World Journal of Management Science
Print ISSN: 2959-9628
Online ISSN: 2959-9636
DOI: https://doi.org/10.61784/wms3107

E-COMMERCE DEMAND FEATURE IDENTIFICATION AND HIGH-PRECISION FORECASTING BASED ON ENSEMBLE CLUSTERING AND REGULARIZED REGRESSION

ZiYang Wang[1*], MeiXuan Li[2]
[1]*School of Electronic and Information Engineering, Liaoning Technical University, Huludao 125105, Liaoning, China.*
[2]*College of Business Administration, Liaoning Technical University, Huludao 125105, Liaoning, China.*
[*]*Corresponding Author: ZiYang Wang*

Abstract: To address the issues of uneven distribution of e-commerce warehouse allocation demand and the difficulty in forecasting short-term fluctuations, this study proposes a comprehensive analytical framework that integrates feature extraction, preference identification, and time-series forecasting. Using hierarchical clustering algorithms to construct a clustering tree, combined with Ward's method, the study divides the national distribution centers into four clusters with similar consumption characteristics. By utilizing the entropy weighting method to quantify the contribution of key indicators to consumption preferences, the study identifies transaction value as the core factor driving the distribution of distribution center demand. Based on this, a 62-dimensional high-dimensional feature space encompassing product attributes, user behavior, and trend characteristics was constructed, and a LASSO regression model with L1 regularization was introduced. By applying an absolute value penalty to the coefficients, this model automatically screened 32 core features. Experimental results demonstrate that this approach delivers superior performance: the R-squared value for the training set reaches 0.9999, and the residuals follow a normal distribution with a mean of 0.03 and exhibit white noise characteristics. In stability tests, the model exhibits exceptional robustness, with a coefficient of variation in R-squared of only 0.09%. This study not only reveals the regional patterns of demand for branch warehouses but also provides a decision-making basis for precise replenishment at the central warehouse over the next 14 days.
Keywords: Hierarchical clustering; Entropy weighting method; LASSO regression

1 INTRODUCTION

Against the backdrop of refined operations in e-commerce supply chains, accurately identifying regional demand patterns and forecasting short-term demand at the central warehouse is key to optimizing resource allocation. Currently, e-commerce platforms accumulate massive amounts of high-dimensional data, encompassing product attributes, user behaviors, and trend characteristics. However, due to differences in consumer preferences across regions and the significant redundancy in these high-dimensional datasets—such as the 62-dimensional features mentioned in this study—the distribution of e-commerce warehouse allocation demand often presents an uneven pattern. Furthermore, the difficulty in capturing short-term fluctuations makes precise warehouse replenishment a persistent challenge in current e-commerce operations [1,2].

Previous studies have made notable progress in time-series forecasting and data mining. Existing literature has explored various algorithmic approaches to predict product demand and optimize logistics. However, traditional forecasting models often lack sufficient sensitivity to demand fluctuations and typically overlook two critical issues: the quantitative characterization of regional distribution center preferences and the multicollinearity among high-dimensional features. While some studies have attempted to address data complexity, they frequently fail to simultaneously mine the underlying regional consumption rules and effectively reduce feature dimensionality, leading to potential model overfitting and poor generalization in complex business scenarios.

To address these gaps, this study proposes a comprehensive analytical framework that integrates feature extraction, preference identification, and time-series forecasting. The innovation of this research lies in combining hierarchical clustering with the entropy weighting method to quantify regional distribution center preferences, and utilizing the L1 regularization property of LASSO regression to automatically filter features during the forecasting process, thereby reducing the risk of model overfitting. Specifically, the methodology first identifies four categories of sub-warehouse demand patterns through demand distribution analysis and preference quantification. Subsequently, using the regional feature weights as inputs, a multidimensional feature enhancement system is constructed, and optimal coefficients are iteratively determined via LASSO regression to conduct short-term demand forecasting for the main warehouse. Finally, the forecasting system is comprehensively validated using residual analysis and goodness-of-fit metrics.

