Analysis of promotional online shopping behavior based on machine learning

Weihao Huang

Software Academy, South China Normal University, Guangzhou, 528225, China

Abstract. Artificial intelligence's widespread use in e-commerce enables precise, personalized services and recommendations through deep data analysis, enhancing user experience and loyalty. Studying user behavior is essential for AI applications, providing businesses with valuable insights into customers' needs and behaviors, ultimately boosting sales efficiency and market competitiveness. This paper aims to investigate the application of artificial intelligence in the e-commerce domain, focusing on user behavior analysis and purchase behavior prediction. Using a user behavior dataset provided by the Tianchi platform, machine learning models, particularly random forest and logistic regression models, are constructed to predict user behavior in purchasing specific product categories. The results demonstrate that the random forest-based machine learning model can predict user purchase behavior effectively. Furthermore, the paper provides statistical data on user behavior and introduces related techniques such as feature engineering, performance metrics, and algorithms. The study highlights that artificial intelligence can offer personalized recommendations and services by analyzing users' historical behavior, interests, and preferences, optimizing merchants' product and service strategies, and ultimately enhancing sales efficiency and user satisfaction.

Keywords: online shopping behavior; machine learning.

1. Introduction

The application of artificial intelligence (AI) in the e-commerce sector has become increasingly widespread. By conducting deep analysis of massive data, AI technologies can provide more accurate and personalized services and recommendations, enhancing users' shopping experience and loyalty. Among these, studying user behavior is a key aspect of AI applications, which offers e-commerce businesses deeper insights into their customers, helping them better understand user needs and behaviors, ultimately improving sales efficiency and market competitiveness.

Specifically, AI technologies can analyze users' historical behavior, interests, and preferences to provide personalized recommendations and services, such as personalized recommendation systems [1] and intelligent customer service [2]. These applications not only increase users' purchase rates and satisfaction but also improve e-commerce businesses' market competitiveness and commercial value. On the other hand, deep analysis of user behavior data can help better understand users' needs, preferences, and behavioral patterns, optimizing product and service strategies for merchants, and increasing sales efficiency and user satisfaction. For example, by predicting user purchasing behavior, businesses can develop targeted marketing strategies for different customer segments, boosting sales and profits. Moreover, AI technologies can help merchants optimize customer service and feedback, such as automated Q&A systems [3] and sentiment analysis [4], enhancing customer satisfaction and service quality. Therefore, the application of AI in e-commerce and the study of user behavior are of great significance, providing more accurate and efficient services and recommendations for e-commerce businesses while also promoting the rapid development of the e-commerce industry.

However, when using AI technologies to analyze and predict shopping behavior, some challenges are faced [5], such as data quality issues, feature extraction problems, and algorithm selection. To better apply AI technologies and increase the commercial value of e-commerce businesses, it is necessary to further study shopping behavior analysis and prediction methods, improving the predictive accuracy and practical application effects of the models.

The purpose of this paper is to investigate how to use AI methods to analyze user shopping behavior data, predict user purchasing behavior, and provide personalized recommendations and

marketing services for e-commerce businesses. Specifically, this paper will employ machine learning algorithms to analyze and model user browsing, clicking, and purchasing behavior data, predicting user purchasing behavior. To enhance the model's predictive accuracy and practical application effects, this paper will explore optimizing feature selection and algorithm choice, and utilizing deep learning models to improve prediction accuracy.

The research questions of this paper have practical significance and application value. By studying shopping behavior analysis and prediction methods, this paper will offer e-commerce businesses an AI-based approach for shopping behavior analysis and personalized recommendations, helping merchants better understand user needs and behaviors, increasing user satisfaction and market competitiveness. Simultaneously, the research methods and techniques of this paper can provide a reference and guidance for data analysis and prediction problems in other fields.

2. Literature review

Consumer purchasing behavior research is an important field, involving the entire process from the motivation of demand to the occurrence of purchasing behavior. According to existing research, the study of user purchasing behavior can be divided into two aspects: one aspect is to analyze the factors influencing user purchasing behavior or the impact of a specific factor on purchase intention; the other aspect is predicting purchasing behavior.

