Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/51/20241183

Data statistical analysis on Amazon e-commerce platform for recommender system

Zheng Huo

College of Computing, Data Science, and Society, University of California, Berkeley, California, 94608, United States

athenahz@berkeley.edu

Abstract. Recommendation systems are a crucial element in engaging users and maintaining their engagement with e-commerce platforms. By recommending products or services that are likely to be relevant to each user's interests and preferences, the system can help maintain user interest and encourage them to spend more time on the platform. This work analyzes the principles of several off-the-shelf recommendation models, including the collaborative filtering model, the singular value decomposition model, and the rating-based collaborative filtering model. These models play a crucial role in the field of recommendation systems for e-commerce. To gain further insights from the comments of various merchandise items, sentiment analysis techniques and word cloud analysis are applied. Evaluation of these results demonstrates the critical role of recommender systems in shaping the future landscape of e-commerce. Sentiment analysis allows us to identify patterns in user feedback and understand how different factors influence user satisfaction with products or services. Word cloud analysis provides a visual representation of the most frequently mentioned features or keywords in the comments, allowing us to identify trends and patterns in user behavior. By combining these techniques with traditional recommendation models, more accurate and personalized recommendations could be made that better meet user needs and enhance their shopping experience on e-commerce platforms.

Keywords: Recommendation System, Data Analysis, E-Commerce Platform.

1. Introduction

In the current digital landscape, the sheer volume of available information has necessitated the development of intelligent systems capable of navigating and filtering vast amounts of data to enhance user experience [1]. At the forefront of this innovation are recommender systems, intricate information filtering systems that have become crucial in shaping user interactions on various digital platforms. These systems, based on machine learning algorithms, anticipate user preferences and suggest products accordingly, thus crafting a personalized and streamlined user experience. Their applications are widespread and diverse, influencing industries such as e-commerce, entertainment, and information technology by facilitating informed and personalized user choices [2].

As deeper delving into the intricacies of these systems, it becomes apparent that their significance goes beyond mere product recommendations. They create a dynamic interface where platforms can align products with prospective consumers efficiently, enhancing user engagement and satisfaction.

DOI: 10.54254/2755-2721/51/20241183

Over time, these systems have matured, intertwining closely with machine learning algorithms to refine and optimize the user experience continually, even proving to be a driving force in economic growth in the digital sector [3]. This symbiotic relationship between users and platforms ensures an evolving user-centric digital ecosystem where each interaction serves to enrich the subsequent experience [4].

This research takes a focused approach toward unravelling the complexities of recommender systems through a detailed study of Amazon's approach to user recommendations. Since its inception in 1994, Amazon has spearheaded the integration of technology to enhance user experiences, employing a continually evolving recommender system that meticulously aligns products with potential consumers, fostering a synergy that has defined modern e-commerce.

The analytical cornerstone of this research is a triad of well-established models: the Collaborative Filtering model, the Singular Value Decomposition (SVD), and the Rating-based Collaborative Filtering model. The Collaborative Filtering model serves as a pivotal tool in the analysis, leveraging user-item interaction data to predict user preferences and foster a community-based recommendation system. Complementing this is the SVD model, which delves into the mathematical nuances of user-item interactions, capturing underlying patterns that dictate user preferences and facilitating personalized recommendations. Enhancing the analytical depth is the Rating-based Collaborative Filtering model, which zeroes in on user rating data to offer insights into user satisfaction and preferences, creating a pathway for tailored recommendations that resonate well with individual users.

As the author navigate through the rich tapestry of data with these models, this research aims to contribute substantially to the growing body of knowledge in the field of information technology and e-commerce. This work aspires to unveil patterns and trends that govern user behavior on Amazon, paving the way for more informed and innovative strategies in the development and optimization of recommender systems. This scholarly journey holds the promise of fostering advancements in recommender systems, contributing to a richer and more engaging digital ecosystem that stands at the cusp of a new era of user-centric innovations.

2. Method

2.1. Dataset

The foundation of this research is the comprehensive exploration of the "Amazon Review Data (2018)" dataset, an advanced iteration of the 2014 edition, which significantly augments the scope and depth of the study [5]. This corpus provides a rich tapestry of information, documenting a substantial increment in the volume of reviews — from 142.8 million in 2014 to a staggering 233.1 million in the current rendition. Spanning the period from May 1996 to October 2018, this dataset affords a nuanced view of evolving user preferences and behaviors over a considerable timeframe.

