Advances in Artificial Intelligence, Big Data and Algorithms
G. Grigoras and P. Lorenz (Eds.)

doi:10.3233/FAIA230929

Research on Product Sales Forecasting Model Based on Machine Learning

Pengcheng WANG[1]
School of Software, Beihang University, China

Abstract. In the cost management of supply chain, inventory management is particularly important. In order to help suppliers better manage inventory and reduce inventory turnover cost, this paper will study the commodity sales forecasting model based on random forest regression algorithm and genetic algorithm in the field of machine learning, so as to dynamically optimize and guide inventory management. First, variance selection method is used for the initial feature selection. Then the data set is trained and predicted using the random forest regression algorithm, and the feature importance is sorted, and the second feature selection is carried out. Then genetic algorithm is used to optimize the hyperparameters of random forest regression algorithm to get the best combination of hyperparameters. Finally, the optimal combination of hyperparameters is used to construct a model for the selected important features, and the sales volume of the product is forecasted. By comparing the evaluation indexes of the model, it is found that using the optimized parameters and important feature data for training, the prediction effect is better.

Keywords. Random forest regression algorithm, Genetic algorithm, Sales forecasting model

1. Introduction

For suppliers, inventory management is a very important issue, the more inventory, the more funds occupied, if the inventory is too small, there will be a break in goods, goods supply is not on, resulting in lower sales. This paper will study relevant algorithms and use historical data to forecast commodity sales, so as to help suppliers upgrade inventory management mode, prepare inventory in advance according to sales forecast results, and reduce inventory costs. Therefore, it is of great significance to construct commodity sales forecasting model based on machine learning algorithm.

There are many machine learning algorithm models that can be used to predict product sales, such as regression trees, support vector machines, the Bagging series of random forests [1-2, 4, 7, 8, 9], and the Boosting series of AdaBoost, XGBoost [10], LightGBM [11], and CatBoosting [3], and some neural network models [5].

Li Shuang improves the random forest algorithm in the machine learning algorithm and proposes a Bayesian optimization method combined with time series segmentation to optimize the random forest model, so as to improve the forecasting effect of commodity sales [8]. Huang Wenyi et al. predicted the sales of O2O takeout business based on incremental random forest algorithm. Firstly, the important features that contribute most to the prediction accuracy are identified by removing the noise features

[1] Corresponding Author. Pengcheng WANG, School of Software, Beihang University, China; E-mail: pch.wang@outlook.com.

to improve the accuracy of the prediction. Secondly, by adding incremental features, the incremental method based on random forest is used to forecast sales volume, and the forecasting error is further controlled [9]. Ji Shouwen et al. took into account the sales characteristics of goods and the trend of data series, combined the two-step clustering algorithm and ARIMA model on the basis of XGBoost model, assigned weights to the prediction results, and obtained the final prediction results through weighted calculation [10]. Deng Tingyan et al. realized Walmart's sales forecast based on LightGBM model. In the feature engineering stage, some features unrelated to the model input are deleted, and then statistical methods are used to extract and classify the features, and statistics such as the mean value and standard difference of the features are obtained as incremental features to improve the prediction effect. Finally, LightGBM model is used for training and prediction [11].

The main research of this paper is to use random forest regression algorithm for feature selection, and then use genetic algorithm [6] for hyperparameter optimization to improve the accuracy of commodity sales prediction.

2. Experimental Analysis

2.1. Data Set

The data set of this experiment comes from the product sales data of Yijia company in 2022. The data set contains 13 features such as product number, year, month, day and price, and finally the sales volume of the day is used as the label.

2.2. Feature Engineering

Through the analysis of the data set, it is found that all the data are complete, there are no missing values, and no special processing is needed. One of the features, "week", is encoded by unique heat. Finally, the variance selection method is used to filter out the features whose eigenvalues have little fluctuation. Finally, we get 7 basic features, plus 7 unique thermal coding features, a total of 14 feature dimensions.

2.3. Build Model

Firstly, according to the ratio of 8:2, the whole data set is divided into two parts, the training set and the test set. Then, RandomForestRegressor with sklearn is used to build a random forest regression model with default parameters for training data, and then make predictions for test set and training set respectively. The prediction result is evaluated by R2 and MAE, and the predicted R2 in the test set is 0.84 and the MAE is 19.49. The R2 and MAE predicted on the training set are 0.97 and 6.97 respectively. The index of prediction results is good, but the index of prediction on the training set is obviously better than that on the test set, and overfitting phenomenon appears.

