ISSN 2616-5767 Vol.7, Issue 5: 111-115, DOI: 10.25236/AJETS.2024.070514

Research on Mobile E-commerce Recommendation Algorithms Based on Logistic Regression Improved Model Features

Guo Jiajie[1,a,*]

[1]*Software Engineering Institute of Guangzhou, Guangzhou, China*
[a]*229559413@qq.com*
[*]*Corresponding author*

Abstract: *With the rapid development of internet technology, the field of e-commerce has experienced unprecedented growth. The rapid expansion of the user base, as well as the explosive increase in the variety and quantity of products, has greatly driven the development of the internet economy. However, this growth has also brought about the issue of "information overload," where users find it difficult to make effective choices and decisions when faced with a vast amount of information. As a result, how to efficiently extract valuable information from users' historical behavior data and, combined with their current context, accurately recommend products that match their needs and preferences has become a significant technical challenge and research topic of interest to both academia and industry. This challenge not only involves providing personalized services and improving user experience but also concerns the enhancement of business conversion rates. In this paper, we combine manually engineered features with location information, apply feature sparsification, and use a large-scale logistic regression model to compare the performance of a basic logistic regression model, random forest model, and GBDT model. The research results demonstrate that the logistic regression model, incorporating manually crafted cross-features and location information, significantly improves the F1 score.*

Keywords: *Feature Engineering, Recommendation algorithms, Recommendation technology*

1. Introduction

With the rapid development of e-commerce and social media platforms, recommendation systems have become a key technology for enhancing user experience and increasing business value. Significant research focuses on how to utilize users' historical behavior data, user profiles, and contextual information to optimize the performance and accuracy of recommendation algorithms.User profile-based recommendation algorithms have received widespread attention. For instance, Huang Yichen (2023) proposed a movie recommendation system that integrates features such as users' interest tags and viewing records to achieve personalized recommendations, improving recommendation accuracy and user satisfaction[1].

In the context of personalized news recommendation systems, Zhai Mei (2023) reviewed several methodologies for optimizing recommendation performance, highlighting the importance of user behavior data and context[2].Xu Wenjian (2023) developed a distributed recommendation algorithm using Hadoop-Mahout, demonstrating its scalability and effectiveness in providing course recommendations[3].Cheng Wanqing (2023) explored the use of multiple machine learning algorithms in water quality index prediction, which also demonstrates the versatility of these models in various fields beyond recommendation systems[4].Wang Shuai (2023) focused on enhancing recommendation systems through personalized data augmentation in self-supervised learning, demonstrating the potential of machine learning in improving system accuracy[5].

In online loan projects, Ouyang Mengqian (2023) applied machine learning algorithms for project evaluation and selection, emphasizing the application of AI-driven models in financial assessment[6].Context-aware recommendation systems have also seen advancements. Cheng Xiufeng (2023) proposed a framework for e-commerce platforms, integrating contextual information to enhance recommendation effectiveness[7].In social media platforms, Huang Chan (2024) designed a content recommendation algorithm using artificial intelligence techniques, illustrating the widespread use of AI in enhancing recommendation accuracy[8].Zhou Chenxi (2023) combined user profiles and link

ISSN 2616-5767 Vol.7, Issue 5: 111-115, DOI: 10.25236/AJETS.2024.070514

prediction methods in a personalized movie recommendation system, showcasing how machine learning algorithms can improve system performance[9].

In cross-border e-commerce, Li Jianbin (2023) explored the impact of product attributes and contextual information on recommendation system performance, showing that GBDT was particularly effective in handling complex recommendation tasks[10].

In the realm of credit assessment, Cao Zaihui (2019) introduced a two-layer classifier model based on ensemble learning, which outperformed single models like SVM and ANN in prediction accuracy[11].

