Information Retrieval

ASSIGNMENT-3 REPORT

Neev Swarnakar 2020390

Report Overview:

This report presents the approach, methodologies, assumptions, and results for creating a product recommendation system based on Amazon reviews focusing on ***smartphones***. The goal was to develop a predictive model using collaborative filtering techniques to recommend relevant items to users.

Methodologies Used:

1. **Data Acquisition and Preprocessing:**
    - Acquired Amazon Reviews Dataset for Electronics category.
    - Selected the smartphone subset for analysis.
    - Preprocessed the data by handling missing values, duplicates, and other data-cleaning tasks.
    - Divided the dataset into train and test data for model evaluation.

2. **Descriptive Statistics:**
    - Calculated key statistics such as the number of reviews, average rating score, unique products, and rating distribution (good, bad, average).
    - Identified the most positively reviewed smartphone and analysed ratings over consecutive years.

3. **Text Preprocessing:**
    - Applied text preprocessing techniques, including HTML tag removal, accented character handling, acronym expansion, unique character removal, lemmatisation, and text normalisation.
    - Transformed the review text into a format suitable for machine learning models.

4. **Exploratory Data Analysis (EDA):**
    - Analysed the top 20 most and least reviewed brands to understand market trends and user preferences.
    - Visualised word clouds for good and bad ratings to identify common sentiments and keywords.

- Created a distribution pie chart for ratings to visualise sentiment distribution.
- Examined trends in review counts over consecutive years to identify patterns and market shifts.

5. **Feature Engineering:**
   - Used the Bag of Words model to transform review text into numerical features for machine learning models.

6. **Machine Learning Models:**
   - Utilised five machine learning models (Random Forest, Support Vector Machine, Multinomial Naive Bayes, Logistic Regression, K-Nearest Neighbors) to predict rating classes (good, bad, average) based on review text.
   - Evaluated model performance using precision, recall, F1-score, and support metrics for each rating class.

7. **Collaborative Filtering:**
   - Constructed a user-item rating matrix and normalised ratings using min-max scaling.
   - Developed user-user and item-item recommender systems using cosine similarity.
   - Evaluated recommender systems' performance using Mean Absolute Error (MAE) for different neighbourhood sizes (N).

8. **Top 10 Products by User Sum Ratings:**
   - Ranked products based on the sum of user ratings to identify top-performing items in the smartphone category.

9. **Documentation and Reporting:**
   - Maintained thorough documentation of code and analysis steps for clarity and reproducibility.
   - Compiled results, insights, and visualisations into a comprehensive report following the specified format for submission.

Assumptions:

1. **Review Quality:**
    - Assumed that the quality of reviews, including their relevance and authenticity, is consistent across the dataset.
    - Assumed that user ratings accurately reflect their satisfaction levels with the products.

2. **Data Completeness:**
    - Assumed that the dataset is complete and representative of the overall smartphone market on Amazon.
    - Assumed that missing values and duplicates have been appropriately handled during preprocessing without significant impact on analysis outcomes.

3. **Rating Classification Threshold:**
    - Set a threshold of >=3 for "Good" ratings and consider the rest as "Bad" ratings for classification purposes.
    - Assumed this threshold effectively distinguishes between positive and negative sentiments in reviews.

4. **Text Preprocessing Impact:**
    - Assumed that text preprocessing techniques such as lemmatisation, special character removal, and text normalisation improve the quality of text data for machine learning models.

5. **Model Performance:**
    - Assumed that the chosen machine learning models (Random Forest, SVM, Naive Bayes, Logistic Regression, KNN) are suitable for sentiment analysis and rating classification based on review text.
    - Assumed that collaborative filtering techniques such as user-user and item-item recommendation systems provide accurate and relevant recommendations.

6. **Word Cloud Analysis:**
    - Assumed that word clouds effectively capture the most commonly used words in good and bad reviews, providing valuable insights into sentiment analysis.

