

Pricing and Replenishment Prediction Model for Vegetable Commodities Based on Mixed Integer Linear Programming Optimization XGBoost

Shan Chen, Yuying Ma[†], Qiutong Wu[†], Yixin Bian^{*}

School of Computer Science and Information Engineering, Harbin Normal University, Harbin, Heilongjiang, China, 150500

^{*} Corresponding Author Email: bianyu79@163.com

[†] These authors contributed equally.

Abstract. In fresh food superstores, the replenishment volume and pricing strategy of vegetables are affected by various factors such as seasonal seasons, supply demand, purchasing costs, and profits of vegetable categories. In this paper, we use the data related to the sales flow details and wholesale prices of each commodity from July 1, 2020 to June 30, 2023 as a dataset, and build a model based on the gradient boosting tree algorithm, XGBoost algorithm [1], to predict the total daily replenishment volume and pricing strategy of each vegetable category in the coming week (July 1-7, 2023) [2]. It is also optimized using a mixed-variable 0-1 linear programming (MILP) [3] model with decision variables including a binary variable of whether to order the vegetable category or not and a continuous variable of the number of vegetable categories ordered to compute the optimal replenishment volume and pricing strategy for July 1, 2023 for each item. Through the model in this paper, under the premise of meeting the market demand for each category of vegetable goods, while keeping the control within a specific range, the superstore can get the optimal replenishment plan and pricing strategy, reasonably allocate the resources of the warehouse space, improve the market competitiveness of the superstore, and maximize the revenue.

Keywords: XGBoost prediction model, mixed integer linear programming, vegetable replenishment pricing strategy.

1. Introduction

Fresh produce supermarkets face a number of challenges when dealing with vegetable commodities. These commodities are characterized by short shelf life and susceptibility to quality, and therefore require scientific and reasonable replenishment and pricing strategies. Supermarkets usually purchase vegetables in the early hours of the morning, but they are unable to specify individual items and purchase prices at the time of purchase, so they need to make decisions under unknown conditions. Supermarkets generally use "cost-plus pricing", which also takes into account wastage and discounted sales. In order to make more informed decisions, supermarkets need to fully understand the market demand and sales trends, and at the same time, rationally control the available varieties and quantities. This paper analyzes the relationship between total sales and cost-plus pricing for each vegetable category in the vegetable sales data based on market research and relevant supermarket sales records, establishes an XGBoost forecasting model for replenishment planning by vegetable category in supermarkets, and gives the daily replenishment totals and pricing strategies for each vegetable category in the coming week (July 1-7, 2023), which can maximize the revenue of supermarkets. On this basis, the practical requirements of life are combined: the vegetable category has limited selling space; the total number of saleable items needs to have a certain interval range (in this paper, it is controlled at 27-33 items); and the order quantity of each individual item has a minimum display quantity requirement (in this paper, 2.5kg is the requirement). Based on the available varieties from June 24-30, 2023 [4], a 0-1 linear programming model based on mixed variables is developed to give the replenishment quantity of individual items and pricing strategy on July 1, 2023, to maximize the revenue of the superstore under the premise of satisfying the need for individual vegetable items in the market as much as possible.

2. Research Methodology

2.1. Data Sources and Preprocessing

This study aims to address the challenges faced by superstores in replenishment planning and pricing decisions in the vegetable category, the data used comes from market research and the data related to the sales flow details and wholesale prices of each commodity of the superstores from July 1, 2020 to June 30, 2023, the data used include: vegetable category commodity information, sales flow details and wholesale price data, and the data of the recent wastage rate. The missing rate of the data is found to be 0. This paper adopts the replacement method for the outliers of the data to prevent the overall data from being affected by the large fluctuation of the data in a short period of time. The sales volume is analyzed using quantile and robust statistical methods for all the measured data. Since there are fewer outliers here, the data was directly processed to remove outliers. Multiple data tables were merged to construct a comprehensive data set. Information from different sources was integrated to provide complete data for subsequent modeling and analysis.

2.2. Principle of XGBoost model [5-7]

The theoretical basis of Extreme Gradient Boosting (XGBoost) is the theory of classification and regression trees, which can solve regression and classification problems very effectively and belongs to the integrated learning model. The algorithm is based on the gradient boosting decision tree (GBDT), improving the calculation of the objective function and reducing the number of operations. The core of XGBoost is to fit the residuals of the prediction results and optimize them to the prediction scores by constructing a new optimization structure for the decision tree, and summing the prediction scores of each tree to be the prediction value of the prediction model.

