Inside Beauty

November 1, 2018

Springboard data science career track capstone project

1 Table of content

2 Introduction

It can be hard to find the right beauty product, so many choices... Although there are a lot of reviews available, people can have different skin types, and usually one won't know whether she/he likes a product until having tried it for a while. If we are buying a product, hoping it can "do something" instead of just looking for a psychological effect, it's probably useful to take a look at its formula. With that, this project is focused on answering the following two questions:

1. How important is product formula in determining a beauty product's price? The price of beauty product varies a lot. There is always a debate on whether it is worthwhile to invest on expensive skin cares or makeups. While this could be a matter of personal belief, customers may be curious to know where their money goes for. Is a product expensive because it has sophisticated ingredients, or because it has attractive packaging, or simply because it is from a high-end brand. In this project, we build a regression model to predict a product's price, and to investigate whether it is the ingredients or other factors that matter the most in determining the price.

2. Given a product's ingredient list, can we more or less know its category? The cosmetic companies certainly never stop coming up with new concepts. Apart from a "face cream", we now have toner, serum, masks, face oil, eye cream, or even mask primer. I have been wondering why people would like to use eye creams: is the skin around eye somehow different from the skin of the rest of the face, and does eye cream has some special ingredients to accommodate that? If not, why would one want to buy an eye cream that is usually of the same price as a face cream but half in size? I am curious to see to what extent machine learning can distinguish the category a product belongs to just by looking at its ingredients. Are all different categories really designed for different uses or do they exist simply because of the marketing strategy of cosmetic companies? This can help people understand the cosmetic products like a chemist and simplify their skin care routine if desired.

2.1 Data Acquisition

Product and ingredient information are scraped from Beautypedia and Paula's choice websites. These are websites run by Paula Begoun and her team, where they constantly post reviews on cosmetic products.

- Product information
  - name: product name
  - category: subcategory of products
  - brand: product brand
  - ingredient: list of ingredients in a product
  - image: product image

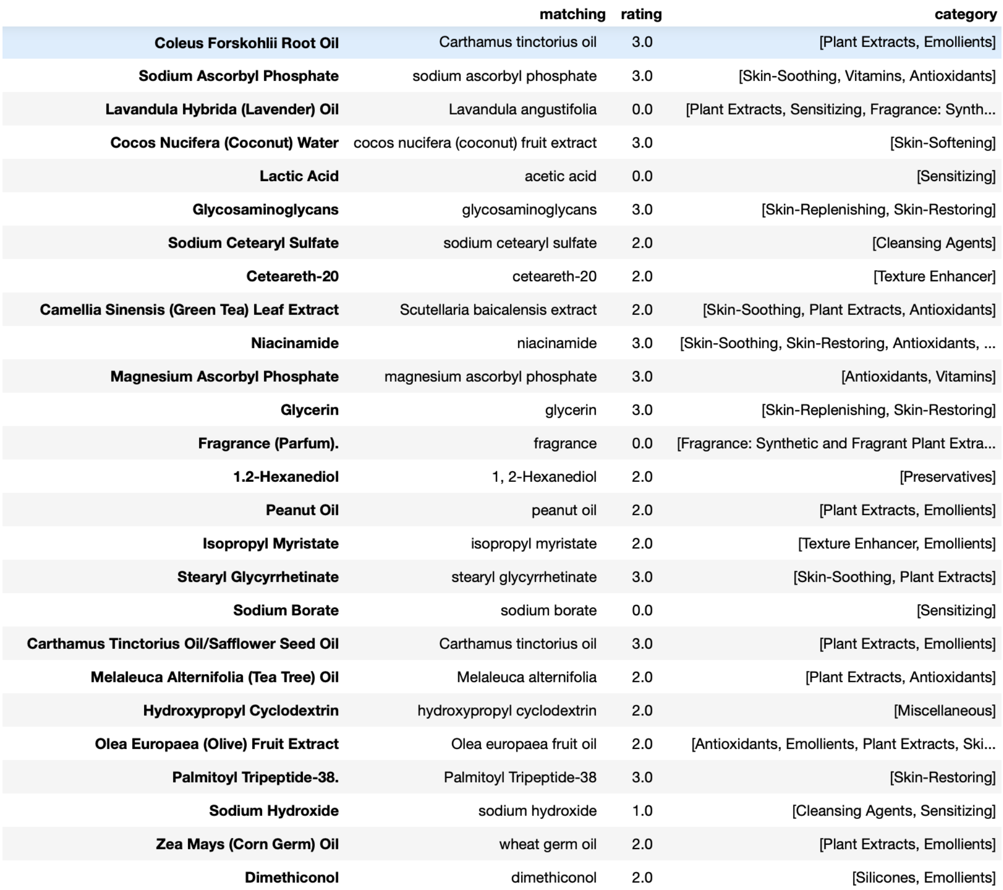
We created three main tables for skin care, body care and makeup products. Products are further divided into subcategories such as moisturizer, serum, sunscreen, exfoliator. . . , and one product may belong to multiple categories. There are 4810, 419, 2513 unique products for skin care, body care and makeups, respectively.

- Ingredient information of 1750 ingredients.
  - name: ingredient name
  - rating: rating of each ingredient according to Paula and her team
  - category: ingredient category – indicate an ingredient's function in products. An ingredient may belong to several categories.

