Analyzing and Predicting Purchase Behavior in the E-commerce Industry

Data Mining & Modeling  
  
  
  
  
  
  
Project By  
Awaghade, Vedanti Vikas  
Jumale, Amey  
Narayanan, Vaidehi  
Shahani, Mehek Kishore  
Sharma, Yash

# Introduction

This project aims to analyze customer behavior in the e-commerce industry by leveraging a dataset of customer reviews and ratings from Sephora, a leading skincare and beauty retailer. Through descriptive, predictive, and prescriptive analyses, we seek to understand purchase behavior, predict customer satisfaction, and provide actionable recommendations for improving business performance.

# Data Gathering

For this project, we selected the "Sephora Products and Skincare Reviews" dataset from Kaggle, a trusted platform for secondary data. The dataset provides detailed customer feedback, including reviews, ratings, and product details, making it ideal for our analysis.

* **Dataset Source**: Kaggle – by nadyinky (File: reviews\_0-250.csv)
* **Attributes Collected**:
  + Review Text: Customer feedback for sentiment analysis.
  + Rating: Numerical score (1–5) reflecting satisfaction.
  + Product Name & ID: Links reviews to specific products.
  + Skin Type: For personalization analysis.
  + Helpful Votes: Measures review usefulness.
  + Submission Time: Date of the review.
  + Price USD: Product price for monetary analysis.
* **Data Size**: The sample file contains 250 reviews, with the full dataset exceeding 5,000 reviews across skincare categories.

# Data Preprocessing

The dataset was preprocessed to ensure quality:

* Removed duplicate entries and irrelevant columns (e.g., unrelated metadata).
* Handled missing values by filling empty review texts with "" and dropping rows with missing critical fields (e.g., ratings).
* Converted submission\_time to datetime format for recency calculations.
* Created a new feature, review\_length, by counting words in each review.

A screenshot of a computer program

AI-generated content may be incorrect.

# Descriptive Analysis

Exploratory Data Analysis (EDA) provided key insights:

## Distribution of Ratings

The average rating is 4.31, with most reviews giving 5 stars, indicating high customer satisfaction.**A graph with different colored bars

AI-generated content may be incorrect.**

## Most Reviewed Products

Products like "Lip Sleeping Mask Intense Hydration with Vitamin C" received the highest number of reviews.

A graph of different colored bars

AI-generated content may be incorrect.

## Top Products by Rating

"Gentle Hydra-Gel Face Cleanser" had the highest average rating.

A graph with different colored bars

AI-generated content may be incorrect.

## Review Length

The average review length is 59.17 words, with a right-skewed distribution (most reviews are short, but some are lengthy).A graph of a distribution of words

AI-generated content may be incorrect.

## Keyword Frequency

Frequently used words in reviews include "skin," "product," and "love," highlighting key customer focuses (Visualization 6: Top 20 Most Frequent Keywords in Reviews)

A chart of a graph

AI-generated content may be incorrect.

# Predictive Analysis

**RFM Analysis**

RFM (Recency, Frequency, Monetary) analysis segmented customers:

* **Recency**: Days since the last review.
* **Frequency**: Number of reviews per customer.
* **Monetary**: Average product price reviewed.  
  This identified high-value customers (frequent reviewers of expensive products).

A screenshot of a number

AI-generated content may be incorrect.

## Sentiment Analysis

Using TextBlob, we calculated sentiment polarity (-1 to +1) for each review:

* Most reviews are positive (peak around 0.5), indicating overall satisfaction.
* Products with negative sentiment were flagged for further investigation.

A graph of a distribution of a number of points

AI-generated content may be incorrect.

## Predictive Model

A Random Forest Classifier predicted whether a customer would recommend a product:

* **Features**: Rating, sentiment, review length, price.
* **Accuracy**: 96.06%, showing strong predictive power.
* **Example Prediction**: A new review with a rating of 4, positive sentiment, and moderate length was predicted as "Will recommend."

## Cluster Analysis

KMeans clustering on RFM features identified three customer segments:

* One-time reviewers.
* Regular reviewers.
* High-value, frequent contributors.

