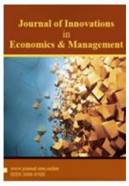
Journal of
Innovations in Economics & Management
journal homepage: www.journal-iem.online



Research article

Personalized Recommendation Research Based on Logistic Regression Algorithm for Amazon Product Reviews

Jinfeng Wang[1] Yuzhi Qi[2]*

1.School of Foreign Languages and Cultures, Chongqing University, Chongqing 401331, China

2.School of Public Administration, Sichuan University, Chengdu 610064, China

*Corresponding author: Jinfeng Wang (Email: 20200849@stu.cqu.edu.cn)

ABSTRACT

This study explores personalized recommendation strategies within Amazon's product review system using the Logistic Regression algorithm. By analyzing user behavior and review data, a predictive model was developed to forecast user preferences for specific products. The research employed extensive real-world data and validated the model's effectiveness and accuracy through empirical analysis. Findings indicate that the proposed personalized recommendation system significantly enhances user experience and increases product sales. The study contributes an effective recommendation algorithm for e-commerce platforms, offering practical implications for enhancing user engagement and optimizing marketing strategies in competitive markets.

Keywords: Logistic Regression algorithm, personalized recommendation, Amazon product reviews, user behavior analysis

Introduction

The evolution of the Internet and the profound integration of e-commerce platforms have been ongoing for a considerable period of time. The substantial quantity of behavioural data generated by online shoppers has emerged as a significant source of data for user preference recommendation systems. E-commerce is defined as the buying and selling of goods and services through the Internet, which encompasses online transactions, electronic payments, and digital interactions between businesses and customers.1 A personalised recommendation system can provide users with tailored product recommendations by analysing their historical behaviour, interests, hobbies and social relationships in real-time. This significantly improves the user experience and enhances the commercial value and revenue of the platform. Among the numerous recommendation algorithms, the Logistic Regression algorithm has garnered considerable interest due to its simplicity, clarity, and efficacy in binary classification problems. The objective of this experiment is to utilise the Logistic Regression algorithm to construct a personalised recommendation system based on the acquisition and analysis of Amazon product review data. By analysing users' ratings and recommendation behaviour, it is possible to predict whether users will recommend a product. It is possible for customers to interact with products on the system, leaving comments or opinions about the products. These interactions can be of great assistance in reaching target customers.2 Such interactions not only enrich the behavioural data of online shopping, but also provide more

Backgrounds

In the contemporary era, the majority of activities have been digitised. 3 The global pandemic of the novel coronavirus (COVID-19) has had a profound impact on individuals and businesses worldwide. The pandemic has prompted consumers to

---

[1] Zhang, X., Guo, F., Chen, T., et al. (2023). A brief survey of machine learning and deep learning techniques for e-commerce research. *J. Theor. Appl. Electron. Commer. Res., 18*. https://doi.org/10.3390/jtaer18040110

[2] Tran, D., & Huh, J. (2022). New machine learning model based on the time factor for e-commerce recommendation systems. *The Journal of Supercomputing, 79*. https://doi.org/10.1007/s11227-022-04909-2

[3] Enhancing E-Commerce Applications with Machine Learning Recommendation Systems

embrace online shopping and altered their purchasing behaviour, making them more emotional and faster in their decision-making process, particularly when it comes to selecting products and online stores. 4 The pervasive integration of Internet technology and e-commerce platforms has become a prevalent trend in global business activities. The advent of e-commerce has profoundly altered the manner in which consumers interact with businesses and market dynamics through online transactions, electronic payments and digital interactions. Nevertheless, the accelerated growth of e-commerce platforms has resulted in a deluge of product information, which has led to a phenomenon known as 'choice paralysis' and 'information overload' among users.5 In order to address this challenge, personalised recommendation systems have emerged.

As users generate a substantial quantity of behavioural data when shopping online, this data not only reflects consumers' shopping preferences and habits, but also becomes an essential data source for personalised recommendation systems. The use of advanced algorithms and dynamic analysis techniques, such as the Logistic Regression algorithm, enables personalised recommendation systems to delve deeply into users' historical behaviour, interests, hobbies and social relationships. This enables the provision of tailored recommendations for each user. Such customised recommendations not only significantly enhance the user experience and satisfaction, but also significantly enhance the commercial competitiveness and profitability of e-commerce platforms.

Personalised recommendation systems, such as those employed by Amazon, Netflix and numerous online businesses, have successfully addressed the issue of information overload and choice paralysis by analysing users’ behavioural data and providing precise content recommendations and shopping suggestions.6 These companies have not only achieved remarkable results in practical applications, but have also promoted personalised recommendation systems through articles, competitions and other means, thereby promoting technological progress and innovation in the industry as a whole.