2.2 Data Cleaning

2.2.1 Ingredient Matching

Different companies may list the same ingredient in different ways. The most common ingredient water, can appear as "water", "water (aqua)", "Water/Aqua/Eau", "purified water". . . in different products. Some companies also like to list ingredient's function in brackets, such as "Cocamidopropyl Betaine (cleansing)". To reduce sparsity of the ingredient features and make use of ingredient information in the ingredient dictionary dataset we obtained, we tried to match all ingredients to the 1750 existing ingredients. We used the SequenceMatcher from python difflib



	matching	rating	category
Coleus Forskohlii Root Oil	Carthamus tinctorius oil	3.0	[Plant Extracts, Emollients]
Sodium Ascorbyl Phosphate	sodium ascorbyl phosphate	3.0	[Skin-Soothing, Vitamins, Antioxidants]
Lavandula Hybrida (Lavender) Oil	Lavandula angustifolia	0.0	[Plant Extracts, Sensitizing, Fragrance: Synth...
Cocos Nucifera (Coconut) Water	cocos nucifera (coconut) fruit extract	3.0	[Skin-Softening]
Lactic Acid	acetic acid	0.0	[Sensitizing]
Glycosaminoglycans	glycosaminoglycans	3.0	[Skin-Replenishing, Skin-Restoring]
Sodium Cetearyl Sulfate	sodium cetearyl sulfate	2.0	[Cleansing Agents]
Ceteareth-20	ceteareth-20	2.0	[Texture Enhancer]
Camellia Sinensis (Green Tea) Leaf Extract	Scutellaria baicalensis extract	2.0	[Skin-Soothing, Plant Extracts, Antioxidants]
Niacinamide	niacinamide	3.0	[Skin-Soothing, Skin-Restoring, Antioxidants, ...
Magnesium Ascorbyl Phosphate	magnesium ascorbyl phosphate	3.0	[Antioxidants, Vitamins]
Glycerin	glycerin	3.0	[Skin-Replenishing, Skin-Restoring]
Fragrance (Parfum).	fragrance	0.0	[Fragrance: Synthetic and Fragrant Plant Extra...
1.2-Hexanediol	1, 2-Hexanediol	2.0	[Preservatives]
Peanut Oil	peanut oil	2.0	[Plant Extracts, Emollients]
Isopropyl Myristate	isopropyl myristate	2.0	[Texture Enhancer, Emollients]
Stearyl Glycyrrhetinate	stearyl glycyrrhetinate	3.0	[Skin-Soothing, Plant Extracts]
Sodium Borate	sodium borate	0.0	[Sensitizing]
Carthamus Tinctorius Oil/Safflower Seed Oil	Carthamus tinctorius oil	3.0	[Plant Extracts, Emollients]
Melaleuca Alternifolia (Tea Tree) Oil	Melaleuca alternifolia	2.0	[Plant Extracts, Antioxidants]
Hydroxypropyl Cyclodextrin	hydroxypropyl cyclodextrin	2.0	[Miscellaneous]
Olea Europaea (Olive) Fruit Extract	Olea europaea fruit oil	2.0	[Antioxidants, Emollients, Plant Extracts, Ski...
Palmitoyl Tripeptide-38.	Palmitoyl Tripeptide-38	3.0	[Skin-Restoring]
Sodium Hydroxide	sodium hydroxide	1.0	[Cleansing Agents, Sensitizing]
Zea Mays (Corn Germ) Oil	wheat germ oil	2.0	[Plant Extracts, Emollients]
Dimethiconol	dimethiconol	2.0	[Silicones, Emollients]

Sample matching results

package for this purpose. The basic idea of SequenceMatcher tries to find the longest continuous matching subsequence that is "junk free" (without e.g. space). SequenceMatcher returns a value between 0 and 1 that can serve as similarity metric between two strings. The matching results tend to "look right" to people but not necessarily be the one with minimal edit.

SequenceMatch provides overall satisfactory matching results with mistakes occasionally. For example, in the above sample matching results, lactic acid is mistakenly matched to acetic acid, zea mays (corn germ) oil is matched to wheat germ oil. Currently, we set a threshold of 0.25 and ingredients with match metric below that will be labelled as unknown ingredients. With 16985 unique ingredients in all three product dataframes, there are only 23 unmatched ingredients. Thus, potentially we have room to increase the threshold of matching to reduce falsely matched cases, although it would still be hard to deal with cases of similar name but very different ingredient pairs such as lactic acid and acetic acid, or Green 6 (a pigment) and green tea. In addition, considering we have ~8000 products and each product typically has ~20 ingredients, it seems that the products share a lot of ingredients.

2.2.2 Further Cleaning and Feature Engineering

Drop categories. We are not going to study products that are not "chemical" products, like makeup brushes, cleaning devices. #### Merge some categories. This allows us to have larger categories, so that we are able to have meaningful modelling with reasonable data size. * Merge

“Eyes”, “Eye Cream & Treatment”, ′Eye Masks" * Merge “Face & Body Sunscreen”, “Water-Resistant Sunscreen”, “Sunscreen” * Merge “Makeup Remover”, “Cleansers”, “Face Wipes” * Merge “Lips”, “Lip Balm”, * Merge “AHA Exfoliant”, “BHA Exfoliant”, “Scrubs”, “Exfoliants” * Merge “Lipstick”, “Lip Gloss”, “Lip Liner” * Merge “Eyeshadow Palette”, “Eyeshadow” * Merge “Waterproof Mascara”, “Mascara” * Merge “Foundation Without Sunscreen”, “Foundation With Sunscreen”, “Foundation”, “BB & CC Cream”, “Tinted Moisturizer”

Split “size” column to a number and unit, do unit conversion as necessary

Compute “average price” It may also be interesting to look at per size price. Some categories can be much expensive when considering the size, such as eye creams, which are typically half size of face moisturizers with similar price range.

Ingredients related features There are two kinds of ingredients: “inactive” and “active”. Inactive ingredients are general ingredients and active ingredients are usually special-functional ingredients such as sunscreen agent. It would be helpful to seperately count these two types of ingredients.

After matching all the ingredients to the ingredient dictionary, we are able to compute quite a number of features: * Number of ingredients of a certain rating (how many ingredient rated as Good/Average etc.) * Number of ingredients belongs to a certain category (how many antioxidants/sunscreen etc.) * Average or weighted rating. For products that list ingredients in descending order of their quantity, we can give ingredients different weights based on their position in the list. * Count some special categories of ingredients. For example, peptides, ingredients called “xxx extract”… * Finally we can make a giant binary matrix indicating all ingredients’ presense in each product. We may need dimensionality reduction techniques to preprocess these features for some machine learning models.

