Academic Journal of Computing & Information Science
ISSN 2616-5775 Vol. 7, Issue 10: 82-88, DOI: 10.25236/AJCIS.2024.071012

Research on retail pricing strategy of supermarket fresh products based on XGBoost model

Litao Zeng, Yang Gao, Haisen Hu

School of Electronics and Automation, City Institute Dalian University of Technology, Dalian, 116600, China

Abstract: *For the problem of wastage caused by the short freshness duration and perishability of vegetable category goods in fresh produce superstores, this investigation constructed an automatic pricing and replenishment decision support module using historical sales of data. The research firstly adopts a Python-based Long Short-Term Memory Network (LSTM) time-series of forecast model to predict the future wholesale prices of the vegetable category, and achieve a high degree of model fit (r2 ≥ 0.93). Subsequently, this research built an XGBoost sales volume prediction model based on sales unit price and wholesale price, in which the model for most of the categories showed significant prediction results. Furthermore, this study formulated an objective solution model with revenue maximization as the goal, and optimized the solution using genetic algorithm to maximize the total income of the superstore in the next seven days. Additionally, through the sensitivity analysis of the model, this study verified the stability of the model output. The results show that the model is more robust to small changes in the input variables. This study provides the theoretical basis and practical guidance for inventory management and revenue optimization in fresh produce superstores.*

Keywords: *Fresh Supermarket, Vegetable Preservation, LSTM, XGBoost, Genetic Algorithm*

1. Introduction

Fresh retail has a great market prospect in the future development, but it is facing a great realistic dilemma. The shelf life of vegetable products is relatively short, easy to deteriorate, rot, most varieties such as unsold on the day, the next day can not be sold again. Therefore, in order to reduce the rate of wastage and wastage, this study will build a model based on the historical sales and demand of each product combined with XGBoost, LSTM and other algorithms to assist supermarkets in making sales strategies and daily replenishment. The data in this paper come from the C question of 2023 Higher Education Society Cup National College Students Mathematical Contest in Modeling.

Ping-Hui Hsu et al. [1] showed that consumers are very sensitive to the remaining shelf life of fresh products, that is, when the purchase time is gradually approaching the shelf life, the perceived value of the products of customers gradually decreases, so the demand rate of the products also decreases. Liu Xinmin et al. [2] show that the freshness and shelf time of fresh agricultural products are positively correlated with the channel sales price, and different consumer types have different choices for price and freshness. Hu Hanli et al. [3] the sales model selection and pricing strategy of the fresh supply chain composed of suppliers and retailers can be divided into decentralized and centralized sales models, and skimming and penetration pricing strategies. Duan Yongrui [4], et al. showed that based on the cost of merchants' investment in preservation technology, this paper studied the joint pricing and inventory decision-making problem of non-instant spoilage products when they were allowed to be out of stock and the out of stock was partially delayed. However, most of these studies are based on the traditional prediction methods of linear regression, such as time series method, analysis method and pattern recognition method, which have limitations when dealing with nonlinear and dynamically changing fresh data [5]. This results in limited forecasting accuracy, which in turn affects the effectiveness of pricing and replenishment strategies.

In this study, the LSTM time series forecasting model was constructed based on Python, and the future wholesale prices of vegetables were predicted by using the historical wholesale price data. Compared with the existing research, the advantage of this study is that the LSTM model can better capture the nonlinear relationship in the time series data and improve the accuracy of prediction. The model evaluation results showed that the fitting degree was good (r2≥0.93). The following models are constructed in this study: The sales forecasting model based on XGBoost, which takes into account

ISSN 2616-5775 Vol. 7, Issue 10: 82-88, DOI: 10.25236/AJCIS.2024.071012

factors such as unit price and wholesale price, and the model evaluation for six categories shows that except for the pepper model (r2 is 0.6438), the other categories have achieved significant effects (r2 is greater than 0.78). A model is established to solve the target of maximization of super revenue, which is solved by genetic algorithm to ensure the maximization of total revenue in the next seven days. This method overcomes the limitation of traditional methods in solving optimization problems. In this study, the reliability and sensitivity of the model are proved by stable model iteration and sensitivity analysis. Through the use of advanced machine learning models, more accurate price and sales forecasts are provided for fresh retailers to provide more scientific pricing and replenishment decision support. In addition, it not only improves the operational efficiency of fresh retail, but also provides a new research direction and methodology for the industry. By optimizing pricing and replenishment strategies, fresh retailers are expected to achieve higher returns and lower wastage.

