Group Members: Sai Hemanth Kilaru, Sri Ram Theerdh Manikyala

Course Title: INFO511 Foundations in Data Science

Term name and year: Spring 2025

Submission: Final Project Report

Instructor’s Name: Angela Cruze

Date of Submission: 05-11-2025

1. Introduction

Understanding what drives consumer recommendations in the electronics industry is crucial for both product development and marketing strategies. This project explores factors that influence recommendation percentage using a structured dataset from Best Buy's electronics catalog. The objective is to identify the strongest predictors of recommendation behavior and detect temporal trends in product releases and pricing. Two research questions guide the analysis, supported by exploratory visualizations and statistical modeling.

2. Dataset Overview

The dataset, retrieved from a public Databricks repository, includes around **1,000 product records** across categories like appliances, video games, and tech accessories. Despite the limited row count, each entry contains over 35 detailed fields such as:

- **Pricing**: initial_price, final_price, discount_percentage
- **Feedback Metrics**: rating, reviews_count, questions_count, recommend_percentage
- **Categorical Info**: root_category, esrb_rating, availability
- **Text Features**: product_description, features, q-a

This granularity makes the dataset well-suited for statistical analysis, trend detection, and natural language insights.

3. Methodology

Data Cleaning & Preprocessing

- Removed symbols from price fields and converted to float
- Parsed release_date and derived year, month
- Imputed missing numerical values using **median**, categorical with **mode**
- Engineered features: discount_percentage, description_length, price_bin

Exploratory Data Analysis (EDA)

- Used **Plotly** for interactive plots: distributions, scatter plots, heatmaps
- Applied **WordCloud** for textual insights
- Built two HTML dashboards (Q1 and Q2) for visualization

Statistical Modeling

- Used **OLS**, **Ridge**, and **Lasso Regression** to model recommendation percentage
- Forecasted future trends using **ARIMA time series modeling**
- Extracted text sentiment using **zero-shot classification (BART)** via Hugging Face

Evaluation Metrics

- **R^2 Score** and **Mean Squared Error (MSE)** for model performance
- **Correlation matrix** to validate feature relationships
- VIF used to check multicollinearity among predictors

4. Research Question 1

What Factors Influence Recommendation Percentage?
(See Dashboard 1:https://drive.google.com/file/d/19gL8whv6y71L0vojGxgRWJm3evmsbyM-/view?usp=sharing)

We explored multiple numeric features to assess their effect on recommend_percentage. Key plots and findings:

- **Correlation Heatmap**: Rating showed the strongest positive correlation with recommendation %; price showed weak correlation.
- **Rating vs. Recommendation %**: A strong upward trend confirmed the influence of user ratings.
- **Price vs. Rating (by Category)**: No clear pattern; pricing did not significantly impact user satisfaction.
- **Bubble Chart (Rating vs. Reviews)**: Products with high ratings and many reviews had larger recommendation bubbles.

Rating is the most influential factor. Review count reinforces trust. Price has minimal direct impact. A rare insight was that **products with negative sentiment consistently had lower recommendation percentages**, highlighting the value of text analysis.

5. Research Question 2

Are There Seasonal or Temporal Patterns in Product Launches and Pricing?
(See Dashboard 2: https://drive.google.com/file/d/1iK_tddnhjKrUumrQX0-WY7nfK3CQA3eW/view)

We examined how product launches and pricing evolved over time.

Key visualizations and insights:

- **Monthly Product Release Trends**: Peaks in April and November align with real-world launch cycles (e.g., back-to-school, holiday prep).
- **Yearly Average Price Trends**: Showed a gradual decline post-2020, likely due to competitive pricing or product bundling.
- **Yearly Average Ratings**: Ratings remained stable, with a slight upward trend — possibly indicating better product-market fit over time.
- **ARIMA Forecast**: Modeled future product launches based on monthly data. While some fluctuations were predicted, the overall trend was stable, suggesting maturity in release cycles.
  There is clear **seasonality** in release timing, and average prices show **declining trends over years**, reflecting market adjustments and pricing strategies.

6. Challenges Faced

- **JSON Fields** (e.g., availability, q_a) needed parsing and flattening before use
- Some products lacked key fields like rating or release_date, requiring imputation
- Visual overload: Limited bar charts and histograms in favor of diverse formats like **bubble plots, parallel coordinates, ARIMA visualizations**

7. Conclusion

This project demonstrates how structured product data can offer valuable insights into customer recommendation behavior and market seasonality. Key findings include the **importance of ratings**, the **amplifying effect of review count**, and **seasonal spikes** in product releases. Statistical modeling validated the strength of these relationships. The project also highlights the use of **language models** in extracting sentiment from product descriptions — a unique blend of structured and unstructured analysis.

8. Future Scope

- Incorporate **customer review text** for sentiment + keyword extraction
- Extend analysis to multiple vendors for broader trends
- Use clustering to identify **product persona groups**
- Forecast **category-specific recommendation growth**
- Real-time dashboard deployment using **Streamlit** or **Tableau**