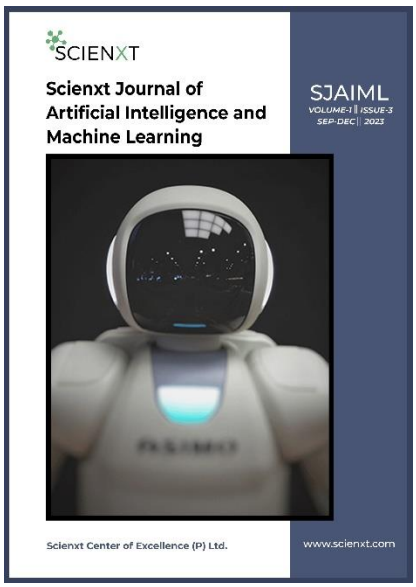


Scienxt Journal of Artificial Intelligence and Machine Learning
Volume-1 || Issue-3 || Sep-Dec || Year-2023 || pp. 1-12

Recommendation system using machine learning algorithm

Kruthika K. S
Department of AIML
Jyothy Institute of Technology
Bengaluru, India

Neethu C. V
Department of AIML
Jyothy Institute of Technology
Bengaluru, India

Ramya B. N
Department of AIML
Jyothy Institute of Technology
Bengaluru, India

**Corresponding Author: Kruthika K. S*
Email: Kruthika.k.sdatta@gmail.com

Abstract:

In this project, the nuanced exploration of an Amazon product dataset is facilitated through the adept use of Python and pandas. The comprehensive codebase spans multiple tasks, including meticulous data cleaning, insightful visualization techniques, rigorous statistical analysis, and the integration of a sophisticated content-based recommendation system. By delving into the intricate relationships within the dataset, the analysis uncovers noteworthy patterns related to product distribution, customer ratings, and inter-feature correlations. The incorporation of diverse visualization methods such as histograms and word clouds adds depth to the project, providing a rich, multifaceted perspective on user preferences and emerging product trends. Overall, this undertaking showcases the versatility of data-driven methodologies in extracting valuable insights from complex datasets in the context of Amazon's extensive product landscape.

Keywords:

Python, Pandas, Content-based recommendation system, Data analysis, Data cleaning, Visualization, Product distribution.

1. Introduction:

In the ever-expanding landscape of modern applications, recommender systems have become integral, offering users personalized suggestions from vast collections of items. These systems, ranging from movie platforms like Netflix to online retailers such as Amazon, play a pivotal role in aiding users in decision-making by predicting item preferences or presenting associated recommendations. The evolution of these systems has been marked by a decade of extensive research, primarily focused on designing innovative recommendation algorithms. For application designers seeking to incorporate recommender systems, a myriad of algorithmic choices exist, necessitating careful consideration based on performance experiments. Designers typically rely on comparative evaluations, assessing candidate algorithms against structural constraints like data type, timeliness, reliability, and resource footprints.

In this dynamic landscape, researchers continuously contribute to the field by proposing novel recommendation algorithms and evaluating their performance against existing approaches. These evaluations involve the application of metrics that quantify the effectiveness of algorithms, yielding numeric scores for comparative analysis. This iterative process not only enhances the diversity of available recommendation techniques but also ensures that the selected algorithm aligns with the specific requirements and constraints of the application, fostering continual innovation and refinement in the realm of recommender systems.

The project seamlessly transitions into Exploratory Data Analysis (EDA), harnessing the power of visualization tools like Matplotlib to unveil compelling insights into the Amazon product dataset. Engaging visualizations depict the distribution of products across main and subcategories, providing a bird's-eye view of market trends. Additionally, the EDA phase delves into customer rating distributions, shedding light on the nuanced sentiment of users towards various products. The project's latter segment integrates machine learning methodologies, employing TF-IDF text vectorization and cosine similarity, to construct a robust content-based recommendation system

Figure. 1: Digital view of recommendation system

2. Literature survey:

The article emphasizes the essential role of recommender systems in managing the vast information on the Internet. It reviews recent developments, addresses challenges, evaluates algorithms, and highlights the interdisciplinary nature of this research. The inclusion of physical aspects illustrates macroscopic behavior. The conclusion underscores the scientific depth of recommendation systems, appealing to a broad range of researchers.

The rapid evolution of the internet has revolutionized daily activities, transforming conventional approaches to tasks like online shopping and communication. Traditional mass production for a single market is no longer viable, necessitating a shift towards customization to meet diverse customer needs in the era of online shopping. However, this surge in product variety poses a challenge for consumers who must navigate a multitude of options. This paper delves into recommender systems as a solution, exploring common techniques and their associated trade-offs to assist customers in making informed choices amidst a plethora of offerings.