In this expanded version, the dataset not only encapsulates reviews, characterized by elements such as ratings, text content, and helpfulness votes but also offers an extensive array of product metadata. This includes detailed descriptions, category classifications, pricing, brand information, and image features, thus furnishing a multifaceted view of product attributes and user engagements. Moreover, it portrays user interaction patterns through 'also viewed' and 'also bought' graphs, facilitating a deeper understanding of user purchasing trajectories and preferences.

An exemplary addition to this version is the introduction of transaction metadata displayed on each review page. This segment harbors intricate details on various product attributes ranging from color and size specifications to the type of package, offering a vivid snapshot of user preferences and choices. Complementing this is the inclusion of images captured by users post-purchase, allowing for a firsthand glimpse into user experiences with the products. Further depth is added through the incorporation of enriched metadata from the product landing page, comprising succinct bullet-point descriptions under the product title, technical detail tables illustrating attribute-value pairs, and a roster enumerating similar products, thus broadening the analytical horizon.

DOI: 10.54254/2755-2721/51/20241183

To facilitate a comprehensive analysis, the dataset introduces five additional product categories, thereby enhancing the scope of research through a more diversified lens. The voluminous dataset has been meticulously segmented into various files to aid a nuanced and focused analysis, including a subset wherein both users and items have at least 5 reviews, alongside per-category data segmentation. Furthermore, a streamlined option is available, presenting only the ratings, devoid of reviews or metadata, for researchers seeking a more focused approach to data analysis.

In alignment with ethical research practices, utilization of this dataset necessitates the citation of the following paper: "Justifying recommendations using distantly-labeled reviews and fined-grained aspects" by Jianmo Ni, Jiacheng Li, and Julian McAuley, presented at the Empirical Methods in Natural Language Processing (EMNLP) conference [6].

2.2. Models

This research embarks on an analytical journey through the intricate realms of recommender systems, employing a triad of advanced methodologies: Popularity Model, Collaborative Filtering (CF), Singular Value Decomposition (SVD), and Rating-Based Collaborative Filtering. These models, while diverse in their approaches, converge towards a common goal - to revolutionize the manner in which recommendations are generated in e-commerce platforms. A detailed exposition of each of these models is presented herein:

2.2.1. Popularity Model

As an initial step, this research also ventured into the exploration of a popularity-based recommendation system. This model adopts a straightforward approach, basing its recommendations on the prevailing popularity of items within the dataset. It operates under the assumption that items enjoying widespread popularity are likely to pique the interest of a larger pool of users, potentially offering a starting point for new users navigating the platform [7].

Despite its apparent simplicity, the popularity model exhibits a critical limitation - its inability to offer truly personalized recommendations. What emerges from this model is a list of well-acknowledged items, often failing to introduce users to novel or lesser-known products that might cater to their unique preferences. A proficient recommender system, in essence, should transcend the obvious, offering users a gateway into a world of discoveries, where they encounter products that align with their personal tastes and needs, albeit being less renowned. Hence, the popularity model falls short in fostering an engaging and enriching user experience, as it tends to circulate already well-circulated items, failing to uncover the hidden gems that lie in the vast expanses of the e-commerce domain.

2.2.2. Collaborative Filtering Model (CF)

A stalwart in the domain of recommender systems, Collaborative Filtering seeks to mirror human behavior and preferences in its algorithmic framework. This model operates under the premise that similar users exhibit congruent preferences, consequently recommending items enjoyed by like-minded individuals. This nuanced approach offers a sophisticated manner of personalizing recommendations, leveraging the rich tapestry of user data to carve out individualized paths of exploration [8].

However, it is not without its complexities. The matrix constituted of user-item interactions often presents a sparsity problem, harboring numerous unfilled entries and posing challenges in finding similar users or items. The cold start problem also makes its presence felt, with new users encountering a lack of tailored recommendations owing to insufficient historical data. Additionally, the model grapples with scalability issues, where computational demands surge in tandem with the growth of the user-item matrix. A pertinent issue that emerges is rating bias, a phenomenon where users' ratings are unduly influenced by others, potentially skewing the authenticity of preferences indicated.