2. Model

2.1 The basic fundamental of XGBoost

XGBoost (Extreme Gradient Boosting) stands for Extreme Gradient Lift Tree. XGBoost is the trump card of ensemble learning methods and has been used by most winners in Kggl data mining competitions. XGBoost performs very well in most regression and classification problems. This section will introduce the algorithm principle of XGBoost in more detail. The general approach is to minimize the loss function of the training data. We use the letter L to represent the loss, as follows: where F is the hypothesis space, the hypothesis space is given for all possible cases that satisfy the objective, given the properties and possible values of the properties. A set of hypotheses with no omissions.

$$\min(f\in F)\frac{1}{N}\sum_{i=1}^{N} L\left(y_i, f(x_i)\right) \tag{1}$$

It is called empirical risk minimization, and the trained model is more complex. When the training data is small, the model is prone to overfitting problems. Therefore, in order to reduce the complexity of the model, the following formula is often used: where J(f) is the complexity of the model [6].

$$\min(f\in F)\frac{1}{N}\sum_{i=1}^{N} L\left(y_i, f(x_i)\right) + \lambda J(f) \tag{2}$$

It is called the structural risk minimization. The model with the structural risk minimization tends to have a better prediction of both the training data and the unknown test data. The generation and pruning of decision tree correspond to empirical risk minimization and structural risk minimization respectively. The generation of decision tree in XGBoost is the result of structural risk minimization.

Compared to the traditional GBDT algorithm, it has some differences, firstly, although both use CART trees as the base learner, XGBoost also supports linear classifiers. Secondly, when optimizing the data, XGBoost performs a quadratic Taylor expansion of the function, adding second-order derivatives to the first-order derivatives. And XGBoost supports column sampling which is used in Random Forest. using column sub-sampling prevents over-fitting even more so than the traditional row sub-sampling (which is the usage of column sub-sampling also supports the traditional row sub-sampling). The usage of column sub-samples also speeds up computations of the parallel algorithm described later [7]. Finally, XGBoost supports parallelism by sorting all the characteristics before training and saving them in a block structure.

XGBoost updates the model by minimizing the loss function and uses gradient boosting to solve the problem. In each iteration, XGBoost calculates the first and second order derivatives of the loss function with respect to the current model prediction, and then constructs a new decision tree based on these derivatives. The XGBoost decision model can be seen as figure 1:

ISSN 2616-5775 Vol. 7, Issue 10: 82-88, DOI: 10.25236/AJCIS.2024.071012

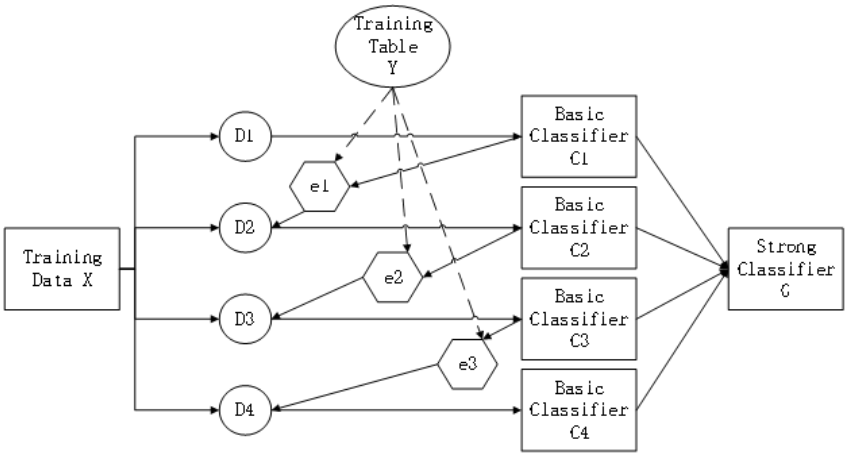

Figure 1: XGBoost decision model

2.2 LSTM

LSTM is a variation of recurrent neural networks (RNNs) [8], where ordinary RNNs undergo a continuous training process in which the number of network layers increases and it is difficult to learn the connection information. Hochreiter (1991) and Bengio, Simard, and Frasconi (1994) analyzed fundamental reasons for the long-term dependencies problem: error signals flowing backward in time tend to either blow up or vanish. In 1997, Sepp Hochreiter and Jürgen Schmidhuber proposed LSTMs to address the inability to artificially lengthen the task due to the tendency of RNNs to suffer from gradient vanishing, and they have been refined and generalized by many in further work, and they are particularly well suited for dealing with data with time-dependencies. The LSTM control unit schematic can be seen as figure 2:

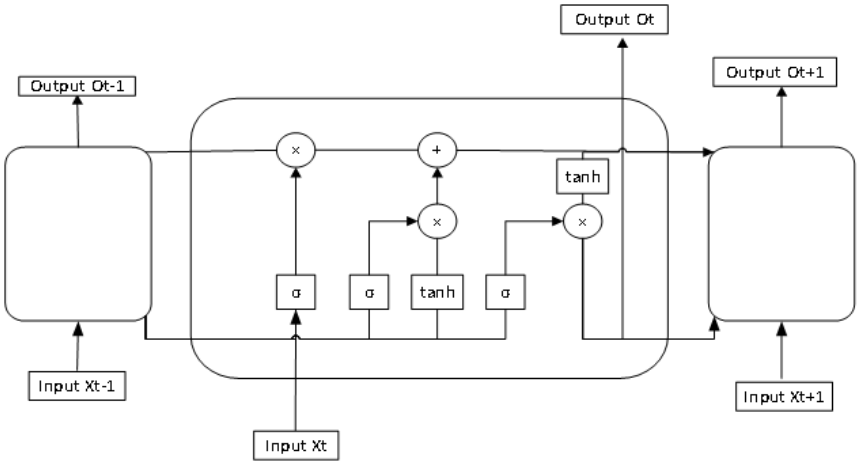

Figure 2: LSTM Control unit diagram

The LSTM neural network structure consists of a cellular state unit, and three gate structures (input gate, forgetting gate, and output gate) The cellular state unit allows for convenient recording of the current state of the moment while the gate functions control the remembering and forgetting of information.

$$\mathrm{f_t} = \sigma\left(W_{\mathrm{f}} \cdot [h_{t-1}, x_t] + b_f\right) \tag{3}$$

$$\mathrm{i_t} = \sigma\left(W_{\mathrm{t}} \cdot [h_{t-1}, x_t] + b_t\right) \tag{4}$$

$$\tilde{C}_{\mathrm{t}} = \tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right) \tag{5}$$

$$C_{\mathrm{t}} = \mathrm{f_t} \times C_{\mathrm{t\text{-}1}} + \mathrm{i_t} \times \tilde{C}_{\mathrm{t}} \tag{6}$$

ISSN 2616-5775 Vol. 7, Issue 10: 82-88, DOI: 10.25236/AJCIS.2024.071012

$$O_t = \sigma（W_O \cdot [h_{t-1}, x_t] + b_O）\tag{7}$$

$$h_t = O_t \times \tanh（C_t）\tag{8}$$

2.3 Genetic algorithm

Genetic algorithm (GA) is an optimization algorithm that is inspired from the natural selection. It is a population based search algorithm, which utilizes the concept of survival of fittest [9]. The core of the genetic algorithm is to select a better chromosome colony through continuous evolutionary evolution, so that the chromosome fitness value continuously decreases and converges thus obtaining the global local optimal solution. In the initial stage, we randomly generate N sequences as the initial cluster of chromosomes, each chromosome in the cluster represents a path solution, and then new chromosomes (X, Y) X+Y=N are generated by means of gene transformation and mutation respectively.