---

[4] Development of Recommendation System in e-Commerce using Emotional Analysis and Machine Learning Methods

[5] Design and implementation of an intelligent recommendation system for product information on an e-commerce platform based on machine learning

[6] Intelligent Classification and Personalized Recommendation of E-commerce Products Based on Machine Learning

When evaluating the relative merits and demerits of each recommendation algorithm, a meticulous analysis is frequently conducted employing a multitude of indicators.7 In accordance with the specific requirements of each application, recommendation systems employ a range of evaluation indicators, including accuracy, recall, F1 score, coverage and user satisfaction, in order to achieve the desired recommendation effect. These indicators facilitate the assessment of the performance of the recommendation system, thereby ensuring that it can achieve the desired effect in practical applications.

Recommendation systems typically present a list of recommendations in one of two ways: through collaborative filtering or content-based filtering. Collaborative filtering is predicated on the behavioural data of users and is therefore well-suited to scenarios where users exhibit similar behaviours or preferences. In contrast, content-based filtering is based on the analysis of historical data and known preferences of users, and is therefore suitable for users with established profiles.8 Both methods have their respective advantages and disadvantages. Collaborative filtering has the potential to uncover hitherto unidentified interests, whereas content-based filtering can provide more accurate recommendations when the user profile is complete.

In particular, in large e-commerce platforms such as Amazon, the analysis of product review data and the application of logistic regression algorithms have become important topics in research and practice. By analysing consumer ratings and recommendation behaviour, researchers can predict whether users will recommend a product, thereby optimising the algorithms and strategies of recommendation systems and improving the accuracy of recommendations and user satisfaction.

The objective of this study is to investigate the potential of utilising Amazon product review data to develop a personalised recommendation system based on the logistic regression algorithm. By applying a range of evaluation indicators to analyse the recommendation system in detail, it can provide more accurate and effective recommendation services for e-commerce platforms, alleviate the problems of user choice paralysis and information overload, and further promote the development and

---

[7] Intelligent Recommendation Method for Product Information of E-commerce Platform Based on Machine Learning Algorithm

[8] PRODUCT RECOMMENDATION SYSTEM USING MACHINE LEARNING THROUGH BIG DATA IN E-COMMERCE WEBSITE

innovation of the e-commerce industry.

The e-commerce industry is currently undergoing a period of rapid development, and the nature of the products purchased by online shoppers is becoming increasingly sophisticated. Moreover, there are indications that their purchasing behaviour is undergoing subtle changes.9 In the digital era, the interaction model between enterprises and consumers has undergone a fundamental transformation, and personalised recommendation systems have become a pivotal link between the two. By continuously optimising and improving recommendation algorithms, it can be anticipated that a more efficient and intelligent consumer experience will be achieved on digital platforms in the future.

Literature review

In the field of e-commerce, recommendation systems have become an important tool for improving the user experience and promoting sales. The continuous development of machine learning technology has led to a notable improvement in the accuracy and efficiency of recommendation systems. This paper reviews the application of machine learning in e-commerce recommendation systems in recent years, with a particular focus on personalised product review recommendation systems based on logistic regression algorithms. Furthermore, the analysis of Amazon product review data is employed to evaluate the relative merits and drawbacks of distinct research methodologies. The application of machine learning in e-commerce recommendation systems

Tran and Huh (2022) proposed a new machine learning model, ML.Recommend, which is based on time factors and combines the Microsoft ML.NET platform. In their study, Farooqi et al. (2022) examined a range of machine learning algorithms and evaluated their suitability for use in e-commerce recommendation systems. They emphasised the potential of machine learning technology in this context.

Chen (2023) put forth the DARR intelligent recommendation model as a solution to the challenges of product information recommendation on e-commerce platforms. The efficacy of this model was corroborated through comparisons with public recommendation datasets. Peng et al. (2023) presented a design and implementation

[9] e-Commerce Personalized Recommendation Based on Machine Learning Technology

scheme for an intelligent recommendation system with the objective of enhancing the accuracy of product recommendations and user experience, and optimising the algorithm. Zhang et al. (2023) conducted a survey on the application of machine learning and deep learning technologies in e-commerce from 2018 to 2023. The survey identified the most recent methods, principal topics, and potential challenges in this field.