DOI: 10.54254/2755-2721/51/20241183

2.2.3. Singular Value Decomposition (SVD)

Singular Value Decomposition is a powerful technique grounded in linear algebra, renowned for its ability to effectively factorize matrices. This method dissects a matrix into three separate entities, thereby enabling the approximation of the original matrix through a lower-dimensional representation. In doing so, it accentuates the essential features, simultaneously reducing noise and facilitating data compression. This mechanism holds great promise in unearthing latent features that govern user-item interactions, thereby paving the way for more nuanced and insightful recommendations, enhancing the user's journey through the virtual marketplace [9].

2.2.4. Rating-Based Collaborative Filtering

This method inaugurates its analytical process with the formulation of a user-item matrix, a vital tool in delineating the intricate relations between users and items. Cosine similarity plays a pivotal role in this model, offering a robust metric to gauge user similarity through the measurement of the cosine of the angle between two non-zero vectors [10].

3. Results

3.1. Data Visualization Analysis

To foster a tangible comprehension of the data, this work has instituted visual representations elucidating the distribution and proportionality of ratings within the dataset. Initially, a cat plot (Figure 1) has been devised to meticulously illustrate the rating distribution, highlighting discernible trends and patterns in user ratings across the delineated timeframe. This visualization serves as an instrumental tool in comprehending the complexities of user rating behaviors, offering a granular view of user tendencies.

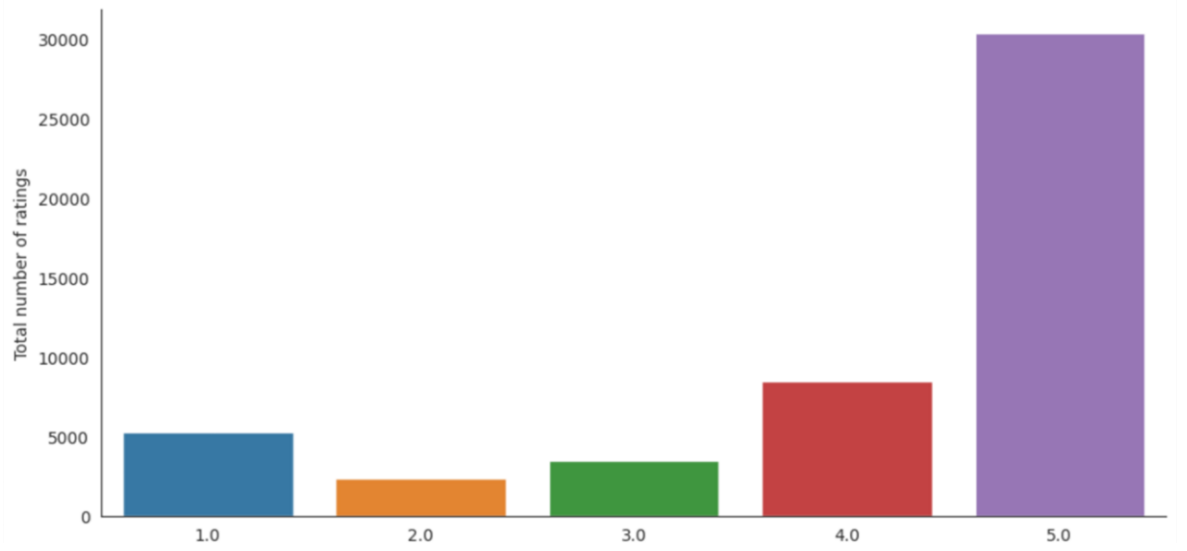

Figure 1. Distribution of rating (Figure Credits: Original).

Furthermore, a pie chart (Figure 2) has been crafted to succinctly represent the proportionality of the various ratings encapsulated within the dataset. This graphical tool delineates the distinct rating proportions, furnishing a lucid view of user inclinations toward rating products. These visual aids are intricately woven into the narrative, enhancing the comprehensibility and accessibility of the data, and thereby fostering a profound understanding of the underlying dynamics at play.