Gene transformation in the initial community randomly selected a and b two groups of chromosomes, in a randomly intercepted section of the gene sequence in b eliminate the same genes, while connecting the remaining genes and then intercepted part of a group to access, to get a new chromosome.

Genetic variation A group of chromosomes is randomly selected in the initial cluster, a random section of gene sequence is intercepted and disrupted, and then connected.

The newly generated N group of chromosomes and the initial N group of chromosomes are put together for fitness sorting, and the N group of chromosomes with the lowest fitness is selected to replace the initial cluster, and the cycle is repeated when the upper limit of the number of cycles is reached, to obtain the optimal solution under a certain meaning.

3. Result

We need to consider the revenue composition of the superstore through several aspects, and we choose to start from the wholesale price as well as the sales volume and the attrition rate.

For the future wholesale price, we first choose to use the LSTM time-series prediction model to predict the future wholesale price based on historical data, according to the LSTM network structure to build the model and calculate each LSTM unit, the input data using the lag of the first order as a feature, the model parameters using the industry's default three-layer LSTM network structure for the initial training, and select the learning rate of 1e -3.The number of neurons in the three layers are 1024, 512, and 64 as the basic hyperparameters, and the parameter optimization method used is the small batch gradient descent method, with the batch set to 64, and the dropout strategy is also introduced in the pooling layer and set to 0.5, which is calculated as follows:

However, since the prediction task is to construct the lag feature of the historical wholesale price as the input, we need to construct the lag feature function to improve the prediction accuracy of the model. The specific formula is as follows:

Where $X_ı$ denotes the eigenvalue at time t, Xtn denotes the eigenvalue at time t-n, and f denotes the lagged feature of the generating function. With the above model we obtain the predicted wholesale prices as shown in the table 1:

Table 1: Forecast wholesale price

Category name	MSE	RMSE	MAE	R2
Capsicum	1.5826	1.2580	0.9416	0.6438
Flower and leaf	01066	0.3266	0.2474	0.7887
Aquatic rhizome	0.6251	0.7907	0.5166	0.9124
Edible fungus	0.6251	0.6666	0.5746	0.8519
Cauliflower	0.4443	1.0338	0.5055	0.8431
Nightshade	0.4007	0.9124	0.5053	0.9054

According to the above table it shows that LSTM has sufficient prediction accuracy, especially in predicting floral and foliar species as well as aquatic rhizomes.

Secondly, for sales volume, we choose to build an XGBoost model for prediction, again using historical data as inputs, with sales unit price and wholesale price as independent variables, and sales volume as the dependent variable, the model parameters are used as default constants, and XGBoost

ISSN 2616-5775 Vol. 7, Issue 10: 82-88, DOI: 10.25236/AJCIS.2024.071012

consists of a loss function as well as a canonical term, with the specific formulas as follows:

$$\text{loss function:} L(\theta) = \sum[l(y_i, \hat{y_i}) + \Omega(f)] \tag{9}$$

The prediction model can be expressed after the iteration as:

$$\hat{y_i}^t = \sum_{i=1}^{n} f_i(x_i) \tag{10}$$

Finally, we arrive at the following result as table 2:

Table 2: Forecast sales

Category name	Mean absolute percentage error	R2_score
Capsicum	14.546	0.944069259
Flower and leaf	12.018	0.930800258
Aquatic rhizome	29.903	0.945991688
Edible fungus	16.972	0.939006393
Cauliflower	20.867	0.935911997
Nightshade	19.543	0.948331309

From the above table we can find that the model fit is more excellent, so the predicted data has some accuracy.

