UC Irvine
Paul Merage
School of Business

Inside Sephora Reviews: Understanding the Drivers of High-Rated Products

Machine Learning for Analytics
Team 18B
Christie Shin
Shivani Vallamdas
Gema Zhu
Vishal Srivastava
Shuai Zhao

December 5, 2025

Table of Contents

Executive Summary

The cosmetics industry has experienced significant growth in recent years, driven largely by increased online engagement and the rising influence of customer-generated reviews. For retailers like Sephora, product ratings and reviews have become essential indicators of customer satisfaction and are widely relied upon by shoppers when making purchase decisions. However, the interpretation of these ratings can be affected by several factors, including the number of reviews a product receives and the attributes associated with the product and its brand. Products with limited review volume often exhibit highly variable ratings, while those with large numbers of reviews tend to show more stable evaluations. Understanding how review volume interacts with product ratings and attributes is increasingly important for Sephora.

Introduction

1.1 Business Idea

Sephora relies heavily on online ratings and review data to ensure that the products it carries truly meet customer expectations. Each product is associated not only with a variety of inherent attributes, such as price and product category, but also by customer-generated signals like review volume and rating. Together, these factors influence how shoppers perceive a product's credibility, quality, and overall value.

Because of this, it is crucial for Sephora to understand which combinations of attributes consistently lead to highly rated products. Gaining insight into these relationships allows Sephora to distinguish between products that are genuinely well-received and those that only seem highly rated because they have too few reviews or inflated ratings. By understanding how product attributes, review volume, and ratings interact to influence brand trustworthiness and perceived product quality, Sephora can make more informed merchandising decisions to ensure that the products it promotes are both reliable and well-received by customers.

Through this analysis, we aim to answer the following questions:

1. Which brands consistently produce products with the highest predicted rating probability?
2. How does the number of reviews influence the probability of a product having a high rating?
3. Does the variation type (full size, mini, value set, etc.) significantly influence whether a product receives a high rating?
4. Which product categories (primary, secondary, tertiary) are most associated with high-rated skincare products?

1.2 Data Source

The data used in this study was sourced from Kaggle under the title *"Sephora Products and Skincare Reviews"*. The dataset was collected via a Python-based web scraper in March 2023 and provides a comprehensive view of Sephora's skincare product attributes and customer review activity.

The dataset contains two main components:

- **Product Data**
  - Information for over 8,000 beauty products, including product, brand names, price, ingredients, and product size and variation type
- **Review Data**
  - Approximately 1 million user reviews across more than 2,000 skincare products, including individual star ratings, and review text

Below is a detailed description of each of the features we used:

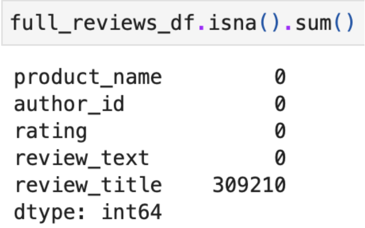
product_name	The full name of the product
brand_name	The full name of the product brand
price_usd	The price of the product in US dollars
primary_category	First category in the breadcrumb section
secondary_category	Second category in the breadcrumb section
tertiary_category	Third category in the breadcrumb section
variation_type	The type of variation parameter for the product
ingredients	A list of ingredients included in the product
highlights	A list of tags or features that highlight the product's attributes (e.g., Vegan, Matte)
new	Indicates whether the product is new or not (1-true, 0-false)
sephora_exclusive	Indicates whether the product is exclusive to Sephora or not (1-true, 0-false)
rating	The average rating of the product based on user reviews
reviews	The number of user reviews for the product
size	The size of the product (oz, ml)

Data and Research Method

2.1 Data Cleaning

First, we appended the 5 csv files of reviews together into a pandas data frame labelled *full_reviews_df*. We subsetted the 4 necessary columns from this dataset and resaved them into the data frame. The columns we used were 'product_name', 'author_id', 'rating', ‘review_text’, and ‘review_title’. We then created a new data frame labelled *missing_reviews* which included all the rows that had missing reviews by using the .isna() function and finding any rows that had empty spaces by using (full_reviews_df[‘review_text’].str.strip() == “”). This dataframe had 1444 rows.

Next, we resaved the *full_reviews_df* by subsetting only the rows that are not missing reviews and that have text in the row. To do this, we used full_reviews_df[‘review_text’].notna() & (full_reviews_df[‘review_text’].str.strip() != “”). After this process, we checked if there were any missing values by using .isna().sum(). Since we are focusing on only the reviews and the review_title column was added just for extra context and not an integral part of our analysis, we kept the rows that had a missing review_title.



```
full_reviews_df.isna().sum()
```

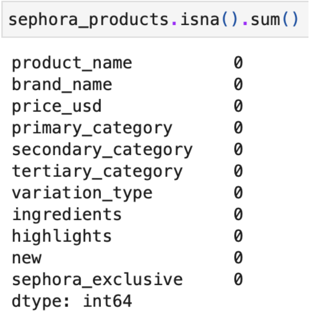
```
product_name         0
author_id            0
rating               0
review_text          0
review_title    309210
dtype: int64
```

Lastly for this data frame, we reset the index to ensure the rows were numbered properly after removing all the missing values. After removing the missing values, the number of rows went from 1,094,411 to 1,092, 967.

Next, we created a data frame for the products dataset by also subsetting columns from the main dataset. We used the following columns:

‘product_name’, ‘brand_name’, ‘price_usd’, ‘primary_category’, ‘secondary_category’, ‘tertiary_category’, ‘variation_type’, ‘ingredients’, ‘highlights’, ‘new’, ‘sephora_exclusive’.

To remove the missing values we created a list of columns called ‘cols_to_check’ which included all the columns from above. We then used .dropna(subset = cols_to_check) to drop the missing values in these columns. The index was then reset as well. After removing the missing values, the number of rows went from 8494 to 4710.

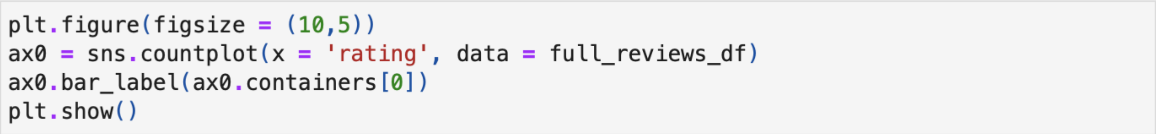


```
sephora_products.isna().sum()
```

```
product_name          0
brand_name            0
price_usd             0
primary_category      0
secondary_category    0
tertiary_category     0
variation_type        0
ingredients           0
highlights            0
new                   0
sephora_exclusive     0
dtype: int64
```

2.2 Exploratory Data Analysis

After cleaning the data we visualized certain parts of the data to get a better understanding of the distribution of the data. The graph below shows the distribution of the ratings of the products. We can see that the data is imbalanced and the majority of the ratings are 4 or 5 star ratings.



```
plt.figure(figsize = (10,5))
ax0 = sns.countplot(x = 'rating', data = full_reviews_df)
ax0.bar_label(ax0.containers[0])
plt.show()
```

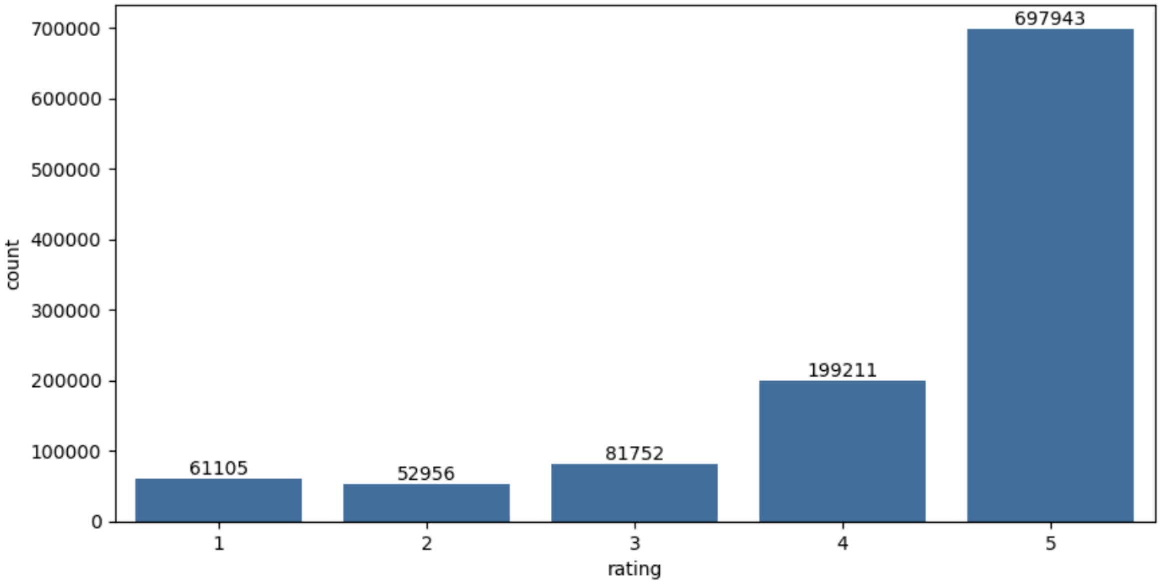


Figure 2.2.1 Distribution of Product Numeric Ratings

Additionally, we graphed the distribution of the primary_category, secondary_category, and the top 10 from the tertiary_category. The tertiary_category column has many values and some of the categories had very little values therefore we decided to look at only the top 10. These 3 categories are from the sephora_products data frame. Please refer to the 3 figures below.

