**Report Final Project**
Data Mining - CSC84040
Fall 2024

# Optimizing Product Recommendations Using a Hybrid Filtering System for Amazon Electronics Dataset


**Prince Osei**
CUNY Graduate Center

**Annapia Borraccino**
CUNY Graduate Center

**Ravi Prakash Tripathi**
CUNY Graduate Center


## 1. Introduction

Recommender systems play a crucial role in the digital era by providing personalized product suggestions to users. These systems drive sales and enhance user experience by tailoring recommendations to individual preferences. They are generally categorized into two main types. The first one is content-based filtering, which uses detailed information about items and user preferences to generate recommendations. However, this approach requires substantial metadata and struggles with sparse datasets. The following one is collaborative Filtering, which relies on user-item interactions, assuming users with similar preferences will like similar items. Its major limitation is the cold-start problem, where the system has insufficient data for new users or items. To overcome these limitations, hybrid recommender systems combine both methods, leveraging their strengths to produce robust recommendations.

## 2. Data & Goal of the project

This project focuses on building a hybrid recommendation system for e-commerce using the Amazon 2023 dataset, which includes 571.54 million product reviews and 33 item categories

with rich metadata. However, in the interest of time, this project only focuses on the Electronics section of the data.

Our project seeks to create an interactive product recommendation system utilizing a weighted hybrid approach. By combining content-based and collaborative filtering techniques, the system aims to address cold-start issues and enhance recommendation diversity. The Amazon 2023 dataset serves as the foundation for this system, offering extensive user reviews, metadata-rich descriptions of items, and a wide range of product categories. The proposal emphasizes optimizing the balance between the two algorithms, handling computational complexity, and ensuring scalability without compromising performance. This project seeks to create an interactive product recommendation system utilizing a hybrid approach. By combining content-based and collaborative filtering techniques, the system aims to address cold-start issues and enhance recommendation diversity. The proposal emphasizes optimizing the balance between the two algorithms, in the attempt of efficiently handling computational complexity and ensuring scalability.

## 3. Methods & Discussion

The first step in developing the recommendation system involved performing Exploratory Data Analysis (EDA) and data cleaning to prepare the dataset for analysis and modeling. We began by inspecting the data to identify missing or inconsistent entries. Key columns, such as prices, ratings, and the number of ratings, were cleaned to ensure accuracy. Non-numeric characters, such as the currency symbol "₹" and commas in prices, were removed, and the columns were converted into appropriate data types. Similarly, entries in the ratings and number of ratings columns that did not conform to numeric formats were filtered and corrected. Additional

preprocessing steps focused on enhancing the dataset's usability and richness. Missing values in essential columns such as 'actual_price' and 'discount_price' were dropped to maintain data integrity. We extracted the manufacturer names from product titles to create a new column, 'manufacturer,' which was positioned as the second column for better clarity. Furthermore, we calculated discounts and discount percentages for each product and added these as new columns to facilitate analysis of pricing trends. During EDA, we examined key statistics to understand the data distribution and identify potential biases. For example, we observed an average customer rating of 4.0 with a standard deviation of 0.32, indicating generally positive feedback. However, outliers, such as ratings of 0.0, prompted further investigation into potential sources of dissatisfaction. Visualizations were used to provide insights into popular brands and customer satisfaction levels. For instance, the top 15 most popular brands were plotted based on product counts, highlighting major contributors like TheGiftCard and boAt.

