Research on Automatic Pricing and Replenishment Decision of Vegetable Products Based on Machine Learning

Zitao Lin[1, *], Ziying Tan[2], Hongjie Ye[3], Man Sun[2]

[1]Faculty of Mathematics and Computer Science, Guangdong Ocean University, Zhanjiang, China, 524088

[2]Zhongge School of Arts, Guangdong Ocean University, Zhanjiang, China, 524088

[3]Faculty of Chemistry and Environment Science, Guangdong Ocean University, Zhanjiang, China, 524088

*Corresponding author: 17708669112@163.com

Abstract. In fresh food supermarkets, due to the short shelf life of vegetable products and the deterioration of products over time, unsold products on the same day cannot be sold the next day. Therefore, supermarkets need to make automatic pricing and replenishment decisions on a daily basis based on historical sales and demand. All product data for this study were provided by the 2023 Higher Education Society Cup National College Student Mathematical Modeling Competition.This study aims to explore the relationship between commodity sales and wholesale prices and pricing using five machine learning methods: neural networks, linear regression, decision tree regression, random forest regression, and XGBOOST regression, in order to predict sales and wholesale prices for the next week. By comparing the performance of five machine learning methods, it was found that the XGBOOST regression prediction model is the most suitable for predicting the daily sales of goods. The linear regression prediction model is the most accurate for predicting the wholesale prices of "chili", "flower and leaf", "edible fungi", "cauliflower", and "eggplant" goods, while the random forest regression prediction model is the most accurate for predicting the wholesale prices of "aquatic rhizome" goods. By constructing a dynamic programming model and using genetic algorithms to solve, automatic pricing and replenishment decisions for various goods within the next week can be obtained.

Keywords: Machine Learning, Predictive Models, Dynamic Programming Models, Genetic Algorithms.

1. Introduction

The shelf life of vegetable products is relatively short, so it is of great significance for supermarkets to make daily automatic pricing and replenishment decisions based on the historical sales and demand of each product to reduce food waste and increase supermarket profits.[1] By analyzing the historical sales data and demand situation of goods, supermarkets can more accurately predict marketdemand, thereby avoiding excessive procurement and inventory backlog, and reducing waste of grain and other commodities.[2] This precise predictive ability directly affects the subsequent security checks and analysis results of the supermarket, and is of great significance for achieving dynamic estimation, load scheduling, and cost reduction [3-4].

The supermarket may have used imprecise pricing and replenishment methods in the past, such as experiential pricing, fixed cycle replenishment decisions, and reliance on manual judgment for replenishment.[5] However, by applying advanced technologies and algorithms, supermarkets are gradually shifting towards more intelligent management methods.[6] For example, using machine learning and data mining techniques, supermarkets can establish models to identify sales patterns and demand trends of goods, thereby accurately predicting market demand every day. In addition, supermarkets can also use automated systems to automatically adjust product pricing strategies and replenishment decisions based on real-time monitoring and analysis, in order to maximize profits.[7]

This article uses five machine learning methods: neural network, linear regression, decision tree regression, random forest regression, and XGBOOST regression to make daily automatic pricing and

replenishment decisions based on the historical sales and demand of various commodities.[8] Propose dynamic programming models for various commodities, and finally solve these models using genetic algorithms to obtain automatic pricing and replenishment decisions for various commodities in the next week.[9-10]

2. Basic principles of five regression prediction models in machine learning

2.1. Neural Network Regression

A prediction method using artificial neural networks to establish a nonlinear relationship between input and output. It analyzes data correlations to make predictions. It simulates the information transmission in the human brain through connections and weights between nodes (neurons). Each neuron processes input from the previous layer, weighted by its corresponding weights, and applies an activation function to generate its output. These outputs are passed to the next layer, ultimately producing the network's output.

Neural network regression typically consists of three main components:

(1) Input layer: Accepts input data and passes it to the hidden layer. Each input feature corresponds to an input node.

(2) Hidden layer: composed of multiple neurons, which can have one or more hidden layers. The hidden layer is used to process input data and extract features.