This article delves into the realm of e-commerce and Recommender Systems (RSs), highlighting their pivotal role in managing information overload in online platforms. Conducting a Systematic Literature Review (SLR), the paper addresses traditional RS methods, categorizing them into Content-Based Filtering (CBF), Collaborative Filtering (CF), Demographic-Based Filtering (DBF), hybrid filtering, and Knowledge-Based Filtering (KBF). The review spans papers published from 2008 to 2019, identifying gaps and significant issues in traditional RS techniques. The comprehensive analysis includes a comparative table of metrics, advantages, and disadvantages, offering valuable insights for future research in the field of e-commerce recommender systems.

The paper discusses limitations in collaborative filtering, particularly in item-based collaborative filtering relying on local resources and user rating matrices, leading to challenges like sparsity and the cold-start problem. The proposed approach introduces an additional database to support item-based collaborative filtering and aims to address data insufficiency issues. This extra database enhances the accuracy of item similarity calculations, providing improved prediction results, and facilitating successful recommendations even in scenarios involving sparse matrices or new items.

The paper addresses a gap in previous research by investigating how consumer preferences influence the accuracy of recommender systems in electronic markets. Introducing a microeconomic model, the study systematically analyzes various structures of consumer preferences and develops a metric for recommendation accuracy. Through simulations, the research evaluates the impact of consumer preference structures on a widely used collaborative filtering algorithm. Results indicate that recommendation accuracy is notably influenced by factors such as the similarity and number of consumer types, as well as the distribution of consumers in specific markets. Surprisingly, the study finds instances where random product recommendations surpass the performance of the collaborative filtering algorithm.

The paper emphasizes the growing significance of E-commerce in the global market and financial transactions, highlighting the need for robust and user-friendly systems through comprehensive analysis and design. The Internet's transformative impact on business operations is acknowledged, particularly in creating avenues for companies and customers to engage in buying and selling goods and services. The focus is on enhancing the efficiency of E-commerce websites, with a specific emphasis on recommender systems. The study provides an overview of the value of recommender systems, with a detailed analysis of Collaborative Recommender System (CF) techniques. CF techniques are explored for their ability to propose relevant products to customers based on their interests, streamlining the search and selection process for customers.

The article underscores the growing urgency for recommendation systems in response to the rapid evolution of e-commerce and the vast influx of internet information. Despite significant progress in research and application over the past decades, challenges like sparsity and cold-start issues persist. The proposed solution is a multi-mode e-commerce recommendation system that addresses these challenges by collecting extensive user information. It employs various recommendation algorithms that learn from each other, culminating in a comprehensive integration of information to enhance the accuracy and completeness of goods or service recommendations for users.

3. Methodology:

3.1. Data collection:

The dataset contains information on over 1000 products sold by Amazon, encompassing various details such as product names, categories, prices, ratings, and customer reviews. This

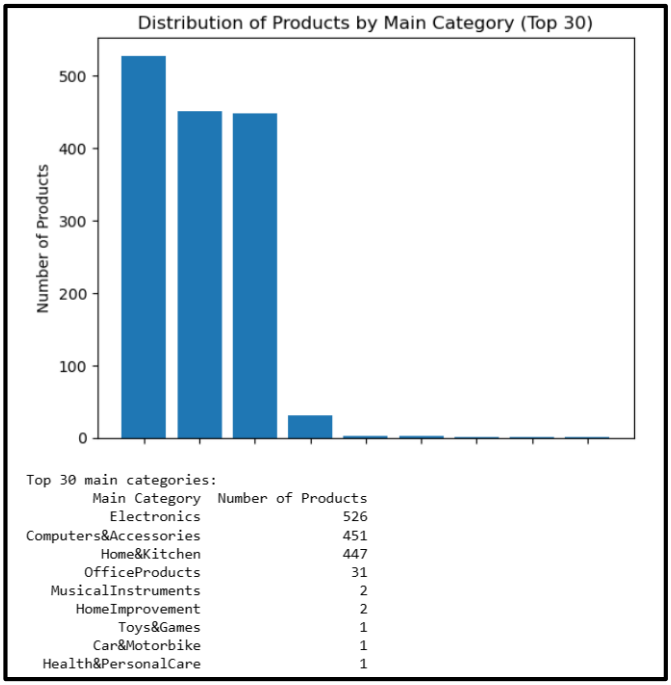
dataset was crucial for the analysis and development of a recommendation system. The dataset was likely sourced from Amazon's product listings and customer reviews, possibly through web scraping or API access. The information collected includes features like 'product_id,' 'product_name,' 'category,' 'discounted_price,' 'actual_price,' 'discount_percentage,' 'rating,' 'rating_count,' 'about_product,' 'user_id,' and more. The data collection process involved gathering diverse product information to facilitate comprehensive analysis and enhance the effectiveness of the recommendation system.

