Vegetable commodity pricing and replenishment strategy based on GA-XGBOOST algorithm

Lijiale Yang*, Chenglong Huang, Shilei Wu

Combat Support Academy, Rocket Force University of Engineering, Shaanxi, China, 710000

*Corresponding author: 18179302335@163.com

Abstract. In this paper, for the problem of vegetable commodity replenishment, sales volume and cost-plus pricing, firstly, we analyzed the distribution law and correlation relationship of sales volume of different categories and individual products of vegetables, then we used ARIMA time series model and wavelet analysis method to solve to get the relationship between sales volume change and time, and found that sales of flower and leaf category fluctuated more stably with the time series, and showed cyclic changes in the first two years; Then K-means clustering algorithm and Spearman correlation coefficient were used to derive the correlation of sales volume between categories and individual products. Through the K-means model, all individual products are divided into four categories, and the more groups, the greater the correlation. In this paper, support vector machine, random forest, linear regression, GA-XGBoost and other methods were used to establish the pricing and total sales volume regression model respectively, and GA-XGBoost was selected as the optimal model by combining the evaluation indexes. Finally, using the GA-XGBoost algorithm based on the sliding window tree, the time series data is converted into regression data to find the optimal replenishment and pricing prediction value of six categories of commodities for one week, and get the optimal value of daily revenue of the superstore. For example, if the sales unit price of the chili pepper category is set at RMB 20.642, the predicted sales quantity and superstore revenue will be 386.918kg and RMB 5296.856, respectively.

Keywords: Pricing of Vegetable Commodities, K-means Clustering, Genetic Algorithm, XGBoost.

1. Introduction

Fresh supermarket refers to the sale of "fresh three products" based supermarkets, that is, fruits and vegetables, meat, aquatic products and other products with freshness. The freshness period of vegetable commodities is particularly short, and its appearance with the increase in selling time and decline, and is very susceptible to the influence of the external environment, resulting in a significant increase in the rate of loss. In addition to the distribution of supermarkets need to ensure a complete, professional cold chain technology to support, as well as in a certain supermarket sales space to construct a reasonable sales mix, but also in the demand for vegetables inaccurate and price fluctuations under the information gap in the development of optimal pricing and replenishment decision-making program, so as to achieve the maximum benefit.

In recent years, many scholars have begun to focus on the replenishment and pricing of fresh food under stochastic demand conditions, Jiang Cherrymei et al[1] proposed a joint decision-making model based on the perceived freshness function of commodities and the quantity loss function of raw materials, and established a joint decision-making model for inventory and pricing of the second level of the cold chain, where the perceived freshness and the price affect the demand together. Xu Xiaofeng et al[2] proposed a combined ARIMA-NARX forecasting model for fresh commodities sales. Firstly, the ARIMA model is used to describe the linear law in the sales volume time series, after that, NARX is used to capture the nonlinear relationship in the ARIMA residuals, and the NARX residuals prediction results are used to correct the ARIMA prediction value. Yang Shuai et al[3] proposed a profit model that incorporates the quality change of fresh food into the retail chain, and considers the pricing, replenishment strategy and shelf allocation strategy of fresh food in a collaborative manner, and finds the optimal pricing optimal solution through search algorithm. Pan Xiaofei et al[4] proposed a method to consider the market demand of fresh food as an additive function of retail price and superstore's freshness effort level and stochastic factors, and constructed a risk-neutral expected profit model of fresh food superstore, and then planned and solved the optimal decision of superstore's freshness effort level, order quantity and retail price. Bi et al[5] proposed a joint inventory control and dynamic pricing method for fresh products based on deep reinforcement learning method, modeled the problem by combining the characteristics of fresh products, then designed the joint inventory control and dynamic pricing algorithms for fresh products based on deep reinforcement learning, and then obtained the joint inventory control and dynamic pricing model based on the DQN method through simulation experiments.

In summary, many literatures have studied the fresh commodities pricing problem and replenishment strategy problem, but fewer scholars have used reinforcement learning based algorithms to study the commodities pricing and replenishment problem, so this paper proposes a commodities replenishment and pricing strategy model based on the improved XGBoost algorithm, and optimizes the XGBoost model based on the genetic algorithm to compare with several other models, and its accuracy is higher than the other models. Its accuracy is higher than other model data, thus confirming the effectiveness of the model in the commodity replenishment and pricing strategy problems.