2.2.3 Image Preprocessing and Logo Image Filtering

We preprocessed all images to size 128 * 128. Some products on Beautypedia do not have real product photos but a logo of the brand. We build a simple classifier trained on hand picked small data set (with 104 logo samples and 283 non-logo samples) to filter these log images. In a sample of predictions, we got 6% false positives and 1% false negatives. We kept 6324 unique non-logo images.

|Predicted logo images |Predicted Non-logo images | |: :|: :| ||

2.3 Exploratory Data Analysis

2.3.1 Visualization

Detailed graphical EDA can be found in this notebook. We explored the following aspects:

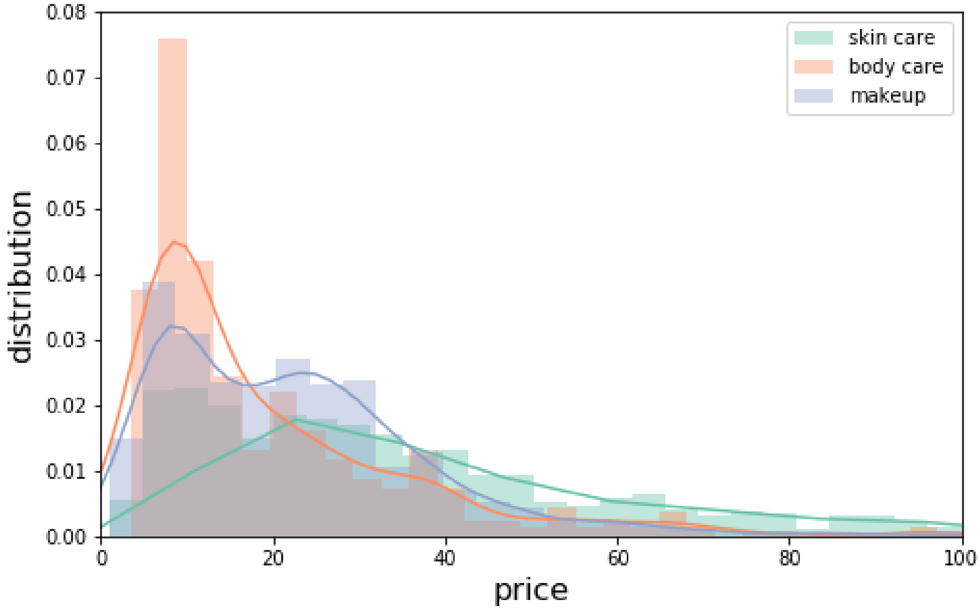
Unique products In the dataset, products belongs to multiple categories are stored in multiple rows. For example, in skin care, there are 5105 rows and 4810 unique products, with 4535 products belongs to a single category.

Missing values Makeups are more likely to have missing ingredient information, especially in categories such as lipstick and mascara, which people care less about the ingredients.

Number of products by category The largest category nighttime moisturizer has 1000+ products. Categories like retinol, skin lightener, body wash, eyeshadow primer have less than 50 products.

Number of products by brand Brand is highly cardinal: there are 224 skin care brands, 107 body care brands and 132 makeup brands.

Price v.s. category In general skin cares can be most expensive, body cares are cheapest, makeups are in the middle. Price is typically right skewed.



Most expensive category is serum (highest median price), cheap categories includes cleanser, lip balm, mascara… #### Price v.s. brand

Price varies a lot with brand. Products from expensive brands can easily cost $200, while brands such as wet'n wide, Colourpop have affordable products below 10 bucks.

Price v.s. ingredient There is no single metric to evaluate a product's formula. However, we can start with examine some simple assumptions: 1. Price is related to the number of ingredients in a product — if a product contains a long list of product, it may be more likely that the company has spend some time on this sophisticated formula and the product might be expensive. 2. Price is related to the quality of ingredients, expensive products may contain more skin-beneficial ingredients such as antioxidant. Since we have rated the ingredient using the ingredient dictionary on Paula's choice website, we can compute the average rating of each product and see if it is related to the price.

We can make scatter plots to get the initial idea of whether there is a correlation between price and number of ingredients or ingredient rating.

There seem to be a trend (although not very strong) that products with more ingredients are more expensive in skin care products. This trend is not clear in body care and makeups, which makes sense: people usually don't expect body care products to have functions such as anti-aging, pore-reducing, anti-acne, etc like for skin cares. It may be more important for a body care product to smell good and have some basic hydration effect, and that does not need a lot of ingredients.

For makeups, people may care more about the packaging, whether the color suit..., again these concerns are not strongly related to ingredient.

From the scatter plots, we don't see a strong correlation between average ingredient rating and product price, even for skin care products. Also, it seems like more ingredient doesn't mean "good" ingredient. However, we haven't consider ingredient's quantity in products. Also, we cannot garantee that our ingredient rating is scientifically correct.

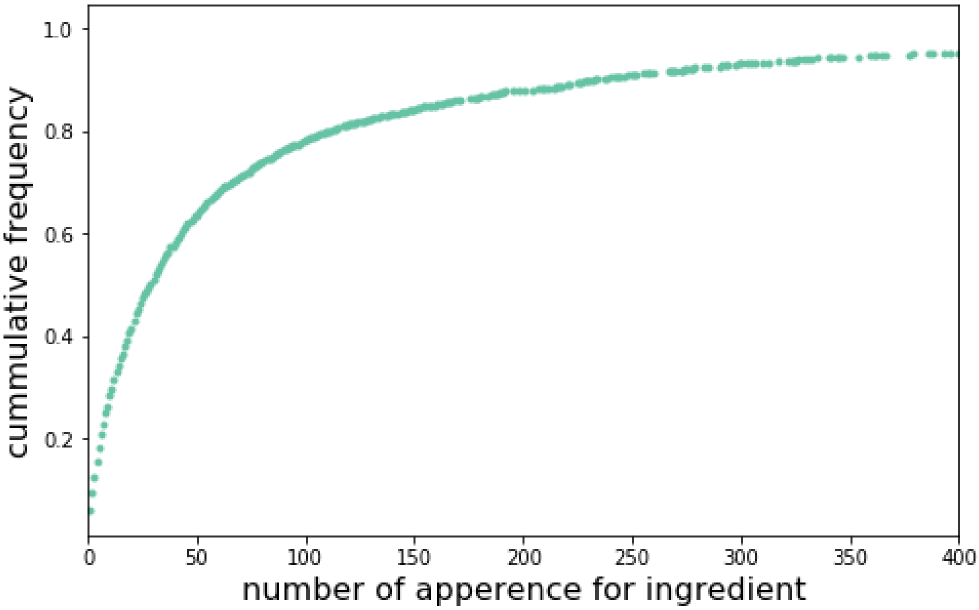
Price v.s. ingredient category We examined the correlation between price and different ingredient categories using the the pearson correlation between price and the count of ingredient for each category in products. We identified the top categories with highest correlation with price: In skin care: Skin-Restoring, Texture Enhancer, Antioxidants, Emollients, Skin-Replenishing. In body care: Antioxidants, Plant Extracts, Skin-Restoring, Hydration, Fragrance. In makeups: Plant Extracts, Skin-Replenishing, Antioxidants, Skin-Restoring, Sensitizing.