In addition, since the random forest regression algorithm carries out feature importance judgment when splitting nodes, the feature importance ranking can be directly obtained after the model is constructed. The importance distribution of features is shown in Figure 1:

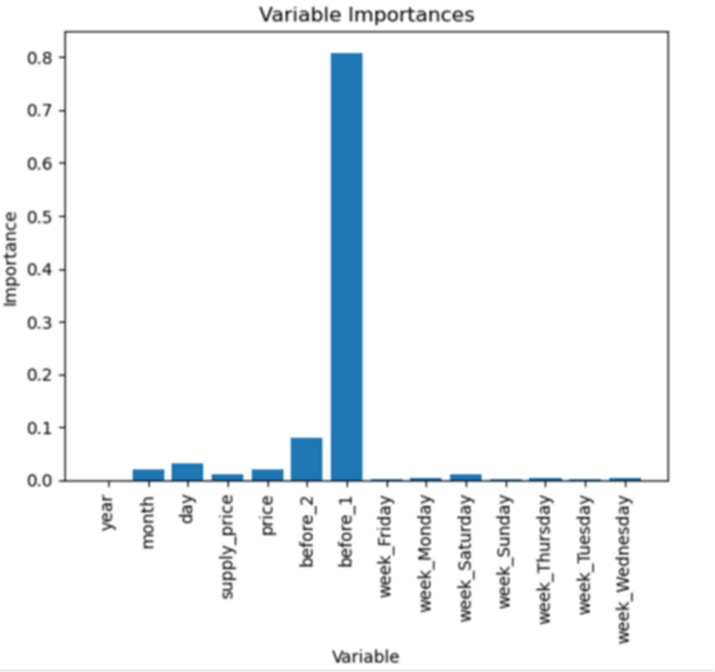


Figure 1. Distribution of feature importance.

Then, the first five most important features were selected according to the importance of the features, the random forest regression model was reconstructed, retraining and repredicting, and the comparative analysis was carried out. The result indicators of commodity sales forecasting at different characteristic latitudes are shown in Table 1:

Table 1. Product sales forecast results index of different characteristic latitudes.

		Total feature	Important feature
Test set	**R2**	0.84	0.86
	MAE	19.49	18.6
Training set	**R2**	0.97	0.97
	MAE	6.97	6.77

The comparison of data in the table shows the importance of features is sorted and selected in the training process of the random forest model, and the important features after selection are used to train and predict the model, and the prediction effect is better, but the overfitting phenomenon still exists.

2.4. Hyperparameter Optimization

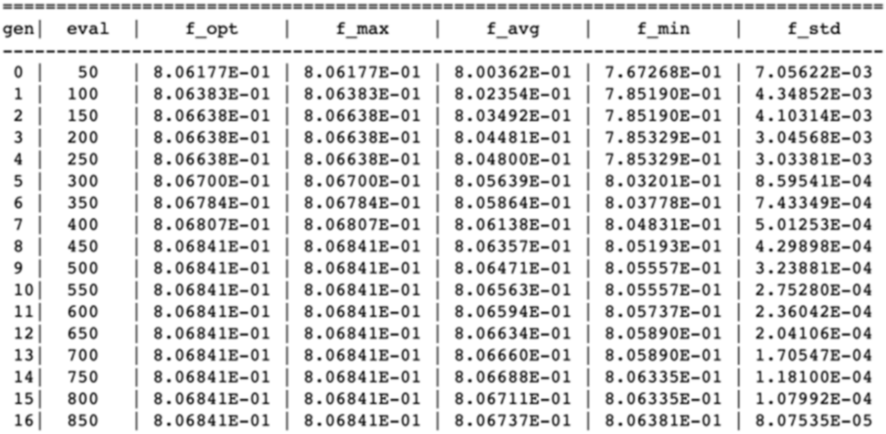
The random forest has more than a dozen hyperparameters, and the prediction effect of different parameter combinations is different. Here, genetic algorithm is selected to optimize the more important 5 parameters. The relevant description and parameter range of the parameters are shown in Table 2:

Table 2. Related description and tuning range of random forest hyperparameters.

Parameter name	Description	Default value	Tuning range
min_samples_split	The minimum number of samples required to split an internal node	2	2~10
min_samples_leaf	The minimum number of samples required for leaf nodes	1	1~10
max_leaf_nodes	Maximum number of leaf nodes	Unlimited	100~10000
max_features	The number of features to consider when finding the best segmentation	1	1~10

First, the values of the above 5 parameters are randomly combined to generate 50 chromosomes, and the initial population is constructed, with each parameter value as a code in the chromosome. Then, a random forest model was constructed for each parameter combination and trained. The average R2 value was calculated using the five-fold cross-validation as the objective function value and fitness of the genetic algorithm, and their fitness was sorted. The elite replication selection strategy was adopted to select the top 50% of the fitness combinations and retain them, and then some of them were randomly selected. Replicate, cross and mutate the parameter values, build a new combination, fill the remaining 50%, form a new population, and continue to evolve. Until the set maximum evolutionary algebra is exceeded, or the objective function has stalled, has approached the optimal value, evolution stops, and the best combination of them is selected as the result of hyperparameter optimization.

