ISSN 2616-5775 Vol. 9, Issue 2: 39-45, DOI: 10.25236/AJCIS.2026.090206

Construction of Big Data Mining and Accurate Prediction Model for Consumer Behavior Based on Machine Learning

Wen Wen*

Department of International Business Administration, Woosong University, Daejeon, 34606, Korea
**Corresponding author*

Abstract: *In the context of the deep empowerment of the digital economy in the transformation of the consumer market, consumer behavior presents digital, multidimensional, and complex characteristics. The value mining of massive consumer behavior data has become the key for enterprises to achieve precise marketing and enhance core competitiveness. Traditional consumer behavior analysis methods have limitations such as low data processing efficiency, insufficient prediction accuracy, and difficulty in capturing complex behavioral associations, which cannot meet the needs of enterprise refined operations. This article takes consumer behavior big data as the research object, integrates consumer behavior theory and big data mining technology, and constructs a complete system of consumer behavior big data processing and accurate prediction models. Firstly, we establish a multi scenario consumer behavior big data system to address the pain points of data heterogeneity and uneven quality through data collection, preprocessing, feature engineering, and dataset partitioning; Secondly, design a three-level prediction architecture of "benchmark model+integrated model+fusion optimization", selecting logistic regression and decision tree as benchmark models, XGBoost, LightGBM, CatBoost as integrated models, and combining grid search and Bayesian optimization to achieve hyperparameter optimization, constructing a model-based fusion strategy to improve prediction performance; Finally, the effectiveness of the model is verified through systematic experiments, and application strategies for the model are proposed in combination with e-commerce and retail scenarios. The experimental results show that the AUC value of the integrated prediction model reaches 0.93, and the F1 score reaches 0.91, significantly better than the single model, which can accurately predict consumer purchase behavior and loss risk. The research results of this article not only enrich the application of machine learning in consumer behavior analysis, but also provide theoretical support and practical reference for precision marketing and user management in enterprises, with important theoretical value and application prospects.*

Keywords: *machine learning, consumer behavior, big data mining, accurate prediction, ensemble learning*

1. Introduction

In the current era of deep penetration of the digital economy, consumer behavior has fully entered the stage of digital transformation [1]. The massive data generated from multiple scenarios such as e-commerce transactions, social interactions, and offline consumption constitute the core carrier of consumer behavior big data, providing solid data support for accurately capturing consumer demand and predicting consumption trends. Traditional consumer behavior analysis methods are limited by their data processing capabilities, making it difficult to uncover potential correlations in multidimensional data. The accuracy and timeliness of predictions cannot meet the practical needs of precision marketing and risk management for enterprises. Building efficient big data mining and prediction models has become an urgent need for industry development and academic research.

Domestic and foreign scholars have conducted extensive research on consumer behavior mining and machine learning prediction [2]. Existing results have confirmed the feasibility of machine learning algorithms in consumer behavior classification and demand prediction, but there are still significant limitations: some studies focus on the application of a single algorithm and lack multi model comparison and optimization; Some studies overlook the core role of data preprocessing and feature engineering, resulting in insufficient model generalization ability; At the same time, existing research lacks sufficient attention to the business applicability of the model, making it difficult to effectively connect theoretical research with practical applications in enterprises.

ISSN 2616-5775 Vol. 9, Issue 2: 39-45, DOI: 10.25236/AJCIS.2026.090206

Based on this, this article takes consumer behavior big data as the research object, integrates big data mining technology and machine learning algorithms, focuses on the core goals of model construction and accurate prediction, clarifies the research content and technical ideas, and breaks through the limitations of traditional analysis methods by optimizing data processing flow and building efficient prediction models. The core innovation of this article lies in constructing a feature system that adapts to the characteristics of consumer behavior data, combining ensemble learning algorithms to improve prediction accuracy, and strengthening the business interpretability and practicality of the model, providing theoretical support and practical reference for enterprise precision marketing and user management, and promoting theoretical innovation and technological application in the field of consumer behavior analysis.

