Information Systems and Economics (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/infse.2023.041012
ISSN 2523-6407 Vol. 4 Num. 10

Pricing and Replenishment Decision of Vegetable Goods Based on LSTM and XG-Boost Models

Zhengyi Luo[1,†], Xin Zou[1,†]

[1]*School of Electronics and Information, Nanchang Institute of Technology, Nanchang, China*
[†]*These authors also contributed equally to this work*

Keywords: Kendall Correlation Coefficient, LSTM, XG-Boost, Goal Planning Model, Monte Carlo Algorithm

Abstract: In general, vegetable commodities do not have a long shelf life, and along with the increase in time the quality of vegetable commodities will be reduced, so many super if the day is not sold out, basically cannot be sold again such commodities, which invariably increases the rate of loss. To this end, this paper analyzes the relevant sales data and develops appropriate pricing and replenishment decisions. First of all, the distribution analysis of vegetables in various categories of goods, based on the Kendall correlation coefficient test for the consistency of the single product test, and then the correlation between the two categories of correlation analysis. Then, a mathematical model is constructed to maximize the revenue of the superstore, using the LSTM time-series prediction model to predict the wholesale price of each category of vegetables in the coming week based on the historical wholesale price, the GBDT sales volume prediction model based on the unit price and wholesale price, the objective function of maximizing the revenue of the superstore is set up, and the total revenue of each category of vegetables in the coming week is obtained through Monte Carlo algorithm solving. Finally, in order to meet the requirements of the total number of individual items and the minimum display quantity, the LightGBM time series prediction model is constructed on the historical wholesale price data to predict the wholesale price of the vegetable category on a single day and establish the objective model for maximizing the revenue of the superstore, which is determined by the daily sales quantity, the sales unit price, and the difference of the wholesale price together with the wastage rate. The proposed model has high solution efficiency and optimality.

1. Introduction

With the improvement of the quality of life, people's dietary requirements are also higher, general vegetable commodities do not have a long shelf life, so along with the increase in sales time, the need for timely replenishment according to the sales volume of vegetable commodities as well as demand, so through the different categories, different single product sample data, we further explore and specify how to replenish the vegetable commodities and pricing.

In this paper, based on the vegetable categories provided by a superstore, the relevant information on individual items, the sales flow of each commodity, the wholesale price, and the recent attrition

rate data, we carry out deep mining based on the data to establish a mathematical model. First, the distribution analysis of each category of vegetable goods, the consistency test of each single product based on Kendall's correlation coefficient test, and then the correlation analysis between two and two of each category. Then, a mathematical model is constructed to maximize the revenue of the superstore, using the LSTM time-series prediction model to predict the wholesale price of each category of vegetables in the coming week based on the historical wholesale price, the GBDT sales volume prediction model based on the unit price and wholesale price, the objective function of maximizing the revenue of the superstore is established, and the total revenue of each category of vegetables in the coming week is obtained through the Monte Carlo algorithm solution. Finally, to meet the requirements of the total number of individual items and the minimum display quantity, the LightGBM time series prediction model is constructed for the historical wholesale price data to predict the wholesale price of vegetable categories on a single day, and the objective model for maximizing the revenue of the superstore is set up, which is determined by the daily sales quantity, the sales unit price, and the difference between the wholesale price and the attrition rate together.

2. Model formulation and solving

2.1 Correlation analysis based on Kendall's correlation coefficient

The sales volume data for each category of vegetables was visualized to obtain the distribution of sales volume for each category of vegetables shown in Figure 1.