Xiao Yi (2021) discussed the application of AI for COVID-19 diagnosis and treatment, underscoring its role in predictive modeling and disease assessment[12].Qi Qinghou (2023) integrated LightGBM and DeepFM into personalized recommendation models for e-commerce systems, further demonstrating the impact of machine learning in recommendation algorithms[13].Cheng Xiufeng (2023) also explored the implementation and evaluation of recommendation systems in knowledge retrieval scenarios, showing how these models can be applied in various non-commercial contexts[14].Liu He (2024) proposed a short video recommendation algorithm based on user preferences, emphasizing the importance of collaborative filtering methods for improving user retention rates[15].Jiang Yaping (2024) conducted research on movie recommendation systems using Spark, showing how the ALS algorithm and Spark's computational power can enhance real-time recommendation capabilities[16].Tian Renjie (2024) explored the application of contrastive learning and negative sampling in recommendation systems, demonstrating its effectiveness in sparse data environments[17].Zhou Yangtao (2023) reviewed the use of deep learning models in personalized learning resource recommendations, highlighting the role of model optimization in improving recommendation results[18].Su Yonghui (2024) designed a recommendation software using collaborative filtering algorithms, which further supports the use of model-based methods in improving recommendation accuracy[19].Ding Yuchen (2023) proposed a dual debiasing collaborative filtering algorithm for handling sparse implicit feedback data, which significantly improves recommendation accuracy in sparse data environments[20].

This paper extends the research by integrating manually engineered features with location data, applying feature sparsification, and employing logistic regression to improve recommendation accuracy.

2. Principles of the Models

2.1. Logistic Regression Model

Logistic regression is a linear model designed for binary classification tasks. Despite its name, it functions as a classification model rather than a regression model. Logistic regression maps the linear combination of input features (i.e., the result of linear regression) to a probability value between 0 and 1 using the sigmoid (logistic) function. This probability is used to predict the likelihood of a sample belonging to a certain class. If the probability exceeds a specified threshold (usually 0.5), the model classifies the sample as positive; otherwise, it is classified as negative. While logistic regression is relatively simple, it remains widely used in recommendation systems due to its speed and ease of training. By integrating complex feature engineering, logistic regression can achieve robust business performance.

2.2. Random Forest

Random Forest is an ensemble learning method that performs classification or regression by constructing multiple decision trees and aggregating their predictions. Specifically, Random Forest involves sampling multiple subsets from the training data (with replacement) and training a decision tree on each subset. Final predictions are made through majority voting for classification tasks or averaging for regression tasks. This method enhances the model's robustness and generalization capability, thereby reducing the risk of overfitting. Random Forest is particularly well-suited for handling high-dimensional data and nonlinear relationships.

2.3. GBDT (Gradient Boosting Decision Trees)

GBDT is an ensemble learning model based on gradient boosting, primarily used for classification and regression tasks. It iteratively builds a sequence of weak learners (usually decision trees), with each successive model correcting the prediction errors of its predecessor. At each iteration, GBDT trains a new tree on the residuals (i.e., the differences between the actual and predicted values), allowing the

ISSN 2616-5767 Vol.7, Issue 5: 111-115, DOI: 10.25236/AJETS.2024.070514

model to gradually improve its predictions over time.

3. Experimental Analysis

3.1. Data and Preprocessing

This study utilizes a publicly available dataset from the Alibaba Tianchi competition on mobile recommendation algorithms. The dataset is built on real user-item interaction data from a mobile e-commerce platform and includes comprehensive behavioral data for 20,000 users along with millions of product records. The dataset is divided into two parts: the first consists of users' mobile behavior data across the full product set, capturing user ID, product ID, behavior type, user location, product category ID, and behavior timestamp; the second part contains a product subset, which includes product ID, location, and category ID.

The dataset exhibits a significant class imbalance, with a positive-to-negative sample ratio of approximately 1:1200, posing a potential risk of model training failure. To mitigate insufficient feature space coverage during random sampling, we first applied K-means clustering to the negative samples. We then performed subsampling within each cluster to ensure a diverse representation of negative samples, which were combined with positive samples to form a more balanced training set.

Some of the constructed features contained missing values (e.g., xx_diff_hours). For the Logistic Regression (LR) model, we trained the model using a dataset where features with missing values were removed. For the Random Forest (RF) and Gradient Boosting Decision Tree (GBDT) models, missing values were handled by assigning them a value of -1. When using the LR model, the inconsistent scales of different features necessitated the use of the StandardScaler from the sklearn library to normalize the feature values.