The dataset $D = \{x_i, y_i\}$ contains m samples, each sample has n features, and the prediction scores of the samples are obtained by summing the scores of each tree with the following expression:

$$\begin{cases} \hat{y} = \sum_{k=1}^K f_k(x_i), f_k \in F, \\ F = \{f(x) = \omega_{q(x)}(q: R^m \rightarrow T, \omega \in R^T)\} \end{cases} \quad (1)$$

Where: k denotes the number of regression trees, F denotes the space of regression trees, $f(x)$ is one of the classes of regression trees, T denotes the number of leaf nodes (features), x_i represents the input parameter, and $\omega_{q(x)}$ denotes the score of each tree after the structural function $q(x)$ maps the input parameter x_i to the respective leaf node T .

XGBoost transforms the optimization problem of the objective function into a problem of solving the minimum of a quadratic function and uses the information of the second-order derivative of the loss function to train the decision tree model. In order to avoid overfitting of the integrated model and to generate a simpler set of numbers, the model complexity is added as a regular term to the objective function species, and the expression of the objective function of XGBoost is as follows:

$$L = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (2)$$

Where: y_i is the actual value of the t th iteration, $\hat{y}_i^{(t-1)}$, $\hat{y}_i$ denotes the corresponding predicted value, $l(y_i, \hat{y}_i)$ denotes the error between the predicted value and the actual value, and $\Omega(f_k)$ is the regular term, which is used to penalize the complexity of the tree.

The result of the t th iteration of the objective function is as follows:

$$L^{(t)} \approx \sum_{i=1}^n L(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + C \quad (3)$$

Where: $f_t(x_i)$ denotes the complexity of the decision tree corresponding to the t th iteration of the variable x_i , C is a constant.

Expanding the loss function using the second order Taylor's formula, the objective function is optimized as:

$$L^{(t)} \approx \sum_{i=1}^n l\left(y_i, \hat{y}_i^{(t-1)} + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_t)\right) + \left(\gamma T + \frac{1}{2} \lambda \|w\|^2\right) \quad (4)$$

Where: γ and λ control the number and fraction of leaf nodes respectively to adjust the complexity of the tree, and g_i and h_i denote are the first and second order derivatives of the mean square loss function respectively.

2.3. Model Evaluation Criteria

In this paper, four performance metrics: coefficient of determination R^2 score, mean absolute error (MAE), root mean square error (RMSE), and accuracy are used as performance evaluation metrics of the model. The calculation formula is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (5)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (6)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (7)$$

$$\text{Accuracy} = \frac{\text{Total Number of Predictions}}{\text{Number of Correct Predictions}} \quad (8)$$

Where: n represents the number of samples for testing and training, y_i , $\hat{y}_i$, $\bar{y}$ are the actual, predicted, and average of the actual values of the data, respectively.

Accuracy is a commonly used evaluation metric in classification problems, which indicates the ratio of the number of samples correctly predicted by the model on the test set to the total number of samples. The higher the Accuracy, the better the overall performance of the model on the test set. RMSE score: is a metric used in regression problems to measure the prediction error of the model. It measures the difference between the actual observations and the model predictions. The formula is as follows. R^2 score is a measure of how well the model explains the variance of the target variable in a regression problem. MAE score is another measure of how well the model explains the mean absolute difference between the predicted values of the model and the actual observed values in a regression problem. The lower the MAE, the lower the mean prediction error of the model, which is an important performance indicator for a regression problem.

3. Model building and solving

3.1. Predicting the replenishment and pricing strategy for the coming week XGBoost model

3.1.1 Analyze the relationship between total sales and cost-plus pricing for each vegetable category

The preprocessed data are grouped and summarized according to the "classification code", and then each group is operated to calculate the total sales volume, the average sales unit price, the average loss rate, the average wholesale price, and the total cost of goods sold. Finally, a data table is generated with columns for "Category Code" and "Cost Plus Pricing", which shows the cost plus pricing for each vegetable category, i.e., the average profit per kilogram sold. This can help in analyzing the economics and pricing strategy of the vegetable categories. The partial table of total sales in relation to cost-plus pricing for each vegetable category is shown in Table 1:

Table 1. Total sales and cost pricing relationship for each vegetable category

	Classification Number	Cost Pricing
0	1011010101	1.986771
1	1011010201	2.540966
2	1011010402	2.477578
3	1011010501	2.791882
4	1011010504	3.310978
5	1011010801	3.647758

3.1.2 Predicting Restocking and Pricing Strategies for the Week Ahead XGBoost Model Construction

3.1.2.1 Feature Engineering

Convert the sales dates in the data from text type to date-time type for time series analysis. Extract features such as year, month, day, and day of the week from them, and normalize the features to ensure that they are on the same scale and to avoid certain features from unduly influencing the model. Some of the data features extracted are shown in Table 2:

Table 2. Demonstration of some data feature extraction

	Sales Date	Category Code	year	month	day	day of week
0	2023-07-01	1011010101	2023	7	1	5
1	2023-07-01	1011010201	2023	7	1	5
2	2023-07-01	1011010402	2023	7	1	5
3	2023-07-01	1011010501	2023	7	1	5
4	2023-07-01	1011010504	2023	7	1	5
5	2023-07-02	1011010101	2023	7	2	6
6	2023-07-02	1011010201	2023	7	2	6
7	2023-07-02	1011010402	2023	7	2	6
8	2023-07-02	1011010501	2023	7	2	6

3.1.2.2 Data Splitting

The data was split into a training set (historical data) and a test set (data from the week after) using date conditions. The classification code, year, month, day and day of the week were selected as features from each dataset. Sales volume (kg) was selected as the target variable for training and testing the model. So that the model can be trained and evaluated. This helps in evaluating the generalization performance of the model and ensures the performance of the model on new data.

3.1.2.3 Eigenvalue Normalization Processing

Min-Max normalization is used in the model to map the range of values of different features to phase to a range of 0 to 1. This helps to improve the performance and stability of the model and ensures that the contribution of individual features to the model is balanced. It does not allow certain features to have an excessive impact on the model's output.

3.1.2.4 Model Prediction and Simulation

Using the optimal combination of parameters obtained in the model training phase and the trained XGBoost regression model, the sales volume in the new prediction dataset is predicted. The steps are as follows: input data: the new dataset X_new (containing classification code, year, month, day and day of the week), which has been processed by feature engineering and feature normalization, is passed to the model for prediction. Prediction process: the model will perform calculations based on the input feature values to arrive at the predicted value of sales volume (kilograms sold) y_new. The total daily replenishment and pricing strategy for the coming week (July 1-7, 2023) are predicted as shown in Table 3 and Table 4:

Table 3. Daily Pricing Strategies for the Week Ahead (July 1-7, 2023)

	Classification Code	Cost plus pricing
0	1011010101	1.98677097
1	1011010201	2.540965731
2	1011010402	2.477578093
3	1011010501	2.791882459
4	1011010504	3.310978402

Table 4. Daily Replenishment Strategy for the Week Ahead (July 1-7, 2023)

	Date of sale	Sort code	Volume (kg)
Sales Date	Classification Code	Sales volume (kg)	
0	2023-07-01	1011010101	202.426056
1	2023-07-01	1011010201	50.112465
2	2023-07-01	1011010402	50.112465
3	2023-07-01	1011010501	26.097593
4	2023-07-01	1011010504	100.265526

The predicted results are visualized as shown in Figure 1:

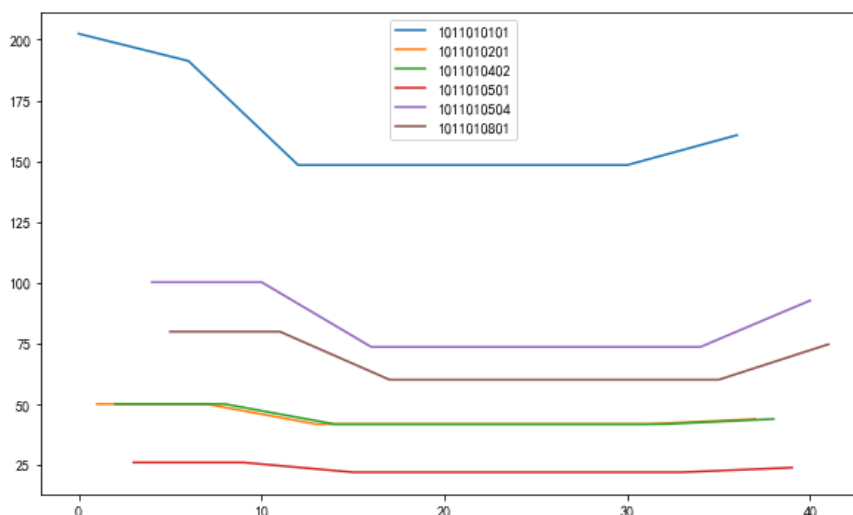


Figure 1. Line graph of the results of the prediction of vegetable pricing and replenishment for the coming week

