**ADVANCED DATA MINING**

**FINAL PROJECT REPORT**

**SKIN CARE PRODUCTS: EDA AND SENTIMENT ANALYSIS**

**CHARISH YADAVALI (02151348)**

**THEOPILUS GUDAPATI (02147220)**

**1.Abstract:**

The skincare industry is rapidly expanding, generating vast amounts of data from customer reviews, product features, and ratings. This project aims to analyze this data to uncover actionable insights that help businesses understand customer preferences and product performance. Using exploratory data analysis (EDA) and sentiment analysis, the study identifies key patterns in pricing, ratings, and customer sentiment. Sentiment polarity scores are computed from product highlights using natural language processing (NLP) techniques, offering a deeper understanding of customer emotions and their impact on ratings.

Machine learning models, including Random Forest Regressor and Logistic Regression, are employed to predict product ratings and classify sentiment as positive or negative. Results reveal strong correlations between customer sentiment and product ratings, with mid-range priced products often receiving higher satisfaction scores. The study demonstrates the effectiveness of combining EDA, sentiment analysis, and predictive modeling to extract meaningful insights from skincare product data, providing a robust framework for businesses to optimize their strategies and improve customer satisfaction.

**2.Introduction:**

The global skincare industry is valued at billions of dollars and continues to grow rapidly due to increasing consumer demand for innovative and effective products. With this growth comes a massive influx of data in the form of product features, customer reviews, and sales figures. Analyzing this data is crucial for brands to understand customer preferences, refine product offerings, and remain competitive in the market.

Customer reviews, in particular, provide valuable feedback about product efficacy, quality, and customer satisfaction. However, the sheer volume of textual data can be challenging to process manually. By applying data analysis and machine learning techniques, this project seeks to uncover actionable insights from customer reviews and product data. Specifically, sentiment analysis can reveal customer emotions, while predictive modeling can identify key factors influencing product ratings.

Through this study, businesses can:

1.Identify patterns in customer feedback to improve their offerings.

2.Understand pricing strategies and their correlation with customer satisfaction.

3.Leverage machine learning models to predict product performance and classify customer sentiment.

This project focuses on a skincare product dataset, employing exploratory data analysis (EDA), sentiment analysis, and predictive modeling to generate insights that benefit both consumers and businesses.

**3.Related Work:**

Sentiment analysis and machine learning techniques have been extensively used in e-commerce and product review domains to derive insights about customer behavior and product performance. Below are some key contributions from the literature:

**Sentiment Analysis**

* **Tools and Techniques**: Sentiment analysis leverages natural language processing (NLP) tools such as TextBlob, VADER, and transformer-based models like BERT to analyze textual data. These tools assign polarity scores to text, indicating positive, negative, or neutral sentiments. Research has shown that sentiment scores often correlate with customer satisfaction and ratings.
* **Applications**: Studies in the e-commerce domain have demonstrated the importance of sentiment analysis in identifying factors influencing customer preferences. For instance, positive sentiment in reviews often correlates with higher ratings, while negative sentiment highlights areas for improvement.

**Exploratory Data Analysis**

* **Visualization**: Visual tools such as histograms, scatter plots, and word clouds have been widely adopted to identify patterns and trends in data. EDA enables analysts to understand data distributions, relationships between features, and key themes in customer feedback.
* **Key Findings**: Previous studies have used EDA to explore the impact of pricing strategies, product descriptions, and customer demographics on sales and satisfaction.

**Predictive Modeling**

* **Regression Models**: Regression techniques, including Random Forest and Linear Regression, have been used to predict product ratings and sales. These models analyze features such as price, reviews, and product attributes to forecast numerical outcomes.
* **Classification Models**: Logistic Regression, Naive Bayes, and Support Vector Machines (SVM) are commonly used for sentiment classification. These models have shown high accuracy in predicting binary outcomes, such as positive or negative sentiment.

**Insights from Previous Studies**

1. **Correlations**: Sentiment polarity often correlates with product ratings, offering insights into customer satisfaction levels.
2. **Feature Importance**: Features like product price, review count, and textual highlights significantly influence customer ratings.
3. **Challenges**: Handling large-scale textual data requires efficient preprocessing, including handling missing values, encoding, and dimensionality reduction.