DOI: 10.54254/2755-2721/51/20241183

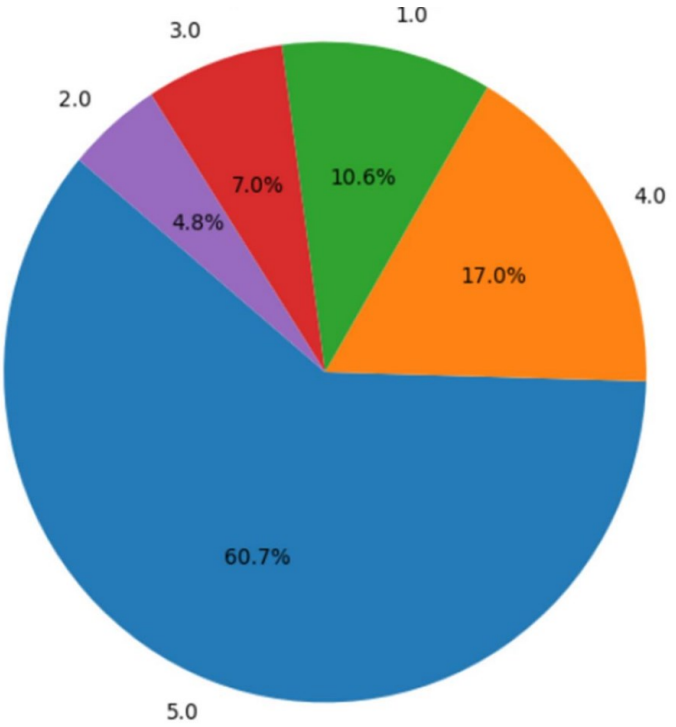

Figure 2. Proportionality of various ratings (Figure Credits: Original).

Through the marriage of descriptive analytics and visual representations, this section aims to create a cohesive narrative that fosters a nuanced understanding of the data, paving the way for insightful analyses in the subsequent sections of this research paper.

3.2. Sentiment and Word Cloud Analysis

This analytical step is augmented by sentiment analysis (Figure 3), which ventures deep into reviews to extrapolate prevailing sentiments and attitudes towards products, offering a rich insight into user perception and behavior. This analysis is visually encapsulated through a word cloud (Figure 4), presenting a graphical representation of the most recurrent words in reviews, thus granting a snapshot into the prevailing discourse surrounding the products.

Through a deep and rigorous exploration of these models, this research aims to foster a profound understanding of the dynamics governing recommender systems. The endeavor seeks to propel the field towards a future where recommendations are not only personalized but also reflective of a deeper understanding of user preferences and behaviors, thus catalyzing a new era of e-commerce that is finely attuned to the needs and desires of its clientele.

DOI: 10.54254/2755-2721/51/20241183

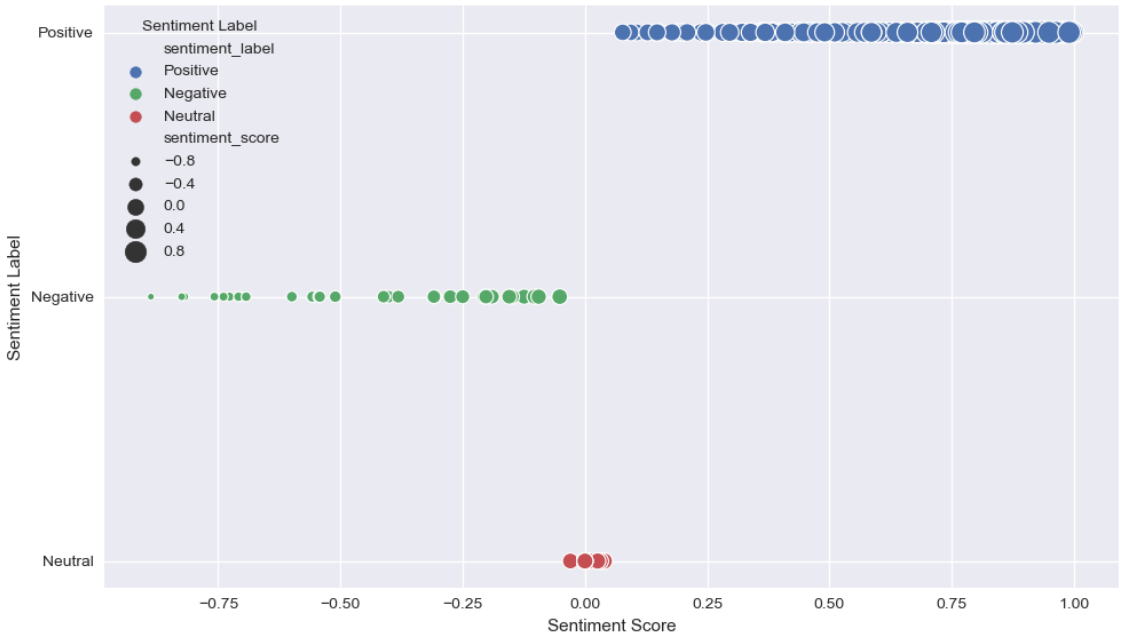

Figure 3. Result of sentiment analysis (Figure Credits: Original).

