Sephora Product & Review Analysis

Presented by Shivani Vallamdas, Vishal Srivastava, Christie Shin, Shuai Zhao, Gema Zhu

BANA 212: Machine Learning Analytics

Table of Contents

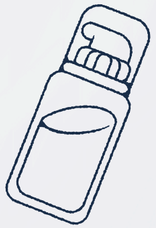
Research Focus

Question: How can Sephora leverage product attributes, number of reviews, and ratings to better promote top brands and build customer trust?

Why it’s important: It helps Sephora promote the right products, elevate trustworthy brands, and reduce risk in merchandising decisions.

Research Questions

Q1 Which brands consistently produce products with the highest predicted rating probability?



Q2 How does the number of reviews influence the probability of a product having a high rating?

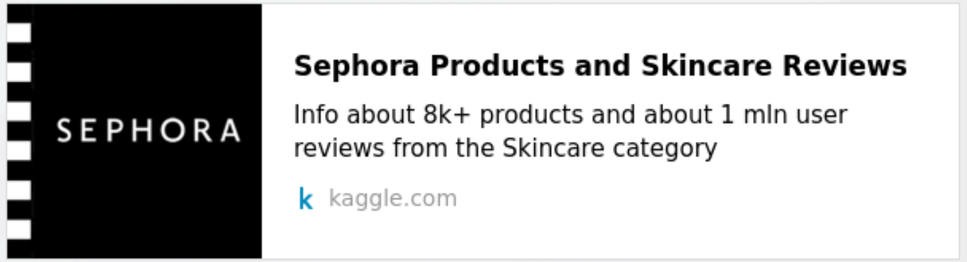
Dataset

Data sourced from Kaggle, titled “Sephora Products and Skincare Reviews”

Collected via PythonScraper in March 2023 and contains the following:

- **Product Data:** Info about all beauty products (over 8,000 products) including product and brand names, prices, ingredients, and ratings

- **Review Data:** User reviews (about 1 million reviews on over 2,000 products) of all products from the Skincare category, including user appearances, and review ratings



Features: ingredients, price, brand, size, primary_category, secondary_category, tertiary_category, variation_type, reviews, highlights, new (product), sephora_exclusive

Data Cleaning

1. Removed all rows that had a missing review from the reviews data frame (1444 rows)

2. Removed all rows that had missing values from the necessary columns from the products data frame (3,784 rows)

3. Reset index for both dataframes

4. Visualized relevant categories of the data frame to better understand the data

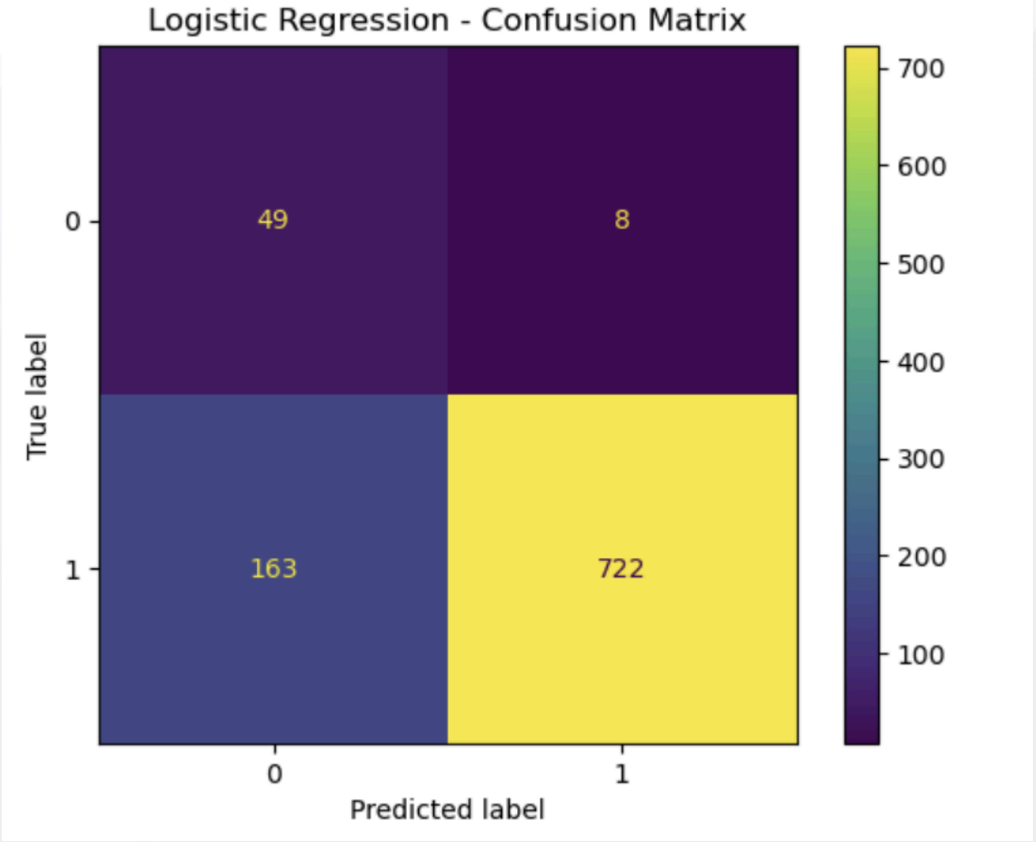
5. Split data into 80% training and 20% testing for ML models