7. **Model Generalization:**
    - Assumed that the performance metrics (precision, recall, F1-score, MAE) accurately assess model performance across different scenarios.

Results and Observations:

These results and observations of this assignment provide a comprehensive understanding of user sentiments, market trends, and model performance in the smartphone product category. They serve as valuable insights for decision-making prospects in product recommendation systems and sentiment analysis. Below are the results obtained:

1. **Data Overview:**
    - Total number of rows for the 'smartphone' product: **18358**
    - Total number of rows after preprocessing: **481**

2. **Descriptive Statistics:**
    - Number of reviews: **481**
    - Average rating score: **4.40**
    - Number of unique products: **428**
    - Number of good ratings: **451**
    - Number of bad ratings: **30**
    - Number of reviews corresponding to each rating:
      1.0: 16
      2.0: 14
      3.0: 44
      4.0: 96
      5.0: 311

3. **Top Brands Analysis:**
    - Top 202 reviewed brands:
      B00VWJOK7M 5
      B013D2ULO6 4
      B018HB1GW4 3
      B016F3M7OM 3
      B01FDPW1NK 3
      B0167Q104K 2
      B00N41UTWG 2
      B00N2KD9KI 2
      B00MUTWLW4 2
      B011CS01P2 2

B011IH6COQ 2
B00MI48ILY 2
B0153RGFG2 2
B0149QBOF0 2
B00NEYHIHM 2
B015WALYMK 2
B00HNJWT9G 2
B00HITWPYA 2
B00BUSDVBQ 2
B00AYAZENY 2

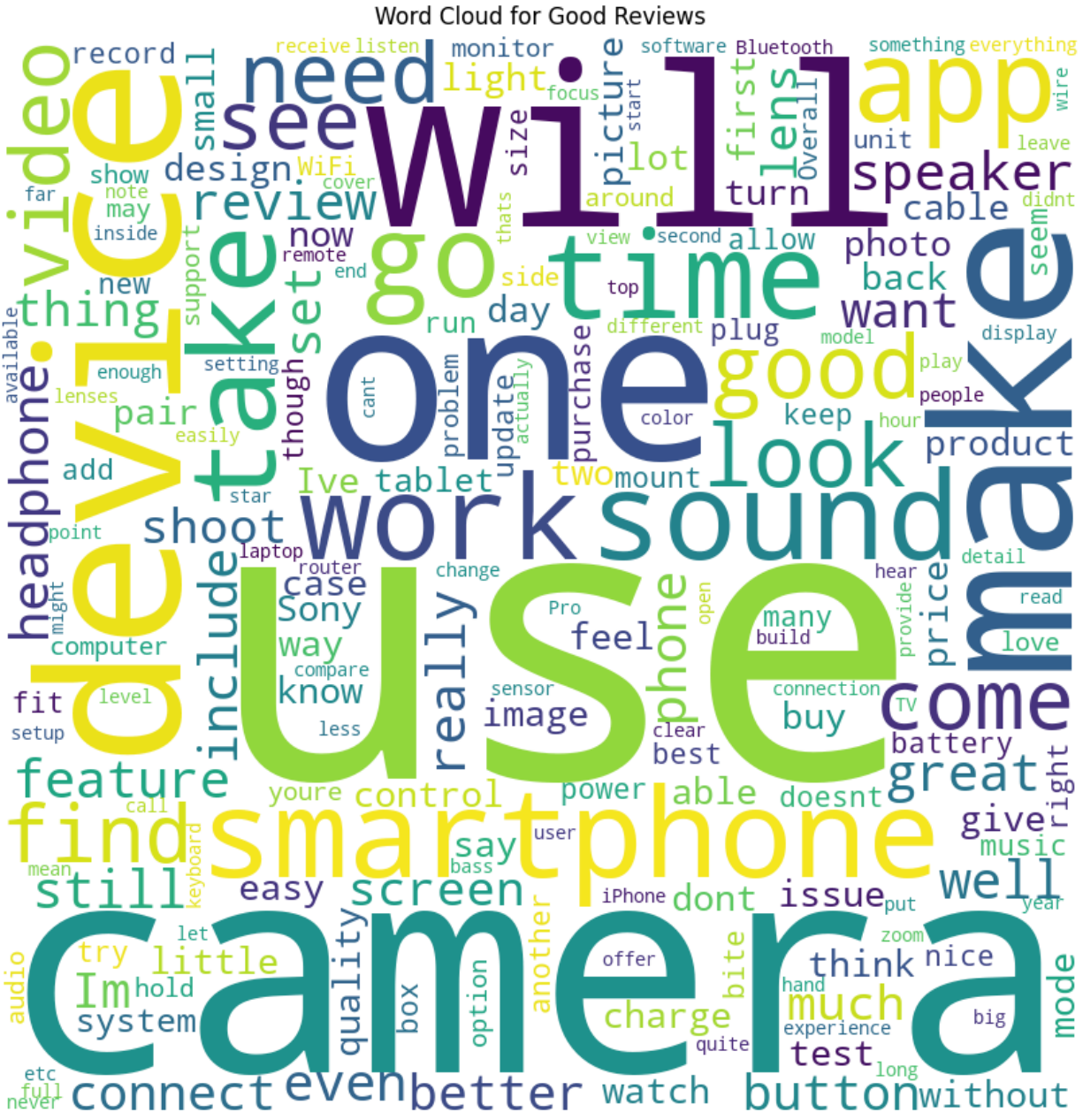
- Top 20 least reviewed brands:
  B00N1BRWLA 1
  B00N0NJEF6 1
  B00MYHOD5A 1
  B00MVRS36S 1
  B00MQOBJHQ 1
  B00MITLPX2 1
  B00MIWRGY6 1
  B00MCCN8E4 1
  B00MBFYUGM 1
  B00M6XTUPU 1
  B00M6UC5B4 1
  B00M1NEUKK 1
  B00LY8JVZ2 1
  B00LTMPOUO 1
  B00LR4OF5Y 1
  B00LP6CFEC 1
  B00LN3LQKQ 1
  B00LJ07JOU 1
  B00LAJQVR6 1
  B01E4I8I2U 1

