Project Title

Analyzing and Predicting Sephora Skincare Trends

Milestone: Project Report

Charan Kumar Domalapati
domalapati.c@northeastern.edu

Signature of Student : Charan Kumar Domalapati

Contents

1. Introduction

1.1 Problem Setting : The beauty and skincare industry is rapidly growing, with consumers increasingly relying on online reviews to make informed purchasing decisions. Sephora, a leading retailer in the beauty space, offers a vast array of skincare products, and the reviews left by customers are crucial in influencing future purchases. Analyzing these reviews can provide valuable insights into consumer preferences, product effectiveness, and overall trends.

1.2 Objective : This project focuses on leveraging customer review data to predict whether a skincare product will be recommended. By understanding the factors that drive recommendations, Sephora can optimize its product offerings, marketing strategies, and customer service initiatives.

1.3 Problem Definition :

Questions Addressed

- What are the key factors in customer reviews that determine whether a product is recommended?
- Can we accurately predict the is_recommended label based on review text and product features?
- How do different machine learning and deep learning models compare in terms of predictive accuracy for this task?
- How do customer ratings correlate with the likelihood of a product being recommended?
- What role do specific product ingredients play in influencing customer recommendations?
- Are there any temporal trends in the recommendation patterns over the years?
- Which product categories (e.g., moisturizers, serums) have the highest rates of recommendation, and why?
- How does the price of a product influence its recommendation rate?
- Can sentiment analysis of review text provide additional insights into product recommendations?
- What are the most common themes or topics in reviews of highly recommended versus non-recommended products?
- How does the availability of products (in stock versus out of stock) impact their recommendation status?
- What impact does product exclusivity (e.g., Sephora exclusives) have on recommendation rates?
- How do customer demographics (if available) influence product recommendations?

3. Data Sources :

The primary datasets utilized in this project are:

Sephora Products Dataset: This dataset offers an in-depth look at a wide range of skincare products available on Sephora, providing extensive details about each item. It includes not only the product descriptions and features but also customer reviews, ratings, and the product variations. This rich dataset is essential for analyzing product performance and understanding consumer preferences in the skincare market.

Sephora Reviews Datasets: This dataset contains detailed customer reviews and ratings for Sephora skincare products. In addition to these features, it also includes a binary indicator showing whether the product is recommended to other users or not. This comprehensive information will be crucial in analyzing customer sentiment and predicting product recommendations.

These datasets are sourced from Kaggle. Link

4. Data Description :

Product Info Dataset:

Number of Rows: 8495

Number of Columns: 27

Description: This dataset offers an in-depth look at a wide range of skincare products available on Sephora, providing extensive details about each item. It includes not only the product descriptions and features but also customer reviews, ratings, and the product variations.

Reviews Dataset:

Number of Rows: 1094411

Number of Columns: 18

Description: This dataset contains detailed customer reviews and ratings for Sephora skincare products. In addition to these features, it also includes a binary indicator showing whether the product is recommended to other users or not. This comprehensive information will be crucial in analyzing customer sentiment and predicting product recommendations.

4.1 Description of some of the key features in the above datasets:

product_id : The unique identifier for the product

brand_name : The full name of the product brand

loves_count : The number of people who have marked this product as a favorite.

rating : The average rating of the product based

reviews : The number of user reviews for the product

ingredients : A list of ingredients included in the product

review_text : It contains the written feedback provided by customers about their experiences with the skincare products. This text includes detailed opinions, observations, and sentiments expressed by the users.

size : The size of the product, which may be in oz, ml, g, packs, or other units depending on the product type

price_usd : The price of the product in US dollars

is_recommended : Indicates whether the product is recommended to the other users or not.(1-Yes, 0 – No)

helpfulness : Indicates whether the product is helpful to the other users or not.(1-Yes, 0 – No)

submission_time : It contains the review submission time which is in yyyy–mm-dd format

new : Indicates whether the product is new or not (1-true, 0-false)

online_only : Indicates whether the product is only sold online or not (1-true, 0-false)

out_of_stock : Indicates whether the product is currently out of stock or not (1 if true, 0 if false)

sephora_exclusive : Indicates whether the product is exclusive to Sephora or not (1 if true, 0 if false)

limited_edition : Indicates whether the product is a limited edition or not (1-true, 0-false)

child_count : The number of variations of the product available

5. Data Preprocessing :

5.1 Data Cleaning

```
df_merge = df_merge.drop(['value_price_usd','sale_price_usd','variation_desc','variation_value'],axis=1)
df_merge[['child_max_price', 'child_min_price']] = df_merge[['child_max_price', 'child_min_price']].fillna(0)
df_merge['rating']=df_merge['rating'].fillna(0)
df_merge['reviews']=df_merge['reviews'].fillna(0)
df_merge['helpfulness']=df_merge['helpfulness'].fillna(0)
df_merge['submission_time'] = pd.to_datetime(df_merge['submission_time'])
df_merge['year'] = df_merge['submission_time'].dt.year
df_merge['tertiary_category'] = df_merge['tertiary_category'].fillna(df_merge['secondary_category'])
df_merge.drop(['submission_time'], axis=1, inplace=True)
df_merge['skin_tone'] = df_merge['skin_tone'].fillna('Unknown')
df_merge['skin_type'] = df_merge['skin_type'].fillna('Unknown')
df_merge['eye_color'] = df_merge['eye_color'].fillna('Unknown')
df_merge['hair_color'] = df_merge['hair_color'].fillna('Unknown')
```

The code snippet is performing data preprocessing on a DataFrame df_merge. It drops unnecessary columns like 'value_price_usd,' 'sale_price_usd,' 'variation_desc,' and 'variation_value.' It then handles missing values by filling them with 0 for numerical columns like 'child_max_price,' 'child_min_price,' 'rating,' 'reviews,' and 'helpfulness.' For categorical columns like 'skin_tone,' 'skin_type,' 'eye_color,' and 'hair_color,' it fills missing values with 'Unknown.' Additionally, the 'submission_time' is converted to a datetime format, and the year is extracted into a new column 'year.' Finally, it fills missing values in 'tertiary_category' with values from 'secondary_category' and drops the original 'submission_time' column.

