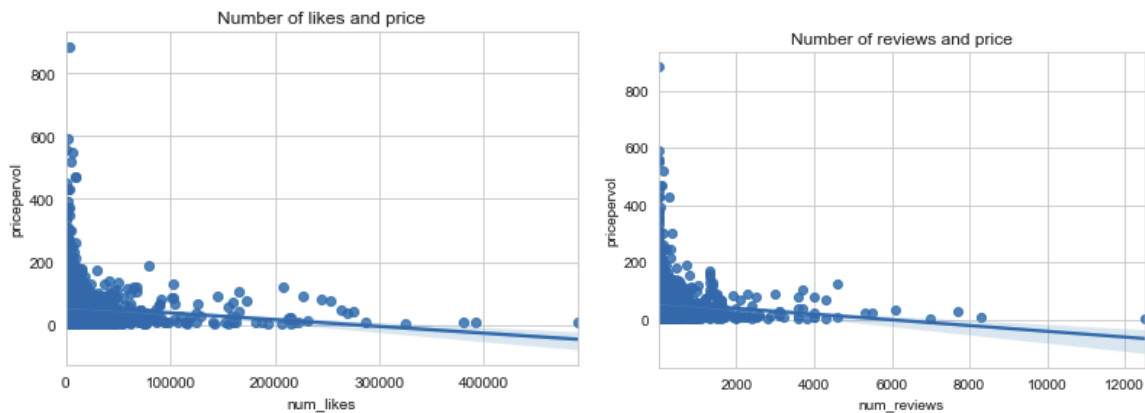


- How does price vary depending on the formulation type?



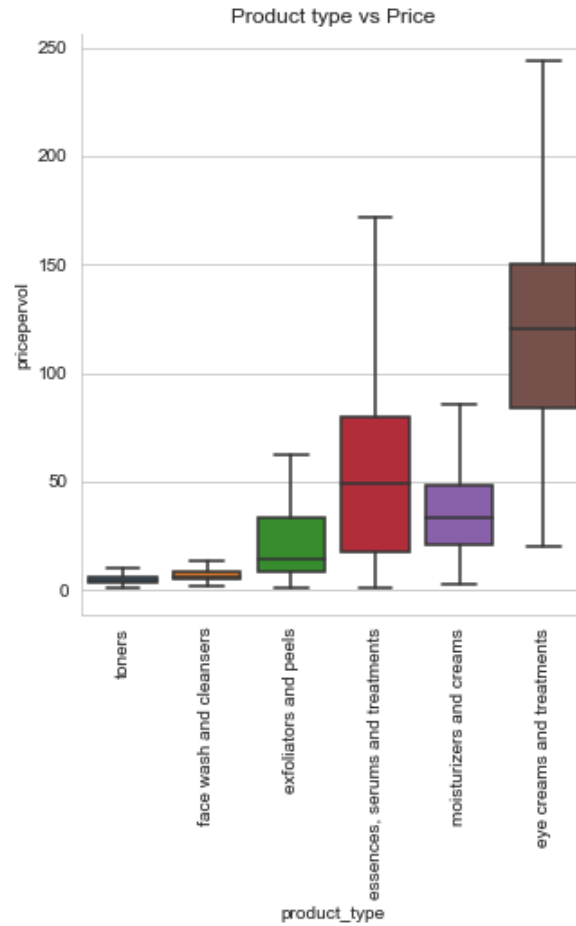
6) Popularity

- Is there a relationship between popularity (num_likes, num_reviews) and the price or rating? No conclusive linear relationship.



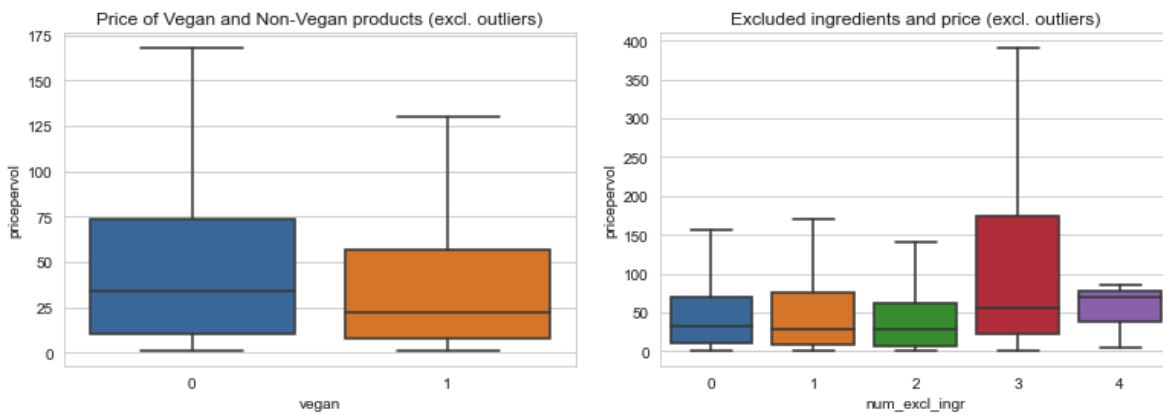
7) Product category

- Price tends to vary with the product category. Eye creams are typically expensive whilst toners and cleansers are consistently cheaper.