For the attrition rate, we chose to use the historical average data because of the small variation in the attrition rate. The result can be seen in the table 3:

Table 3: The rate of commodity loss

Category name	Loss rate
Capsicum	7.745065
Flower and leaf	12.63664
Aquatic rhizome	9.995515
Edible fungus	8.758815
Cauliflower	10.482727
Nightshade	6.637586

After arriving at the above parameters, we finally choose to construct the target model for maximizing the revenue of the superstore and choose to use genetic algorithm to solve the problem, and the formula for maximizing the revenue of the superstore is as follows.

$$\text{Maixmize} = \sum_{i=1}^{n7} \text{sales}_{\text{unit}_{\text{price}(x_i)}} \times \left(\frac{100 - \text{Loss}_{\text{rate}}}{100}\right) \times (x_i - \text{Wholesale}_i) \tag{11}$$

$$x_i > \text{Wholesale}_i, \forall i \in \{1,2,\cdots,7\} \tag{12}$$

The main decision variable in the model is the unit sales price per day, denoted as x_i. This is the output of the model and represents the recommended sales price for day i. The model also contains an important constraint that the sales unit price needs to be greater than the wholesale price [10], in order to solve the model, we finally chose to use a genetic algorithm, first of all, we need to encode the chromosomes first, each chromosome can be a seven-dimensional vector representing the degree of sales in the next seven days, such as X1, X2, X3...X7, where each represents the sales price of day i. The initial population generation is then performed. Day's sales price, followed by initial population generation, randomly generating multiple chromosomes and evaluating the fitness of each chromosome with the following equation:

$$\text{Sales}_{unit_{price_{(x_i)}} \times ((100 - \text{Loss}_{\text{rate}})/100)} \times (x_i - \text{Wholesale}_i) \tag{13}$$

Subsequently the chromosomes with higher fitness were selected and cross-paired finally terminating after a certain number of iterations and deriving the final data, we chose to set the overall size to 50, the crossover probability to 0.5, the variance probability to 0.3, and the number of iterations to the default of 50, obtaining the figure 3 iterative graph:

ISSN 2616-5775 Vol. 7, Issue 10: 82-88, DOI: 10.25236/AJCIS.2024.071012

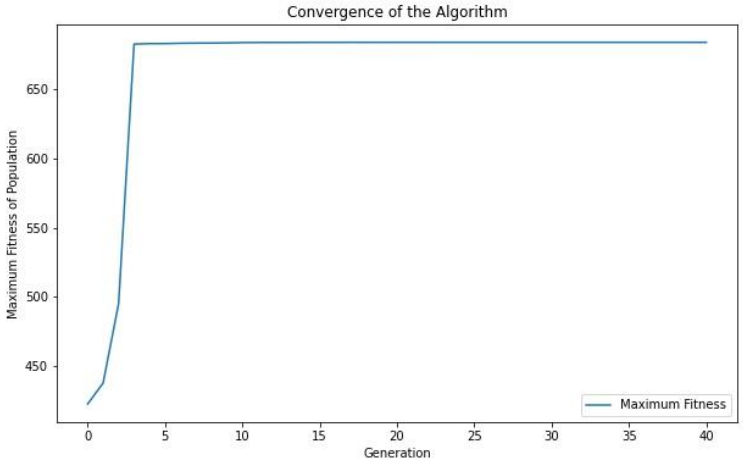

Figure 3: Iterative graph

From the graph, we can find that the revenue of the daily superstore gradually converges and tends to a stable value, which indicates that the planning is oriented to the optimal solution, and finally we can get the total revenue of the next seven days is 41398.93 yuan, and here we list the results of the incoming chili pepper class as shown in the table 4:

Table 4: Chili sales for the next week

Next week	Sales unit price	sales volume	excess revenue	category
first day	20.64177154	386.9182739	5296.856349	capsicum
second day	14.02944355	137.2070007	1085.298374	capsicum
Third day	6.734661718	149.3578491	259.1630142	capsicum
Fourth day	17.62982325	292.8843689	3440.239811	capsicum
Fifth day	6.733832662	132.2991028	245.9464716	capsicum
Sixth day	12.14915634	204.6310272	1354.07374	capsicum
Seventh day	7.550443411	166.344223	366.6259935	capsicum

4. Conclusion and Outlooks

In view of the different demands of different users for vegetables in different time periods, and the loss of vegetables will affect the price based on time, this paper takes the cost plus pricing method as a reference, because there is a certain relationship between the sales volume of vegetables. First, fluctuations in vegetable prices can be attributed to the direct and interaction effects of demand, supply, import and export, and the direct effect is generally positive and larger than the interaction effect.