Liu (2022) tested eight hypotheses through a model, and the results demonstrated that customer income level, online shopping experience, product price, product quality, recommendation relevance, credit rating, and service quality significantly and positively influence shopping willingness, which in turn affects customer shopping behaviour. Dey et al. (2022) proposed a method for designing and implementing recommendation systems using machine learning on e-commerce websites, with the aim of creating an information environment for users to assist customers in easily locating the products they require.

Xu et al. (2024) proposed a personalised recommendation system using BERT models and a nearest neighbour algorithm, which was optimised for the operability and scalability of the eBay e-commerce platform. Rahman et al. (2024) propose an intelligent recommendation system based on element-level collaborative filtering to construct a comprehensive and functional product recommendation system. Mykhalchuk et al. (2021) developed a recommendation system using sentiment analysis and machine learning methods to address the issue of enhancing e-commerce functions and provide a methodological reference for designing and developing recommendation systems.

These studies illustrate the diverse applications of machine learning in e-commerce recommendation systems, encompassing fundamental algorithmic research, technical implementation on specific platforms, and in-depth analysis of user experience and behaviour. Each paper contributes valuable knowledge and practical experience from different perspectives, thereby providing a solid foundation for future research.

Application of personalised recommendation systems in product review analysis

In their paper on personalised recommendation systems, Chen and He discuss the application of Siamese networks to unsupervised visual representation learning. The researchers discovered that simple Siamese networks are capable of learning meaningful representations even in the absence of negative pairs, large batches, or momentum encoders. This finding is of significant importance for the comprehension of the optimal utilisation of user data in recommendation systems.

In the context of product review analysis, Purcell and Neubauer discuss the potential of digital twin technology in agriculture. They propose the use of high-fidelity modelling and bidirectional data flow to improve the accuracy and efficiency of data analysis. Although this paper is focused on the field of agriculture, the methods proposed are instructive for understanding how to improve the accuracy and efficiency of product review analysis through technological means.

Page et al. introduce PRISMA 2020, an updated reporting guideline for systematic reviews and meta-analyses. This guideline is of paramount importance for ensuring the quality and transparency of literature reviews, particularly when dealing with voluminous and intricate product review data.

In order to assess the performance of recommender systems, Gattis et al. examined the relationship between parents' beliefs about infant care and their actual behaviour. Although this study was focused on the parenting domain, its methodology: comparing self-reported beliefs with observed behaviour – can be applied to assess the consistency between user feedback and actual usage behaviour of recommender systems.

Finally, McKibbin and Fernando examine the global macroeconomic impact of the COVID-19 pandemic. Although this study does not directly address recommender systems, it provides insight into the evaluation and adjustment of technical systems in the context of global events, which is crucial for optimising recommender systems in an uncertain economic environment.

Descriptive analysis

This report aims to explore the relationship between user rating, recommendation behavior, review length, brand rating distribution, useful number of reviews and other features through statistical analysis of Amazon product review data, and make predictions based on Logistic regression model. Through visual analysis, we can intuitively understand the data characteristics and provide data support for the construction of personalized recommendation system.

Data preprocessing

Data preprocessing is the basis of effective analysis. We read the data file from the specified path and clean and process the data. The dataset contains fields such as user ID, product ID, rating, recommendation status, and review text. In order to ensure the integrity and accuracy of the data, we select key fields such as id, asins, reviews.rating, reviews.doRecommend and reviews.text, and process the missing values. In addition, we calculated the length (number of characters) of each comment for subsequent analysis.

Statistical analysis

Figure 1-6 show the statistical analysis of the data respectively.

