Research on Vegetable Sales Prediction and Strategy of Supermarkets Based on XGBoost and Random Forest Modeling

Kaiwen Zhou[†]
China University of Mining & Technology, Beijing
Beijing, China
kw3253975912@163.com

Xiang Li[†]
China University of Mining & Technology, Beijing
Beijing, China
lixiang237617@163.com

Tong Sun[†]
China University of Mining & Technology, Beijing
Beijing, China
13716810187@163.com

Bingchun Luo[†]
China University of Mining & Technology, Beijing
Beijing, China
2110630415@student.cumtb.edu.cn
† These authors also contributed equally to this work

2023 International Conference on Applied Physics and Computing (ICAPC) | 979-8-3503-5924-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICAPC61546.2023.00076

***Abstract*—This paper mainly focuses on the forecasting and strategy development of vegetable sales in supermarkets. Statistical tools were used to organize and categorize the sales volume and sales unit price of each category, and descriptive statistics and data visualization were completed to observe and analyze the sales distribution pattern of each category and single product of vegetables. First, the association and correlation between individual items and between categories were analyzed by the Spearman correlation coefficient test. Secondly, the optimization model, XGBoost model, and random forest model were constructed to predict the sales volume and price of vegetables in six categories, and the performance of the models was verified by several evaluation indexes. Finally, the objective planning model for maximizing the benefits of the superstore was established, the PSO algorithm was used for optimization, and the daily replenishment quantity and pricing strategy for the coming week were formulated for each category, which provided the superstore with decision-making references with practical value.**

Keywords—XGBoost, Random Forest, Goal Programming, PSO Algorithm

I. Introduction

In the marketing and sales arena, pricing and replenishment strategies for time-sensitive commodities, particularly fresh vegetables, are critical to ensuring that superstores maximize profits. The sales and availability of such commodities are often constrained by their unique timeliness, uncertain replenishment status, and time-related changes in supply and demand [1, 2]. This paper focuses on the pricing and replenishment strategies of vegetable commodities in fresh food superstores, aiming to address the following core issues: first, by analyzing the possible correlations between the categories and individual products of different vegetable commodities as well as the distribution pattern of sales volume, to reveal their interrelationships. Second, a replenishment plan is formulated based on the category as a unit to explore the relationship between total sales volume and cost-plus pricing, and to provide suggestions for the total daily replenishment volume and pricing strategy for the coming week, to maximize the superstore's revenue. Finally, a replenishment plan is formulated for individual items, and this paper takes the available varieties from June 24 to 30, 2023 as an example, and provides suggestions for the replenishment volume and pricing strategy for individual items on July 1, aiming to maximize the revenue of the superstore under the premise of meeting the market demand [3, 4].

II. Category and Dish Correlation Analysis

A. Sales Frequency Distribution Patterns of Categories and Individual Products

There is a significant variation in the number of subdivided types of vegetable categories in the marketing sector. To quantitatively measure this variation, this paper summarizes and presents the distribution of the number of different vegetable categories based on the classification codes and classification names provided. The results show that the smallest number of individual items is in the cauliflower category, followed by eggplant and aquatic roots and tubers, while the largest number of individual items is in the leafy vegetables category.

By further analyzing the sales data, this paper categorized the single-item codes as shown in Fig. 1, counted the sales frequency of different categories of vegetables, and visualized the sales proportion of each category through a box plot. The ordering of sales frequency is not the same as the ordering of the category single-item categories, showing that variety does not directly lead to an increase in purchase frequency.

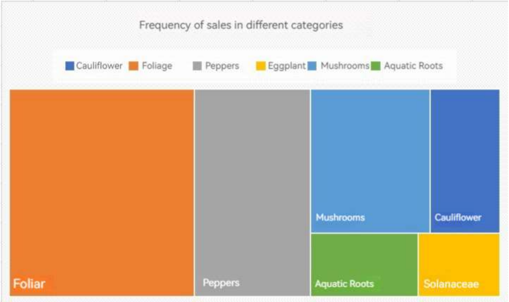

Fig. 1. Number of individual items in each category

Authorized licensed use limited to: Liberty University. Downloaded on October 28,2024 at 03:47:01 UTC from IEEE Xplore. Restrictions apply.

This paper explores factors that may influence sales frequency, including the daily demand for certain vegetable items and geographical eating habits. For example, in the Yunnan region, edible mushrooms are more popular than in other regions of the country, and the sales frequency is correspondingly higher.