Final number of rows in data frames:

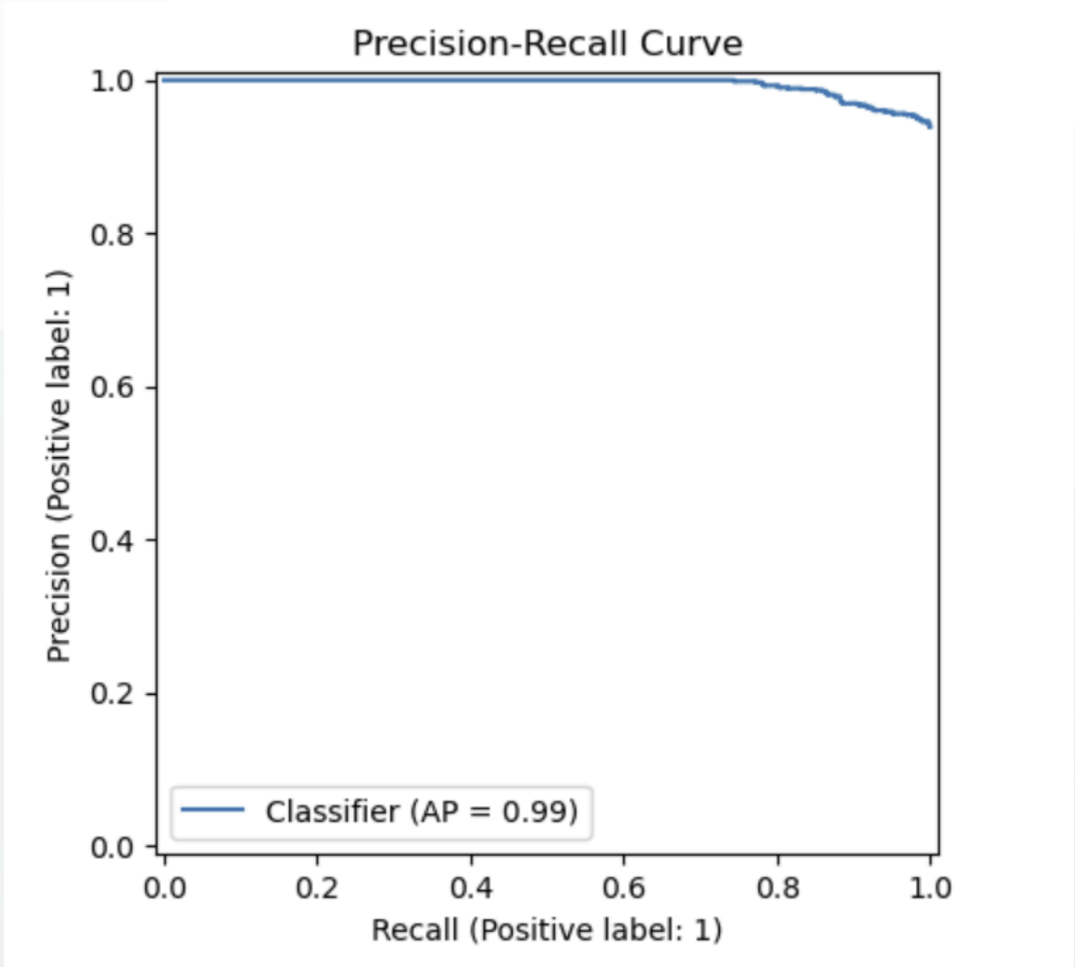
Product Data: 4,710 rows
Review Data: 1,092,967 rows