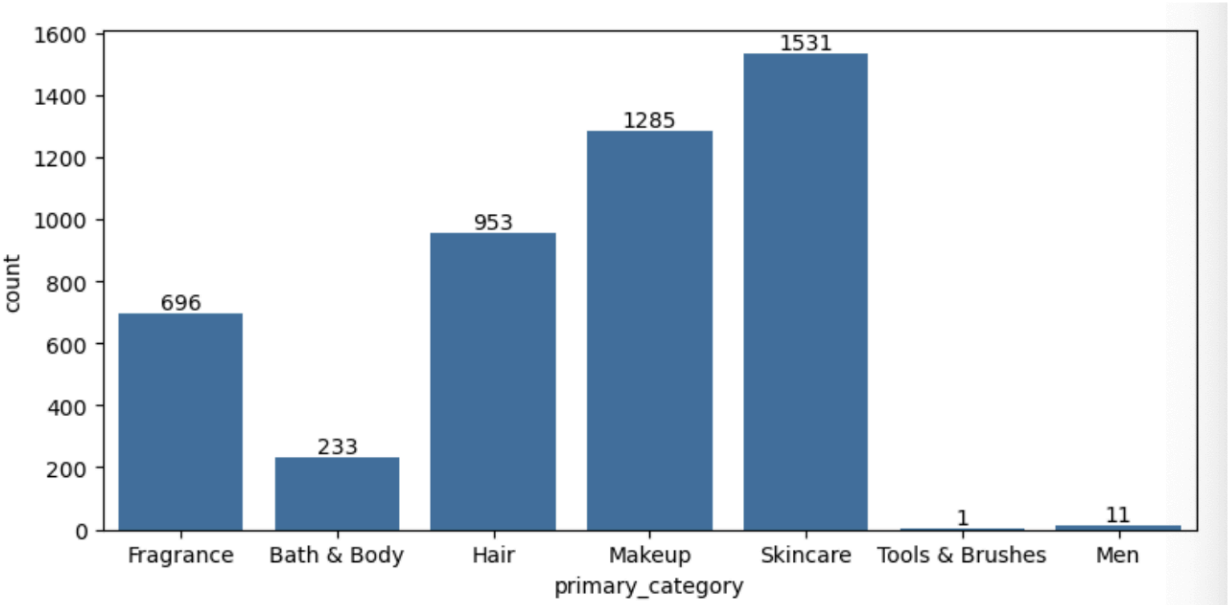


Figure 2.2.2 Distribution of ‘primary_category’

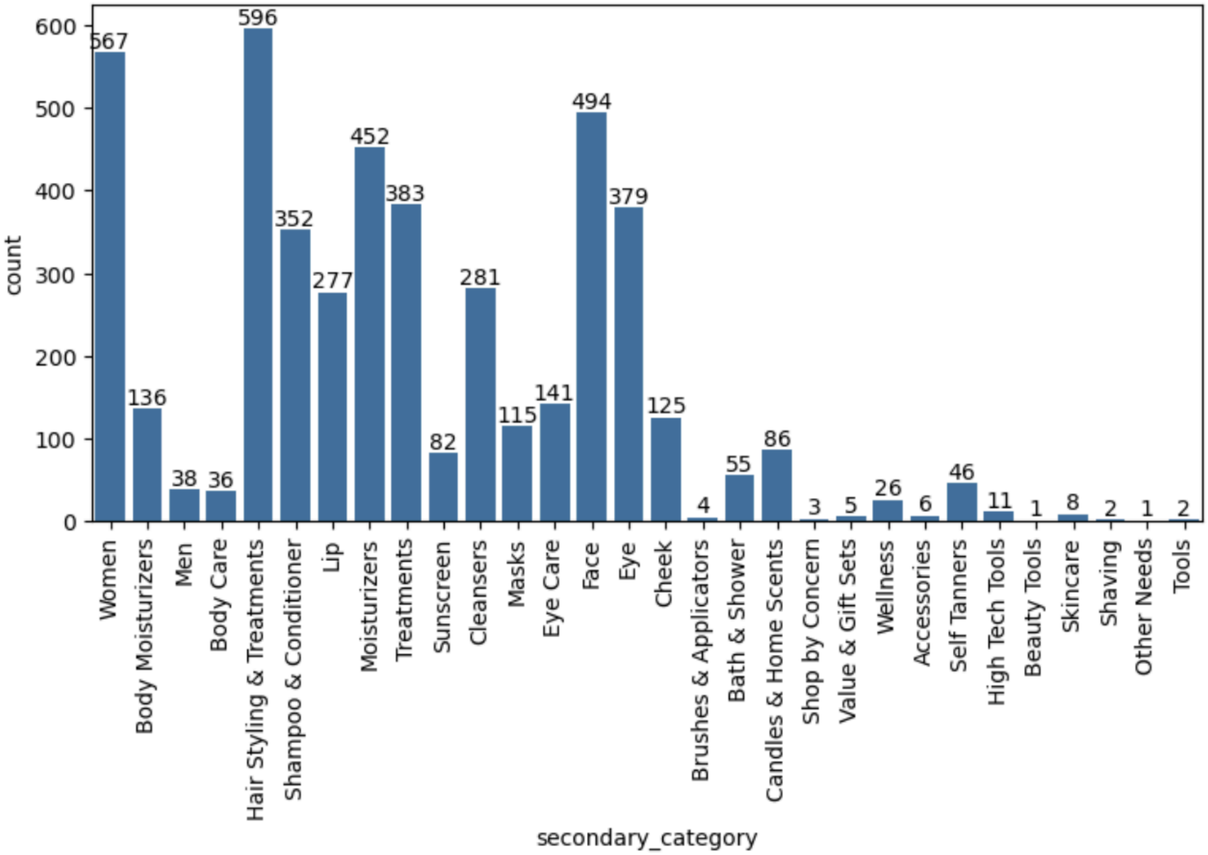


Figure 2.2.3 Distribution of ‘secondary_category’

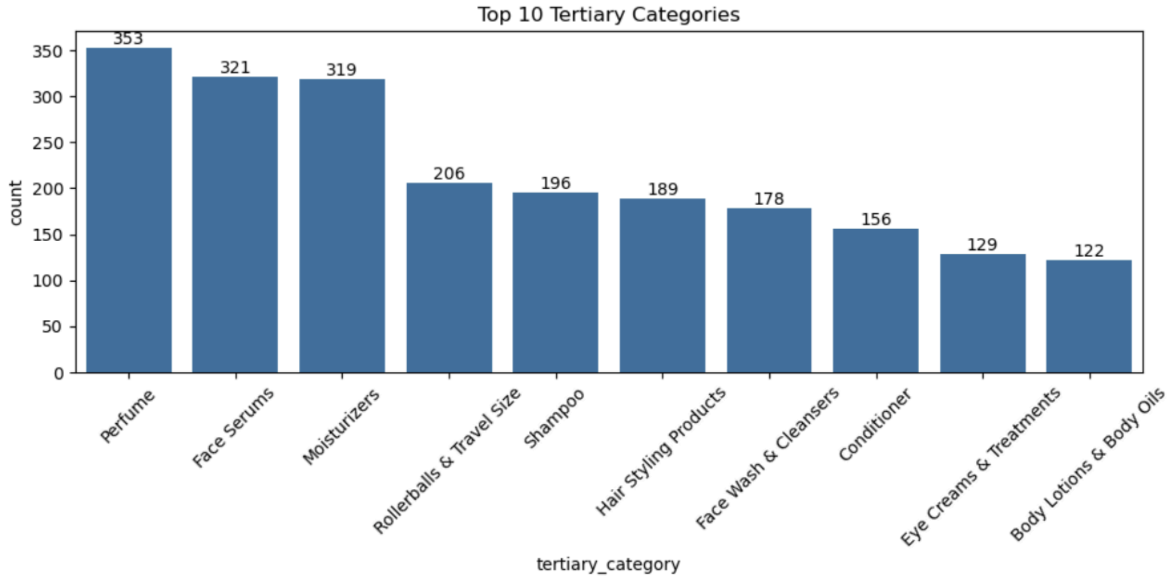


Figure 2.2.4 Distribution of 'tertiary_category'

Furthermore, the figure below shows the distribution of products that were new (1) and products that were not new (0).

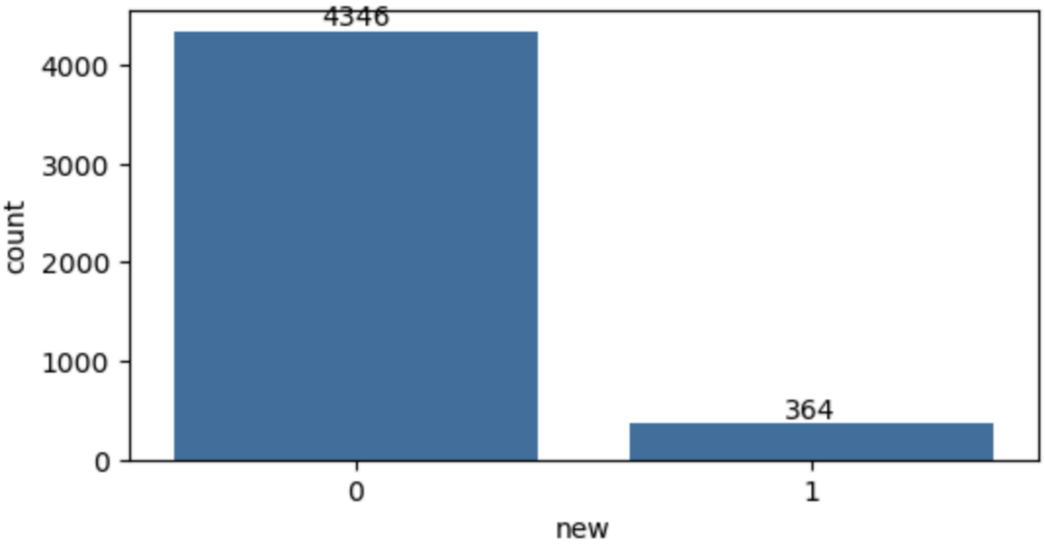


Figure 2.2.5 Distribution of 'new' products

The figure below shows the distribution of products that are sephora_exclusive (1) and products that are not sephora_exclusive (0).