To further clarify the sales frequency of different individual products, this paper counted the sales frequency of all individual products and produced a Nightingale rose diagram as shown in Fig. 2 for display. The results show that the highest sales frequency is Wuhu green pepper in the chili category, followed closely by broccoli in the cauliflower category. In addition, the sales frequency of enoki mushrooms in the aquatic root category was also high, while water chestnuts, which had the lowest sales volume, also belonged to the aquatic root category. These results provide valuable references for understanding people's purchasing habits of vegetable individual items in the region.

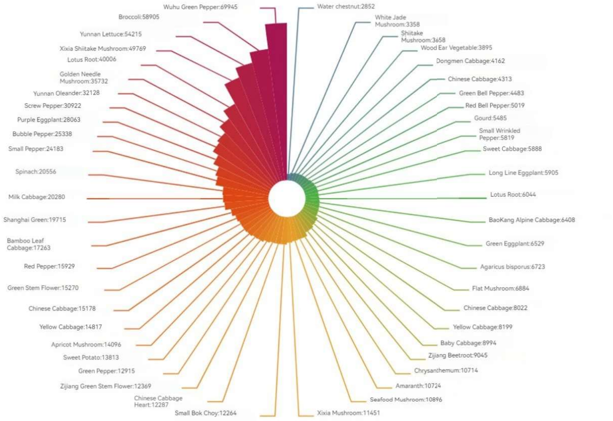

Fig. 2. Sorting of sales frequency of individual products

After an in-depth discussion of sales frequency distribution, it is important to turn to the analysis of sales volume distribution. Sales frequency shows the frequency of purchase but does not reveal the amount of purchase "volume", which is critical to the formulation of replenishment strategy. For this reason, this paper on the sales volume data for centralized statistics and box plot analysis is shown in Fig. 3. First, the box plots of single sales volume for all vegetable categories show that the values of each category overlap, and the median difference does not exceed 0.5 (Kg), indicating that the single purchase volume is generally similar. In particular, the higher single sales volume is for aquatic roots and tubers, while the lower one is for edible mushrooms, but the lower sales frequency ranking reveals that certain categories have high sales frequency but not much actual sales volume. This provides an important basis for evaluating pricing gains and replenishment strategies. Second, the box line graph of total daily sales volume indicates that although the single sales volume is similar, there is a significant difference in the cumulative daily sales volume due to the difference in sales frequency. In particular, the flower and leaf category has the largest sales volume, highlighting the importance of timely replenishment. In summary, through an exhaustive analysis of sales volume, this paper provides a more comprehensive perspective for the evaluation of pricing and replenishment strategies. At the same time, these findings provide a useful reference for developing more precise sales and replenishment strategies in the future.

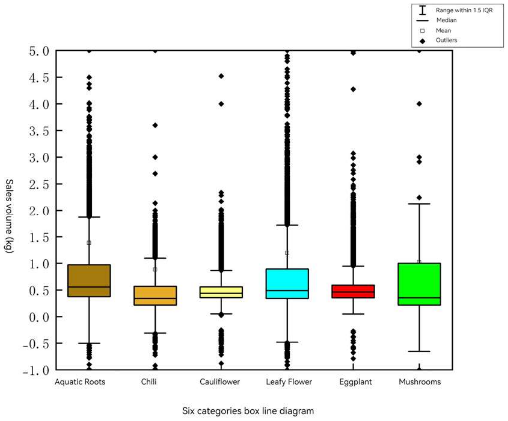

Fig. 3. Box line diagram of single sales volume by category

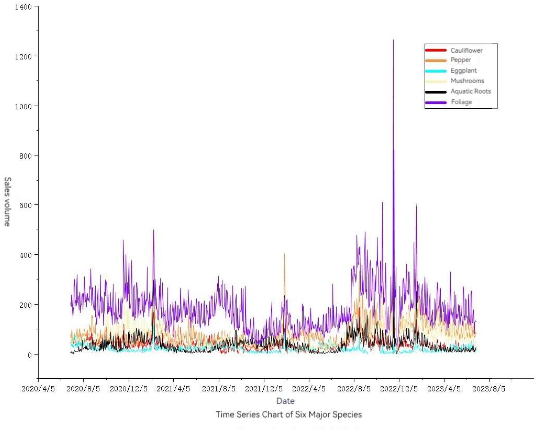

Fig. 4. Distribution of sales volume of the six categories concerning date

Analysis of Fig. 4 shows that the difference in sales of different single products of the same category is very large, such as in the chili pepper single product, certain single products such as the combination of pepper series, chili pepper combo packs almost no one to buy, as well as nearly half of the edible fungi single product almost no purchase record, may be because although the purchase order contains the single product, because of the scarcity of sources of goods superstores seldom stocked, or the favorite crowd is smaller resulting in fewer purchases. The higher sales volume of all

Authorized licensed use limited to: Liberty University. Downloaded on October 28,2024 at 03:47:01 UTC from IEEE Xplore. Restrictions apply.

items included Wuhu green peppers, enoki mushrooms, and broccoli.

