Trend Prediction and Empirical Analysis of Pork Futures Prices Based on Machine Learning

Yiyang Sun *, Yixiao Gong, Yu Pan, Changjun Yu
School of Mathematics and Statistics, Nantong University, Nantong, China, 226019
* Corresponding Author Email: syy77178818@163.com

Abstract. The live hog futures market has faced abnormal price cycles and severe fluctuations in recent years, making traditional forecasting methods less effective. Accurate price prediction is crucial for market participants to make informed decisions. This study employs the XGBoost model to predict live hog futures price trends. It uses price indices from 1 to 28 days prior as independent variables and closing prices as dependent variables to analyze feature importance. The significant variables identified are then used to develop a prediction model. The study finds that prices from 1 to 3 days prior, 5 days prior, and 22 days prior are significant for prediction. The developed model achieves a Mean Absolute Percentage Error (MAPE) of 0.11% on the training set and 1.71% on the test set. It forecasts a volatile trend for the next 20 days. The study innovatively integrates the findings with traditional trading indicators (e.g., "5-day moving average" and "20-day moving average") and demonstrates their relevance in the live hog futures market. This approach provides a new perspective for predicting price trends in complex and dynamic markets.

Keywords: XGBoost Model, Pig Cycle, Live Hog Futures Market, Price Trend, Futures Price Forecasting.

1. Introduction

In recent years, the fluctuations in pork market prices have shown new characteristics, characterized by a shortened upswing cycle and an extended downturn cycle. This deviation from the traditional pattern has posed a severe challenge to the stability of the pork market. As an important component of agricultural products, live hog futures play a crucial role in stabilizing market prices and mitigating risks[1]. However, the high degree of uncertainty in their prices has brought significant difficulties in forecasting future prices.

The academic community has made certain progress in the research on pork price fluctuations and predictions. Zhu Zengyong has refined the analysis of the "pig cycle," dividing it into rational and irrational phases of production capacity reduction, providing a new framework for cyclical analysis[2]. Hou Senyang and Shi Jing have applied the H-P filter model to systematically study the mechanisms of pork price changes[3]. In terms of prediction methods, Zhang Yingying constructed an ARIMA model to predict pork prices, verifying the applicability of traditional time-series models in this field[4][5]. With the development of computing technology, machine learning, and deep learning methods have gradually been applied in this area. Jiang Baichen developed an improved SVM model[6], Cao Huiru constructed a Bi-RNN-LSTM model[7], and Hu Chun'an and Jiang Wei proposed the VMD-BO-BiLSTM model. These studies have shown that advanced algorithms have great potential in price prediction[8]. However, research on machine learning for live hog futures still needs to be further explored.

This study focuses on the prediction of live hog futures prices, aiming to explore efficient data analysis and model construction methods to improve prediction accuracy. The main objectives include identifying the significant factors affecting live hog futures prices in the current complex market environment[9] and developing a high-precision prediction framework based on feature engineering and model optimization.

From a theoretical perspective, this study will enrich the empirical research system for predicting live hog futures prices, providing new perspectives and methodological support for the development of related theories. From a practical standpoint, by constructing a high-precision prediction model and offering reasonable suggestions, this study aims to provide decision-making support for participants in the live hog futures market, helping them more effectively cope with price fluctuations and market risks.

This study employs an empirical analysis approach, systematically collecting macroeconomic data and live hog market data. It uses time-series feature analysis and feature importance assessment techniques to screen for key influencing factors. Innovatively, the study introduces the concept of "daily moving averages" for model optimization. The transformed data are then input into the XGBoost model for training and prediction, with the goal of achieving more accurate forecasts of live hog futures prices.

2. Data Processing and Model Principles

The data in this article is sourced from the market data provided by https://www.tdx.com.cn/. When establishing the predictive model, the closing price was chosen as a key reference indicator to more accurately depict the future trend of the main pork futures prices. The variables selected include the pig-feed ratio and the per-head output value of pigs raised by small-scale and large-scale farming operations, which reflect the economic benefits of the breeding industry.