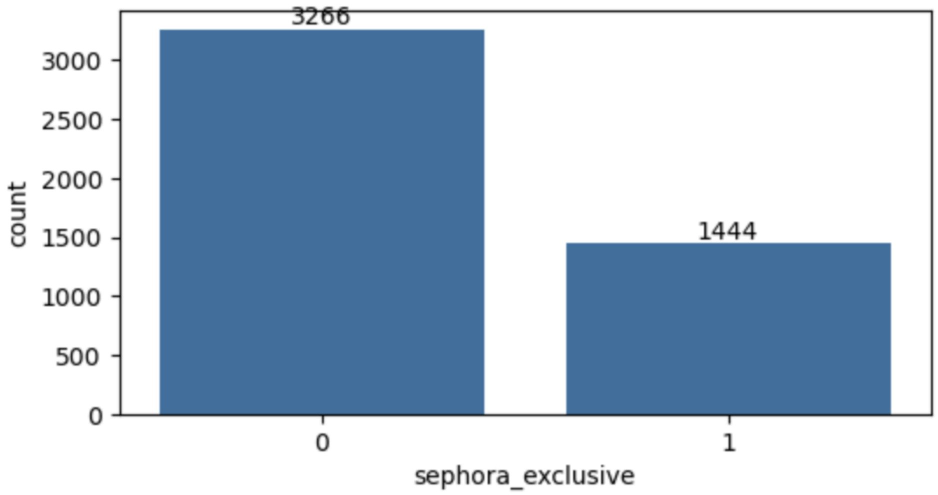


Figure 2.2.6 Distribution of ‘sephora_exclusive’ products

Lastly, the figure below shows the distribution of variation_type of the products.

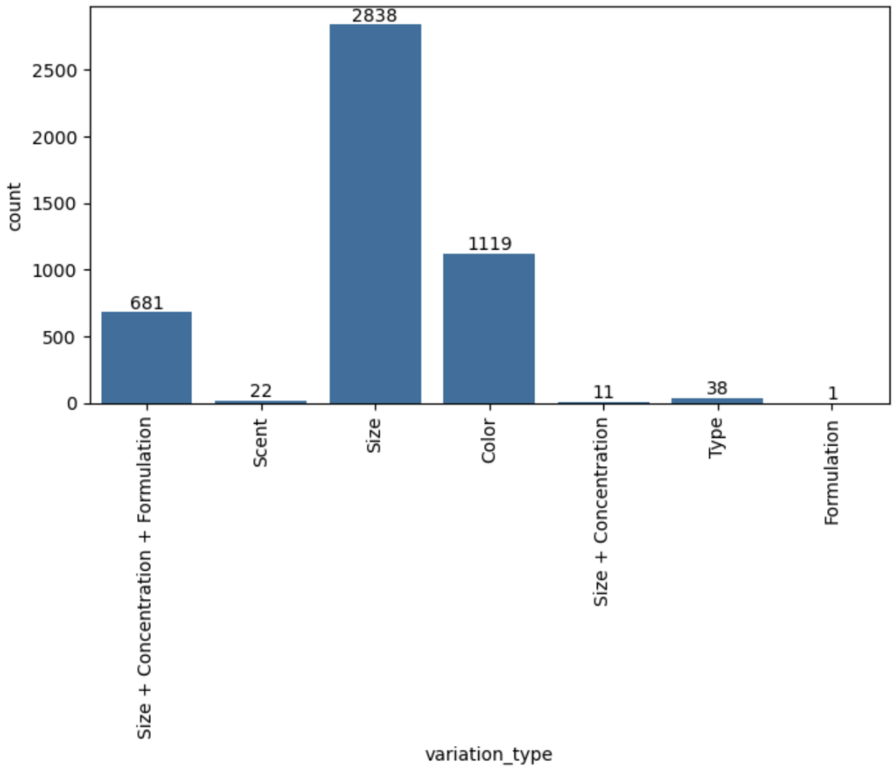


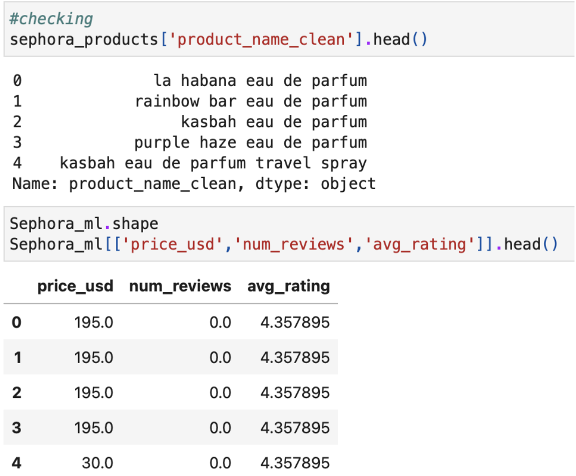
Figure 2.2.7 Distribution of ‘variation_type’ of products

The exploratory data analysis helped us better understand the shape of our data and where our biases could potentially lie.

2.3 Train-test Split

First, we cleaned the column ‘product_name’ in both data frames by converting all values to strings using .astype(str), converting all text to lowercase using .str.lower(), removing extra spaces from the beginning and end of each value using .str.strip(), and removing quotation marks both straight and curly using .str.replace(r’[“””]’, ‘’, regex=True). We placed these clean product names into a new column titled ‘product_name_clean’.

Second, we used the groupby() function to group all the reviews by product name using the ‘product_name_clean’ column. We then created 2 new columns from the grouped data. ‘Avg_rating’ includes the mean of all ratings per product and ‘num_reviews’ includes the total number of ratings per product. After this step, we merged this data into the product data frame and created a new data frame from this combined data labelled Sephora_ml. Lastly, we filled any missing review counts with 0 so this could be used in the future for our ML models. Any missing reviews in ‘num_reviews’ was filled with 0 instead of NaN and the missing values in ‘avg_rating’ were filled with the median of the average rating.



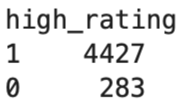
```
#checking
sephora_products['product_name_clean'].head()
```

```
0              la habana eau de parfum
1            rainbow bar eau de parfum
2                 kasbah eau de parfum
3            purple haze eau de parfum
4    kasbah eau de parfum travel spray
Name: product_name_clean, dtype: object
```

```
Sephora_ml.shape
Sephora_ml[['price_usd','num_reviews','avg_rating']].head()
```

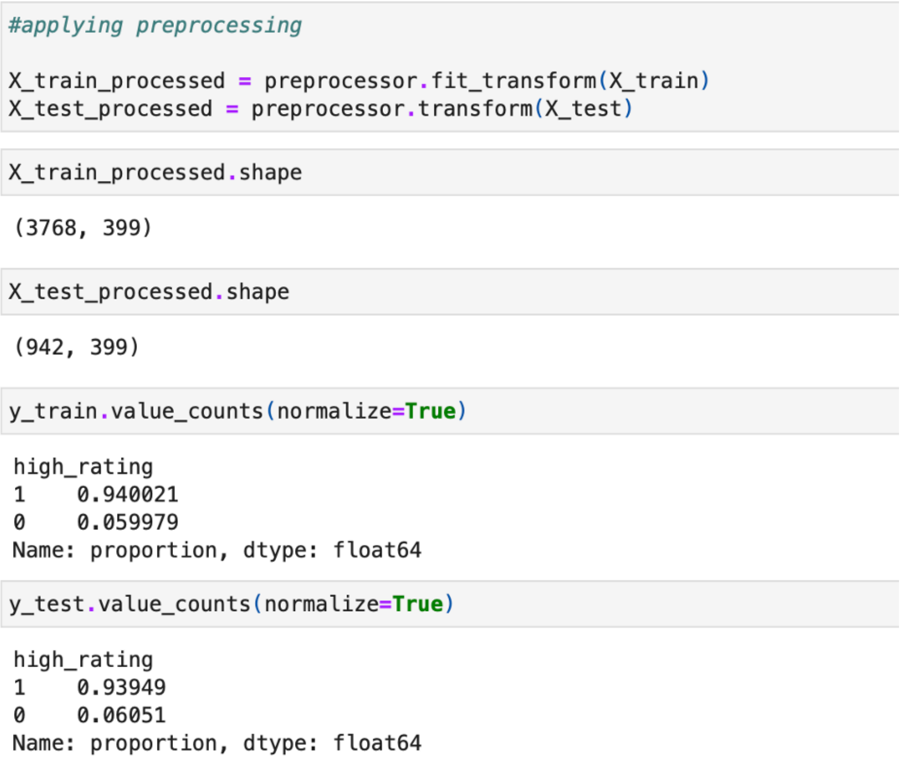
	price_usd	num_reviews	avg_rating
0	195.0	0.0	4.357895
1	195.0	0.0	4.357895
2	195.0	0.0	4.357895
3	195.0	0.0	4.357895
4	30.0	0.0	4.357895

To set up the final modeling, we created a column called ‘high_rating’ that included the products that have an average rating higher than 4. A high rating was given a 1, otherwise it was given a 0.



```
high_rating
1    4427
0     283
```

Lastly, to begin splitting the data, we subsetted the necessary features that we need for the model and assigned it to X. The necessary features were ‘price_usd’, ‘num_reviews’, ‘brand_name’, ‘primary_category’, ‘secondary_category’, ‘tertiary_category’, ‘variation_type’, ‘new’, ‘sephora_exclusive’. We assigned the ‘high-rating’ column to Y. We also separated the numerical and categorical columns as ‘num_features’ and ‘cat_features’ . We then built the preprocessing pipeline using StandardScaler and OneHotEncoder from sklearn.preprocessing and ColumnTransformer from sklearn.compose. To actually split the data, we used train_test_split from sklearn.model_selection. The test_size was 0.2 and the random_state was equal to 42. The stratify was equal to y. Lastly we applied preprocessing to X_train and X_test.



```
#applying preprocessing

X_train_processed = preprocessor.fit_transform(X_train)
X_test_processed = preprocessor.transform(X_test)
```

```
X_train_processed.shape
```

```
(3768, 399)
```

```
X_test_processed.shape
```

```
(942, 399)
```

```
y_train.value_counts(normalize=True)
```

```
high_rating
1    0.940021
0    0.059979
Name: proportion, dtype: float64
```

```
y_test.value_counts(normalize=True)
```

```
high_rating
1    0.93949
0    0.06051
Name: proportion, dtype: float64
```

2.4 Cross-Validation on Preprocessing Data

To ensure robust model evaluation and prevent overfitting, we applied a five-fold stratified cross-validation to both the logistic regression and random forest models. First, we constructed a full modeling pipeline by embedding the data preprocessing steps directly into the cross-validation procedure. The preprocessing pipeline included standardization of numerical features using StandardScaler and one-hot encoding of categorical features using OneHotEncoder. The pipeline was fit only on the training portion of each fold, and the learned transformations were then applied to the corresponding variation fold. This approach fully prevents data leakage, since no information from the validation fold was used during the preprocessing or model fitting.