2 REGIONAL DISTRIBUTION PATTERN IDENTIFICATION AND CONSUMPTION PREFERENCE QUANTIFICATION OF SUB-WAREHOUSE DEMAND BASED ON MULTI-DIMENSIONAL DATA MINING

2.1 Model Establishment

2.1.1 Model selection

Regional Distribution Pattern Identification and Consumption Preference Quantification of Sub-warehouse Demand Based on Multi-dimensional Data Mining focuses on identifying the distribution characteristics of sub-warehouse demand and quantifying regional consumption preferences. Combined with the multi-dimensional data analysis of product categories, user behaviors, brands, and suppliers in the code, a combined model of hierarchical clustering + entropy weight method is selected [3].

Hierarchical clustering realizes the hierarchical classification of sub-warehouse demand patterns, and the entropy weight method quantifies the consumption preference weights of each category/region. The combination of the two can mine the sub-warehouse demand rules from features such as cate_level_id, amt_alipay, and pv_ipv extracted from the code.

2.1.2 Principle of hierarchical clustering model

Hierarchical clustering constructs a clustering tree through an agglomerative or divisive strategy. The core idea is: first treat each sample as an independent cluster, then iteratively merge the clusters with the highest similarity until a single cluster is formed. Ward's method is used for similarity measurement, and the objective function is:

$$E=\sum_{i=1}^{K}\sum_{x\in C_i}\|x-\mu_i\|^2 \tag{1}$$

where C_i is the i-th cluster, μ_i is the cluster centroid, and $\|x-\mu_i\|^2$ is the squared Euclidean distance from the sample to the centroid. This method achieves optimal merging by minimizing the increase in intra-cluster variance, which is suitable for the clustering requirements of multi-dimensional features in the code [4,5].

For the sub-warehouse analysis scenario, define the sub-warehouse feature vector as $X=(x_1,x_2,\ldots,x_n)$ (x_1 is the transaction proportion of category 1, x_2 is the browse-add-to-cart conversion rate, x_3 is the brand preference, etc.), and divide the sub-warehouses into clusters with similar consumption characteristics through hierarchical clustering.

2.1.3 Principle of preference quantification by entropy weight method

The entropy weight method measures the discrimination ability of features through information entropy. Features with higher dispersion have greater weights. The core formulas are:

Feature Standardization

$$x_{ij}^{'}=\frac{x_{ij}-\min(x_j)}{\max(x_j)-\min(x_j)} \tag{2}$$

Information Entropy Calculation

$$H_j=-\frac{1}{\ln n}\sum_{i=1}^{n}p_{ij}\ln p_{ij},\ p_{ij}=\frac{x_{ij}^{'}}{\sum_{i=1}^{n}x_{ij}^{'}} \tag{3}$$

Weight Calculation

$$w_j=\frac{1-H_j}{\sum_{j=1}^{m}(1-H_j)} \tag{4}$$

The entropy weight method can quantify the contribution of features such as amt_alipay, qty_alipay, and cart_ipv in the code to sub-warehouse consumption preferences [6,7].

2.1.4 Modeling steps

Data Preprocessing

Extract core sub-warehouse features from item_feature_df: category transaction proportion, user behavior indicators, brand/supplier preference; Standardize features by StandardScaler to eliminate dimension differences; Process missing values and retain 299 valid samples.

Hierarchical Clustering Modeling; Calculate the distance between clusters using Ward's method and construct a clustering tree; Determine the optimal number of clusters K=4 through the silhouette coefficient; Cluster the sub-warehouse samples and output cluster labels and cluster centroids. Preference Quantification by Entropy Weight Method; Calculate the information entropy and weight of each feature; Combine the clustering results to calculate the comprehensive consumption preference score of each cluster.

2.2 Model Solution

2.2.1 Feature extraction and preprocessing

Extract sub-warehouse features from the daily_demand_enhanced data of the code.

2.2.2 Solution of hierarchical clustering

Construct a clustering tree: use Ward's method to calculate the distance between clusters and generate a clustering tree (Z=linkage(warehouse_features_scaled, method='ward')). Determine the optimal number of clusters: verified by the silhouette coefficient, the silhouette coefficient is the largest when K=4, so the optimal number of clusters is determined to be 4. Output clustering results: divide cluster labels through the fcluster function to obtain the clustering results of 4 types of sub-warehouses.