In this experiment, a total of 17 generations were evolved, with 50 combinations in each generation, and a total of 850 groups of hyperparameter combinations were evaluated for fitness. Figure 2 shows the maximum, minimum, average and standard deviation of the objective function values in each generation. It can be seen that the values of f_opt and f_max are the same due to the adoption of the elite replication selection strategy. Each generation selects the maximum objective function value. The fitness evaluation times and objective function values of each generation are shown in Figure 2:



```
==============================================================================
gen|  eval  |     f_opt     |     f_max     |     f_avg     |     f_min     |     f_std
------------------------------------------------------------------------------
 0 |   50   | 8.06177E-01 | 8.06177E-01 | 8.00362E-01 | 7.67268E-01 | 7.05622E-03
 1 |  100   | 8.06383E-01 | 8.06383E-01 | 8.02354E-01 | 7.85190E-01 | 4.34852E-03
 2 |  150   | 8.06638E-01 | 8.06638E-01 | 8.03492E-01 | 7.85190E-01 | 4.10314E-03
 3 |  200   | 8.06638E-01 | 8.06638E-01 | 8.04481E-01 | 7.85329E-01 | 3.04568E-03
 4 |  250   | 8.06638E-01 | 8.06638E-01 | 8.04800E-01 | 7.85329E-01 | 3.03381E-03
 5 |  300   | 8.06700E-01 | 8.06700E-01 | 8.05639E-01 | 8.03201E-01 | 8.59541E-04
 6 |  350   | 8.06784E-01 | 8.06784E-01 | 8.05864E-01 | 8.03778E-01 | 7.43349E-04
 7 |  400   | 8.06807E-01 | 8.06807E-01 | 8.06138E-01 | 8.04831E-01 | 5.01253E-04
 8 |  450   | 8.06841E-01 | 8.06841E-01 | 8.06357E-01 | 8.05193E-01 | 4.29898E-04
 9 |  500   | 8.06841E-01 | 8.06841E-01 | 8.06471E-01 | 8.05557E-01 | 3.23881E-04
10|  550   | 8.06841E-01 | 8.06841E-01 | 8.06563E-01 | 8.05557E-01 | 2.75280E-04
11|  600   | 8.06841E-01 | 8.06841E-01 | 8.06594E-01 | 8.05737E-01 | 2.36042E-04
12|  650   | 8.06841E-01 | 8.06841E-01 | 8.06634E-01 | 8.05890E-01 | 2.04106E-04
13|  700   | 8.06841E-01 | 8.06841E-01 | 8.06660E-01 | 8.05890E-01 | 1.70547E-04
14|  750   | 8.06841E-01 | 8.06841E-01 | 8.06688E-01 | 8.06335E-01 | 1.18100E-04
15|  800   | 8.06841E-01 | 8.06841E-01 | 8.06711E-01 | 8.06335E-01 | 1.07992E-04
16|  850   | 8.06841E-01 | 8.06841E-01 | 8.06737E-01 | 8.06381E-01 | 8.07535E-05
```

Figure 2. Fitness evaluation times and objective function value of each generation of genetic algorithm.

Using these data to draw a line graph, it can be clearly seen that with the iterative evolution of the population, the maximum objective function value becomes larger and larger, and eventually tends to stabilize, approaching the optimal value. The average value of objective function of each generation population is also increasing and finally approaching the optimal objective function value. The changing trend of the maximum and average value of each generation is shown in Figure 3:

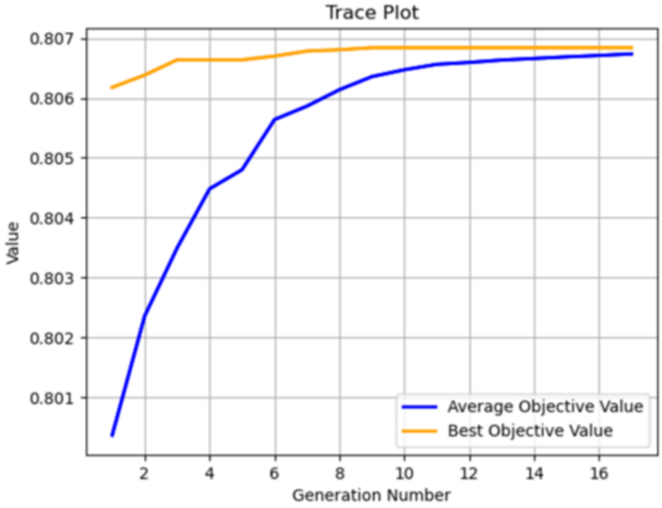


Figure 3. Variation trend of maximum and average value of objective function in each generation of genetic algorithm.

