Capstone Project Report

Product Performance & Customer Sentiment Analysis on Amazon

Team Members:

- **Shivathmika Kalvala**

- **Ahmed Tawfeeq P S**

- **Shubham Singh Chouhan**

-

Batch Code: DA464S46

Project Guide: Komilla Bhatia

1. Executive Summary

- The project analyzes Amazon product sales and customer feedback to uncover patterns in customer behavior and identify high-performing products.
- More than 1,000 Amazon products were assessed based on pricing, ratings, reviews, and discounts.
- Tools like Python, SQL, Excel, and Tableau were used for EDA, modeling, and visualization.
- Products with high discounts still maintained high ratings, indicating trust and satisfaction.
- Key insights guide businesses in optimizing pricing, product promotion, and inventory management.

2. Introduction

Problem Statement: With millions of products on Amazon, customers struggle to identify high-quality products worth purchasing.

Objectives:

- Analyze customer sentiment through ratings and reviews.
- Evaluate the impact of discounts on product performance.
- Help sellers identify top-rated and top-selling products.

3. Data Description

Source: Amazon product dataset with over 1,000 entries.

Industry: E-commerce/Retail.

Duration: Snapshot data (single time point).

Fields Used: Product name, category, price (actual/discounted), rating, review count, customer feedback.

Capstone Project Report

4. Data Preprocessing

- Cleaned missing/incomplete entries.
- Converted prices and ratings to numeric formats.
- Encoded categorical data (product categories).
- Engineered features like discount %, sentiment score.
- Scaled price data for modeling.
- Split data into training/test sets (e.g., 80:20).

5. Tools and Techniques Used

- Python: Data cleaning, Z-tests, modeling (Logistic, Linear Regression, KNN, K-Means).
- SQL: Data extraction, filtering (category & price range).
- Excel: Pivot tables, charts (sales, ratings).
- Tableau: Dashboards for sales performance and discount impact.

6. Exploratory Data Analysis (EDA)

- Top-selling product: Redmi 9 Activ
- Highest-rated product: Aquadpure Copper + Mineral Water Filter- Categories with highest sales:

  Electronics, Computers & Accessories
- Discounted products contribute >97% of total sales.
- Visualization: bar charts, pie charts, histograms.

7. Modeling

- Model | Purpose | Accuracy
- Linear Regression | Predict customer ratings | 90%
- Logistic Regression | Classify products | 74%
- KNN | Classify product categories | 84%
- K-Means Clustering | Segment products by price | Silhouette Score: 71%

- Justified model choice based on problem type.
- Applied feature selection and standard model evaluation metrics.

8. Results and Evaluation

- Linear regression gave highest predictive accuracy.

- KNN useful for automating category classification.
- K-Means helped identify pricing clusters.
- Evaluation based on accuracy, Silhouette score, etc.

9. Business Implementation of ML Models

- Linear Regression: Predict satisfaction -> improve product development.
- Logistic Regression: Classify products -> automate catalog management.
- KNN: Automate tagging -> better search experience.
- K-Means: Price segmentation -> targeted marketing strategies.

10. Summary of Insights

- Electronics dominate sales.
- Discounts don't reduce ratings.
- Top products have both high ratings and sales.
- SQL search enhances user experience via category and price filters.

11. Conclusion

- Strong correlation between discounts and sales.
- High-rated products like Aquadpure Water Filter reinforce brand trust.
- Discounts can be leveraged without reducing product perception.
- Modeling validates data-driven product strategies.

12. Limitations and Future Scope

- Dataset is static; time-series data would improve trend analysis.
- No customer demographics included.
- Future improvements: Add personalization, expand features, deploy models in real-time.

13. Appendices

- Screenshots of SQL code, Python Z-test, ML outputs.
- Pivot tables used in Excel for dashboards.

- Tableau visual links (if interactive version available).

14. Bibliography and References

- Amazon Dataset (Kaggle/UCI/Other)
- Python: Scikit-learn, Pandas, Matplotlib
- SQL (MySQL)
- Tableau Public
- Academic papers/articles on discount pricing and sentiment analysis
- AI tools used: ChatGPT, Python Notebooks