Logistic Regression

```
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
```



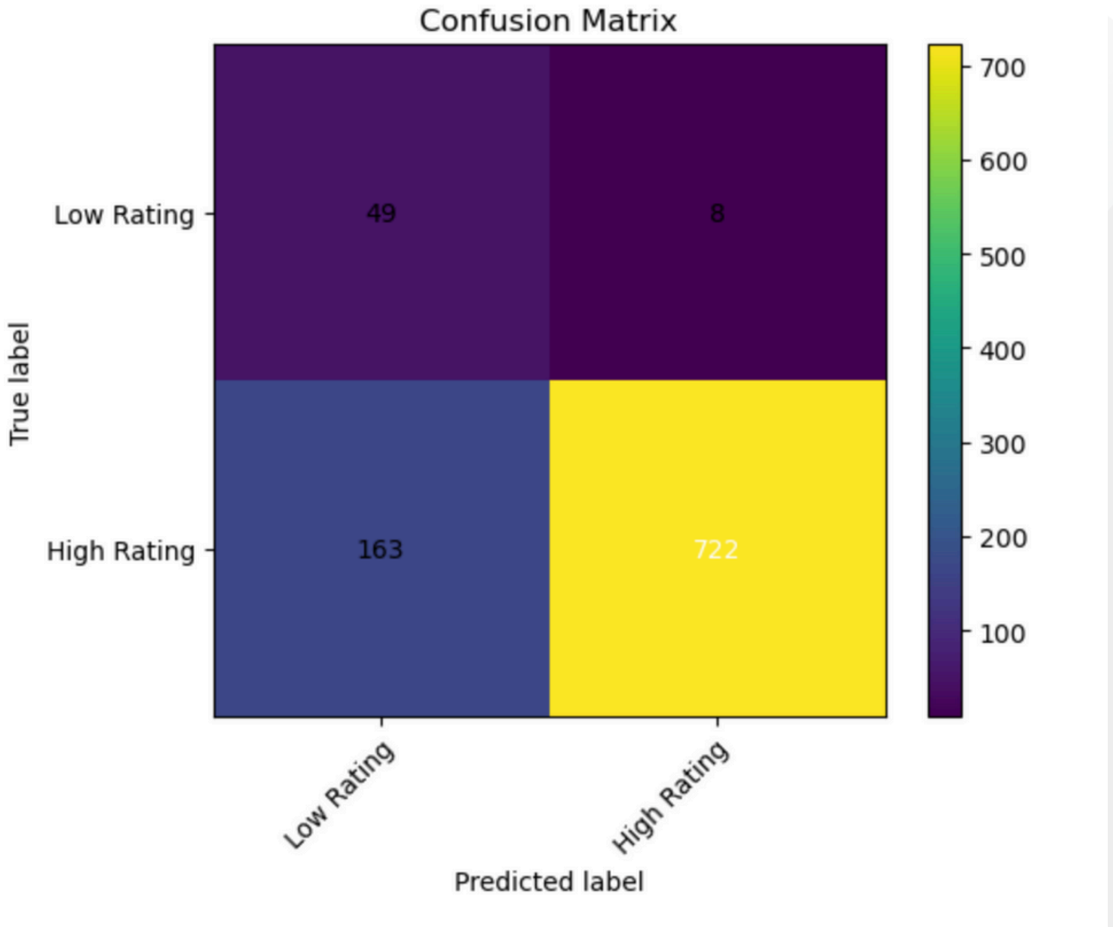
Logistic Regression



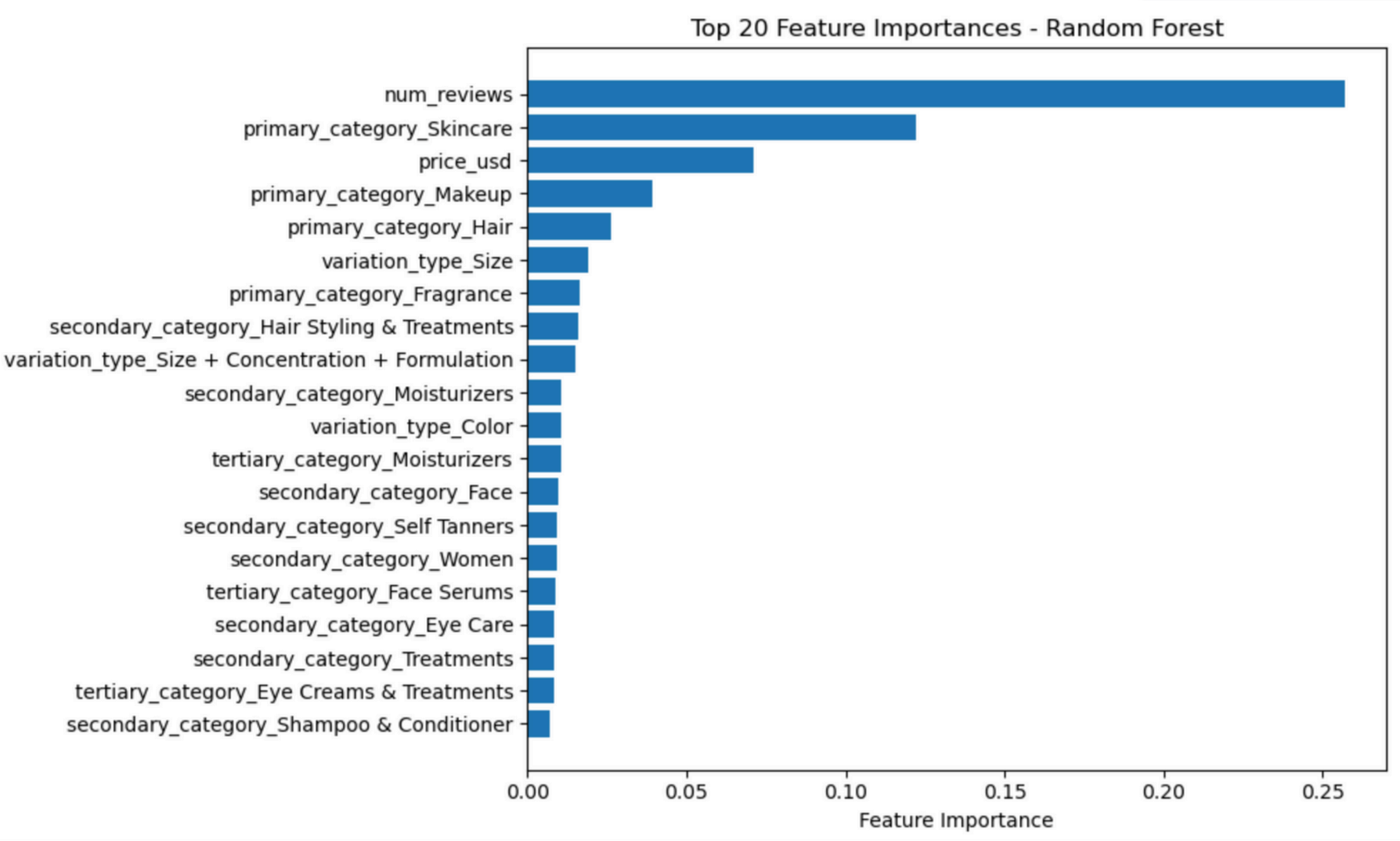
	feature	coef	odds_ratio
35	cat__brand_name_Caudalie	2.361428	10.606082
201	cat__brand_name_StriVectin	2.111456	8.260263
50	cat__brand_name_Dermalogica	2.061516	7.857873
334	cat__tertiary_category_Face Oils	1.967304	7.151367
262	cat__primary_category_Makeup	1.766808	5.852144
119	cat__brand_name_Kate Somerville	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

