Application of LightGBM and LSTM combined model in vegetable sales forecast

Application of LightGBM and LSTM combined model in vegetable sales forecast

Zhang He[1, a], Sun Yu[1*]

[1]Yunnan Normal University, Kunming, Yunnan Province, 650500, China

*Corresponding author's e-mail: sunyu_km@hotmail.com

[a]zhaoban@ynnu.edu.cn

Abstract: In view of the nonlinear and linear influence of vegetables in sales forecasting, the previous single model could not fully explore the variation law of vegetables in sales, a combined model based on LightGBM and LSTM is proposed, and the significant and abstract characteristics affecting sales forecasting are explored respectively by combining the advantages of the two models. First, the LightGBM model and the LSTM model based on intensive learning are modeled and analyzed, and then the two models are weighted array by the error reciprocal method for sales forecasting. The experimental results show that the proposed combination model based on LightGBM and LSTM is more accurate than the single model, and the prediction results of the model for vegetable short-term sales have provided important reference value for the strategic marketing of enterprises.

1. Introduction

With the advent of the era of artificial intelligence, companies are facing various cost challenges. Although the modern management of the catering industry has become increasingly mature, the large backlog of dishes has led to rot and waste, thereby reducing profits, or insufficient supply of dishes. Satisfying the needs of consumers has led to a decline in turnover and customer satisfaction. Nowadays, many supermarkets use manual forecasting methods, and sales staff judge the quantity of vegetables they need to purchase the next day based on their own experience. Because the sales of vegetables are affected by many factors, the accuracy of human objective judgments is very low, which has resulted in insufficient supply of certain dishes and shortage of goods. Therefore, if an enterprise wants to stand out in the market, it must gradually develop from a traditional sales management model to an informationized and automated management model.

Common sales forecasts can be divided into linear forecasts [1-3] and nonlinear forecasts [4-6]. For example, Ramos et al. [7] used ARIMA and state-space models to predict the future sales volume of commodities, and Yan Bo et al. [8] used ARMA-based sales forecasting methods. However, sales forecasts are not only related to historical sales, but also some external Abstract features are also relevant. With the widespread application of machine learning in the prediction of time series problems, for example, Bi Liyuan et al. [9] proposed a database query cost prediction based on recurrent neural networks. This method uses the self-learning ability of each neuron to realize the input and output of sample data. This method realizes the non-linear relationship between the input and output of sample data through the self-learning ability of each neuron. It has certain non-linear mapping ability in prediction, but the general ability is relatively weak.

On the basis of the above research, this paper conducts an experiment on the sales volume of a

certain vegetable in a large fresh vegetable supermarket for two consecutive years, and proposes a prediction method based on the LightGBM and LSTM combined model. First, the original time series sales data are standardized. After data preprocessing, LightGBM and LSTM neural networks were used to model the time series data of potato vegetable sales in the supermarket in the past three years. Finally, the average absolute percentage error (MAPE) of the vegetable sales prediction based on the combined model of LightGBM and LSTM is 0.061. The proposed LightGBM-LSTM combined prediction model has higher performance in vegetable sales prediction than a single LightGBM and LSTM model. Forecast accuracy rate.

2 Related theories and methods

2.1LightGBMmodel

The LightGBM model [10] is an ensemble learning model based on Gradient Facilitated Decision Tree (GBDT) [11]. The key point of the model is to accumulate all tree models as the output result. The LightGBM model is optimized in the original GBDT algorithm, which solves the problem of the difficulty of training large amounts of data in GBDT. The traditional GBDT-based algorithm such as Xgboost [12] uses a pre-sorting method for prediction. First, all values on the data set are pre-sorted according to the features, and the best segmentation point on the feature is found by traversing the entire data set. This greatly increases the time complexity and memory usage. As shown in Figure 1, because the LightGBM model uses the histogram algorithm, the continuous eigenvalues are discretized into N integers and a histogram with a unit of N is constructed. When traversing the data, the discrete eigenvalues are used as an index in the histogram. Calculate the statistics in the middle, and finally find the optimal split point according to the discrete value of the histogram. The LightGBM model also uses the idea of mutually exclusive feature bundling (EFB), which combines some mutually exclusive features to reduce the dimension of features. Finally, the LightGBM model adopts the leaf-wise leaf growth strategy with depth limitation. Unlike the GBDT algorithm, it does not need to search and split each layer of leaves. In this way, the operating efficiency of the computer is greatly improved.

