迴歸分析 – Sephora 產品與護膚品評論分析報告

楊書暐 – M132040015

研究動機

購買保養品時，需要根據自己的狀況，去做搜尋查找相似的用戶回饋或詢問專業意見。本次模型想要通過蒐集網頁中資料建立預測系統來提供保養品建議，來省略大量查找的時間。

資料簡介

資料來源：https://www.kaggle.com/datasets/nadyinky/sephora-products-and-skincare-reviews

本研究資料源自於 Sephora 線上商店，包含超過 8,000 種美容產品的資訊（如品牌、產品名稱、價格、成分、評等）及超過 100 萬筆護膚品類用戶評論（包含用戶的外觀特徵和評論評等）。經篩選後，研究主要採用用戶評論相關資料，不包含所有產品。

- 資料欄位：40 欄
- 資料總數：1100554 筆

變數介紹

變數名稱	變數簡介	資料型態
產品類		
product_id	產品的唯一識別碼	類別變數
product_name	產品全名	類別變數
brand_id	品牌的唯一識別碼	*數值變數
brand_name	產品品牌的全名	類別變數

loves_count	將此產品標記為最愛的人數	數值變數
*rating	根據使用者評論得出的產品平均評分	數值變數
reviews	使用者對產品的評論數量	數值變數
Size	產品的大小，可能是盎司、毫升、克、包或其他單位，視產品類型而定	類別變數
variation_type	產品變異參數的類型(例如：尺寸、顏色)	類別變數
variation_value	產品變異參數的特定值（例如：100 mL、Golden Sand）	類別變數
variation_desc	產品變異參數的說明(例如：最白肌膚的色調)	類別變數
Ingredients	產品所含成分的清單，例如：[「產品變異 1:」，「水、甘油」，「產品變異 2:」，「滑石粉、雲母」]，或者如果沒有變異[「水、甘油」]	類別變數
price_usd	產品價格（美元）	數值變數
value_price_usd	產品的潛在節省成本，在網站上的正價旁邊顯示	數值變數
sale_price_usd	產品的銷售價格（美元）	數值變數
limited_edition	表示產品是否為限量版(1 – 真, 0 – 假)	二元變數
New	表示產品是否為新產品(1 - 真, 0 – 假)	二元變數
online_only	表示產品是否只在線上銷售（1 - 真，0 - 假）	二元變數
out_of_stock	表示產品目前是否缺貨（1 - 真，0 - 假）	二元變數
sephora_exclusive	表示產品是否為 Sephora 獨家專賣（1 - 真，0 - 假）	二元變數
Highlights	突顯產品屬性的標籤或	類別變數

	特徵清單（例如['Vegan', 'Matte Finish'])	
primary_category	痕跡導航部分中的第一個類別	類別變數
secondary_category	痕跡導航部分中的第二個類別	類別變數
tertiary_category	痕跡導航部分中的第三個類別	類別變數
child_count	可提供的產品變化數量	數值變數
child_max_price	產品變異中的最高價格	數值變數
child_min_price	產品變異中的最低價格	數值變數
用戶評價類		
author_id	網站上評論作者的唯一識別碼	類別變數
*rating	作者對產品的評分，以 1 到 5 分為標準	計數變數
is_recommended	表示作者是否推薦該產品 (1 - 真, 0 – 假)	二元變數
helpfulness	該評論的所有評分與正面評分的比率：有用性 = 總回饋次數 / 總回饋次數	數值變數
total_feedback_count	使用者針對該評論所留下的回饋總數（正面與負面評價)	數值變數
total_neg_feedback_count	對評論給予負面評價的使用者人數	數值變數
total_pos_feedback_count	對評論給予正面評價的使用者人數	數值變數
submission_time	評論張貼在網站上的日期，格式為 yyyy-mm-dd	日期變數
review_text	作者撰寫的評論正文	類別變數
review_title	作者撰寫的評論標題	類別變數
skin_tone	作者的膚色(例如：白皙、黝黑等)	類別變數
eye_color	作者眼睛的顏色（如棕	類別變數

	色、綠色等）	
skin_type	作者的皮膚類型 (例如混合性、油性等)	類別變數
hair_color	作者的髮色（例如棕色、赤褐色等）	類別變數

目標

此次研究想要根據用戶提供的信息來建立一個對是否推薦(is_recommended)進行預測的模型。

變數處理

此次研究並不包含文本分析的部分，因此有關評論的內容將不會使用。另外"is_recommended"有些許缺失值(附錄)，將會丟棄相對應的資料，並且有些項目的缺失值過多，將會直接丟棄。
丟棄後剩餘資料為：926423 筆