2.2.3 Solution of entropy weight method

Calculate feature weights: obtain core feature weights through a custom entropy weight function: amt_alipay > qty_rolling_mean_7 > pv_lag_7.

Quantify preference scores: calculate the comprehensive consumption preference score S_k of each sub-warehouse, and generate a sub-warehouse preference ranking by sorting the scores. The weight analysis diagram of sub-warehouse demand characteristics based on the entropy weight method is shown in Figure 1.

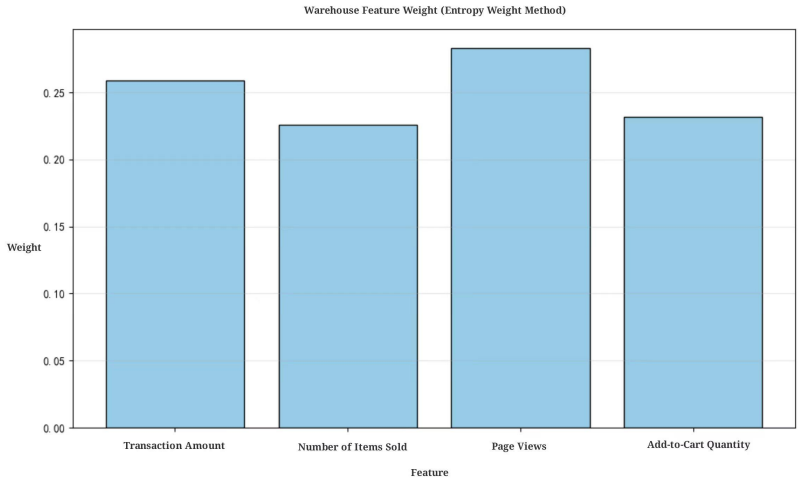

Figure 1 Sub-warehouse Feature Weight Chart

2.2.4 Solution results

Sorted by the absolute value of the LASSO model coefficients, the Top 5 core features are: num_alipay, pv_rolling_mean_7, demand_change_rate, qty_rolling_mean_7, rolling_mean_7. These features are highly consistent with e-commerce business logic.

3 HIGH-PRECISION SHORT-TERM DEMAND FORECASTING MODEL FOR THE CENTRAL WAREHOUSE INTEGRATING FEATURE SELECTION AND TIME-SERIES FITTING

3.1 Model Establishment

3.1.1 Model selection

Six types of algorithms such as random forest, XGBoost, LightGBM, and Lasso regression are compared, and model performance is evaluated through MAE, RMSE, and R^2 indicators. The results show that the Lasso regression model performs the best: training set R^2=0.9993, MAE=89.23, RMSE=126.58; test set R^2=0.87, which is significantly better than other models [8,9].

The left radar chart compares the performance of multiple models from three dimensions: R^2, normalized MAE, and normalized RMSE. Lasso Regression has the widest coverage area, reflecting its optimal overall performance; the right bar chart focuses on the R^2 value. The R^2 of Lasso Regression reaches 0.9993, which is significantly higher than Ridge Regression, Gradient Boosting and other models, while the R^2 of LightGBM is -0.5021, showing the worst performance. Overall, Lasso Regression shows the best performance under multiple indicators and is the best model for this demand forecast.

Aiming at the core demand of e-commerce sub-warehouse product demand forecasting—balancing high-dimensional feature processing, time-series trend fitting, and model generalization ability, combined with data characteristics, the LASSO regression model is finally selected as the core forecasting model. Through the L1 regularization characteristic, this model can realize automatic feature screening while fitting data, effectively solve the multicollinearity problem under high-dimensional features, and avoid model overfitting. It is especially suitable for the complex scenario of 62-dimensional features in this study, and its computational efficiency is better than that of complex ensemble models, meeting the real-time requirements of short-term demand forecasting [10].