People would hope that these categories of ingredients are also the good ingredients. By ranking the ingredient based on their averge rating (table below), we can see that some categories that highly correlated with price, such as antioxidants, skin-restoring, emollients are also "good" categories, while categories such as fragrance and sensitizing ingredients are actually not good for skin — they might be there to make products more pleasant to use.

ingredient category	averge rating
Anti-Acne	3.000000
Skin-Restoring	2.808824
Vitamins	2.600000
Skin-Replenishing	2.558559
Sunscreen Actives	2.478261
Exfoliant	2.423077
Skin-Soothing	2.380000
Antioxidants	2.261044
Emollients	2.236467
Film-Forming Agents	2.200000
Hydration	2.138889
Skin-Softening	2.121212
Emulsifiers	2.038462
Slip Agents	2.000000
Thickeners	2.000000
Silicones	2.000000
Film-Forming/Holding Agents	2.000000
Coloring Agents/Pigments	2.000000
Texture Enhancer	1.990909
Thickeners/Emulsifiers	1.981982
Cleansing Agents	1.809524
Absorbent	1.783784
Plant Extracts	1.677557
Miscellaneous	1.653846
Uncategorized	1.642857
Scrub Agents	1.545455
Preservatives	1.538462

ingredient category	averge rating
Sensitizing	0.155172
Fragrance: Synthetic and Fragrant Plant Extracts	0.138686

Ingredient Frequency We plot the cumulative distribution of ingredient frequency, where the x axis is number of products containing an ingredient, and y axis is the percentage of ingredient that appear less than x times. The most interesting ingredients would be those that are neither too rare nor too common. In text data, people often set an upper bound ("max_df") and a lower bound ("min_df") while counting words, we may also consider doing that while selecting features. "min_df" should be when the curve starts to increase steeply, "max_df" should be when the curve starts to become flat.



2.3.2 Statistical testing

In graphical EDA, we have seen many factors can contribute to cosmetic products' price. We can also use statistical test to evaluate the significance of these variables.

Product Category and Brand The Anova tests for category and brand give large F-test statistic any close-to-zero p-values. This suggest at least one category or brand have different price distribution. In addition, we conduct pairwise t-test on top six brand/category to compare their price. We can visualize the p-values with heatmaps. The color encodes the logorithm of p-values when comparing the average price of two brands/categories. Red block indicates the row group is more expensive than the column group, blue indicates that row group is cheaper than the column group. Also, darker color means the two brand/category is easy to tell apart (distinguishable), while near-white blocks are where there is no statistical significant. From the heatmap, it is clear to see that serums are the most expensive among the six categories and lip products are the cheapest. Also, Clarins and Shiseido are the most expensive brands of the six, and hard to distinguish among themselves, Neutrogena is the cheapest.

|log (p_value) for comparing price of two categories |log (p_value) for comparing price of two brands | |: :|: :| ||

Slope test for number of ingredients In the EDA we have done earlier, we found price seem to be positively correlated with number of ingredient, and the effect is stronger in skin care products. We will examine the statistic significance using slope tests (t-test on slopes) on the number of ingredient.

Test results for skin care products:

Test results for body care products:

Test results for makeup products:

Surperisingly, we see that even in body care and makeup products, the number of ingredients have statistical significant with price, although the p-values are larger compare to that in skin care products. It could be the case of small practical significance but large statistical significance. The r-values are small, so the underlining correlation may be buried under large variance. Also, the number of active ingredient in skin care is negatively correlated with price, the reason might be that the products containing active ingredient such as sunscreens are cheaper categories.

Slope test for ingredient rating We apply stope test for average rating of inactive and active ingredients, and two different weighted rating of inactive ingredients (w1: reciprocal weight of ingredient position in list, w2: exponential decay with ingredient postion in list). From the statistic test, we see that the rating for inactive ingredients do has statistical significance with price, which is not obvious on the rating-price scatter plot we have in EDA earlier. Again, it could be the case of small practical significance but large statistical significance, the slopes are large (meaning the fitted lines are almost vertical), and for makeup and body care, the slopes are negative.

Test results for skin care products:

Test results for body care products:

Test results for makeup products:

F-test for ingredient category count We used statsmodel to perform linear regression on ingredient category count features. The resulting F-statistic is 68.03 and the corresponding p-value is near-zero, which indicates at least one of the features should have none zeros slope. In addtion, we obtained the t-statitic for each feature, which can help us identify possible features that have significant influence on price. The top 5 ingredient categories that have largest (absolute) t-statistic/smallest p-values are

category	slope	t-value	p-value
Skin Restoring	6.4777	14.847	0.000
Antioxident	2.6827	9.790	0.000
Texture_Enhancer	2.1385	8.810	0.000
Plant Extract	-1.6197	-7.593	0.000
Fragrance	2.2986	5.880	0.000

These match the categories we find in early EDA that have large correlation with price. However, it is interesting to see that the correlation between price and plant extract alone is possitive, but when we do linear regression with all categories here, plant extract has a negative slope.