8) Other

- Vegan products, contrary to the initial hypothesis, do not necessarily cost more than non-vegan products. In fact, the average price for products marked as vegan are slightly lower.
- Where the product excludes only 1-2 less desirable ingredients, there is no marked difference in price but there is a slight premium to products excluding at least 3 of these.



- award also highly correlated to both number of reviews and likes, indicating popularity
- high correlation for several skin types with products for oily and combination skin being most related
- some correlation between:
 - o Salicylic acid and acne blemishes and to some extent, pores
 - o Loss of firmness and anti-aging
 - o Dark spots, vitamin c although low
 - o Dryness with hyaluronic acid
 - o Dryness, loss of firmness
 - o Dullness/uneven texture, vitamins c, aha
 - o Pores, salicylic acid and AHA
- Those targeting acne/blemishes tend to have a higher rating and are less expensive than average too.
- Skincare products that target dark circles tend to have a higher rating than average.

Modelling and approach

The dataset was split 70/30 between training and test sets before modelling.

Various machine learning techniques were explored in the project, including Logistic Regression, Support Vector Machines (SVM), Decision Trees, Random Forest, XGBoost and Gradient Boosting.

For each of the techniques, the methodology was as follows:

1. Trained the base model (using default parameters)
2. Tune the parameters to check for performance improvement (used nested cross-validation)
3. Derive insights to understand how it is predicting affordability and what features are important

- I have experimented using machine learning techniques (SVM, logistic regression, Random Forest, PCA) for feature selection. This proved effective as model performance was either maintained or improved despite using a smaller dataset (i.e. 108/175 features using SVM for feature selection and 125/175 features using Logistic Regression).

The initial project did not consider the ingredients as a feature to the model. However, upon some deliberation, I extended the project to see if the use of certain ingredients could be more predictive of pricing or whether this will only introduce noise. The model results for the main and extended project will be presented.

Model performance

Cross validated performance, based on accuracy, precision, recall and f1 score, was compared across the models. Generally, the model accuracy has ranged between between 70-80%.

Here are the cross-validated performance results for the models (without using ingredients).

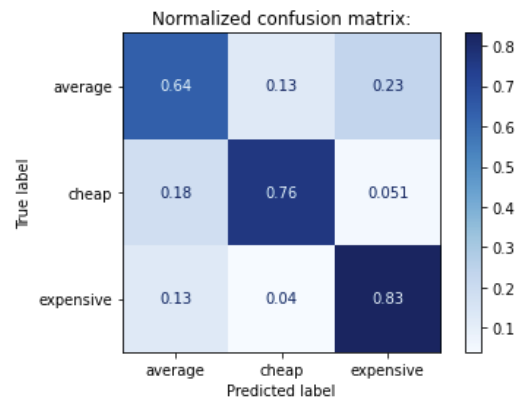
Description	Accuracy (CV test) mean	Accuracy (CV test) std	Accuracy (CV train) mean	Accuracy (CV train) std	Recall (CV test) mean	Precision (weighted, CV) mean	F1 (weighted, CV) mean
SVM (linear) - scaled training set, selected f...	0.766066	0.034404	0.829062	0.007109	0.766066	0.770590	0.766034
SVM - scaled training set, linear kernel	0.754318	0.035590	0.842418	0.011534	0.754318	0.757986	0.754644
SVM - scaled training set, selected features u...	0.752173	0.040834	0.825858	0.008697	0.752173	0.757186	0.752488
LR - scaled SVM feature set	0.750063	0.040442	0.817576	0.007022	0.750063	0.754237	0.749815
LR - scaled and selected (based on LR chosen f...	0.745796	0.046211	0.816775	0.003005	0.745796	0.748574	0.744658
LR - scaled training set, base (saga)	0.738304	0.035777	0.831465	0.009348	0.738304	0.740529	0.736470
SVM - PCA features, linear	0.735078	0.027628	0.790338	0.013018	0.735078	0.739356	0.734686
LR - PCA scaled training set, base model	0.721208	0.036772	0.786326	0.006776	0.721208	0.723839	0.719981
RF - unscaled training set, tuned parameters	0.721186	0.035091	0.887824	0.007065	0.721186	0.723094	0.719929
RF - unscaled training set, tuned parameters	0.719064	0.029854	0.999733	0.000534	0.719064	0.724041	0.718977