2. Data analysis of vegetable commodity sales

Data source: http://www.mcm.edu.cn/, the dataset contains commodity information of 6 vegetable categories and the sales flow details and wholesale price of each commodity of a superstore from July 1, 2020 to June 30, 2023, part of the data is shown in Table 1 below.

Table 1 Sales flow details and wholesale prices for each product

	Item Name	code	name	time	Sales (kg)	price tag (RMB/kg)	Discounts or sales
0	beefsteak lettuce	1011010101	philodendron	2020-10-30	0.743	5.0	sales
1	beefsteak lettuce	1011010101	philodendron	2020-10-30	0.502	5.0	sales
2	beefsteak lettuce	1011010101	philodendron	2020-10-30	0.301	5.0	sales
3	beefsteak lettuce	1011010101	philodendron	2020-10-30	0.382	5.0	sales
4	beefsteak lettuce	1011010101	philodendron	2020-10-30	0.261	5.0	sales
…	…	…	…	…	…	…	…
878503	Wofeng Sunshine Seafood Mushroom	1011010801	edible mushroom	2022-07-28	1.000	1.5	Discounts
878504	Wofeng Sunshine Seafood Mushroom	1011010801	edible mushroom	2022-07-28	1.000	2.5	sales
878505	Wofeng Sunshine Seafood Mushroom	1011010801	edible mushroom	2022-07-28	1.000	1.5	Discounts
878506	Wofeng Sunshine Seafood Mushroom	1011010801	edible mushroom	2022-07-29	1.000	2.5	sales
878507	Wofeng Sunshine Seafood Mushroom	1011010801	edible mushroom	2022-07-29	1.000	2.5	sales

2.1. Modeling of category and item distribution patterns

We will use the approach of making time series plots to visualize the distribution pattern of categories and individual items of vegetable goods. Therefore, the aggregation approximation is performed with the number of subsequences as 1 and 10, respectively.

(1) Analyze the distribution pattern of categories from Figure 1 and Figure 2 can be seen: six categories are tending to flatten, but do not rule out the occasional few days there will be a surge in sales, but the overall distribution is still relatively uniform; which flower and leafy vegetables sales compared to the other five types of vegetables have a clear difference, has been in the leading position in sales, and aquatic roots and tubers are the opposite at the bottom of the state.

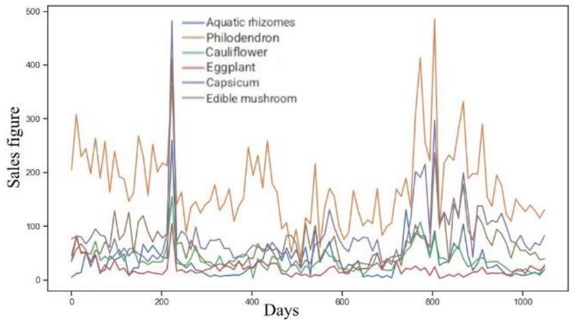

Fig. 1 Change in category sales when the number of subsequences is 1

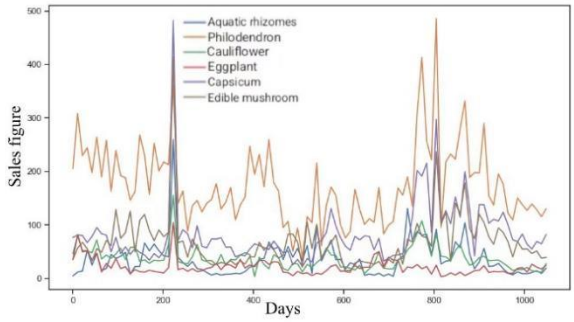

Fig. 2 Change in category sales when the number of subsequences is 10

(2) Analyze the distribution pattern of single products we mainly selected the six highest sales volume and the most representative of the single product, the purpose of doing so is that it allows us to focus on the most important vegetable varieties sold by the superstore, which is conducive to the understanding of the customer's personal preferences and needs, and provides an important guide to the superstore's inventory management and the procurement of fresh commodities; the second is that through the analysis of the sales volume of a small number of single products can greatly simplify the data Secondly, by analyzing the sales volume of a few individual products, we can greatly simplify the complexity of the data and make the analysis more intuitive, which helps us to quickly extract key information and quickly understand the general situation of vegetable sales. From Figures 3 and 4, we can see that in the first year, except for Wuhu peppers and Chinese cabbages, which have obvious increases, the other four items tend to be stable in the whole time series; however, there are still cases of soaring sales on individual days; in the last two years, the sales curves of the six individual items converge with each other, which means that their sales have remained more or less the same.