3.2. Data preparation:

The data preparation phase involved several essential steps. Firstly, a data inspection was conducted to identify any missing values, duplicates, or inconsistent data. This step ensures the dataset's readiness for analysis. Subsequently, data cleaning was performed to address errors, inconsistencies, and irrelevant information, enhancing the dataset's reliability. Following cleaning, data transformation was employed to make the dataset more suitable for analysis, including operations such as scaling, normalization, and feature engineering. Finally, the prepared data was saved in a new file to prevent overwriting the original dataset, allowing for future reference and reproducibility.

3.3. Exploratory data analysis (EDA):

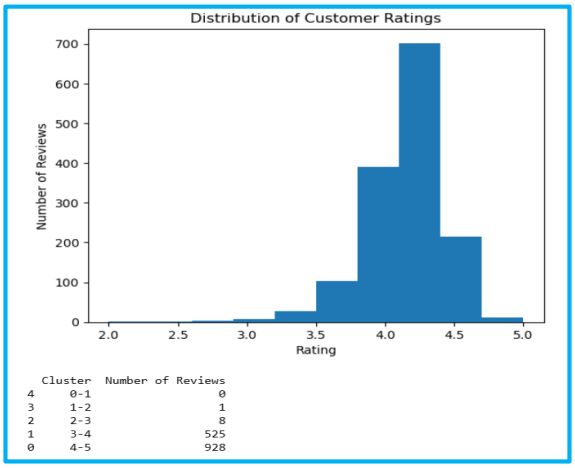
During EDA, the dataset was explored to gain insights into the distribution of products across categories, customer ratings, and reviews. This involved analyzing the distribution of products by both main and sub-categories using bar plots, visualizing the distribution of customer ratings through a histogram, and examining reviews through word clouds or frequency tables. The EDA process aimed to uncover patterns, trends, and anomalies in the data, providing a foundation for subsequent analysis and decision-making.



Top 30 main categories:

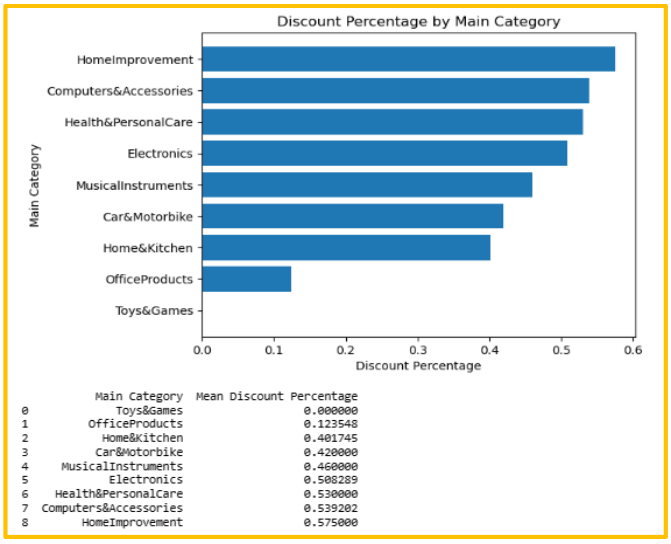
Main Category	Number of Products
Electronics	526
Computers&Accessories	451
Home&Kitchen	447
OfficeProducts	31
MusicalInstruments	2
HomeImprovement	2
Toys&Games	1
Car&Motorbike	1
Health&PersonalCare	1

Figure. 2: Bar plot of distribution of products by main category



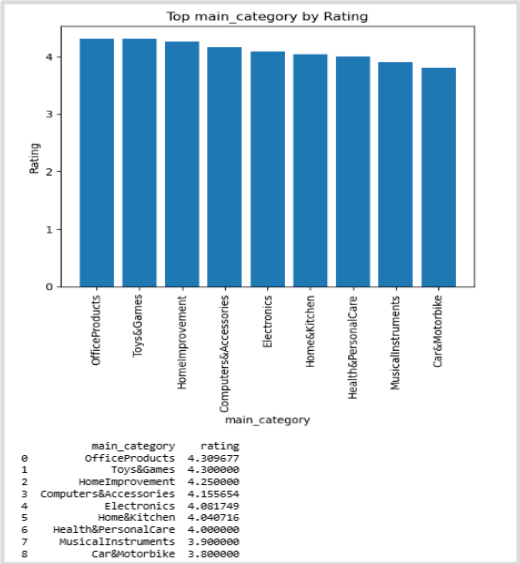
	Cluster	Number of Reviews
4	0-1	0
3	1-2	1
2	2-3	8
1	3-4	525
0	4-5	928