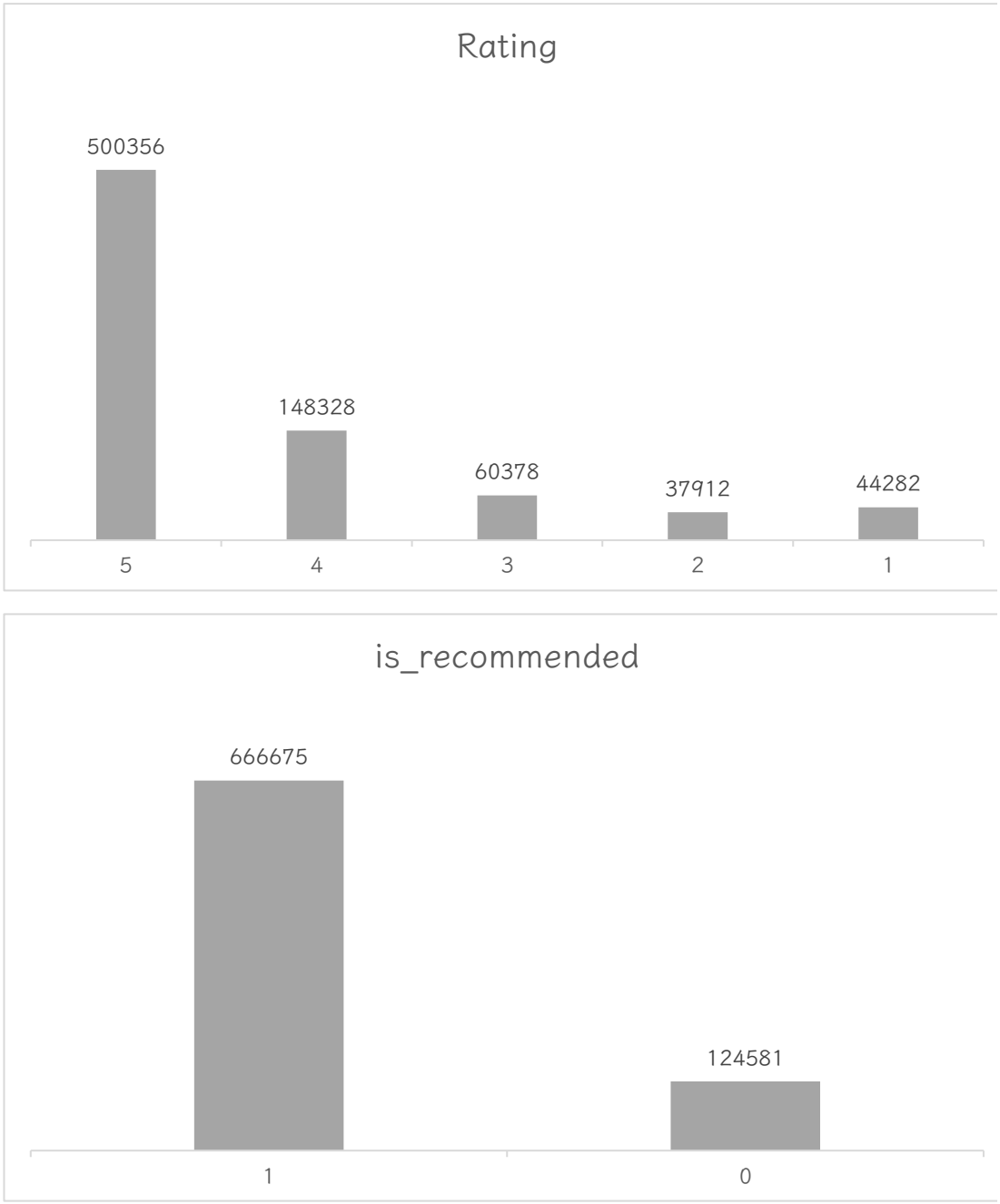
現丟棄變數：

- variation_desc
- sale_price_usd
- value_price_usd
- child_max_price
- child_min_price
- helpfulness
- review_title
- author_id
- review_text