The best model was then tested on the test set:

Without ingredients: Support vector machine (SVM) using a linear kernel trained on 108 features only

- Accuracy score on test: 74.1%
- Average ROC AUC score: 80.8% (cheap - 83.9%, expensive - 84.4%, average - 74.1%)
- F1 score: 74%

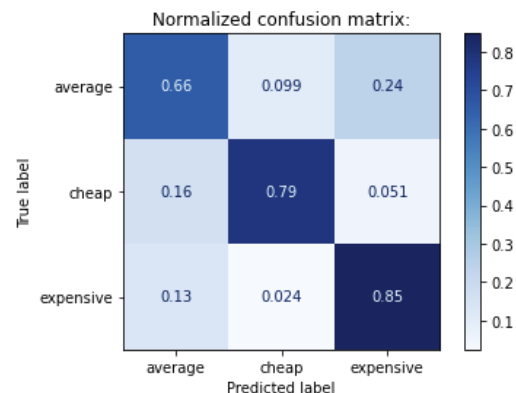
- Matthews correlation coefficient: 61.4%



- Most important features found are:
 - Product type: coherent with initial findings during EDA
 - Numerous brands
 - Lotion formulation, Salicylic Acid, sensitive skin type, others
- Whilst some factors like number of reviews, number of likes and number of excluded ingredients had little to no sway on predicting product affordability, contrary to the initial hypothesis.
- Patterns in the highly and least predictive features have some overlap with what was found through logistic regression with top predictive features consisting mostly of brands and product type.
- In the SVM, 55 features had no predictive power and the majority of these were also brands. Another observation is that this is close to the number of optimal features based on logistic regression models.
- Number of reviews and ratings have low predictive power in this model.

(extended project scope) **With ingredients**: XGBoost trained on the full feature set

- Accuracy score on test: 76.1%
- Average ROC AUC score: 82.3% (cheap - 86.1%, expensive - 85.0%, average - 75.7%)
- F1 score: 76%



With the XGBoost model, number of likes and reviews play a significant role in affordability. It also considers product type, rating, formulation, publishing of clinical results, selected ingredients and brands as useful.

Here are some of the important features for the XGBoost model:

- Brands: The Ordinary, The Inkey List, Sephora collection, Clinique
- Ingredients: glyceryl laureate, dimethicone, capric triglycerides, niacinamide, vitamin C, linalool, hexanediol, soybean oil, ci titanium dioxide, etc.
- Skincare concerns: Anti-aging, loss of firmness, Dryness

The 2% uplift in accuracy confirms the initial hypothesis that the use of certain ingredients are predictive of pricing.

The confusion matrix reveals that not surprisingly, the model has better predictive power for the cheap and expensive categories, compared to the average.

It was also interesting to see how other models were making predictions. For example, while the decision tree classifier had lower accuracy, the simplicity and intuitiveness of the insights found are worth noting. This model makes a classification mainly using the product type and brand. It also considers other features such as number of likes, number of reviews, formulation type, and rating at the latter part of the decision tree. Some insights found are:

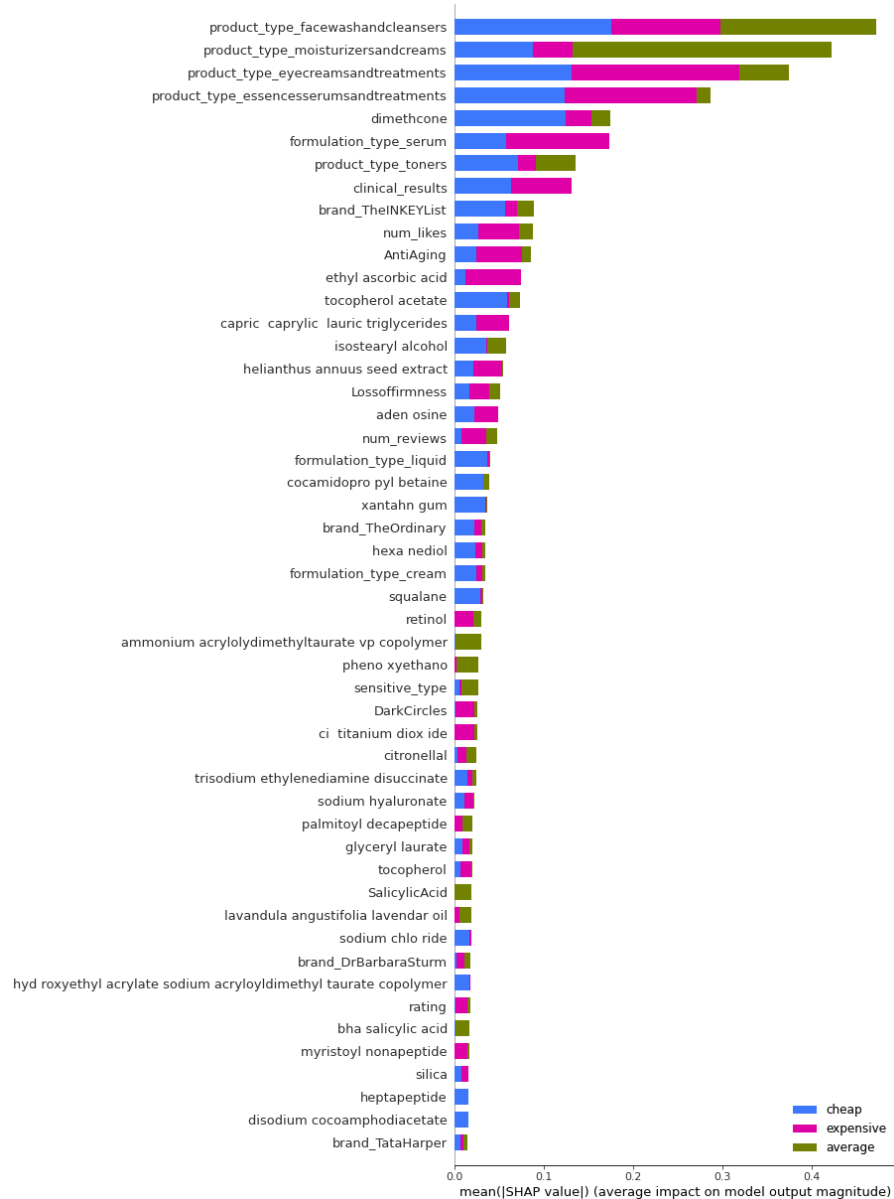
- Eye creams are usually expensive. However, we do find averagely priced products too (typically from Laneige and Mario Badescu).
- Moisturizers and creams fall under the average price category except if it is from Dr. Barbara Sturm, La Mer and Charlotte Tilbury, in which case, they would be expensive! You can find cheap ones at Sephora though.
- Face wash / Cleansers and toners are affordable with prices in the cheap-average range. Toners at Milk Makeup or Mario Badescu tend to be more expensive than its competitors. For face wash, we can expect higher price points from brands like La Tata Harper, La Mer, Eve Lom and Dr. Barbara Sturm.
- Generally, you can expect low prices from The Ordinary serums.
- Serums are typically expensive except if you get them from The InkeyList which you can expect to be cheap.
- Face wash and cleansers are usually cheap or priced averagely.

Conclusion

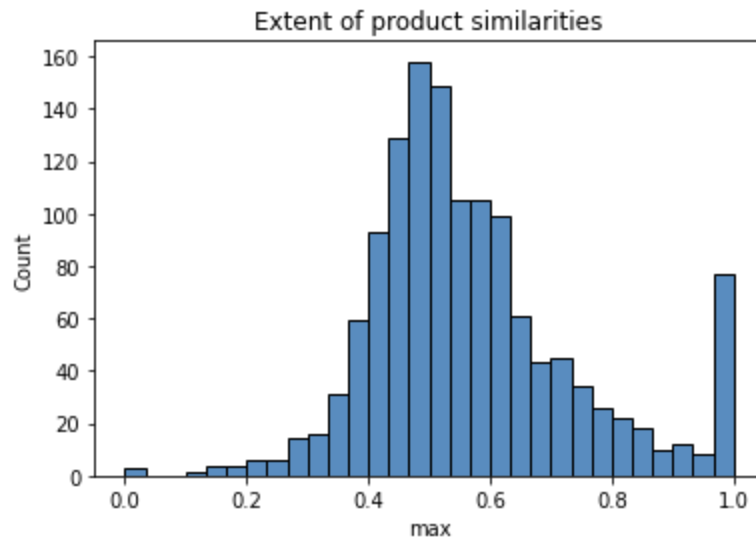
Overall, product type tends to be the most important predictor of pricing, consistently found as the top (if not one of the top) features across various models. Other factors like formulation type, clinical results, skin concern being targeted, specific ingredients and brands are also influential in product affordability.