B. Correlation analysis

The Spearman correlation coefficient has wider applicability conditions than the Pearson correlation coefficient [5]. Therefore, Spearman correlation coefficient is directly used. Since n is greater than 30, which is a large sample, the test value is calculated, and the corresponding p-value is compared with 0.05. The heat map of the correlation coefficient for the calculated categories is shown in Fig. 5.

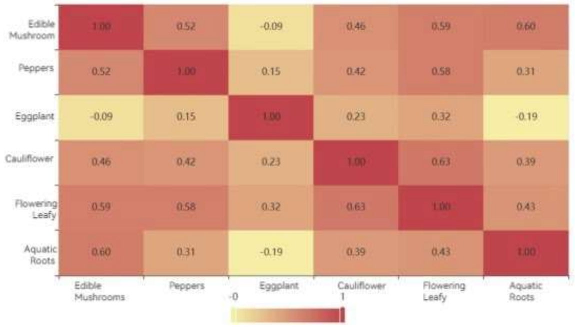

Fig. 5. Heat map of Spearman's correlation coefficient for six categories

Concluding:

(1) Most of the correlations among the six individual categories showed positive correlations;

(2) The capsicum category showed a strong positive correlation with aquatic rhizomes and a weak correlation with the eggplant category;

(3) The edible mushrooms category showed moderate to strong positive correlation with words also tired and aquatic rhizomes;

(4) Cauliflower species showed a strong positive correlation with phloem species and a weak correlation with eggplant species.

III. CATEGORY REPLENISHMENT AND PRICING OPTIMIZATION

A. Modeling

The demand function reveals the consumer's response to changes in the price of goods and is key to analyzing the impact of superstore pricing. In this paper, the impact of markup cost pricing on the total sales of each vegetable category is explored through a constant elasticity demand function (double logarithmic model). After data processing to obtain the average price of each commodity, the model is applied to fit the model, and part of the results are shown in Fig. 6 to Fig. 9, and the results of the flower and leaf category are plotted using Origin software, which provides an important reference for the pricing and replenishment strategy of the superstore.

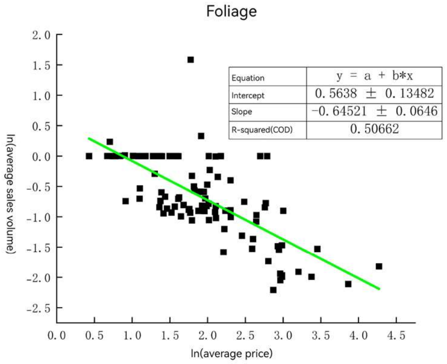

Fig. 6. Fitted plot of floral and foliar species

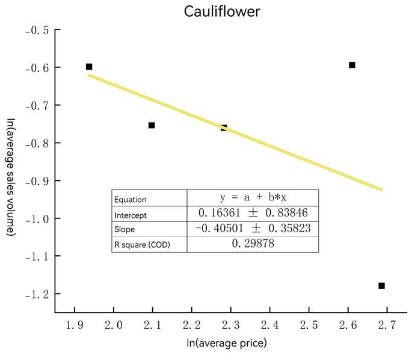

Fig. 7. Fitted plot for cauliflower species.

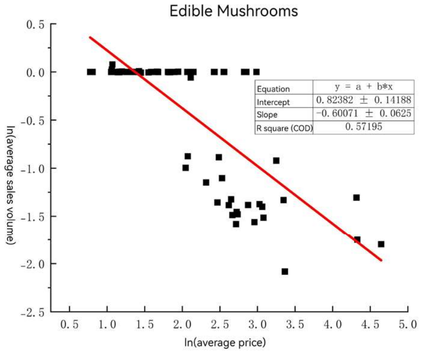

Fig. 8. Fitting plot for edible mushrooms

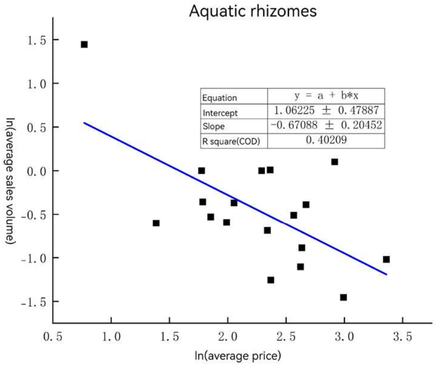

Fig. 9. Fitted plot of aquatic.