In the study of factors influencing consumer [6] purchasing behavior, some scholars explore external and internal factors, including product characteristics, payment security, cultural aspects, and consumer interests. Research also examines the effects of demographic factors, online shopping popularity, and merchant reputation on purchasing behavior. In addition, researchers analyze the impact of online marketing on online shopping user behavior from a psychological perspective, such as the mediating role of perceived value and risk.

In terms of predicting user purchasing behavior, most research is carried out using machine learning methods [7]. Typically, traditional machine learning models are used for predictions first, followed by the introduction of new concepts or methods to improve the models, aiming to achieve better prediction results than conventional models.

In conclusion, consumer purchasing behavior research is a multidimensional, interdisciplinary field that involves both in-depth analysis of internal and external factors influencing purchasing behavior and prediction of purchasing behavior. Previous research has provided valuable insights for understanding consumer behavior, improving the quality of products and services, optimizing marketing strategies, and enhancing merchant competitiveness.

Machine learning is an AI-based technology that trains algorithms to learn patterns and trends from historical data to predict future events. Machine learning can use user behavior information from historical data, such as browsing, favoriting, and purchasing, to construct predictive models. Common machine learning algorithms include logistic regression, random forests, and neural networks.

Compared to traditional manual modeling methods, machine learning can automate model construction and parameter optimization more effectively, handling large amounts of data and complex relationships with higher accuracy and efficiency. Machine learning can also improve model performance and accuracy through continuous learning and adjustments, offering better adaptability and robustness. Therefore, applying machine learning in predicting user behavior holds significant advantages and is a primary focus of current research.

3. Data source and analysis

We will use the Taobao User Behavior dataset [8] provided by the Tianchi platform for this study. The dataset contains all behaviors of approximately one million random users between November 25, 2017, and December 3, 2017, including clicks, purchases, adding to cart, and likes. Each row in the dataset represents a user behavior, consisting of user ID, item ID, item category ID, behavior type,

and timestamp. The dataset is relatively large, with the original data containing 987,994 users, 4,162,024 items, and 100,150,807 behaviors, providing us with a rich data resource for our research.

We first checked the original data for any missing information. After our investigation, there were no missing values in the original dataset, so we did not need to deal with missing or abnormal values. However, there were a small number of data points outside the date range of November 25, 2017, to December 3, 2017, so we filtered the dataset to exclude invalid data. After excluding 55,576 invalid data points, we retained 100,095,231 data points. The final data used is as follows (Table 1).

Table 1. The data used in the article

Data	Count	Value Range
user_id	987,991	(1 - 1,018,011)
item_id	4,161,138	(1 - 5,163,070)
category_id	9,437	(80 - 5,162,429)
behavior_type	4	('pv', 'buy', 'cart', 'fav')
timestamp	777,600	('2017-11-25 00:00:00' - '2017-12-03 23:59:59')

We counted all pv, buy, cart, fav behaviors in the dataset by hour and plotted them in the following image.