Exploring the SHAP values allows us to go a level deeper in understanding, apart from which features are important, which certain features are more commonly associated with a particular class. For example:

- Xanthan gum, heptapeptide, and squalane for cheap products
- Ammonium acryloyldimethyltaurate, salicylic acid, and phenoxyethanol for average products
- Dark circles, titanium dioxide, and retinol for expensive products

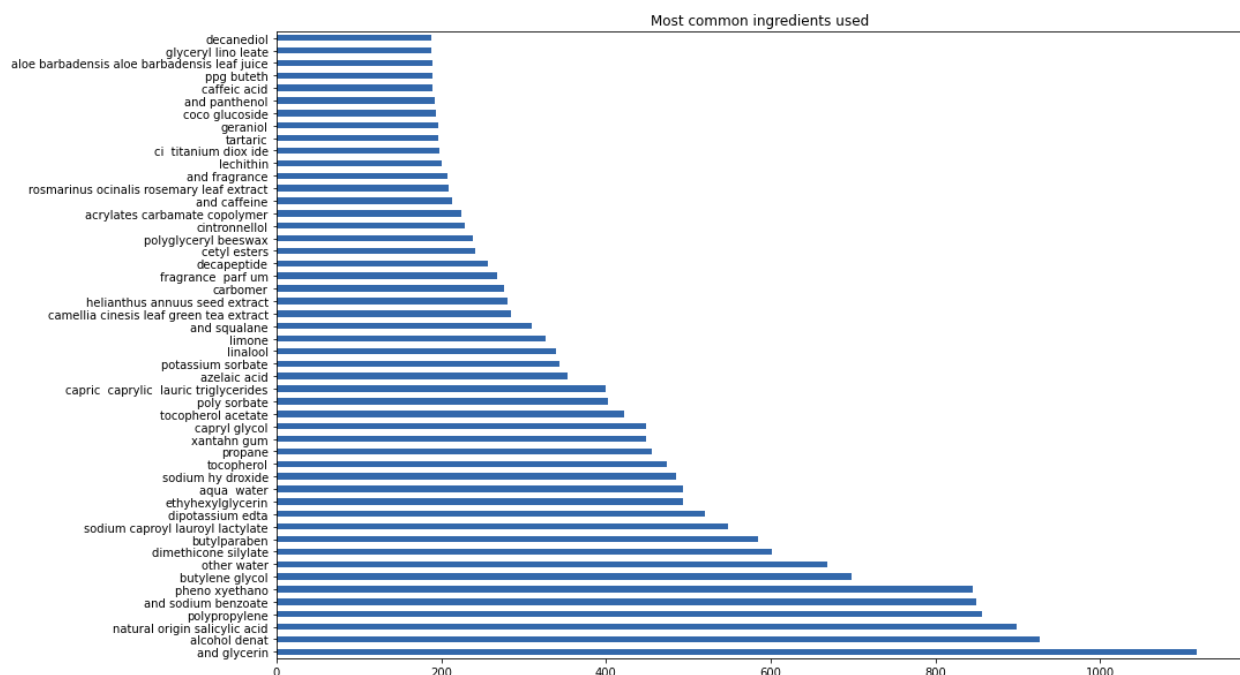


Complementing this with domain knowledge will be helpful in validating the initial insights. In the project extension of analysing ingredient formulation, it was also found that on average, the closest product (based on ingredients only) is 56.9% similar.



Out of 1338 products analyzed, there are 815 products (60.91%) for which there is a comparable competitor (based on a 50.0% similarity in the ingredients list).

Additionally, there are also ingredients often across product types. Some of these are glycerin, alcohol, salicylic acid, polypropylene.



Assumptions and limitations

In establishing the product similarity, the proportion of ingredients is also an important factor as the concentration of an ingredient can change the product effectiveness. However, this level of detail is not publicly available and is considered out of the project scope.

Another limitation is considering only the products available in Sephora. While there is a wide variety of products, the majority of the brands are still European and American. There is limited coverage for Asian products which may contain other types of ingredients. For example, snail mucin and ginseng are more common in Asian products. As such, the model generalisability may be limited.

Next steps

- The model is currently being deployed on Streamlit for easier user interaction with the model. In terms of the model deployment, it would be more helpful for the user to see a short description of the ingredient to allow them to interpret the findings more quickly.
- The model can be improved by increasing the training set. As noted earlier, it may be helpful to diversify the observations to include products used in other continents like Asia and Africa.