(3) Output layer: Generate the final prediction result based on the output of the hidden layer. The number of nodes in the output layer depends on the requirements of the problem.

The mathematical formulas in neural network regression involve the following aspects:

Linear combination of input and weight:

$$Z^l = w^l x^{l-1} + b^l \tag{1}$$

Activation function (usually using nonlinear functions):

$$a^l = f(Z^l) \tag{2}$$

Among them, represents the output of the input layer, represents the weight matrix, represents the deviation vector, represents the activation function, represents the input of the hidden layer, and represents the output of the hidden layer.

Forward propagation: For node calculation between input layer and hidden layer:

$$Z^L = w^L a^{L-1} + b^L \tag{3}$$

$$\hat{y} = f(Z^L) \tag{4}$$

In summary, the symbols represent output from the previous layer, weight matrix, deviation vector, activation function, input to the output layer, and the network's predicted output. Backpropagation updates weights by calculating the gradient of the loss function, typically MSE or MAE, on the weights. Using MSE as the loss function, the loss for each training sample (x, y) is defined as:

$$L = (y - \hat{y})^2 \tag{5}$$

Among them, y is the true value and the predicted value of the neural network. By calculating the gradient of the loss function on the weight, we can use gradient descent or other optimization algorithms to update the weight to minimize the loss function.

2.2. Linear regression

Linear regression is a statistical model for continuous variables, assuming a linear relationship between target (dependent) and input (independent) variables. It aims to minimize prediction errors by finding the best-fitting line (or hyperplane) using the least squares method, which minimizes the squared differences between predicted and actual values. The model comprises:

(1) Input variable (feature): represents the characteristics or attributes of the observed sample.

(2) Weight (coefficient): used to measure the contribution of each feature to the target variable.
(3) Bias (intercept): used to adjust the difference between the predicted value and the target variable.
(4) Target variable (dependent variable): represents the continuous variable to be predicted.

For simple linear regression (with only one input variable), its mathematical formula can be expressed as:

$$y = b_0 + b_1 x \tag{6}$$

In multiple linear regression, y represents the predicted target variable, x is the input variable, b is the bias (intercept) term, and w is the weight (coefficient) term. The mathematical formula is:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n \tag{7}$$

In linear regression, y is the predicted target variable, x is the input variable, b is the bias term, and w is the weight term indicating the influence of x on y. The model aims to minimize prediction errors by adjusting w and b. The least squares method is commonly used to determine the optimal w and b values.

2.3. Decision Tree Regression

Decision tree regression uses decision tree models to predict continuous variables. It constructs a decision tree to segment input features and estimate target variable values at each leaf node. This method assumes regularity in target variable means and variances within each feature's value range.

The decision tree regression model consists of the following structures:

(1) Decision Node (Internal Node): Segments are performed based on a certain condition of the input feature to determine the flow of samples to different sub nodes.
(2) Leaf node (terminal node): A node without further segmentation, providing predicted values for the sample.
(3) Segmentation condition: The decision node performs binary segmentation based on the value of a certain feature, dividing the sample into different sub nodes.
(4) Predicted value: The estimated value of the target variable given by the leaf node.

The mathematical formula for decision tree regression can be expressed as:

$$y = f(x) \tag{8}$$

In decision tree regression, y represents the predicted target variable, x is the input feature value, and the prediction function is obtained by segmenting x using the decision tree model. The prediction process involves traversing the decision tree from the root node, following segmentation conditions, and ultimately reaching a leaf node. The predicted value at the leaf node serves as the estimated target variable value. This entire process relies on the trained decision tree model.

2.4. Random Forest Regression

Random forest regression combines multiple decision trees for regression prediction. It randomly selects samples and features to build trees and averages their predictions. This method incorporates randomness in both sample selection and feature evaluation, increasing model diversity. The model consists of multiple decision trees, a prediction rule, and an integration method. Random forest regression lacks a single mathematical formula.However, the prediction process of random forest regression can be represented by the following methods:

$$y = f(x) = \frac{1}{n}\sum(f_i(x)) \tag{9}$$

In random forest regression, y represents the predicted target variable, x is the input feature value, is the estimated target value from the i-th tree, and n is the total number of trees. Predictions are based on averaging or voting from all trees, which mitigates overfitting and enhances model stability and generalization.

