mathematics

MDPI

Article

Price Prediction for Fresh Agricultural Products Based on a Boosting Ensemble Algorithm

Nana Zhang [1], Qi An [2], Shuai Zhang [3] and Huanhuan Ma [4,*]

[1] College of Economics & Management, Zhejiang University of Water Resources and Electric Power, Hangzhou 310018, China; zhangnn@zjweu.edu.cn
[2] School of Engineering, The Open University of China, Beijing 100039, China; anq@ouchn.edu.cn
[3] College of Business Administration, Capital University of Economics and Business, Beijing 100070, China; shuaizhang@cueb.edu.cn
[4] College of Economics and Management, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
* Correspondence: huanhuan_270@nuaa.edu.cn

Abstract: The time series of agricultural prices exhibit brevity and considerable volatility. Considering that traditional time series models and machine learning models are facing challenges in making predictions with high accuracy and robustness, this paper proposes a Light gradient boosting machine model based on the boosting ensemble learning algorithm to predict prices for three representative types of fresh agricultural products (bananas, beef, crucian carp). The prediction performance of the Light gradient boosting machine model is evaluated by comparing it against multiple benchmark models (ARIMA, decision tree, random forest, support vector machine, XGBoost, and artificial neural network) in terms of accuracy, generalizability, and robustness on different datasets and under different time windows. Among these models, the Light gradient boosting machine model is shown to have the highest prediction accuracy and the most stable performance across three different datasets under both long-term and short-term time windows. As the time window length increases, the Light gradient boosting machine model becomes more advantageous for effectively reducing error fluctuation, demonstrating better robustness. Consequently, the model proposed in this paper holds significant potential for forecasting fresh agricultural product prices, thereby facilitating the advancement of precision and sustainable farming practices.

Keywords: boosting ensemble learning algorithm; light gradient boosting machine; fresh agricultural products; price predictions

MSC: 62M10; 68M07

check for updates

Academic Editors: Jian Luo, Zheming Gao and Xin Yan

Received: 20 November 2024
Revised: 23 December 2024
Accepted: 26 December 2024
Published: 28 December 2024

Citation: Zhang, N.; An, Q.; Zhang, S.; Ma, H. Price Prediction for Fresh Agricultural Products Based on a Boosting Ensemble Algorithm. *Mathematics* **2025**, *13*, 71. https://doi.org/10.3390/math13010071

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The time series of agricultural prices is short and exhibits irregular cyclicality that can be linked to economic performance and productivity [1]. Future price, production, and consumption levels are relatively volatile and easily affected by various factors, complicating farmers' efforts to optimize production strategies, policymakers' development of effective agricultural policies, and traders' and consumers' ability to make informed decisions [2]. The scientific and effective forecasting method is conducive to accurately predicting the price of agricultural commodities, evading market risk, correctly guiding agricultural production, increasing agricultural income, and accomplishing government macroeconomic regulation [3–6]. Fresh agricultural products are distinct from general

agricultural commodities because they are regional, temporal, and perishable. In recent years, parts of China have witnessed the problem of fresh agricultural products being unsalable stock and a roller coaster ride of agricultural product prices [7]. Especially since the outbreak of COVID-19 in early 2020, the prices of fresh products such as vegetables, fruits, and aquatic products have been fluctuating significantly. Price prediction of fresh agricultural products has become a public concern, as well as a focus on management decision-making and scientific research. Therefore, it is of considerable importance to carry out scientific and accurate analysis and predict the prices of fresh agricultural products through intelligent and practical algorithms.

Nowadays, mainstream price forecasting methods widely applied to heterogeneous applications are quantitative forecasting methods. They refer to the econometric method, time series analysis method, and intelligent prediction method. In recent years, machine learning prediction methods have been increasingly applied to price prediction. The common methods refer to artificial neural networks (ANNs), support vector machines (SVMs), random forests (RFs), and decision trees (DTs). They are widely applied to the price prediction of oil, energy, stocks, and credit risk assessment [8–15] and compared with the performance of multiple datasets [16]. Luo et al. [9,11,12,14,15] have significantly improved SVMs and applied them effectively to power and credit prediction tasks. In contrast, machine learning prediction methods are rarely applied to price prediction of agricultural products [1,2,4,6,17–23], with the focus predominantly on elementary machine learning models such as ARIMA, ANN, SVM, and DT. Moreover, the existing research primarily concentrates on crops, grain products, and commodities [1,2,6,17,18,20,21], with only a few studies addressing the price prediction of specific fresh agricultural products like tomatoes [19], corn [22], and bananas [23]. In reality, the category of fresh agricultural products is extensive, encompassing fruits, vegetables, meats, and seafood, among others. There is a notable dearth of research that leverages a variety of machine learning models to predict the prices of multiple categories of fresh agricultural products and investigates which models within the machine learning paradigm are more suitable for fresh price prediction tasks. Additionally, using only simple machine learning models cannot overcome each model's inherent limitations, so the prediction results' performance and robustness still have opportunities to be improved. Machine learning, especially derived integrated learning, is better for dynamic data stream processing [24].

It is well known that short-term market price forecasting has been a difficult problem for a long time because of too many factors that cannot be accurately predicted [19]. In managing nonlinear data with complex factors and volatility, models like ARIMA, DT, RFs, SVMs, XGBoost, and ANNs have limitations. ARIMA struggles with strong nonlinear patterns and seasonality. DTs risk overfitting and are sensitive to data transformations. RFs face challenges in feature selection and computational demands. SVMs are sensitive to parameter choices and computationally intensive for large datasets. XGBoost may require careful parameter tuning and is less robust against outliers. ANNs require substantial computational resources, lack interpretability, and are prone to overfitting, mainly when dealing with noisy data. Price fluctuations of fresh agricultural products consistently exhibit repeated ups and downs, instability, and either benign or vicious cycles. These characteristics define highly nonlinear data with significant oscillations. However, there is a paucity of studies integrating machine learning techniques in fresh agricultural economics. Furthermore, existing methods such as ARIMA, ANN, SVM, and DT exhibit deficiencies when predicting the prices of fresh agricultural products. Consequently, researchers consistently face the challenge of selecting the most appropriate prediction method to enhance the accuracy and robustness of forecasts. Considering these factors, this paper selects and constructs a more suitable machine learning prediction method, the Light GBM (gradient

boosting machine) model, by comparing it with six other models in fresh agricultural applications. The objective is to determine which machine learning models are better suited for fresh price prediction tasks. The Light GBM model is based on the Boosting integrated learning algorithm and was first applied to the price prediction of fresh agricultural products. We also compare the Light GBM model with the traditional time series algorithm and state-of-the-art machine learning algorithm in the short-term, medium-term, and long-term time windows to verify the effectiveness of the Light GBM algorithm. A real daily dataset was used in numerical experiments, which contains the price of bananas, beef, and crucian carp from February 2019 to June 2019. The model proposed in this paper holds significant potential for forecasting fresh agricultural product prices, thereby facilitating the advancement of precision and sustainable farming practices.

The remainder of the study is organized as follows: Section 2 presents the modeling methodologies of different forecasting schemes. In Section 3, we describe the design structure for these six models. Real data on the prices of three fresh agricultural products are collected to verify different models. It also contains forecasting comparisons for one classical time series forecasting model and five machine learning models. Some practical implications and our research findings are addressed in Section 4

2. Materials and Methods

This section introduces seven price prediction methods to be used in this paper.

2.1. The ARIMA Method

ARIMA model stands for Autoregressive Integrated Moving Average model. It forecasts future trends based on a number of lagged observations of time series, which is the most common statistical model used for time series forecasting. The ARIMA method is popular for its simplicity as it requires only endogenous variables without exogenous variables being required. However, it also has some limitations. It assumes that the seasonal pattern in time series data is either inherently strong and stable (stationary) over time or becomes stable after being differentiated. Additionally, it can only capture linear relationships, making it unsuitable for modeling nonlinear dynamics.