Chi-square test for individule ingredient We will use chi-square chi square homogeneity test on some "interesting" ingredients — ingredients that appear in a considerable number of products but are not too common. We do see chi-square test identify a few ingredients which cosmetic companies advertise a lot, such as piptides, retinol, xxx extract…

Top 20 ingredient with smallest p-values:

Quite a number of ingredients have pretty small p-values. However, we should keep in mind that when the features are colinear, adding them all together to the machine learning model may not be very helpful because there aren't much new information.

2.4 Machine Learning

2.4.1 tSNE with ingredient features

Before doing supervised learning to predict product categories, we attempted to run some tSNE plots with ingredient features, and see if products belonging to the same category come close together in tSNE plot. It turns out it's not easy to separate different categories in a tSNE plot. The reason may be that there are too much noise with all the ingredient count. Those ingredient that are less relevant to product category contribute equally to the tSNE model, making it hard to reveal the pattern related to product category. However, after pruning individuel ingredients (by choosing ingredients with high chi2 statistics with category) and add other ingredient-related features such as count and rating, we are able to see a vague cluster of cleansers (red dots), and a cluster which is mostly a mixture of sunscreen and daytime moisturizer (purple and brown dots):

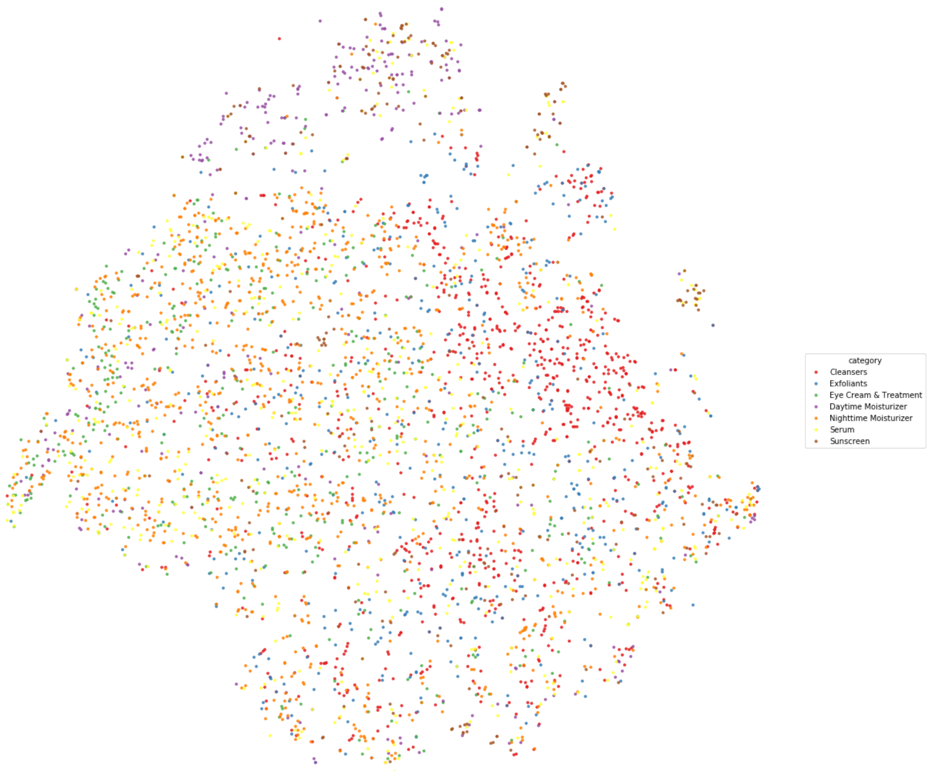
Still, the tSNE plot looks quite noisy. Eye creams, nighttime moisturizer and serums tend to mix together, which may not be a surprise as they are similar after all. We would want to do supervised learning to see if we can distiguish product category further.

2.4.2 Product category classification with ingredient features

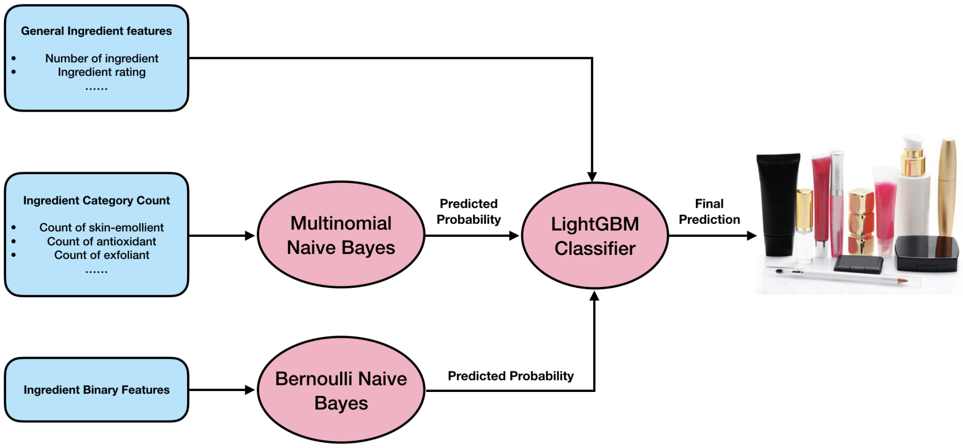
We attempted to predict a product's category using only ingredient related features. This can provide us insight over which categories are similar in their formula and and help people understand chemical aspect of cosmetic products. It will be a multilabel classification problem as a product may belong to multiple categories, for example "Even Skin Vitamin C Cream" from brand Trish McEvoy is both nighttime moisturizer and vitamin C product. We can do one vs rest classification to tackle the multilabel problem. Scikit-learn's OneVsRestClassifier is the perfect wrapper for this job. It can take any regular classifier as estimator to do one vs rest classification for each class under the hood and return the prediction for all classes.