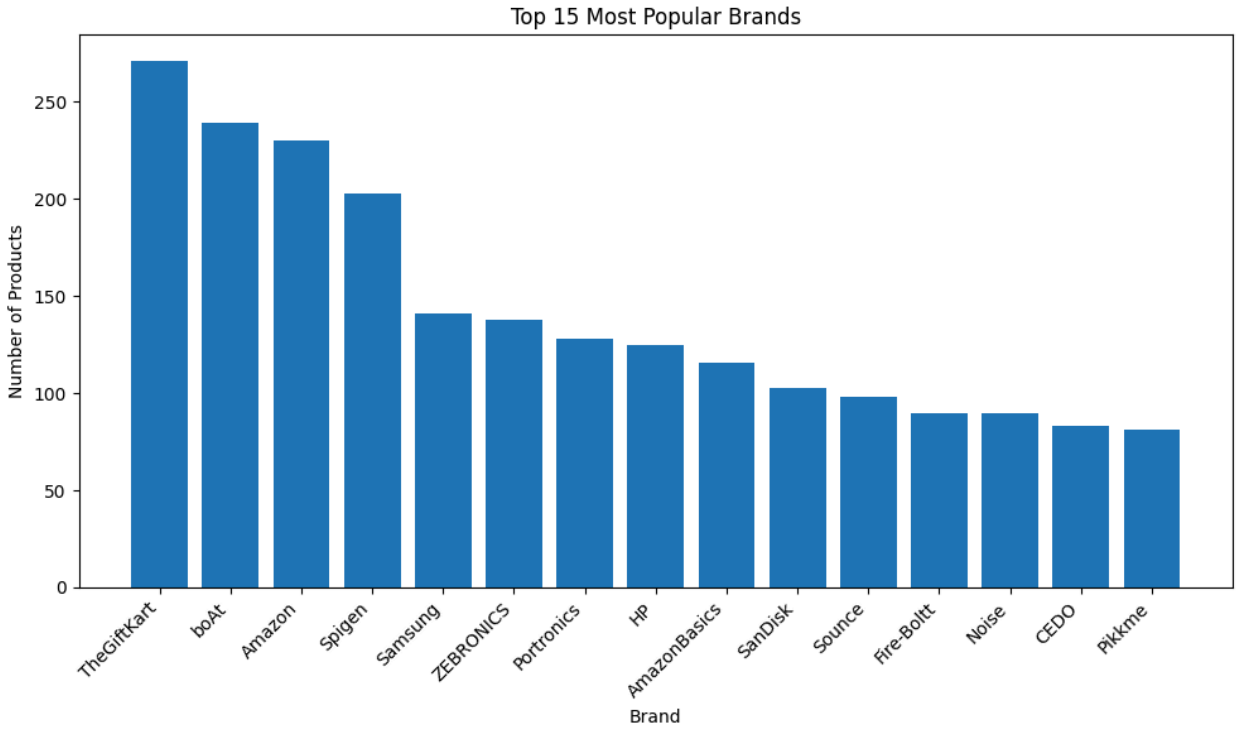


Figure 1. Most popular brand highlighted by EDA

To address the potential influence of fake or paid reviews, we focused on brands with significant purchase volumes. This filtering process allowed us to identify reliable trends by analyzing only those brands with more than 50 purchases. For these brands, we computed average ratings to ensure that the analysis reflected genuine customer satisfaction. The cleaned and processed dataset served as a robust foundation for building the recommendation system, ensuring high-quality inputs for subsequent evaluation.

The second stage of the project involved feature engineering and the implementation of advanced content-based and collaborative filtering models. Numerical features, such as ratings and discount percentages, were normalized using MinMaxScaler to ensure consistency in scaling. For content-based filtering, descriptive text features were combined from multiple columns, including product name, manufacturer, main category, and sub-category, creating a rich textual dataset for analysis. Two approaches were implemented for content-based filtering. The first approach utilized the TF-IDF vectorizer to extract numerical representations of textual data. The TF-IDF matrix, optimized with parameters such as n-gram range and document frequency thresholds, enabled semantic similarity calculations using cosine similarity. This resulted in a content similarity matrix, providing a foundational structure for generating recommendations.

The second approach employed a more sophisticated method using the BERT transformer model. Product descriptions were tokenized and embedded into high-dimensional vectors using the BERT tokenizer and model. These embeddings captured contextual relationships in the data, allowing for more nuanced similarity calculations. To optimize computation, the embeddings were generated in batches and processed using GPU acceleration when available. The resulting

similarity matrix was converted into a DataFrame for integration into the recommendation system.

For collaborative filtering, an item-item matrix was created based on product ratings and manufacturer attributes, addressing the absence of user identifiers. Dimensionality reduction was performed on the matrix using Truncated Singular Value Decomposition (SVD), reducing it to 50 latent components. This improved computational efficiency while preserving significant patterns in the data. Collaborative similarity was computed using cosine similarity on the reduced matrix, resulting in a collaborative similarity DataFrame.

The hybrid recommendation system combined scores from both content-based approaches and collaborative filtering. By assigning adjustable weights, such as 50% for TF-IDF or BERT and 50% for collaborative filtering, the system leveraged the strengths of each method. The system ranks and returns the top 5 suggestions for a selected product while evaluating their average similarity and diversity to ensure meaningful results.

```
Hybrid Recommendations for 'Apple 20W USB-C Power Adapter (for iPhone, iPad & AirPods)':
1. Apple 30W USB-C Power Adapter
2. Apple 5W USB Power Adapter (for iPhone)
3. Apple MagSafe Charger (for iPhone, AirPods Pro, AirPods with Wireless Charging Case)
4. Apple AirPods (3rd Generation) with Lightning Charging Case
5. Apple AirPods Pro (2nd Generation)

Average Similarity of Recommendations: 0.8707
Diversity of Recommendations: 0.2382
```

Figure 2. Evaluating the hybrid system using TF-IDF.

```
Hybrid Recommendations for 'Apple 20W USB-C Power Adapter (for iPhone, iPad & AirPods)':
1. Apple 5W USB Power Adapter (for iPhone)
2. Apple MagSafe Charger (for iPhone, AirPods Pro, AirPods with Wireless Charging Case)
3. Apple AirPods (3rd Generation) with Lightning Charging Case
4. Apple 30W USB-C Power Adapter
5. Apple iPhone 13 (128GB) - Starlight

Average Similarity of Recommendations: 0.9798
Diversity of Recommendations: 0.0288
```

Figure 3. Evaluating the hybrid system using BERT.