類別變數處理

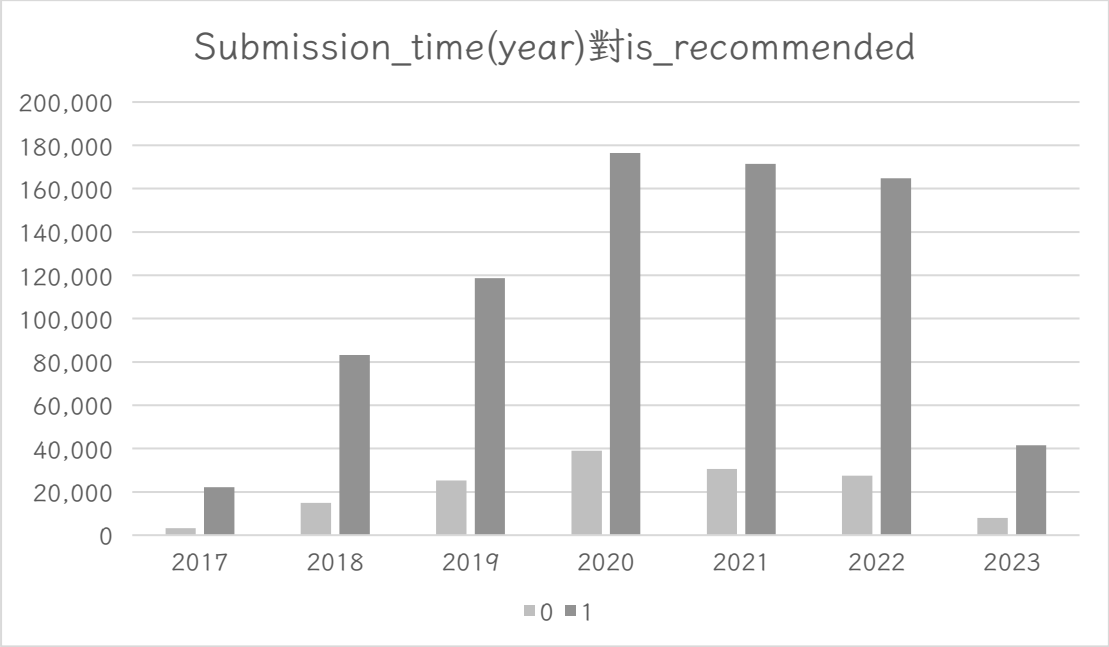
由於我們的目標為“is_recommended”，可以看到多數人傾向推薦商品(圖一)，但是我們的是想要對商品進行推薦評價，因此我們將會在*處理完相關類別變數後，將以商品為根據，將數值及二元變數取平均值。

*確保單一商品的類別變數只會有單一唯一值



(圖一) 用戶對商品評分及是否推薦圖

時間(submission_time)



(圖二) 年份長條圖

從上圖可以看到，對年份來說，推薦與否的趨勢都是呈現出相同狀況，在這裡認為時間對是否推薦的影響不大，因此將不會使用。

商品在網頁中的分類

- primary_category：再刪除完缺失值後，發現只剩下單一值"Skincare"，故無法使用。
- secondary_category：13 類
- tertiary_category：42 類

可以觀察到有些 secondary_category 中是沒有 tertiary_category 的，為了後續處理，將直接刪除對應的列。剩餘的 1971 筆缺失值，由於對資料影響不大，也直接做刪除。在刪除過後，可以發現有一類是*只有一項商品，並且只有 15 則評價，所以也做刪除。

secondary_category	tertiary_category	Product_name
Shop by Concern	Anti-Aging	Lotus Youth Preseve Rescue Mask Mini

*唯一商品

個人特質

變數名稱	唯一值	缺失值
瞳色 eye color	6	35689
膚色 skin tone	14	34495
膚況 skin type	4	15963
髮色 hair color	7	50127