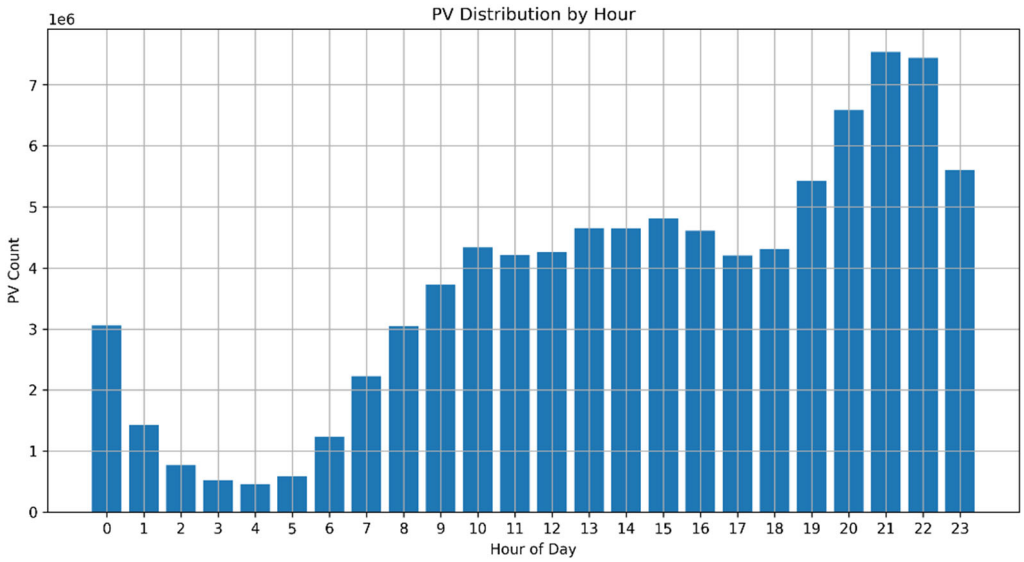

Fig. 1 PV distribution by hour

It can be observed that user activity changes over time, showing a distinct periodicity. User activity is low during the early morning hours (0-6 am) and gradually increases over time. In the evening hours (18-23), user activity reaches its peak, especially during the 21-22 hour period, where the number of all types of behaviors reaches its highest value.

For various behavior types, the number of PVs (page views) is significantly higher than for other behavior types. This indicates that when users use the platform, the number of times they browse product pages accounts for the vast majority. Compared to other behaviors, users are more inclined to browse products rather than immediately buy, add to cart, or favorite. This is consistent with our intuition.

Although the total number of Buy (purchase), Cart (add to cart), and Fav (favorite) behaviors is significantly lower than PV (page views), their distribution over time still maintains a similar periodic pattern. In the evening hours, the number of purchases is the highest, especially during the 20–22 hour period, where the number of purchase behaviors significantly increases. This may be because most users are more willing to shop during their leisure time in the evening.

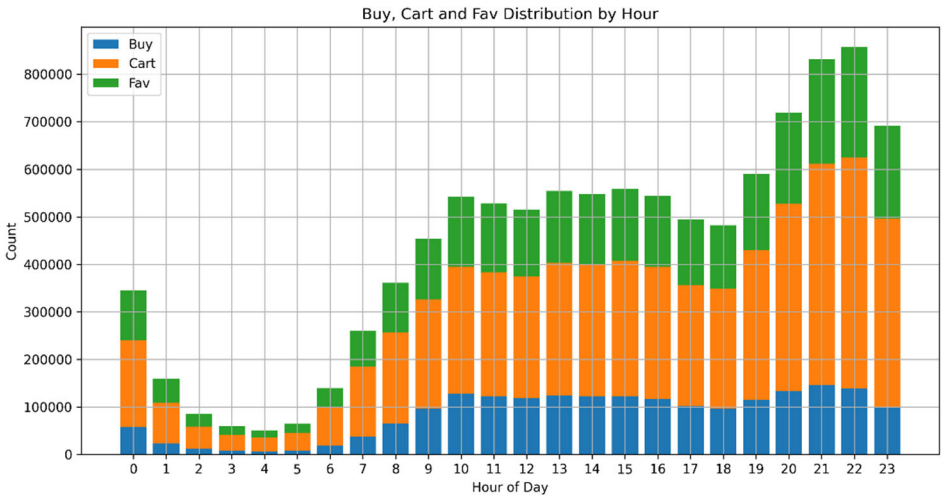

Fig. 2 Buy, Cart and Fav distribution by hour

Looking at the number of Buy (purchase) behaviors, the overall purchase conversion rate is relatively low. This may indicate that when users browse products, they may add items to their cart or favorite them, but they do not immediately make a purchase. This also means that to some extent, e-commerce platforms can increase sales by improving the conversion rate.

4. Feature processing and model construction

Feature construction involves combining domain-specific knowledge and real-world scenarios to explore more accurate and comprehensive indicators and characteristics of the research object based on raw data. This process creates new features that have a positive impact on model training and practical significance. In the original dataset, basic features usually include user ID, product ID, category ID, behavior type, and behavior time. In addition to these basic features, we need to use specific methods to construct more valuable sample features.