  Most positively reviewed headphone: **B00006B81E**

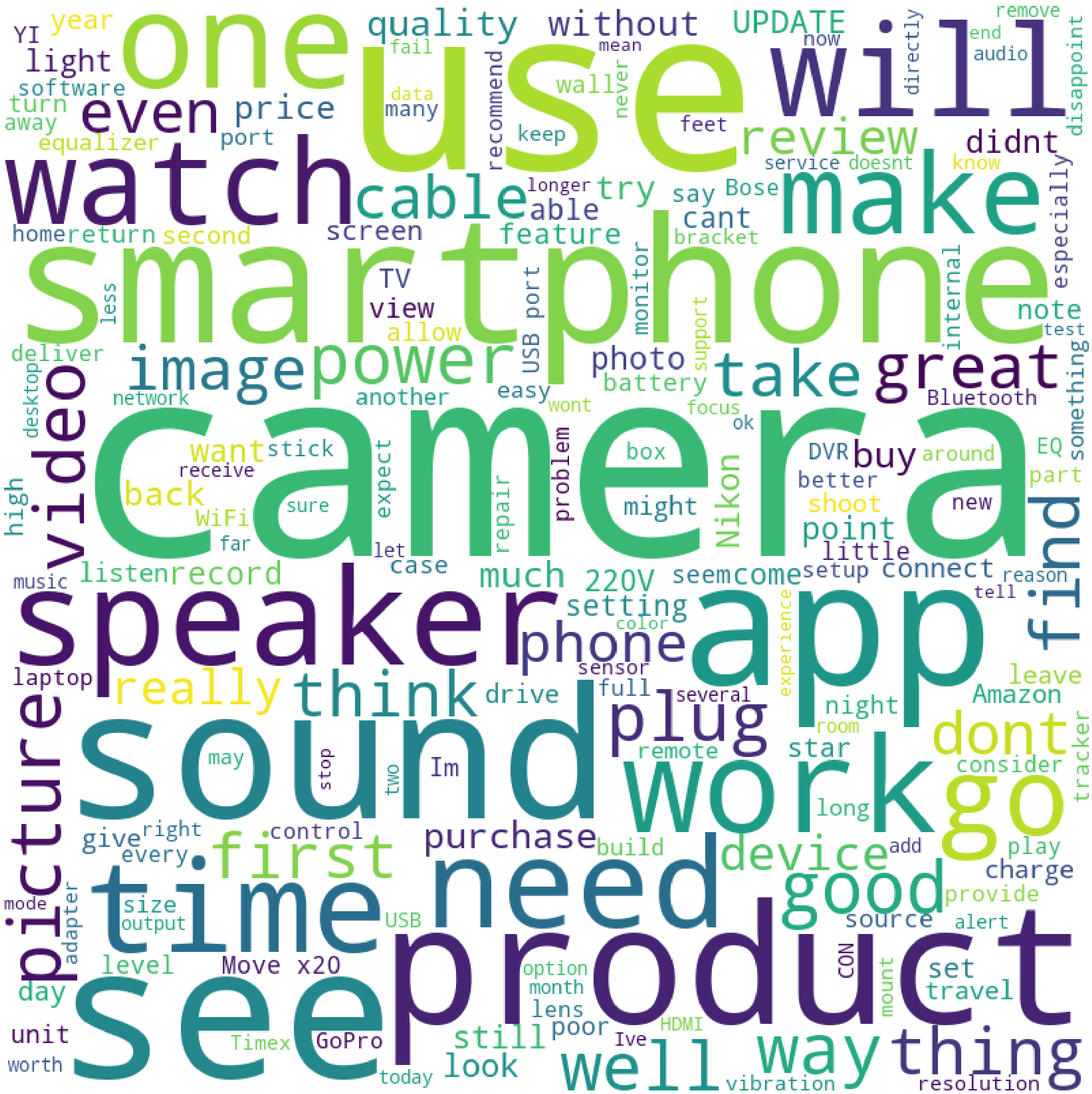
4. **Rating Trends Over Time:**
   - Count of ratings for the product over five consecutive years:
     Year No. of reviews
     2010 1
     2011 8