The model simulation results are shown in Figure 2:

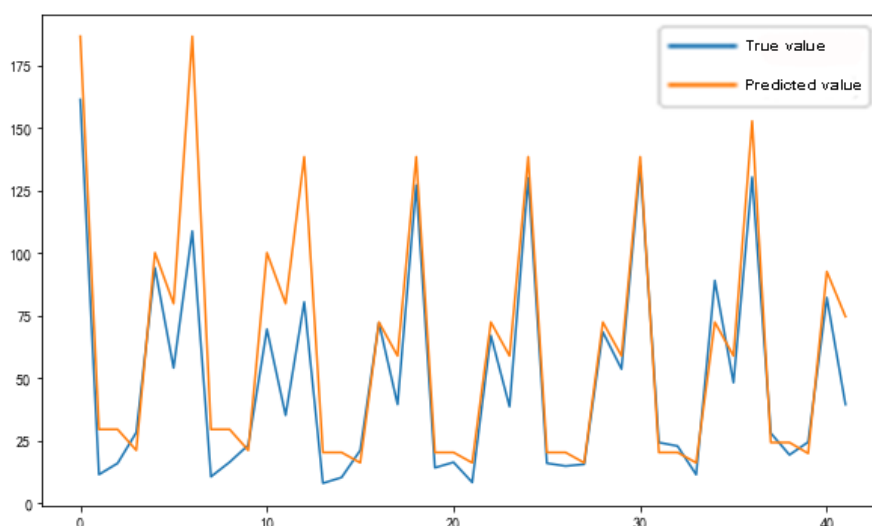


Figure 2. Comparison of model predicted values and real values

The prediction model performance parameters in this paper are shown in Table 5:

Table 5. XGBoost prediction model performance parameters

Training set	Test set			
Accuracy	R^2	MAE	RMSE	Accuracy
0.705	0.744	14.151	21.077	0.744

As can be seen from the table:

The R^2 score of 0.744 indicates that the model explains the target variable relatively well. The MAE score of 14.151, indicates that the mean absolute error of the model is about 14.151, the mean prediction error of the model is about 21.077, the training set accuracy is 0.705 and the test set accuracy is 0.744. The model has a relatively high classification accuracy on the test data, about 74.4% of the test samples were correctly classified. Taken together, the model performs relatively well on the test data and is able to predict the target variable more accurately.

3.2. Mixed Integer Linear Programming (0-1 MILP) Optimized Ordering Decision Model for Vegetable Category

3.2.1 XGBoost model training and prediction

The model building and training is basically similar to the process described above, the goal of the model is to predict the replenishment volume and pricing strategy of individual items on July 1, 2023 based on the vegetable sales data from June 24-30, 2023, replacing the dataset with features such as: date of sale, classification code, year, month, day, week, etc. to train the model in order to perform analysis and Prediction. Initial predicted replenishment volumes and pricing strategies for July 1, 2023 were obtained through the XGBoost model. The results of the predicted replenishment volume and pricing strategy for individual items are shown in Table.6 and Table.7:

Table 6. Day of July 1, 2023 Pricing Strategy

	Individual Product Codes	Cost plus pricing
0	102900005115250	6.002908
1	102900005115762	1.210498
2	102900005115779	2.383783
3	102900005115786	1.521768
4	102900005115823	2.431887

Table 7. July 1, 2023 Day Replenishment Strategy

	Date of sale	Category code	Sales volume (kg)
0	2023-07-01	102900005115250	10.322021
1	2023-07-01	102900005115762	11.384210
2	2023-07-01	102900005115779	12.258148
3	2023-07-01	102900005115786	15.197799
4	2023-07-01	102900005115823	8.416905

3.2.2 Mixed Integer Linear Programming (0-1 MILP) Optimized Ordering Decision Model for Vegetable Category [8-9]

3.2.2.1 Model Construction

The objective of this model is to determine which vegetable categories should be ordered, and the quantity to be ordered to maximize the total profitability in a given time horizon, and there are two sets of mixing variables in this problem: indicating whether or not to order the vegetable category i , which can take the value of 0 or 1. If $x_{[i]} = 1$, it means to order it; if $x_{[i]} = 0$, it means to not order it. Continuous variable $y_{[i]}$: denotes the number of ordering for vegetable category i . It can take any real value, but it is limited by some linear constraints. In this paper, a mixed integer linear programming (MILP) approach is used to solve this problem [10].