2.5. XGBOOST regression

XGBoost (eXtreme Gradient Boosting) regression is an ensemble learning method based on gradient lifting algorithms used to establish regression models. It improves on the basis of the decision tree by optimizing the objective function and introducing regularization techniques to improve the performance of the model. Its principle is based on the following key ideas:

(1) Loss function optimization: Using gradient lifting algorithm, iteratively optimize the loss function to gradually build a powerful regression model.

(2) Decision Tree Integration: Based on the decision tree, continuously adjusting the prediction results of the model by serially adding more decision trees.

(3) Regularization: Introducing regularization terms to avoid overfitting and control the complexity of the model.

The XGBoost regression model consists of the following structures:

(1) Multiple Decision Trees: A weak classifier based on decision trees that iteratively constructs multiple decision tree models.

(2) Guided sampling: By randomly selecting subsets of training samples and features, sampling with dropout is performed to construct each decision tree.

(3) Prediction rule: Each decision tree is segmented based on input features and an estimate of the target variable is given at the leaf nodes.

Loss function: By optimizing the loss function, the model is updated based on the difference between the predicted results and the actual values.

XGBoost regression is based on the gradient lifting algorithm, where the loss function uses second-order Taylor expansion. Specifically, for the t-th iteration, the objective function of XGBoost regression can be expressed as:

$$Obj(t) = \sum(L(y_i, \hat{y}_i(t-1))) + \Omega(f_t) \tag{10}$$

The objective function at the t-th iteration,, consists of the loss function , which measures model fit, and a regularization term to control complexity. True and predicted values for the i-th sample after t-1 iterations are and, respectively. The decision tree model from the t-th iteration is. By continuously optimizing the objective function, the model gradually improves prediction performance to achieve the optimal regression model.

3. Results and Discussion

3.1. Establishment of Regression Prediction Simulation Model

Use these five methods to predict the relationship between Y (daily sales) and X (pricing and wholesale prices) for various commodities. After experiments, the average absolute percentage error and fitting degree of the test set for these prediction algorithms are shown in the Table 1 to 2:

Table 1. Percentage Error

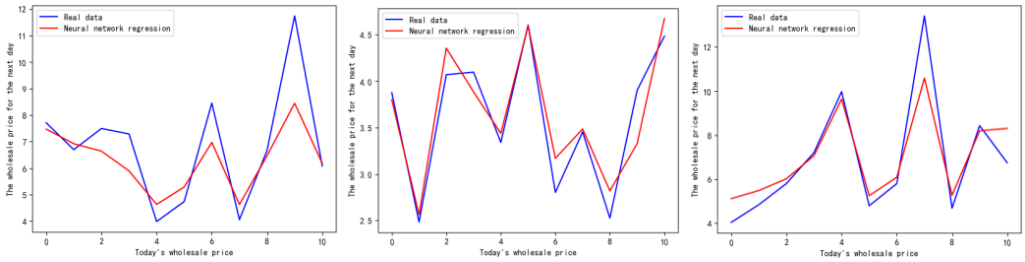
Methods	peppers	Florifolias	Aquatic rhizomes	Edible fungi	florescent vegetables	Solanaceae
neural network	54.687	44.175	121.647	66.05	73.865	86.263
linear regression	54.591	42.816	121.374	59.561	73.622	86.894
Decision Tree Regression	0	0	0	0	0	0
Random Forest Regression	19.39	18.725	40.212	23.995	25.218	26.29
XGBOOST regression	14.546	12.018	29.903	16.972	20.867	19.543