Figure. 3: Histogram of distribution of customer rating



	Main Category	Mean Discount Percentage
0	Toys&Games	0.000000
1	OfficeProducts	0.123548
2	Home&Kitchen	0.401745
3	Car&Motorbike	0.420000
4	MusicalInstruments	0.460000
5	Electronics	0.508289
6	Health&PersonalCare	0.530000
7	Computers&Accessories	0.539202
8	HomeImprovement	0.575000

Figure. 3: Horizontal bar plot of discount per of customer rating



	main_category	rating
0	OfficeProducts	4.309677
1	Toys&Games	4.300000
2	HomeImprovement	4.250000
3	Computers&Accessories	4.155654
4	Electronics	4.081749
5	Home&Kitchen	4.040716
6	Health&PersonalCare	4.000000
7	MusicalInstruments	3.900000
8	Car&Motorbike	3.800000

Figure. 4: Bar plot of top main category by rati

3.4. Visualization:

The insights derived from visualizations played a pivotal role in understanding the dataset. Bar plots were utilized to analyze the distribution of products by both main and sub-categories, offering insights into popular and less popular product categories. Histograms provided a visual representation of the distribution of customer ratings, shedding light on overall customer satisfaction. Additionally, word clouds were generated to highlight the most common words in reviews, offering qualitative insights into customer sentiments and preferences. Visualization techniques enhanced the interpretability of the dataset and facilitated more informed decision-making.

Figure. 4: World cloud based on the review text in the dataset

3.5. Recommendation system:

The recommendation system here is likely to employ collaborative filtering based on the cosine similarity measure.

In collaborative filtering, user-item interactions are utilized to make recommendations.

Cosine similarity is a metric used to measure the cosine of the angle between two non-zero vectors. In the context of recommendation systems, these vectors often represent user or item preferences. The formula for cosine similarity between two vectors,

Cosine Similarity (A, B) = A · B / ||A|| · ||B||.

Where:

- A · B represents the dot product of vectors A and B
- ||A|| and ||B|| denote the Euclidean norms (magnitude) of vectors A and B, respectively

The formula can also be expressed in terms of the vectors' components:

Cosine Similarity (A, B) = Σ (Ai * Bi) / √ (Σ (Ai) ^2 Σ (Bi) ^2)

Here:

- n is the number of dimensions (features) in the vectors
- Ai and Bi represent the components of vectors A and B in the i-th dimension

The resulting similarity score ranges from -1 (perfect dissimilarity) to 1 (perfect similarity). A score of 0 indicates orthogonality (no similarity). In the context of recommendation systems, higher cosine similarity scores between user vectors suggest greater similarity in preferences, leading to more accurate and personalized recommendations.

The system likely utilized the TF-IDF Vectorizer to transform product descriptions into numerical feature vectors, facilitating the calculation of cosine similarity. Recommendations were then made based on the similarity scores between users, with the top products suggested to users based on their historical preferences. This approach allows for personalized recommendations, enhancing user experience and engagement.

4. Result:

This project focused on analyzing data from Amazon's product dataset and implementing a recommendation system. With over 1000 products, the dataset underwent thorough cleansing, addressing issues like missing values and duplicates, and ensuring compatibility by converting string data to numerical formats. Exploratory Data Analysis provided insights into product distribution trends, customer ratings, and reviews, and the implemented recommendation system utilized collaborative filtering with cosine similarity, offering personalized product suggestions to enhance user engagement within the platform.

