Vegetable Price Forecasting Based on ARIMA Model and Random Forest Prediction

Huishan Zhang[1], Zihao Yan[2, *]

[1]School of Electronic and Information Engineering, Anhui University, Hefei, China, 230031

[2]School of Integrated Circuits, Anhui University, Hefei, China, 230031

*Corresponding author: 3124000474@qq.com

Abstract. Given the trouble that the freshness length of vegetable commodities is incredibly brief and the fine will deteriorate with the expansion in promoting time, we elevate our arithmetic processing based totally on the present records to remedy the fundamental coefficients; utilize the Pearson correlation coefficient evaluation to get the visualization of the warmness map and inside the relationship between the whole quantity of income of every class and the fee plus pricing relationship; due to a large amount of data, multi-feature modeling is proposed using random forests to merge several aspects while predicting revenues, which provides good resistance to noise for most datasets and is less likely to fall into overfitting. Then, the random forest model is applied to predict the revenue volume; the wholesale charges of saleable small products are predicted by the ARIMA model; the complete revenue corresponding to the constant revenue charge of each category is calculated; the procedure is optimized with the help of constraints; and finally, the multi-feature prediction of vegetable demand is made according to the random forest model, which in turn gives the replenishment and pricing strategy. Through the above analysis, vegetable supermarkets can make higher pricing and inventory choices to maximize profits.

Keywords: Vegetable Pricing Strategy, ARIMA Model, Random Forest Prediction.

1. Introduction

In fresh food superstores, vegetable items have a short shelf life, and their quality deteriorates with increasing sales time. To maintain quality and freshness, supermarkets make daily replenishment decisions based on historical sales and demand. Pricing is usually based on "cost-plus pricing", and discounts are offered for damaged and deteriorating merchandise. Market demand analysis plays an important role in replenishment and pricing decisions. Supermarkets need to accurately understand consumer demand to avoid oversupply or overpricing. There is a correlation between the sales volume of vegetables and the time of year, and vegetables are more abundantly available between April and October. In addition, due to the limited sales space in supermarkets, a reasonable sales mix becomes very important. Supermarkets need to reasonably mix different types of vegetable commodities according to consumer demand and commodity availability to meet the purchasing needs of different consumers.

To learn about the replenishment and pricing selection model for vegetable products, Chun-Chin Wei mounted a new multi-period model to decide on a couple of order replenishment choices for a product in a short-sale length [1], which tested the retailer's income function, and supplied the manufacturer's and the complete channel's income characteristic in the provide chain problem. Susana Garrido Azevedo et al. examined the sorts of superior applied sciences used at some stage in the manufacturing furnish chain that underpin the primary strategies of the Supply Chain Operations Reference Model (SCOR) [2]. Also identifies a set of sustainability overall performance warning signs (environmental, monetary, and social) relevant to the evaluation of Supply Chain four (SC4.0). Steffen studied the top-of-line manipulation trouble of pricing and replenishment in serial stock systems [3], and the most excellent trajectories of inventory, replenishment charge, and retail rate are derived by way of the use of a section layout and a formal synthesis procedure.

With the premise of meeting the market demand for every class of vegetable commodities, this paper proposes a polynomial becoming method to predict the market demand. Firstly, the wholesale

cost of on-hand objects is anticipated with the aid of the capability of ARIMA model, and then the earnings volume is estimated with the aid of the skill of Random Forest Model. On this basis, the assumption that "the earnings extent of the previous week is same to the market size can be geared up to the prediction" is proposed, and the polynomial turning into approach is used to portray the market demand from twenty-fourth to thirtieth June 2023, and the market demand on the eighth day is predicted. Finally, primarily based on the prediction consequences and the attainable market size, the 27 most fantastic sellable persona merchandise have been selected, and the corresponding pricing technique and replenishment approach have been developed. Through the model test, it can be proved that this scheme has the right stability.