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
```

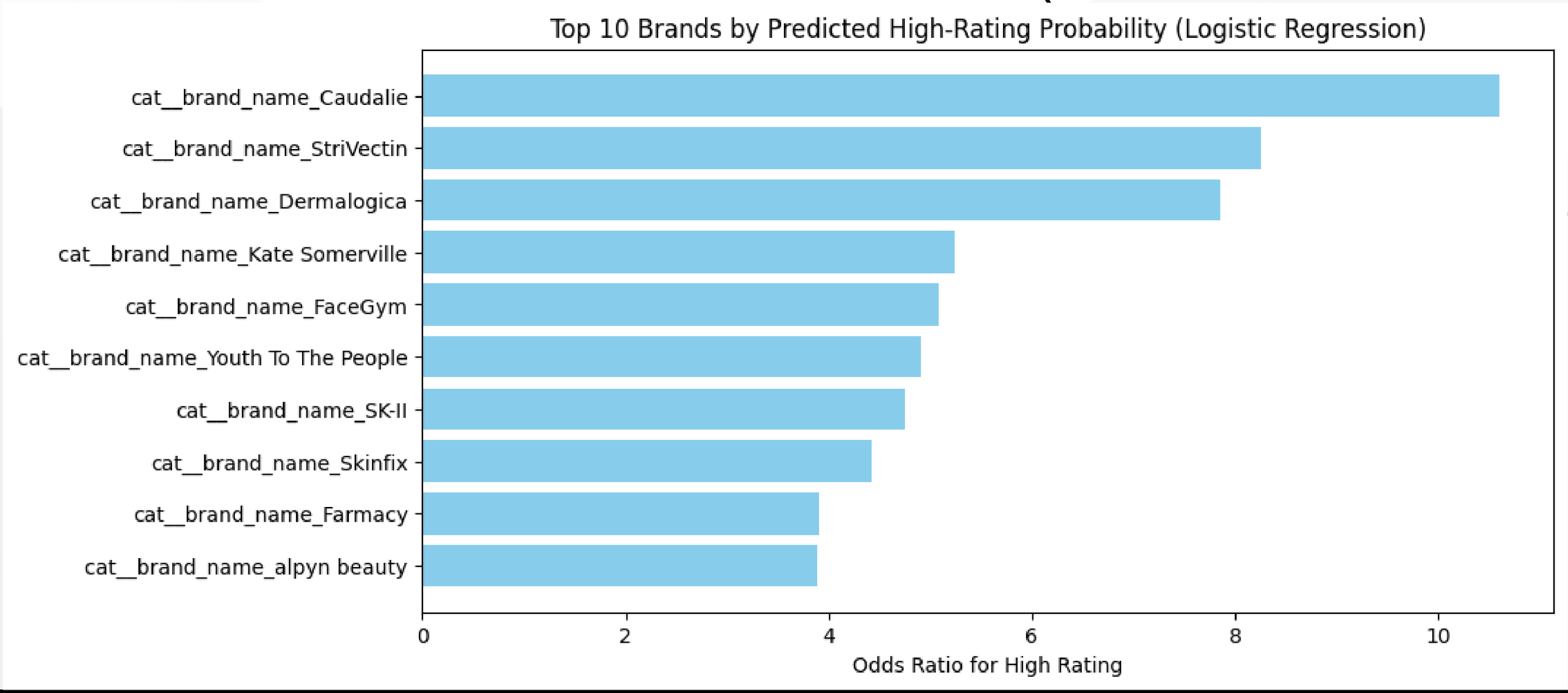


Random Forest



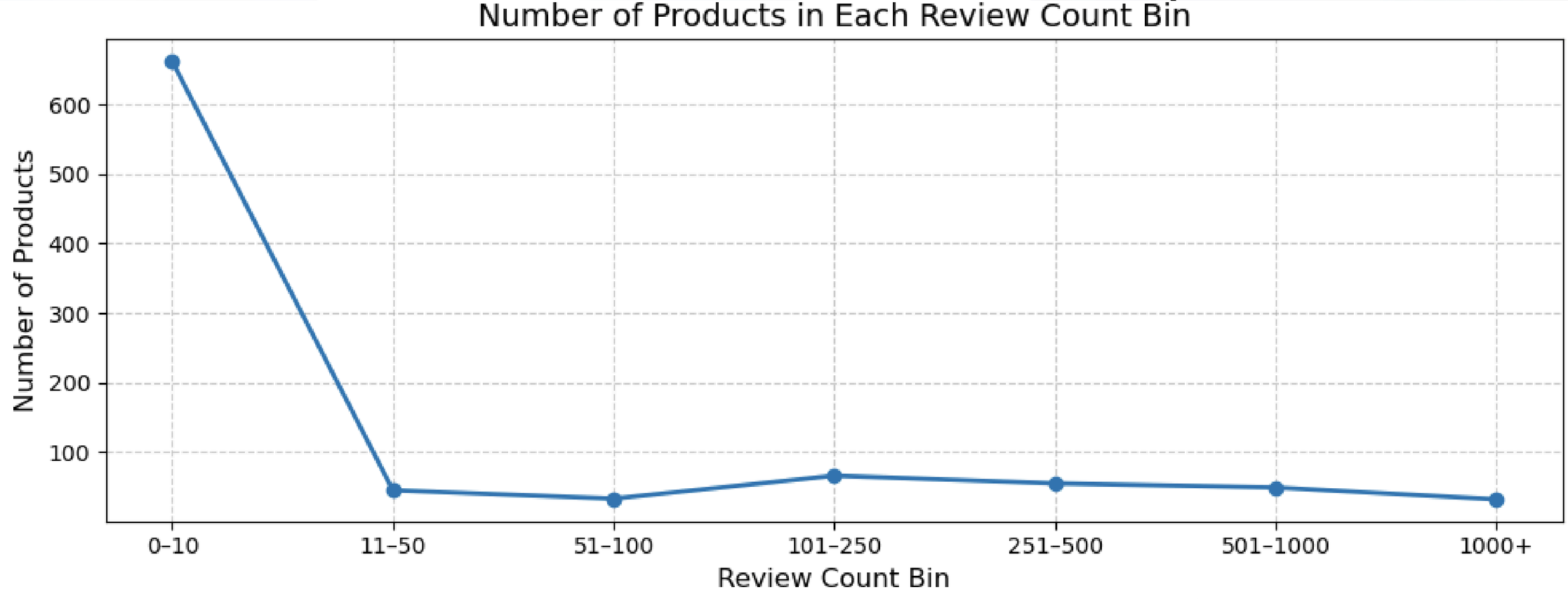
Q1: Results

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



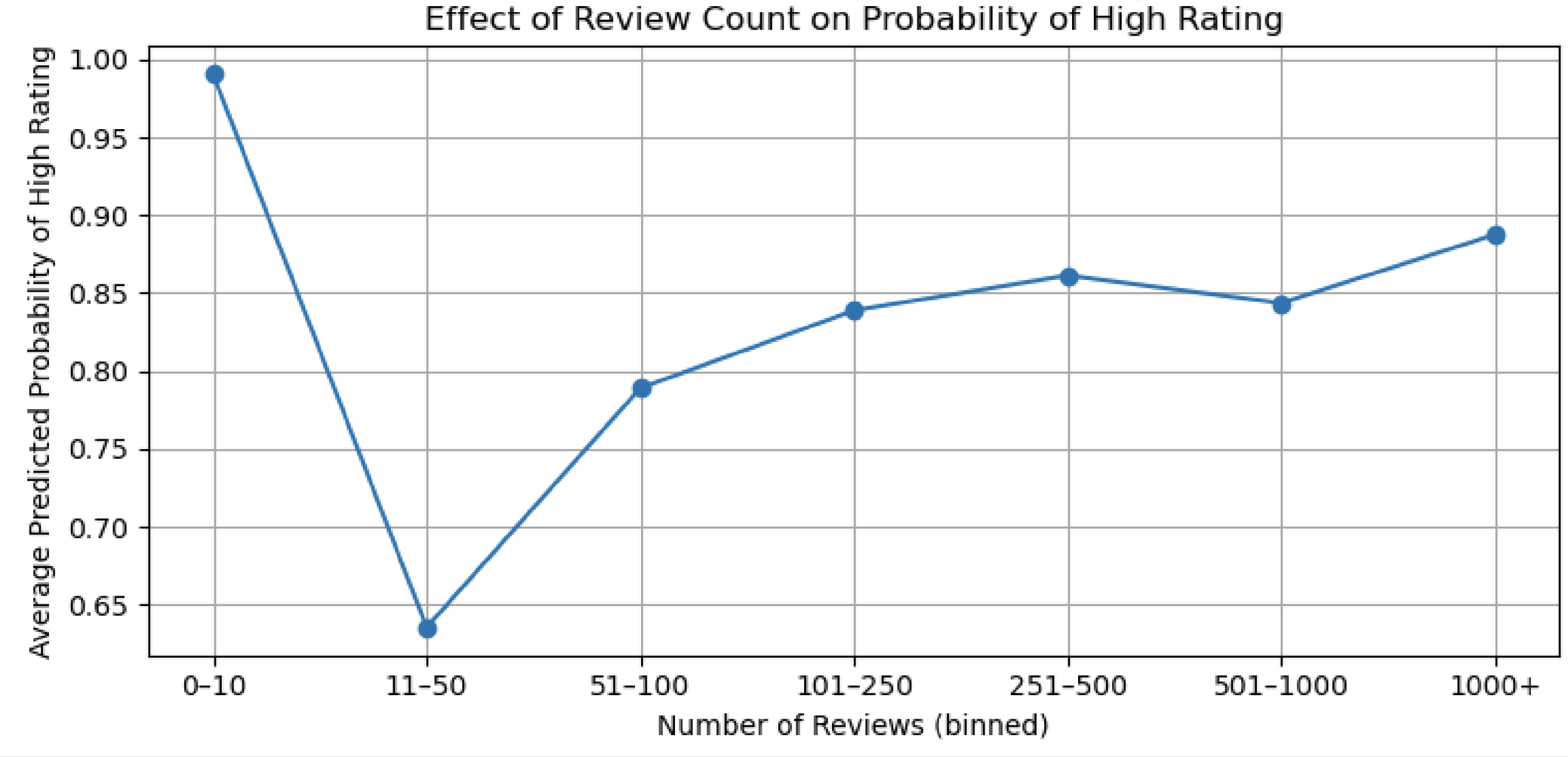
Q2: Results

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



Q2: Results

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



Recommendations

1. **Prioritize High-Performing Brands in Merchandising & Marketing**
    a. Feature top-performing brands more prominently in:
        i. Homepage placements
        ii. “Best of Skincare”/”Highly - Rated” collections
        iii. Email & App marketing campaigns

2. **Actively Pursue Review Generation from Customers**
    a. Motivate customers to leave reviews by:
        i. Post-purchase review incentives (points, samples)
        ii. Early access programs for reviewers
        iii. Pre-generated survey review prompts for under-reviewed products

Limitations

- Majority of the ratings are between 4-5 thus leading to a skewed model
- One-hot encoding creates very high-dimensional features, making the model harder to interpret and potentially less stable
- Most products fall in the 0-10 review range while only a small number receive hundreds or thousands of reviews.This imbalance directly affects how the model interprets rating probabilities
- Product ratings are subjective and often influenced by brand popularity and consumer expectations, which may not reflect true product quality

Thank You!

Presented by Shivani Vallamdas, Vishal Srivastava,
Christie Shin, Shuai Zhao, Gema Zhu

BANA 212: Machine Learning Analytics