The ingredients are like the words in documents, thus many techniques people use for text data can be applied here. We can create a "bag of ingredients" matrix, where the matrix elements are binary indicators of whether a certain ingredient exists in a certain product. As in text data, Naive Bayes can be a good baseline model. In this case, we can use Bernoulli Naive Bayes to deal with binary ingredient features. In addition, we also have ingredient category available in after finding the matching ingredients in the ingredient dictionary. Thus, we can have "bag of ingredient categories" features that counts how many ingredients of an ingredient category are in a product. These features can be used with Multinomial Naive Bayes.

Apart from the count features and binary features, there are also some general ingredient features that could potentially be useful, such as number of ingredients. We can make use of these features though model stacking: the naive bayes models will serve as first layer models, then the


category
Cleansers
Exfoliants
Eye Cream & Treatment
Daytime Moisturizer
Nighttime Moisturizer
Serum
Sunscreen

tsne



flow_chart

predicted probabilities can be joined with general ingredient features to feed in the final model. The training pipeline is as follows:

We focused on 16 categories which have more than 100 products in the dataset.

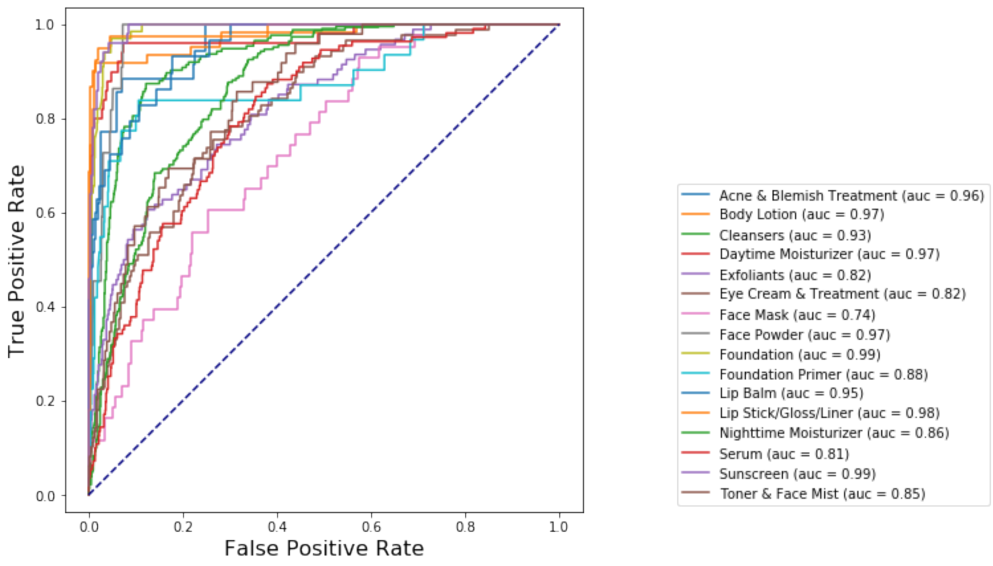
Bernoulli Naive Bayes with binary ingredient features The Bernoulli Naive Bayes serves as a good baseline model. It achieves 0.3724 Hamming score on training set with cross validation, and 0.3588 on test set. The auc (area under curve) score for all 16 categories are larger than 0.78. Also, it is straight forward to use Bernoulli Naive Bayes to identify the key ingredients that are mostly associated with each category. This is done by predicting on individual ingredient ("hypothetical" products with just one ingredient, the feature matrix will be an identity matrix with dimension N_ingredient). The results makes sense in general, we are able to identify:

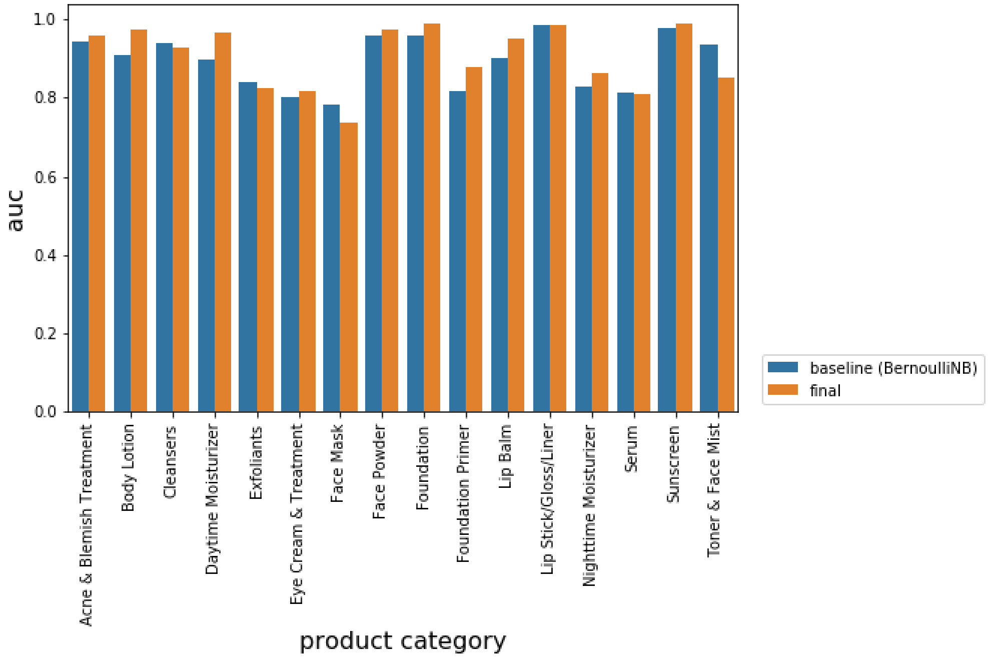
- Acne Treatment products: benzoyl peroxide, BHA(beta hydroxy acid)
- Cleansers: solium xxx (sodium salt of fatty acids)
- Exfoliants: AHA(alpha hydroxy acid), BHA(beta hydroxy acid)
- Lipsticks: Coloring Agents/Pigments
- Sunscreens: octocrylene, homosalate... (common sunscreen agents)