Table 2. Degree of Fit

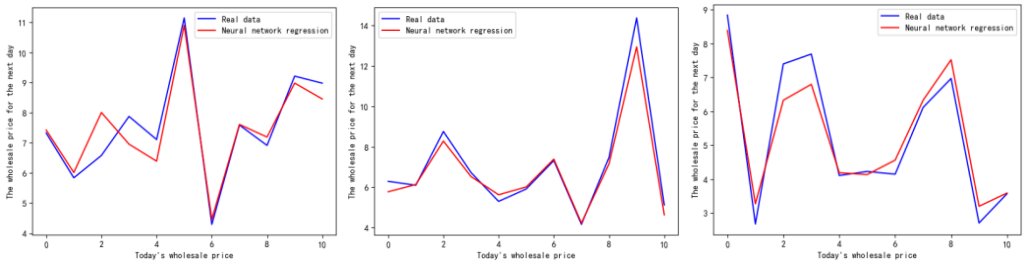
Methods	peppers	Florifolias	Aquatic rhizomes	Edible fungi	florescent vegetables	Solanaceae
neural network	0.038	-0.256	0.097	0.013	0.116	0.017
linear regression	0.023	0.056	0.123	0.099	0.124	0.033
Decision Tree Regression	1	1	1	1	1	1
Random Forest Regression	0.86	0.826	0.871	0.861	0.865	0.877
XGBOOST regression	0.944	0.931	0.946	0.939	0.936	0.948

3.2. Establishment of Time Series Prediction Simulation Model

Build a time series prediction model using these five methods and test the fitting degree of various products.The prediction and testing of the fitting degree of each regression model for various products are shown in Fig 1 to 5.

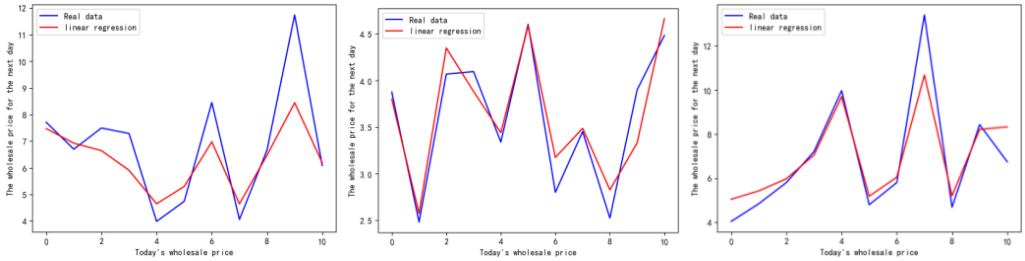


(a) Fit of chili products (b) Fit of flower and leaf products (c) Fit of aquatic rhizome

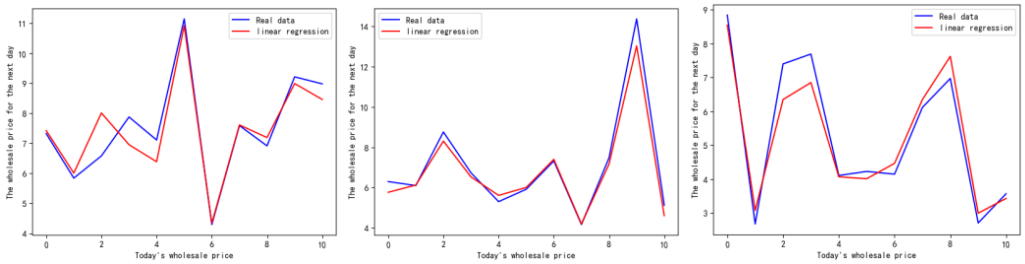


(d) Fit of edible fungi (e) Fit of cauliflower products (f) Fit of eggplant products