2. Relevant theories and technical foundations

2.1 Core theory

The core theory is the logical foundation of this study, mainly covering consumer behavior theory and big data mining theory, which support each other and jointly define the research boundary and clarify the research direction. Consumer behavior theory, as the foundation for analyzing the laws of consumer decision-making, focuses on the behavioral characteristics and intrinsic driving mechanisms of consumers throughout the entire process of demand generation, information collection, purchase decision-making, user experience, and feedback evaluation. Its core branches include rational behavior theory, planned behavior theory, and technology acceptance model. This theory provides clear guidance for big data mining of consumer behavior, identifying the key dimensions of user behavior, consumer preference associations, and decision influencing factors that need to be captured, avoiding the blindness of data mining, and ensuring that the mined behavioral characteristics are highly consistent with consumer demand.

The core goal of big data mining theory is to transform the value of massive heterogeneous data. Its core logic is to screen, process, and analyze consumer behavior data from multiple scenarios and dimensions through standardized and systematic processes, extracting potential correlations, behavioral patterns, and development trends hidden in the data. Its core process covers five key steps: data collection, preprocessing, feature extraction, pattern mining, and result application. It provides a standardized framework for efficient processing of consumer behavior big data, effectively solving the pain points of single dimension, low efficiency, and insufficient association mining in traditional data processing, and laying a high-quality data foundation for the construction of subsequent machine learning models.

2.2 Key technologies

The key technology is the core support for the implementation of accurate prediction models, with a focus on machine learning algorithms and model evaluation techniques. The synergy between the two ensures the prediction accuracy and practicality of the model. Machine learning algorithms, as the core of model construction, combine the high-dimensional and nonlinear characteristics of consumer behavior data to select algorithm types with strong adaptability, mainly including traditional classification algorithms and ensemble learning algorithms. Logistic regression (LR), as a benchmark classification model, has the advantages of high computational efficiency and strong interpretability, and is suitable for preliminary classification and prediction of consumer behavior; Random Forest (RF) uses a multi decision tree ensemble learning strategy to effectively alleviate overfitting problems, improve model generalization ability, and efficiently process high-dimensional features in consumer behavior data; Ensemble learning algorithms such as XGBoost and LightGBM optimize model performance through gradient boosting strategies, accurately capturing nonlinear correlations and complex patterns in consumer behavior data, significantly improving prediction accuracy, and are the core algorithms of this model construction.

Model evaluation technology is used to scientifically and objectively measure the performance of predictive models, providing clear direction for model optimization. This study selected multidimensional evaluation indicators, including accuracy, precision, recall, F1 score, and area under the curve (AUC). AUC focuses on evaluating the model's generalization and discrimination abilities, while precision and recall are used to balance prediction accuracy and comprehensiveness. The collaboration of multiple indicators ensures the objectivity and comprehensiveness of model performance evaluation, providing a reliable basis for subsequent model hyperparameter optimization and

ISSN 2616-5775 Vol. 9, Issue 2: 39-45, DOI: 10.25236/AJCIS.2026.090206

improvement.

3. Consumer behavior big data processing and model construction

3.1 Construction of big data system

The construction of a big data system for consumer behavior is a prerequisite for accurate model prediction. The core revolves around four major stages: data source collection, preprocessing, feature engineering, and dataset partitioning, ensuring the integrity, accuracy, and availability of data and adapting to the input requirements of machine learning algorithms [3]. The data source adopts a multi scenario fusion collection mode, focusing on selecting e-commerce platform transaction data, user browsing, clicking, adding, and purchasing data, user profile data, and offline consumption related data, covering the entire consumer behavior chain. E-commerce transaction data includes core indicators such as order amount, payment method, and purchase frequency, while user behavior data covers dimensions such as visit time, click path, and interaction frequency. User profile data includes basic information such as age, gender, consumption level, and region. Multi-dimensional data fusion can comprehensively capture consumer behavior characteristics and consumption preferences.