5.2 Cleaning Text Data

```
    text = re.sub(r'[^A-Za-z\s]', '', text)
    text = text.lower()
    text = re.sub(r'http\S+', '', text)  # Remove URLs
    text = re.sub(r'@\w+|#\w+', '', text)  # Remove mentions and hashtags
    tokens = word_tokenize(text)
    stop_words = set(stopwords.words('english'))
    tokens = [word for word in tokens if word not in stop_words]
    # Lemmatize the tokens
    lemmatizer = WordNetLemmatizer()
    tokens = [lemmatizer.lemmatize(word) for word in tokens]
    return ' '.join(tokens)

text_data = df_text['text_data'].apply(clean_text)
print(text_data)
```

The code snippet is a text preprocessing function that cleans and prepares text data for analysis. It converts the text to lowercase, removes non-alphabetical characters, URLs, mentions, and hashtags. The text is then tokenized into individual words, and stop words (common words like "and" or "the") are removed. After that, the remaining tokens are lemmatized, which means they are reduced to their base or root form. Finally, the cleaned and processed text is rejoined into a single string. This function is applied to a DataFrame column, text_data, to prepare the text for further analysis or modeling.

5.3 Splitting the Original Data for Unseen data

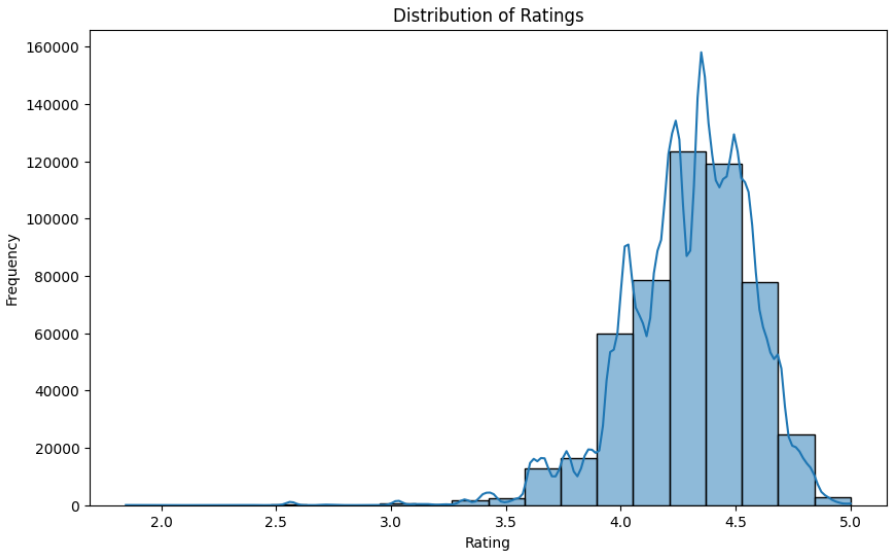
Splitting the 5% of Original data for unseen data to test the models in the end

```
sample_size = int(0.05 * len(df_merge))

# Randomly sample 5% of the data
unseen_data = df_merge.sample(n=sample_size, random_state=42)

# Remove the sampled data from the original dataframe to create the training set
df_merge = df_merge.drop(unseen_data.index)
print("Training data shape:", df_merge.shape)
print("Unseen data shape:", unseen_data.shape)
```

```
Training data shape: (520638, 36)
Unseen data shape: (27402, 36)
```

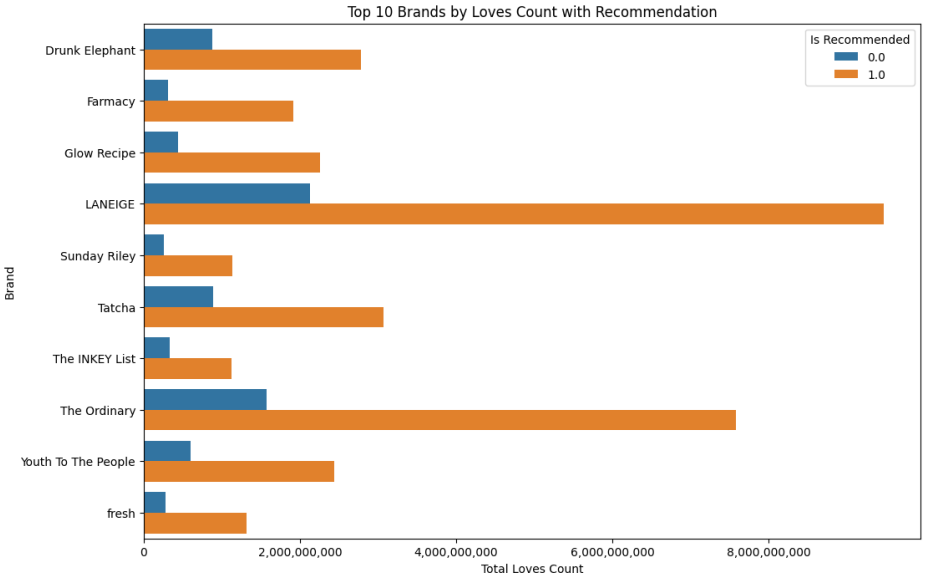
6. Exploratory Data Analysis(EDA) :