Fig 1. Fit Analysis of Neural Network Regression Processing Data

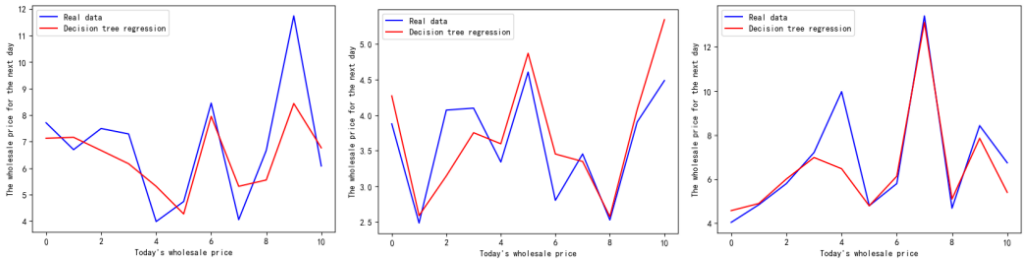


(a) Fit of Pepper Products (b) Fit of Flower and Leaf Products (c) Fit of Aquatic Rhizome Products

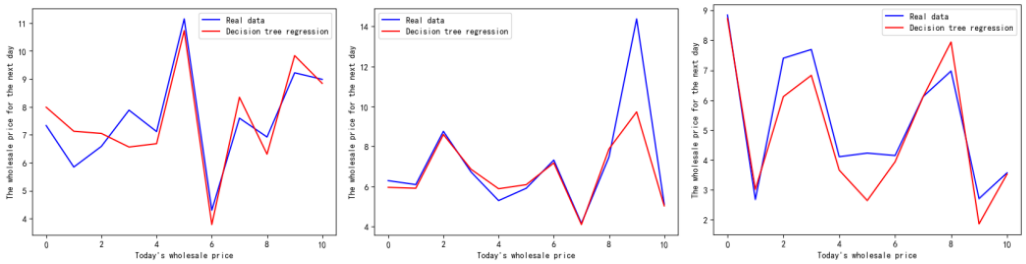


(d)Fit of Edible Fungi Products (e)Fit of cauliflower products (f)Fit of eggplant products

Fig 2. Fit Analysis of Linear regression Processing Data

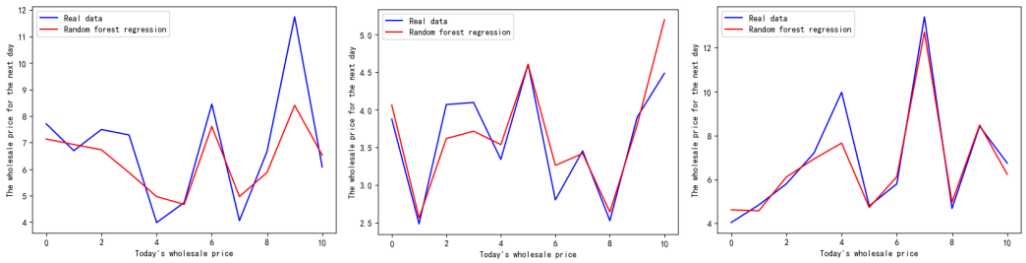


(a) Fit of Chili Pepper Products (b) Fit of Flower and Leaf Products (c) Fit of Aquatic Rhizome Products

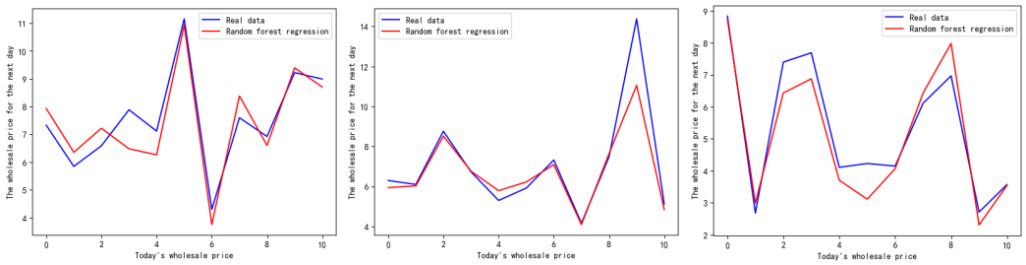


(d) Fit of Edible Fungi Products (e) Fit of cauliflower products (f) Fit of eggplant products