Data preprocessing is the key to improving data quality. For the collected heterogeneous and multidimensional data, three major operations are carried out in sequence: data cleaning, missing value processing, and outlier removal: using deduplication algorithms to delete duplicate data and avoid the interference of data redundancy on model training; By using mean imputation, median imputation, and interpolation methods, targeted processing of missing values for different types of data is carried out to ensure data integrity; Using the box plot method to identify abnormal transactions, clicks, and other abnormal data, combined with business logic to determine whether to exclude them, in order to avoid misleading model training with abnormal data. Feature engineering is the core of mining the potential value of data and improving model performance. Based on consumer behavior theory and big data mining logic, a multidimensional feature system is constructed, including user features, product features, behavior features, and temporal features. User features focus on consumption ability and preferences, while product features focus on category and popularity. Behavioral features mine user behavior associations, and temporal features capture the temporal patterns of consumption behavior. At the same time, through feature encoding, normalization processing, and cross feature construction, unstructured and semi-structured data are transformed into structured features to reduce algorithm computational complexity. Finally, the processed dataset is divided into a training set, a validation set, and a testing set in a ratio of 7:2:1. The training set is used for model parameter training, the validation set is used for hyperparameter optimization and model adjustment, and the testing set is used for final model performance validation to ensure the rationality and scientificity of the dataset division.

3.2 Prediction model construction

The core goal of constructing a predictive model is to accurately capture consumer behavior patterns and improve prediction accuracy. Combined with the machine learning algorithms mentioned earlier, a three-level model architecture of "benchmark model+ensemble model+optimization improvement" is designed, which balances the interpretability and predictive performance of the model, and adapts to the high-dimensional and nonlinear characteristics of consumer behavior big data. Firstly, establish a benchmark model and select logistic regression and decision tree as the basic classification models. Logistic regression, with its advantages of high computational efficiency and strong interpretability, is used to preliminarily capture the linear correlation of consumer behavior and complete the basic prediction task; Decision trees can effectively mine nonlinear correlations between features and are suitable for processing high-dimensional behavioral features, but they suffer from overfitting and insufficient generalization ability. They serve as a benchmark model for performance comparison with subsequent ensemble models.

On the basis of the benchmark model, an ensemble learning model is built as the core prediction model, and three gradient boosting algorithms, XGBoost, LightGBM, and CatBoost, are selected. These algorithms effectively alleviate overfitting problems and accurately capture complex correlations and potential patterns in consumer behavior data through multi weak classifier ensemble and gradient descent optimization strategies. For three types of integrated models, hyperparameter optimization is carried out separately, using a hybrid optimization strategy of grid search combined with Bayesian optimization to optimize core hyperparameters such as learning rate, tree depth, and number of leaf nodes, and determine

ISSN 2616-5775 Vol. 9, Issue 2: 39-45, DOI: 10.25236/AJCIS.2026.090206

the optimal hyperparameter combination for each model to improve model prediction accuracy and generalization ability. In order to further optimize the model performance, a model fusion strategy is constructed. The prediction results of the three integrated models are weighted and fused, and the weight is allocated based on the performance of each model in the verification set. The prediction advantages of different models are taken into account to remedy the limitations of a single model. At the same time, in response to the temporal characteristics of consumer behavior, the concept of temporal modeling is introduced to improve the model and enhance its ability to capture the temporal patterns of consumer behavior. Ultimately, an accurate prediction model adapted to consumer behavior big data is formed to achieve precise prediction of consumer purchasing behavior, demand preferences, and churn risk, as shown in Figure 1.

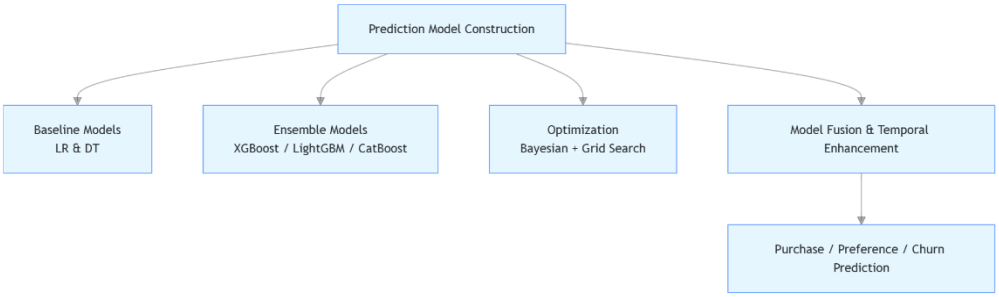

Figure 1. Framework of prediction model construction

4. Experimental verification and application analysis

4.1 Experimental design

The core objective of the experimental design is to accurately test the predictive performance of the model and ensure the reproducibility of the experiment, clarifying the four core contents of the experimental environment, parameter settings, data preparation, and experimental process. The experimental environment uses Windows 10 operating system, Intel Core i7-12700H CPU, 32GB memory, NVIDIA RTX 3060 graphics card, Python 3.9 software development environment, and relies on open source libraries such as Scikit learn, XGBoost, LightGBM to achieve model training and prediction. Matplotlib and Seaborn libraries are used to visualize the experimental results.