6.1. Distribution of Ratings



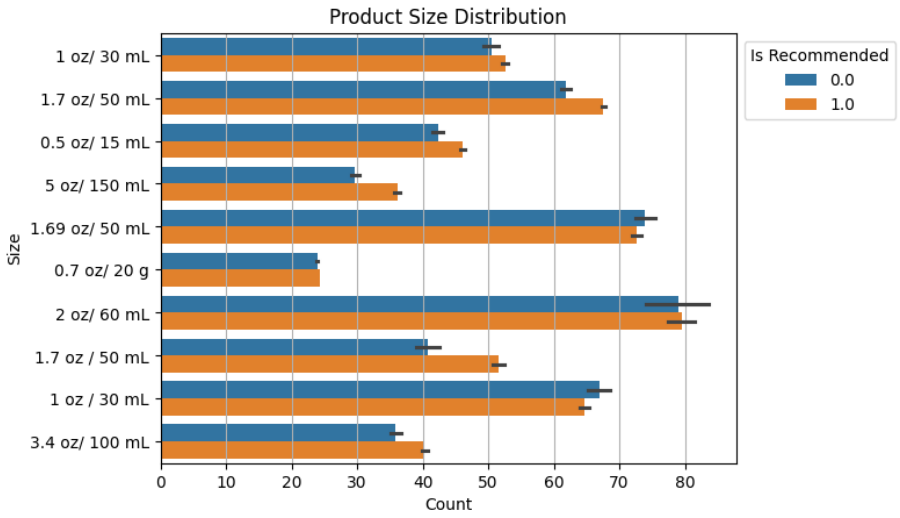
The graph shows a right-skewed distribution of ratings, with the majority of ratings clustered between 4.0 and 4.5. The highest frequency of ratings is around 4.5, indicating that users tend to give higher ratings. Low ratings below 3.0 are rare, suggesting overall positive feedback in the dataset.

6.2. Top 10 brands by total loves count



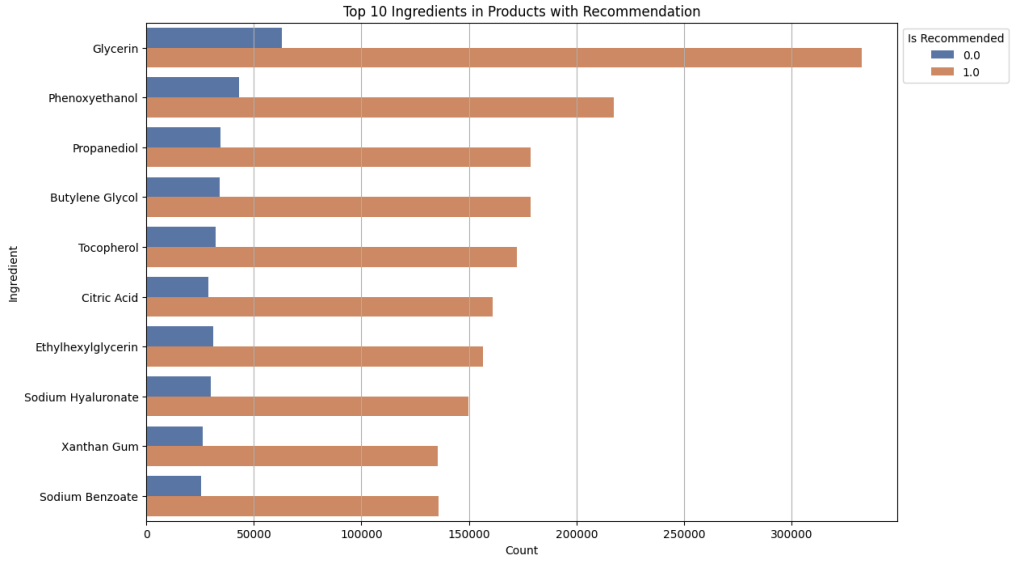
The graph illustrates that the top 10 brands by loves count are predominantly associated with recommended products, with LANEIGE and The Ordinary leading by a significant margin. Most brands have minimal loves count for non-recommended products, highlighting that recommended products are much more popular. LANEIGE and The Ordinary, in particular, stand out as the most loved and recommended brands.