For this dataset, we can build features based on user purchase preferences and behaviors. This paper analyzes user behavior characteristics from a statistical perspective, constructing a series of derived features based on the four types of user behaviors (click, purchase, add to cart, like) and user attributes. (Table2)

Table 2. Feature processing

Feature	Description
1-1	Number of times user clicks on products
1-2	Number of unique products clicked by user
1-3	Number of unique product categories clicked by user
2-1	Number of times user adds products to cart
2-2	Number of unique products added to cart by user
2-3	Number of unique product categories added to cart by user
3-1	Number of times user purchases products
3-2	Number of unique products purchased by user
3-3	Number of unique product categories purchased by user
4-1	Number of times user likes products
4-2	Number of unique products liked by user
4-3	Number of unique product categories liked by user

As these features come from different dimensions, their units and scales vary, potentially leading to an unreasonable distribution of feature weights during model training, thereby affecting the training results. To eliminate the impact of scale differences, it is necessary to normalize the features, unifying

their scales and units. Common normalization methods include Min-Max normalization and Z-score normalization. In this dataset, we use Min-Max normalization for feature normalization, resulting in a sparse matrix, with missing feature values filled with zeros.

Logistic Regression [9] is a linear model widely used for binary classification problems. It uses a logistic function (usually the Sigmoid function) to map linear regression output to probability space, generating a probability output representing the likelihood of a sample belonging to a specific category. Logistic Regression aims to find the best-fitting decision boundary, dividing the input feature space into two regions corresponding to the positive and negative classes.

In this paper, we use the Logistic Regression class from the scikit-learn library to implement the logistic regression model. The parameters of the Logistic Regression class include penalty, solver, max_iter, and C. The penalty is the regularization term to prevent overfitting and improve generalization ability, with options being L1 and L2. L2 regularization typically meets most needs. The solver determines the optimization algorithm for the logistic regression loss function, with options being newton-cg, lbfgs, liblinear, sag, and saga. Max_iter represents the maximum number of iterations for the algorithm, while C is the inverse of the regularization coefficient, inversely proportional to the regularization strength. Through manual parameter tuning, the main settings for Logistic Regression in this paper are penalty=L2, solver=sag, max_iter=100, and C=1.0.

Random Forest [10] is an ensemble learning method that combines multiple decision trees to build a powerful classification or regression model. Random Forest generates a set of different decision trees by bootstrap sampling of the training data and randomly selecting feature subsets. The final prediction of the Random Forest is obtained by voting on the prediction results of these decision trees (for classification problems).

The Random Forest model is implemented using the RandomForestClassifier class in the sklearn library, with main parameters including criterion, max_features, n_estimators, max_depth, and min_samples_split. Criterion measures the quality of classification, with options being gini and entropy; gini uses gini impurity to measure classification quality, while entropy uses information gain. Max_features represent the maximum number of features a single decision tree can use; n_estimators is the number of trees to be built before the final vote, with more trees generally leading to better predictive performance; max_depth is the maximum depth of the decision tree; min_samples_split is the minimum number of samples required to continue splitting a node, limiting further subdivision of the subtree. After manual parameter tuning, the main settings for Random Forest in this paper are criterion=gini, max_features=auto, n_estimators=100, max_depth=5, and min_samples_split=2.

5. Evaluation indicators and methods

To ensure the reliability of evaluation results, we typically use cross-validation and test sets to assess model performance, avoiding overfitting and underfitting issues. For classification tasks, we often use a series of evaluation metrics to measure model performance. These metrics include precision, recall, F1 score, and AUC.

Precision (Precision = TP / (TP + FP)): Precision is the proportion of actual positive cases among samples predicted as positive. It is the ratio of true positives (TP) to all samples predicted as positive (i.e. true positives and false positives). A higher precision indicates a higher proportion of actual positive cases among samples predicted as positive.

Recall (Recall = TP / (TP + FN)): Recall is the proportion of actual positive cases that are predicted as positive. It is the ratio of true positives to all actual positive samples (i.e. true positives and false negatives). A higher recall indicates a higher proportion of correctly predicted positive cases among actual positive samples.