For each fold, the models were trained on 80% of the data and validated on the remaining 20%. It was also rotated across all five folds so that each observation became a validation point exactly once. We then evaluated the model performance using accuracy, precision, and recall. The cross-validated scores were then aggregated to compute the mean and standard deviation for each metric, which served as an estimate of the model stability.

The five fold accuracy values were visualized using bar plots with a dotted line referencing the average accuracy, as seen in *Figure 2.4.1* and *Figure 2.4.2*.

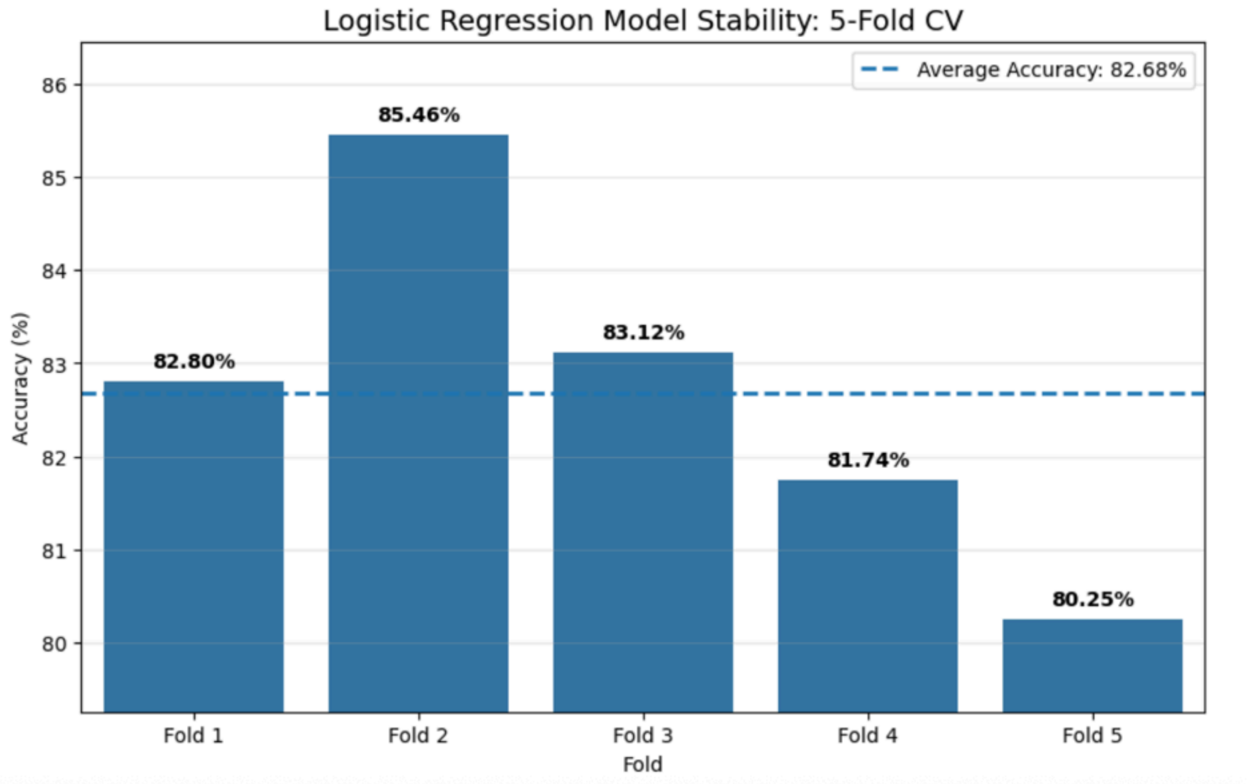


Figure 2.4.1 Distribution of five fold accuracy for Logistic Regression Model

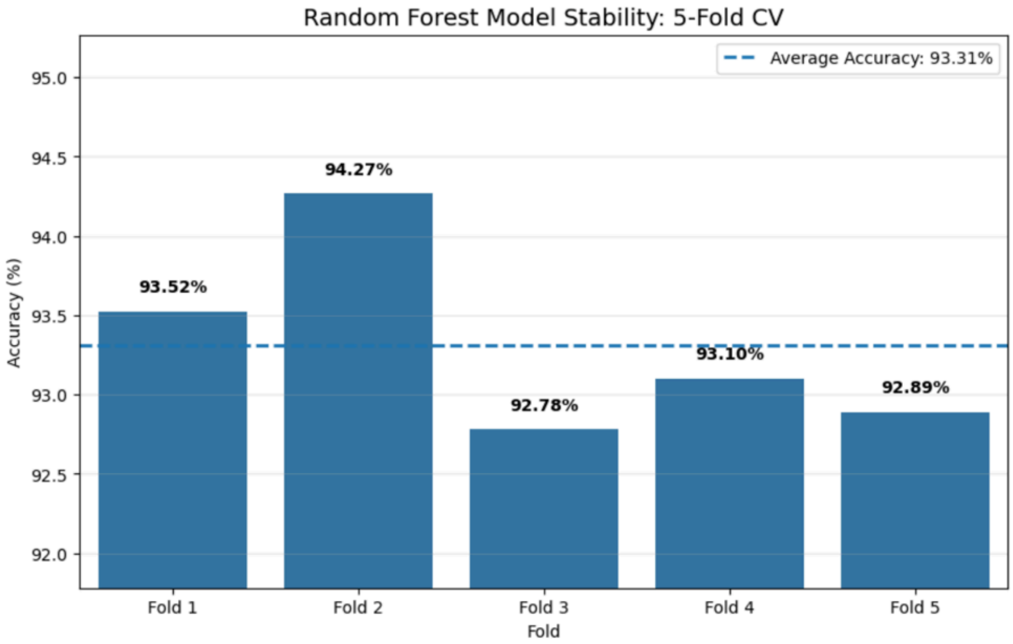
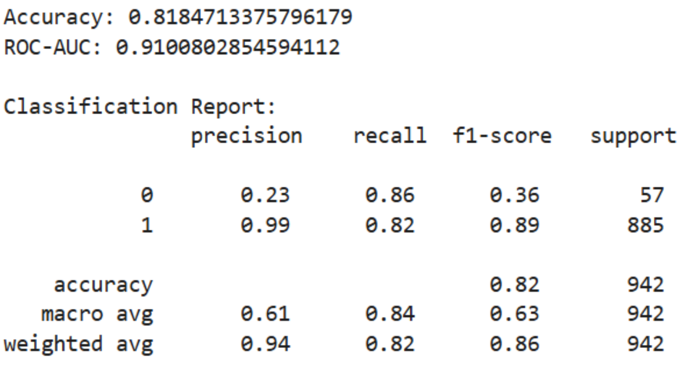


Figure 2.4.2 Distribution of five fold accuracy for Random Forest Model

Models

3.1 Logistic Regression

We first applied Logistic Regression as our baseline model due to its simplicity and interpretability. The model achieved an accuracy of 82% and a ROC–AUC of 0.91, indicating that it is reasonably effective at distinguishing between high-rated and low-rated products.



```
Accuracy: 0.8184713375796179
ROC-AUC: 0.9100802854594112

Classification Report:
               precision    recall  f1-score   support

           0       0.23      0.86      0.36        57
           1       0.99      0.82      0.89       885

    accuracy                           0.82       942
   macro avg       0.61      0.84      0.63       942
weighted avg       0.94      0.82      0.86       942
```

Figure 3.1.1 Classification Report for Logistic Regression

However, the classification report and confusion matrix reveal a clear imbalance in performance across classes. The model predicts high-rated products with extremely high precision (0.99) but performs poorly on low-rated products. This is primarily due to the strong class imbalance in our data, because low-rating products make up only a small fraction of the dataset, leaving the model with too few negative examples to learn meaningful patterns.

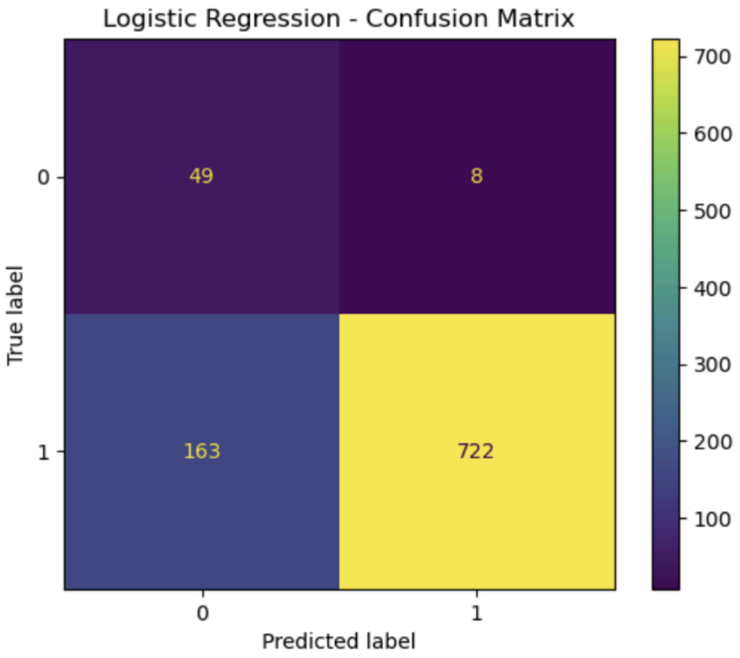


Figure 3.1.2 Confusion Matrix for Logistic Regression

The Precision-Recall curve further highlights this issue. Precision remains very high across most recall thresholds, confirming that the model is overly confident in predicting the majority class (high ratings) while struggling to generalize to the minority class. Despite these limitations, Logistic Regression still provides meaningful insights into feature associations.