6.3. Product Size Distribution



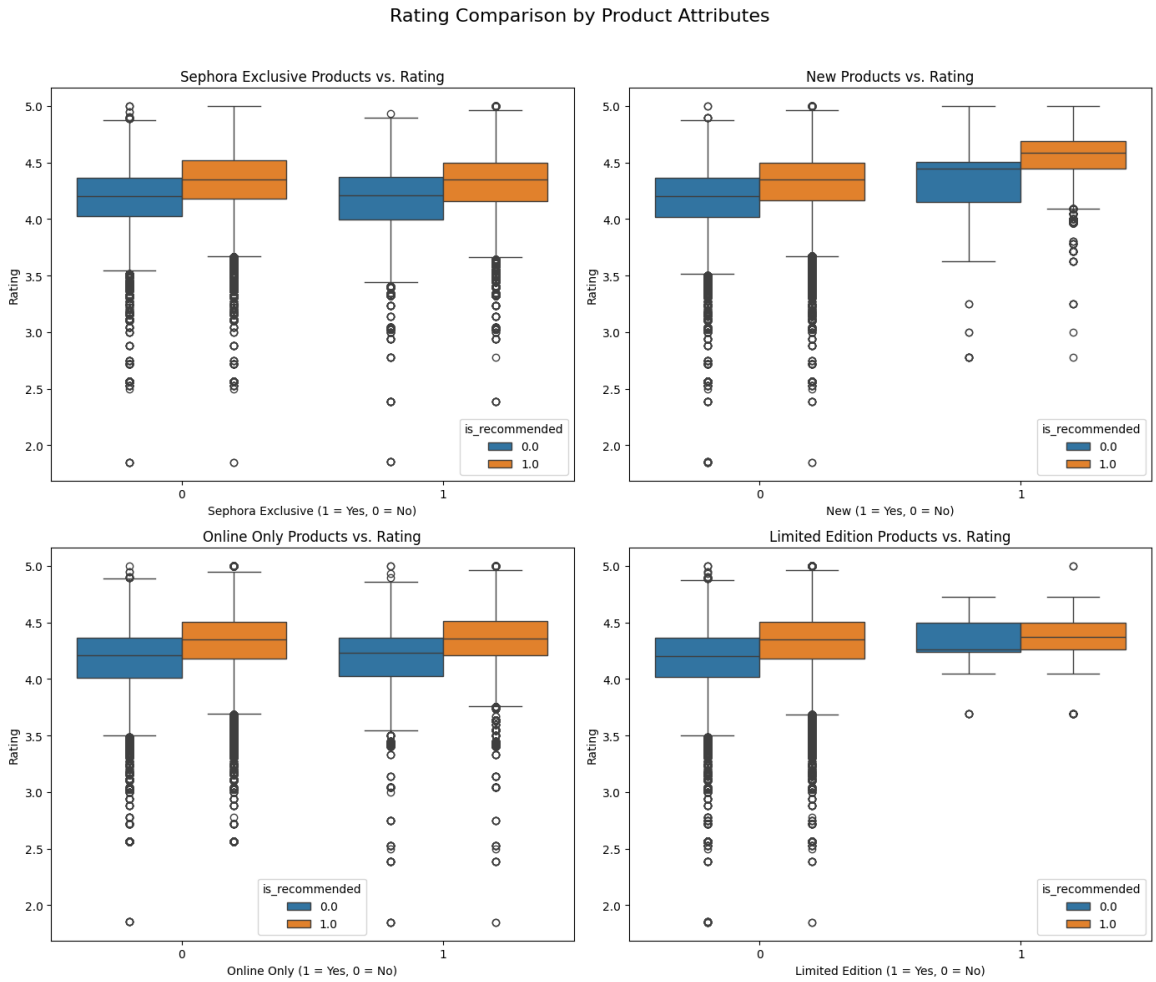
The graph depicts the distribution of product sizes with a comparison between recommended (orange bars) and non-recommended (blue bars) items. It shows that most product sizes have a nearly even split between recommended and non-recommended products, with slight variations. Notably, the 1.7 oz/50 mL and 1 oz/30 mL sizes have higher counts, indicating these sizes are more commonly available or popular, but the recommendation status is fairly balanced across all sizes.

6.4. Top 10 Ingredients in Products with Recommendation



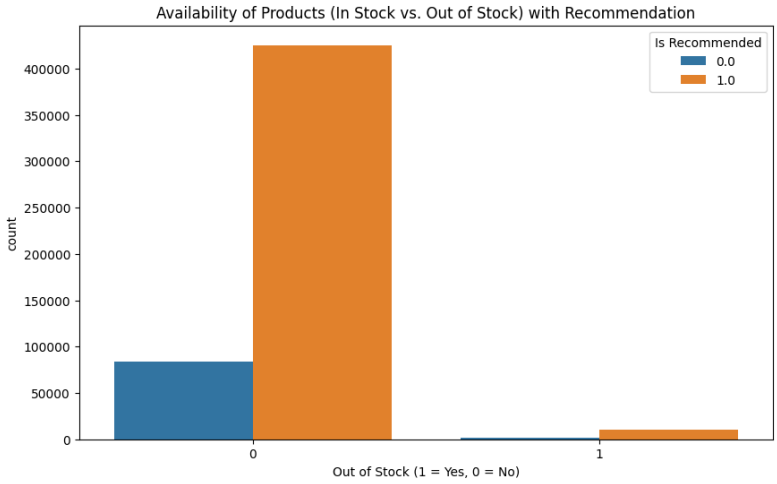
The graph shows the top 10 ingredients used in products, comparing the counts of recommended (orange bars) versus non-recommended (blue bars) products. Glycerin stands out as the most common ingredient, overwhelmingly found in recommended products. Across all ingredients, the recommended products far outnumber the non-recommended ones, indicating that these ingredients are generally favored in products that receive positive recommendations.