# Multiple Regression Analysis

A multiple linear regression model predicted ratings based on:

* **Independent Variables**: Sentiment, review length, price.
* **Dependent Variable**: Rating.
* **Results**:
  + R-squared: 0.098 (9.8% of rating variance explained).
  + Sentiment had the strongest impact (coefficient: 1.8649, p < 0.001).
  + Review length and price had smaller but significant effects.
* **Residual Analysis**: Residuals showed a striped pattern due to discrete ratings, with no major violations of regression assumptions.

A graph with a blue line

AI-generated content may be incorrect.

## Regression Summary

|  |  |
| --- | --- |
| **Metric** | **Value** |
| R Squared | 0.098 |
| Adjusted R-Squared | 0.098 |
| Observations | 74697 |
| F-Statistic | 279.0 (P<0.001) |

This means that about 9.8 percent of the variation in ratings can be explained by the three predictors. While this value is relatively low, it is typical when analyzing human behavior, where many unmeasured factors can influence outcomes.

Coefficient Interpretation

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Coefficient** | **Significance**  **(p-value)** | **Interpretation** |
| Intercept | 3.825 | p < 0.001 | The baseline rating when all predictors are zero |
| Sentiment | 1.8649 | p < 0.001 | Higher sentiment (more positive tone) is strongly associated with higher ratings |
| Review Length | 0.001 | p < 0.001 | Longer reviews tend to slightly increase ratings |
| Price (USD) | 0.0007 | p < 0.001 | Products with higher prices are associated with slightly higher ratings |

Sentiment is the most influential factor in the model, with the largest and most significant effect on rating.

## Residual Analysis

The residuals plot (differences between actual and predicted ratings) showed a striped pattern, which is expected because ratings are often given as whole numbers (e.g., 1 to 5 stars). This pattern indicates that:

* The model fits well across the range of predicted values
* There is no strong sign of non-constant variance (heteroscedasticity)
* The model does not suffer from major violations of regression assumptions

A graph of a graph with blue and red lines

AI-generated content may be incorrect.

## Coefficient Visualization

The bar chart of regression coefficients visually confirmed the numerical results. Sentiment had the largest impact on rating, followed by review length and price. All predictors had positive coefficients.

A graph with purple lines and red dots

AI-generated content may be incorrect.

# Interpretation and Recommendations

**Key Findings**

* High ratings (average 4.31) indicate strong customer satisfaction, but sentiment analysis revealed areas for improvement in specific products.
* Sentiment is the most significant predictor of ratings, followed by review length and price.
* RFM and clustering identified high-value customers who can be targeted for loyalty programs.

**Recommendations**

1. **Product Improvement**: Focus on products with negative sentiment by addressing common complaints (e.g., "moisturizer" issues).
2. **Customer Engagement**: Target high-value customers with personalized offers based on their skin type and purchase history.
3. **Review Quality**: Encourage detailed reviews, as longer reviews correlate with higher ratings.
4. **Pricing Strategy**: Higher-priced products tend to receive slightly better ratings, suggesting a focus on premium product lines.

**7. Visualizations**

The analysis includes several visualizations:

* **Distribution of Ratings**: Histogram showing a peak at 5 stars.
* **Top 10 Most Reviewed Products**: Bar plot highlighting popular products.
* **Top 10 Products by Rating**: Bar plot of highest-rated products.
* **Rating by Skin Type**: Box plot showing rating variations.
* **Review Length Distribution**: Histogram with a right-skewed pattern.
* **Top 20 Keywords**: Bar plot identifying frequent terms like "skin," "product," and "love."
* **Sentiment Distribution**: Histogram showing a positive skew.
* **Regression Coefficients**: Bar plot emphasizing sentiment’s impact.
* **Residuals vs Fitted Values**: Scatter plot confirming model fit.
* **Actual vs Predicted Ratings**: Scatter plot validating regression predictions.

# Conclusion

This project successfully analyzed customer behavior in the e-commerce skincare industry using the Sephora dataset. Descriptive analysis revealed high satisfaction levels, predictive models accurately forecasted recommendations, and regression analysis highlighted sentiment as a key driver of ratings. The insights and recommendations can help Sephora enhance customer satisfaction and business performance.