This project builds upon these foundations by integrating sentiment analysis and machine learning into a unified workflow. The dataset used in this study provides a comprehensive view of skincare products, offering opportunities to explore new dimensions in sentiment and rating predictions.

**4.Problem definition:**

The skincare industry produces a vast amount of data, including product details, customer reviews, and ratings. Extracting actionable insights from this data is challenging due to its unstructured nature. This project aims to address the following problems:

1. Understanding the relationship between features like price, reviews, and sentiment polarity with product ratings.
2. Classifying customer sentiment as positive or negative based on product highlights.
3. Predicting product ratings using machine learning models to provide data-driven recommendations for businesses.

By tackling these challenges, the project seeks to help brands optimize product offerings, enhance customer satisfaction, and improve decision-making strategies.

**5.Methods and Methodology:**

This project adopted a systematic methodology to analyze skincare product data, beginning with data preprocessing. The dataset was cleaned by handling missing values through imputation techniques, addressing outliers in numerical features like price, and encoding categorical variables for analysis. Sentiment polarity scores were computed from product highlights using the TextBlob library, converting textual data into numerical sentiment values ranging from -1 (negative) to +1 (positive). Additionally, new features such as price ranges were created to enable deeper analysis.

Exploratory Data Analysis (EDA) was performed to uncover patterns and relationships within the data. Visualizations like histograms revealed the distribution of ratings and prices, while scatter plots highlighted correlations between sentiment polarity and ratings. Word clouds identified frequent themes in product descriptions, such as "hydrating" and "lightweight," indicating key customer preferences. This step provided critical insights into the overall structure of the dataset and informed the modeling approach.

Two machine learning models were implemented to achieve the project objectives. The Random Forest Regressor was used to predict product ratings based on features such as price, reviews, and sentiment polarity. Logistic Regression was employed to classify customer sentiment as positive or negative. Both models were trained using an 80/20 train-test split to ensure robust performance evaluation. Hyperparameters were tuned for optimal results, and feature importance analyses were conducted to identify key drivers of predictions.

Model evaluation focused on assessing accuracy and reliability. For the Random Forest Regressor, metrics such as R² Score and Mean Absolute Error (MAE) were used to measure prediction accuracy and error rates. Logistic Regression performance was evaluated through accuracy scores and confusion matrices, which provided insights into the model’s ability to distinguish between positive and negative sentiment. This comprehensive methodology ensured actionable insights were derived from the data to support business decision-making.

**6.Experimental settings:**

**1. Computing Environment**

* **Hardware**:
  + Processor: Intel Core i7 (or equivalent).
  + Memory: 16GB RAM.
  + Storage: SSD for faster data processing.
* **Software**:
  + Python (Version 3.8+).
  + Jupyter Notebook: For interactive coding and analysis.

**2. Tools and Libraries**

* **Data Manipulation**:
  + pandas and numpy: For data cleaning, manipulation, and feature engineering.
* **Visualization**:
  + matplotlib and seaborn: For histograms, scatter plots, and heatmaps.
  + WordCloud: For text-based visualizations of product highlights.
* **Sentiment Analysis**:
  + TextBlob: To compute sentiment polarity scores.
* **Machine Learning**:
  + scikit-learn: For implementing Random Forest Regressor and Logistic Regression models.

**3. Dataset Configuration**

* **Features**:
  + Numerical: price\_usd, reviews, sentiment polarity.
  + Categorical: primary\_category, brand\_name (encoded as numerical values).
* **Targets**:
  + Regression: rating (predicting product ratings on a scale of 1–5).
  + Classification: sentiment\_label (1 for positive sentiment, 0 for negative sentiment).
* **Data Split**:
  + Training Set: 80% of the dataset.
  + Testing Set: 20% of the dataset.

**4. Model Training and Parameters**

1. **Random Forest Regressor**:
   * Parameters:
     + n\_estimators=100: Number of decision trees.
     + max\_depth=None: No restriction on tree depth.
   * Evaluation Metrics:
     + R² Score.
     + Mean Absolute Error (MAE).
2. **Logistic Regression**:
   * Parameters:
     + Regularization: Default C=1.0.
   * Evaluation Metrics:
     + Accuracy.
     + Confusion Matrix.

**5. Reproducibility**

* **Random Seed**: A fixed seed (random\_state=42) was used to ensure consistent data splitting and parameter initialization.
* **Code Modularity**: The analysis was structured into reusable code blocks for data preprocessing, modeling, and evaluation.