2. Model

2.1. The structure of ARIMA model

ARIMA model prediction is primarily based on the autocorrelation between the time collection of data, from the almost random time sequence of information to summarise the time-dependent traits and shape of the data. ARIMA is an aggregate of the autoregressive manner (AR), algorithmic transferring common manner (MA), and the distinction algorithm. The difference Algorithm collectively represents a blended model.

The AR model makes use of the historical statistics before every time node to linearly enlarge the node data, which in flip generates an autoregressive illustration between the time collection data. Thus, the records Y_t at the modern second t can be represented as a linear aggregate of $p$preceding historic data:

$$Y_t = \nu + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \epsilon_t \ , \tag{1}$$

Where α_i is the autocorrelation coefficient; ϵ_t is the error between the linear combination of historical data and the data at the current moment; and ν is a constant term.

The MA model is mainly used to reduce the error caused by the random fluctuation part of the time-series data by expressing the data Y_t, at the current moment t as the accumulation of the error generated in the autoregressive process:

$$Y_t = \kappa + \sum_{i=1}^{q} \phi_i \epsilon_{t-i} + \epsilon_t \ , \tag{2}$$

where ϕ_i is the coefficient of the deviation value after adding weights; q represents the current data Y_t which is expanded into a q-order linear combination of errors generated by the autoregressive process; and κ is a constant.

The AR model requires the time collection of information to be smooth. In the case of non-smoothness of the time-series data, it is quintessential to function a couple of distinction calculations on the data. The parameter d is used to denote the wide variety of differencing required when the time collection statistics are changed into an easy series.

Combining the AR model with the MA model that is to obtain the expression of the ARIMA (p, d, q) three-parameter model:

$$D^{(d)} Y_i = \mu + \sum_{i=1}^{p} \alpha_i Y_{i-i} + \sum_{i=1}^{q} \phi_i \epsilon_{i-i} \ , \tag{3}$$

Where $D^{(d)} Y_i$ indecates the d-order difference sequence of time series data Y_i; μ means a constant. The flow of short-term prediction of urban rail transit stations using ARIMA models is shown in Figure 1.

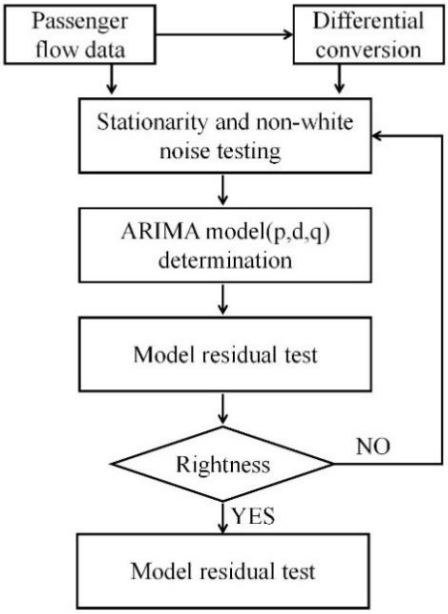

Figure 1 ARIMA model prediction process

2.2. The structure of the random forest model

Random forests and support vector machines can be used in regression problems to avoid overfitting and have good predictive performance. However, when dealing with complex data and large datasets, the parameters may need to be properly tuned for optimal performance [4-7].

The random forest algorithm randomly draws k training samples from the original sample set and obtains a new training set containing k samples (the training sets are independent of each other and the elements can be duplicated) without sample substitution [8], which can be obtained as $\{h_1(X),h_2(X),.....,h_k(X)\}$, where X is the input (experimental parameters of Mo-Nb alloy hot compression: strain ε, deformation temperature T, strain rate s), and Y is the output (flow stress σ). The edge function of the predictor is then calculated or plotted in the random forest classifier and can be expressed as [9]:

$$margins, Y = av_k Ih_k X = Y - maxav_k Ih_k X = j, \quad (4)$$

Where I function means the objective function and ask is the computational average. The edge function represents the difference between the average weight that correctly classifies X (input vector) to Y (output vector) by the integrated classifier and the maximum weight that incorrectly classifies X (input vector) as a vector of other categories. The larger the value of the margin function, the higher the classification reliability, the better the generalization of the model, and the better the prediction ability. When the number of decision trees increases indefinitely, the generalization error of the random forest algorithm will have an upper bound, so as long as the number of decision trees in the random forest model is guaranteed to be large enough, the generalization error is a fixed value, so the random forest can achieve good prediction for new sample sets and will not be overfitting [10].