3.2. Feature Engineering

We first engineered user-side features to develop a detailed user profile. For this study, only user behavior data from the two weeks preceding the prediction date was considered. Specifically, user profile features included the total number of user behaviors within this two-week period, the total number of different behavior types, the user’s purchase rate, and the time required for the user to make a purchase after clicking.

Next, we constructed product-side features based on user behavior data from the same two-week period. These features included the total number of behaviors for each product, the number of different behavior types, the product’s purchase rate, and the time from the initial click to the final purchase.
Considering the critical role that product category information plays in shaping user purchase intent, we also developed category-level features. These included the total number of behaviors within each category during the two weeks before the prediction date, the number of different behavior types, the category's purchase rate, and the time from click to purchase.

The baseline Logistic Regression (Base LR) model utilized these single-dimension features. However, the improved Logistic Regression (LR) model introduced additional cross-features to capture more nuanced relationships. These cross-features included the ranking of the number of users interacting with a product within its category, behavior ranking, and sales volume ranking. Moreover, the total number of behaviors for each product and category during the two-week period was considered, along with the different types of behaviors.

Recognizing the importance of location information in recommendation systems, we incorporated location data into the model. We applied Geohash encoding, a spatial encoding technique that converts geographic coordinates into hierarchical string codes. The final cross-feature was a binary (0/1) feature based on matching user and product location codes, further enriching the model’s ability to represent spatial relationships.

3.3. Model Selection and Construction

The user profile features, product profile features, and product category information were selected as inputs to the models. A basic Logistic Regression model served as the baseline, while Random Forest and GBDT models were used for comparison. Additionally, cross-features, including location information, were introduced to the Logistic Regression model to enhance its predictive capabilities.

ISSN 2616-5767 Vol.7, Issue 5: 111-115, DOI: 10.25236/AJETS.2024.070514

3.4. Evaluation and Comparison of Prediction Results

The model performance was evaluated using the F1 score, a widely used metric in recommendation systems. The F1 score is the harmonic mean of Precision and Recall, balancing both metrics. Precision refers to the proportion of true positives among the samples predicted as positive, while Recall represents the proportion of true positives identified out of all actual positive samples. By balancing these two metrics, the F1 score provides a comprehensive evaluation of model performance, particularly in cases of imbalanced positive and negative sample distributions.

The study utilized three months of data, and a pipeline was built in Spark to construct both the primary features and the additional cross-features. Each feature set was built using data from the two weeks prior to the prediction date to generate the training and validation sets. The test set, consisting of data from the most recent week, was used to evaluate the generalization ability of the models. Table 1 presents the evaluation metrics of the four models across various datasets.

Table 1: Comparison Table of Four Model

Model	Dataset	F1 Score	Precision	Recall
Base LR	Training	0.7	0.72	0.68
	Validation	0.68	0.7	0.66
	Test	0.67	0.69	0.65
Random Forest	Training	0.85	0.87	0.83
	Validation	0.74	0.75	0.73
	Test	0.73	0.74	0.72
GBDT	Training	0.88	0.89	0.87
	Validation	0.77	0.78	0.76
	Test	0.76	0.77	0.75
Improved LR	Training	0.81	0.82	0.8
	Validation	0.79	0.8	0.78
	Test	0.78	0.79	0.77

Based on the prediction results from the test set, we observe that the Logistic Regression model, enhanced with location information and various cross-features, outperforms in both precision and recall, leading to a higher F1 score. This demonstrates that manually engineered cross-features and location data offer valuable support in enabling the model to capture underlying data patterns more effectively.

4. Conclusions

This study summarizes the application and advantages of an improved feature-based logistic regression model in mobile e-commerce recommendation algorithms. As e-commerce platforms continue to grow, the issue of information overload has increasingly affected user experience and business conversion rates. This paper proposes leveraging user behavior data, product information, and location data, combined with feature engineering and model optimization, to enhance recommendation system performance.