6.5. Rating Comparison by Product Attributes



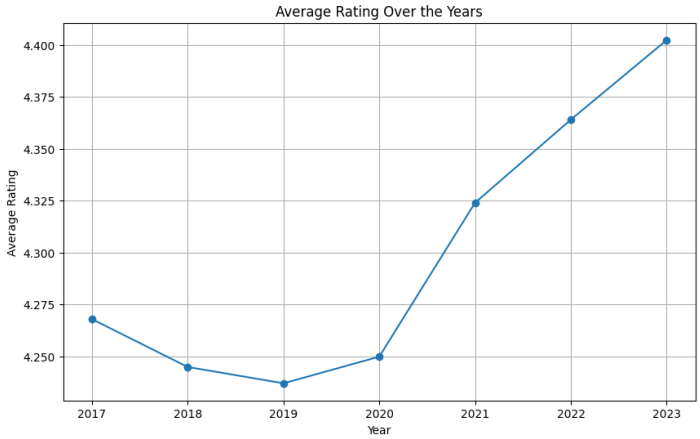
The box plots compare ratings across four product attributes: Sephora exclusive, new products, online-only products, and limited edition products, based on whether they are recommended (orange) or not (blue). For all attributes, recommended products tend to have higher median ratings, indicating they are generally rated more favorably. The box plots also show that recommended products have less variability in ratings, suggesting more consistent positive feedback. New products and limited edition products, in particular, show a clear distinction, with recommended items having significantly higher ratings. Non-recommended products exhibit more variability and lower ratings, particularly in the Sephora exclusive and online-only categories. This analysis highlights that recommendation status strongly correlates with higher and more consistent ratings across these product attributes.

6.6. Availability of Products (In Stock vs. Out of Stock) with Is Recommended



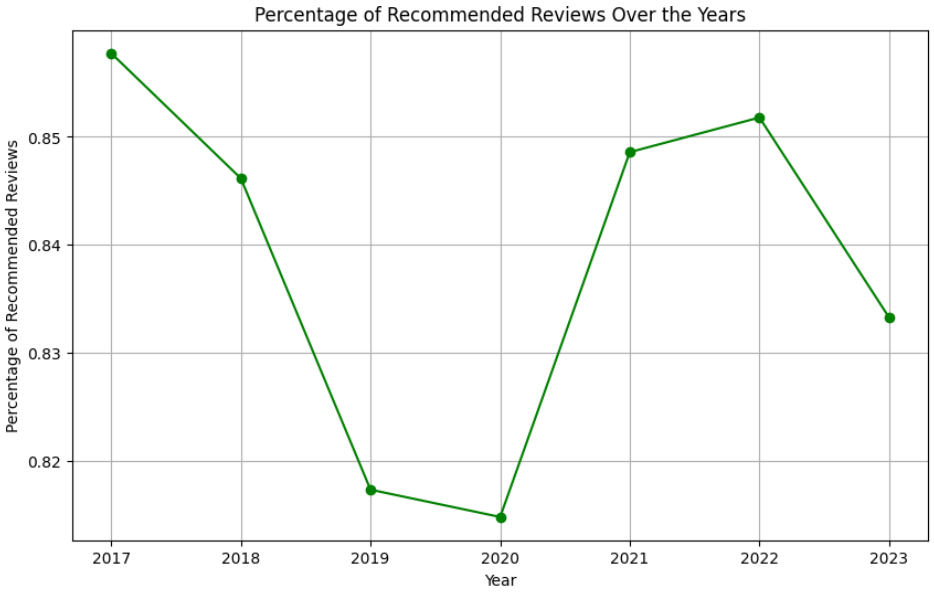
The bar chart illustrates the availability of products (in stock versus out of stock) with a comparison of their recommendation status. The majority of products are in stock (indicated by '0' on the x-axis), and among these, a large portion is recommended (orange bar). Out-of-stock products (indicated by '1') are far fewer, with a slightly higher proportion of recommended products compared to non-recommended ones. This suggests that recommended products are more likely to be in stock, possibly due to higher demand, while out-of-stock items are rare, but still more often recommended when they do occur.

6.7. Average Rating Over the Years of Sephora



The line graph depicts the trend in average ratings from 2017 to 2023. The average rating slightly declined from 2017 to 2019, reaching its lowest point in 2019. However, starting in 2020, there has been a consistent and significant increase in average ratings each year, peaking in 2023. This upward trend suggests that products have been receiving increasingly higher ratings over the past few years, indicating possible improvements in product quality or customer satisfaction.