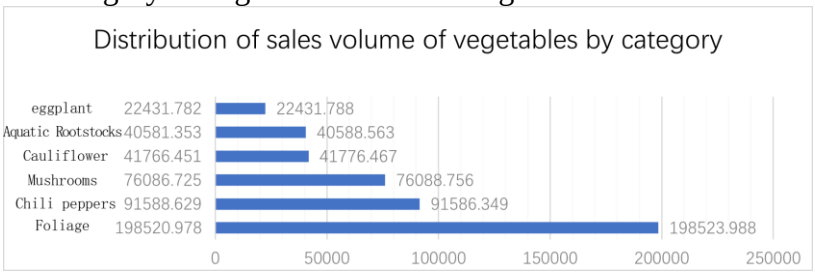

Figure 1: Distribution of sales volume of vegetables by category

It can be seen that "foliage" has the highest sales volume and is the most widely distributed among all categories of vegetables. The sales volume of "eggplant" is the lowest, but second, only to "foliage", "pepper", and "edible mushrooms" sales volume is also as high as 91,586 and 76,088. The sales volume is also as high as 91586 and 76088. Then analyzing the distribution of sales volume of vegetable single product, taking the distribution of sales volume of a single product in the top 5, we can see that the sales volume of "Wuhu green pepper" is the highest, the sales volume of "net root" is located in the second, which is the least distributed. The one with the lowest sales volume is "Xixia Shiitake Mushroom". The specific distribution is shown in Figure 2.

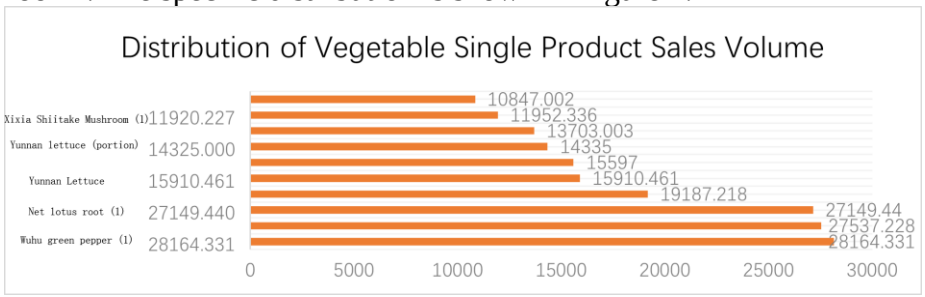

Figure 2: Distribution of sales volume of individual vegetable products

The sales trend analysis of each category of vegetables was carried out to make a relevant trend

graph indicating the distribution of the overall vegetable sales volume by category and by individual product, as shown in Figure 3.

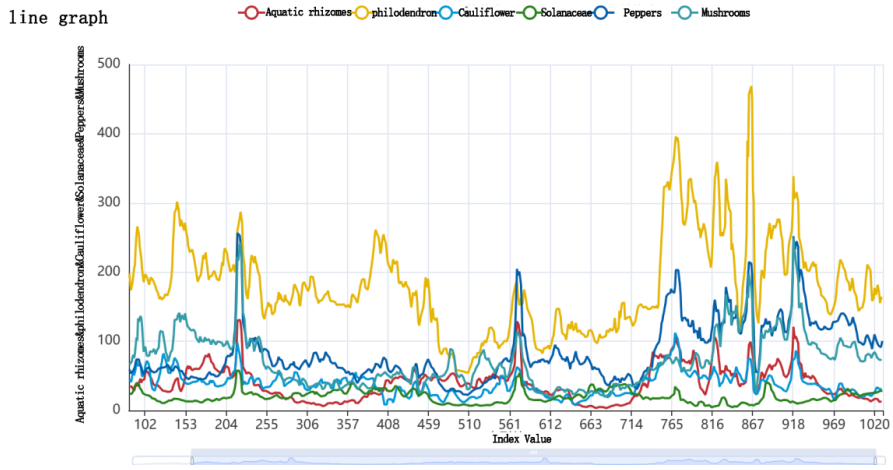

Figure 3: Trends in sales of vegetables by category over time

As can be seen in Figure 3, aquatic roots and tubers, foliage and edible mushrooms have higher sales volumes within the sales timeframe, usually tending to be at their highest values at the three time points 204, 561 and 867. While cauliflower and eggplant sales were lower, both would be in an upward trend at the three specific time points mentioned above. The more stable changes were in cauliflower and peppers.