2.2. The Decision Tree

The decision tree was proposed by Quinlan J. R. in 1986 [25], which is one of the basic machine learning methods. It is commonly used to solve a wide range of problems in classification and prediction. Decision trees adopt a tree structure that recursively partitions the feature space. In the case of classification, each node in the tree structure continues to split to maximize purity until the nodes only contain single-class samples. In the case of regression, the score and gain at each node of the tree structure are calculated to determine the best split. These pure nodes are called leaf nodes. When a test sample is input into a decision tree, it can be traced down to the leaf node. One disadvantage of the decision tree is that the model is unstable. The growth of the structure of the optimal decision tree is sensitive to a small change in the data, leading to imprecise predictions.

2.3. The Random Forest

Random forest (RF) is an advanced classification algorithm proposed by Breiman that uses multiple weak classifiers to form a strong classifier [26]. Random forest uses bootstrap (self-service method) resampling technology to replace samples from training samples to form different training sets. These training sets are used to train different decision trees [27]. Decision trees are voted according to the classification results to form the random forest. The random forest has a robust performance in dealing with missing and unbalanced data. However, its performance on classification and regression is desirable. The speed of model

training and classification regression is fast. Random forest has been applied in the fields of stock price forecasting, electricity price forecasting, and energy price forecasting. The convergence speed of random forests does not depend on the number of noise variables but is related to the number of strong features [28]. The prediction of RF is based on averaging the predictions of the individual trees. The formula is as follows:

$$\widehat{y_i} = \frac{1}{N}\sum_{i=1}^{N} h_i(x) \tag{1}$$

where $h_i(x)$ is the prediction of the i-th tree for input x.

2.4. The Support Vector Machine

Cortes and Vapnik proposed a support vector machine (SVM) based on statistical learning theory [29]. A support vector machine is a machine learning model that is applied to binary classification problems. SVM is based on the principle of structural risk minimization and uses kernel functions to map data from low-dimensional to high-dimensional space. It can effectively solve the dimensionality explosion and high computational complexity problems. Meanwhile, it shows great advantages in solving small sample, nonlinear, and high-dimensional problems. As to limitations, this method is sensitive to noise and is difficult to generalize to classification to more than two classes or regression problems. For regression, SVM is known as support vector regression (SVR). The prediction formula for a new input x is given by Equation (2):

$$f(x) = \sum_{i=1}^{vn_s} (\alpha_i - \alpha_i^*)K(x_i, x) + b \tag{2}$$

where vn_s is the number of support vectors, α_i and α_i^* are the Lagrange multipliers, $K(x_i, x)$ is the kernel function, and b is the bias term.

2.5. The XGBoost

XGBoost is an ensemble learning algorithm that combines the results of multiple weak learned through a gradient-boosting decision tree to obtain prediction results. It is initially from a research project [30]. The classification of ensemble learning is mainly based on the relationship between the individual learners that make up the ensemble algorithm, which can be roughly divided into three types: Boosting, Bagging, and Stacking. XGBoost is an optimized distributed gradient boosting algorithm. These simple individual learners may be trained to learn the dataset using different algorithms to obtain their own prediction results. All the predicted results are merged by using the following fusion strategy: After the current decision tree is approximated by a number of weak learners, the next additional learner is trained to correct the residuals in regression or errors in classification in the previous trees so they can be continuously reduced. The prediction result of the entire model is the sum of predictions from each weak learner. The implementation of the algorithm was widespread for the efficiency of computing time and memory resources. As compared to Light GBM, its disadvantages include no early stopping, slower training, and decreased accuracy.

The objective function XGBoost aims to minimize is as follows:

$$Obj = \sum_i l(y_i, \widehat{y_{Xl}}) + \sum_k \varphi(f_k) \tag{3}$$

where $l(y_i, \widehat{y_{Xl}})$ is the loss function that measures the difference between the true label y_i and the predicted label $\widehat{y_{Xl}}$, $\varphi(f_k)$ is the regularization term that penalizes the complexity of the model, with f_k representing the complexity of the k-th tree.

2.6. The Artificial Neural Network

Artificial neural network (ANN) methodologies represent computational paradigms that emulate the functional dynamics of neuronal interconnections in the human cerebral cortex [31]. These models encompass neurons, stratified architectures (input, concealed, and output layers), and adjustable parameters such as weights and biases. Their operational mechanism entails forward propagation and backpropagation processes; during forward propagation, input data is processed through layers using activation functions to introduce nonlinearity. The prediction formula for a regression problem at the output layer is as follows:

$$\widehat{y} = f(\sum_{i=1}^{n} \omega_i x_i + b_A) \tag{4}$$

where $\widehat{y}$ is the predicted output, ω_i represents the weight, x_i represents the inputs, and b_A is the bias term.

During training, artificial neural networks are refined to enhance their functional efficacy. This approach is extensively employed across various domains, including image recognition, natural language processing, speech recognition, and recommendation systems, leveraging its robust autonomous learning capabilities and proficiency in addressing nonlinear complexities. Nonetheless, it confronts challenges such as the high computational expenses of training and the limited interpretability of the resulting models.

2.7. The Light GBM

The Light GBM algorithm is a gradient decision tree-based framework proposed by Microsoft in 2017 [32]. It is a state-of-the-art gradient boosting framework optimized for distributed computing and excels in various machine learning tasks, including ranking, classification, and prediction. The algorithm's key innovations lie in its histogram-based approach, leaf-wise tree growth strategy, native support for categorical features, and feature optimization techniques.

The histogram-based approach discretizes continuous features into k bins, significantly reducing memory usage and computational complexity, as illustrated by the following formula:

$$x_{Li} \rightarrow histogram_bin(x_{Li}, k) \tag{5}$$

where x_{Li} represents the value of the i-th feature, $histogram_bin(x_{Li}, k)$ is a function that discretizes a continuous floating-point feature value x_{Li} into one of k integer values. This process is called histogram binning. Histogram binning is a key step in the Light GBM algorithm, which simplifies data representation by converting continuous feature values into discrete integers, helping to improve the efficiency and accuracy of the algorithm. k represents the number of bins in the histogram, which determines how many different intervals the data will be divided into.

Light GBM generates trees by leaf-wise strategy, selecting the node with the most significant gain from the current leaf for segmentation, loops, and iterates, and growing a deeper decision tree. This strategy is formulated as follows:

$$node_to_split = argmax_{j \in leaves}(gain(j)) \tag{6}$$

where $node_to_split$ refers to the node that is to be split, and $gain(j)$ represents the gain of the j-th leaf node.

In order to prevent overfitting, a threshold is introduced to limit and prevent the occurrence of overfitting problems. In the case of the same number of splits, the leaf-wise strategy can reduce errors and obtain better accuracy. The leaf growth strategy is shown in Figure 1.

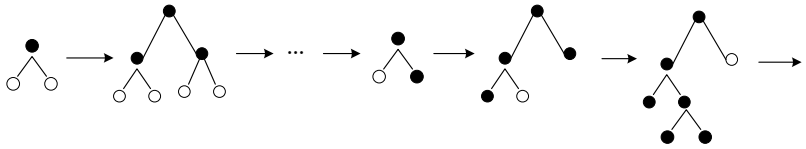


Figure 1. Leaf growth strategy.

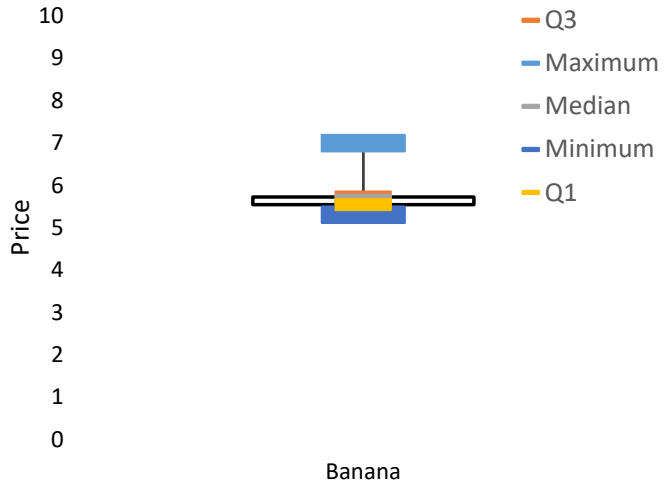
Light GBM has the advantages of high training efficiency, low memory usage, high accuracy, support for parallel and GPU, and the ability to handle large-scale data. It has become a new hotspot in various types of machine learning research and is a common sharp weapon in many application fields.