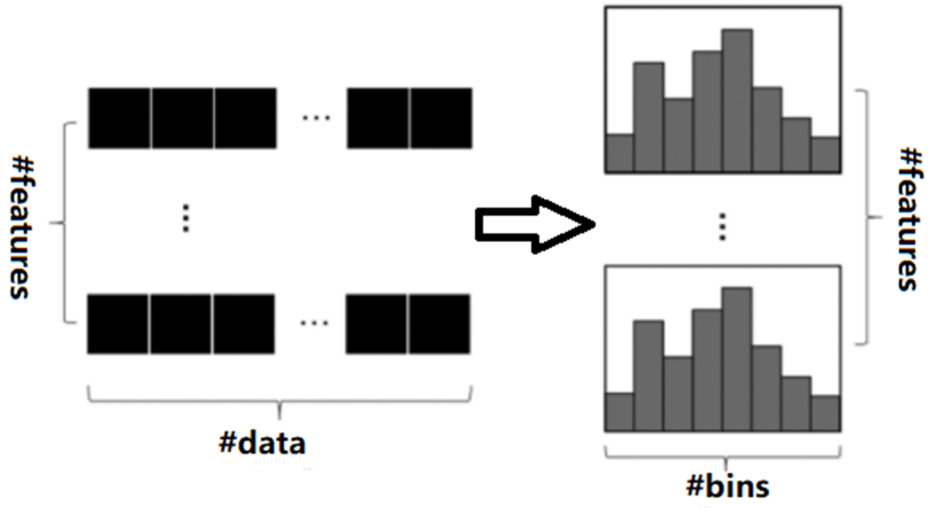

Figure 1 LightGBM decision tree algorithm

2.2 LSTM model

In order to solve the problems of gradient disappearance and gradient explosion in the conventional recurrent neural network RNN, the existing RNN is improved, and the long short-term memory network (LSTM) is proposed [13]. As a special neural network, LSTM can train a model of data features through continuous a large number of iterations to complete classification or prediction tasks.

The traditional neural network is usually composed of an input layer, a hidden layer and an output layer. However, LSTM designs a special structural unit on the original RNN. There are mainly three stages of forgetting, selection memory, and output. The stage can selectively add or remove the information of each time node. The structure of LSTM memory cells is shown in Figure 2:

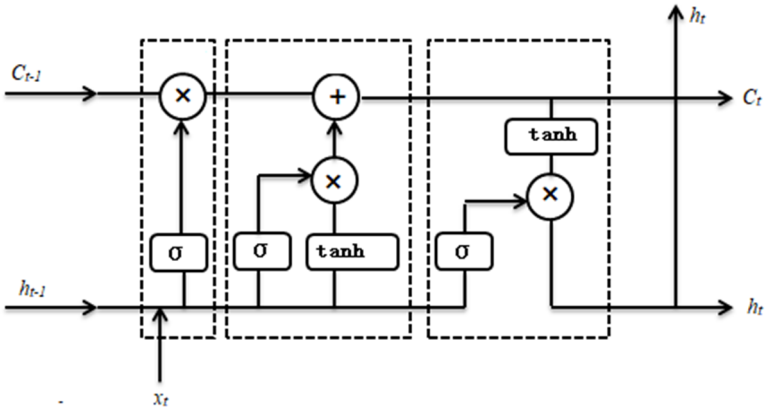

Figure2 LSTM memory cell structure

The forgetting stage is also called the forgetting gate. It determines how much information about C_{t-1} at the last time point needs to be retained to the current time point. It is usually implemented with the Sigmoid function. The value range of Sigmoid is from 0 to 1. When the Sigmoid value increases When the time, the amount of information allowed to pass is also increasing. The calculation expression of the forget gate is as follows:

$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f) \tag{1}$$

When the cell runs through the whole process in a way of fixed and changing information, the calculation formula of the current cell state C_t is:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{2}$$

Selection memory is also called input gate, its function is mainly to store some important information x_t, and decide whether to transmit it backward. The tanh function is usually used to determine which information is stored in the unit state. The calculation expression is as follows:

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i) \tag{3}$$

$$\tilde{C}_t = \tanh(w_c[h_{t-1}, x_t] + b_c) \tag{4}$$

The output stage is also called the output gate. Its function mainly determines which current information h_t will be output. The calculation expression is as follows:

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o) \tag{5}$$

The final output is:

$$h_t = \sigma * \tanh(c_t) \tag{6}$$

In the above formula, W_f, W_i, W_O, W_c *is* Weight matrix for each state, f_t, i_t, o_t, is activation vector value of each state at time t, b_f, b_i, b_o is the bias term in each state. σ is activation function for sigmoid; tanh is the hyperbolic tangent function.