2012 21
2013 23
2014 26

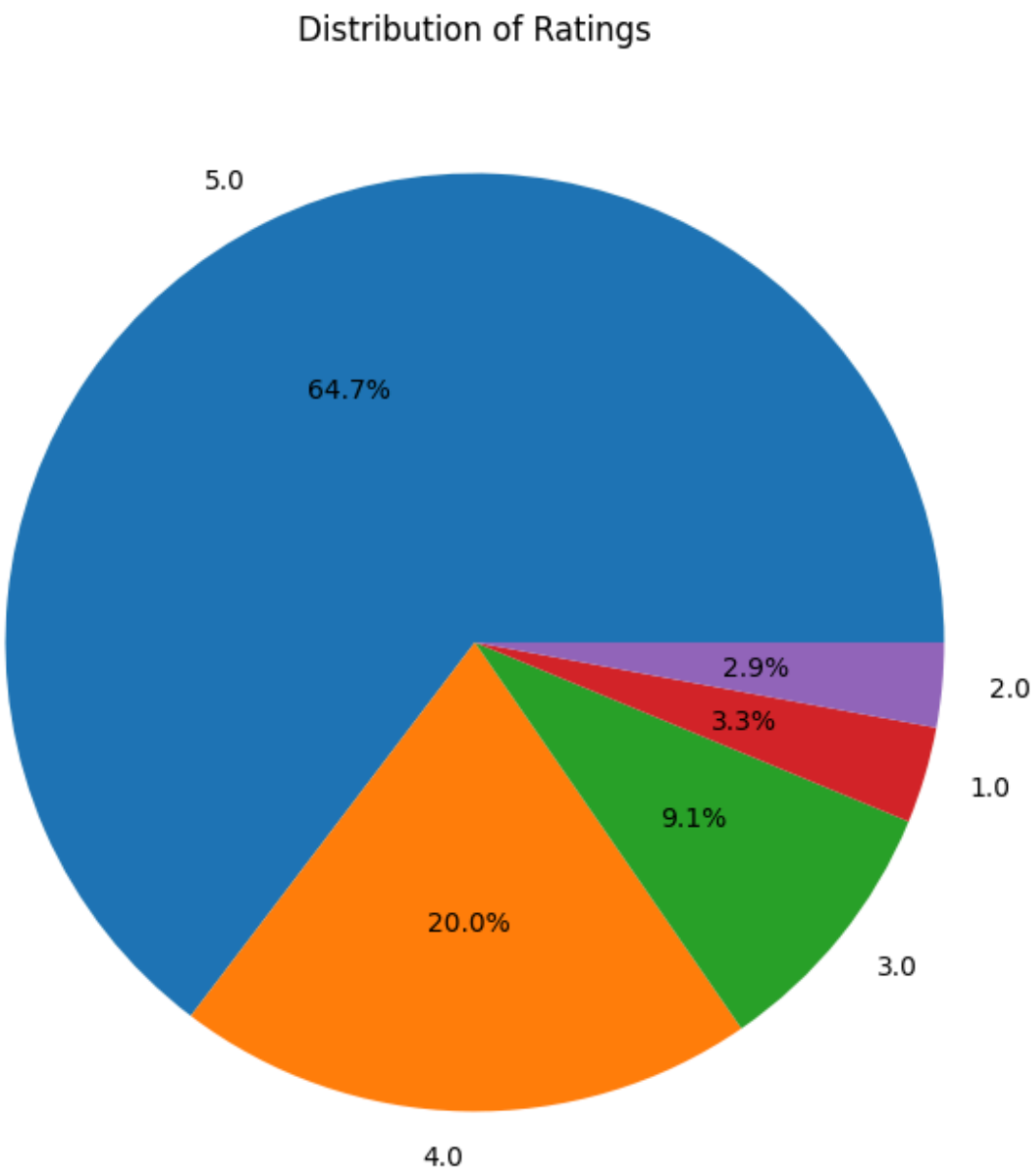
5. Word Cloud Analysis:



Word Cloud for Bad Reviews



6. **Distribution of Ratings:**
    - Pie chart for Distribution of Ratings vs. No. of Reviews:



    - **Good: 451 reviews.**
    - **Average: 44 reviews.**
    - **Bad: 30 reviews.**

7. **User Engagement Insights:**
    - Year with the maximum reviews: **2016**
    - Year with the highest number of customers: **2016**

- **Insights:** Indicates peak activity and user interest in smartphones during 2016.

8. **Machine Learning Model Performance:**
   - Random Forest:

	precision	recall	f1-score	support
Average	1.00	0.10	0.18	10
Bad	0.00	0.00	0.00	6
Good	0.88	1.00	0.93	105
accuracy			0.88	121
macro avg	0.62	0.37	0.37	121
weighted avg	0.84	0.88	0.82	121

   - Support Vector Machine:

	precision	recall	f1-score	support
Average	0.00	0.00	0.00	10
Bad	0.00	0.00	0.00	6
Good	0.87	1.00	0.93	105
accuracy			0.87	121
macro avg	0.29	0.33	0.31	121
weighted avg	0.75	0.87	0.81	121

   - Multinomial Naive Bayes:

	precision	recall	f1-score	support
Average	0.00	0.00	0.00	10
Bad	0.00	0.00	0.00	6
Good	0.87	1.00	0.93	105
accuracy			0.87	121
macro avg	0.29	0.33	0.31	121
weighted avg	0.75	0.87	0.81	121