The figures indicate that the recommendations achieved a good similarity score. However, since the primary goal of the project was to create a system akin to those used in retail, greater emphasis should be placed on the diversity score. Diversity is crucial for encouraging customers to explore a wider range of products from different brands or items that complement their previous purchases. Unfortunately, the hybrid system fell short in this regard. While BERT provided substantial contextual understanding, it appeared to overly reduce diversity. In contrast, TF-IDF managed to retain a reasonable level of diversity, making it the preferred approach. Consequently, we decided to focus on TF-IDF and explore ways to enhance its performance further. In fact, this led to building a diversification function to balance similarity and diversity in the final recommendations. This enhances the recommendation system by balancing similarity and diversity to create a more engaging and varied suggestion list. It iteratively selects items from a pool of candidates ranked by their hybrid similarity scores and evaluates them based on two criteria: similarity to the user's preferences and dissimilarity from the items already selected. By prioritizing candidates with the highest combined scores, the function ensures that recommendations are relevant yet diverse, encouraging users to explore a broader range of products. This approach not only avoids redundancy but also aligns with retail goals by promoting cross-selling and exposing users to complementary items or different brands. Although this method may slightly reduce similarity scores, it significantly improves the variety of recommendations, fostering a richer user experience. Evaluation metrics, including average similarity and diversity, demonstrated the system's effectiveness in generating relevant and engaging recommendations.

```
Diverse Recommendations for 'Apple Lightning to USB Cable (2m):
1. Apple 30W USB-C Power Adapter
2. Apple Airpods (3rd Generation) with Lightning Charging Case
3. Apple Lightning to USB Camera Adapter, USB 3.0 OTG Cable for iPhone/iPad to Connect...
4. Apple iPhone 13 (256GB) - Midnight
5. Apple EarPods with 3.5mm Headphone Plug

Average Similarity of Recommendations: 0.7092
Diversity of Recommendations: 0.3064
```

Figure 4. Final evaluation of the hybrid system using TF-IDF and diversification.

## 4. Basic Interface

For demonstration purposes, we chose to build a very simple interactive interface.This serves as a preliminary design to provide users with an intuitive and engaging way to explore product recommendations. Users can begin by entering keywords to search for products, and the system provides a list of matching items from which they can select one. Once a product is chosen, users can specify their preferred content-based similarity method, either TF-IDF or BERT, to generate recommendations. The hybrid system then combines content-based and collaborative filtering to suggest products, presenting the results along with evaluations of their relevance and diversity. This approach not only highlights the system's flexibility in utilizing different content similarity techniques but also ensures an engaging experience by enabling users to actively participate in the recommendation process.

## 5. Limitations and Future Work

The project successfully demonstrated the potential of a hybrid recommendation system; however, several limitations were identified. One key challenge was achieving a higher diversity score in recommendations, as the system sometimes favored similar products, limiting the variety of suggestions offered to users. Additionally, the representation of manufacturers in the recommendations was uneven, with some brands being underrepresented. Another limitation was

the reliance on current features and preprocessing methods, which, while effective, may not fully capture the nuances of user preferences and product relationships. To address these limitations, future work should focus on improving the system's ability to generate diverse recommendations, ensuring a broader range of products and brands are represented. Refining feature extraction and preprocessing methods can further enhance the system's understanding of product relationships. Exploring alternative embedding models, such as RoBERTa or domain-specific transformers, offers the potential for more nuanced content-based similarity calculations, leading to more sophisticated and tailored recommendations. These advancements will help the system better align with the goals of retail applications, fostering greater user engagement and satisfaction.

## 6. Conclusion

In conclusion, this project successfully explored the development of a hybrid recommendation system that leverages the strengths of content-based and collaborative filtering techniques. By implementing advanced methods like TF-IDF and BERT for content similarity, along with dimensionality reduction for collaborative filtering, the system demonstrated its capability to provide personalized and relevant product recommendations. The integration of a diversification function further enhanced the system’s ability to balance similarity and variety, aligning with the goals of retail applications to engage users and promote product discovery. Despite some limitations, such as challenges in achieving higher diversity and ensuring balanced brand representation, the project laid a good foundation for future improvements. With further refinement of features, preprocessing methods, and the exploration of more advanced embedding models, this system has significant potential to evolve into a robust, scalable, and highly effective tool for personalized recommendations in e-commerce.