3.Combination Model of Vegetable Sales Data Forecast

Based on LightGBM and LSTM combined model architecture can be divided into four steps:

Step 1: In order to better train the model, first perform data preprocessing on the original sales data of different dishes, including outlier detection, outlier detection, discrete feature processing, etc. Finally, through max-min normalization processing, the predicted sequence $S_n=\{S_1, S2,\ldots S_i\}$ is obtained. The data is linearly mapped to the range of 0-1, the specific formula is as follows:

$$S = \frac{S_i - S_{\min}}{S_{\max} - S_{\min}} \tag{7}$$

Among them: Si is the data that needs to be standardized, S_{min} is the minimum value in the sales data, and Smax is the maximum value in the sales data.

Step 2: Put the processed prediction sequence Sn into the LightGBM model. The learning rate of LightGBM model is 0.1, the maximum depth is 8, and num_leaves is 50.

Step 3: Constructing an LSTM model for modeling in non-feature engineering and mining data that cannot be directly observed or statistically abstracted. Two hidden layers are built in the LSTM model. The number of neurons in the first layer is 128, and the number of neurons in the second layer is 256. A Dropout layer is added after the BN layer to enhance the generalization of the model and prevent excessive Fit, and set the Dropout ratio to 0.3, the epoch to 500, and the BatchSize to 16, and the Adam optimization algorithm for network training.

Step 4: After obtaining the prediction results of the two models, use the reciprocal error method to determine the weight of the two models [14], the formula is as follows:

$$d_t = \omega_1 d^1{}_t + \omega_2 d^2{}_t, t = 1,2,3,\ldots n \tag{8}$$

$$\omega_1 = \frac{\delta_2}{\delta_1 + \delta_2} \tag{9}$$

$$\omega_2 = \frac{\delta_1}{\delta_1 + \delta_2} \tag{10}$$

Where $d^1{}_t$, $d^2{}_t$ are the predicted values of LightGBM and LSTM respectively, ωi is the weight coefficient, and δ_1, δ_2 are the errors between models.

4. Experimental results and analysis

4.1 Experimental data source and experimental platform

The experimental platform used in the experiment is as follows: the local host uses Intel core i5-6300U, 8GB memory and Windows 10 operating system; the integrated development environment is PyCharm 2016. The running platform of the experiment uses the keras framework based on deep learning.

The experimental data selected in this paper comes from the daily sales data of vegetables in Kunming Wenyuan Fresh Supermarket. The completeness of the comprehensive data and the local preference for vegetables. Six kinds of vegetables are selected from the sales categories for the selection of experimental samples. They are potato, tomato, cucumber, green onion, cabbage, carrot. Weather data is obtained by Python crawler to get the daily historical weather in Chenggong District of Kunming City during this time period. The time span is from December 20, 2017 to December 25, 2019. Use December 20, 2017 to August 20, 2019 as the training data sample, August 21, 2019 to December 15, 2019 as the test sample, December 15, 2019 to December 25, 2019 For the verification sample. The supermarket is surrounded by universities, middle schools, government agencies, and various stores. The sales data of each vegetable is shown in Figure 3. In the graph, the vertical coordinate represents sales volume, unit for kg, horizontal coordinate represents time, unit for days.