F1 Score (F1 Score = 2 * (Precision * Recall) / (Precision + Recall)): The F1 score is the harmonic mean of precision and recall, providing a comprehensive measure of both. A higher F1 score indicates that both precision and recall are high. The F1 score is particularly useful when dealing with imbalanced samples, as it considers the predictive performance of both positive and negative cases.

AUC: AUC is the area under the Receiver Operating Characteristic (ROC) curve. The ROC curve depicts the relationship between true positive rates (TPR) and false positive rates (FPR) at different classification thresholds. AUC values range from 0 to 1, with values closer to 1 indicating better classifier performance. AUC considers all possible classification thresholds, making it suitable for evaluating the overall performance of a classifier across various thresholds.

6. Experimental results and analysis

The aim of this paper is to construct a machine learning model to predict whether users will purchase a specific product category in the future, using a dataset of anonymized historical behavior data from approximately 1 million users provided by the Tianchi platform. The analysis is conducted by product category, taking into account factors such as user behavior types and purchase preferences while creating a series of new features. Due to the differences in units and dimensions among various features, normalization is applied.

This study employs Python 3 as the main tool, utilizing libraries such as numpy, pandas, and scikit-learn to support various machine learning models, including logistic regression and random forest. The raw data is read from CSV files using pandas and then converted to pickle format. Pickle, a module in Python's standard library, enables the efficient processing of data by serializing Python objects into binary format or restoring them as Python objects.

As the research objective pertains to binary classification, user-product category pairs with purchasing behavior are considered positive samples, while those without are negative samples. The entire dataset is randomly divided into training and testing sets at a ratio of 8:2.

Logistic regression and random forest models were applied to the test set for classification prediction, and precision, recall, F1 score, and AUC values were recorded. Comparing the precision, recall, and F1 scores of the two models revealed that their precision was 58.17% and 62.46%, respectively, while recall was 19.95% and 25.48%, resulting in overall low F1 scores. This indicates that both classifiers are cautious when classifying, only assigning samples to the positive class with high confidence, which leads to high precision but low recall. Given the highly imbalanced nature of the samples and the focus on user behavior research, the emphasis is placed on prediction accuracy rather than comprehensiveness, making this an acceptable outcome. Based on AUC values, logistic regression and random forest models show an AUC of 0.6081 and 0.6448, respectively, indicating that both models are effective and feasible for predicting user purchasing behavior.

Table 3. Experimental results

Model	Precision	Recall	F1 Score	AUC Score
Logistic Regression	0.581671026	0.199526219	0.301629251	0.608071045
Random Forest	0.624592236	0.254763067	0.368815497	0.644823133

The results indicate that employing an ensemble learning-based random forest model for predicting user purchasing behavior improves the F1 score and AUC value compared to logistic regression, demonstrating the effectiveness of ensemble algorithms for enhancing prediction accuracy. The random forest model boasts the advantages of ensemble learning, providing better classification performance, generalization ability, learning capacity, computational speed, and the ability to efficiently process high-dimensional data while reducing the risk of overfitting.

In this experiment, the performance of logistic regression and random forest models improved compared to other models. The random forest model achieved an F1 score of 0.3688, a 22.29% increase compared to the logistic regression model, and an AUC value of 0.6448, a 6.04% increase. These results demonstrate that the random forest model effectively improves the performance of the learning algorithm, enhances the model's generalization ability, reduces the risk of overfitting, and is feasible for optimizing user purchasing behavior prediction.

In this experiment, we applied machine learning models, such as logistic regression and random forest, to predict user shopping behavior. Through the use of these models, we discovered that machine learning models can mine user shopping behavior and potentially provide valuable information for recommendation systems and marketing services. In practical scenarios, these models have significant application value, which will be discussed from the perspectives of recommendation and marketing services.

