Product Recommendation System Based on Amazon Review

Objective

The goal of this project is to build a predictive model for forecasting user ratings and evaluating the usefulness of reviews. The system employs collaborative filtering techniques to recommend relevant products and personalizes user experiences.

---

1. Data Pre-processing

Data Loading

- The Amazon Electronics 5-core dataset and associated metadata were downloaded and loaded into Pandas DataFrames.
- Metadata was kept in a separate DataFrame.

Data Cleaning

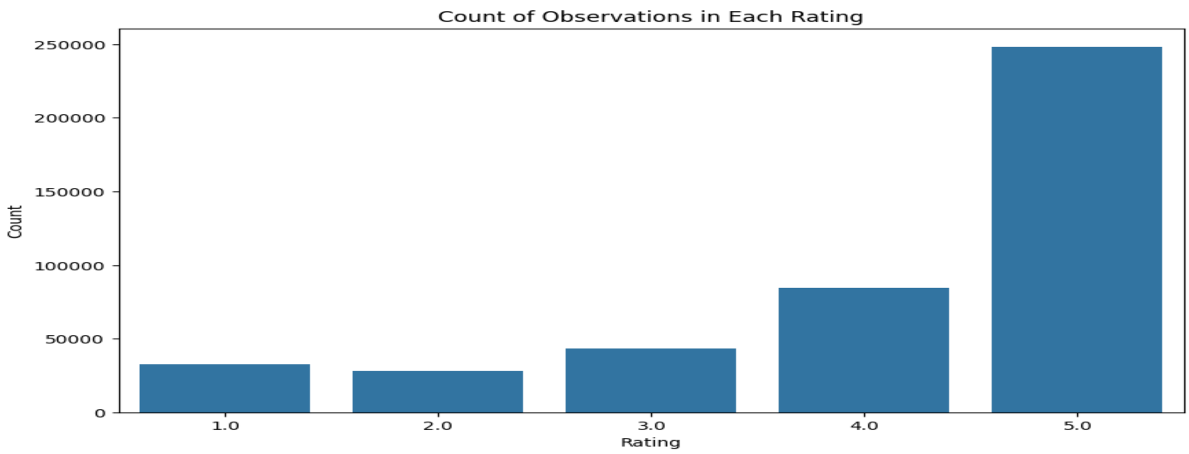
- Handled missing values by dropping rows with `NaN` values.
- Removed duplicates and unnecessary columns.
- Focused on a specific category: **Headphones**.

---

2. Exploratory Data Analysis

Descriptive Statistics for Headphones

- Total Reviews: **411,152**
- Average Rating: **4.11**
- Number of Unique Products: **26,849**
- Good Ratings (≥3): **353,373**
- Bad Ratings (<3): **57,779**
- Distribution of Ratings:
  - 1-star: **30,991**
  - 2-star: **26,788**
  - 3-star: **40,752**
  - 4-star: **79,149**
  - 5-star: **233,472**

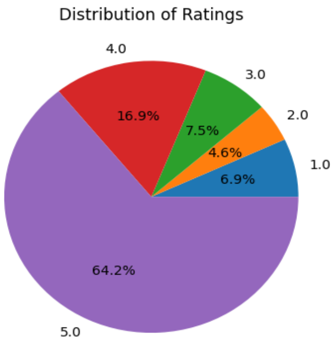


Key Insights

- Most Reviewed Brand: **Sony**
- Most Positively Reviewed Product: **Sony (ASIN: B000053ZF1)** with an average rating of **5.0**.

Visualization Highlights

- Pie chart of ratings distribution.



- Word Clouds for "Good" and "Bad" reviews.

- Yearly trends of reviews showing 2015 as the year with maximum reviews.

---

3. Text Preprocessing

- **Steps Applied**:
  - Removed HTML tags.
  - Handled accented characters.
  - Expanded acronyms.
  - Removed special characters.
  - Lemmatized text.
  - Normalized text.
- Saved the cleaned review text for further analysis.

---

4. Feature Engineering

- Created three vectorized representations for the review text:
  1. **Bag of Words (BoW)**.
  2. **TF-IDF**.
  3. **Hashing Vectorizer**.

---

5. Classification Models

Target Classes

- Good (Rating > 3).
- Average (Rating = 3).
- Bad (Rating < 3).

Models Evaluated

- Multinomial Naive Bayes
- Logistic Regression
- Linear SVC
- Random Forest

Performance Metrics

- Evaluated using Precision, Recall, F1-Score, and Support.

Best Model

- **Logistic Regression**:
  - F1-Score: **0.94** for "Good" class.
  - Accuracy: **88%**.

6. Collaborative Filtering

User-User Recommender

1. Created a User-Item matrix.
2. Normalized ratings using Min-Max Scaling.
3. Used Cosine Similarity to find the top N (10-50) similar users.
4. Used K-Folds Validation to compute MAE for predictions.

Item-Item Recommender

- Followed similar steps as User-User recommendations.

Results

- Plotted MAE against N (10, 20, 30, 40, 50) for both systems.
- **User-User MAE** was slightly better for smaller N, while **Item-Item** performed better for larger N.

Top Recommendations

- Generated the top 10 product recommendations based on predicted ratings.

```
Shape of User-Item sparse matrix: (735, 3238)
                   User Ratings  Predicted Ratings
Recommended Items
B000068O3L                  3.0           0.021694
B00013BKS2                  3.0          -0.001160
B00008Z1QI                  3.0          -0.001160
```

Conclusion

This project successfully built a comprehensive recommendation system for electronics, particularly headphones, leveraging both supervised learning models and collaborative filtering. The system demonstrated robust performance and provided actionable insights for personalization in e-commerce.

Let me know if you need refinements or specific sections expanded!