Fig 3. Fit Analysis of Decision Tree Regression Processing Data

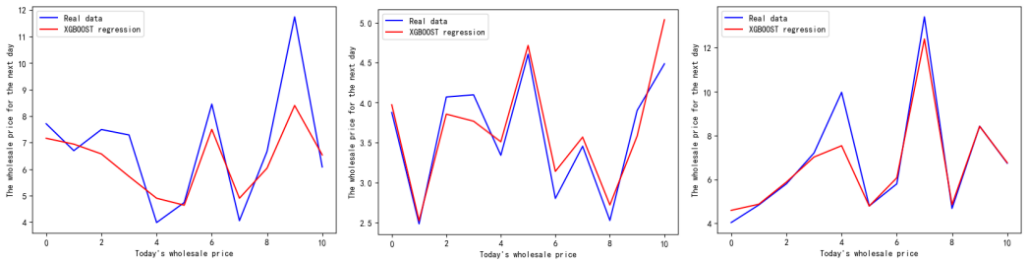


(a)Fit of chili products (b) Fit of flower and leaf products (c) Fit of Aquatic Rhizome Products

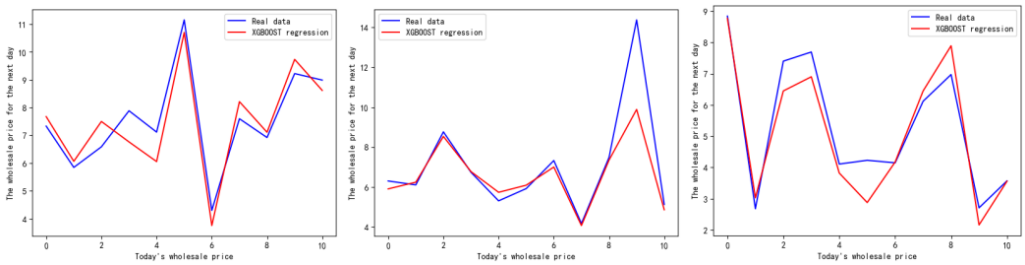


(d) Fit of Edible Fungi Products (e) Fit of cauliflower products (f) Fit of eggplant products

Fig 4. Fit Analysis of Random Forest Regression Processing Data



(a) Fit of chili products (b) Fit of flower and leaf products (c) Fitting Degree of Aquatic Rhizome Products



(d) Fitting Degree of Edible Fungi Products (e) Fit of cauliflower products (f) Fit of eggplant products

Fig 5. Fit Analysis of XGBOOST Regression Processing Data

3.3. Analysis of experimental results

(1) Model selection

The XGBOOST regression prediction model is the most suitable for predicting the sales volume of products on that day; for predicting the wholesale prices of goods in the next week, the linear regression prediction model is the most accurate for predicting "chili", "flower and leaf", "edible fungi", "cauliflower", and "eggplant" goods. The random forest regression prediction model is the most accurate for predicting "aquatic rhizome" goods. The following are the wholesale prices of various goods in the next week.As shown in Table 3.

Table 3. Wholesale Price Forecast for Various Commodities in the Next Week

Time	peppers	Florifolias	Aquatic rhizomes	Edible fungi	florescent vegetables	Solanaceae
1	4.67	3.26	11.52	6.84	7.96	4.96
2	4.71	3.32	10.78	7.08	7.89	5.01
3	4.75	3.37	10.48	7.25	7.83	5.04
4	4.79	3.41	9.22	7.38	7.77	5.07
5	4.83	3.45	9.65	7.48	7.71	5.11
6	4.87	3.49	9.07	7.55	7.65	5.13
7	4.91	3.52	8.02	7.61	7.59	5.16

(2) Automatic pricing and replenishment decision-making

Next, we construct dynamic programming models for various commodities, with the objective function being: maximum profit=(pricing wholesale price) * sales volume * (100 loss rate)/100; The constraint condition is: pricing>wholesale price; The required values for the solution are pricing, sales volume, and maximum profit.