Firstly, in terms of recommendation systems, by predicting user shopping behavior, we can better understand users' interests and needs, providing more personalized product recommendations. This will help increase user activity on the platform, enhance user satisfaction, and improve loyalty. Recommendation systems based on user purchase behavior prediction can update recommended content in real-time, ensuring users receive the latest and most relevant product recommendations during each platform visit. Moreover, by predicting purchasing behavior, the recommendation system can provide merchants with targeted product ranking and display strategies, improving product exposure and conversion rates.

Secondly, regarding marketing services, accurate user shopping behavior predictions can offer targeted marketing strategies for e-commerce platforms. For example, for potential buyers, the platform can send promotional materials such as coupons and event notifications to increase purchasing intent and conversion rates. Simultaneously, by analyzing prediction results, the platform can achieve refined marketing, avoiding sending excessive irrelevant information to uninterested users, reducing user fatigue. Additionally, purchase behavior prediction can help the platform allocate advertising resources reasonably, achieving more efficient advertising placement. By predicting, the platform can target advertising to user groups with high purchasing potential, improving ad click-through and conversion rates.

7. Conclusion

Using the Taobao user behavior dataset provided by the Tianchi platform, this paper constructs a machine learning model to predict whether users will purchase a specific product category in the future. The analysis is conducted by product category, considering user behavior types and purchase preferences, creating a series of new features and normalizing these features. The experiment employed logistic regression and random forest models, finding that the random forest model demonstrates better performance in predicting user purchasing behavior. The experimental results show that the ensemble learning-based random forest model is feasible for predicting user purchasing behavior.

In the future, more feature engineering techniques can be employed in this field to extract more representative features and improve the prediction performance of the model. Advanced machine learning and deep learning methods, such as support vector machines, neural networks, and XGBoost, can also be explored to further enhance prediction accuracy. Additionally, the prediction results can be utilized to optimize personalized recommendations and marketing strategies on e-commerce platforms, increase conversion rates, or combine with other data sources such as user social media behavior and geographic location to uncover potential user shopping behavior patterns, providing more value for recommendation systems and marketing strategies.

References

[1] Zhou, Xujuan, et al. "The state-of-the-art in personalized recommender systems for social networking." Artificial Intelligence Review 37 (2012): 119-132.

[2] Fu, Min, et al. "ICS-Assist: Intelligent customer inquiry resolution recommendation in online customer service for large E-commerce businesses." Service-Oriented Computing: 18th International Conference, ICSOC 2020, Dubai, United Arab Emirates, December 14–17, 2020, Proceedings 18. Springer International Publishing, 2020.

[3] Diaconita, Irina, Christoph Rensing, and Stephan Tittel. "Getting the information you need, when you need it: a context-aware Q&A system for collaborative learning." Open Learning and Teaching in Educational Communities: 9th European Conference on Technology Enhanced Learning, EC-TEL 2014, Graz, Austria, September 16-19, 2014, Proceedings 9. Springer International Publishing, 2014.

[4] Hussein, Doaa Mohey El-Din Mohamed. "A survey on sentiment analysis challenges." Journal of King Saud University-Engineering Sciences 30.4 (2018): 330-338.

[5] De Bruyn, Arnaud, et al. "Artificial intelligence and marketing: Pitfalls and opportunities." Journal of Interactive Marketing 51.1 (2020): 91-105.

[6] Bhullar, Arshan, and Pushpinder Singh Gill. "Future of mobile commerce: an exploratory study on factors affecting mobile users' behaviour intention." International Journal of Mathematical, Engineering and Management Sciences 4.1 (2019): 245.

[7] Koehn, Dennis, Stefan Lessmann, and Markus Schaal. "Predicting online shopping behaviour from clickstream data using deep learning." Expert Systems with Applications 150 (2020): 113342.

[8] Tianchi, Xiaomiaomeng. "User Behavior Data from Taobao for Recommendation" Tianchi, https://tianchi.aliyun.com/dataset/649?t=1680319691970&lang=en-us. Accessed 24 April 2023.

[9] Hilbe, Joseph M. Logistic regression models. CRC press, 2009.

[10] Biau, Gérard, and Erwan Scornet. "A random forest guided tour." Test 25 (2016): 197-227.