6.8. Percentage of Recommended Reviews Over the Years

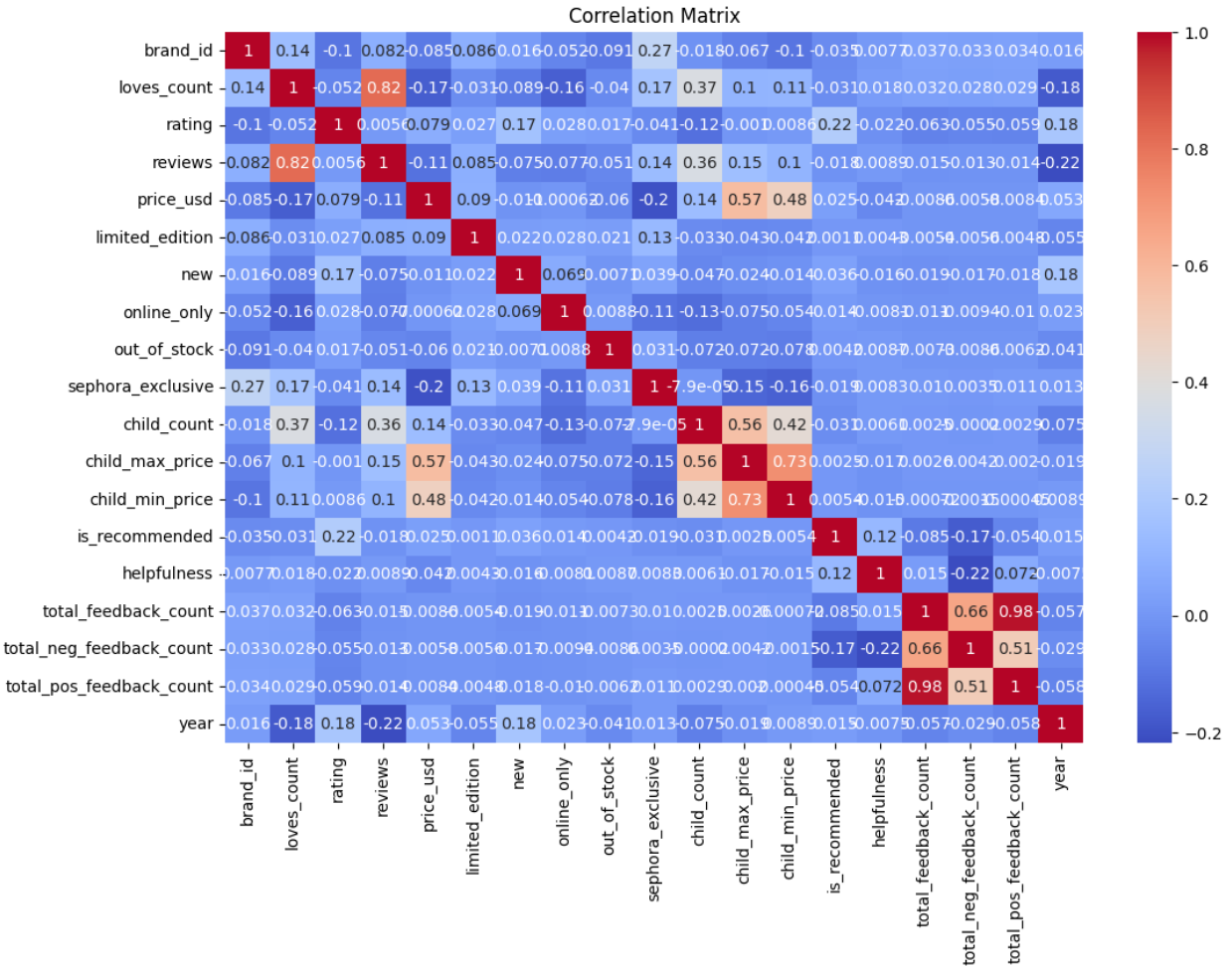


The line graph illustrates the percentage of recommended reviews from 2017 to 2023. The percentage initially declined from 2017 to 2019, reaching its lowest point in 2019. However, there was a significant rebound in 2020, with the percentage increasing steadily and peaking in 2022. In 2023, the percentage of recommended reviews dropped again, though it remained higher than in 2019. This fluctuation suggests varying levels of product satisfaction and recommendation rates over the years, with a notable improvement after 2019, followed by a decline in 2023.

6.9. price vs rating

The scatter plot shows the relationship between price and rating, with each point representing a product. The data points are scattered across a wide range of prices, showing no clear trend that higher or lower prices lead to better ratings. The ratings mostly cluster between 4.0 and 4.5 across different price points, indicating that price does not significantly impact the average rating. However, there are some outliers, particularly at lower and higher prices, where ratings can be unusually high or low. This suggests that both expensive and cheap products can receive a wide range of ratings, emphasizing that other factors besides price influence product satisfaction.