我們最後需要將根據商品做推薦，這些個人特質不好做處理，因此直接做刪除不使用。

剩餘類別

variation type	variation value	size	ingredients	highlights

剩餘的這五項變數，透過*Cramer's V 計算能得到他們對商品"Product_id"是一對一的關係。因此不須做處理。

合併資料

將這些類別變數都確定後，就可以以商品為標的來做合併，我們將剩下的五個數值變數取平均，使得我們的資料變成 1998 筆商品的資料。

建立模型

目標變數：取平均後的 is_recommended

選擇變數

除去目標變數後，我們剩下了*23 個變數，其中仍然有 5 個變數是有缺失值的，由於這五項變數都為類別變數且填入缺失值對模型影響不大，因此不使用這 5 個變數。

變數名稱	範例
highlights	['Clean at Sephora', 'Best for Dry, Combo, Nor... (後續過長)
variation_value	5 oz/ 150 mL
variation_type	Size
size	5 oz/ 150 mL
ingredients	['Aqua (Water), Coco-Glucoside, Butylene Glyco... (後續過長)

另外，商品名稱 "product_name" 及商品 id "product_id" 也與目標變數有一對一關係，這裡也不做使用。

初步的線性模型

首先我們使用了剩餘的 16 個變數，將資料標準化和拆分為 8:2 的訓練及測試資料後，直接建立一個初步的模型。

變數	coef	std err	t	P>|t|
const	0.8168	0.001	654.121	0
rating	0.1246	0.001	86.312	0
total_feedback_count	2.39E+04	3.04E+04	-0.785	0.432
total_neg_feedback_count	6321.363	8050.179	0.785	0.432
total_pos_feedback_count	1.85E+04	2.35E+04	0.785	0.432
brand_name	0.0049	0.002	3.229	0.001
price_usd	0.0004	0.001	0.329	0.742
brand_id	-0.0006	0.001	-0.419	0.676

child_count	0.0015	0.001	1.116	0.265
limited_edition	0.0034	0.001	2.699	0.007
loves_count	-0.0025	0.002	-1.457	0.145
new	-0.0016	0.001	-1.209	0.227
online_only	-0.0008	0.001	-0.627	0.531
out_of_stock	3.33E-05	0.001	0.026	0.979
reviews	0.0031	0.002	1.77	0.077
sephora_exclusive	0.0033	0.001	2.527	0.012
tertiary_category	0.0034	0.001	2.621	0.009

可以發現在 P-value 小於 0.1 的標準下，我們會剩下 6 個變數。可以使用這 6 個變數建立一個更小的模型。

調整後模型

變數	coef	std err	t	P>ltl
const	0.8168	0.001	644.18	0
rating	1.26E-01	1.00E-03	89.548	0
brand_name	0.0054	0.001	3.94	0
limited_edition	3.00E-03	1.00E-03	2.357	0.019
reviews	0.0035	0.001	2.76	0.006
sephora_exclusive	0.0036	0.001	2.84	0.005
tertiary_category	0.003	0.001	2.323	0.02

R-squared : 0.904	F_statistic : 1743	AIC : -3888	Skew : -0.528
Adj. R-squared : 0.903	Log-Likelihood : 1950.8	BIC : -3853	Kurtosis : 34.696