Authorized licensed use limited to: Liberty University. Downloaded on October 28,2024 at 03:47:01 UTC from IEEE Xplore. Restrictions apply.

The graphical results exhibit an inverse relationship between average sales volume and average price for most vegetable categories, which is consistent with daily experience. However, some of the data points are poorly fitted and may be influenced by the particular period. The anomaly is the eggplant category, where the sales volume is positively related to the price, possibly due to less data or specific attributes of the vegetable.

B. Replenishment and pricing optimization

1) XGBoost-based category sales prediction

XGBoost is a tree enhancement algorithm that specializes in handling sparse data and parallel computation and is suitable for sorting and regression problems, with a speed of at least ten times that of existing gradient boosting [6]. Its objective function consists of a loss function, which describes the degree to which the model fits the data, and a regular term, which avoids over-complexity of the model leading to overfitting, and improves generalization. The prediction model is updated after the tenth iteration. XGBoost improves the applicability of XGBoost by obtaining the second-order derivative form of the function through Taylor expansion, separating the selection of the loss function and the optimization of the model algorithm. With sales volume as the dependent variable and sales unit price and wholesale price as the independent variables, the XGBoost model is constructed, and the parameters are set as default constants, through which the impact of pricing on sales volume can be analyzed in depth to provide data support for pricing strategy.

As Table I takes the prediction results of the flower and leaf category as an example. The results of the table run out of Python are organized into Excel to make it more simple and intuitive.

TABLE I FORECAST RESULTS OF SALES VOLUME OF FOLIAGE AND FLOWERS

Category	Next week	Quantity sold (kg)
Foliage	Day 1	434.9904
	Day 2	220.8845
	Day 3	237.004
	Day 4	428.7252
	Day 5	345.6129
	Day 6	196.8927
	Day 7	257.9579

Secondly, the average absolute percentage errors MAPE for the test sets cauliflower, aquatic rootstock, foliar, eggplant, edible mushroom, and chili pepper were obtained as 20.792, 28.076, 13.124, 19.438, 16.084, and 14.855, respectively. The corresponding r2 for the test sets were 0.9360, 0.9459, 0.9310, 0.9482, 0.9391, 0.9440. It is found that the average absolute percentage errors of the test sets are mostly below 20%, and the r2 scores are all higher than 0.9, which achieves a more satisfactory model-fitting effect.

2) Random Forest-based forecasting of future wholesale prices

Random forests train decision trees through put-back sampling and random subspace methods, which reduce overfitting, are easy to parallelize, and accelerate large-sample training. For wholesale price prediction, a specific function model is constructed, where Ti is the decision tree and n is the number of decision trees. Facing the lagging problem of predicted data in multi-step time series forecasting, a lagging feature, i.e., historical wholesale price, is introduced to improve the prediction accuracy of wholesale price P at the future (t+1) time.

The results of solving the problem with the example of the foliage category are shown in Table II.

TABLE II FORECAST RESULTS OF SELLING PRICE OF FOLIAGE AND FLOWERS

Category	Next week	Price (yuan/kg)
Foliage	Day 1	8.29524
	Day 2	4.897983
	Day 3	6.400685
	Day 4	8.329545
	Day 5	7.575759
	Day 6	7.005502
	Day 7	7.075985

3) Targeted planning for maximizing superstore revenue

To maximize the revenue of the superstore, it is necessary to develop a target planning model considering the attrition rate. This model aims to maximize the total revenue for the next seven days, where the number of sales is predicted by the previous XGboost model. By taking into account the incoming loss rate, quality of the product, and wholesale price, as well as using the XGboost model to predict the number of sales, this model provides the superstore with a well-defined objective and structured planning scheme for maximizing revenue. Objective function:

$$\max z=\sum_{i=1}^{7} XS_i \times \left(\frac{100-SHL}{100}\right) \times (X_i-PF_i) \qquad (1)$$

Particle swarm optimization algorithm is a population intelligent optimization algorithm that initializes a set of solutions randomly, and then iterates, continuously updating these solutions in the search space, so that the whole population as a whole is adjusted in the direction of better fitness values, and eventually expects to find the optimal solution of the problem within a limited number of iteration steps [7]. The flow is shown in Fig. 10.