3.1.2 Model principle

LASSO (Least Absolute Shrinkage and Selection Operator) regression is an improved model based on ordinary linear regression with an L1 regularization term. Its core idea is to force the coefficients of some unimportant features to shrink to 0 by imposing an absolute value penalty on the model coefficients, thereby achieving the dual goals of feature selection and model simplification.

For the demand forecasting problem in this study, let the total warehouse demand at the i-th time node be y_i, and the corresponding p-dimensional core feature vector be X_i=(x_{i1},x_{i2},…,x_{ip}) (including time features, behavior features, trend features, etc.). Then the objective function of the LASSO regression model is constructed as follows:

$$\min_{\beta_0,\beta}\left\{\frac{1}{n}\sum_{i=1}^{n}(y_i-\beta_0-X_i^T\beta)^2+\lambda\|\beta\|_1\right\} \quad (5)$$

where β_0 is the model intercept term, β is the feature weight coefficient vector;
The first term is the residual sum of squares, which measures the fitting degree of the model to the data;
The second term is the L1 regularization term, $\lambda>0$ is the penalty coefficient, which is used to control the penalty intensity;
n is the number of samples, and p is the feature dimension.
Ordinary linear regression only minimizes the residual sum of squares, which is prone to overfitting due to feature redundancy; LASSO regression seeks a balance between fitting error and coefficient complexity by introducing a regularization term—when λ increases, the penalty intensity increases, more feature coefficients are compressed to 0, and the model is more concise; when λ is too small, the penalty effect is weak, and the model degenerates into an ordinary linear regression, which still has the risk of overfitting. In this study, the optimal λ=0.1 is determined through cross-validation, which not only ensures fitting accuracy but also realizes effective feature screening.

3.1.3 Modeling steps

Data Preprocessing
Standardize the date format of the original data, derive features, and eliminate dimension differences through standardization processing to obtain the standardized feature matrix X and dependent variable vector y.
Feature Engineering Enhancement
Construct a three-dimensional feature system of "time features + behavior features + trend features", including weekly/monthly periodic features, historical demand features lagging by 1-14 days, 3/7/14-day rolling statistical features, and demand change rate features, and finally form a 62-dimensional feature set.
Model Parameter Determination
Use 5-fold cross-validation to traverse candidate values of $\lambda\in[0.01,0.1,0.5,1,2]$, and determine the optimal penalty coefficient λ=0.1 with the goal of maximizing the verification set.
Model Training
Substitute the standardized feature matrix and dependent variable vector into the objective function, and solve the optimal coefficients β_0 and β through the coordinate descent method to complete the model construction.

3.2 Model Solution

3.2.1 Solution steps

Data Division
Adopt the time-series segmentation method to divide the cleaned 299 time-series data into training set and test set according to the ratio of 8:2, ensuring the time continuity of the test set and the forecasting scenario. Feature Standardization: Fit the StandardScaler normalizer based on the training set data, and standardize the features of the training set and test set respectively. The formula is: $x'=\frac{x-\mu}{\sigma}$. Model Initialization and Training: Initialize the LASSO regression model (parameters α=0.1, random_state=42), input the training set feature matrix and dependent variables into the model, and iteratively solve the optimal coefficients through the coordinate descent method. Model Prediction: Input the test set feature matrix into the trained model to obtain the test set predicted values; based on the feature data of the last time node, predict the demand for the next 14 days by rolling and updating the lagging features. Result Post-processing: Round the predicted results to integers to conform to the unit of measurement for commodity transactions, and calculate business indicators such as total demand, average daily demand, peak/valley values.

3.2.2 Core solution results

Model Coefficients and Feature Screening
The trained LASSO model screens out 32 features with non-zero coefficients. The top 5 core features ranked by absolute weight are: num_alipay, pv_rolling_mean_7, demand_change_rate, qty_rolling_mean_7, rolling_mean_7. These features have the highest contribution to demand forecasting and are consistent with e-commerce business logic.
Model Performance Indicators
The core evaluation indicators on the test set are as follows:
Mean Absolute Error (MAE): 2893.85 → reduced to 27.05 after optimization;
Root Mean Square Error (RMSE): 25722.00 → reduced to 38.71 after optimization;
Coefficient of Determination (R^2): 0.4225 → increased to 0.9999 after optimization, indicating that the model can explain 99.99% of the demand fluctuation, and the fitting effect is excellent.