2.1. Data Standardization

For live hog futures data, proper normalization can help the model more effectively analyze the intrinsic characteristics of the data. Normalization generally involves mapping the data to a fixed range, usually the [0, 1] interval, to eliminate the influence of data dimensions and units. This allows data with different units or magnitudes to be effectively compared[10]. It not only converts the data into pure scalars but also simplifies the computational process, thereby enhancing the efficiency and accuracy of the model in processing data. Min-max normalization is a commonly used method, with the transformation function as follows,

$$\boldsymbol{X'} = \frac{\boldsymbol{X} - \boldsymbol{X_{min}}}{\boldsymbol{X_{max}} - \boldsymbol{X_{min}}} \tag{1}$$

Where X_{max} is the maximum value of the sample data, and X_{min} is the minimum value of the sample data.

2.2. The Principle of the XGBoost Model

XGBoost is a machine learning algorithm based on Gradient Boosting Decision Trees (GBDT). Its core idea is to generate a new decision tree model on the basis of all previous trees in each iteration so as to minimize the residuals or errors from the previous iteration, thereby enhancing the overall performance of the model. The final prediction of the model is the sum of the predictions from all decision trees, that is

$$\boldsymbol{\hat{y}_i} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{f_k(x_i), f_k \in F} \tag{2}$$

Where K is the number of decision trees, and F is the set of all decision trees.

3. Empirical Study of Pork Futures Price Prediction and Market Trend Analysis

3.1. Empirical Study of Pork Futures Price Prediction and Market Trend Analysis

The feature engineering of this model mainly consists of two parts: lag processing and moving average. Firstly, the closing price of the label data is lagged to create a list representing the previous day's settlement price, which facilitates the model in capturing the autoregressive nature of the sequence, that is, the influence of the previous day's value on the subsequent day. Secondly, to smooth the time series, moving average feature engineering is employed. Based on the characteristics of futures prices, the window size is set to 7. In this paper, 28 moving average features with different window sizes are created to provide the model with trend information across different ranges, taking into account the temporal dimension comprehensively.

The model involves principles related to decision trees, which have been mentioned in the preceding text. The model is initialized by specifying the learning rate, the maximum depth of the trees, and the subsample size. Building on the feature engineering and time-series cross-validation discussed earlier, decision trees are constructed step by step, and the model is improved through gradient boosting.

3.2. Analysis of experimental results

The Mean Absolute Percentage Error (MAPE) is selected as the error evaluation metric. The error evaluation of the final model training is 0.0668. The results are visualized in Table 1,

Table 1. Deviation of the Original Prices

Evaluation Metrics(%)	Training Set	Test Set	Full Dataset
MAE	20.15	258.058	67.98

This paper finds that the model performs well on the training set but only moderately on the prediction set. This suggests that the model suffers from a certain degree of overfitting. Moreover, the results of MAPE are almost consistent with those of MAE, and the prediction deviations are shown in Table 2.