The experimental parameter settings are based on the previous hyperparameter optimization results, and the core parameters of each model are determined as follows: logistic regression learning rate of 0.1, maximum iteration times of 1000; The maximum depth of the decision tree is 8, and the minimum number of sample partitions is 2; The learning rates of XGBoost, LightGBM, and CatBoost are all set to 0.05, with tree depths of 6, 8, and 7, and leaf nodes of 32, 64, and 64, respectively. The remaining parameters are set to default values. The experimental data is selected from the consumer behavior dataset constructed in Section 3.1, using a 7:2:1 division ratio. The training set is used for model parameter fitting, the validation set is used for real-time parameter adjustment to avoid overfitting, and the test set is used for final performance testing. The core tasks of the experiment are consumer purchasing behavior prediction and user churn risk prediction, ensuring that the experimental tasks are in line with actual consumption scenarios. The experimental process is as follows: data loading and preprocessing verification, benchmark model and integrated model training, model prediction, performance indicator calculation, result visualization, recording experimental data and parameters throughout the process to ensure the reproducibility of the experiment, as shown in Figure 2.

ISSN 2616-5775 Vol. 9, Issue 2: 39-45, DOI: 10.25236/AJCIS.2026.090206

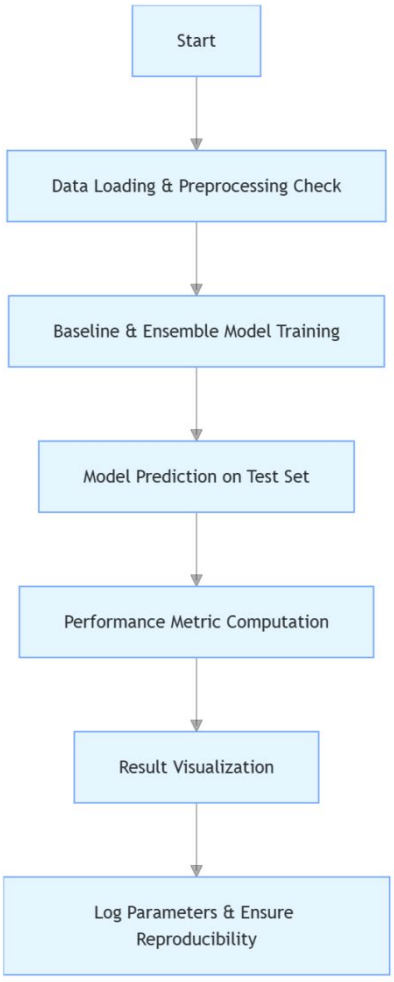

Figure 2. Experimental design flowchart

4.2 Result analysis

The analysis of experimental results is carried out from two dimensions: model performance comparison and feature importance interpretation. The multi-dimensional evaluation indicators specified in the previous text are used to comprehensively and objectively test the predictive performance of the model. The comparison results of model performance show that the overall performance of the ensemble learning model is significantly better than the benchmark model: the AUC values of logistic regression and decision tree are 0.72 and 0.78, respectively, while the AUC values of XGBoost, LightGBM, and CatBoost reach 0.89, 0.91, and 0.92, respectively, and the F1 scores are 0.87, 0.89, and 0.90, respectively, all significantly higher than the benchmark model, indicating that the ensemble learning model can more accurately capture the nonlinear correlations of consumer behavior and has stronger generalization ability; The AUC value of the integrated model after model fusion is increased to 0.93, and the F1 score is 0.91, which further verifies the effectiveness of the model fusion strategy, and can give full play to the advantages of each single model to make up for the limitations of the single model.