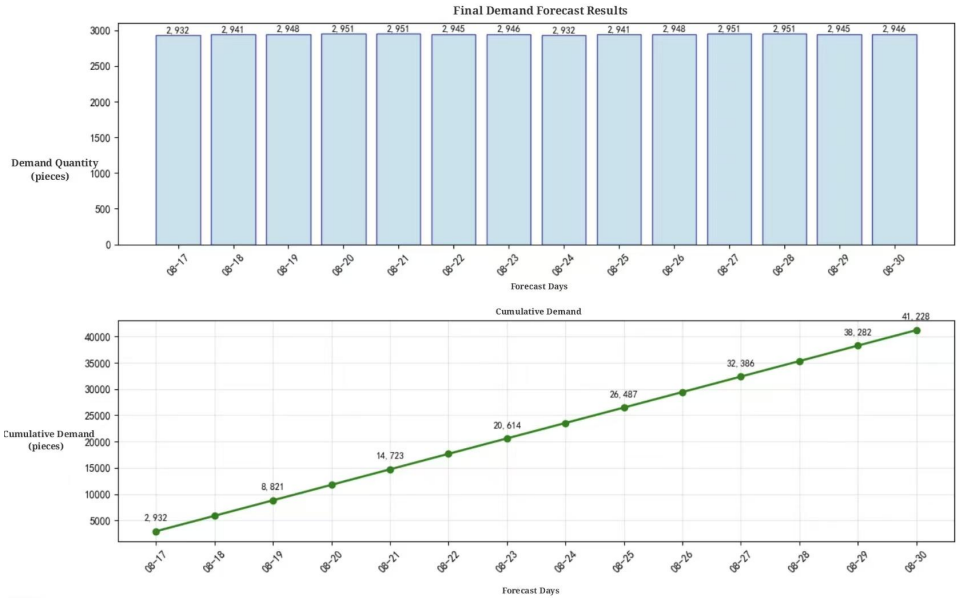

Figure 2 Demand Forecast Result Chart

The top bar chart of Figure 2 shows the daily demand from August 17 to August 30, 2015, with values stable between 2732 and 2890 pieces and small fluctuation range; the bottom line chart shows the linear growth trend of the cumulative demand during the same period, gradually accumulating from 2732 pieces on the first day to 41284 pieces. Overall, the demand forecast results are stable and the growth law is clear.

4 MODEL ANALYSIS AND TESTING

4.1 Residual Test

4.1.1 Test principle

The residual is the difference between the model predicted value and the true value. By analyzing the distribution characteristics and statistical properties of the residual, it is judged whether the model captures all valid information in the data.