The Kendall rank correlation coefficient is defined as [1]: let the number of elements of both sets X, and Y be M. The two attendant variables obtaining the ith value are denoted by X_i and Y_i, respectively, which are calculated as shown below:

When there is no X_i=X_j or Y_i=Y_j in the set.

$$\tau = \frac{C-D}{\frac{1}{2}N(N-1)} \tag{1}$$

where C denotes the number of pairs of elements in XY possessed in the same order. D denotes the number of pairs of elements in XY possessed in the reverse order.

When the above relationship exists in the set.

$$\tau = \frac{C-D}{\sqrt{(T_0-T_1)(T_0-T_2)}} \tag{2}$$

By analyzing the sales volume of the categories of vegetables again as a whole, it is possible to calculate Kendall's coefficient of concordance value of 0.772 and the overall data significance is 0.000, which shows significance at the level. Thus, there is a high degree of concordance in the category sales volume of vegetables. The correlation analysis between the two categories of vegetables can be continued.

In the correlation analysis of sales volume of different categories, the Kendall correlation coefficient model was first established to analyze the sample data in two groups. The median, mean, standard deviation, skewness, and other indicators of different categories are statistically indicated indirectly for correlation. The normality test was conducted for different categories of commodities and all the results showed that all the P-values were 0.000***, which indicates that aquatic roots and tubers, foliage, cauliflower, eggplant, chili peppers, and edible mushrooms in the above table have a very high level of statistical significance. So, it is necessary to take the Spearman correlation coefficient further as the method does not have a high requirement for the distribution of original variables.

The heat map of vegetable category correlations is shown in Figure 4. Aquatic rootstocks have moderate to high positive correlation with foliage and eggplant; cauliflower shows a positive correlation with chili peppers, while there is no correlation with edibles, which are weak; eggplant is similar to cauliflower and shows a strong correlation with chili peppers and cauliflower, but also a weak correlation with edibles; chili peppers have weak correlation with edibles and foliage; and aquatic rootstocks have weak correlation with foliage and foliage; and aquatic rootstocks have high positive correlation with foliage, chili peppers category had a high positive correlation. From the overall data, it can be seen that the correlation between most of the indicators is positive, which means that in most cases the sales volume of vegetable categories is synchronized with the increase and decrease.