The study compares logistic regression, random forest, and GBDT models, showing that the logistic regression model, through the introduction of manually crafted cross-features and geographic information, significantly improves key metrics such as F1 score, precision, and recall. This highlights its potential in large-scale data processing and recommendation systems. The study not only provides a new technical approach to improving personalized recommendations on mobile e-commerce platforms but also lays the groundwork for future research in recommendation systems.

The main contribution of this study is validating that the improved feature-based logistic regression model can significantly enhance recommendation accuracy in complex scenarios, effectively addressing the issue of information overload. This enables more precise product recommendations, ultimately improving user experience and the commercial value of the platform.

References

[1] Huang Yichen. Design and Implementation of a Movie Recommendation System Based on User Profiles [J]. Journal of Tongren University, 2023.
[2] Zhai Mei. A Review and Discussion on Personalized News Recommendation Systems[J]. Computer

ISSN 2616-5767 Vol.7, Issue 5: 111-115, DOI: 10.25236/AJETS.2024.070514

Science & Exploration, 2023.
[3] Xu Wenjian. Distributed Course Recommendation Algorithm Based on Hadoop-Mahout[J]. Computer Engineering and Design, 2023.
[4] Cheng Wanqing. Construction and Evaluation of Water Quality Index Prediction Models Based on Multiple Machine Learning Algorithms[J]. Journal of Environmental Sciences, 2023.
[5] Wang Shuai. Self-supervised Sequential Recommendation Algorithm Based on Personalized Data Augmentation[J]. Journal of Computer Applications, 2023.
[6] Ouyang Mengqian. Evaluation and Selection of Online Small Loan Projects Based on Machine Learning Algorithms[J]. E-commerce Research, 2023.
[7] Cheng Xiufeng. Research on Context-Aware E-commerce Platform Recommendation System Framework [J]. Journal of Management Science, 2023.
[8] Huang Chan. Design of a Content Recommendation Algorithm for Social Media Platforms Based on Artificial Intelligence[J]. Computer Programming Skills & Maintenance, 2024.
[9] Zhou Chenxi. Research on Personalized Movie Recommendation Based on User Profile and Link Prediction[J]. Computer Science & Exploration, 2023.
[10] Li Jianbin. Cross-border E-commerce Recommendation Algorithm Based on Product Attributes and Context[J]. E-commerce Research, 2023.
[11] Cao Zaihui, Yu Dongxian, Shi Jinfa, Zong Sisheng. Application of Two-layer Classifier Models to Personal Credit Assessment[J]. Control Engineering of China, 2019, 26(12): 2231-2234.
[12] Xiao Yi, Liu Shiyuan. The Application and Value of Artificial Intelligence Technology in the Diagnosis and Treatment of COVID-19[J]. Chinese Journal of Medical Imaging Technology, 2021, 29(4): 289-291.
[13] Qi Qinghou. Research on Personalized Recommendation Models for E-commerce Business Systems Integrating LightGBM and DeepFM[J]. Computer Engineering, 2023.
[14] Cheng Xiufeng. Simulation Implementation and Evaluation of Recommendation Systems in Knowledge Retrieval Scenarios[J]. Journal of Computer Applications Research, 2023.
[15] Liu He, Liu Chunsheng. Research on Short Video Recommendation Algorithms Based on User Preferences in the Context of Big Data[J]. Broadcast Television Network, 2024.
[16] Jiang Yaping. Research and Application of Key Technologies in Movie Recommendation Systems Based on Spark[J]. Computer Programming Skills & Maintenance, 2024.
[17] Tian Renjie. Research on Graph Contrastive Learning Recommendation Algorithms Based on Mixed Negative Sampling[J]. Journal of Computer Science, 2024.
[18] Zhou Yangtao. A Review on Personalized Learning Resource Recommendation Based on Deep Learning[J]. Educational Science Research, 2023.
[19] Su Yonghui. Design and Application of Campus Association Recommendation Software Based on Collaborative Filtering Algorithm[J]. Journal of Information Technology, 2024.
[20] Ding Yuchen. Dual Debiasing Collaborative Filtering Recommendation Algorithm for Sparse Implicit Feedback Data[J]. Computer Science and Technology, 2023.