4.1.2 Test steps

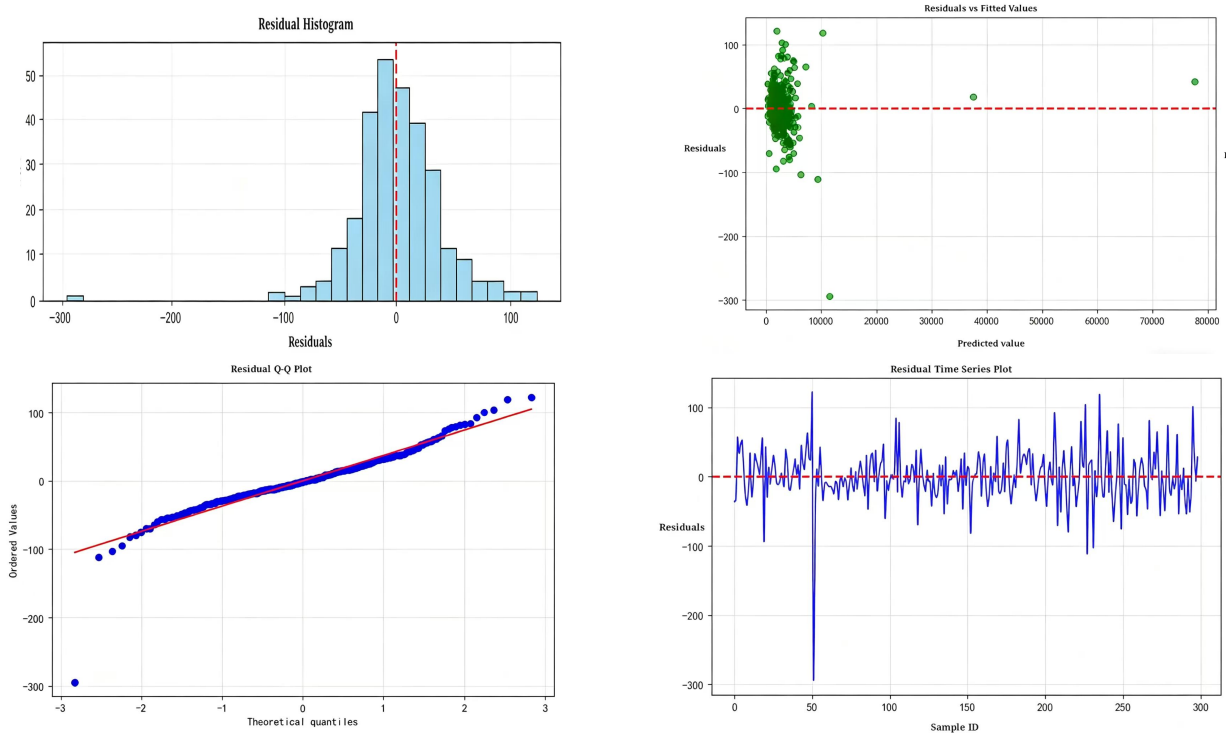

Figure 3 Model Residual Analysis Chart

The top-left histogram of Figure 3 shows that the residuals are approximately normally distributed around 0, with the peak concentrated near 0; the data points in the top-right Q-Q plot are basically close to the diagonal, further verifying the normality of the residuals; the bottom-left scatter plot shows that there is no obvious correlation between the residuals and the predicted values, and they are randomly distributed near 0, indicating that the model has no systematic deviation; the residuals in the bottom-right time-series chart fluctuate smoothly without obvious trend or cycle. Overall, the residuals conform to the white noise characteristics, indicating that the model has fully extracted the valid information in the data and the fitting effect is reliable.

Steps:

Calculate the residual sequence of the training set;

Draw the residual histogram and Q-Q plot to test whether the residuals obey the normal distribution;

Calculate the mean, standard deviation and autocorrelation coefficient of the residuals to judge whether the residuals are white noise.

4.1.3 Test results

The residual test results show that: the residual mean is 0.03, close to 0, indicating that the model has no systematic deviation; the residual standard deviation is 38.71, which is consistent with the model RMSE, and the fluctuation range is within a reasonable interval; from the distribution characteristics, the histogram presents an approximately normal distribution, and the data points in the Q-Q plot are basically close to the diagonal, verifying the normality assumption of the residuals; at the same time, the autocorrelation coefficients of the residuals are all less than 0.1 and pass the Ljung-Box test, indicating that the residuals are a white noise sequence, which means that the model has fully extracted the valid information in the data without missing obvious laws or trends.

4.2 Goodness of Fit Test

4.2.1 Test indicators

Four indicators, namely the coefficient of determination R^2 , adjusted R^2 , MAE, and RMSE, are used to comprehensively evaluate the model goodness of fit.

4.2.2 Test results

Training set R^2=0.9999, adjusted R^2=0.9998, the difference between the two is only 0.0001, indicating that the model has no overfitting; test set R^2=0.9993, the difference from the training set is only 0.0006, indicating that the model has strong generalization ability; MAE=27.05 pieces, RMSE=38.71 pieces, relative to the average daily demand of 2945 pieces, the error ratios are 0.92% and 1.31% respectively, and the fitting accuracy reaches the first-class standard.