Consequently, this study initially processed the data to establish the relationship between sales volumes of various vegetable categories and their distribution, aiming to identify an optimal combination for vegetable sales. Inappropriate pricing could potentially reduce consumer willingness to make purchases. The increase in vegetable price volatility is affected by inherent time series, natural disasters and the uniqueness of the vegetable itself, and the substitution and integration of different vegetables. Recent examples of garlic, ginger, bean and onions show that the price of vegetables has risen abnormally. The high vegetable price could negatively affect the consumers. This paper then constructs a predictive model using historical data to forecast sales volume and wholesale prices. This is done in order to establish a more rational pricing strategy, develop an optimal pricing and replenishment strategy, identify varying consumer demands for vegetables at different time periods, and assist supermarkets in maximizing profits. The genetic algorithm is selected as the method for solving the optimal solution of supermarket revenue, primarily due to: The simulation ability of neural network based on GA is worse than BPNN. But its generalization ability is good, predicting accuracy is better than BPNN. At the same time, it also avoids the waste of vegetables, which helps consumers to buy their own satisfactory goods, reduce unnecessary waste, and improve the income of supermarkets.

For the ever-changing business environment, we also need to continuously conduct in-depth research and exploration in the hope of constructing more accurate and efficient models, so we need to integrate

ISSN 2616-5775 Vol. 7, Issue 10: 82-88, DOI: 10.25236/AJCIS.2024.071012

deep learning technology in order to incorporate more practical algorithms, and we need to conduct more analysis of consumer behavior in order to facilitate a more accurate prediction of potential revenue. All in all, the key to staying competitive in the many challenges of future R&D is to continue to improve and expand our models with the help of new technologies and methodologies.

References

[1] Hsu P, Teng H, Wee H. Optimal lot sizing for deteriorating items with triangle-shaped demand and uncertain lead time[J]. European J. of Industrial Engineering, 2009, 3(3):247-260.
[2] Liu Xinmin, Yan Xiuxia, Fu Kaiying, et al. Research on pricing strategy of fresh agricultural products in dual-channel supply chain from the perspective of multi-dimensional collaboration [J]. Research of Business Economics, 2020, (11):151-154.
[3] Hu Hanli, Cao Yu, Wu Kan. Research on sales model selection and pricing of fresh supply chain based on pre-sale [J]. Operations Research and Management, 2022, 31(11):128-134.
[4] Duan Yongrui, Lei Wei, Li Guiping. Non-instant spoilage inventory and pricing strategy considering preservation investment [J]. Journal of Systems Management, 2019, 28(04):732-741.
[5] Song Zhilan, Article Review. Fresh product pricing and inventory strategy under the background of new retail [J]. Logistics Technology, 2021, 40(07):89-94.
[6] Yang Z, Zhang Q, Zhang R, et al. Transverse Vibration Response of a Super High-Speed Elevator under Air Disturbance[J]. International Journal of Structural Stability and Dynamics, 2019, 19(9):25.
[7] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System [J]. CoRR, 2016, abs/1603.02754
[8] Klein I. Smartphone Location Recognition: A Deep Learning-Based Approach [J]. Sensors, 2019, 20(1): 214-218.
[9] Sourabh K, Singh S C, Vijay K. A review on genetic algorithm: past, present, and future. [J]. Multimedia tools and applications, 2020, 80(5):31-36.
[10] Bo Y, Zhuo C, Xinni W, et al. Influence of logistic service level on multichannel decision of a two-echelon supply chain[J]. International Journal of Production Research, 2020, 58(11):3304-3329.