Finally, the optimal super parameter values of random forest regression optimized by genetic algorithm are: 266, 9, 2, 2676, 9. These values are then used to rebuild the random forest regression model, retraining and repredicting, and the obtained results are compared with the model built with default parameters, as shown in Table 3:

Table 3. Forecasting results of different hyperparameters of commodity sales.

		Default parameter	Optimization parameter
Test set	R2	0.86	0.91
	MAE	18.6	10.23
Training set	R2	0.97	0.95
	MAE	6.77	8.57

The data in the table means that using the parameters selected by the genetic algorithm, the prediction effect of the test set is improved, while the prediction effect of the training set is decreased, which avoids overfitting to a certain extent. At the same time, the parameters optimized by genetic algorithm can enhance the model ability of generalization, and the optimization effect is remarkable.

2.5. Fitting Effect

The predicted value can be plotted on the line chart of the real value of the data set, and the fitting effect between the predicted value and the real value can be visually observed. The fitting effect is shown in Figure 4:

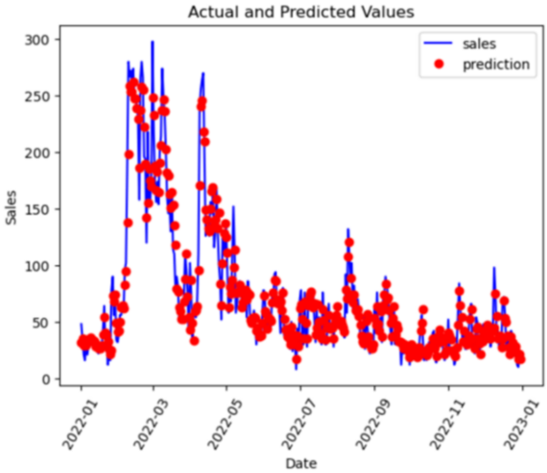


Figure 4. A fitting rendering of the predicted and true values.

3. Conclusion

Through experimental analysis, the prediction effect of random forest regression model is effectively improved after using random forest regression algorithm for feature selection and genetic algorithm for hyperparameter optimization. Suppliers can reduce inventory costs by managing inventory in advance based on forecasted commodity sales.

References

[1] Naik Hruthvik, Yashwanth Kakumanu, Suraj P, Jayapandian N. Machine Learning based Food Sales Prediction using Random Forest Regression. ICECA 6 (2022), 195-198.
[2] Chavare Ranveer, Joshi Rushikesh, Wagh Om, Vaishale Aditya, Ingale Aditya. Car Sales Price Prediction using MLR, Random Forest and Support Vector Machine. ICONAT (2023).
[3] Ding Jingyi, Chen Ziqing, Xiaolong Li, Lai Baoxin. Sales Forecasting Based on CatBoost. ITCA 2 (2020), 636-639.
[4] Adetunji Abigail Bola, Akande Oluwatobi Noah, Ajala Funmilola Alaba, Oyewo Ololade, Akande Yetunde Faith, Oluwadara Gbenle. House Price Prediction using Random Forest Machine Learning Technique. Procedia Computer Science 199 (2021), 806-813.
[5] TaiCheng Wei, Yanbing Liu, Mimi Zhang, Shenshen Liu, Ning Li. Commodity sales forecasting based on spatiotemporal graph neural network. Computer system application 32 (2023), 52-65.
[6] Yuandong Cheng, Yiwei Yang, Jun Yan. Research on path planning based on hybrid adaptive elite genetic algorithm. Journal of Hubei University for Nationalities 41 (2023), 51-57.
[7] Ying Chen. Prediction of auto credit default based on grid search and random forest. Technology and industry 23 (2023), 116–121.
[8] Li Shuang. Sales Forecasting Model of E-commerce Activities Based on Improved Random Forest Algorithm. ICCGIV 2 (2022), 195-198.
[9] Huang Wenyi, Xiao Qin, Dai Hongyan, Yan Nina. Sales Forecast for O2O Services - Based on Incremental Random Forest Method. ICSSSM 15 (2018).
[10] Ji Shouwen, Wang Xiaojing, Zhao Wenpeng, Guo Dong. An application of a three-stage XGboost-based model to sales forecasting of a cross-border e-commerce enterprise. Mathematical Problems in Engineering (2019).
[11] Deng Tingyan, Zhao Yu, Wang Shunxian, Yu Hongjun. Sales Forecasting Based on LightGBM. ICCECE (2021), 383-386.