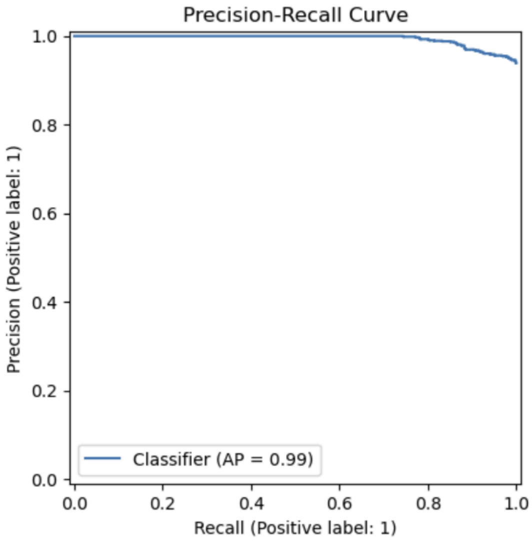
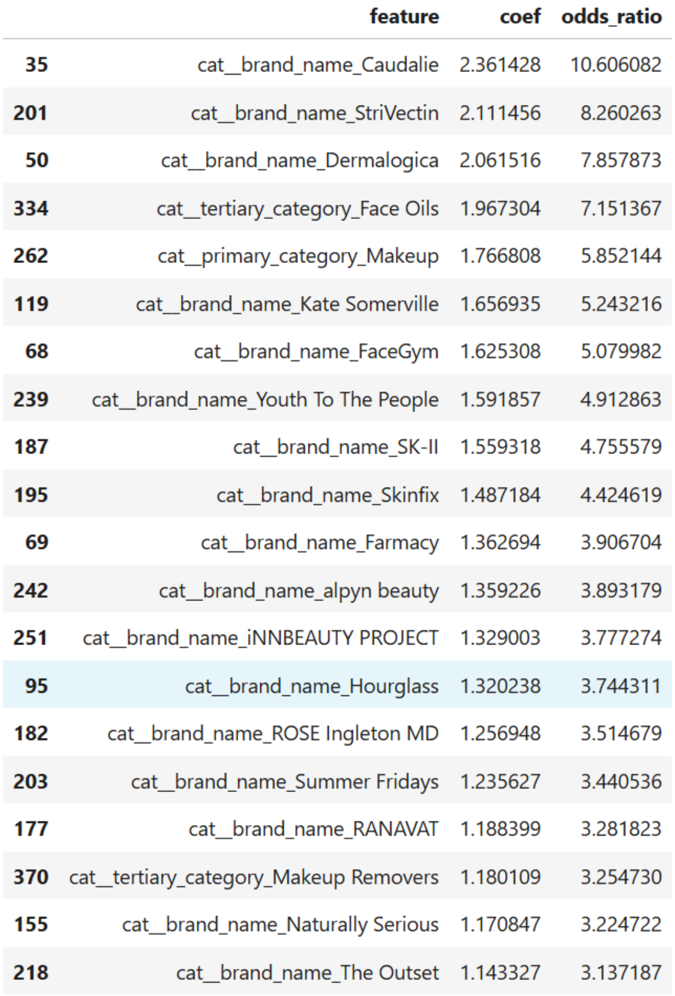


Figure 3.1.3 Precision-Recall Curve

From the coefficient table, we observe strong positive effects from several brands such as Caudalie, StriVectin, and Dermalogica, which may be driven by high customer trust, consistent product performance, and strong brand reputation. In addition, certain tertiary categories, most notably Face Oils and Makeup Removers, show high odds ratios, indicating that products within these subcategories are several times more likely to receive high ratings compared to the baseline category.

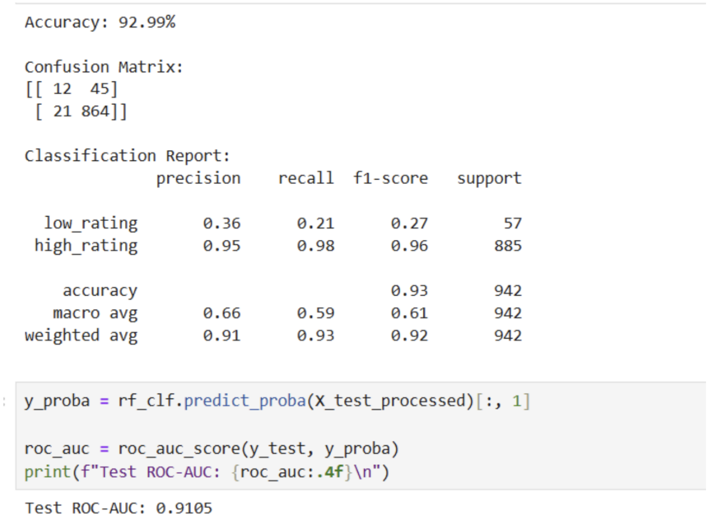
However, Logistic Regression is ultimately limited by its linear structure. It cannot adequately capture the complex, non-linear relationships and interactions present in our high-dimensional, one-hot encoded dataset. These limitations motivate our transition to more flexible models, specifically Random Forest, for deriving the final results.



	feature	coef	odds_ratio
35	cat__brand_name_Caudalie	2.361428	10.606082
201	cat__brand_name_StriVectin	2.111456	8.260263
50	cat__brand_name_Dermalogica	2.061516	7.857873
334	cat__tertiary_category_Face Oils	1.967304	7.151367
262	cat__primary_category_Makeup	1.766808	5.852144
119	cat__brand_name_Kate Somerville	1.656935	5.243216
68	cat__brand_name_FaceGym	1.625308	5.079982
239	cat__brand_name_Youth To The People	1.591857	4.912863
187	cat__brand_name_SK-II	1.559318	4.755579
195	cat__brand_name_Skinfix	1.487184	4.424619
69	cat__brand_name_Farmacy	1.362694	3.906704
242	cat__brand_name_alpyn beauty	1.359226	3.893179
251	cat__brand_name_iNNBEAUTY PROJECT	1.329003	3.777274
95	cat__brand_name_Hourglass	1.320238	3.744311
182	cat__brand_name_ROSE Ingleton MD	1.256948	3.514679
203	cat__brand_name_Summer Fridays	1.235627	3.440536
177	cat__brand_name_RANAVAT	1.188399	3.281823
370	cat__tertiary_category_Makeup Removers	1.180109	3.254730
155	cat__brand_name_Naturally Serious	1.170847	3.224722
218	cat__brand_name_The Outset	1.143327	3.137187

Figure 3.1.4 Logistic Regression Coefficients and Odds Ratios

3.2 Random Forest



```
Accuracy: 92.99%

Confusion Matrix:
[[ 12  45]
 [ 21 864]]

Classification Report:
              precision    recall  f1-score   support

  low_rating       0.36      0.21      0.27        57
 high_rating       0.95      0.98      0.96       885

    accuracy                           0.93       942
   macro avg       0.66      0.59      0.61       942
weighted avg       0.91      0.93      0.92       942
```

```
y_proba = rf_clf.predict_proba(X_test_processed)[:, 1]

roc_auc = roc_auc_score(y_test, y_proba)
print(f"Test ROC-AUC: {roc_auc:.4f}\n")
```

```
Test ROC-AUC: 0.9105
```

Figure 3.2.1 Classification Report for Random Forest

The model was first trained on our training set and subsequently evaluated on a separate test set, where the model reached an accuracy of about 92.99%. As shown in *Figure 3.2.1*, the F-1 score for the ‘high_rating’ class is 0.96, indicating that the model is highly effective at correctly identifying products with high ratings. In contrast, the F-1 score for the ‘low_rating’ class is substantially lower at 0.27, demonstrating that the model has difficulty detecting products with low ratings. This disparity reflects the underlying class imbalance within the dataset, where the number of highly-rated products significantly outnumbers the low-rated ones. However, the macro F1 score of 0.61 provides a more balanced perspective, revealing considerable variation in performance across the two classes. These results highlight the limitations of relying solely on accuracy and underscores the need for additional strategies to address class imbalance.

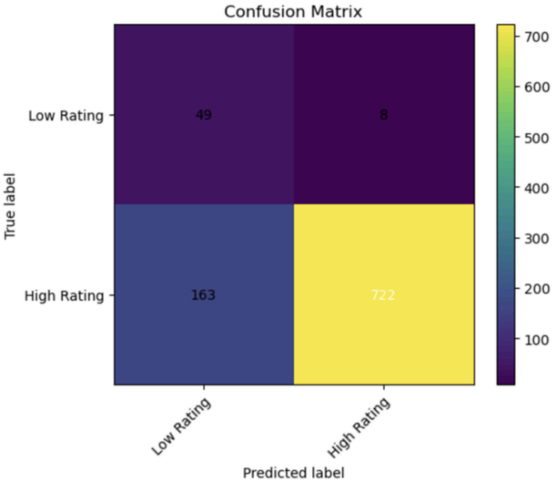


Figure 3.2.2 Confusion Matrix for Random Forest

The confusion matrix above indicates that the model is highly effective at identifying high-rating products, correctly classifying 722 of them. This strong performance is particularly relevant to the central objective of our research, which focuses on uncovering the product features that drive high customer ratings.

Although the model shows lower accuracy on the minority low-rating class, this does not hinder our primary objective. Since the model consistently recognizes high-rated products, the patterns it learns from the data provide meaningful and reliable insight into the factors most strongly associated with achieving high ratings.