3. Numerical Experiments

Empirical studies using machine learning commonly have two main phases. The first one addresses the selection of relevant variables and models for the prediction, separating a portion of the data for the training and validation of the models, thus optimizing them. The second phase applies the optimized models to the data intended for testing, thus measuring predictive performance.

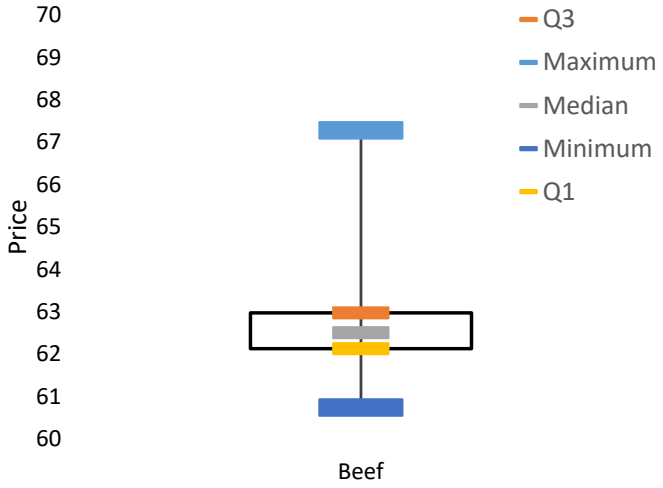
3.1. The Data Source

To evaluate the performance of the Light GBM model in the price prediction for three types of fresh agricultural products, the datasets that we consider in our analysis are the daily average prices of bananas, beef, and crucian carp. These types of food are selected from the main fresh agricultural products, including fruits, vegetables, meat, and seafood, respectively. The time series is on a daily basis and covers the time period from 1 February 2019 to 30 June 2019, as shown in Figure 2. The data source is published by the official website of the Ministry of Agriculture and Rural Affairs of the People's Republic of China. The research data is authoritative and reliable. At the same time, the research data is public and can be accessed and obtained through the URL http://www.moa.gov.cn/ (accessed on 30 June 2023). Research data can be accessed and listed systematically by anyone at any time.

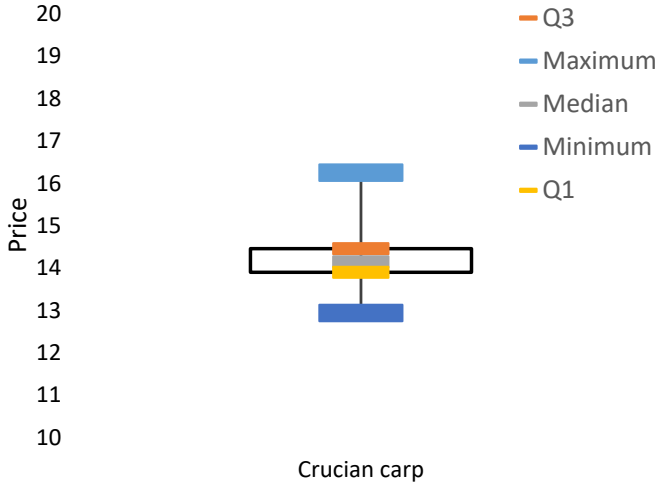


(**a**) Boxplot for the daily average wholesale price of bananas.

Figure 2. *Cont.*



(**b**) Boxplot for the daily average wholesale price of Beef.



(**c**) Boxplot for the daily average wholesale price of crucian carp.

Figure 2. Boxplots for the daily average wholesale price of fresh agricultural products.

The three box plots depicted in Figure 2 correspond to the daily average wholesale prices of bananas, beef, and crucian carp, respectively. These plots are constructed based on the calculated values of maximum, minimum, median, the first quartile (Q1), and the third quartile (Q3) for each fresh good. Given the substantial price variability among these items, the vertical axis range was adjusted to 10 to facilitate a comparative analysis of pricing trends. Analyzing Figure 2 reveals that the sequence of interquartile ranges (IQR = Q3 − Q1) for bananas, crucian carp, and beef is bananas, crucian carp, and beef, respectively. Notably, beef exhibits the widest IQR, suggesting a more dispersed price distribution compared to the other two commodities.

3.2. The Problem Statement

To construct a prediction model for the prices of fresh agricultural products. We assume that the price on day d is a function of the price of the past t days; the function is as shown in Equation (7).

$$P_d^k = F_M(P_{d-1}^k, P_{d-2}^k, \cdots, P_{d-t}^k) \tag{7}$$

where p_d^k is the output value, which represents the price of product k on day d. k represents the type of fresh agricultural products, and t indicates any integer less than d. In this model, we want to make predictions based on the previous prices in the past week, and thereby, we set $t = 7$. p_{d-1}^k , which represents the price of the product k on day d-1 and so

on. F_M represents the function of different prediction models. M represents the underlying model type.

The basic steps for using these models to predict the prices of fresh agricultural products are illustrated in Figure 3.

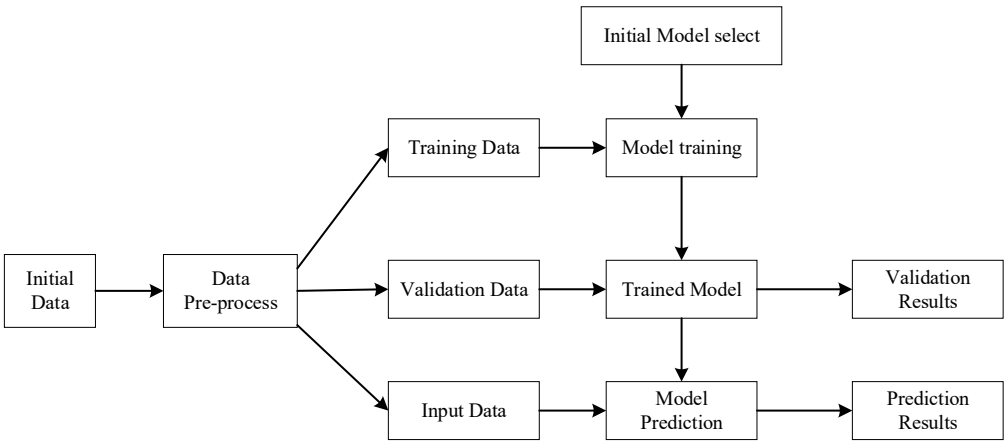


Figure 3. Basic flow chart of price prediction of fresh agricultural products.

In general, most machine learning algorithms do not directly support category features but instead convert category features into one-hot encoding, that is, multi-dimensional 0/1 features, which increase the space and time requirements during model training. In contrast, the Light GBM model directly supports category features without conversion and adds the decision rules of category features to the decision tree algorithm. The calculation process of the Light GBM algorithm is shown in Figure 4.