In this paper, the Random Forest Algorithm (RFA) is used for multi-feature modelling. Wholesale price, promoting price, seasonal facets, and temporal facets such as whether or not it is a working day are merged collectively to predict the income volume, firstly, the class facts are divided into a coaching set and check set, then the facts are educated on the coaching set and examined on the check set, and the take a look at consequences are sooner or later outputted. The unique go with the flow chart is proven in Figure 2 below.

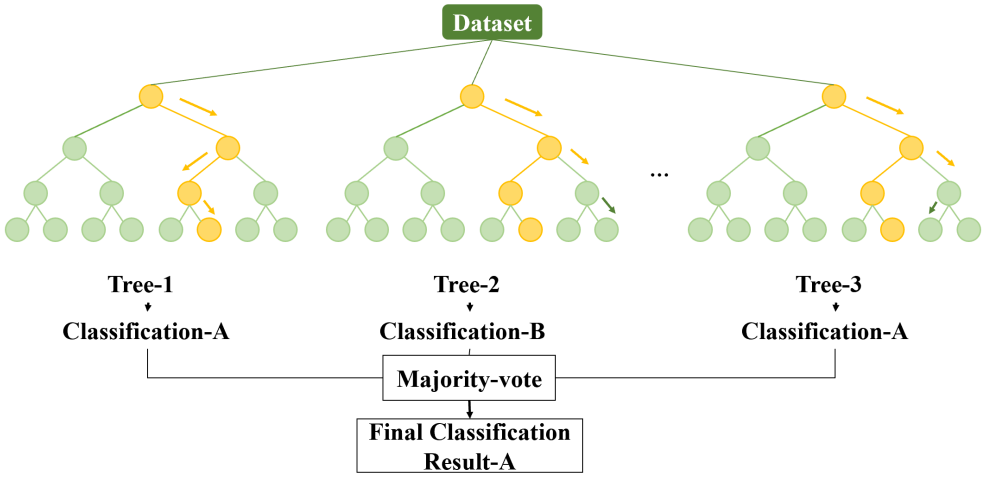

Figure 2 Flowchart of Random Forest algorithm

3. Results

3.1. Data preprocessing

The number of purchases for all categories is shown in Figure 3, and the comparison reveals that there is a big difference between the box line graph and the histogram of purchase frequency, and it is seen from the box plot that the higher sales volume is for aquatic roots and foliage, and the lower one is for edible mushrooms. A comprehensive analysis of the histogram and box line graph shows that although some categories and individual items have a high frequency of sales, the actual number of purchases is not much, and the core of the decision to operate is the sales volume, which will be analyzed and discussed in the subsequent part of this paper.

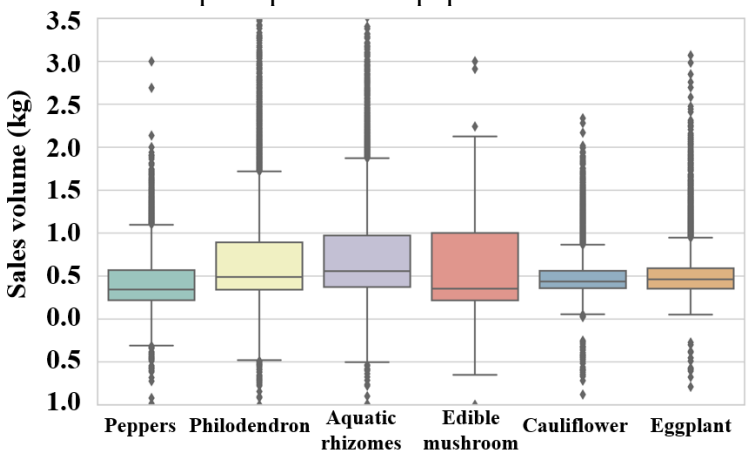

Figure 3 Boxplot of the number of purchases across all categories

A visual analysis of the individual items under each category, shown in Figure 4, is a box line diagram of the individual items in the chili category, which shows that the higher sales are of Wuhu green peppers and red peppers, followed by screw peppers and green sharp peppers. There is a large difference in sales of different chili individual items, and there are some chilies that almost no one buys.