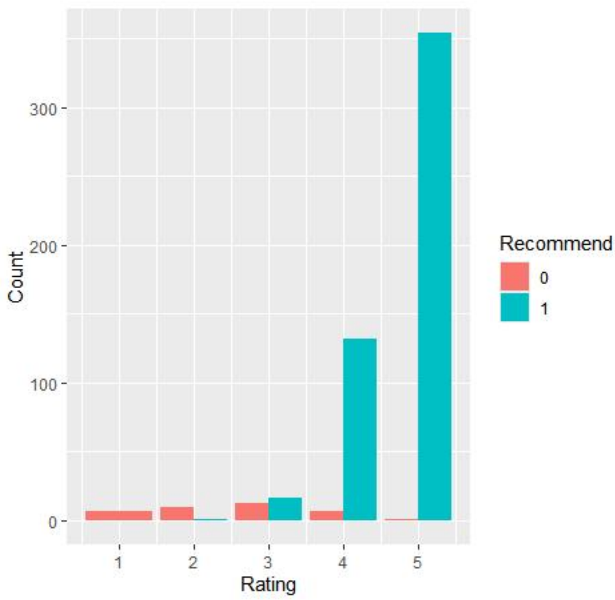


Figure 1 Relationship between Rating and Recommendation

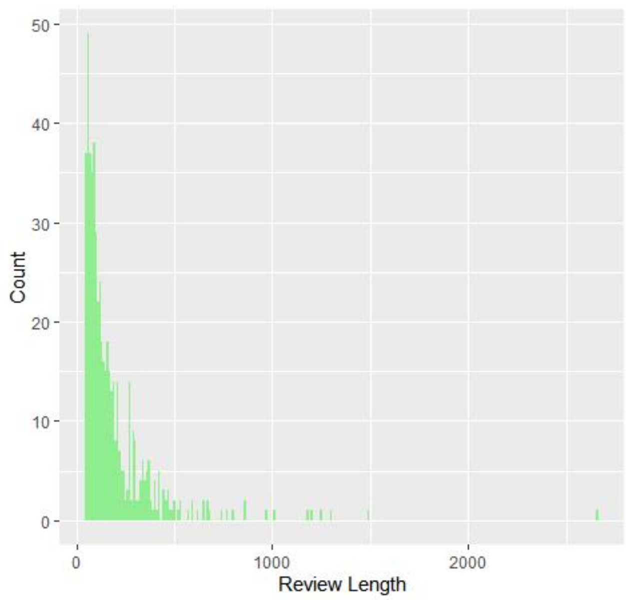


Figure 2 Distribution of Review Lengths

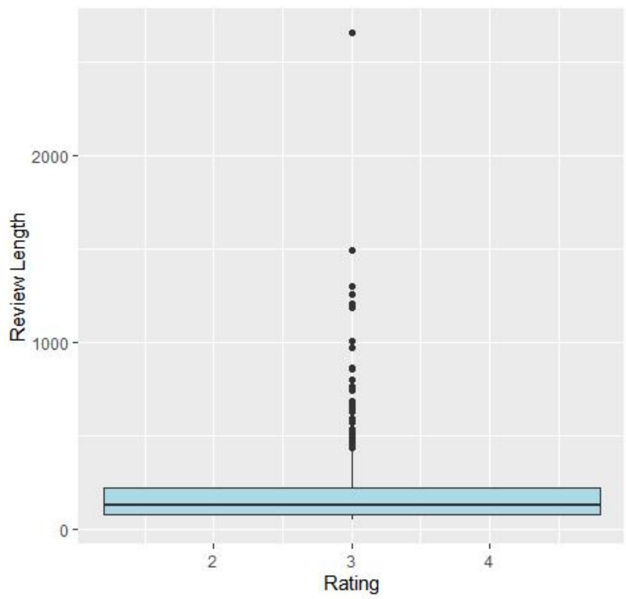


Figure 3Relationship between Rating and Review Length

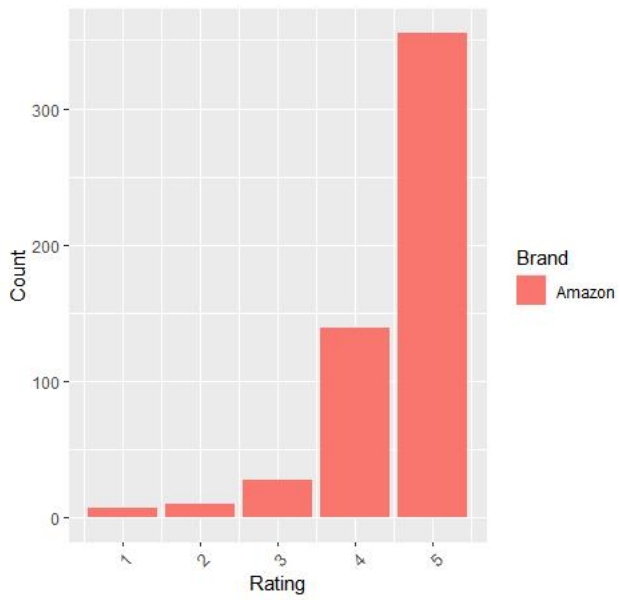


Figure 4 Rating Distribution by Brand

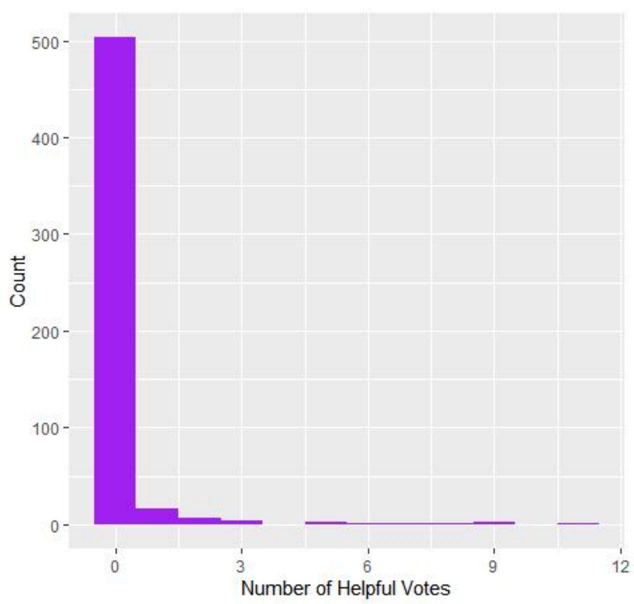


Figure 5 Distribution of Number of Helpful Votes

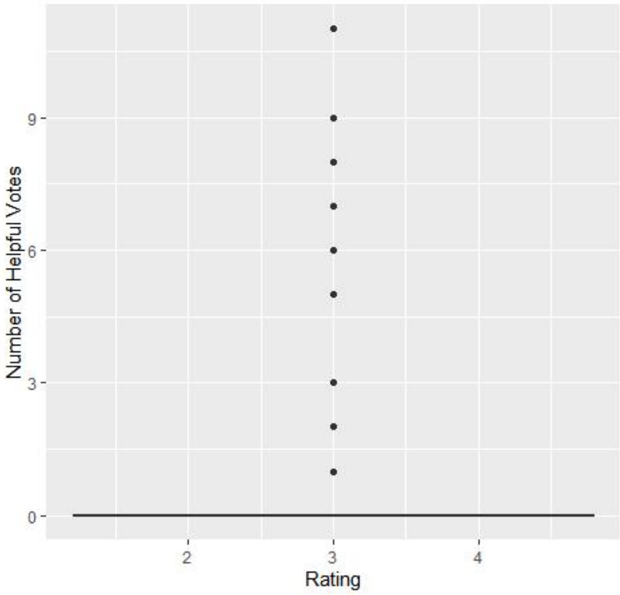


Figure 6 Relationship between Rating and Number of Helpful Votes

1. Comment length distribution

As can be seen from the comment length distribution map, the length of most comments is concentrated in the short range, especially the length of comments less than 100 characters. This shows that users tend to be short and to the point when writing reviews. However, we also noticed that a small number of reviews were more than 1,000 characters long and even reached more than 2,000 characters, which may be because these users have more detailed reviews or opinions about the product.

2. Relationship between rating and recommendation

The relationship between rating and recommendation shows the distribution of whether users recommend or not under different ratings. It can be clearly seen that the percentage of users recommending products increases significantly when the ratings are 4 and 5, while fewer users recommend products with low ratings (1, 2, and 3). This result is in line with the expectation that the more satisfied users are with a product, the more inclined they are to recommend it.

3. The relationship between rating and comment length

The relationship between rating and comment length shows the distribution of comment length under different ratings. While most comments are short in length, there is a wider range of comment length at a score of 3, with many outliers. This may indicate that users may be more detailed about their reasons and opinions when giving a moderate rating, while an extreme rating (1 or 5) may be more direct and brief.