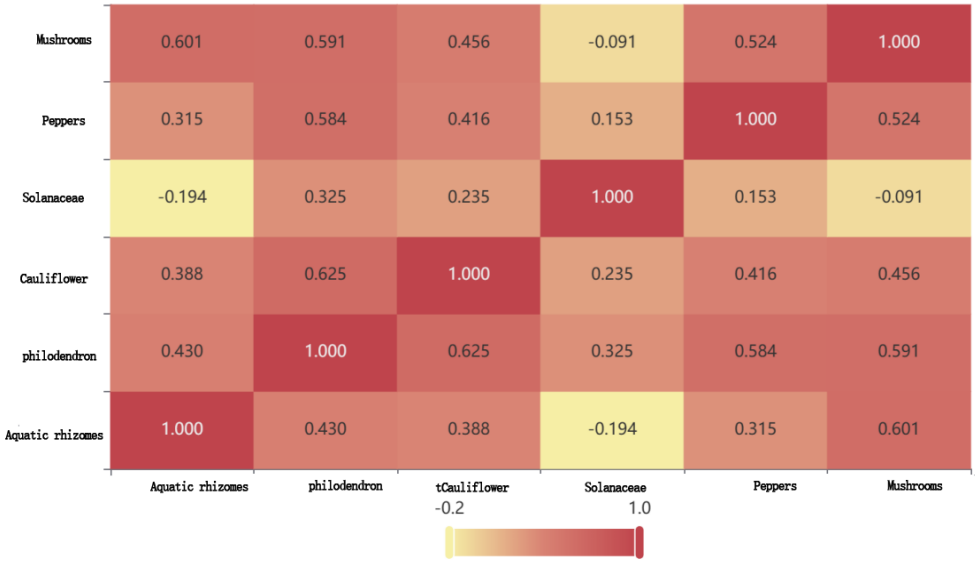

Figure 4: Heat map of correlation of each category of vegetables

In order to further analyze the distribution of the sales of vegetable individual items, we carry out the internal consistency test for the individual items under different categories. For each category, we can know whether there is consistency in the sales of each single product under different time. Since the Kendall's correlation coefficient takes the range of 0-1, if the indicator is closer to 1, it means that the value of the variable is more strongly correlated with me. The closer it is to 0 indicates that the consistency is Weaker, the specific consistency test parameters are as follows. Aquatic roots and stems: 2.5890488784906006e-08, flowers and leaves: 1.2336500355402410e-08, cauliflower: 6.2777095660995080e-08, eggplant: 6.7461120159238439e-08, peppers: 3.5519960901967642e-08, Edible mushrooms: 9.1545083406969823e-09.

It is easy to see from the results that the sales trend for each item may be affected by factors such as seasonality, market activity, inventory, and change. All of these factors can lead to inconsistencies in the resulting sales of individual items. Thus from the consumer's point of view, the consumer can search for the effects of many different factors that lead to errors in the final sales figures. It can also mean a lack of stability and consistency in the daily sales pattern.

2.2 Forecasting and developing replenishment totals and pricing strategies based on GBDT regression

The sample data were categorized according to the date of sale and the mean value, the mean value of the wholesale price and the rate of wastage was found, and then the corresponding unit price of sale, the mean value, the mean value of the wholesale price, and the total volume of sales were used.

Find the average unit price difference per kilogram sold as a percentage of the wholesale price.

Firstly, the future wholesale price is predicted based on LSTM, LSTM mainly includes forgetting gate, input gate and output gate, according to the network structure of LSTM, set F_f, F_i, F_c, F_o, A_f, A_i, A_c, A_o as the model with corner markers, and the formula of LSTM unit is:

$$f_t = \sigma(F_f[h_{t-1}, x_t] + A_f) \tag{3}$$

$$i_t = \sigma(F_i[h_{t-1}, x_t] + A_i) \tag{4}$$

$$\partial_t = tanh(F_C[h_{t-1}, x_t] + A_c) \tag{5}$$

$$C_t = f_t C_{t-1} + i_t \partial_t \tag{6}$$

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

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

For the above prediction model, it is necessary to construct lagged features whose inputs are historical wholesale prices. That is, lagged data from the past day, week or month is created as a new feature. That is, the data from days 1, and 2 is used to predict day 3 and so on. The lagged feature formula is shown below.

$$X_{t-n} = f(F_t, F_{t-1}, \cdots, F_{t-n+1}) \tag{9}$$

where F-t denotes the eigenvalue at time t, F-(t-n) denotes the eigenvalue at time t-n, and f denotes the winning generating function of the lagged feature. The dynamic changes of the time series are incorporated into the features to improve their prediction accuracy.

The GBDT model is established [2-3]. Firstly, the loss function $S(y, f(x))$ is defined, and the minimization loss function can be obtained after specific iterations of the GBDT algorithm.

$$c_{m,j} = argmin \sum_{x_i \epsilon R_{m,j}} L(y_i * f_{m-1}(x_i) + c) \tag{10}$$

where $C_{m,j}$ is the best-fit value of the jth node region of the mth regression tree for loss function minimization. After updating the prediction results, the predicted value of the mth regression tree is obtained $f_m\ (x_i)$.