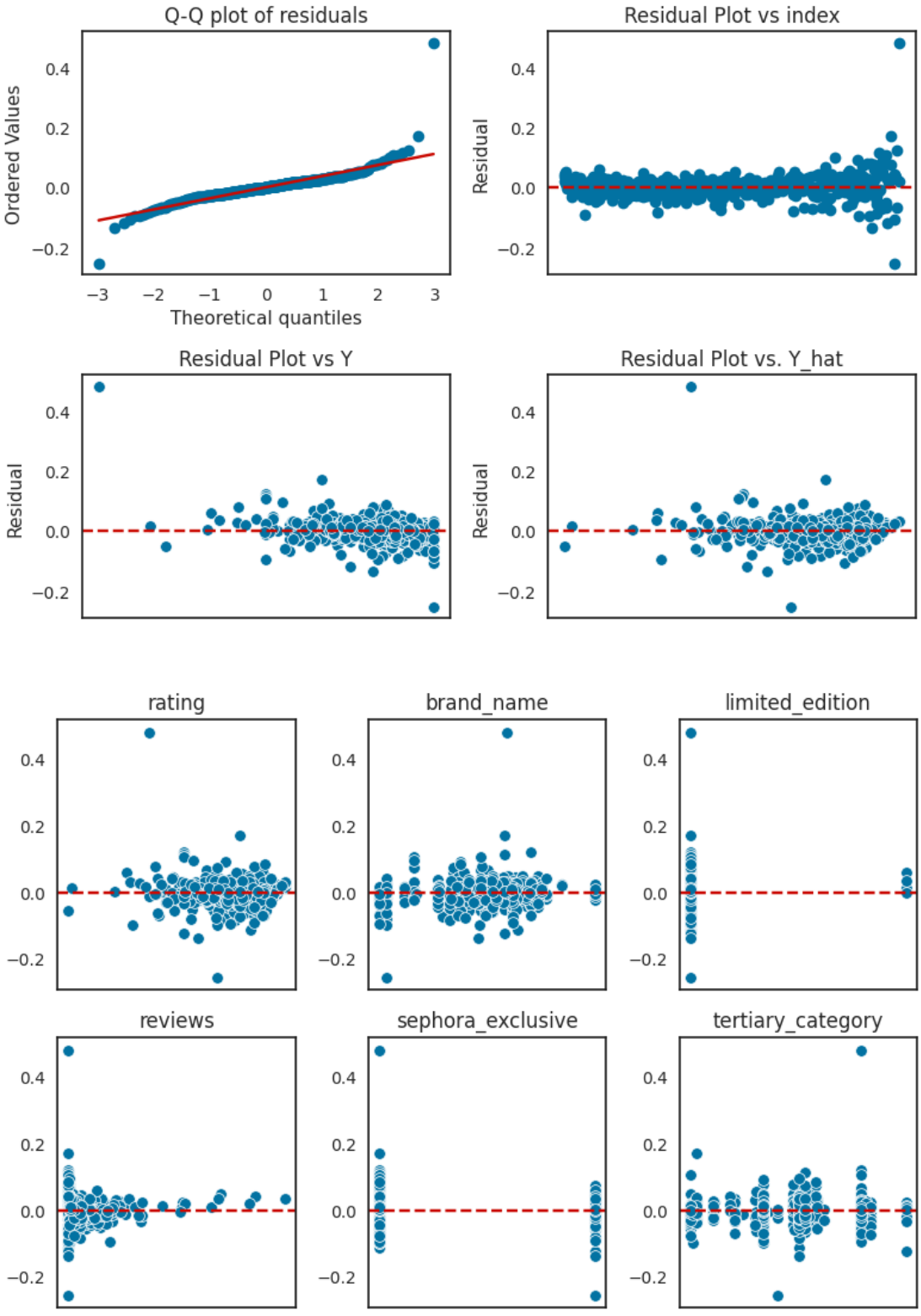
我們可以看到模型解釋力非常高，但是鋒度(Kurtosis)卻是偏高的。

測試集 ANOVA

	SS	df	MS	F_statistic
Coefficients	8.0577	6	1.3429	1763.2120
Residuals	0.8462	1111	7.6165e-4	

根據上表，我們可以得出這個調整後的模型對測試集也能很好的適應。

殘差檢定(對測試集)



在殘差檢定上的表現也相當好，不過鋒度就如同 QQ-plot 呈現出的一樣。

結論與未來展望

- **結論**：

根據篩選變數後的模型，我們可以透過品牌、評分、是否為限量款等等，來預測是否推薦該產品，不過鋒度過高需要找辦法解決。

- **未來展望**：

希望能透過更多有關個人特質的變數以及原始的資料來做模型以取得更個人化的預測。

附錄 – 唯一值與缺失值總表

變數名稱	唯一值	缺失值
variation_desc	935	1091034
sale_price_usd	88	1090576
value_price_usd	174	1069365
child_max_price	222	644796
child_min_price	208	644796
helpfulness	3767	567735
review_title	364105	316797
hair_color	7	232911
eye_color	6	215771
skin_tone	14	176682
is_recommended	2	174131
tertiary_category	118	161894
skin_type	4	117700
highlights	4417	115729
variation_value	2729	64644
variation_type	7	52560
size	2055	44661
ingredients	6538	22843
review_text	969419	7587
product_name	2334	6143
brand_name	142	6143

total_feedback_count	676	6143
total_neg_feedback_count	259	6143
total_pos_feedback_count	590	6143
submission_time	5317	6143
author_id	503216	6143
rating	5	6143
price_usd	221	6143
reviews	1556	278
secondary_category	41	8
new	2	0
loves_count	7436	0
out_of_stock	2	0
primary_category	9	0
limited_edition	2	0
sephora_exclusive	2	0
product_id	8494	0
child_count	55	0
brand_id	304	0
online_only	2	0