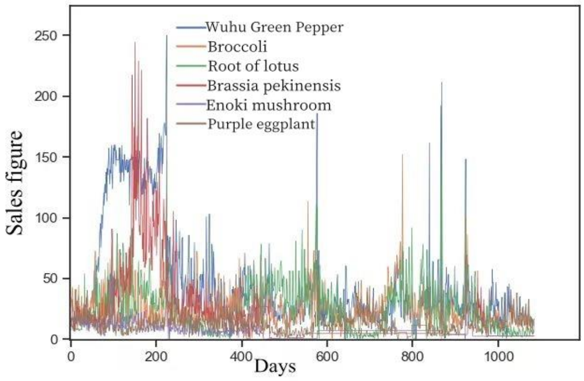

Fig. 3 Change in single product sales when subsequence is 1

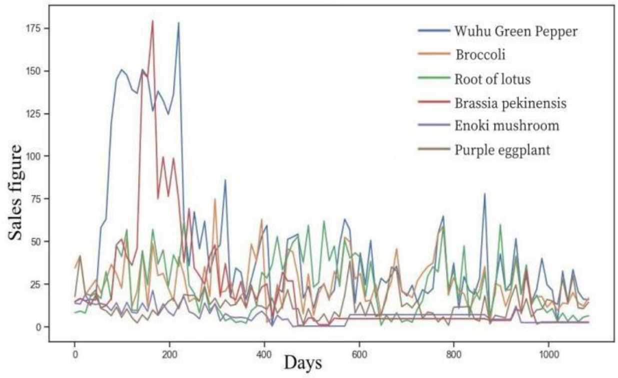

Fig. 4 Change in single product sales when subsequence is 10

2.1.2 Accuracy analysis based on ARIMA modeling

In order to verify the accuracy of the time series plots of the categories, we introduce ARIMA (Autoregressive Moving Average Model) to perform a validation fit [6]. We choose the flower and leaf category as a representative to analyze. It is easy to know that the customer demand for the vegetable category is stable, and thus the sales of the superstore is a stable set of data that can capture the pattern. The model fit values of the time series model for the foliage category are measured as shown in Figure 5 below:

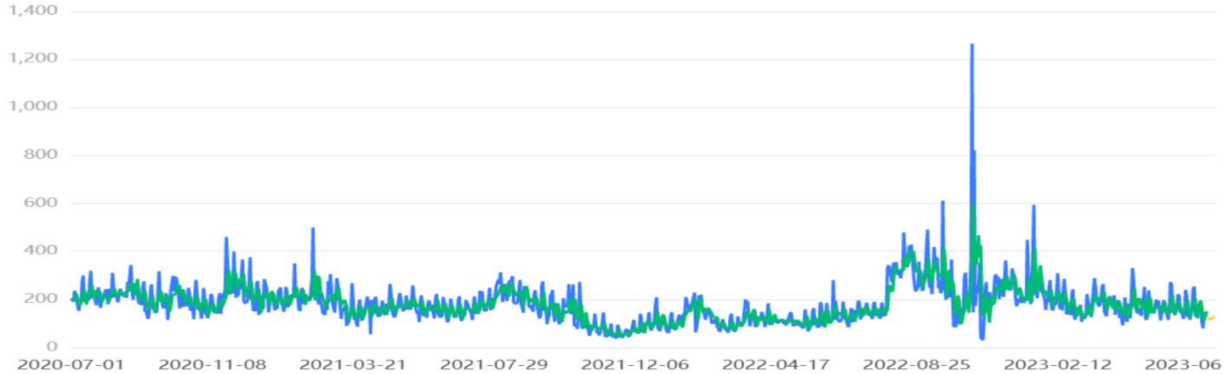

Fig. 5 ARIMA model fit for floral and foliage classes

It can be seen that within this time series is able to coincide both in the overall pattern of distribution and in the presence of mutations, which is a good argument for the accuracy of the model in the time series plot.