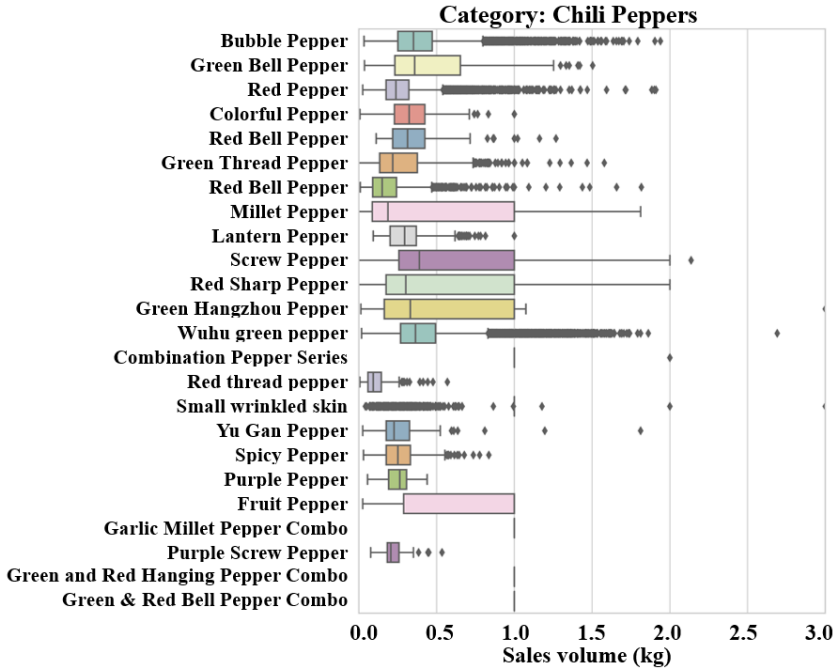

Figure 4 Chili Pepper Single Product Box Line Chart

This paper first calculates the cost margin of cost-plus pricing for each product, weights the category wholesale price, sales price, and cost margin according to the sales volume of each product, and then analyzes the relationship between category sales and cost-plus pricing using Spearman's correlation test to analyze the relationship between the total sales volume and cost-plus pricing for each vegetable category, and the relationship between the profit-sales volume of each category can be visualized as shown in Figure 5 below:

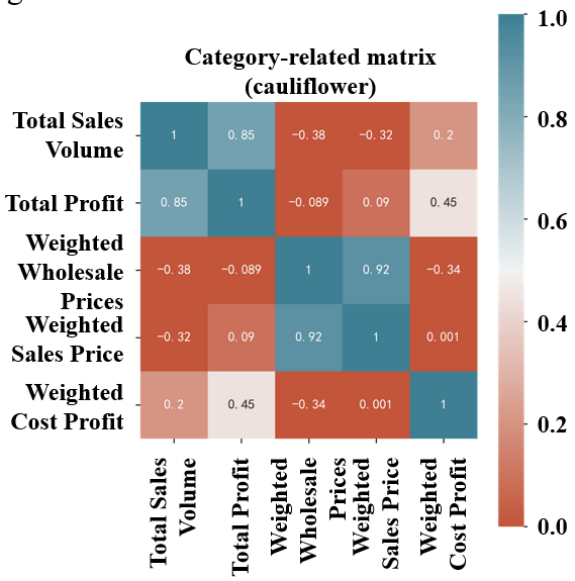

Figure 5 Visualisation of heat maps

3.2. Results obtained from the ARIMA model

The wholesale charge of every class is derived, so this paper predicts the subsequent seven-day style by constructing an ARIMA time sequence model. Due to the fact the charge base of one-of-a-kind classes is no longer the same, it is integral to educate the ARIMA model for special classes separately. In this paper, the chili pepper class as an example, imports the complete time income facts obtained, studies the impact of the time sequence plot above, by the time collection plot can essentially decide that the sequence is now not clean, and then via the autocorrelation plan in Figure 6 to confirm a little:

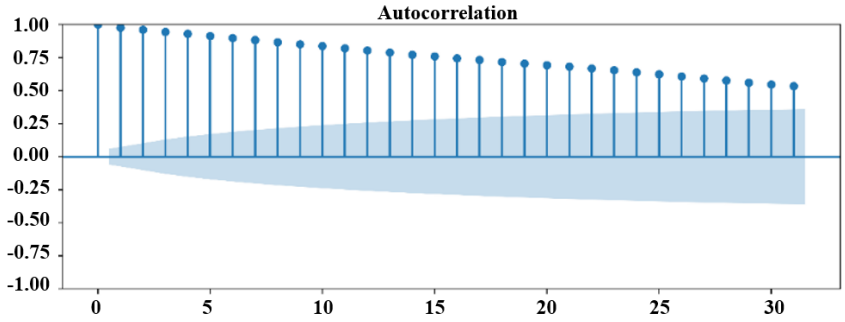

Figure 6 Chilli category autocorrelation chart

Looking from right to left in the autocorrelation diagram, the autocorrelation coefficient slowly decays to zero as the time series order increases, which does not meet the requirement of a smooth series. The ADF test is performed on the time series, and the unit root test statistic corresponds to a p-value that is significantly greater than 0.05, and the time series is finally judged to be non-stationary.

Next, we perform a first-order difference on the time series data, and then again perform a smoothness test on the differenced series; if it does not meet the smoothness, then perform a second-order difference and then perform the test. Repeat the above process to finally get a determined ARIMA with three parameters, denoted as ARIMA(p, d, q). p represents the autoregressive order, which represents the data for the period we need to do the autoregression. d represents the number of differencing done when the time becomes smooth, and q represents the moving average order. Next, the partial autocorrelation plot of Figure 7 is plotted to determine the remaining parameters:

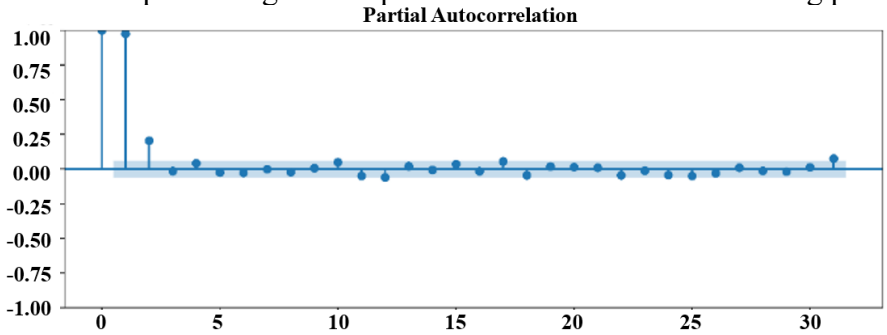

Figure 7 Chilli category biased autocorrelation map

Through the partial autocorrelation diagram, we can see that there is more than one sort of parameter to determine, so we can traverse every p, and q parameters to get the corresponding AIC and BIC which obtained as shown in equation 5 and 6, and different data to decide the parameters. pic (Akaike data content), and BIC (Bayesian statistics content) are the standards for judging the model, and the smaller the price suggests that the model is better, and its calculation method is:

$$AIC=2k-2\ln(L) \tag{5}$$

$$BIC=k\times\ln(n)-2\ln(L) \tag{6}$$

In addition, we need to establish the relationship between price and sales volume. Here, the least squares method is used to establish a linear regression model as shown in Figure 8, and observe the model effect. It can be observed from the model that the sales price model of chili peppers conforms to a normal distribution, but R2 = 0.022 in the demand price curve, which hardly meets the demand for accuracy of prediction, and then the random forest model is introduced for further prediction analysis.