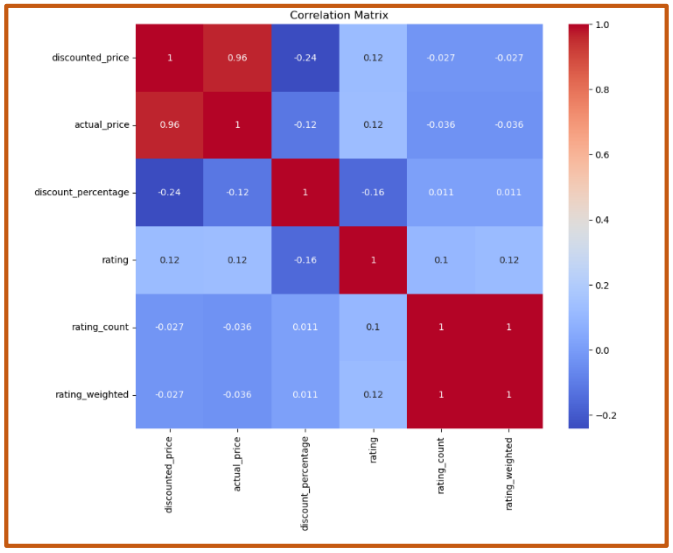


Figure. 5: Correlation matrix

	Id Encoded	recommended product	score recommendation
0	893	FYA Handheld Vacuum Cleaner Cordless, Wireless...	1.000000
1	893	Eureka Forbes Active Clean 700 Watts Powerful ...	0.347227
2	893	AMERICAN MICRONIC- Imported Wet & Dry Vacuum C...	0.317435
3	893	INALSA Upright Vacuum Cleaner, 2-in-1,Handheld...	0.310397
4	893	INALSA Vaccum Cleaner Handheld 800W High Power...	0.298866

Figure. 6: List of recommendations for a product with product_id 893 with score

5. Conclusion:

In the realm of online retail, the project embarked on a comprehensive exploration of Amazon's vast product landscape, employing data-driven methodologies to extract valuable insights and enhance user experiences. The meticulous data collection process involved harnessing a diverse dataset encompassing crucial product attributes, user interactions, and reviews. This served as the foundation for subsequent stages, ensuring the relevance and richness of the analysis.

Looking ahead, the project aims to advance its capabilities in online retail analysis. By building upon a foundation of diverse and relevant data, the focus shifts to refining the recommendation system. Plans include exploring advanced algorithms, integrating real-time monitoring, and incorporating technologies like natural language processing for a deeper understanding of customer sentiment. The project also envisions leveraging predictive analytics and trend forecasting to stay ahead of evolving market dynamics. These enhancements strive to ensure ongoing innovation, keeping Amazon at the forefront of the online retail landscape and delivering an even more enriched user experience.

Table. I: System configuration details

Hardware Requirements	• A computer with a multi-core CPU • High-end graphics card (GPU) • Sufficient RAM would be required to train deep learning models on large datasets
Software Requirements	• Windows 32/64- bit operating System
Platform	• Windows 32/64-bit Operating System
Programming Language/Tools	• Python • HTML • CSS • Flask

6. References:

(1) Lü, L., Medo, M., Yeung, C. H., Zhang, Y. C., Zhang, Z. K., & Zhou, T. (2012). Recommender systems. Physica A: Statistical Mechanics and its Applications, 519(1), 1-49.

(2) Shani, Guy & Gunawardana, Asela. (2011). Evaluating Recommendation Systems. 10.1007/978-0-387-85820-3_8. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P. (eds) Recommender Systems Handbook. Springer, Boston, MA.

(3) Sivapalan, Sanjeevan & Sadeghian, Alireza & Rahanam, Hossein & Madni, Asad. (2014). Recommender Systems in E-Commerce. 10.13140/2.1.3235.5847.

(4) P. M. Alamdari, N. J. Navimipour, M. Hosseinzadeh, A. A. Safaei, and A. Darwesh, "A Systematic Study on the Recommender Systems in the E-Commerce," in IEEE Access, vol. 8, pp. 115694-115716, 2020, doi: 10.1109/ACCESS.2020.3002803.

(5) Huang, T. C., Chen, Y. L., & Chen, M. C. (2016). A novel recommendation model with Google similarity. Decision Support Systems, DECSUP 1272. DOI: 10.1016/j.dss.2016.06.005

(6) Sebastian Köhler, Thomas Wöhner, and Ralf Peters. "The impact of consumer preferences on the accuracy of collaborative filtering recommender systems." *Journal of Business Economics* (2016) DOI: 10.1007/s12525-016-0232-3.

(7) Farah Tawfiq Abdul Hussien et al. "Recommendation Systems for E-commerce: An Overview." Journal of Physics: Conference Series 1897 (2021) 012024. doi:10.1088/1742-6596/1897/1/012024

(8) Xiaohong, C. (2012). Research of E-commerce Recommendation System Based on Multi-mode. In: Qu, X., Yang, Y. (eds) Information and Business Intelligence. IBI 2011. Communications in Computer and Information Science, vol 267. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-29084-8_9.