2.1.3 Periodicity analysis based on discrete wavelet transforms

To further illustrate the periodic variation of the sales time series graph, the model based on discrete wavelet are its periodicity is verified. Edible mushrooms are selected as an example, as shown in Figure 6 below:

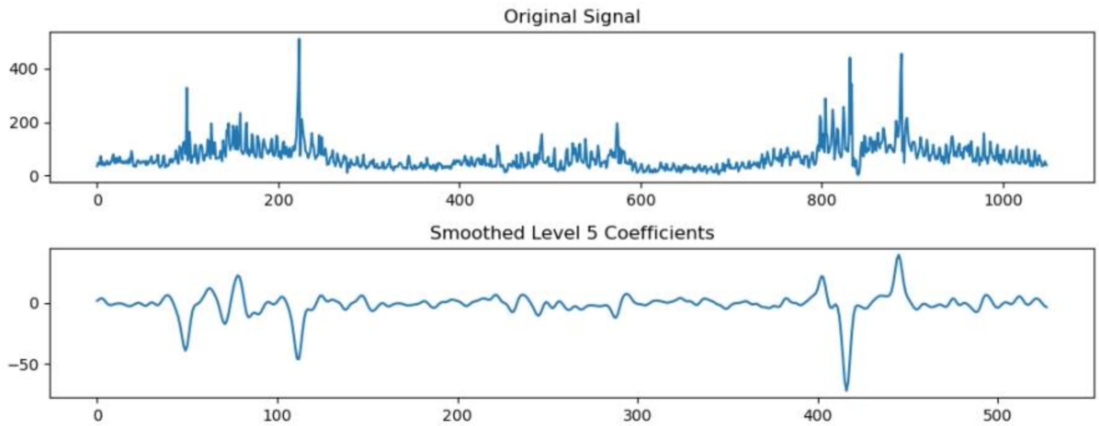

Fig. 6 Plot of wavelet analysis data for floral and foliage classes

As can be seen from the figure, the data image can be divided into three segments according to the time series, the first two segments are consistent showing cyclical changes, while the third segment is disordered, according to the almanac analysis may be affected by social factors such as epidemics, which led to a large fluctuation in sales. This analysis validates the cyclical variation of the sales time series model.

2.2. Analysis of the relevance of each of the categories and individual products

2.2.1 Spearman correlation coefficient

Spearman's correlation coefficient is a statistical measure of the degree of non-linear association between two variables. The formula for calculating this coefficient is shown below:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{1}$$

The correlation coefficients between categories were calculated and converted into a heat map as shown below:

Figure 7 Heat map of category correlation coefficients

From Figure 7, it is easy to see that there are three positive correlation coefficients of more than 0.5 with other categories for both foliar and edible mushrooms, while eggplant has negative correlation with edible mushrooms and aquatic rhizomes, and the rest of the correlations can be identified with the help of different color shades.

2.2.2 K-means clustering algorithm

K-mean clustering algorithm is an algorithm that analyzes operations through an iterative method, and the basic idea of the K-mean clustering algorithm is to select the number of clusters K that need to be classified, and randomly select K data objects as the initial clustering centers from a dataset that contains n data objects [10]. Then, the similarity between other data objects is measured by calculating the distane between them and these K initial clustering centers. The formula for this algorithm is shown in Equation 2: Results

$$J(C_1, C_2, \ldots, C_K) = \sum_{j=1}^{k} \sum_{i}^{n_j} (x_i - c_j)^2 \tag{2}$$

Through cluster analysis, four cluster groupings were obtained, with the same group indicating strong correlation and different groups indicating weak correlation. The contour coefficient plot is shown in Figure 8 below, and in the results obtained, it can be summarized into the following four categories: first, home-style dishes with high sales, such as turnip peppers, Chinese cabbage, broccoli, net lotus root, and enoki mushrooms. The second is seasoned dishes with high sales volume, mostly used as side dishes, such as millet pepper, screw pepper, red bell pepper, and Yu Gan pepper. Third, seasonal dishes, only a certain period of time sales are higher, the frequency is larger. Fourth, other categories of dishes, and the other three categories of correlation is not strong, the sales volume is low.