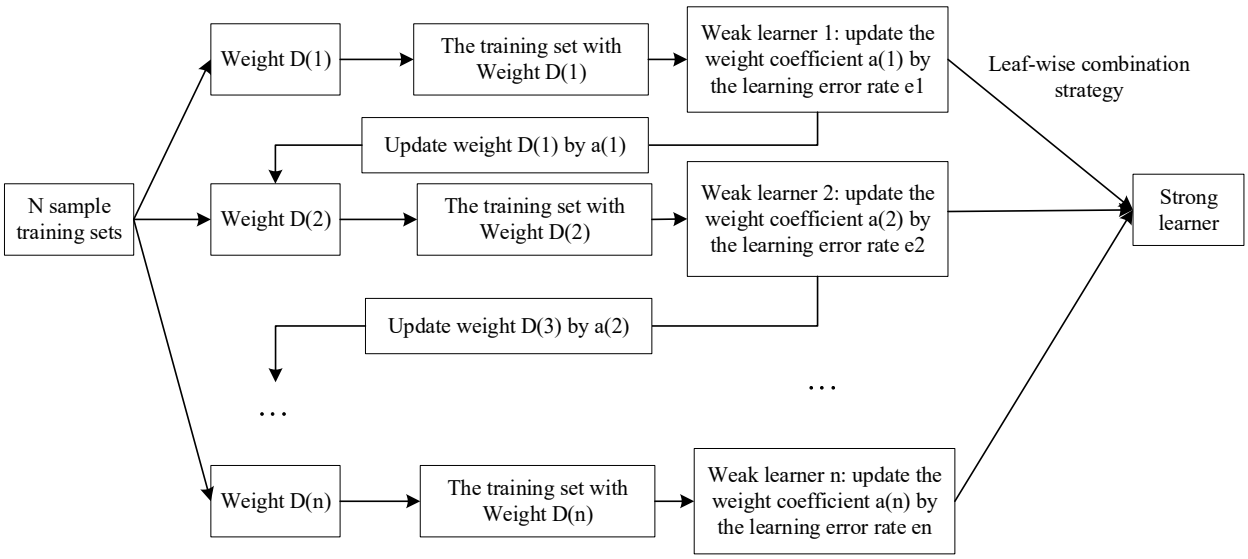


Figure 4. The prediction process of the Light GBM method.

Then, we build the Light GBM model in the following steps:

Step 1: Initialize the weights of the training samples so that each sample has the same weight, and train these samples to obtain the initial learner.

Step 2: Calculate the classification error rate of these samples and the coefficient of the learner. Adjust the weight of the samples according to the misclassification rate so that these misclassified samples can obtain more attention from subsequent learners. The adjustment is conducted in a way that the weight of the correctly classified sample is reduced, and the weight of the wrongly classified sample is increased.

Step 3: Repeat the second step for N loops to obtain N learners. Integrate these N learners with different weights through the leaf-wise tree growth strategy to obtain a strong learner.

3.3. The Method Comparison

This paper compares a wide range of methods for predicting the price of bananas, beef, and crucian carp, including a classical time series forecasting method ARIMA and six machine learning methods (DT, RF, SVM, XGBoost, ANN, and Light GBM). All these models are coded in Python programming language and evaluated in terms of accuracy (Supplementary Materials).

3.4. The Parameters Evaluation

In machine learning, hyperparameters are parameter values that are set before the learning process begins. Since the selection of hyperparameters directly impacts model performance, how to select and optimize hyperparameters is essential. In order to better compare the performance between models and ensure the accuracy and reliability of the final prediction results, we need to determine the hyperparameters of seven different algorithms, respectively, as listed in Table 1.

Table 1. All hyperparameters of every method.

Methods	Hyperparameters	Definition
ARIMA	p, q, and d values	p indicates lag observations for current value prediction; d refers to the number of times the data values have been differenced; q represents the size of the moving average part of the model.
DT	min_samples_split, min_samples_leaf	min_samples_split and min_samples_leaf specify the minimum number of samples required to split an internal node and to be at a leaf node, respectively.
RF	min_samples_leaf, n_estimators, min_samples_split	n_estimators parameter specifies the number of trees in the random forest; The parameters min_samples_split and min_samples_leaf are the same as above.
SVM	C and gamma	C controls the trade-off between achieving a low error on the training data and minimizing the norm of the weights. The gamma defines the influence of a single training example.
XGBoost	learning_rate, n_estimators, max_depth	learning_rate adjusts prediction changes per boosting step. n_estimators sets the number of trees in the ensemble. max_depth is the maximum depth of each tree in the ensemble.
ANN	ω, b_A	ω represents the weight, and b_A is the bias term.
LightGBM	learning_rate, n_estimators, max_depth, num_leaves, min_child_samples, min_child_weight, feature_fraction, bagging_fraction	learning_rate governs the model's prediction adjustment per gradient step. n_estimators denotes the number of boosting iterations or trees. max_depth is each tree's maximum depth. num_leaves and min_child_samples are the maximum leaves per tree and minimum data in a child node. min_child_weight specifies the minimum instance weight sum in a child. The parameters feature_fraction and bagging_fraction specify the fraction of features and data to be used for each tree and boosting round.

This paper utilizes a 10-fold cross-validation technique to select the hyperparameters of the model. The original dataset is divided into a training set and a test set. The training set is for training the model, and the test set is for evaluating the accuracy of the results. In order to make full use of the dataset in the selection of algorithm parameters, the dataset is randomly divided into ten equal-sized subsets. Each time, nine of these ten subsets are used as training data, and the remaining one becomes the test dataset. Since there are ten available subsets to be used as test sets each time, the whole cross-validation process loops ten times. Then, for each combination of parameters, we calculate the average of 10 cross-validation results. Finally, we select the parameter combination with the best cross-validation performance as the optimal parameters to be preset for the corresponding algorithm. The cross-validation process is illustrated in Figure 5.

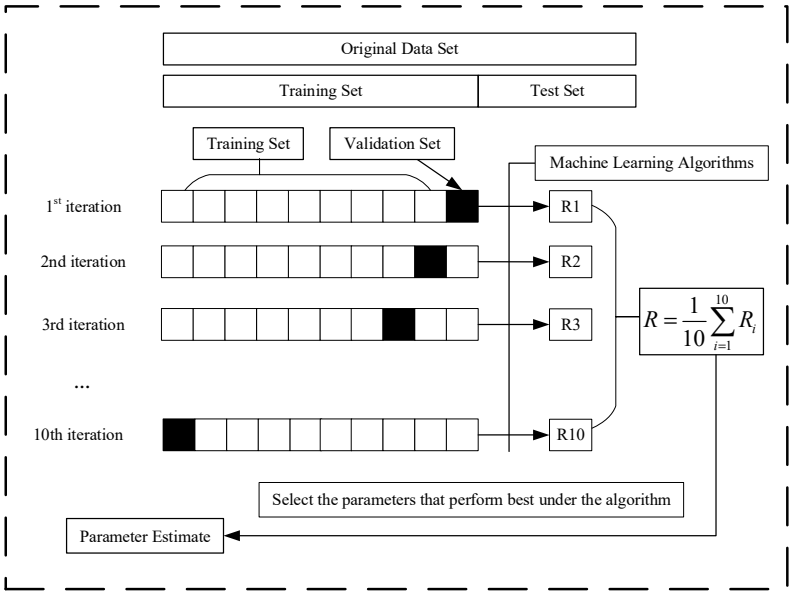


Figure 5. Process of 10-fold cross-validation technique.

3.5. The Experimental Design

To compare the proposed algorithm with other methods in the field of fresh agricultural product price prediction, the performance of the Light GBM integrated learning algorithm is evaluated from three perspectives: accuracy, generalization, and robustness.

(1) First, we compare the prediction results from the Light GBM model with the results from the AMIRA model, DT model, RF model, SVM model, XGBoost model, and ANN model to see which method predicts the prices closest to the actual prices.
(2) Second, we conduct this comparison of price prediction for bananas, beef, and Crucian carp datasets. The purpose of using different datasets is to verify the generality of the proposed model.
(3) Finally, we divide each dataset into three time windows, namely short-term window (0–90 days), medium-term window (0–120 days), and long-term window (0–150 days). By observing the performance of the Light GBM model under changing time windows with the same data input, the model robustness can be verified.

The prediction accuracy is examined via the root mean square error (RMSE), mean absolute percentage error (MAPE), and mean square error (MSE). The formulas of RMSE, MAPE, and MSE are illustrated in Equations (6)–(8).