**7.EXPERIMENTAL RESULTS AND ANALYSIS:**

**IN STOCK:**

**A screenshot of a computer program

Description automatically generated**

**A comparison of a graph

Description automatically generated with medium confidence**

In this analysis, we first looked at the number of products in stock, which totaled 7,868. Among these, we identified the top 5 in-stock products based on attributes such as price, rating, and reviews. For example, the Fragrance Discovery Set had a rating of 3.63 and 11 reviews, making it one of the popular products.

We also categorized the in-stock products. The categories with the highest product counts were Skincare, Makeup, and Hair, with Skincare alone having over 2,200 products. When looking at price ranges, most products fell into the $20–50 range, making it the most affordable and diverse price category. These insights help businesses focus on the most stocked categories and their pricing strategies.

**OUT OF STOCK:**

**A screenshot of a computer program

Description automatically generated**

**A screenshot of a graph

Description automatically generated**

we analyzed the out-of-stock products, which totaled 626. The top 5 out-of-stock products were also highlighted. For instance, the African Beauty Butter Mini Gift Set had a high rating of 3.56 but was out of stock, indicating strong demand.When categorized, Makeup had the highest number of out-of-stock products, followed by Fragrance and Skincare. Price analysis showed that most out-of-stock products were in the $20–50 range, highlighting that mid-range products are in high demand and often sell out quickly. This information is crucial for inventory management and restocking strategies.

**IN STOCK VS OUT OF STOCK ANALYSIS:**

**A screenshot of a graph

Description automatically generated**

This graph compares in-stock and out-of-stock products across categories. Green bars represent in-stock items, while red bars indicate out-of-stock items. Categories like Skincare, Makeup, and Hair have the highest number of in-stock products but also show significant out-of-stock counts, reflecting strong demand. Smaller categories like Tools & Brushes and Gifts have minimal out-of-stock items, likely due to lower demand or better stock management. This analysis highlights the need for improved inventory strategies, especially in high-demand categories like Makeup and Skincare, to minimize stockouts and meet customer expectations.

**Average Sentiment by Category and Sentiment vs. Ratings:**

**A screenshot of a computer code

Description automatically generated**

**A screenshot of a computer code

Description automatically generated**

**A screenshot of a graph

Description automatically generated**

the average sentiment by category and the correlation between sentiment polarity and ratings. Categories like Hair and Skincare had the highest average sentiment, reflecting positive customer feedback in these areas. Meanwhile, categories like Makeup and Gifts had slightly lower sentiment, indicating room for improvement.

On the right, we analyzed the relationship between sentiment polarity and ratings. The scatter plot demonstrates a positive correlation, with higher sentiment scores often corresponding to higher ratings. This emphasizes that customer sentiment is a reliable predictor of product satisfaction.

**Top Positive and Negative Sentiment Highlights:**

A screenshot of a computer

Description automatically generated

Here we extracted the top positive and negative sentiment highlights to better understand customer feedback. Positive highlights often referenced attributes like being award-winning, long-wearing, and free from harmful chemicals. For instance, the Madagascar Vanilla Perfume Oil Rollerball had a perfect sentiment score of 1.0.

Conversely, negative highlights frequently mentioned issues like unsatisfactory finishes or poor performance. Products like the Lip Fetish Sheer Colour Balm scored lower, with sentiment polarity of -0.166. These insights can guide brands in improving their product descriptions and addressing common customer concerns.

**Random Forest Regression Results:**

**A screenshot of a computer program

Description automatically generated**

**A graph showing a number of blue dots

Description automatically generatedA graph of a distribution of a forest

Description automatically generated**

Here, we analyzed the performance of the Random Forest Regressor for predicting ratings. On the left, the scatter plot compares actual vs. predicted ratings, showing limited predictive power. The R² score was very low at 0.0128, indicating that the model struggled to explain the variance in ratings.On the right, the residuals distribution is shown. The residuals are mostly concentrated around zero, but the spread indicates inconsistency in predictions. This suggests the need for additional features or more advanced models to improve the predictive performance.