top 10 ingredients associated with each category

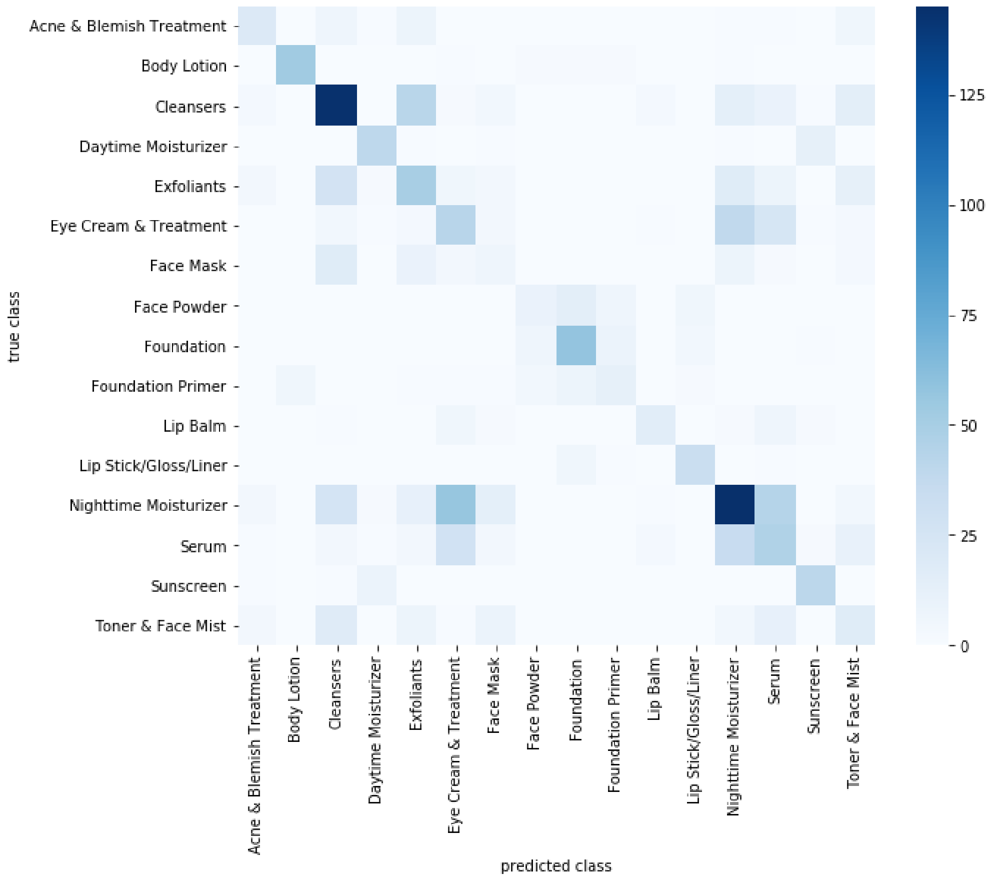
Multinomial Naive Bayes with ingredient category count features We achieved 0.2707 Hamming score using ingredient categories alone. The results is not good enough by itself, but the predictions can be useful features for the final model.

Final Prediction with LightGBM We stacked general ingredient features (number of ingredient, average/weighted rating of ingredients in products) and the predictions of the above two models to train the final model. The final model is based on gradient boost decision tree with lightgbm package. The final Hamming score is 0.4837 for cross-validation on training set and 0.5111 on test set, which is a significant improvement compare to Naive Bayes. We also improved the auc score of 10 product categories. We do note that there are five categories that the final more complicated model cannot beat naive bays: cleansers, exfoliants, face mask, serum and toner.





We can also visualize the model predictions using confusion matrix. In confusion matrix, diagonal elements count the cases when our model correctly predicts a class (true possitives), while off-diagonal elements count the cases when the true class is the row class but the model mistakenly predicted the column class (false positives). We tweeked the definition of confusion matrix for multilabel classification, cases when true classes are both A and B and the model predicts B won't count as false positive in the (A,B) element.



The confusion matrix is in general diagonal dominant, which indicates our model prediction is reasonable. We can also identify the pairs of categories that our model get confused with:

True Class	Top miss predictions
Acne & Blemish Treatment	Exfoliants, Cleansers, Toner & Face Mist
Body Lotion	Foundation Primer, Foundation, Face Powder
Cleansers	Exfoliants, Toner & Face Mist, Nighttime Moisturizer
Daytime Moisturizer	Sunscreen, Nighttime Moisturizer, Face Mask
Exfoliants	Cleansers, Nighttime Moisturizer, Toner & Face Mist
Eye Cream & Treatment	Nighttime Moisturizer, Serum, Cleansers
Face Mask	Cleansers, Exfoliants, Nighttime Moisturizer
Face Powder	Foundation, Foundation Primer, Lip Stick/Gloss/Liner
Foundation	Foundation Primer, Face Powder, Lip Stick/Gloss/Liner
Foundation Primer	Foundation, Body Lotion, Face Powder
Lip Balm	Serum, Eye Cream & Treatment, Sunscreen
Lip Stick/Gloss/Liner	Foundation, Serum, Foundation Primer
Nighttime Moisturizer	Eye Cream & Treatment, Serum, Cleansers
Serum	Nighttime Moisturizer, Eye Cream & Treatment, Toner & Face Mist
Sunscreen	Daytime Moisturizer, Cleansers, Acne & Blemish Treatment
Toner & Face Mist	Cleansers, Serum, Face Mask

Overall, the results do agree with our life experience. The model gets confused on similar categories. For example, daytime moisturizers often have sunscreen ingredients in it, so sometimes our model cannot distiguish sunscreens and daytime moisturizer. Nighttime moisturizer, eye creams and serum are another group that our model get confused a lot in real life, they are all products that are supposed boost hydration and may have some special functions such as anti-aging, reduce hyperpigmentation… It is interesting to see that face masks got confused with cleansers, Exfoliants, and nighttime moisturizer. This is because there are typically two types of face masks: cleansing mask, which may have similar ingredient like cleansers and exfoliants. Another is the so called "sleeping mask", which you can wear overnight, they are typically like a heavy nighttime moisturizer.