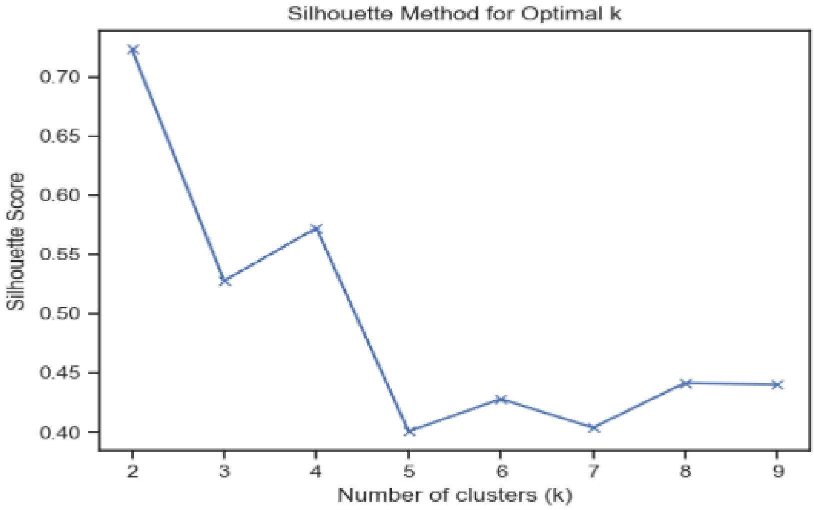

Fig. 8 Plot of k-means contour coefficients

3. Vegetable Pricing and Replenishment Strategy Creation

We used linear regression, support vector machine regression, random forest regression [8], and GA-XGBoost regression model[9] to solve them respectively, and the fixed coefficients, average absolute percentage errors, and average relative errors of the four models were calculated according to the evaluation indexes, and the results were obtained as follows in Table 2.

Table 2 Indicator evaluation results of the four models

Mould	data set	Mean absolute percent error	CoD	Relative absolute errors
linear regression	training set	54.627	0.034	0.584
	Cross-validation set	54.584	0.032	0.561
	test set	54.591	0.023	0.546
SVR	training set	43.731	0.061	0.467
	Cross-validation set	43.852	0.059	0.435
	test set	43.861	0.051	0.439
random forest	training set	19.427	0.831	0.189
	Cross-validation set	19.363	0.849	0.199
	test set	19.390	0.860	0.194
XGBoost	training set	14.637	0.863	0.148
	Cross-validation set	14.552	0.951	0.141
	test set	14.546	0.944	0.145

From the comparative analysis of the four models in the above table, it can be concluded that the GA-XGBoost model is better than the random forest and, SVR and linear regression models both in terms of the coefficient of determination, the average absolute percentage error, and the average relative error. So this paper decides to use XGBoost based as the regression model for analyzing the total sales and pricing.

3.1. Sales volume, pricing fit

This paper statistically analyzes the distribution of sample points on each pricing for total daily sales of chili peppers, foliage, aquatic roots and tubers, edibles, cauliflower, and eggplant, respectively, in order to derive a pattern. The following figures show the graphical representation of the distribution of the eggplant and pepper categories as shown in Figures 9 and 10:

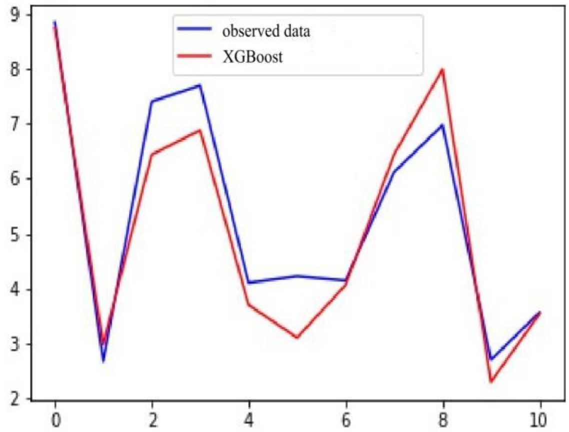

Fig. 9 Fitted curve for eggplant

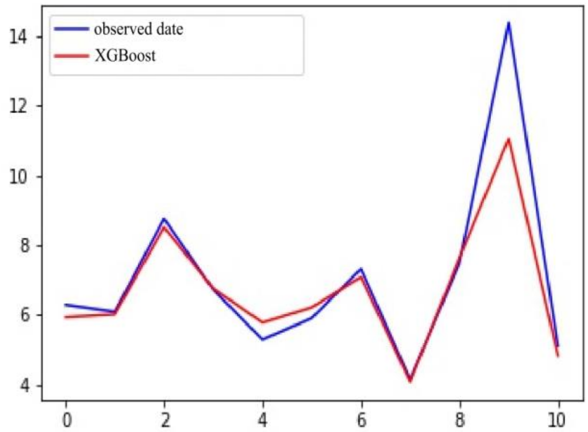

Fig. 10 Fitted curves for chili peppers

3.2. Modeling of merchandise replenishment and pricing

On the SW-tree, data items are arranged according to a pre-defined full-order relationship. When new streaming data arrives, its schema information can be incrementally updated to the sliding window tree, and it is not necessary to scan the historical streaming data several times to determine the support number arrangement relationship of data items [10]. Combined with the commodity category, using the constructed GA_XGBoost regression prediction model based on SW-tree, the 1083 daily data in the data processing for predictive analysis, and several iterations to predict the future week's sales unit price, sales volume, and hypermarket revenue final results. Due to the limitations of this paper, only the results of the first two days of data are shown in Table 3 below.

Table 3 Predictive analysis data for the first two days for the six categories

category	Day 1			Day 2		
	price tag	quantities	earnings	price tag	quantities	earnings
capsicum	20.642	386.918	5296.856	14.029	137.207	1085.298
philodendron	25.666	278.003	5385.950	8.318	490.651	2058.749
aquatic rhizomes	19.104	20.009	136.609	13.135	113.701	240.850
edible mushroom	12.427	237.810	871.353	13.773	110.557	485.416
cauliflower	12.522	60.307	237.904	13.532	63.389	318.080
eggplant	7.921	46.433	151.465	7.930	44.588	134.722

4. Conclusions

In this paper, for the replenishment problem and pricing problem of fresh commodities, a prediction model based on GA-XGBoost is established, and prediction simulation is carried out using existing historical data. The results prove that the GA-XGBoost commodity replenishment and pricing prediction model is better than the models (algorithms) such as linear regression, random forest, and SVR in terms of prediction accuracy. Therefore, the GA-XGBoost prediction model can effectively and accurately help formulate strategies for commodity pricing and replenishment problems.

In the actual merchandise pricing and replenishment prediction, it will be affected by many factors. Therefore, superstore enterprises should utilize the GA-XGBoost model to develop appropriate strategies for merchandise replenishment and pricing in light of their own reality. Meanwhile, the GA-XGBoost model involves the selection of multiple parameters of XGBoost and genetic algorithm, which requires appropriate tuning and optimization to obtain better results.

References

[1] Jiang Yingmei,Mou Jinjin. Joint decision making of inventory and pricing of fresh processed products based on perceived freshness[J]. Highway Transportation Science and Technology,2020,37(3):151-158.

[2] XU Xiaofeng,YU Lean,LIN Ziru,et al. Short-term sales mix prediction of fresh commodities based on feature fusion[J]. Journal of Management Science,2022,25(12):102-123.

[3] YANG Shuai,HUANG Xiangmeng,WANG Junbin. Research on joint optimization strategy of shelf allocation and pricing for fresh food[J]. Supply Chain Management,2022,3(8):49-59.

[4] PAN Xiaofei,XIE Zhiheng,WANG Shuyun. Optimization decision of fresh food superstore preservation effort and pricing considering loss aversion[J]. Highway Transportation Science and Technology,2022,39(6):177-185,190.

[5] BI Wenjie,ZHOU Yubing. Research on joint inventory control and dynamic pricing of fresh products based on deep reinforcement learning[J]. Computer Application Research,2022,39(9):2660-2664.

[6] SONG Ge, LI Xiaosuan, WANG Kewei. Application of ARIMA and SVM combined model in the prediction of novel coronavirus pneumonia[J]. Chinese Journal of Hospital Infection,2022,32(1):151-155.

[7] WANG Sen,LIU Chen,XING Shuaijie. A review of research on K-means clustering algorithm[J]. Journal of East China Jiaotong University,2022,39(5):119-126.

[8] Cao, Taoyun. A study on the importance of variables based on random forest[J]. Statistics and Decision Making,2022,38(4):60-63.

[9] WANG Junfeng,ZENG Bi,GUO Zhixing. A monocular visual SLAM approach for ARM isomorphic processors[J]. Software Guide,2023,22(2):67-74.

[10] WU Xin,LI Jun,MAO Wisdom,et al. Prediction of e-commerce users' purchase behavior based on GA-XGBoost[J]. Journal of Zhejiang Wanli University,2022,35(4):86-92.