**Logistic Regression Results:**

**A screenshot of a computer program

Description automatically generated**

**A blue and white graph

Description automatically generatedA graph with purple lines and numbers

Description automatically generated**

For sentiment classification, Logistic Regression was applied. On the left, the confusion matrix shows the model's performance in predicting positive and negative sentiment. While it performed well on positive sentiment, with 551 correctly predicted cases, it struggled with negative sentiment, only predicting 78 correctly. The histogram on the right represents the predicted probability distribution for positive sentiment. The model shows a clear clustering of probabilities, but further tuning or additional data could enhance its performance for balanced sentiment prediction.

**Support Vector Machine (SVM) Results:**

**A screenshot of a computer program

Description automatically generatedA blue squares with white text

Description automatically generatedA graph with a red line

Description automatically generated**

Support Vector Machine (SVM) was also used for sentiment classification. The confusion matrix on the left highlights similar trends as Logistic Regression, with better performance on positive sentiment (523 correctly predicted cases) compared to negative sentiment (106 correct predictions).The decision function distribution on the right shows how SVM classified sentiments, with most predictions clustering in specific decision boundaries. Overall, SVM performed slightly better than Logistic Regression in handling imbalanced sentiment data."

**Model Comparison: Accuracy Scores:**

**A screenshot of a computer

Description automatically generated**

"Finally, we compared the accuracy scores of the three models. Random Forest Regressor had the lowest score due to its poor predictive power, while Logistic Regression and SVM performed comparably, with accuracy scores of 62.9% and 59.7%, respectively. This highlights that while sentiment classification models performed reasonably well, there is room for improvement in the regression model to better predict product ratings."

**8.FUTURE WORK:**

This project provides valuable insights into skincare product data through sentiment analysis and predictive modeling, but there are several opportunities for further exploration and improvement. Expanding the scope of sentiment analysis to include multilingual reviews would allow for a broader understanding of customer sentiment across different regions and demographics, addressing the linguistic diversity of global markets. Additionally, employing advanced natural language processing (NLP) techniques, such as transformer-based models like BERT or GPT, could help capture more nuanced emotions and contextual meanings in product highlights and reviews, significantly improving sentiment classification accuracy.

Clustering techniques could also be utilized to group products based on features such as pricing, sentiment, and ratings, enabling businesses to identify niche markets and tailor marketing strategies to specific customer segments. Incorporating time-series analysis to examine temporal trends in customer reviews and ratings would further enhance understanding of how customer sentiment evolves over time, especially in response to product launches or marketing campaigns. Enriching the dataset with external data, such as competitor pricing, promotional offers, and seasonal trends, would provide a more comprehensive analysis of factors influencing customer satisfaction.

Lastly, deploying the trained machine learning models into a recommendation system or business intelligence tool would enable real-time predictions for new products, aiding in dynamic decision-making. This would allow businesses to leverage predictive insights in practical applications, enhancing customer satisfaction and optimizing product offerings in real time.

**9.CONCLUSION:**

This project analyzed skincare product data to uncover insights into customer preferences, sentiment, and product ratings. Using exploratory data analysis (EDA), patterns in pricing, ratings, and sentiment polarity were identified, providing a deeper understanding of customer feedback. Sentiment analysis revealed the overall positive or negative sentiment associated with product highlights, enabling sentiment classification.

The machine learning models implemented in this study provided foundational insights but with varying levels of predictive performance. The Random Forest Regressor yielded an R² Score of 0.01, indicating limited explanatory power for predicting product ratings based on features such as price, reviews, and sentiment polarity. This result suggests that additional features or advanced modeling techniques may be required to improve rating prediction accuracy.

The Logistic Regression model achieved an accuracy score of 62.9% in classifying sentiment as positive or negative. The model exhibited strong recall for positive sentiment (93%), indicating its ability to correctly identify positive cases, but lower performance for negative sentiment (recall: 12%). This imbalance suggests the need for further tuning, balanced datasets, or alternative classification models to enhance performance for both sentiment classes.

Despite the challenges in model performance, the analysis highlights the potential of integrating sentiment analysis and predictive modeling to extract meaningful insights from customer reviews. Future work could address model limitations by exploring advanced techniques, incorporating additional features, and experimenting with alternative algorithms for better predictive accuracy and classification balance.

**10.REFERENCES:**

<https://www.kaggle.com/code/melissamonfared/skincare-products-eda-sentiment-analysis/notebook>

<https://www.researchgate.net/publication/379007374_Statistical_and_sentiment_analysis_based_on_comments_of_skin_care_products>