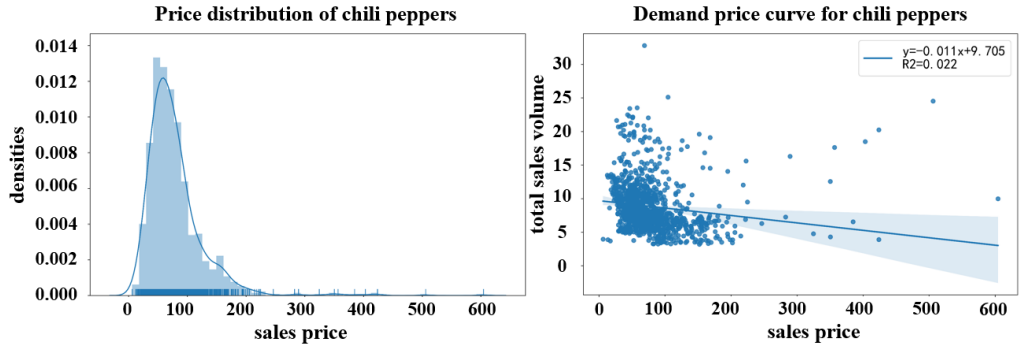

Figure 8 Price distribution and demand rate curve of Chilli class, (a) Price distribution of chili peppers, (b) Demand price distribution curve

3.3. Results obtained from the random forest model

In this paper, when setting the parameters of Random Forest, N estimators, Criterion, Max depth, and Min samples leaf are set to 100, Gini, None, and 1. Since Random Forest can recognize different categories, only one model needs to be trained to get the results, and the results of the training set and the test set are as follows 0.9422 and 0.6315.

It can be found that the effect of the random forest model is significantly improved after the introduction of other features. Compared with the single linear regression, although the MAE score on the test set is 0.6315, this result is satisfactory enough. The research team has tried to find other characteristics to explain the sales volume of each category, but no valid results have been obtained. Therefore, derived wholesale prices (predicted by the ARIMA model) are used, combined with time characteristics such as season and whether or not the working day is used to predict sales. After predicting the sales volume, calculate the total profit corresponding to the fixed sales price. Therefore, this problem is transformed into an optimization problem.

$$\max_{\hat{s}^j} \hat{t}^j * (s^j - w^j) \tag{7}$$

$$s.t\, w^j \le s^j \le 3w^j \tag{8}$$

$$m^j \hat{t}^j \le b^j \tag{9}$$

Where $\hat{t}^j$ represents the predicted sales volume, $\hat{s}^j$ represents the pricing strategy, $\hat{w}^j$ represents the predicted wholesale price given by the ARIMA model, m^j represents the wear rate of category j, and b^j represents the replenishment strategy.

In the first constraint, because random forest may produce some abnormal changes when it inputs unseen eigenvalues, for example, a very, very large price may predict negative or very large sales volume, this paper will set the optimal pricing strategy within the range of three times the cost; In the second constraint, since the random forest only predicts the sales volume, there will be some losses in the process of transportation and sales, and this loss should also be taken into account in the whole process.

The optimization function is a linear optimization problem, which can be solved by selecting a simple linear optimization package or using a grid search strategy. As a result, the final purchase results are obtained, part of which are shown in Table 1. In this table, Cat_name indicates category name, Wei_who_price means Weighted wholesale price; wor_day stands for the working day, 1 means not a working day, 0 represents not a working day; Cat_code stands for Category Name_code; Rep is Replenishment; Pri_str stands for Pricing Strategy; Pro_rev stands for Projected revenue.