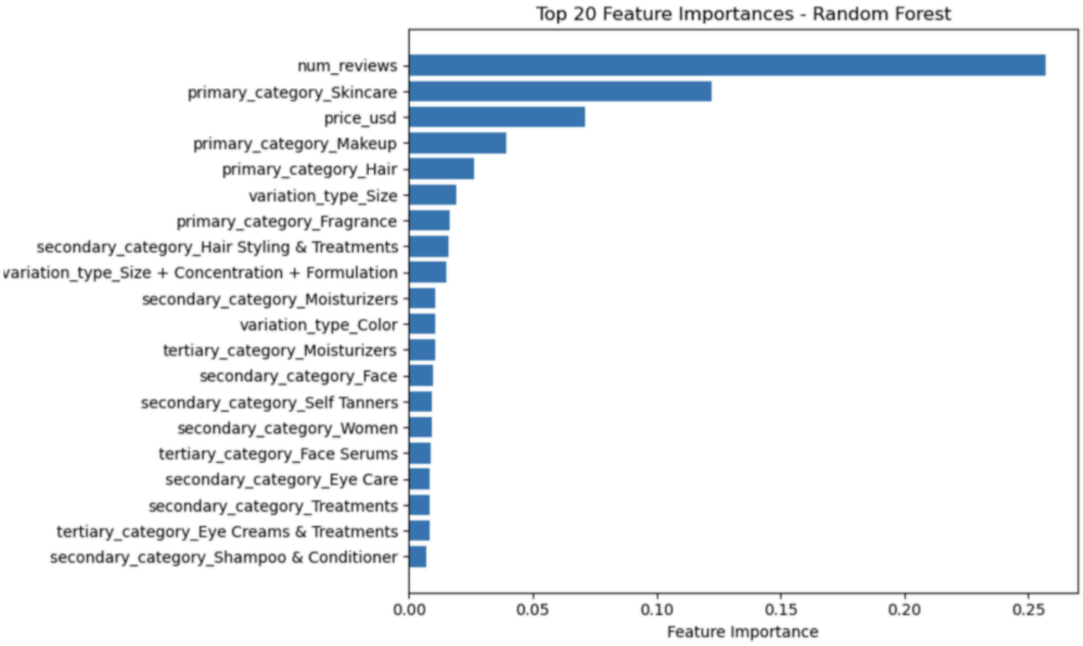


Figure 3.2.3 Top 20 Feature Importance

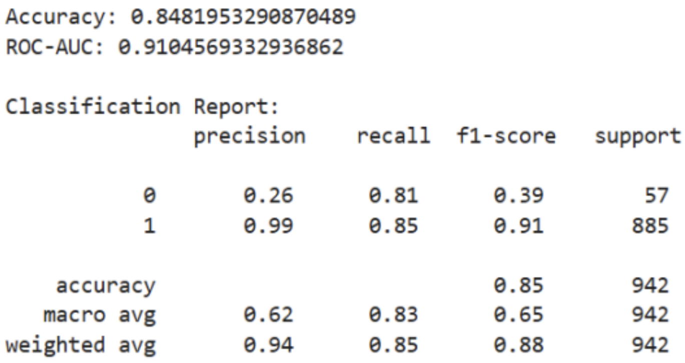
The feature importance plot above further clarifies which variables have the greatest influence on predicting whether a product receives a high rating (4 or 5 stars). The number of reviews stands out as the single most influential feature. This does not mean that more reviews cause higher ratings; rather, it means that review count helps the model distinguish between products with stable, reliable ratings and those with only a few early, potentially biased reviews. Products with low review counts tend to have more volatile ratings, whereas high-review products tend to have more consistent ones. As a result, the Random Forest model relies heavily on the number of reviews to structure its splits and improve prediction accuracy.

Following review count, the next most important features include the primary product category, especially whether the item falls within skincare and the product's price. This shows that the model depends on broader product characteristics and customer engagement signals to make accurate predictions, while more detailed attributes such as secondary and tertiary categories play a secondary role.

Results and Analysis

4. Benchmark

Because our dataset contains a significant class imbalance with high-rating products vastly outnumbering low-rating products, we applied additional methods to see if it improves model performance. We first implemented SMOTE, which is Synthetic Minority Oversampling Technique) for both the Logistic Regression and Random Forest models. SMOTE generates synthetic examples of the minority class, allowing the model to reduce bias toward the majority class.



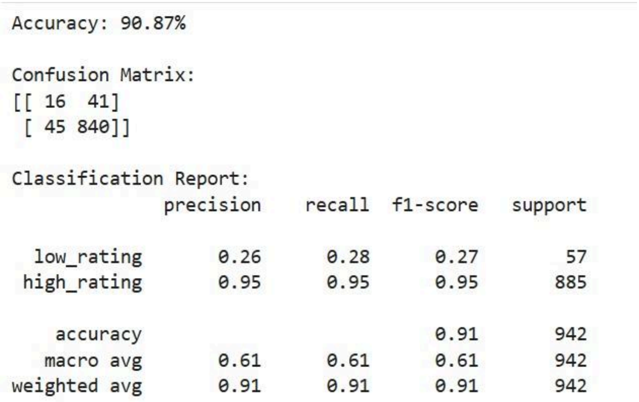
```
Accuracy: 0.8481953290870489
ROC-AUC: 0.9104569332936862

Classification Report:
              precision    recall  f1-score   support

           0       0.26      0.81      0.39        57
           1       0.99      0.85      0.91       885

    accuracy                           0.85       942
   macro avg       0.62      0.83      0.65       942
weighted avg       0.94      0.85      0.88       942
```

Figure 4.1 Logistic Regression after SMOTE

As you can see in *Figure 4.1*, SMOTE slightly improved the performance of logistic regression by balancing the classes and helping the model learn a more stable decision boundary. Accuracy increased from 81.8% to 84.8%, and the minority-class F1-score improved modestly. However, the ROC-AUC remained nearly the same, indicating that the overall discriminative power of the model did not fundamentally change.

Below is a summary of all of our **Logistic Regression** models' performances:

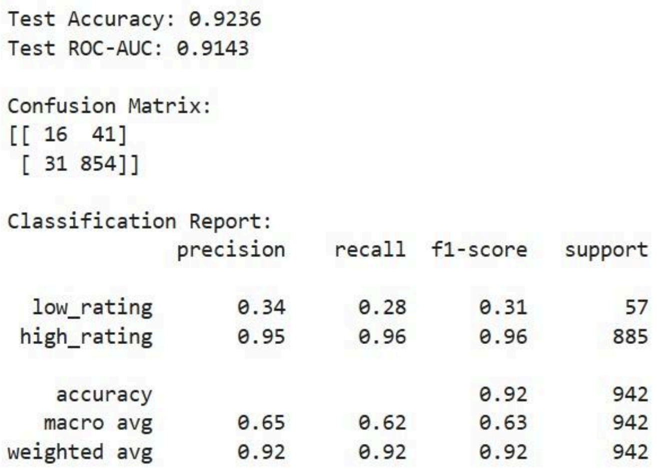
Method for low-rating products	Precision	Recall	F1 Score	ROC-AUC
Logistic Regression (Balanced Class Weights)	0.23	0.86	0.36	0.91
Logistic Regression (SMOTE)	0.26	0.81	0.39	0.91



```
Accuracy: 90.87%

Confusion Matrix:
[[ 16  41]
 [ 45 840]]

Classification Report:
              precision    recall  f1-score   support

  low_rating       0.26      0.28      0.27        57
 high_rating       0.95      0.95      0.95       885

    accuracy                           0.91       942
   macro avg       0.61      0.61      0.61       942
weighted avg       0.91      0.91      0.91       942
```

Figure 4.2 Random Forest after SMOTE

In contrast, as shown in *Figure 4.2,* Random Forest did not benefit from SMOTE in our experiments. Because Random Forest already handles class imbalance naturally and relies on nonlinear splits, synthetic oversampling created unrealistic examples and degraded performance.



```
Test Accuracy: 0.9236
Test ROC-AUC: 0.9143

Confusion Matrix:
[[ 16  41]
 [ 31 854]]

Classification Report:
              precision    recall  f1-score   support

  low_rating       0.34      0.28      0.31        57
 high_rating       0.95      0.96      0.96       885

    accuracy                           0.92       942
   macro avg       0.65      0.62      0.63       942
weighted avg       0.92      0.92      0.92       942
```

Figure 4.3 Random Forest after RandomizedSearchCV

To improve the Random Forest model without relying on synthetic oversampling, we turned over to hyperparameter optimization. After applying RandomizedSearchCV, the optimized model delivered slightly more balanced results. The search explored variations in the number of trees (100–500), maximum depth (None to 20), minimum samples required for splits (2–10) and leaf nodes (1–4), as well as different feature-sampling strategies (“sqrt,” “log2,” 0.3, 0.5).

While the overall accuracy (92.36%) and ROC–AUC (0.9143) remained similar, the model improved recall for low-rating products (0.28, up from 0.21) and increased F1-score (0.31, up from 0.27), indicating moderately better detection of minority-class cases without sacrificing performance on the high-rating class (precision = 0.95, recall = 0.96, F1 = 0.96).

This tuning process allowed the model to better capture non-linear interactions in the high-dimensional feature space and improved overall predictive performance. The optimized Random Forest provided the most reliable estimates among all models.

Below is a summary of all the **Random Forest** models' performance:

Method	Precision	Recall	F1 Score	ROC-AUC
Random Forest (Balanced Class Weights)	0.95	0.98	0.96	0.91
Random Forest (SMOTE)	0.99	0.85	0.91	0.91
Random Forest (RandomizedSearchCV)	0.95	0.96	0.96	0.91

5.*Research Question Results*

1. Which brands consistently produce products with the highest predicted rating probability?

Information gained from the logistic regression model indicate that the brand name is a main factor that can be used to foresee whether a Sephora skincare product will get a high rating (4 or 5 stars). To see which popular skincare brands have the highest predicted rating probability, we sorted their odds ratio values estimated from the logistic regression model. As shown in *Figure 5.1.1*, Caudalie is the brand that clearly exceeds an odds ratio of 10, while StriVectin and Dermalogica are the brands with odds ratios almost 8 indicating that they have a high probability of obtaining positive customer ratings as compared to the reference brand ("19-69").The coefficients and odds ratios of the model represent the brands which have the strongest positive association with a higher probability of a customer giving a positive rating. Among the top brands, Caudalie, StriVectin, and Dermalogica are the ones hitting the highest odds ratios with 10.6, 8.3, and 7.9 correspondingly. It implies that a high rating of these brand products will be about 8–10 times more than the reference brand. The "reference" or "base brand" in logistic regression is the one which all other brands are compared to. As determined by the model, it is the brand "19-69" in this case which is the reference for all the other brands. Technically, the odds ratio of each brand tells how many times the chance of getting high ratings is increased (or decreased) compared to the base brand.