4. Rating distribution of each brand

The brand rating distribution map shows the distribution of user ratings under different brands. In this dataset, the ratings of the "Amazon" brand are concentrated in the 4 and 5 points, which indicates that users have a high overall evaluation of the brand's products. The low number of low ratings (1, 2 and 3) further confirms the quality and customer satisfaction of the Amazon brand products.

5. Comment availability distribution

The distribution of comments shows that the number of times most comments are considered useful is zero. This may be because users don't always flag comments as useful when they read them. A small number of reviews received more useful flags, indicating that they may provide more detailed and valuable information.

6. The relationship between ratings and reviews

The relationship between ratings and useful number of reviews shows the distribution of useful number of reviews under different ratings. It can be seen that although the number of reviews useful with a rating of 3 is higher in some cases, overall there is no significant linear relationship between the rating and the number of reviews useful. This may be because users do not just rely on ratings when evaluating whether a review is useful, but also consider the level of detail and specific opinions of the content of the review.

Through the above analysis, we can draw the following conclusions:

1. Comment length distribution: Most comments are short, with the most comments under 100 characters. However, there are a small number of very detailed long reviews.

2. Relationship between rating and recommendation: The proportion of users recommending products increases significantly under high scores (4 and 5), while fewer users recommend products under low scores (1, 2 and 3).

