Capstone Project - Interim Report

Contents

Industry Review

Current Practices and Background Research

Overview of the Skincare Industry:
The global skincare industry has seen significant growth in recent years, driven by an increasing awareness of skincare and a demand for products tailored to individual needs. Consumers are now more interested in personalized skincare solutions, focusing on products that cater specifically to their skin type, concerns, and lifestyle preferences. Trends such as clean beauty and the use of technology for skincare personalization have become highly prominent.

Current Practices in Skincare Recommendations:
Currently, skincare recommendations are provided through various methods, including dermatology consultations, online quizzes, and machine learning algorithms used by e-commerce platforms. Companies leverage consumer data such as skin type, concerns (e.g., acne, dryness, or pigmentation), and even regional environmental factors to suggest products. Some brands use sophisticated AI-driven systems to analyze this data and offer tailored recommendations to enhance customer satisfaction and loyalty.

Role of Technology:
Technology plays a crucial role in modern skincare recommendations. AI and machine learning are utilized for skin analysis, personalized recommendations, and product suggestions. Examples include virtual skin analysis tools and mobile apps with skin scanning capabilities, which assess skin conditions through images and recommend products accordingly. Many platforms are also using ingredient-based recommendations by matching product ingredients to specific skin needs.

Literature Survey

1. Publications

- **Hu et al. (2018)**: Explored content-based recommendation systems for beauty products, focusing on utilizing user preferences, product types, and ingredients. The study showed how personalization could enhance recommendation relevance.
- **Kim and Lee (2019)**: Proposed a collaborative filtering model tailored to beauty products, addressing the sparsity issue by integrating social data to predict user preferences.
- **Zhang et al. (2021)**: Developed a hybrid recommendation model that combines collaborative and content-based filtering, which improved cold start issues for new products and users by leveraging both user and product features.

Dataset and Domain

Data Dictionary

- **Product**: Full name and description of each skincare product.
- **Price**: Price of the product in the specified currency, such as INR.
- **Rating**: Customer rating of the product on a scale from 0 to 5.
- **Brand**: Name of the brand that manufactures or distributes the product.
- **Product Type**: Category or type of the skincare product (e.g., Face Wash, Face Gel).
- **Packing**: Details about the packaging, indicating quantity and unit (e.g., 200 ml, 100 g).
- **Packing (ml)**: Standardized quantity in millilitres to provide consistency across entries.

Variable Categorization

- **Number of Rows**: 13,752
- **Number of Columns**: 7

Numeric Variables**:**

- Count: 3
- Columns: Price, Rating, Packing (ml)

Categorical Variables**:**

- Count: 4
- Columns: Product, Brand, Product Type, Packing

Data Pre-Processing

Handling Missing Values:
During the data cleaning process, the **Rating** column had a significant number of missing values (2,225 out of total entries). To address this issue, the following steps were taken:

1. **Brand-Specific Mean Imputation**: The missing values in the Rating column were initially replaced with the mean rating for each respective Brand. This approach ensures that the ratings are more reflective of the product's brand.
2. **Overall Mean Imputation**: After applying the brand-specific mean, some entries still had missing values, likely due to brands that did not have any ratings available. These remaining missing ratings were then replaced with the overall mean rating across all products. By applying these techniques, we ensured that the Rating column was complete, allowing for more accurate and reliable analysis in subsequent steps.

Project Justification

Project Statement:
This project aims to develop a personalized skincare recommendation system using machine learning to suggest products based on factors like skin type, concerns, and preferences. A key challenge was handling missing values in the Rating column. To address this, we used brand-specific and overall mean imputation to ensure accurate and reliable analysis for generating recommendations.

Complexity Involved**:**

1. **Data Imputation**:
   The Rating column had 2,225 missing values, which were handled using brand-specific and overall mean imputation strategies, ensuring data consistency without bias.
2. **Data Consistency**:
   Ensuring consistent and accurate data across columns like Price and Product Type was crucial for building a reliable recommendation system.
3. **Recommendation System Development**:
   After cleaning, the challenge was to balance multiple factors for personalized product recommendations.