The analysis of feature importance shows that the core features affecting the prediction results are purchase frequency, browse add purchase conversion rate, and the most recent purchase time in consumer behavior characteristics, with weights accounting for 18.6%, 16.3%, and 14.8%, respectively; The consumption hierarchy and regional characteristics in the user profile, as well as the category popularity in the product features, also have a strong influence, with a weight ratio of over 10%. This result is highly consistent with consumer behavior theory, confirming the rationality of the feature engineering mentioned earlier. At the same time, it provides clear guidance for enterprises to accurately capture

ISSN 2616-5775 Vol. 9, Issue 2: 39-45, DOI: 10.25236/AJCIS.2026.090206

consumer demand and optimize data collection priorities. In addition, the experiment found that the accuracy of the model slightly decreased in predicting low-frequency consumer users, mainly due to the sparse behavior data of such users, which pointed out the direction for future model improvement.

4.3 Model application

Based on the actual business scenarios of retail, e-commerce and other industries, clarify the core application path of the model, focus on three core scenarios: precision marketing, user churn warning, and product recommendation, and propose targeted application strategies. In precision marketing scenarios, based on model prediction results, consumers are divided into three categories: high purchase intention, medium purchase intention, and low purchase intention. Differentiated marketing plans are developed for different levels of users. High intention users push personalized coupons and limited time activities, medium intention users push product details and usage evaluations, and low intention users push interest related product information to improve marketing conversion rates and reduce marketing costs.

In the scenario of user churn warning, a model is used to predict potential churn users, analyze their behavioral characteristics and reasons for churn, and provide support for enterprises to formulate retention strategies. For example, for users who consume frequently but have recently decreased activity, exclusive welfare and care information is pushed to maintain user stickiness. In the context of product recommendation, combining the predicted user needs, preferences, and behavioral patterns of the model, a personalized recommendation system is constructed to achieve "matching of people and goods" and improve product exposure and purchase conversion rates. At the same time, the implementation steps and precautions for the application of the model should be clarified, and suggestions for real-time data updates and dynamic optimization of the model should be proposed. This ensures that the model can adapt to the dynamic changes in consumer behavior and exert its practical value in the long run.

5. Conclusion

This article focuses on the systematic research of machine learning based consumer behavior big data mining and accurate prediction model construction. Based on the actual needs of enterprise precision marketing in the context of the digital economy, it integrates consumer behavior theory, big data mining technology, and machine learning algorithms to complete core research tasks such as big data system construction, prediction model construction, experimental verification, and application analysis. Finally, the following conclusions are formed.

This article successfully constructs a big data processing system that adapts to consumer behavior characteristics. Through multi scenario data fusion collection, standardized preprocessing, and targeted feature engineering, it solves the pain points of consumer behavior data heterogeneity and uneven quality; Design a three-level architecture of "benchmark model+integrated model+fusion optimization", and experimentally verify that the AUC value of the fused model reaches 0.93, which is significantly better than a single benchmark model and an independent integrated model, and can accurately predict consumer purchasing behavior and churn risk; At the same time, the core behavioral characteristics that affect prediction performance have been identified, and the proposed model application strategy can provide effective support for precision marketing in e-commerce and retail enterprises, achieving an organic combination of theoretical research and practical application.

Subsequent research can expand the sources of multi scenario data, integrate online and offline consumption data, and enhance the comprehensiveness and adaptability of the model; Introducing temporal deep learning algorithms to optimize the model's ability to capture dynamic consumer behavior; At the same time, privacy computing technology can be combined to deepen feature mining and model optimization while ensuring data security, further improving the interpretability and business implementation of the model, and promoting the deep application of machine learning technology in the field of consumer behavior analysis.

References

[1] Han H R, Jiao H F, Wang R R, Dai L Y. Progress and Prospects of the Research on Shopping Behavior of Urban Residents [J]. Progress in Geography,2011, (8):1006-1013.
[2] Zhang F, Zhang L N, Li J J. Prediction of User Consumption Behavior Based on Machine Learning

ISSN 2616-5775 Vol. 9, Issue 2: 39-45, DOI: 10.25236/AJCIS.2026.090206

Combination Model [J]. Technology of Io T& AI, 2022, (2):19-27.
[3] Qin G L. The Impact and Mechanism of Artificial Intelligence Autonomy on User Perception and Consumption Behavior—Method Based on Meta-Analysis [J]. E-Commerce Letters, 2025, (1):3806-3812.