The model mainly includes the following key steps:

(1) Data preparation: the time period of June 24-30, 2023 was selected from the sales data for decision making. In this time period, the saleable individual items were extracted and the total sales volume, sales unit price, wastage rate, and wholesale price information were calculated for each vegetable category. These data are crucial for modeling and constraint definition.

(2) Determine the decision variables: the two main decision variables are x_i and y_i , which indicate whether to order vegetable category i and the quantity (kg) ordered, respectively. The types of these variables need to be determined according to the nature of the problem, x_i is a binary variable, y_i is a continuous variable.

(3) Determine the objective function: the goal of this paper is to maximize the total profit, define the objective function, which consists of the sum of the product of the cost-plus pricing of each vegetable category and the quantity ordered. The mathematical expression of the objective function is as follows:

$$\sum_i \text{Cost} - \text{plus pricing}_i \times \text{Order Quantity}_i \quad (9)$$

Add constraints:

$$\begin{cases} 27 \leq \sum_i x_i \leq 33 & \text{Total number of goods ordered constraints} \\ y_i \geq 2.5 \times x_i & \text{Minimum Cargo Display Requirement} \\ y_i \leq \text{Sales Forecast}_i \times x_i & \text{Forecasted sales constraints} \end{cases} \quad (10)$$

Solve the model:

Use a linear programming solver or a mixed integer linear programming solver to solve the model to find the optimal values of the decision variables to maximize the objective function. The final July 1, 2023 partial display of the vegetable unit replenishment and pricing strategy is shown in Table 8:

Table 8. Optimal vegetable individual item replenishment quantity and pricing strategy for July 1, 2023

	Item Code	Order Quantity	Cost Plus Pricing	Order Status
0	102900005115250	10.32	6.00	1
1	102900005115762	11.38	1.21	1
2	102900005115779	12.26	2.38	1
3	102900005115786	15.20	1.52	1
4	102900005115823	8.42	2.43	1
5	102900005115908	0.00	2.27	0
6	102900011006948	0.00	3.54	0

3.2.2.2 Model Simulation

The model simulation results are shown in Figure 3:

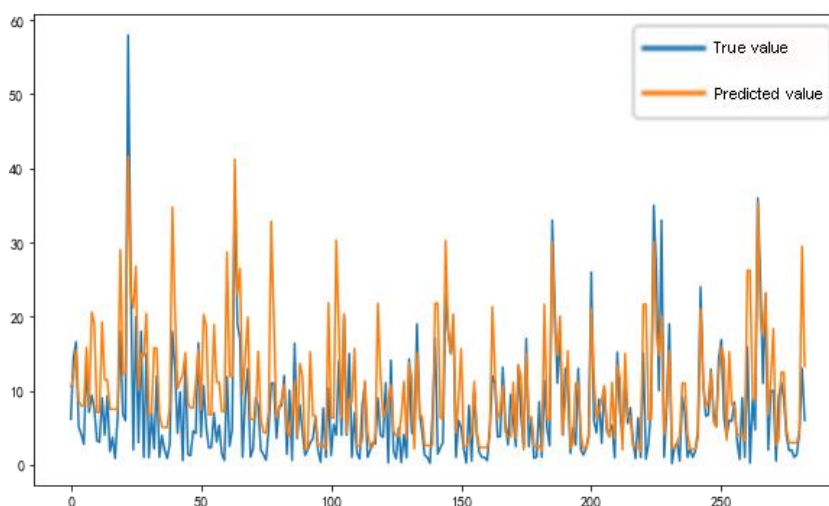


Figure 3. XGBoost model simulation results

The model performance evaluation metrics are shown in Table 9 below:

Table 9. Mixed integer linear programming optimized XGBoost prediction model performance parameters

Training set	Test set			
Accuracy	R^2	MAE	RMSE	Accuracy
0.705	0.331	4.323	6.152	0.613

As can be seen from the table:

The model is able to explain about 33.1% of the variance of the target variable, the average absolute error of the model is about 4.323, the average prediction error of the model is about 6.152, and the model is able to correctly predict about 61.3% of the test data. In summary, the model performs well on the training set but poorly on the test set, which may be due to overfitting because of the large amount of data or the complexity of the data relationships.