By solving the dynamic programming model using genetic algorithm, we obtained the pricing, sales volume, and maximum profit of various products in the next week, that is, automatic pricing and replenishment decisions for various products. As shown in Table 4 to 5.

Table 4. Automatic pricing of various products:

Time	peppers	Florifolias	Aquatic rhizomes	Edible fungi	florescent vegetables	Solanaceae
1	11.89	8.32	14.62	11.93	12.52	10.29
2	13.83	8.32	13.14	11.29	13.56	10.95
3	12.33	15.22	14.62	9.94	13.56	7.49
4	20.79	18.01	12.75	11.44	13.56	7.49
5	19.07	18.88	13.14	16.56	13.56	10.27
6	22.93	8.31	13.14	11.29	12.32	9.66
7	13.82	8.31	11.24	11.29	11.04	7.49

Table 5. Replenishment decisions for various commodities:

Time	peppers	Florifolias	Aquatic rhizomes	Edible fungi	florescent vegetables	Solanaceae
1	178.71	510.11	37.63	175.47	58.19	32.11
2	229.17	508.95	109.65	417.65	61.91	24.52
3	305.26	309.45	51.15	204.15	63.83	39.41
4	415.78	258.55	66.67	160.26	66.28	37.85
5	223.07	275.13	131.81	101.26	83.02	18.38
6	485.81	478.37	148.37	218.67	66.89	30.39
7	262.51	431.77	77.72	210.34	74.25	38.04

4. Conclusion

Machine learning plays a crucial role in supermarket pricing and replenishment decisions, helping to optimize pricing, demand forecasting, inventory management, and real-time decision-making, thereby improving sales revenue and customer satisfaction. Different machine learning methods have their own advantages and limitations: neural networks are suitable for complex problems, but require a large amount of computational resources; linear regression is simple and fast, suitable for simple linear relationships; decision tree regression is easy to understand but easy to overfit; random forests reduce overfitting risks by integrating multiple decision trees, but the training time is longer; XGBOOST has high predictive accuracy and generalization ability, suitable for large-scale datasets and high-dimensional feature spaces.

References

[1] Liu Yuchen. Research on the Joint Decision Problem of Dynamic Pricing and Inventory Control [D]. Shanghai Maritime University.

[2] Zhao Yaya, Zhang Zeren, Dai Yongfu. Prediction of Agricultural Product Sales Range Based on Machine Learning [J]. The Computer Age, 2020(7): 4.

[3] N. Ahmed Mahoto, R. Iftikhar, A. Shaikh, Y. Asiri, A. Alghamdi et al., "An intelligent business model for product price prediction using machine learning approach," Intelligent Automation & Soft Computing, vol. 30, no.1, pp. 147–159, 2021.

[4] Sujjaviriyasup T, Pitiruek K. Agricultural product forecasting using machine learning approach. [J]. International Journal of Mathematical Analysis, 2013, 7(37): 1869-1875.

[5] Chen Z, Goh H S, Sin K L, et al. Automated Agriculture Commodity Price Prediction System with Machine Learning Techniques [J]. 2021.

[6] Nicoleta T, Stavros S, Manos R. Data-Driven Decision Making in Precision Agriculture: The Rise of Big Data in Agricultural Systems [J]. Journal of Agricultural & Food Information, 2019.

[7] Wang Xinwu and Jin Peiyun. A Price Prediction Method for Agricultural Products in Longdong Based on Set Empirical Mode Decomposition and Autoregressive Neural Network [J]. Journal of Lanzhou Institute of Technology, 2022, 29(4): 85-88.

[8] Xu Qigang. A Price Prediction Model for Agricultural Products in Shandong Province Based on Improved KNN-BPNN Algorithm [D]. Jinan University,2015.

[9] Huang Yu, Chen Ting. Establishment of a BP neural network prediction model based on agricultural product price prediction [J]. New Business Weekly,2018.

[10] Kong Congcong. Research and implementation of agricultural product price prediction technology based on recurrent neural network [D]. University of Chinese Academy of Sciences.