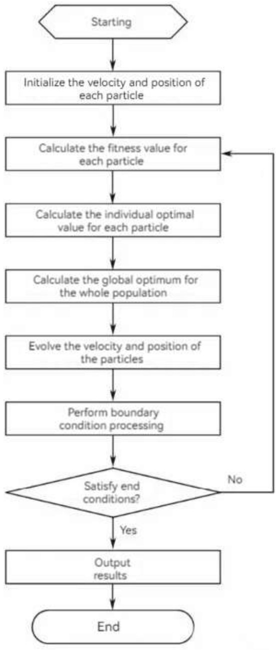
Starting
Initialize the velocity and position of each particle
Calculate the fitness value for each particle
Calculate the individual optimal value for each particle
Calculate the global optimum for the whole population
Evolve the velocity and position of the particles
Perform boundary condition processing
Satisfy end conditions?
No
Yes
Output results
End

Fig. 10. Particle swarm algorithm flow

Firstly, initialize the unit sales price and assign an initial

Authorized licensed use limited to: Liberty University. Downloaded on October 28,2024 at 03:47:01 UTC from IEEE Xplore. Restrictions apply.

speed to this set of values; secondly, use the objective function to calculate the fitness value of the unit sales price; thus calculate the optimal value of the sales pricing for each category, and thus get the optimal value of the objective function to maximize the objective function; every time the optimal value of the objective function is obtained the pricing has to be updated accordingly; if the new pricing satisfies the boundary conditions set in advance, then the result is outputted after reaching the number of iterations; conversely. If the boundary conditions are not met, then return to the use of the objective function to calculate the adaptive value of the sales unit price of this step to adjust. Solve for the total return using the flower and foliage category as an example as follows shown in Table III.

TABLE III FOLIAGE REVENUE

Category	7th	Price (yuan/kg)	Quantity sold (kg)	Proceeds (yuan)
Foliage	1	8.29524	434.9904	1826.005
	2	4.897983	220.8845	310.8266
	3	6.400685	237.004	597.4793
	4	8.329545	428.7252	1803.526
	5	7.575759	345.6129	1216.611
	6	7.005502	196.8927	581.7779
	7	7.075985	257.9579	851.1857

IV. INDIVIDUAL REPLENISHMENT AND PRICING OPTIMIZATION

A. Random forest-based forecasting of individual product sales prices

In the model construction session, the random forest model continued to be used, and the time-series data from June 24-30, 2023 were first processed to transform the raw time-series data into a data format suitable for regression models. Subsequently, the dataset was split into a training set and a test set for model training. The performance of the model was quantified and future values were predicted by calculating the evaluation metrics such as mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R2). The results are shown in Table IV for the example of floral and foliage species.

TABLE IV UNIT PRICE OF SINGLE PRODUCT SALES IN THE FLOWER AND FOLIAGE CATEGORY

Category	Single product	Price
Foliage	Spinach (servings)	11.29
	Wood Ear Vegetable	8.06
	Milk Bok Choy	6.12
	Shanghai bok choy	7.02
	Baby Cabbage	6.25
	Amaranth	5.31
	Yunnan Lettuce	9.72
	Yunnan Lettuce (portion)	9.31
	Yunnan Oil Wheat Lettuce (portion)	7.52
	Bamboo Leaf Lettuce	5.63

B. Objective planning model for maximizing returns to superstores under multiple constraints

The optimization and forecasting of individual items is with the added constraint that the sales volume must be greater than 2.5, as well as a limit on the total number of individual items.

The decision variable is the Price on day i.

The new objective function is:

$$R(k, PF) = S(x, PF) \times (\frac{100 - L}{100}) \times (x - PF) \quad (2)$$

The new constraints are: $x_i > PF_i$, $S(x, PF) > 2.5$

Taking the foliage class as an example this paper uses the PSO optimization algorithm to optimize the results as shown in Table V.