Table 2. Deviation of Predicted Percentages

Evaluation Metrics(%)	Training Set	Test Set	Full Dataset
MAPE	0.11%	1.71%	0.43%

The prediction visualization is shown in Figure 1,

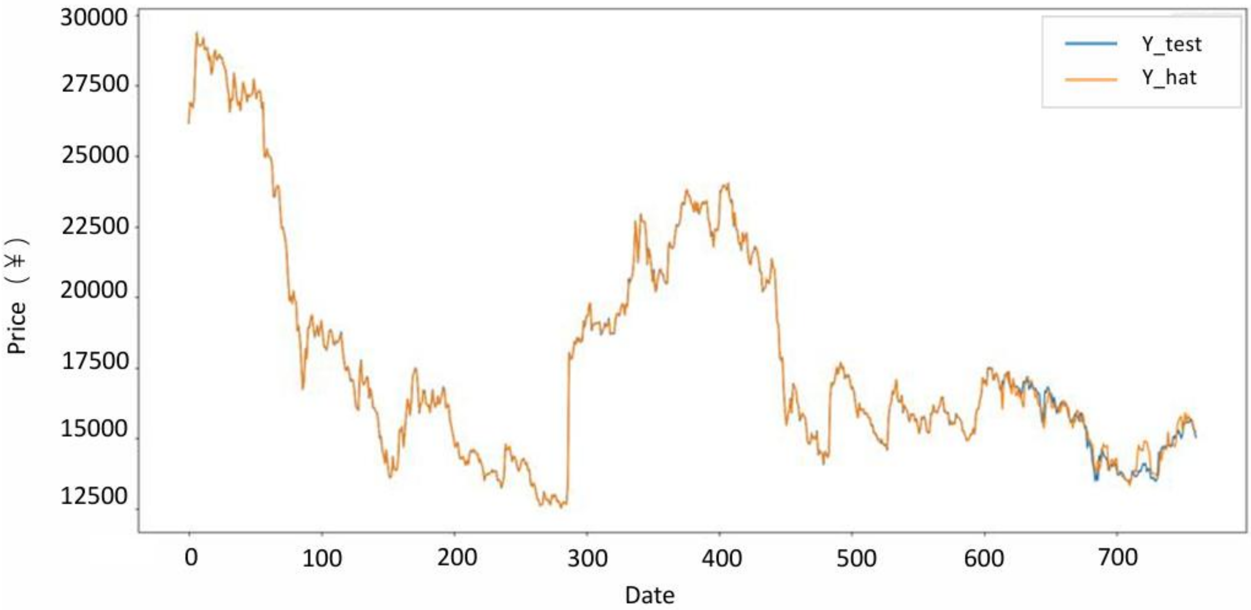

Figure 1. Prediction Effect of the XGBoost Model

Through feature selection and feature engineering, the XGBoost model is able to successfully capture market dynamics and achieve outstanding performance during the training phase, specifically demonstrated by a mean absolute percentage error of 0.0668. The metrics clearly prove a significant statistical relationship between the selected features and the target variable. Moreover, the visualization of the prediction results provides an intuitive illustration of the model's strong inferential capability during the validation period, further confirming its predictive accuracy and practicality.

3.3. Results Analysis

3.3.1. Analysis of Feature Importance

To better understand which specific factors influence the live hog futures market and the relative importance of these factors, a feature importance analysis was conducted on the closing prices of live hog futures from 1 to 28 days prior, based on feature selection and feature engineering. The relative importance ranking of different input factors was further obtained through visualization using the F-score, as shown in Figure 2.