$$f_m(x_i) = f_{m-1}(x_i) + \sum_{j=1}^{j} c_{w,j}\theta, x \in R_{m,j} \tag{11}$$

The GBDT regression prediction value $F(x)$ is then obtained by summing the $C_{m,}$ and j values over the same leaf node region as

$$F(x) = f_M(x_i) = f_0(x) + \sum_{m=1}^{M} \sum_{j=1}^{J} C_{m+j}\theta, x \in R_{m+j} \tag{12}$$

For the time series prediction using lag one stage as a feature, the model parameters are initially trained using the LSTM network structure, and the optimization method is the small batch gradient descent method. The evaluation parameters of the prediction results are obtained, as shown in Figure 5.

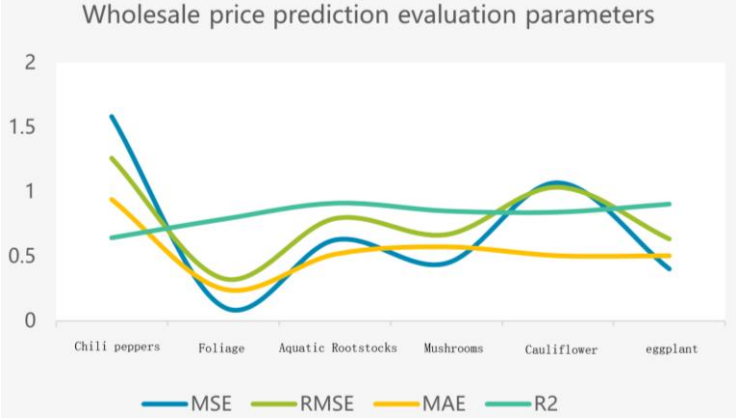

Figure 5: Parameters for evaluation of wholesale price forecast

It can be visualized in the figure that the R2 values of all six vegetable categories exceeded 0.60 and reached a maximum of 0.9124. This indicates that the predicted data of aquatic root and tuber categories have a high accuracy, suggesting that the model has a good predictive ability for all categories.

When using the GBDT prediction model, the sales unit price and wholesale price are the independent variable data, while the sales volume is the dependent variable. The above modeling scheme can be obtained by the smaller the data consisting of the three indicators of the training set MSE, RSME and MAE, the higher the accuracy of the model. In the case of comparing the predicted value with the average value, the closer the result is to 1, the higher the accuracy of the model. And the R2 value of the established GBDT model is close to 1, which can indicate that the model fitting effect is more excellent.

Making a decision that is appropriate for that sales environment requires consideration of a variety of factors such as market competition, consumer demand, and the seasonality of the product being sold, and we intend to determine the optimal sales price of the vegetables in order to maximize the benefits [4]. Let the objective function W:

$$E_{\pi} = pS(p,q,e) - wq + s(q - S_q) - ke^2/2 = p(q - \int_0^{q=d(p,e)} F(\varepsilon)d\varepsilon) - wq + s\int_0^{q=d(p,e)} F(\varepsilon)d\varepsilon - ke^2/2 \quad (13)$$

To ensure the reasonableness as well as the feasibility of the sales price, the modeling also needs to stipulate that the wholesale price must be higher than the sales unit price to ensure that the superstore will not sell goods at a loss in sales. That is, $Y > p$.

In deciding to maximize profit gain, look for a sales unit price strategy X_1, X_2, X_3...., X_7 to maximize the cumulative return for the coming week, however, the daily returns are all determined by the number of sales, the unit price of sales, and the combination of the wholesale price of sales and the attrition rate. The above planning model is solved by the Monte Carlo algorithm [5]. Finally, integration and multiple iterations are performed to obtain the total daily replenishment and pricing strategy for all vegetable categories for the coming week.

After constructing the above target planning model, the loss rate needs to be parameterized i.e., historical average loss rate [6].

The loss rate of leafy vegetable commodities is extremely low, so the validation of the effect of the goal planning model is carried out by making a heuristic convergence diagram for leafy vegetables, as shown in Figure 6.