3. Relationship between rating and comment length: Comments with a rating of 3 have a wide range of variation, possibly because users are more likely to describe their opinions in detail when they are moderately rated.

4. Rating distribution of each brand: The rating of "Amazon" brand is concentrated in 4 and 5 points, and users have a high overall evaluation of its products.

5. Distribution of comments useful number: most comments have a useful number of zero, and a few comments get more useful markers, indicating that these comments may provide more valuable information.

6. The relationship between ratings and the useful number of reviews: There is no significant linear relationship between ratings and the useful number of reviews, and users consider more factors when evaluating the usefulness of reviews.

In summary, through the analysis of Amazon product review data, we not only understand the user's rating and recommendation behavior, but also find the impact of characteristics such as review length and useful number on user reviews. In the future work, it can be further combined with other data sources, such as user purchase history and browsing history, to build a more comprehensive user portrait and improve the accuracy and personalization of the recommendation system.

Method

With the development of Internet technology and the popularity of e-commerce platforms, a large amount of behavioral data left by users when shopping online has become an important data source for recommendation systems. The personalized recommendation system provides users with personalized product recommendations by analyzing users' historical behaviors, interests and social relationships, which greatly improves the user experience and the commercial value of the platform. Among many recommendation algorithms, Logistic regression algorithm has attracted much attention because of its simplicity, ease of understanding and excellent performance in binary classification problems. This experiment aims to use Logistic regression algorithm to build a personalized recommendation system based on Amazon product review data, and predict whether users will recommend a certain product by analyzing users' ratings and recommendation behaviors.

Data collection

This experiment uses Amazon product review data set, which contains a large number of users' ratings of products, review text, recommendation or not. The dataset contains 1597 records involving 27 fields, Include id (unique identification of the record), asins

(ASIN of the product), brand (brand), categories (categories), colors (colors), date Added (date of data addition), date Updated (date of data update), dimension (dimension), ean (EAN). International Commodity Code), keys, manufacturer, manufacturer Number, name, prices, review.date, review. dorecomme nd (recommended or not), review. Num helpful (number of useful reviews), review.rating (rating), review. Source urls (review source URL), review.text (review text), review.title (review title) user City, user Province, Review. username, sizes, UPC, weight and other fields.

Data preprocessing

Data preprocessing is one of the key steps to build a recommendation system. In this experiment, the data file is read first and the data is preliminarily checked. By looking at the basic information of the dataset and the first few lines of data, we found that the dataset contains some missing values and incomplete records. In order to ensure the integrity and accuracy of the data, we selected the key fields of UserId, ProductId, review. rating, review. doRecommend and review. text for analysis, and cleaned and processed the data.

During the data cleansing process, we delete records that contain missing values, ensuring that every piece of data is complete. In addition, we binarize the review.dorecommend field, converting it to 0 and 1 representations, where 1 means recommended and 0 means not recommended. This step lays the foundation for subsequent model training and prediction.

To evaluate the performance of the model, we randomly shuffled the data set and divided it into a training set and a test set in an 80% ratio. The training set is used to train the Logistic regression model, and the test set is used to evaluate the predictive performance of the model. The specific division method is as follows:

First, we set up random seeds to ensure repeatability of the results. Then, the data set is randomly scrambled to avoid the influence of data order on model training. Next, the partition index of the training set and the test set is calculated, with the first 80% of the data as the training set and the last 20% as the test set.

Model training and prediction

After data preprocessing and partitioning, we used the training set data to train the Logistic regression model. Logistic regression is a statistical method widely used in binary classification problems to predict the probability of dependent variables by transforming Logistic functions on linear combinations of independent variables. In this experiment, the independent variable is product review. rating, and the dependent variable is whether it is recommended or not.

After the training, we use the trained model to make predictions on the test set data. The predicted result is the probability that the user recommends. We classify the predicted result with a probability greater than 0.5 as recommended (1), otherwise it is classified as not recommended (0).

Evaluation

To evaluate the performance of the model, we use metrics such as confusion matrix, accuracy, and ROC curve. The confusion matrix shows how the model's predictions compare to the actual situation, It includes the number of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) cases. The accuracy of the model can be calculated through the confusion matrix, that is, the proportion of the model's predictions that are correct. In addition, we plotted the ROC curve and calculated the AUC value. ROC curve showed the Sensitivity and Specificity of a model under different thresholds, the AUC closer to 1, the better the model performance.