TABLE V FLOWER AND FOLIAGE CATEGORY SINGLE PRODUCT REVENUE

Category	Single product	Price (yuan/kg)	Quantity sold (kg)	Proceeds (yuan)
Foliage	Spinach (servings)	11.29	8.66	72.73
	Wood Ear Vegetable	8.06	5.93	29.08
	Milk Bok Choy	6.12	8.81	24.43
	Shanghai bok choy	7.02	10.95	46.74
	Baby Cabbage	6.25	17.06	30.14
	Amaranth	5.31	10.26	22.91
	Yunnan Lettuce	9.72	45.11	167.12
	Yunnan Lettuce (portion)	9.31	71.62	438.59
	Yunnan Oil Wheat Lettuce (portion)	7.52	47.56	235.56
	Bamboo Leaf Lettuce	5.63	28.3	88.8

In addition to this, the maximization of revenue also takes into account the impact of several other aspects:

(1) The effect of seasonal factors, as certain vegetables may be more popular in a particular season, data collected including demand, sales, and price of different vegetables in different seasons can be done for correlation analysis.

(2) Competitors' sales strategies. This includes the vegetable prices and promotional activities of other superstores in the same region. This can help supermarkets understand the competitive landscape of the market and develop differentiated replenishment and pricing strategies.

(3) Supermarket inventory. Includes the inventory and on-sale quantities of vegetable types, as well as the storable time of different vegetables. This can help predict which vegetables will need to be discounted and sold quickly due to overstocking, or which vegetables may need to be heavily restocked due to shortages.

(4) Consumer Behavior and Market Trends: This includes the collection of data on customers' purchasing behavior, such as the number of times vegetables are purchased, the volume of purchases, and the mix of purchases. This can help supermarkets understand the purchasing habits and preferences of different customer groups, to replenish and price in a personalized way; as well as information on consumers' health awareness, dietary trends, and requirements for food safety, etc. This can help supermarkets understand the market trends of the vegetable category and the changes in consumer demand, to adjust the replenishment plan and pricing strategy according to the market trends and consumer preferences.

In addition to the above four aspects, superstores can also collect supplier data to understand the quality, price, and other

Authorized licensed use limited to: Liberty University. Downloaded on October 28,2024 at 03:47:01 UTC from IEEE Xplore. Restrictions apply.

information of vegetable categories and individual products provided by different suppliers to provide a basis for replenishment decisions. Or collect the impact of online grocery shopping software on the offline superstore industry, predict the possible impact of online platforms, and make reasonable pricing of stocking strategies, etc.

V. Conclusion

This paper demonstrates the sales frequency, sales volume, and its time distribution pattern through in-depth data visualization, and provides accurate sales volume and price prediction for superstores, to give reasonable strategies and suggestions. Although the model is simple to implement and has parallel processing capability, it shows shortcomings in dealing with the abnormal data of cigar class and fails to fully consider external factors such as weather and natural disasters. To improve the model, strategies such as eliminating outliers, redefining the wholesale price, and establishing a hierarchical analysis model are proposed. Future promotion directions include more comprehensive environmental analysis, iterative optimization combined with machine learning techniques, and further optimization of time series analysis to enhance the dynamics and predictive accuracy of the model to provide continuous and accurate support for the long-term operation of superstores.

References

[1] Huang Lianruo, Cen Zhongdi. Research on pricing and inventory decision-making of fresh food e-commerce enterprises under double efforts[J]. Journal of Zhejiang Wanli College,2023,36(05):21-28.DOI:10.13777/j.cnki.issn1671-2250.2023.05.016.

[2] Yu-Ying Song. Research on optimization of marketing strategy of MH fresh food supermarket [D]. Henan University of Finance and Economics and Law,2023.DOI:10.27113/d.cnki.ghncc.2023.000443.

[3] Gu, S. H. A study on dynamic pricing of fresh products in H-retailers considering freshness variation[D]. Donghua University, 2023. DOI:10.27012/d.cnki.gdhuu.2023.001318.

[4] 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(08): 49-59. DOI:10.19868/j.cnki.gylgl.2022.08.005.

[5] Yu ZT, Bei YJ, Zhang W et al. A study on the current situation of physical health, mental health and social adaptability of students in higher vocational colleges and the correlation of Spearman's scale[J]. Sports Vision,2023(12):7-11.

[6] CAO Xinzhi, SHEN Junshu, WANG Jie. Construction of a prediction model for immune thyroid function abnormality based on XGBoost algorithm[J]. Chinese Journal of Health Information Management, 2023,20(05):833-837.

[7] XU Wenxin, RUI Jia, LI Yuanxun, et al. Optimal design of Kevlar mesh shell based on particle swarm optimization algorithm[J]. Journal of Qinghai University, 2023, 41(05): 102-108. DOI:10.13901/j.cnki.qhwxxbzk.2023.05.014.

Authorized licensed use limited to: Liberty University. Downloaded on October 28,2024 at 03:47:01 UTC from IEEE Xplore. Restrictions apply.