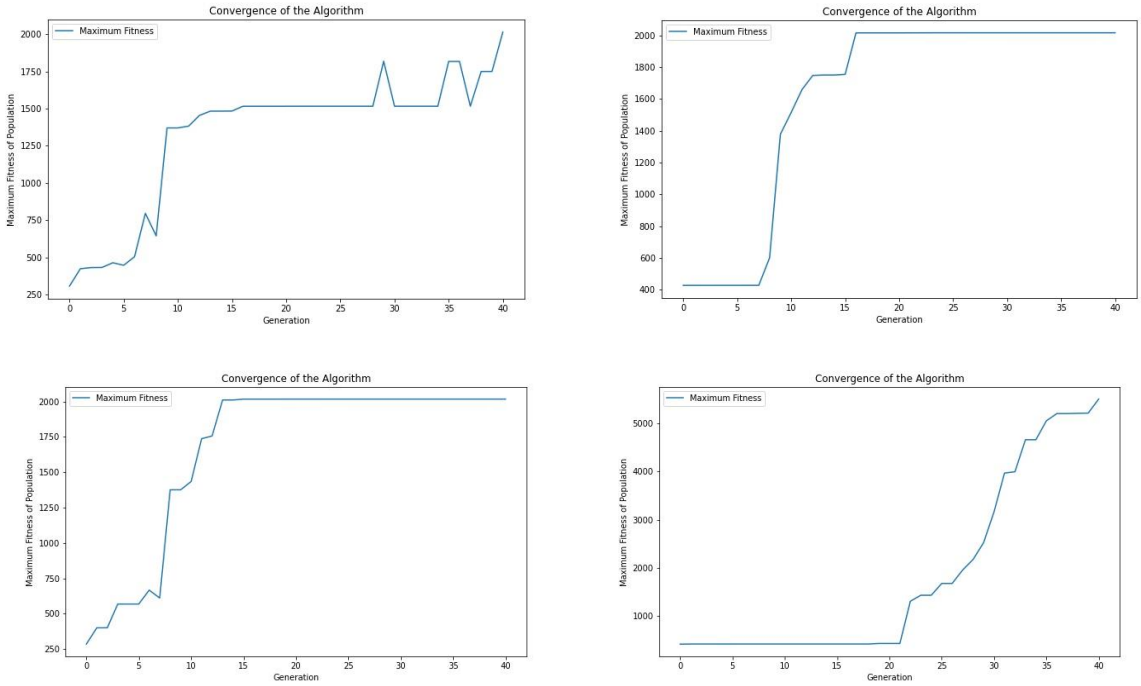

Figure 6: Convergence plot of daily superstore returns for the flower and foliage category

For the foliage class in the future one-week iteration convergence chart as an example, in the above figure, the vertical axis represents the daily super revenue, and the horizontal axis represents the number of iterations, which can be intuitively based on the example of the figure with the number of iterations to increase, the daily super revenue also showed an upward trend, indicating that the planning objectives taken by the model to achieve the optimal.

The total daily replenishment quantity and pricing strategy are formulated. In the case of the planning objective model has been proved to achieve the optimal solution, which can be obtained for each vegetable category in the next week when the supermarket revenue is maximized sales unit price and sales volume, we will sell the number of sales as the daily replenishment, sales unit price as the basis of pricing strategy, and ultimately arrived at the total revenue of each vegetable category, that is, to obtain the maximization of profits for 47,745.626.

2.3 Optimization model of vegetable single product sales and pricing based on XG-Boost

The LightGBM sales volume prediction and XGBoost model [7] are used to predict future wholesale prices, which are then solved by the hypermarket revenue maximization objective planning with the Pso algorithm. We can calculate the sales unit price and sales volume of all vegetable items at the time of hypermarket revenue maximization on July 1, 2023, and use these data as the basis for the total daily replenishment and pricing strategy. Where the goal of maximizing superstore revenue is planned:

The relationship between wholesale price, selling price, and quantity sold is intricate, and to balance these factors and maximize the merchant's profit, a planning solution model is used to illustrate the problem.