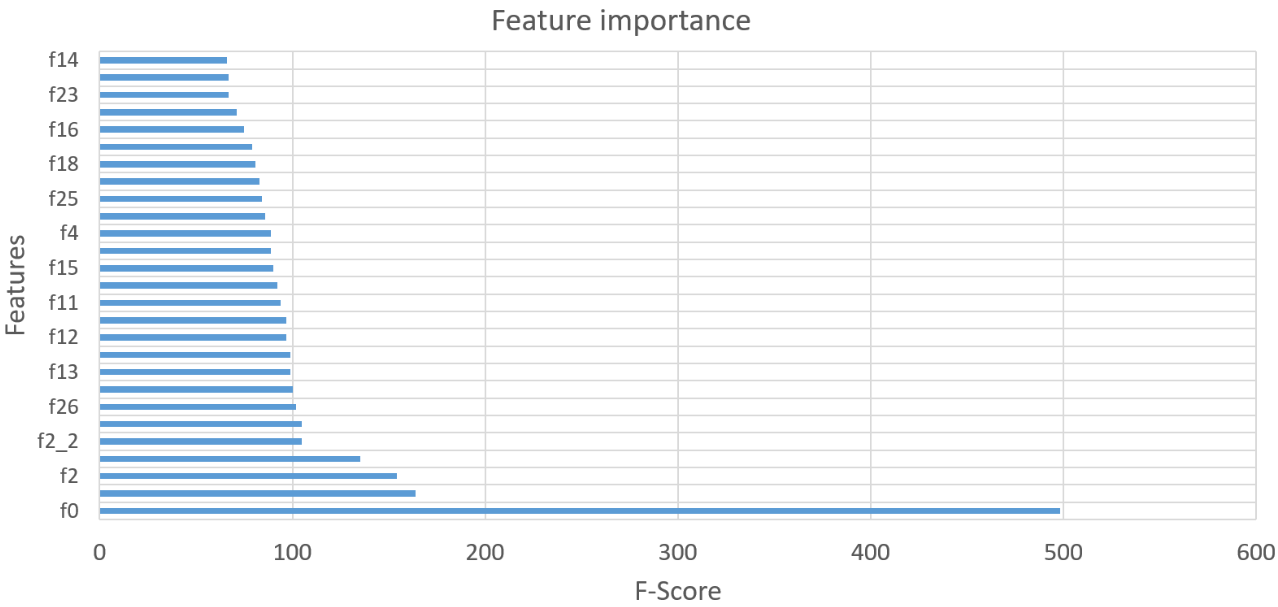

Figure 2. Feature Importance Ranking of Closing Prices from 1 to 28 Days Ago

In the time series analysis used in this paper, "lag" refers to the time interval between the target variable or feature and its previous value. For example, if the value from yesterday is used to predict today's outcome, then yesterday's data is considered "lag 1" relative to today. For instance, the notation f0 represents lag_1, and so on. The length of the bar indicates the importance of the feature for predicting the target. It is evident that the prices from 1, 2, 3, 5, and 22 days ago are important independent variables.

Some traders in the futures market prefer to use the Turtle Trading Rule—where the futures price is considered bullish when the 5-day moving average crosses above the 20-day moving average, and bearish when the 5-day moving average crosses below the 20-day moving average—the features selected in this study are consistent with real-world practices.

To better analyze the results, this paper introduces the concept of "daily moving average": the average of daily closing prices over a certain period, representing the average trend of the futures. Daily moving averages are typically used to judge the long-term trend of stock prices and can also assist in decision-making. In addition, based on this model and combined with time series analysis, the overall trend is predicted. The following figure shows the 7-day moving average plotted according to the futures market data selected in this paper, as shown in Figure 3.