Table 1 Pricing and replenishment strategies

Cat_name	Date of sale	Wei-pri	Wor_day	Holiday	Season	Cat_code	Rep	Pri_Str	Pro_rev

Capsicum	2023-07-01	3.30	0	1	1	4	77.86	6.55	253.09
Capsicum	2023-07-07	3.30	1	0	1	4	68.48	6.26	202.18
Solanaceae	2023-07-01	4.50	0	1	1	3	41.13	7.96	141.95
Solanaceae	2023-07-07	4.58	1	0	1	3	28.12	7.87	92.56
Cauliflower	2023-07-01	7.79	0	1	1	2	51.20	12.89	260.87
Cauliflower	2023-07-07	7.81	1	0	1	2	36.20	12.91	184.76

4. Conclusions and Outlooks

In this paper, the ARIMA model and the random forest model are used to predict vegetable prices and give pricing strategies. By introducing other features, the effectiveness of the random forest model is significantly improved, and the MAE score of the testing machine is 0.6315. Using the wholesale price predicted by the ARIMA model, after combining the seasonal and other factors to predict the sales volume, we calculate the fixed sales price and the corresponding total profit, and transform the problem into an optimization problem, and finally give the optimal pricing strategy.

Since the ARIMA model can only capture linear relationships in nature and cannot capture nonlinear relationships, the fit is not absolutely ideal for this paper. To address these issues, this paper proposes the following outlook: data preprocessing methods such as logarithmic transformation can be used in the ARIMA model to deal with non-smooth data. For nonlinear relationships, higher-order autoregressive moving average models, such as the NARIMA model, can be used to introduce exponential growth and so on to enhance the nonlinear fitting ability of the model. In random forest models, the overfitting problem can be mitigated by increasing the number of training samples to reduce the overfitting of the model. The risk of overfitting in random forest models can be reduced by removing some too deep or unnecessary decision trees through pruning techniques. The negative impact of noisy data on random forest can be reduced by removing the noisy data by using methods such as data cleaning and outlier removal.

References

[1] Wei, C.-C.; Chen, L.-T. Supply Chain Replenishment Decision for Newsvendor Products with Multiple Periods and a Short Life Cycle. Sustainability, 2021, 13, 12777.

[2] Azevedo S G, Pimentel C M O, Alves A C, et al. Support of advanced technologies in supply chain processes and sustainability impact[J]. Applied Sciences, 2021, 11(7): 3026.

[3] Jørgensen, S., & Kort, P. M. Optimal pricing and inventory policies: Centralized and decentralized decision-making. European Journal of Operational Research, 2002, 138(3), 578–600

[4] Khorshidvand B, Soleimani H, Sibdari S, et al. Revenue management in a multi-level multi-channel supply chain considering pricing, greening, and advertising decisions[J]. Journal of Retailing and Consumer Services, 2021, 59: 102425.

[5] LIU Jian-yi, LU Jiang, CHEN Yi-zhao, et al. Study on prediction model of liquid holds up based on random forest algorithm[J]. Chemical Engineering Science, 2023, 268: 118383.

[6] BERTIN Takoutsing, GERARD B.M. Heuvelink. Comparing the prediction performance, uncertainty quantification, and extrapolation potential of regression kriging and random forest while accounting for soil measurement errors[J]. Geoderma, 2022, 428: 116192.

[7] MICHAEL Parzinger, LUCIA Hanfstaengl, FERDINAND Sigg, et al. Comparison of different training data sets from simulation and experimental measurement with artificial users for occupancy detection using machine learning methods Random Forest and LASSO[J]. Building and Environment, 2022, 223: 109313.

[8] GENG Pei, GONG Xiao-tao, CHEN Wen-jing, et al. Hot Deformation Behavior of Zr-2.5Nb[J]. Journal of Netshape Forming Engineering, 2022, 14(6): 65-70.

[9] ZHOU Jian, DAI Yong, TAO Ming, et al. Estimating the mean cutting force of conical picks using random forest with salp swarm algorithm[J]. Results in Engineering, 2023, 17: 100892.

[10] LEO Breiman. Schapire. Random Forests[J]. Machine Learning, 2001,45: 5-32.

[11] GAO Wei, XU Fan, ZHOU zhi-hua. Towards convergence rate analysis of random forests for classification[J]. Artificial Intelligence, 2022, 313: 103788.