Project Outcome**:**

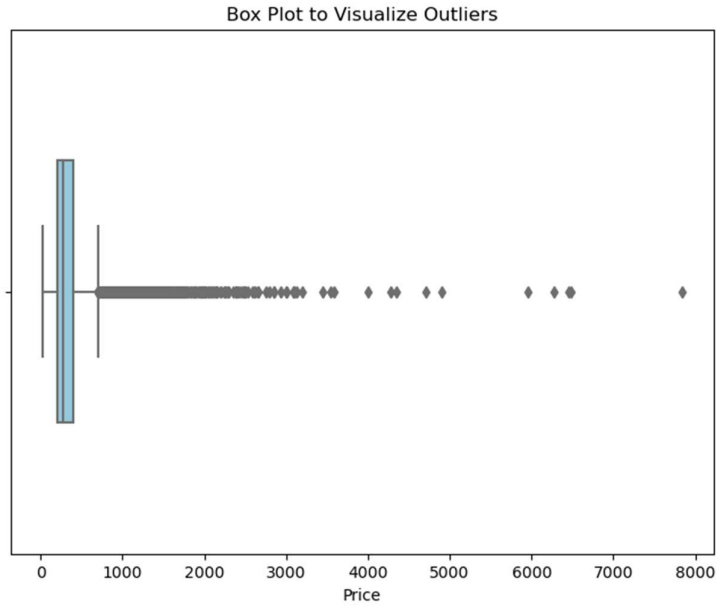
1. **Commercial Value**:
   The recommendation system enhances customer engagement, increases conversion rates, and fosters brand loyalty, providing a competitive advantage in the e-commerce skincare market.
2. **Academic Value**:
   The project demonstrates the application of machine learning in personalized product recommendations, offering insights into data cleaning, recommendation algorithms, and AI-driven personalization.
3. **Social Value**:
   The system empowers consumers to make informed choices, improving skin health and well-being by providing tailored product suggestions.

Data Exploration and Preprocessing

As part of our exploration of the skincare product dataset, we have performed initial preprocessing tasks and conducted exploratory data analysis (EDA). The dataset contains key features such as:

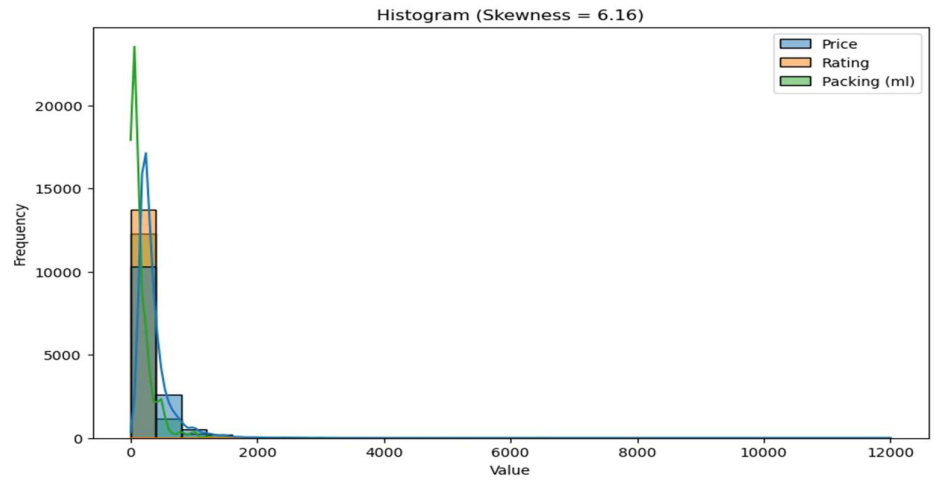
- **Price**
- **Rating**
- **Title** (which has been preprocessed using text cleaning techniques)
- **Packing**
- **Brand**
- **Product Type**

We have capped the **Price** variable at an upper limit of **3000** to mitigate the impact of extreme values and outliers.



The current skewness of the **Price** feature is **6.16**, which indicates a positively skewed distribution.

Skewness of the data: 6.16



To address this skewness, we will apply a **log transformation** on the **Price** feature to bring the distribution closer to normal, which should improve the performance of certain models (e.g., linear regression, clustering).
Additionally, **scaling** will be applied to ensure the transformed data is on a comparable scale to other features.

Feature Engineering

Future steps for feature engineering include:

Transformations:

- We will apply scaling and transformations on numerical features like **Price** and **Rating**. Since the data is skewed, applying a log transformation on these variables may help in reducing skewness. We can also explore the **Box-Cox** or **Yeo-Johnson** transformations for further skewness reduction.
- For the text data (from the **"Title"**), we've already applied **TF-IDF** to convert the raw text into meaningful numerical representations. However, we might explore the use of **word embeddings** (e.g., **Word2Vec** or **GloVe**) in the future to capture deeper semantic relationships between words and improve recommendation accuracy.