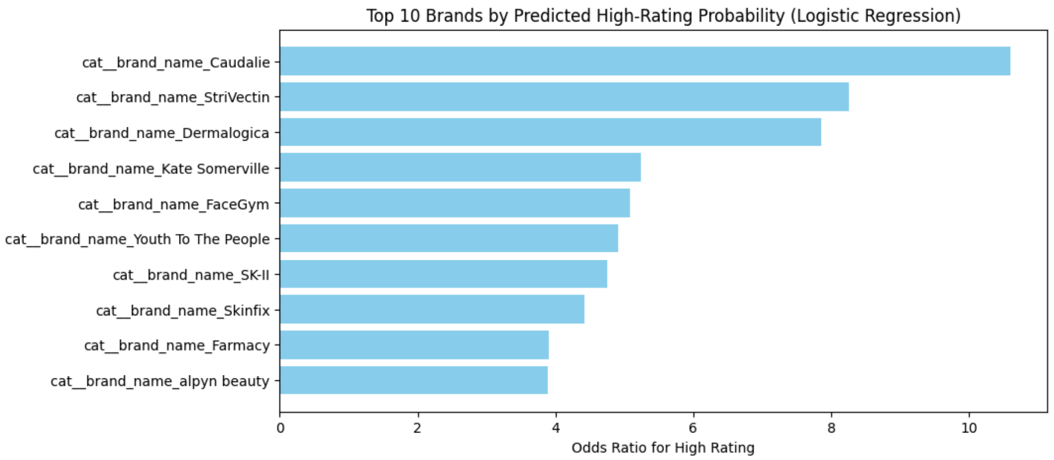


Figure 5.1.1 Top 10 Brands by Predicted High-Rating Probability using Logistic Regression

In addition to that, a number of other brands such as Kate Somerville, FaceGym, Youth To The People, SK-II, and Skinfix were also found to have strong positive coefficients (odds ratios ranging from 4 to 5) that lead to this conclusion. These brands, therefore, can be considered as the ones that always provide excellent customer satisfaction and earn great consumer trust. The general trend points out that prestigious, top-notch skincare brands are likely to receive more positive comments and be considered as more trustworthy.

2. How does the number of reviews influence the probability of a product having a high rating?

The number of reviews has a clear and non-linear influence on the probability of a product receiving a high rating. Most products in the dataset fall into the 0–10 review range, which creates a strong long-tail pattern where only a small group of products accumulate hundreds or thousands of reviews (**Figure 5.1.3**). This distribution matters because products with very few reviews tend to exhibit inflated positivity—often due to early enthusiastic reviewers—resulting in an artificially high predicted probability of receiving a 4–5 star rating.

As products begin accumulating more feedback, the model's predicted probability adjusts. Products in the 11–50 review range show a sharp drop in probability because the broader mix of ratings provides a more realistic assessment of product quality. However, once a product surpasses this early stage, the relationship becomes positive again. Products with 50 to 500 reviews show steadily increasing predicted probabilities, and those with more than 1,000 reviews regain a high predicted probability, reflecting that consistently popular products tend to maintain strong ratings over time (**Figure 5.1.2**).

Overall, review volume acts as a stabilizing indicator of quality. Few reviews inflate positivity, moderate reviews introduce balance and reduce predicted probability, and high review counts signal sustained customer satisfaction. This pattern shows that the number of reviews is an important explanatory feature in predicting whether a product is likely to be highly rated

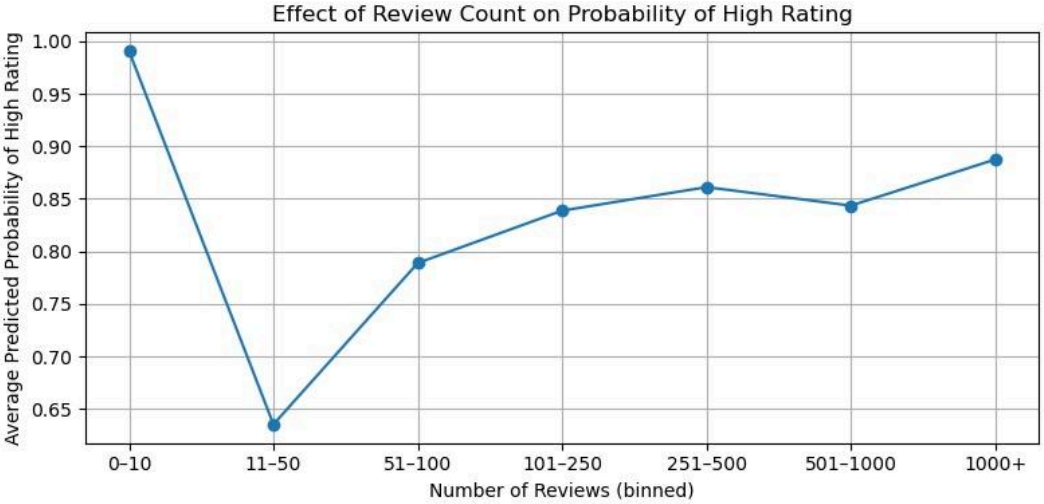


Figure 5.1.2

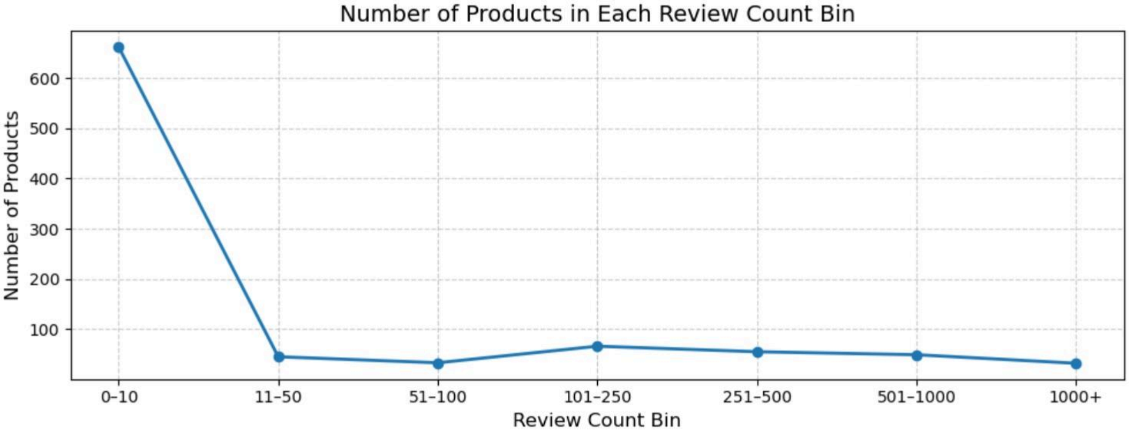


Figure 5.1.3

3. Does the variation type (full size, mini, value set, etc.) significantly influence whether a product receives a high rating?

Results from the logistic regression model indicate that variation type does influence the likelihood of a product receiving a high rating, although the strength and direction of the effect vary by specific format. As shown in *Figure 5.1.4,* the strongest positive effect is observed for products categorized as “Size + Concentration + Formulation,” which show an odds ratio of 2.80, meaning these products are nearly 3 times more likely to receive a high rating compared to the baseline variation type. Products with a simple “Size” variation also increase the likelihood of a high rating, with an odds ratio of 1.54. In contrast, several variation types are associated with lower probabilities of high ratings, including Scent (odds ratio = 0.59), Color (odds ratio = 0.79), and Size + Concentration (odds ratio = 0.71), indicating reduced odds of receiving a 4–5 star rating. The effect strengths for these negative variation types range from 0.21 to 0.41, suggesting moderate practical impact. Overall, these results demonstrate that variation type does have a statistically meaningful influence on high ratings, particularly for formats related to size and formulation, while scented and color-based variations tend to reduce the probability of high customer ratings.

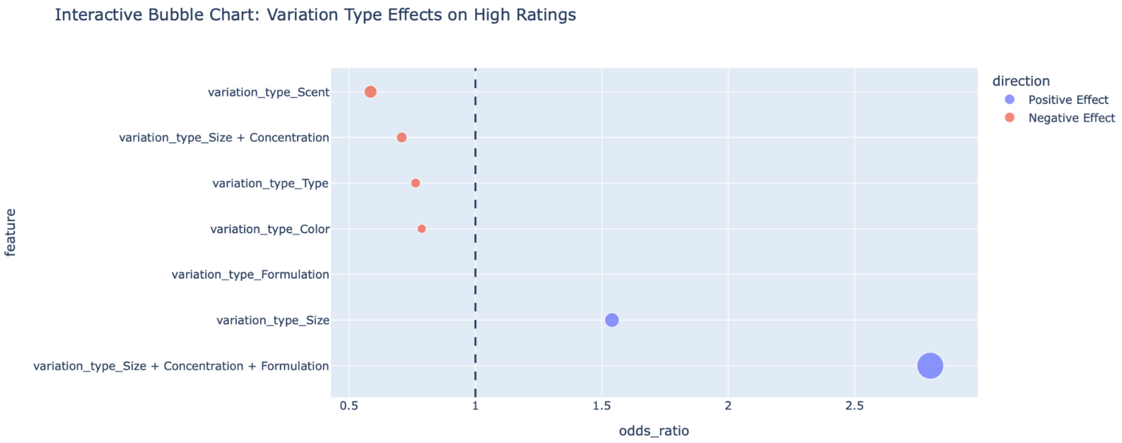
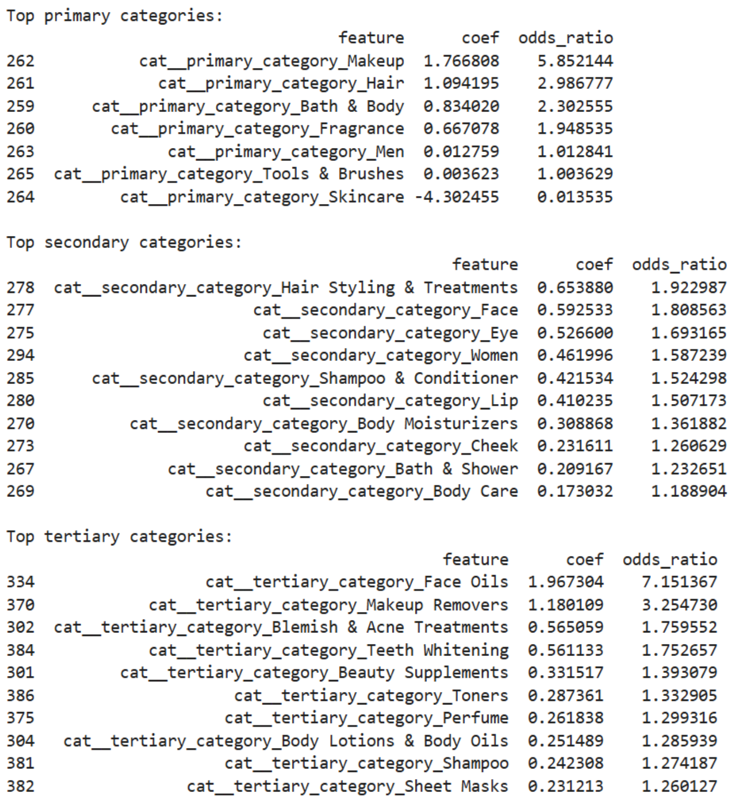