- Logistic Regression:

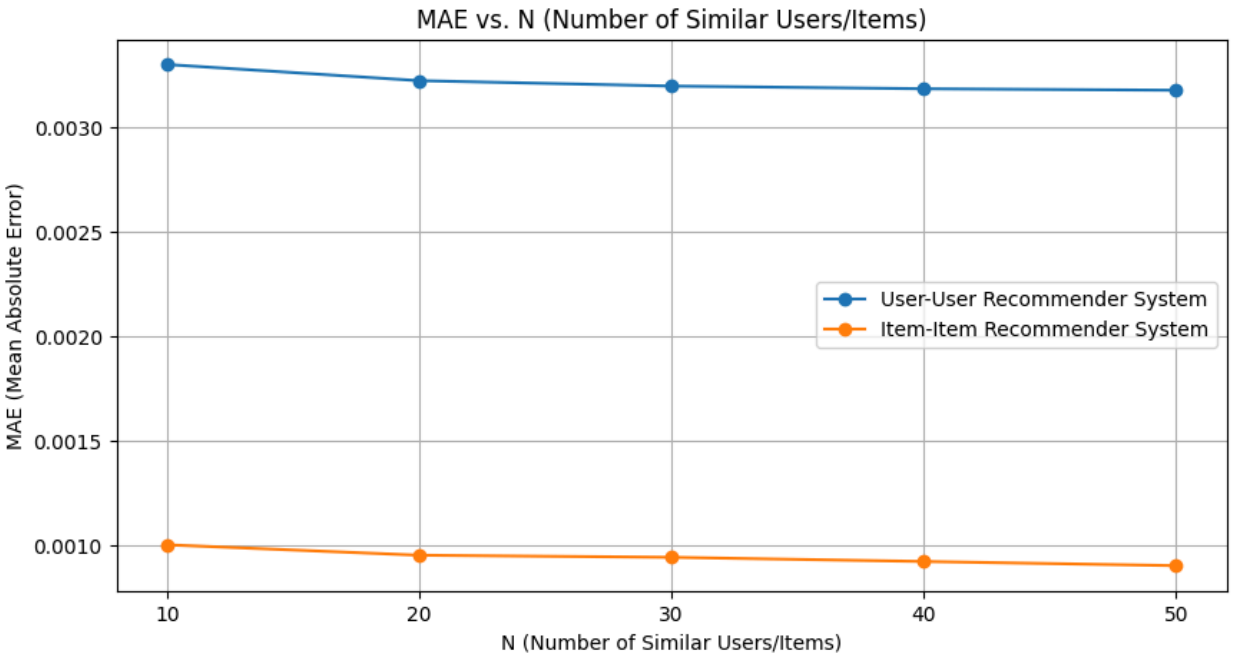
	precision	recall	f1-score	support
Average	0.00	0.00	0.00	10
Bad	0.00	0.00	0.00	6
Good	0.87	1.00	0.93	105
accuracy			0.87	121
macro avg	0.29	0.33	0.31	121
weighted avg	0.75	0.87	0.81	121

- K-nearest Neighbours:

	precision	recall	f1-score	support
Average	0.00	0.00	0.00	10
Bad	0.00	0.00	0.00	6
Good	0.87	0.98	0.92	105
accuracy			0.85	121
macro avg	0.29	0.33	0.31	121
weighted avg	0.75	0.85	0.80	121

9. **Top 10 Products by User Sum Ratings:**
   - List of products ranked by the sum of user ratings:

B00VWJOK7M 4.75
B013D2ULO6 3.75
B018HB1GW4 3.00
B00009K79U 2.00
B00RY1Z9NQ 2.00
B00T85PH2Y 2.00
B00ZDWGFR2 2.00
B004YDUZ22 2.00
B00MI48ILY 2.00
B00N2KD9KI 2.00



- **Insights:** Identifies top-performing products based on user feedback and ratings.