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(R_t - P_t)^2} \quad (8)$$

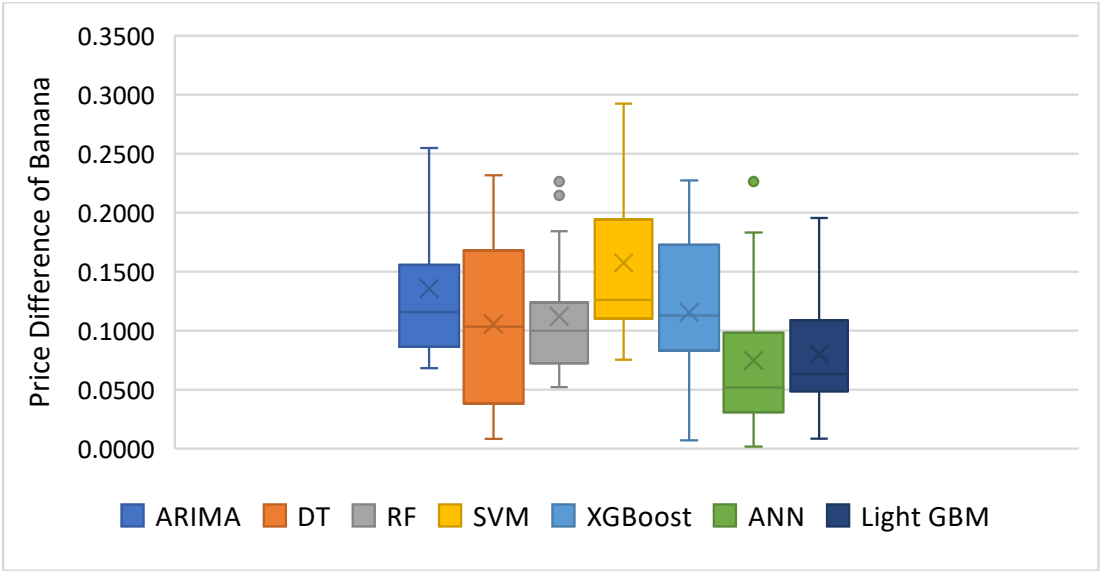
$$MAPE = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{R_t - P_t}{R_t}\right| \times 100\% \quad (9)$$

$$MSE = \frac{1}{N}\sum_{t=1}^{N}(R_t - P_t)^2 \quad (10)$$

where N represents the sample size; Rt represents the actual value, Pt represents the predicted value. Obviously, the smaller the RMSE, MAPE, and MSE values are, the higher the prediction accuracy becomes.

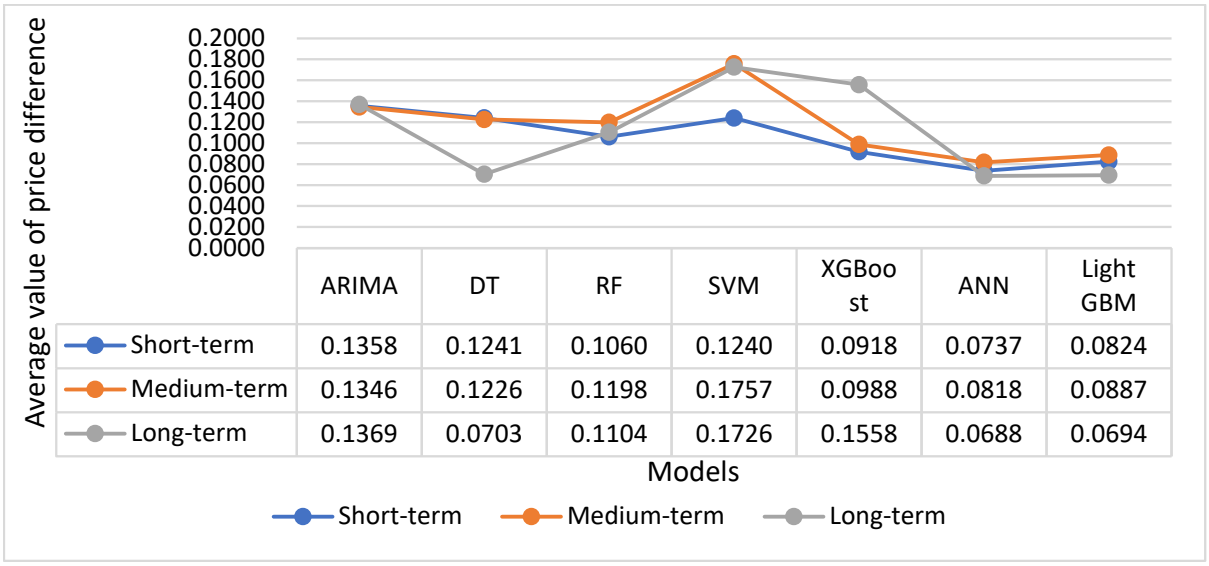
3.6. The Experimental Results and Discussion

In this section, we present a comparative analysis of the absolute deviations of the predicted values from the actual prices for the ARIMA, DT, RF, SVM, XGBoost, ANN, and Light GBM models for predicting prices of bananas, beef, and Crucian carp in changing time windows. Boxplots Figures 6a, 7a and 8a depict a comparative analysis of the absolute deviations of the predicted values from the actual prices. Figures 6b, 7b and 8b illustrate the mean absolute errors between the predicted and actual values for seven models across the same set of data on the three time windows. From these three boxplots, it is evident that the Light GBM model exhibits superior performance for the crops of Beef and Crucian carp in terms of IQR. Although the concentration of absolute prediction errors for beef is slightly inferior to that of the ARIMA model, the overall absolute error values are significantly lower than those of the ARIMA model, as detailed in Figure 7b, the average absolute errors of the Light GBM model over different time windows (0.1232, 0.1348, 0.1168) are markedly lower than those of the ARIMA model (0.4184, 0.4120, 0.4180), indicating superior precision. For the bananas dataset, the Light GBM model demonstrates slightly inferior performance to the RF model in terms of the IQR and slightly inferior performance to the ANN model in terms of the average absolute error (Figure 6b). However, both the RF and ANN models exhibit outliers. This observation can be attributed to the fact that both the RF and ANN models are more prone to overfitting, especially when dealing with datasets with complex underlying patterns. However, the Light GBM model generally exhibits greater robustness in this regard.



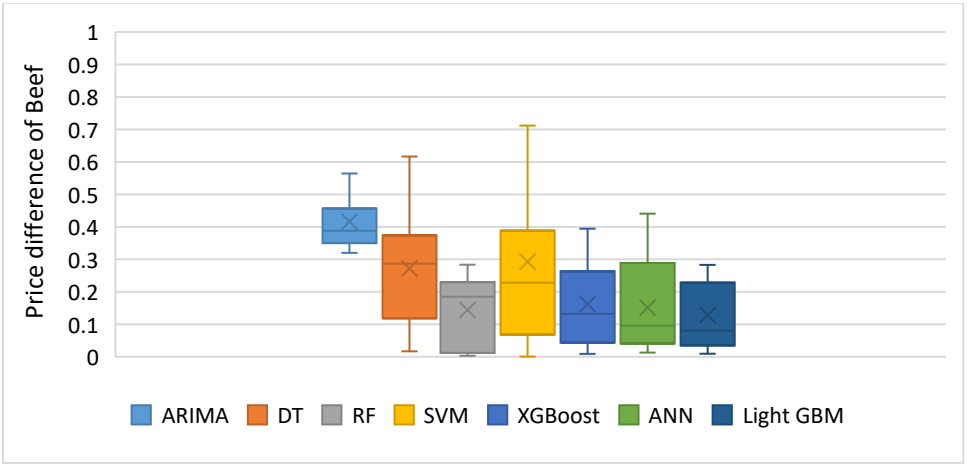
(**a**) Boxplot for seven models' absolute value of the error for bananas.