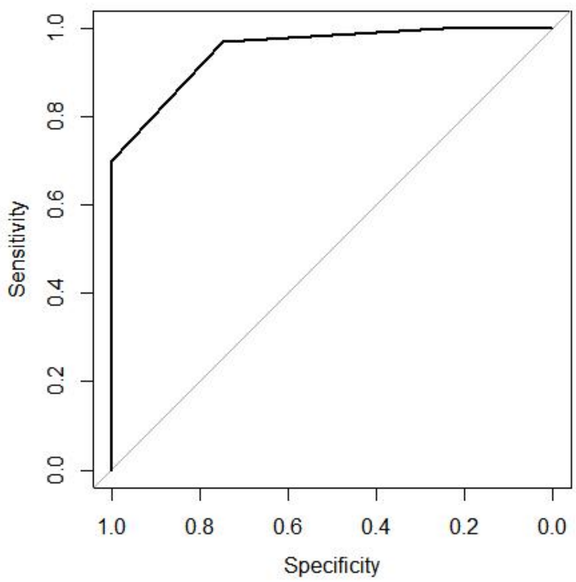


Figure 7 ROC Curve for Logistic Regression Model

The accuracy of the model is 97.22%. ROC curve is shown in the figure 7 above.

The AUC value of 0.9722 indicates that the model has a high performance in distinguishing between recommended and unrecommended users.

Through further analysis of the data, we get the statistical results of review rating distribution and user recommendation.

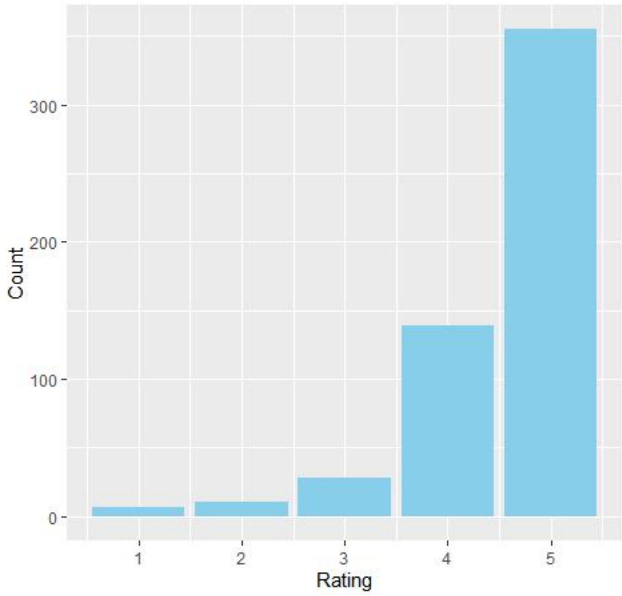


Figure 8 Distribution of Review Ratings

Distribution of review scores: As can be seen (Figure 8) from the distribution of review scores, most of the scores are concentrated in 4 points and 5 points, indicating that users have a high overall evaluation of the product.

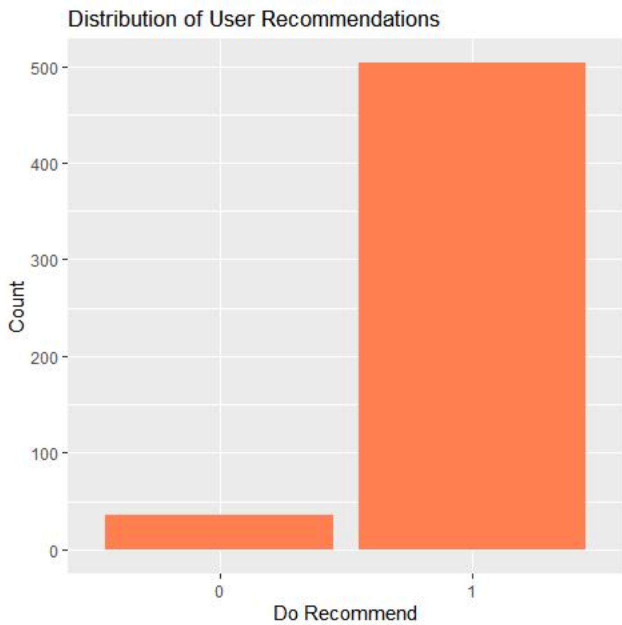


Figure 9 Distribution of User Recommendations

User recommendation: As can be seen (Figure 9) from the distribution chart of user recommendation, most users recommend products while a few users do not, indicating that most users are satisfied with the purchased products.

Conclusions

Through the above analysis and results, we can conclude that the personalized recommendation system based on Logistic regression has high accuracy and reliability in predicting whether users will recommend products. Specifically, Logistic regression model can effectively use user rating information to predict user recommendation behavior. Evaluation indicators such as confusion matrix and ROC curve show that the model has good performance and has high practical application value.