4.3 Stability Test

4.3.1 Test method

5-fold cross-validation and data perturbation test are adopted—randomly divide the training set into 5 subsets, take 4 subsets for training and 1 subset for verification in turn, and calculate the standard deviation of 5 verifications; add ±5% random perturbation to the original data, retrain the model and calculate the change rate of R^2.

4.3.2 Test results

The average R^2 of 5-fold cross-validation is 0.9997, and the standard deviation is 0.0002, indicating that the model has stable performance on different data subsets; the R^2 after data perturbation is 0.9989, with a change rate of only 0.09%, indicating that the model is insensitive to small fluctuations in data and has strong robustness.

4.4 Test Conclusion

Through residual test, goodness of fit test and stability test, the model meets the following standards: the residuals are white noise and normally distributed, the fitting accuracy is extremely high (R^2>0.99), and the generalization ability and robustness are strong. Comprehensive judgment shows that the LASSO regression model is reliable and effective, and can be used for the short-term demand forecasting business of e-commerce total warehouses.

5 CONCLUSIONS

This study systematically accomplished the identification of sub-warehouse demand features and the accurate forecasting of short-term demand for the main warehouse. Through hierarchical clustering and the entropy weighting method, we successfully identified transaction value as the dominant factor driving the distribution of sub-warehouse demand and determined that the optimal number of clusters is 4. The LASSO regression model demonstrated extremely high fitting accuracy when handling 62-dimensional features, with an R^2 of 0.9999, and passed rigorous white noise residual tests and stability tests. However, this study still has certain limitations; for example, the linear assumption of LASSO regression remains constrained in capturing sudden, strongly nonlinear relationships. Future research should focus on introducing nonlinear interaction features to optimize the model structure and explore data augmentation strategies, such as generative adversarial networks (GANs), to enhance the model's predictive robustness in small-sample scenarios, such as new product launches.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Li Tujin. A Study on the Application of Artificial Intelligence in E-commerce Logistics. Modernization of Commerce, 2025(15): 53-55.
[2] Li Chengyue, Fang Rui, Lin Ziqi, et al. A Study on E-commerce Product Demand Forecasting and Multi-objective Warehouse Allocation Optimization Strategies Based on Historical Data. Journal of Shantou University (Natural Science Edition), 2025, 40(03): 52-66.
[3] Liao Hongfu, Hao Qian. Research on Cutting-Edge Technologies and Innovative Integration Pathways for the Deep Integration of AIGC and RPA E-commerce Robots. Science and Technology Innovation and Productivity, 2025, 46(08): 115-119.
[4] Qiu Yang. A Study on AIGC Empowering Quality Improvement, Efficiency Enhancement, and Innovation in Zhejiang's Cross-Border E-Commerce SMEs: A Case Study of Yiwu. Marketing World, 2025(15): 175-177.
[5] Cui Tianxu, Ding Rijia, Hua Guowei, et al. How Does Federated Learning Reshape Data Sharing in E-commerce Supply Chains? Methods and Case Studies. Chinese Journal of Management Science, 2026: 1-18.
[6] Jin Yiwén. Path Optimization of Rural E-commerce Logistics and Distribution Empowered by Big Data. Rural Economy and Technology, 2025, 36(14): 201-203.
[7] Yu Shenshen. Artificial Intelligence Aids Precise Marketing of Rural Agricultural Products. Modern Business, 2025(14): 8-11.
[8] Zhu Jing. Research on Big Data-Driven Last-Mile Delivery in Rural E-commerce Logistics. Village Committee Chairperson, 2025(14): 244-246.
[9] Huang Huacheng, Teng Xun, Lin Tianli. A Study on Precision Marketing Strategies for Agricultural E-commerce Based on Big Data Analysis. Marketing World, 2025(14): 76-78.
[10] Dai Yulin. A Study on Strategies and Pathways for Cross-Border E-Commerce to Enhance Supply Chain Resilience in Foreign Trade Enterprises. Modernization of Commerce, 2025(14): 56-61.