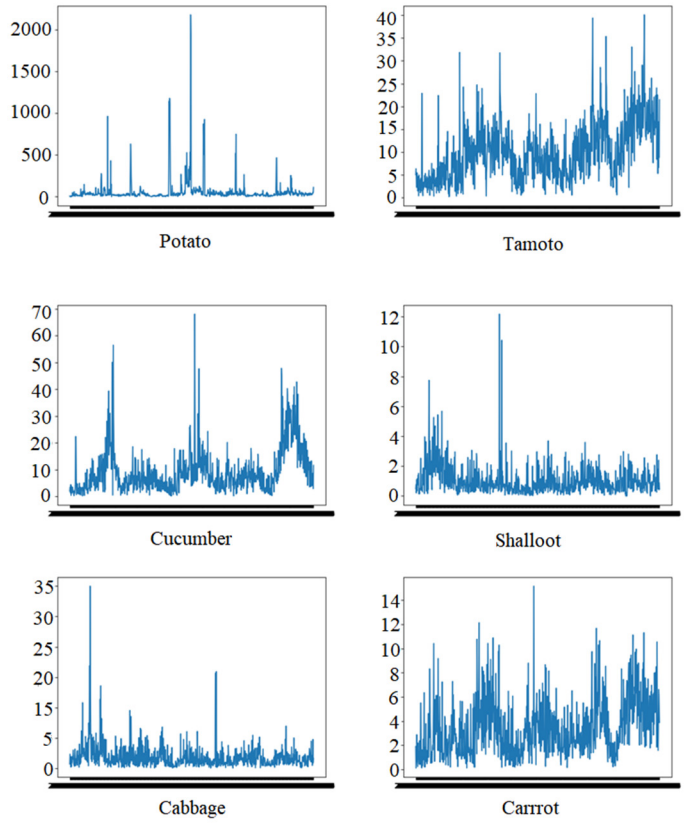

Figure3 Historical sales data of each vegetable

4.2 Experimental evaluation index

In order to effectively verify the ability of the LightGBM-LSTM combined model in vegetable sales forecasting, the average absolute error (MAPE) is selected as the absolute error between the true sales value and the predicted sales value between the models. Its expression is as follows:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - y^1{}_i}{y_i}\right| \tag{11}$$

Among them, y_i is the true value of vegetable sales, and $y^1{}_i$ is the predicted value of vegetable sales. The smaller the MAPE value, the more accurate the model's prediction of vegetable sales.

4.3 Analysis of prediction results

In order to verify the vegetable sales prediction performance of the LightGBM-LSTM combined model, this article first uses the LightGBM and LSTM single models to predict the sales, and finally uses the LightGBM-LSTM combined model to predict the sales. The comparison results of the actual sales and predicted sales of various models are shown in the figure. As shown in 4, the results of MAPE values of various models are shown in Table 1.