Figure 5.1.4 Bubble Chart of the Odds Ratio of the Variation Type of Products

4. Which product categories (primary, secondary, tertiary) are most associated with high-rated skincare products?

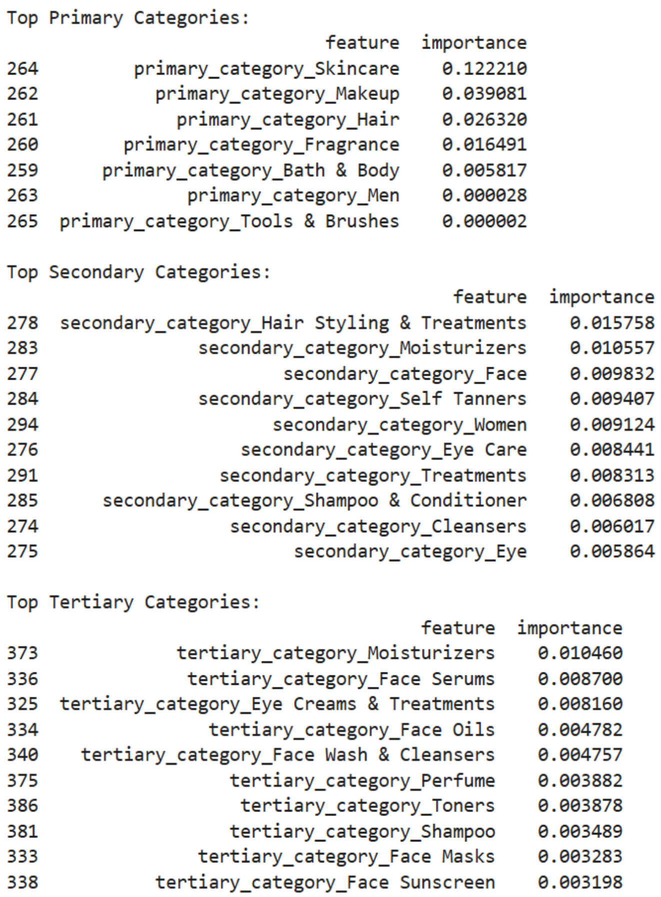
Using Logistic Regression, we examined how different category levels relate to high-rated skincare products. As shown in *Figure 5.1.5* below, at the primary level, Makeup (odds ratio ≈ 5.85) and Hair (≈ 2.99) show positive associations relative to the baseline category. At the secondary level, categories such as Hair Styling & Treatments (≈ 1.92) and Face (≈ 1.81) display elevated coefficients. The tertiary level exhibits the strongest effects, with Face Oils (≈ 7.15) and Makeup Removers (≈ 3.25) showing several-fold increases in the likelihood of receiving high ratings. Several brands, including Caudalie (≈ 10.6), StriVectin (≈ 8.26), and Dermalogica (≈ 7.86), also demonstrate substantial positive associations. While informative, these results are constrained by the linear assumptions of Logistic Regression and its difficulty modeling high-dimensional interactions.



```
Top primary categories:
                                  feature      coef  odds_ratio
262           cat__primary_category_Makeup  1.766808    5.852144
261             cat__primary_category_Hair  1.094195    2.986777
259      cat__primary_category_Bath & Body  0.834020    2.302555
260        cat__primary_category_Fragrance  0.667078    1.948535
263              cat__primary_category_Men  0.012759    1.012841
265  cat__primary_category_Tools & Brushes  0.003623    1.003629
264         cat__primary_category_Skincare -4.302455    0.013535

Top secondary categories:
                                              feature      coef  odds_ratio
278  cat__secondary_category_Hair Styling & Treatments  0.653880    1.922987
277                      cat__secondary_category_Face  0.592533    1.808563
275                       cat__secondary_category_Eye  0.526600    1.693165
294                     cat__secondary_category_Women  0.461996    1.587239
285      cat__secondary_category_Shampoo & Conditioner  0.421534    1.524298
280                       cat__secondary_category_Lip  0.410235    1.507173
270          cat__secondary_category_Body Moisturizers  0.308868    1.361882
273                     cat__secondary_category_Cheek  0.231611    1.260629
267              cat__secondary_category_Bath & Shower  0.209167    1.232651
269                  cat__secondary_category_Body Care  0.173032    1.188904

Top tertiary categories:
                                             feature      coef  odds_ratio
334                  cat__tertiary_category_Face Oils  1.967304    7.151367
370            cat__tertiary_category_Makeup Removers  1.180109    3.254730
302  cat__tertiary_category_Blemish & Acne Treatments  0.565059    1.759552
384            cat__tertiary_category_Teeth Whitening  0.561133    1.752657
301         cat__tertiary_category_Beauty Supplements  0.331517    1.393079
386                     cat__tertiary_category_Toners  0.287361    1.332905
375                    cat__tertiary_category_Perfume  0.261838    1.299316
304   cat__tertiary_category_Body Lotions & Body Oils  0.251489    1.285939
381                    cat__tertiary_category_Shampoo  0.242308    1.274187
382                cat__tertiary_category_Sheet Masks  0.231213    1.260127
```

Figure 5.1.5 Random Forest Coefficients and Odds Ratios

Random Forest provides a more accurate and flexible assessment of category importance. Based on *Figure 5.1.6* below, at the primary level, Skincare is the most influential predictor of high ratings (importance ≈ 0.12), followed by Makeup (≈ 0.039) and Hair (≈ 0.026). Secondary-level importance scores highlight Hair Styling and Treatments (≈ 0.0158), Moisturizers (≈ 0.0106), and Face (≈ 0.0098). Tertiary-level predictors such as Moisturizers (≈ 0.0105), Face Serums (≈ 0.0087), and Eye Creams and Treatments (≈ 0.0082) further indicate that hydration- and treatment-focused categories play leading roles in driving positive ratings. These rankings are more stable and better reflect the underlying non-linear structure of the data.



```
Top Primary Categories:
                              feature  importance
264        primary_category_Skincare    0.122210
262          primary_category_Makeup    0.039081
261            primary_category_Hair    0.026320
260        primary_category_Fragrance    0.016491
259      primary_category_Bath & Body    0.005817
263             primary_category_Men    0.000028
265  primary_category_Tools & Brushes    0.000002

Top Secondary Categories:
                                          feature  importance
278  secondary_category_Hair Styling & Treatments    0.015758
283               secondary_category_Moisturizers    0.010557
277                       secondary_category_Face    0.009832
284               secondary_category_Self Tanners    0.009407
294                      secondary_category_Women    0.009124
276                   secondary_category_Eye Care    0.008441
291                 secondary_category_Treatments    0.008313
285     secondary_category_Shampoo & Conditioner    0.006808
274                   secondary_category_Cleansers    0.006017
275                        secondary_category_Eye    0.005864

Top Tertiary Categories:
                                     feature  importance
373            tertiary_category_Moisturizers    0.010460
336             tertiary_category_Face Serums    0.008700
325  tertiary_category_Eye Creams & Treatments    0.008160
334               tertiary_category_Face Oils    0.004782
340    tertiary_category_Face Wash & Cleansers    0.004757
375                 tertiary_category_Perfume    0.003882
386                  tertiary_category_Toners    0.003878
381                 tertiary_category_Shampoo    0.003489
333              tertiary_category_Face Masks    0.003283
338          tertiary_category_Face Sunscreen    0.003198
```

Figure 5.1.6 Random Forest Feature Importance by Category Level

Although Logistic Regression offers interpretable associations, its linear form limits its ability to capture complex rating patterns. Random Forest achieves higher predictive performance and produces more reliable category rankings across all levels. Therefore, Random Forest is selected as the final basis for identifying which primary, secondary, tertiary, and brand categories are most strongly associated with high-rated skincare products.

Conclusion

6.1 Recommendations

Based on the findings of our analysis, we propose two primary recommendations for Sephora.

Recommendation 1: Enhance Promotion of High-Performing Brands

Sephora can strategically highlight and promote high-performing brands. This may include the following:

- Enhance homepage visibility
- Curated “Highly Rated” product collections
- Targeted email or app-based marketing campaigns that feature top-performing products

Recommendation 2: Strengthen Review Generation to Improve Rating Reliability

Sephora can implement initiatives to encourage even more review generation, as many products currently receive very few reviews, resulting in unstable and less reliable ratings. This can be addressed by the following:

- Offer post-purchase incentives
- Send in-app review reminders
- Develop early-reviewer programs to increase review volume and strengthen rating consistency

6.2 Limitations

Despite the value of these insights, several limitations should be acknowledged. The majority of products in the dataset have ratings clustered between 4 and 5 stars, creating a skewed distribution that makes it challenging for models to learn the characteristics associated with lower ratings. In addition, the one-hot encoding process produced a high-dimensional features space, particularly for brand and category variables, which adds complexity and may reduce model interpretability and stability. The imbalance in review volume also presents challenges, as most products receive between 0 and 10 reviews while only a small subset have hundreds or thousands. This uneven distribution makes it difficult for models to generalize reliably across all products.

Furthermore, although we applied SMOTE and hyperparameter optimization through RandomizedSearchCV to mitigate class imbalance and improve performance, the results remained largely consistent with the baseline models. This suggests that the underlying data limitations of the strong rating skew constrain the extent to which model performance can be meaningfully improved.

Finally, product ratings are inherently subjective and may be influenced by brand reputation, customer expectations, or marketing exposure, meaning that a high rating does not necessarily reflect objective product quality.

Overall, combining Logistic Regression with Random Forest allowed us to examine both predictive performance and interpretable feature patterns, offering meaningful insights into the attributes most strongly associated with highly rated products on Sephora, even within the constraints of the dataset.

References

NadyInky. (2023). *Sephora Products and Skincare Reviews*. Kaggle. https://www.kaggle.com/datasets/nadyinky/sephora-products-and-skincare-reviews