6.10. Correlation Matrix

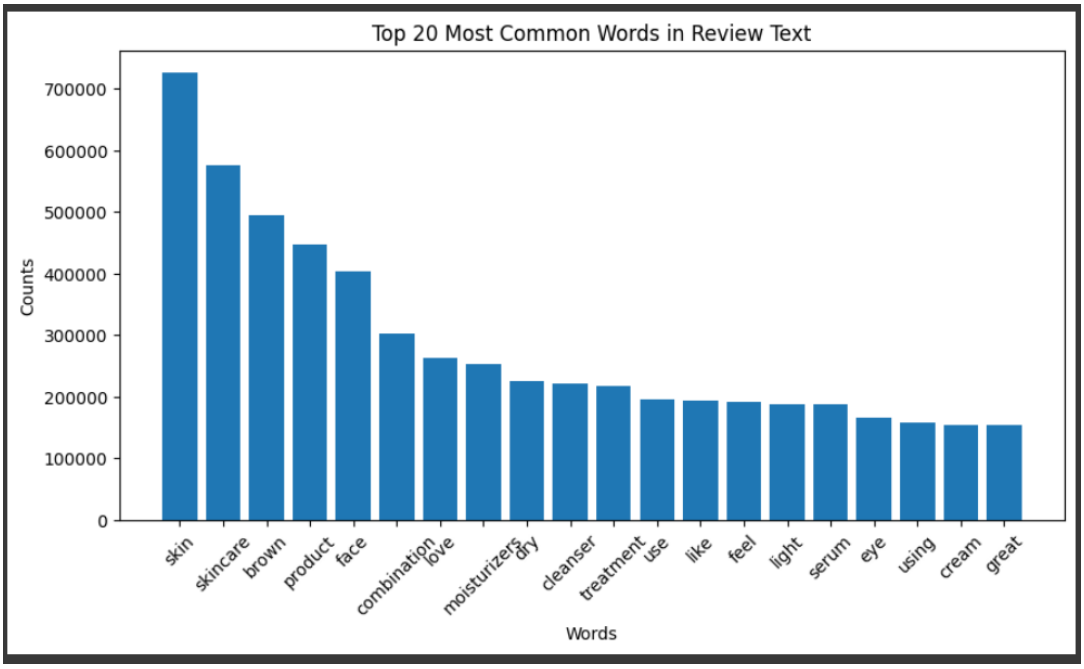


By observing the correlation matrix with a focus on "total_feedback_count" and "total_pos_feedback_count," it's clear that these two features are highly correlated, with a correlation coefficient of 0.98. This high correlation suggests that "total_feedback_count" can largely be explained by "total_pos_feedback_count," indicating redundancy between these variables. Since "total_feedback_count" is being removed due to its high correlation, this will help in reducing multicollinearity in my analysis and potentially improve model performance by focusing on "total_pos_feedback_count" as a representative feature.

6.11. Word Cloud

The word cloud represents the most frequently occurring words in the review text, with larger words indicating higher frequency. Key terms like "skin," "brown," "skincare," "product," and "moisturizer" are prominently featured, suggesting that discussions around skin and skincare products are central themes in the reviews. Other notable words such as "face," "cream," "dry," and "serum" also indicate common concerns or product types that users are reviewing. This visualization provides a quick overview of the main topics and concerns that are most prevalent in the review dataset.

6.12. Top 20 most common words in Review Text



The bar chart shows the top 20 most common words in the review text, with "skin" being the most frequently mentioned word, followed by "skincare," "brown," and "product." These words suggest that the reviews focus heavily on skin-related products and concerns. Terms like "combination," "moisturizer," and "cleanser" indicate specific product types or skin conditions that are often discussed. The presence of words such as "love," "feel," and "like" suggests that users frequently express their opinions and experiences with these products, emphasizing personal satisfaction and product effectiveness. This chart highlights the key themes and topics that dominate the review discussions.

8. Model Approaches:

For Numerical Data :

- Logistic Regression
- Naïve Bayes
- Random Forest
- Neural Network (MLP classifier)

For Text Data:

- Logistic Regression
- Naïve Bayes
- Feedforward Neural Network with 2 hidden layers each of 64 nodes
- LSTM Neural Network

For text Data:

In Logistic Regression and Naïve Bayes we will be vectorizing the text data

```
# Vectorize the text data
vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(text_data)
y = df_text['is_recommended'].astype(int)
```

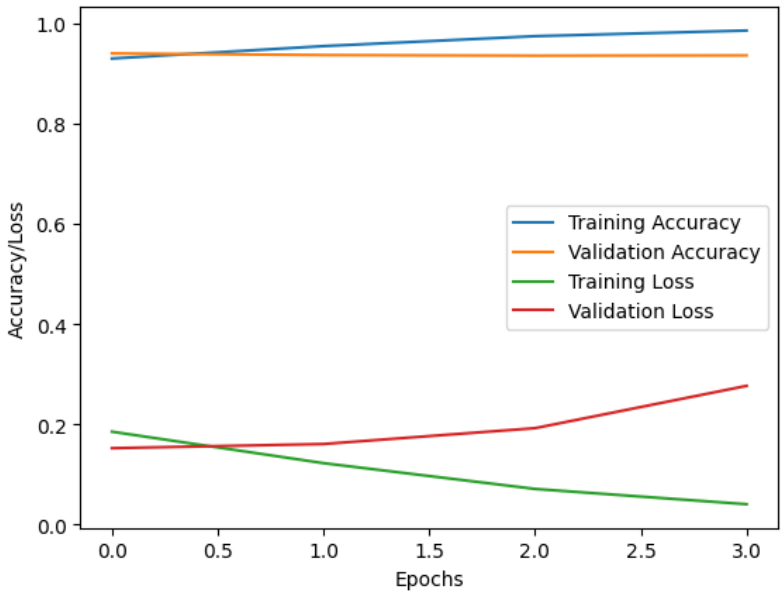
In Feedforward Neural Network and LSTM Neural Network we wil be tokenizing the text data and we will be finding padded sequences

```
# Tokenization
tokenizer = Tokenizer(num_words=20000)
tokenizer.fit_on_texts(text_data)
X = tokenizer.texts_to_sequences(text_data)
vocab_size = len(tokenizer.word_index) + 1  # Adding 1 because of reserved index 0
```

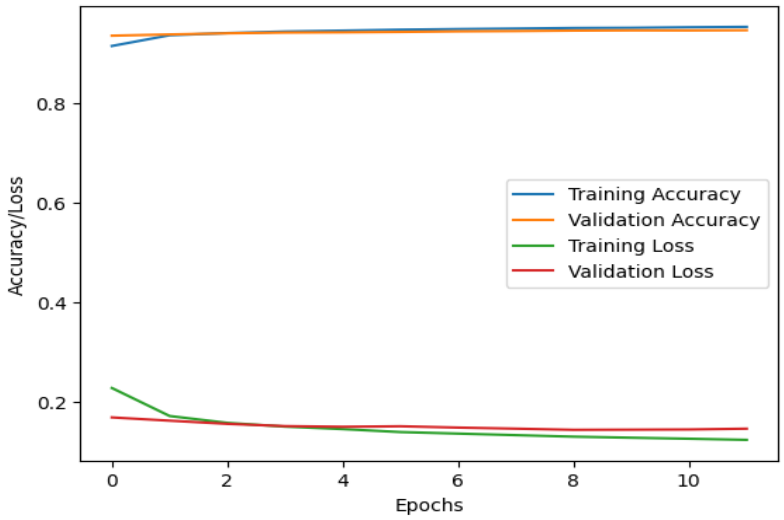
```
# Padding sequences to ensure uniform input length
maxlen = 200
X = pad_sequences(X, padding='post', maxlen=maxlen)
y = df_text['is_recommended']
```

Finding epoch value :

For Feedforward Neural Network I choose epoch = 2 where I have checked the accuracy is good and loss is minimal for both validation and training data.