This experiment builds a personalized recommendation system based on Amazon product review data, and verifies the effectiveness of Logistic regression algorithm in such problems. However, there are still many directions worth exploring in the research and application of recommendation system. Future work can consider the following aspects:

Multi-source data fusion: Combining the user's browsing history, purchase history, social network and other multi-source data, to build a more comprehensive user portrait, improve the accuracy and personalization of the recommendation system.

Deep learning methods: Deep learning models, such as convolutional neural networks (CNN) and recurrent neural networks (RNN), are applied to process more complex user behavior data and text data to further improve recommendation performance.

Online learning and real-time recommendation: Develop a recommendation system model that can be updated and adaptive in real time to reflect the latest behaviors and preferences of users, improve the real-time performance of recommendations and user satisfaction.

Privacy protection and security: A privacy protection mechanism is introduced into the recommendation system to ensure the security and privacy of user data and enhance users' trust in the recommendation system.

In short, personalized recommendation system has a wide range of application prospects in e-commerce, social networking, content distribution and other fields.

With the development of data analysis and machine learning technology, recommendation system will become more intelligent and efficient.

References

[1] Chen, Q. (2023). Intelligent recommendation method for product information of e-commerce platform based on machine learning algorithm. In *2023 2nd International Conference on Artificial Intelligence and Autonomous Robot Systems (AIARS)*. https://doi.org/10.1109/AIARS59518.2023.00099

[2] Chen, X., & He, K. (2021). Exploring simple Siamese representation learning. https://doi.org/10.1109/CVPR46437.2021.01549

[3] Dey, R., Nanda, I., Islam, M. S., et al. (2022). Product recommendation system using machine learning through big data in e-commerce website. In *International Conference on Power Electronics & Sustainable Development (ICPESD)*. https://doi.org/10.26480/icpesd.03.2022.229.233

[4] Farooqi, R. A., Kesarwani, S., Shakeeb, M., et al. (2022). Enhancing e-commerce applications with machine learning recommendation systems. *International Journal of Scientific Research in Science, Engineering and Technology*. https://doi.org/10.32628/ijsrset122935

[5] Gattis, M., Winstanley, A., & Bristow, F. (2022). Parenting beliefs about attunement and structure are related to observed parenting behaviours. https://doi.org/10.1080/23311908.2022.2082675

[6] Liu, L. (2022). E-commerce personalized recommendation based on machine learning technology. *Mobile Information Systems*. https://doi.org/10.1155/2022/1761579

[7] McKibbin, W., & Fernando, R. (2021). The global macroeconomic impacts of COVID-19: Seven scenarios. https://doi.org/10.1162/asep_a_00796

[8] Mykhalchuk, T., Zatonatska, T., Dluhopolskyi, O., et al. (2021). Development of recommendation system in e-commerce using emotional analysis and machine learning methods. In *2021 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and

Applications (IDAACS)*. https://doi.org/10.1109/IDAACS53288.2021.9660854

[9] Page, M. J., McKenzie, J. E., Bossuyt, P. M., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. https://doi.org/10.1371/journal.pmed.1003583

[10] Peng, N., Xiao, X., Di, W., et al. (2023). Design and implementation of an intelligent recommendation system for product information on an e-commerce platform based on machine learning. https://doi.org/10.1117/12.3013353

[11] Purcell, W., & Neubauer, T. (2023). Digital twins in agriculture: A state-of-the-art review. https://doi.org/10.1016/j.atech.2022.100094

[12] Rahman, M. S., Sarkar, T. D., Mitasha, U. T., et al. (2024). E-commerce-based smart recommendation system using element-by-element collaborative filtering following with the machine learning technology. In *2024 International Conference on Communication, Computing and Internet of Things (IC3IoT)*. https://doi.org/10.1109/IC3IoT60841.2024.10550386

[13] Tran, D., & Huh, J. (2022). New machine learning model based on the time factor for e-commerce recommendation systems. *The Journal of Supercomputing, 79*. https://doi.org/10.1007/s11227-022-04909-2

[14] Xu, K., Zhou, H., Zheng, H., et al. (2024). Intelligent classification and personalized recommendation of e-commerce products based on machine learning. *ArXiv, 2024*, abs/2403.19345. https://doi.org/10.48550/arXiv.2403.19345

[15] Zhang, X., Guo, F., Chen, T., et al. (2023). A brief survey of machine learning and deep learning techniques for e-commerce research. *J. Theor. Appl. Electron. Commer. Res., 18*. https://doi.org/10.3390/jtaer18040110