4. Conclusion

In this paper, we make full use of the historical sales data and information related to the sales flow details and wholesale prices of each item from July 1, 2020 to June 30, 2023, analyze the sales pattern and price influencing factors, and use the XGBoost prediction model to predict the replenishment plan and pricing strategy for the coming week (July 1-7, 2023); considering the relationship between supply and demand, we use the mixed integer linear programming method to ensure that the total number of items on sale can be controlled at 27-33 and the order quantity of each item meets the minimum display quantity of 2.5kg under the constraints. Taking into account the supply and demand relationship, the mixed integer linear programming method can ensure that the total number of single products on sale is controlled at 27-33, and the order quantity of each single product meets the constraint of the minimum display quantity of 2.5kg. Based on the data of the available varieties from June 24-30, 2023, the replenishment quantity of each product and pricing strategy for July 1 are calculated to maximize the revenue of the superstore on the condition of meeting the market demand for each category of vegetable products as much as possible. The model can be adjusted to different time periods and data to adapt to market changes and fluctuations in demand. However, the use of the model and linear programming methods in this paper requires a large amount of historical data and related information for training and solving, and the performance of the model may be degraded if the data is insufficient or of low quality. The model combines a number of factors is more complex and requires more computational resources and time for training and solving. XGBoost model and mixed integer linear programming method both involve some parameter settings and adjustments, different parameter choices may lead to different results and need to be carefully optimized. For non-specialists, the complexity of the model may lead to difficulties in understanding and interpreting, which may affect the superstore manager's trust in the decision.

References

- [1] Xiupeng S, Yiik D W, Michael Z L, Chandrasekar P, Chen C, et al. A feature learning approach based on XGBoost for driving assessment and risk prediction. [J], Accident; analysis and prevention, 2019, 129 (): 170-179.
- [2] Giulia Di Teodoro, Marta Monaci, Laura Palagi. Unboxing Tree ensembles for interpretability: A hierarchical visualization tool and a multivariate optimal re-built tree [J], arXiv (Cornell University), 2024, 12.
- [3] Djordje Dugošija, Aleksandar Savić, Zoran Maksimović. A new integer linear programming formulation for the problem of political districting [J], Annals of operations research, 2020, 288 (1): 247-263.
- [4] Milon B, Atanu S, Md. A I, Arafat J N, Shamim A, et al. Predicting Stock Market Price: A Logical Strategy using Deep Learning [J], 2021 IEEE 11th IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE), 2021: 218-223.

- [5] Wei D, Yimiao H, Barry L, Guowei M, et al. XGBoost algorithm-based prediction of concrete electrical resistivity for structural health monitoring [J], Automation in Construction, 2020, 114 ().
- [6] Haonan Sun, Bocheng Zhang, Nian Liu. Imitation learning-based online optimal scheduling for microgrids: An approach integrating input clustering and output classification [J], IET RENEWABLE POWER GENERATION, 2023.
- [7] Amal Asselman, Mohamed Khaldi, Souhaib Aammou. Enhancing the prediction of student performance based on the machine learning XGBoost algorithm [J], Interactive Learning Environments, 2021: 1-20.
- [8] Leilei M, Chaoyong Z, Yaping R, Biao Z, Chang L, et al. Mixed-integer linear programming and constraint programming formulations for solving distributed flexible job shop scheduling problem [J], Computers & Industrial Engineering, 2020, 142 (C): 106347.
- [9] Mauricio Granada-Echeverri, Eliana M. Toro, Jhon Jairo Santa. A Mixed Integer Linear Programming Formulation for The Vehicle Routing Problem with Backhauls [J], International Journal of Industrial Engineering Computations, 2019, 10 (2): 295-308.
- [10] Pei-Ying W, Chiao-Ting C, Jain-Wun S, Wang T, Szu-Hao H, et al. Deep Learning Model for House Price Prediction Using Heterogeneous Data Analysis Along with Joint Self-Attention Mechanism [J], IEEE Access, 2021, 9: 55244-55259.