Figure 6. *Cont.*

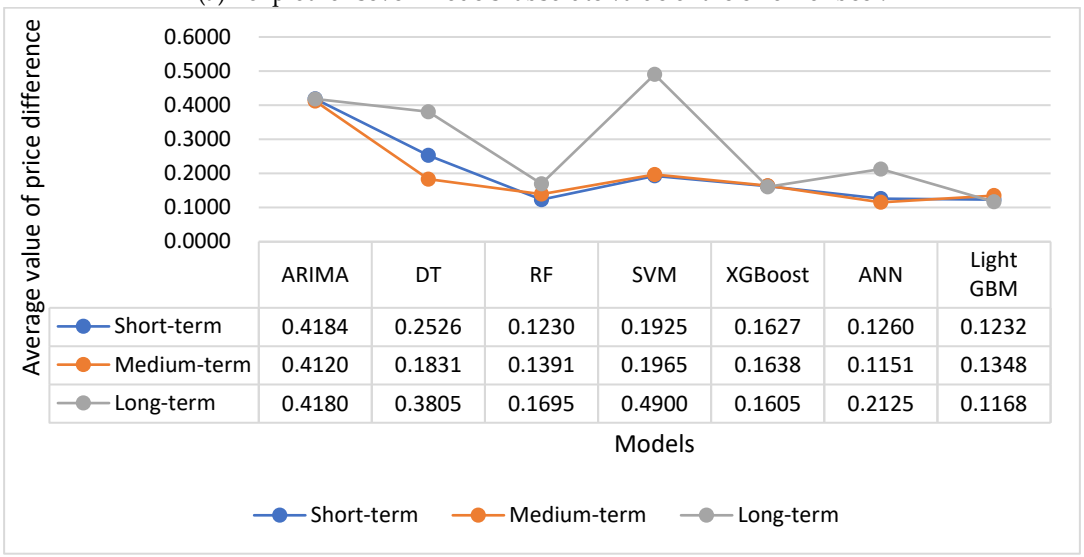


(**b**) The average seven models' absolute value of the error for bananas.

Figure 6. Boxplot The seven results under three time windows for bananas.

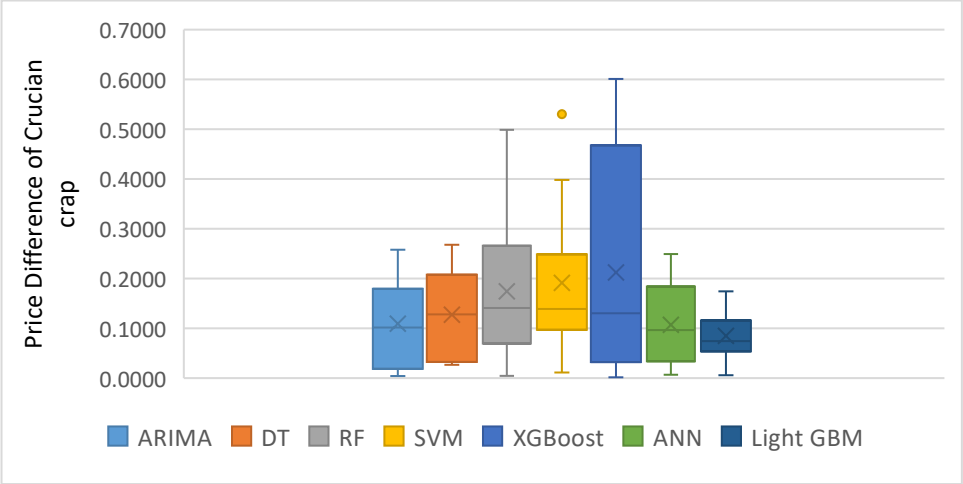


(**a**) Boxplot for seven models' absolute value of the error for beef.

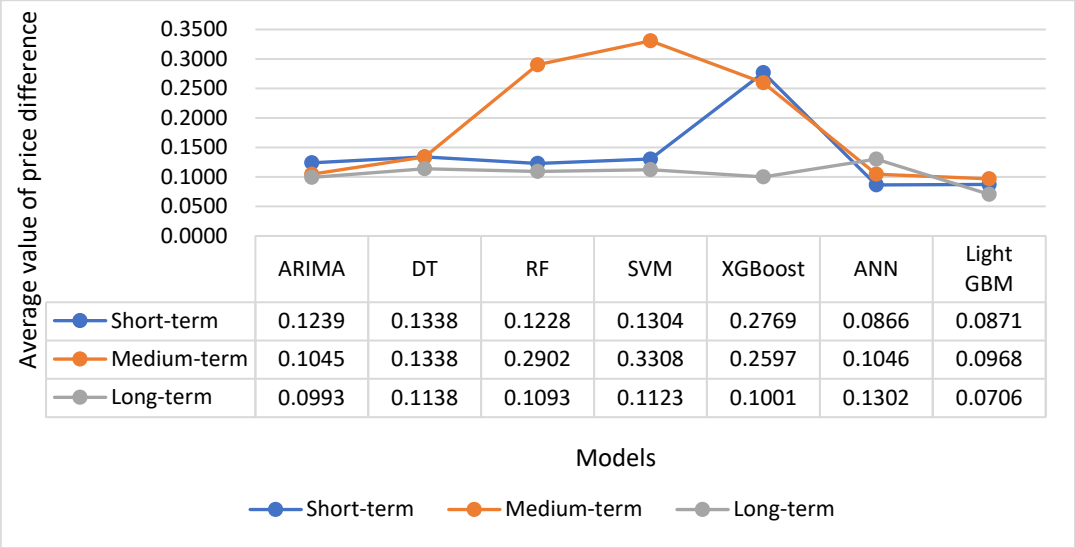


(**b**) The medium-term compared results for beef.

Figure 7. Boxplot The seven results under three time windows for beef.



(**a**) Boxplot for seven models' absolute value of the error for Crucian crap.



	ARIMA	DT	RF	SVM	XGBoost	ANN	Light GBM
Short-term	0.1239	0.1338	0.1228	0.1304	0.2769	0.0866	0.0871
Medium-term	0.1045	0.1338	0.2902	0.3308	0.2597	0.1046	0.0968
Long-term	0.0993	0.1138	0.1093	0.1123	0.1001	0.1302	0.0706

(**b**) The medium-term compared results for Crucian crap.

Figure 8. The six results under threetime windows for Crucian crap.

We also compared the prediction performance of all the methods under changing time windows. Table 2 shows the root mean square error, mean absolute percentage error, and absolute percentage error of all the methods on the three datasets under different time windows. In each column, the bold value is the minimum value.

According to the above charts, the experimental results are analyzed from model accuracy, generalization, and robustness perspectives.

From Table 2, we have found the Random Forest performing well for the crops of beef in the short-term, medium-term, and long-term windows, second only to Light GBM, with RMSE values of 0.1607, 0.1751, and 0.1955, respectively. This is because the RF is not sensitive to outliers. After all, it mitigates errors through the consensus of multiple decision trees. This also explains why the RF model performs better in the absolute error in Figure 2 yet falls short of Light GBM in terms of RMSE. In contrast, ANN may exhibit greater susceptibility to outliers, which performed better for the crops of bananas in the short-term window and medium-term window than the other two, with RMSE values of 0.0968 and 0.1125. The crops of these three datasets perform better in the short-term window, with MAPE values of 1.3725, 0.2027, and 0.6270, respectively. However, these models' mean values of the errors were consistently higher than those of Light GBM. The Light GBM model performs nearly best according to the measurement metrics for three

datasets. This model's RMSE values were calculated as 0.1037, 0.1561, and 0.0957 in the short-term window, 0.1005, 0.1660, and 0.1159 in the medium-term window, and 0.0730, 0.1586, 0.0757 in the long-term window, respectively. Likewise, this model's MAPE values gave better results in three Time Windows and three datasets than the other five models, calculated as 1.6188, 0.2093, 0.6170 in the short-term window, 1.7018, 0.1809, 0.6471 in the medium-term window, and 1.2497, 0.1646, 0.4418 in the long-term window, respectively. They are all higher than the Light GBM model. MSE is no exception. It can be concluded that the prediction accuracy of the Light GBM model is much higher than that of the other models, and the performance is better, which can significantly improve the accuracy of price prediction.

Table 2. The RMSE, MAPE, and MSE results of all the methods under different time windows and datasets.