For Feedforward Neural Network I choose epoch = 13 where I have checked the accuracy is good and loss is minimal for both validation and training data.



9. Performance Evaluation :

Evaluation Metrics for Numerical Data :

Metric	Logistic Regression	Naïve Bayes	Random Forest	Neural Network (MLP Classifier)
Validation Accuracy	0.84	0.81	0.84	0.85
Validation Precision	0.85	0.86	0.87	0.87
Validation Recall	0.99	0.93	0.95	0.98
Validation F1 Score	0.91	0.89	0.91	0.92
Test Accuracy	0.84	0.81	0.84	0.85
Test Precision	0.85	0.85	0.87	0.86
Test Recall	0.99	0.93	0.95	0.98
Test F1 Score	0.91	0.89	0.91	0.92

Evaluation Metrics for Text Data :

Metric	Logistic Regression	Naïve Bayes	FeedForward Neural Network	LSTM Neural Network
Test Accuracy	0.94	0.85	0.94	0.94
Test Precision	0.95	0.85	0.96	0.96
Test Recall	0.98	1.00	0.96	0.97
Test F1 Score	0.96	0.92	0.96	0.97

10. Project Results :

I have taken one unseen 5% data from the original data and I used that unseen data to test my models. After getting accuracy from every model I took a mode of the models predictions and compared with the target variable that is is_recommended .

Confusion Matrix:

[[836 3585]

 [462 22519]]

Evaluation metric:

Accuracy: 0.85

Precision: 0.86

Recall: 0.98

F1 Score: 0.92

Specificity: 0.19

False Positive Rate: 0.81

False Negative Rate: 0.02

Interpreting the results:

The evaluation metrics indicate that the model performs well overall, with an accuracy of 85% and a high precision of 86%, meaning that most of the products predicted as recommended are indeed recommended. The recall is particularly strong at 98%, showing the model is very effective at identifying recommended products. The F1 score of 0.92 reflects a good balance between precision and recall. However, the low specificity (0.19) and high false positive rate (0.81) suggest that the model struggles with correctly identifying non-recommended products, often misclassifying them as recommended.

11. Impact of the Project Outcomes :

- **Enhanced Product Offerings:** Sephora can refine its product line by focusing on the ingredients and attributes that are most likely to lead to positive customer recommendations.

- **Personalized Marketing:** The predictive models enable Sephora to create more targeted marketing campaigns, suggesting products that align with individual customer preferences.

- **Informed Product Development:** Insights from the analysis can guide the development of new skincare products tailored to consumer demands and preferences.

- **Improved Customer Satisfaction:** By understanding and catering to the factors that drive product recommendations, Sephora can enhance overall customer satisfaction.

- **Increased Sales:** Products that are more likely to be recommended are also more likely to be purchased, potentially leading to higher sales figures.

- **Competitive Advantage:** The ability to predict product recommendations gives Sephora a strategic edge in the competitive skincare market.

- **Data-Driven Decision Making:** The project promotes a data-driven approach to business decisions, improving the accuracy and effectiveness of Sephora's strategies.

- **Optimized Inventory Management:** Understanding which products are more likely to be recommended can help Sephora manage inventory more efficiently, ensuring popular products remain in stock.

- **Enhanced Customer Trust:** Accurate predictions and personalized recommendations can increase customer trust in Sephora's product suggestions, fostering brand loyalty.

- **Scalable Insights:** The methodologies developed in this project can be scaled and applied to other product categories beyond skincare, broadening Sephora's overall business impact.

12. Challenges:

1. High Dimensionality: Selecting relevant features from 36 can lead to overfitting if not managed properly.

2. Numerical vs. Text Data: Balancing contributions from both types of data is challenging and crucial for model accuracy.

3. Computational Time: Training complex models on 500,000 rows requires significant time and computing power.

4. Model Interpretability: Complex models like Neural Networks are difficult to interpret, reducing trust in predictions.

5. Overfitting in Text Models: Text data is prone to overfitting, needing careful regularization and validation.

6.Text Data Preprocessing: Poor or excessive text preprocessing can significantly impact model performance.

13. Conclusion :

The project successfully identified key factors influencing whether a Sephora skincare product is recommended, leveraging both numerical and text data from customer reviews. The analysis demonstrated that customer ratings, ingredient composition, and product attributes like exclusivity and availability significantly impact recommendations. The predictive models, particularly those using advanced machine learning and deep learning techniques, achieved high accuracy, with a notable precision and recall, highlighting the models' effectiveness in predicting product recommendations. However, challenges such as balancing numerical and text data contributions, managing high-dimensionality, and preventing overfitting were critical to the project's success. Overall, the project provides valuable insights into consumer preferences and lays the groundwork for improving product offerings and marketing strategies at Sephora.