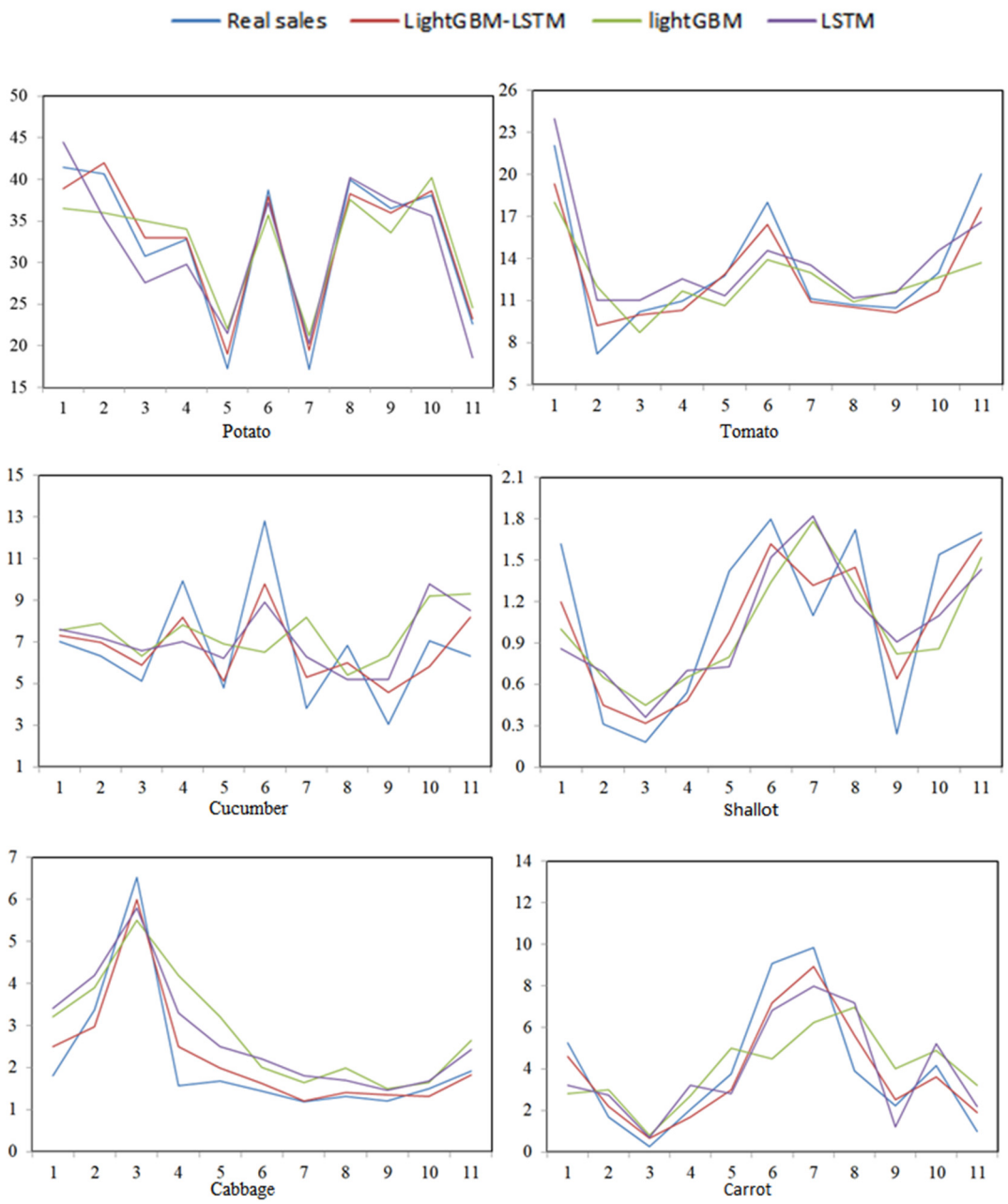

Figure4 Comparison of predicted and actual values of different models

Table 1 Comparison of MAPE values of different models

vegetables	LightGBM	LSTM	LightGBM-LSTM
potato	0.099	0.216	0.058
tomato	0.159	0.224	0.038
cucumber	0.273	0.323	0.105
shallot	0.416	0.618	0.279
Cabbage	0.432	0.518	0.169
carrot	0.526	0.622	0.321

It is not difficult to see from Figure 4 that the predicted value of the combined model sales is closest to the trend of the actual sales curve. From Table 1, it can be seen that the combined model

LightGBM-LSTM proposed in this paper has the smallest MAPE, with an average of 0.161. From a single model, LightGBM model is slightly better than LSTM model. In terms of vegetable types, tomatoes have the smallest MAPE and carrots have the highest. In terms of comprehensive evaluation indicators, the predicted value of the combined model is closer to the real sales volume, which can meet the daily management needs of the enterprise.

The figure shows the SHAP value of the important features of the LightGBM model. The red dots represent the points with higher sales volume, and the blue dots represent the points with lower sales volume. The more the red and blue are distinguished, the more important the features are. It can be seen from the table that the weather has a greater impact on sales, and the other factors are holidays, product names and other factors.