Metric	TW	Short-Term			Medium-Term			Long-Term			
	DataSet	Bananas	Beef	Crucian	Bananas	Beef	Crucian	Bananas	Beef	Crucian	Ave.
RMSE	ARIMA	0.1553	0.3930	0.1526	0.1553	0.3930	0.1342	0.1553	0.3929	0.1342	0.2295
	DT	0.1357	0.2754	0.1633	0.1341	0.3665	0.1633	0.1077	0.3541	0.1529	0.2059
	RF	0.1209	0.1607	0.1614	0.1275	0.1751	0.1358	0.1183	0.1955	0.1241	0.1466
	SVM	0.1420	0.2597	0.1572	0.1873	0.2201	0.3553	0.1839	0.6519	0.1141	0.2524
	XGBoost	0.1081	0.2168	0.2007	0.1415	0.2442	0.1163	0.1284	0.1938	0.1107	0.1623
	ANN	0.0968	0.1665	0.1181	0.1125	0.1998	0.1203	0.0872	0.2478	0.1541	0.1448
	LGBM	0.1037	0.1561	0.0957	0.1005	0.1660	0.1159	0.0730	0.1586	0.0757	0.1161
MAPE%	ARIMA	2.5942	0.6144	0.9827	2.5942	0.6144	0.6892	2.5941	0.6144	0.6892	1.3319
	DT	1.9191	0.4064	0.9208	2.2607	0.4977	0.9208	1.3116	0.4546	0.7842	1.0529
	RF	1.8614	0.2675	0.8459	2.1571	0.1832	0.7365	2.0247	0.1836	0.8185	1.0087
	SVM	2.2889	0.3094	0.8895	3.2511	0.3164	2.2616	3.1943	0.7878	0.7688	1.5631
	XGBoost	1.6198	0.2579	1.1544	2.1874	0.3681	0.7623	2.2704	0.4606	0.6121	1.0770
	ANN	1.3725	0.2027	0.6270	1.5245	0.1848	0.7147	1.2767	0.3417	0.8954	0.7900
	LGBM	0.0867	0.2013	0.6170	0.0823	0.1809	0.6471	1.2497	0.1646	0.4418	0.4088
MSE	ARIMA	0.0241	0.1544	0.0233	0.0241	0.1544	0.0180	0.0241	0.1544	0.0180	0.0661
	DT	0.0184	0.0758	0.0267	0.0180	0.1344	0.0267	0.0116	0.1254	0.0234	0.0511
	RF	0.0146	0.0258	0.0260	0.0163	0.0307	0.0184	0.0140	0.0382	0.0154	0.0222
	SVM	0.0202	0.0675	0.0247	0.0351	0.0485	0.1263	0.0338	0.4250	0.0130	0.0882
	XGBoost	0.0117	0.0470	0.0403	0.0200	0.0597	0.0135	0.0165	0.0376	0.0123	0.0287
	ANN	0.0112	0.0277	0.0139	0.0127	0.0399	0.0145	0.0076	0.0614	0.0237	0.0236
	LGBM	0.0107	0.0244	0.0092	0.0101	0.0276	0.0134	0.0053	0.0252	0.0057	0.0146

Note: For conciseness, Light GBM is abbreviated as LGBM, Time Window and Average is represented by Ave and TW.

Table 3 summarizes the average RMSE, MSE, and MAPE results of all the methods on three different datasets to provide insights into model generalizability. The average RMSE of the Light GBM model on the bananas, beef, and crucian carp datasets are 0.0924, 0.1602, and 0.0958, respectively. The average MAPE values of the Light GBM model on the three datasets are 1.5234, 0.1849, and 0.5686, respectively. The average MSE values of the Light GBM model on the three datasets are 0.0087, 0.0257, and 0.0094, respectively. It has a better prediction precision than the other models on any dataset. We can conclude that the prediction accuracy of the Light GBM model can be generalized to different datasets. The model performance is relatively stable since it is not contingent on the dataset.

Regarding robustness, under the situation of variable time window data input, the RMSE trend of the Light GBM model and the other control models is shown in Figure 9. As can be seen from the figure, based on the RMSE evaluation index, the Light GBM model has significantly better prediction results than the control model under three different time windows: short-term, medium-term, and long-term. Simultaneously, as the time window gradually expands and the amount of data increases, the time-series relationship between the data are more complicated. Based on the MAPE evaluation metric, the Light GBM

model demonstrates significantly superior predictive performance to the control model across three distinct temporal spans: short-term, medium-term, and long-term, with its longest-term forecast exhibiting the most optimized results. Compared with the other control models, the Light GBM model has more obvious advantages. It can be seen that the Light GBM model can more wisely adopt the time series of data relationships. However, under the long-term time window, the prediction performance of the control models is not ideal. It can be concluded that the Light GBM model shows stronger adaptability in time series data prediction and has stronger robustness.

Table 3. The average RMSE, MAPE, and MSE results of all the methods under different datasets.

	Dataset	LGBM	ARIMA	DT	RF	SVM	XGBoost	ANN
RMSE	Bananas	0.0924	0.1553	0.1258	0.1222	0.1710	0.1260	0.0989
	Beef	0.1602	0.3929	0.3320	0.1771	0.3772	0.2183	0.2047
	Crucian carp	0.0958	0.1403	0.1599	0.1404	0.2089	0.1426	0.1308
MAPE	Bananas	1.5234	2.5942	1.8305	2.0144	2.9114	2.0259	1.3912
	Beef	0.1849	0.6144	0.4529	0.2114	0.4712	0.3622	0.2431
	Crucian carp	0.5686	0.7870	0.8753	0.8003	1.3066	0.8429	0.7537
MSE	Bananas	0.0087	0.0241	0.0160	0.0150	0.0297	0.0161	0.0105
	Beef	0.0257	0.1544	0.1119	0.0316	0.1803	0.0481	0.0430
	Crucian carp	0.0094	0.0198	0.0256	0.0200	0.0547	0.0220	0.0174

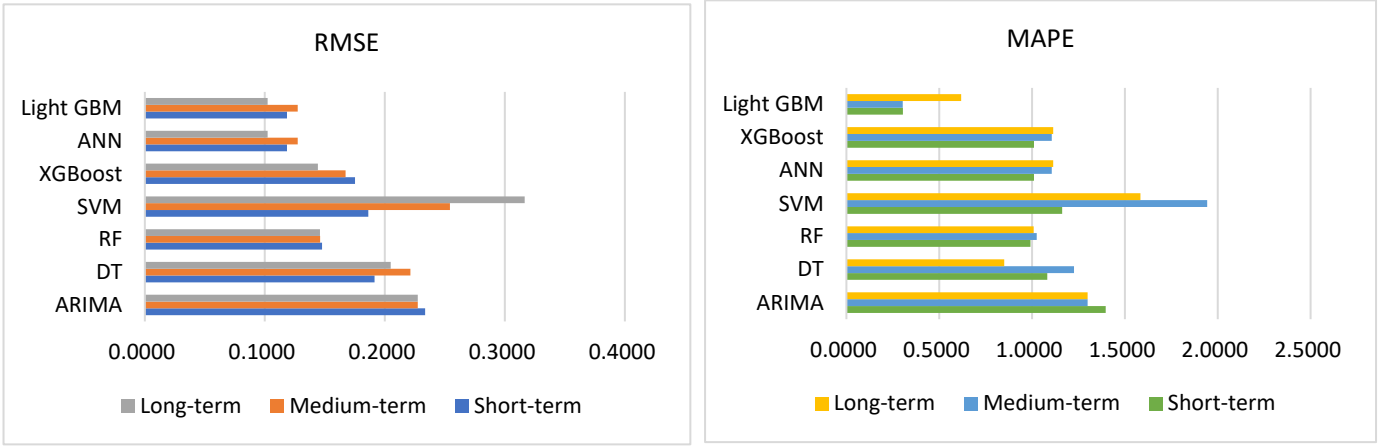


Figure 9. The average RMSE and MAPE values of different models under the three time windows.

4. Conclusions

Examining agricultural prices and their volatility is paramount, considering their extensive influence on external debt, fiscal stability (reflected in government budget balances), economic growth potential, inflationary outlook, and the balance of trade. This paper presents a Light GBM pricing prediction model based on the Boosting ensemble learning algorithm to predict prices for fresh agricultural products under varying time windows. We conduct experiments on three different fresh agricultural datasets: bananas, beef, and crucian carp. Model performance is benchmarked against five state-of-the-art models: ARIMA, decision tree, random forest tree, SVM, and XGBoost. Different time window settings, including short-term window, mid-term window, and long-term window, are also considered. The proposed model is evaluated in three dimensions: accuracy, generalizability, and stability:

(1) Accuracy: the Light GBM model exhibits relatively low prediction error compared with the other five traditional models, shows better data fitting characteristics, and provides higher prediction accuracy.