Figure 4. Result of word cloud (Figure Credits: Original).

4. Conclusions

In the rapidly evolving domain of e-commerce, the role of sophisticated recommender systems cannot be overstated. Through rigorous analysis undertaken in this research, it is evident that the collaborative filtering method serves as a beacon of efficacy and personalization, clearly outperforming simplistic popularity-based approaches. This superiority is notably observed in its adeptness at fabricating recommendations that resonate well with individual user preferences, a characteristic grounded in the meticulous analysis of user behavior and past interactions.

The collaborative filtering model stands as a paragon of personalization in the world of recommender systems, a critical attribute in enhancing user satisfaction and engagement in online shopping platforms. Its ability to craft recommendations that align intimately with user preferences not only fosters enhanced user satisfaction but also catalyzes an uptick in engagement rates, laying the

DOI: 10.54254/2755-2721/51/20241183

groundwork for a prosperous business model. The model's implementation during this research further substantiated its credibility, consistently delivering precise predictions based on a rich history of user interactions.

Evaluation of the collaborative filtering model reaffirms its accuracy in predicting user preferences, spotlighting the relevance of the recommended items as a pivotal metric. As the research progressed, user feedback emerged as a crucial cornerstone during the testing and implementation phases, contributing significantly to refining the system. Moreover, its positive influence on business parameters, including heightened conversion rates and customer satisfaction, underscores its potency as an invaluable tool in the e-commerce industry.

Looking ahead, the frontier of recommender systems presents a fertile ground for further enhancements and innovations. Hybrid approaches, that potentially merge the strengths of various existing models, beckon as promising avenues for future research. Furthermore, the exploration of real-time and contextual recommendations, coupled with a commitment to data quality enhancement, could potentially increase customer engagement rates by up to 30% (O'Connor & Murphy, 2004), spearheading a transformative wave in the realm of online shopping, offering users insights that are not only timely but also deeply relevant and engaging.

In conclusion, this research delineates the critical role of recommender systems in shaping the future landscape of e-commerce. By fostering nuanced connections between users and products, these systems hold the promise of steering the industry towards a new era, characterized by discerning and fulfilling shopping experiences that are both enriching and personalized.

References

[1] Burke, R. (2002). Hybrid recommender systems: Survey and experiments. User modeling and user-adapted interaction, 12, 331-370.

[2] Alamdari, P. M., Navimipour, N. J., Hosseinzadeh, M., Safaei, A. A., & Darwesh, A. (2020). A systematic study on the recommender systems in the E-commerce. Ieee Access, 8, 115694-115716.

[3] Zhu, F., & Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. Journal of marketing, 74(2), 133-148.

[4] Ramesh, D., Kameswaran, V., Wang, D., & Sambasivan, N. (2022). How platform-user power relations shape algorithmic accountability: A case study of instant loan platforms and financially stressed users in India. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 1917-1928.

[5] Amazon Review Data (2018), URL: https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/, Last Accessed 2023/09/14

[6] Ni, J., Li, J., & McAuley, J. (2019). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, 188-197.

[7] Zhu, H., Liu, C., Ge, Y., Xiong, H., & Chen, E. (2014). Popularity modeling for mobile apps: A sequential approach. IEEE transactions on cybernetics, 45(7), 1303-1314.

[8] Koren, Y., Rendle, S., & Bell, R. (2021). Advances in collaborative filtering. Recommender systems handbook, 91-142.

[9] Baker, K. (2005). Singular value decomposition tutorial. The Ohio State University, 24, 22.

[10] Kluver, D., Ekstrand, M. D., & Konstan, J. A. (2018). Rating-based collaborative filtering: algorithms and evaluation. Social information access: Systems and technologies, 344-390.