Objective function

Given a unit sales price x and a wholesale price W, our objective is to maximize the profit π. The profit is calculated from the quantity sold, the unit sales price, and the wholesale price, minus the cost due to wastage.

The objective function is:

Where S(x Wholesale) is the number of sales predicted by the XGBoost model based on the selling unit price x and the wholesale price W.

The selling unit price must be higher than the wholesale price:

$$X > \text{Wholesale} \tag{14}$$

The number of sales must be greater than 2.5.

$$S(x\ Wholesale) > 2.5 \tag{15}$$

Decision Variables: the main decision variable of 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 on day i.

Parameters and Data g Model The input parameters to the model include a 7-day wholesale price (denoted Wholesale) and a loss rate (denoted $Loss_{rate}$). The Wholesale price is the price at which the merchant purchased the merchandise from the supplier, and the $Loss_{rate}$ represents the percentage of merchandise lost during storage, transportation, and sale. There is also important input data, the number of daily sales predicted based on the unit sales price and wholesale price, which is predicted by an advanced XGBoost machine learning model.

Objective Function: The objective is to maximize the merchant's profit, which consists of the number of units sold, the unit price of sales, and the wholesale price, and requires subtracting the costs incurred due to wastage. This formula considers the complex relationship between quantity sold, wastage rate, and price to maximize profit.

Constraints: two core constraints are set by the model to ensure that the sales price is reasonable and feasible:

1) The selling price must not be lower than the wholesale price to ensure that the merchant will not lose money.

2) The number of sales must be greater than 2.5 to ensure a certain level of sales.

The goal is to find a sales unit price strategy x that maximizes the superstore's revenue on July 1st.

Prediction of sales volume. Since there are as many as 33 single-species categories, we averaged the results of the model evaluation. The results are shown in Figure 7. The r2 value is 0.84, and when r2 is closer to 1 means the model is more accurate, it can be found that the r2 is close to 1 in the average model evaluation, which indicates that the model applies to all categories.

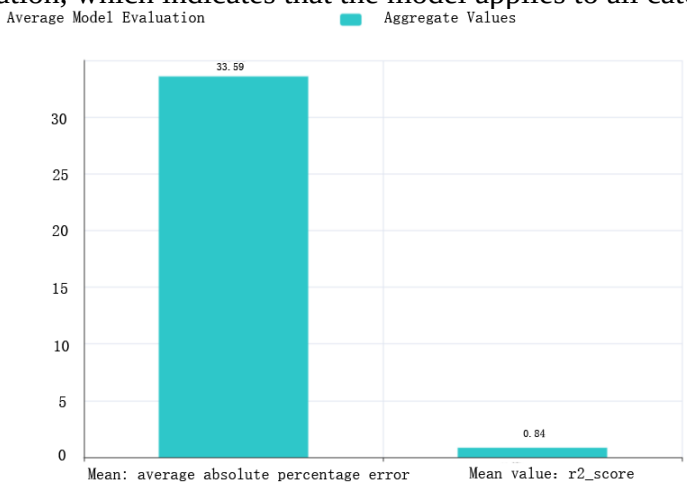

Figure 7: Average model evaluation

The XGBoost categorical wholesale price prediction is shown in Figure 8. As can be seen in the figure, the average single-item R2 values are all over 0.6428, which means that the model can explain

64.28% of the variance in the data of this category, indicating that the model has relatively good prediction ability for all categories.