(2) Generalizability: the light GBM model achieves the highest prediction precision while stable prediction errors over different datasets. The model can adapt to different kinds of datasets, which has strong generalizability.

(3) Robustness: the prediction performance of the light GBM model is consistent in changing time window settings. As the length of the time window increases, more advantages can be brought by the light GBM model. Thus, the model can deal with price prediction in various periods of time and display strong stability.

To sum up, this paper actively explores the application of the ensemble learning algorithm to prediction pricing in the fresh agricultural products industry and verifies the accuracy, generalizability, and robustness of the Light GBM model, provides a practical framework for predicting and estimating fresh prices that are adaptable to different datasets. Because of the sophistication of the price data in the food industry, future research may consider more latent factors that can impact the price and apply them as features in the model in order to reach more accurate and reliable prediction results. Noteworthy, the Light GBM cannot perform well with a small number of samples, sample size must be determined. This study provides a practical framework for predicting and estimating grain prices using ML algorithms that are adaptable to different regions.

Supplementary Materials: The following supporting information can be downloaded at https://github.com/Nina-Zhang666/paper-s-code/blob/main/Supplementary%20material-scripts.txt, (accessed on 30 June 2023).

Author Contributions: Conceptualization, N.Z. and Q.A.; methodology, N.Z.; software, S.Z.; validation, N.Z., H.M. and Q.A.; formal analysis, N.Z.; investigation, S.Z.; resources, H.M.; data curation, S.Z.; writing—original draft preparation, N.Z.; writing—review and editing, Q.A.; visualization, S.Z.; project administration and writing-revised manuscript, H.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was financially supported by the National Natural Science Foundation of China (grant number 72202093).

Data Availability Statement: The original data presented in this study are openly available and can be accessed and obtained through the URL http://www.moa.gov.cn/ (accessed on 30 June 2023). Research data can be accessed and listed systematically by anyone at any time.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kyriazi, F.; Thomakos, D.D.; Guerard, J.B. Adaptive learning forecasting, with applications in forecasting agricultural prices. *Int. J. Forecast.* **2019**, *35*, 1356–1369. [CrossRef]
2. Mahmud, I.; Das, P.R.; Rahman, M.H.; Hasan, A.R.; Shahin, K.I.; Farid, D.M. Predicting Crop Prices using Machine Learning Algorithms for Sustainable Agriculture. In Proceedings of the 2024 IEEE Region 10 Symposium (TENSYMP), New Delhi, India, 27–29 September 2024; IEEE: Piscataway Township, NJ, USA, 2024; pp. 1–6.
3. Zeng, L.; Ling, L.; Zhang, D.; Jiang, W. Optimal forecast combination based on PSO-CS approach for daily agricultural future prices forecasting. *Appl. Soft Comput.* **2023**, *1*, 109833. [CrossRef]
4. Li, B.; Ding, J.; Yin, Z.; Li, K.; Zhao, X.; Zhang, L. Optimized neural network combined model based on the induced ordered weighted averaging operator for vegetable price forecasting. *Expert Syst. Appl.* **2021**, *168*, 114232. [CrossRef]
5. Wang, L.; Feng, J.; Sui, X.; Chu, X.; Mu, W. Agricultural product price forecasting methods: Research advances and trend. *Br. Food J.* **2020**, *122*, 2121–2138. [CrossRef]
6. Zhang, Y.; Na, S. A Novel Agricultural Commodity Price Forecasting Model Based on Fuzzy Information Granulation and MEA-SVM Model. *Math. Probl. Eng.* **2018**, *2018*, 2540681. [CrossRef]
7. Wang, C.; Zhao, A.; Zhao, Y. Design and Implementation of Agricultural Product Prices Short-Term Forecasting System. In *Proceedings of 2013 World Agricultural Outlook Conference*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 15–27.
8. Tay, F.E.H.; Cao, L. Application of support vector machines in financial time series forecasting. *Omega* **2001**, *29*, 309–317. [CrossRef]

9. Luo, J.; Hong, T.; Gao, Z.; Fang, S.C. A robust support vector regression model for electric load forecasting. *Int. J. Forecast.* **2023**, *39*, 1005–1020. [CrossRef]
10. Zhang, N.; Lin, A.; Shang, P. Multidimensional k-nearest neighbor model based on EEMD for financial time series fore-casting. *Phys. A Stat. Mech. Its Appl.* **2017**, *477*, 161–173. [CrossRef]
11. Luo, J.; Yan, X.; Tian, Y. Unsupervised quadratic surface support vector machine with application to credit risk assessment. *Eur. J. Oper. Res.* **2020**, *280*, 1008–1017. [CrossRef]
12. Zhou, J.; Tian, Y.; Luo, J.; Zhai, Q. A kernel-free Laplacian quadratic surface optimal margin distribution machine with application to credit risk assessment. *Appl. Soft Comput.* **2023**, *133*, 109931. [CrossRef]
13. Patel, J.; Shah, S.; Thakkar, P.; Kotecha, K. Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. *Expert Syst. Appl.* **2015**, *42*, 259–268. [CrossRef]
14. Luo, J.; Hong, T.; Fang, S.C. Benchmarking robustness of load forecasting models under data integrity attacks. *Int. J. Forecast.* **2018**, *34*, 89–104. [CrossRef]
15. Luo, J.; Hong, T.; Yue, M. Real-time anomaly detection for very short-term load forecasting. *J. Mod. Power Syst. Clean Energy* **2018**, *6*, 235–243. [CrossRef]
16. Gusain, K.; Gupta, A.; Popli, B. Transition-Aware Human Activity Recognition Using Extreme Gradient Boosted Decision Trees. In *Advanced Computing and Communication Technologies*; Springer: Singapore, 2018; pp. 41–49.
17. Shao, Y.E.; Dai, J.-T. Integrated Feature Selection of ARIMA with Computational Intelligence Approaches for Food Crop Price Prediction. *Complexity* **2018**, *2018*, 1910520. [CrossRef]
18. Shahwan, T.; Odening, M. Forecasting Agricultural Commodity Prices using Hybrid Neural Networks. In *Computational Intelligence in Economics and Finance: Volume II*; Chen, S.-H., Wang, P.P., Kuo, T.-W., Eds.; Springer: Berlin/Heidelberg, Germany, 2007; pp. 63–74.
19. Li, G.; Xu, S.; Li, Z. Short-Term Price Forecasting For Agriproducts Using Artificial Neural Networks. *Agric. Agric. Sci. Procedia* **2010**, *1*, 278–287.
20. Liu, Y.; Duan, Q.; Wang, D.; Zhang, Z.; Liu, C. Prediction for hog prices based on similar sub-series search and support vector regression. *Comput. Electron. Agric.* **2019**, *157*, 581–588. [CrossRef]
21. Chen, L.; Li, Z. Research on the price analysis and prediction method of agricultural products based on logistics information. *Clust. Comput.* **2019**, *22*, 14951–14957. [CrossRef]
22. Wu, B.; Wang, Z.; Wang, L. Interpretable corn future price forecasting with multivariate time series. *J. Forecast.* **2024**, *43*, 1575–1594. [CrossRef]
23. Chow Suliyaa, M.; Prahadeeswaran, M.; Karthick, V.; Sumathi, C.S.; Santosh, G.P. Machine Learning Application in Agricultural Price Forecasting: Lessons from Banana Markets in Tamil Nadu, May 2024. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4835723 (accessed on 19 November 2024).
24. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [CrossRef]
25. Quinlan, J.R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
26. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
27. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
28. Biau, G. Analysis of a random forests model. *J. Mach. Learn. Res.* **2012**, *13*, 1063–1095.
29. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
30. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boost-ing System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
31. Zupan, J.; Gasteiger, J. Neural networks: A new method for solving chemical problems or just a passing phase? *Anal. Chim. Acta* **1991**, *248*, 1–30. [CrossRef]
32. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A highly efficient gradient boosting decision tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 4–9 December 2017; pp. 3149–3157.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.