Scaling the Data:

- **Price** and **Rating** will be scaled using **MinMaxScaler** or **StandardScaler** to ensure that these features are within a comparable range for clustering and model training. This is particularly important for algorithms like **KMeans**, which are sensitive to the scale of the data.
- Additionally, we plan to scale any other numerical features (e.g., product attributes) that might be added to the dataset in the future.

Feature Selection:

- We will evaluate the importance of each feature using techniques such as **Variance Inflation Factor (VIF)**, feature importance from tree-based models (e.g., **Random Forest**), or **correlation analysis** to determine which features should be retained for model building.
- We may also conduct feature engineering for categorical variables such as **Brand** and **Packing**, potentially converting them into one-hot encoding or embedding vectors.

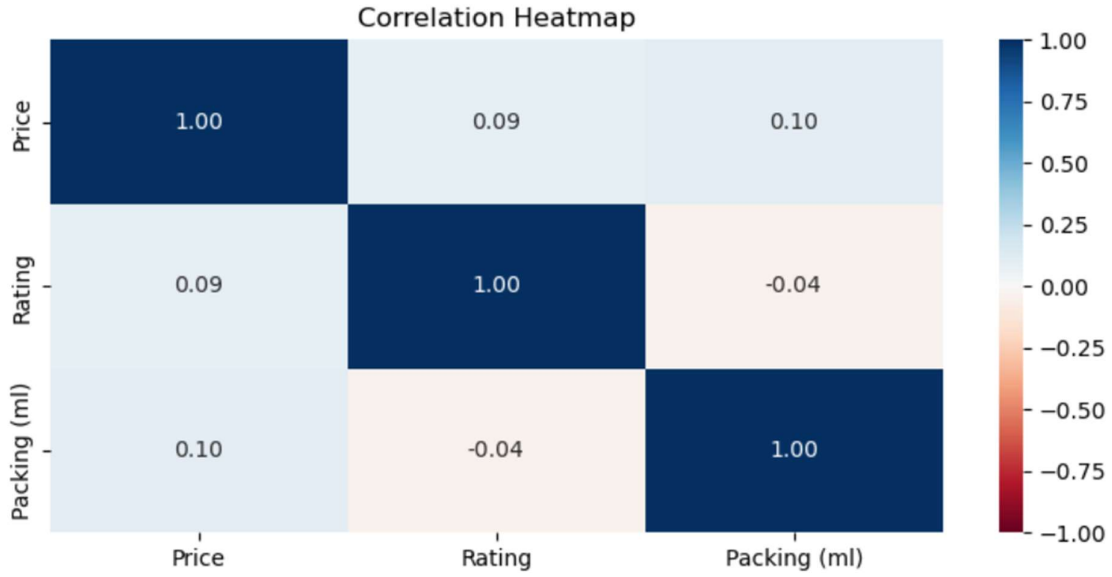
Dimensionality Reduction:

- Since we are working with high-dimensional data (especially after transforming text data into **TF-IDF** vectors), we will consider dimensionality reduction techniques like **Principal Component Analysis (PCA)** or **t-SNE** to reduce the number of features while retaining as much variance as possible. This can help improve clustering performance and make the model more efficient.

Clustering Assumptions:

PCA (Principal Component Analysis)**:**

1. **Multicollinearity**: PCA assumes that there is little to no multicollinearity among features. We will check for correlations between features, especially after transformation (e.g., TF-IDF), and remove or combine highly correlated variables if needed.



KMeans Clustering**:**

2. **Presence of Outliers**: KMeans is sensitive to outliers because it uses the mean of clusters to form the centroid. We will check for and handle outliers using methods like **IQR (Interquartile Range)** or **z-score** before applying clustering.
3. **Scaling**: KMeans requires features to be scaled to ensure that all features contribute equally to the clustering process. We will apply scaling (MinMax or StandardScaler) on numerical features, including **Price** and **Rating**.
4. **Conversion to Numerical Data**: KMeans requires numerical input, so any categorical variables (such as **Brand** or **Product Type**) need to be encoded using one-hot encoding or other appropriate methods. We will ensure that any categorical features are properly handled before applying the algorithm.