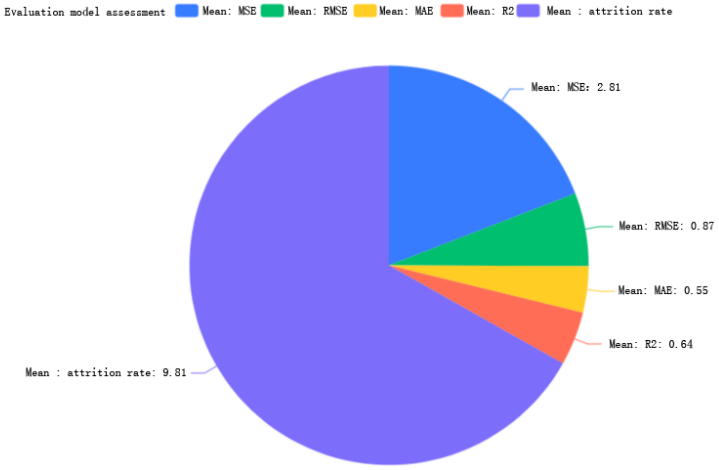

Figure 8: Average model evaluation

The pricing strategy of the superstore and the daily replenishment quantity. By performing the calculations for the above steps, we can find the unit price and quantity sold for each item when the superstore's profit reaches its maximum value on July 1st. Since there is no inventory data provided in the question, we use the sales quantity as the daily replenishment quantity and the sales unit price as the pricing strategy of the item. For these 33 categories, we can find the total revenue of $2949.456 on July 1st.

3. Conclusion

The perishable nature of vegetables determines that the superstore should adopt the "ready-to-sell" sales model, and merchants often have to make replenishment pricing decisions based on historical sales data. This paper analyzes the relevant sales data and develops appropriate pricing and replenishment decisions. First of all, the distribution analysis of various categories of vegetables, is based on the Kendall correlation coefficient test for the consistency of the single product test, and then the correlation analysis between the two categories. Then, the mathematical model for maximizing the revenue of the superstore is constructed. The LSTM time series prediction model is used to predict the wholesale price of each type of vegetables in the coming week based on the historical wholesale price, and then the GBDT sales volume prediction model is used to predict the sales volume of each type of vegetables in the coming week based on the unit price and wholesale price. Finally, the objective function of maximizing the revenue of the superstore is set, and the total revenue of all kinds of vegetables in the coming week is obtained through Monte Carlo algorithm. Finally, to meet the requirements of the total number of individual items and the minimum display quantity, the LightGBM time series prediction model is constructed on the historical wholesale price data to predict the wholesale price of the vegetable category on a single day and establish the objective model for maximizing the revenue of the superstore, which is determined by the daily sales quantity, the sales unit price, and the difference of the wholesale price together with the wastage rate. The proposed model has high solution efficiency and optimality.

References

[1] Hou X, Fang G, Improved ID3 algorithm based on Kendall coefficient 1, Science and Technology Innovation, 2021.
[2] Li S, Research on accurate prediction of heavy overload of transformer based on GBDT algorithm, Modern Information Technology, 2023(4).

[3] Zha W, Dong Y, Jiang Z, Liu Y, A prediction model of manganese content at the endpoint of vacuum self-consumption ingot casting based on GBDT algorithm, Metallurgy in China, 2023 (7).
[4] Chen X, Guo C, Yuan C, Huang W, Liu Z, Intelligent Air Quality Monitoring and Governance Department Based on Monte Carlo Algorithm, Information Recorded Materials, 2022(9).
[5] Pan X, Xie Z, Wang S, Optimal decision making for fresh food superstore preservation efforts and pricing considering loss aversion, Highway Transportation Science and Technology, 2022(6).
[6] Gao Y, Wang W, Wang J. An integrated fuzzy hierarchical classification of LightGBM food safety risk early warning model: a case study of meat products, Food Science, 2021(1).
[7] Zhang X, He L, Zheng J. Research on Sales Forecasting Based on Gray Correlation Analysis and XGBoost Model. Software Journal, 2020, 19(09).