Checkout these two miss-predicted face maskes (the first one I have used it as moisturizer personally).

- Clinique Moisture Surge™ Overnight Mask, predicted as Eye Cream & Treatment and Nighttime Moisturizer:

ingredient: Water, Glycerin, Cetyl Alcohol, Dimethicone, Glyceryl Polymethacrylate, Butyrospermum Parkii (Shea Butter), Cetyl Ethylhexanoate, PEG-8, Glycereth-28, Sucrose, Sorbitan Stearate, PEG-100 Stearate, Trehalose, Mangifera Indica (Mango) Seed Butter, Hypnea Musciformis (Algae) Extract, Gellidiela Acerosa (Algae) Extract, Olea Europaea (Olive) Fruit Extract, Triticum Vulgare (Wheat Bran) Extract, Cladosiphon Okamuranus Extract, Astrocaryum Murumuru Seed Butter, Cetearyl Alcohol, Aloe Barbadensis Leaf Water, PEG-75, Caffeine, Pantethine, Sorbitol, Butylene Glycol, Oryzanol, Bisabolol, Panthenol, Phytosterols, Tocopheryl Acetate, Caprylyl Glycol, Sodium Hyaluronate, Hexylene Glycol, Carbomer, Potassium Hydroxide, Dextrin, Disodium EDTA, Phenoxyethanol, Red 4, Yellow 5

This mask has moisturizing ingredients such as glycerin, shea butter, and a number of antioxidant. This makes it similar to a moisturizer and eye cream.

- Clinique Pep-Start Double Bubble Purifying Mask, predicted as Cleansers.

ingredient: Water, Disiloxane, Cocamidopropyl Betaine, Sodium Cocoyl Isethionate, Decyl Glucoside, Glycerin, Pentylene Glycol, Acrylates Copolymer, Sodium Chloride, Morus Nigra (Mulberry) Root Extract, Scutellaria Baicalensis Root Extract, Vitis Vinifera (Grape) Fruit Extract, Palmitoyl Tetrapeptide-7, Palmitoyl Tripeptide-1, Acetyl Glucosamine, PEG-6 Caprylic/Capric Glycerides, Ethylhexylglycerin, Butylene Glycol, Polysorbate 20, PEG-150 Pentaerythrityl Tetrastearate, Potassium Hydroxide, Coconut Acid (Coconut Derived), Acrylates/C10-30 Alkyl Acrylate Crosspolymer, Carbomer, Xanthan Gum, Disodium EDTA, Phenoxyethanol, Red 33

The cocamidopropyl betaine and sodium cocoyl isethionate in this product are surfactants, which makes it easy to be confused with cleansers.

2.4.3 Price Regression

We built a LightGBM regression model to predict a product's price with both ingredient and non-ingredient features, and assess the relative importance of ingredients in determining price versus other factors such as brand and packaging.

The non-ingredient features include: * Brand * Product category * Size (includes numerical size and size unit) * Packaging (CNN model prediction with product images)

We apply target encoding on brand, product category and size unit. The packaging feature is the prediction of CNN model with product images. The price prediction with image features alone has RMSE = $34.7 for out-of-bag prediction on training set. To ensure no leakage in the model, we use the same fold for CNN model and final prediction.

Ingredient features include: * General features such us number of inactive/active ingredient, average ingredient rating. * Count of ingredients for each ingredient category * 50 selected individual ingredient binary features by chi2 test with price. * tf-idf counts on binary ingredient matrix followed by NMF (non-negative matrix factorization), select the top 50 components as features.

Model pipeline:

The table below summarized the RMSE, MAE and explained variance when using non-ingredient features or ingredient features alone, and final prediction with all features. The ingredient features are not as powerful as non-ingredient features. The five non-ingredient features alone achieved MAE = $10.875. Adding 171 ingredient features only improved the results slightly. But of course, we can always exploit more about ingredient and keep improving our model. The most powerful features from LightGBM's feature importance are brand and product category.

	non-ingredient features only	ingredient features only	all features
RMSE($) (train cv)	22.280	27.445	21.393
RMSE($) (test)	18.629	24.803	17.100
MAE($) (train cv)	12.322	17.130	11.692
MAE($) (test)	10.875	16.298	10.257
explained variance (train cv)	0.610	0.408	0.640
explained variance (test)	0.683	0.438	0.732

2.4.4 Conclusion

We have seen our machine learning model can more or less predict the category of a cosmetic product just using the ingredient information. It is able to associate surfactant with cleanser, coloring agents/pigment with lipsticks, AHA(alpha hydroxy acid) or BHA(beta hydroxy acid) with exfoliator, sunscreen agents with daytime moisturizers and sunscreen products… Our model also helps us identify categories that are similar ingredient-wise, such as eye creams, moisturizers and serums. Based on the machine learning results, it would make more sense for a customer to use a sleeping mask as a nighttime moisturizer, but it would not be wise to use foundation as sunscreens.

From the EDA we have done, we do see some ingredient categories that positively correlated with price are the "good" categories according to beautypedia's expert rating, which is reassuring. But also keep in mind that things like fragrance are also making products expensive, not because it is good to the skin, but because it makes the product pleasant to use. In the price regression, we found ingredient features do have predictive power in determining price, although not as powerful as factors such as brand and product category currently.

We do note there are several limitations in the current model. For example, currently our model is not doing well in classify toners, while in reality, people can easily tell apart a toner based on its texture. The limitation stems from the lacking of quantity information in the ingredient lists. If we know the percentage of water in a product, we can probably do better in distinguishing toners and other products. In the future, we may work on the following aspects to further improve the current model: * Explore other methods for ingredient matching, such as algorithms based Levenshtein distance. * Finding more data from other resourses can help improve the models. Currently, many brands and categories do not have enough products. * Improve the price prediction using product images.

Finally, I want to say that while I believe understanding the formula of cosmetic products and having a knowledge of what's inside the products can make people more rational while spending their money, and have a realistic expectation in the products, I also think it is totally fine to buy products not for need, but for fun!