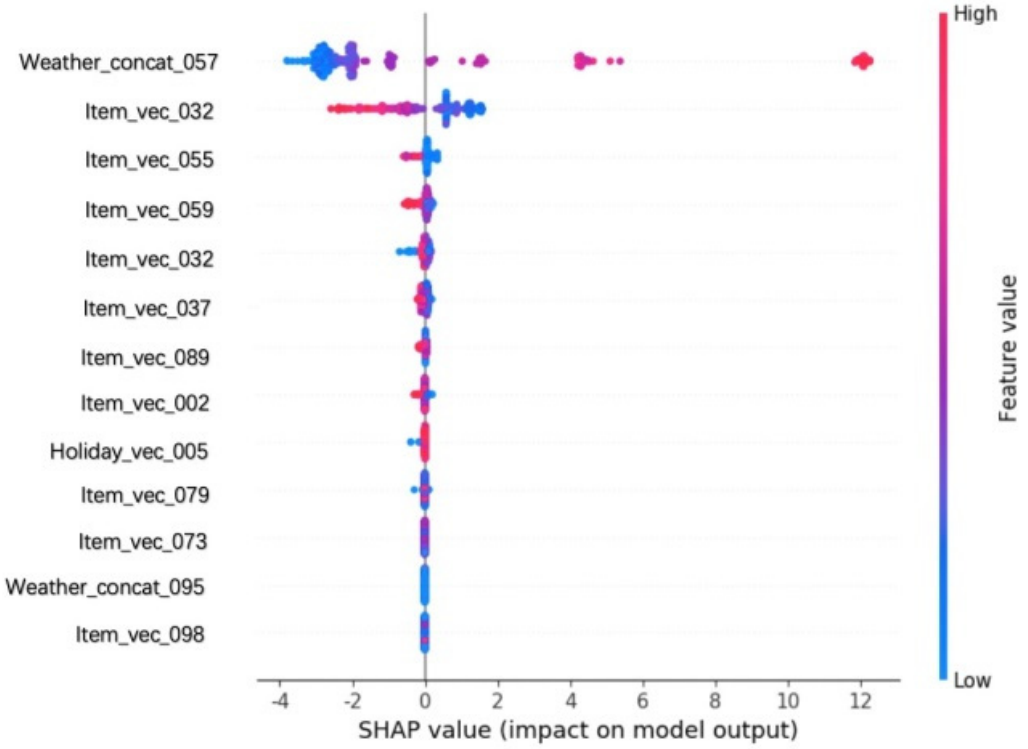

Figure5 SHAP value of important features of LightGBM model

5. concluding remarks

The sales volume of vegetables is affected by many factors. A single model has its own advantages and disadvantages, and it is difficult to achieve predictive results. This paper proposes a prediction method based on the LightGBM-LSTM combined model, and uses LightGBM to mine sales. Use the LSTM model to mine the nonlinear relationship in sales, and then use the reciprocal error method to weight the two prediction results and merge them to finally obtain the predicted value of vegetable sales, which facilitates the management of the enterprise and saves the cost of the enterprise. In the follow-up work, the model is applied to other types of vegetable predictions to improve the stability of the model. In terms of the model, consider doing more optimizations in the feature engineering of the LightGBM model, and introduce attention to the LSTM model Force mechanism to find a better modeling method.

References

[1] Luo Yanhui, Lu Yonggui, Li Bin. ARMA-based hybrid cigarette sales forecasting model[J]. Computer Application Research, 2009, 26(07): 2664-2668.

[2] Arunraj N S, Ahrens D, Fernandes M. Application of SARIMAX model to forecast daily ales in food retail industry[J]. International Journal of Operations Research amp; Informat Systems,

2016, 7(2):1-21.
[3] Hong Peng, Yu Shiming. Vending machine sales forecast based on time series analysis[J]. Computer Science, 2015, 42(S1): 122-124.
[4] Kaneko Y, Yada K. A deep learning approach for the prediction of retail store sales [C] IEEE. International Conference on Data Mining Workshops. IEEE, 2017:531-537.
[5] Sheng Wenshun, Zhao Hanchi, Sun Yanwen. Sales forecasting model based on improved genetic algorithm to optimize BP neural network [J]. Computer System Applications, 2019, 28(12): 200-204.
[6] Liu Youhong. Research on the Application of Neural Networks in the Sales Forecast System of Chain Enterprises [D]. Chongqing University of Technology, 2017.
[7] Ramos P, Santos N, Rui R. Performance of state space and ARIMA models for consumer retail sales forecasting[J]. Robotics and Computer-Integrated Manufacturing,2015,34:151-163.
[8] Yan Bo, Li Guohe, Li Xu. ARMA-based sales forecasting method and system implementation [J]. Computer and Modernization, 2014(05): 131-135. [9] Bi Liyuan, Wu Sai, Chen Gang, Shou Li Dan, Chen Ke,Hu Tianlei.Database query cost prediction based on recurrent neural network[J].Journal of Software,2018,29(03):799-810.
[10] KE G, MENG Q, FINLEY T, et al.LightGBM: a highly efficient gradient boosting decision tree [C]// Proceedings of the 2017 Annual Conference on Neural Information Processing Systems.New York: Curran Associates Inc.,2017 :3146-3154.
[11] FRIEDMAN J H. Greedy function approximation: a gradient boosting machine [J]. The Annals of Statistics,2001,29(5):1189-1232.
[12] CHEN T, GUESTRIN C.XGBoost: a scalable tree boosting system [C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016:785-794.
[13] GREFF K, SRIVASTAVA R K, KOUTNÍK J, et al. LSTM: A Search Space Odyssey[J].IEEE Transactions on Neural Networks & Learning Systems, 2015, 28 (10) :2222-2232.
[14] Chen Zhenyu, Liu Jinbo, Li Chen, Ji Xiaohui, Li Dapeng, Huang Yunhao, Di Fangchun, Xinxingyu, Xu Lizhong. Ultra-short-term power load forecasting based on the combined model of LSTM and XGBoost[J]. Power System Technology, 2020, 44(02)):614-620.