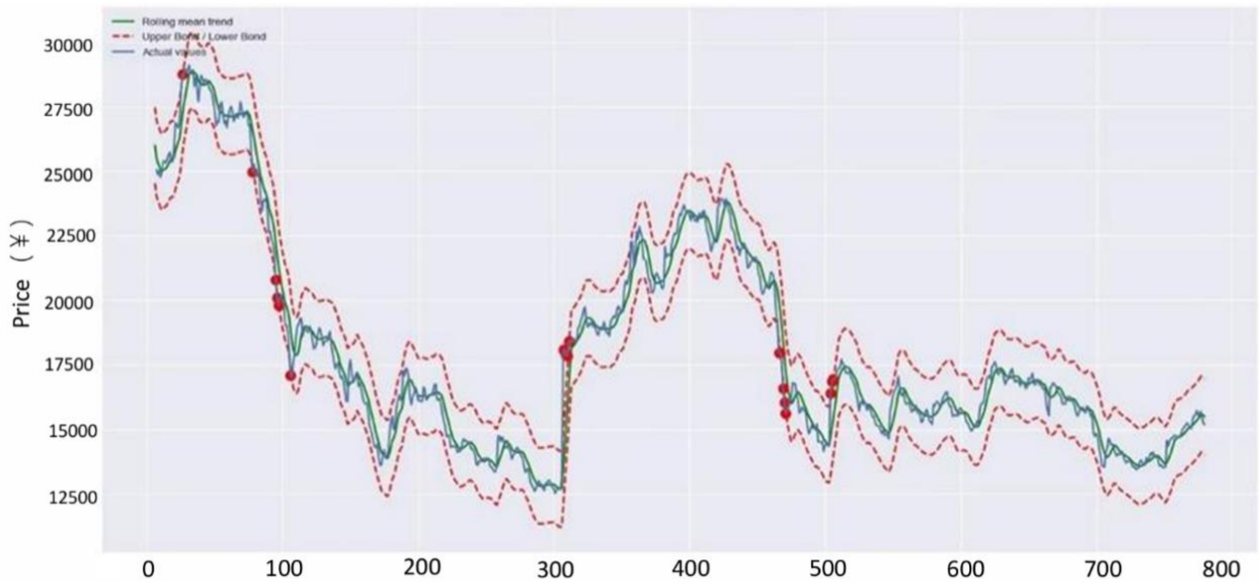

Figure 3. 7-Day Moving Average Chart

In the figure, the green line represents the moving average of the data, and the red dashed lines indicate the upper and lower bounds, which facilitate the identification of outliers. Based on the feature selection results from XGBoost (Figure 15) and in combination with the Turtle Trading Rule, we use the closing prices from 1 to 3 days ago, the average price from 5 days ago, and the average price from 20 days ago as features to predict the closing price of the current day. Compared with the original model, this model has fewer features, is more interpretable, and has a smaller generalization error. The feature importance ranking chart is shown in Figure 4.

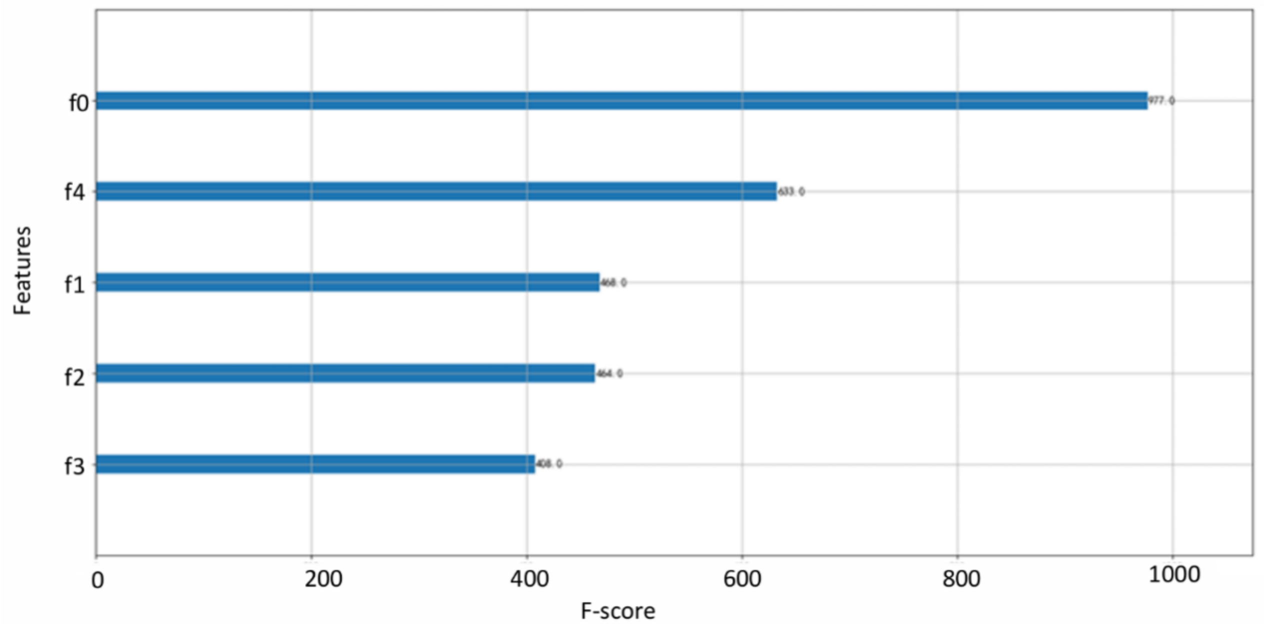

Figure 4. Feature Importance Ranking Chart

Among them, f0 represents the closing price of yesterday, and f4 represents the 20-day moving average. This paper finds that the 20-day moving average is more important than the 5-day moving average. This has a certain guiding significance for futures trading.

3.3.2. Future Prices Prediction

This paper uses a model based on the characteristics of moving averages to predict prices for the next 20 days. The predicted prices for the next 5 days are shown in Table 3.

Table 3. Forecasted Prices for the Next 5 Days (Partial)

Days	Forecast Value
Day 1	17059.48
Day 2	17326.19
Day 3	17249.37
Day 4	17063.27
Day 5	16776.77

This paper predicts that the market trend over the next 20 days will show an upward movement with fluctuations. The forecast results are visualized in Figure 5.

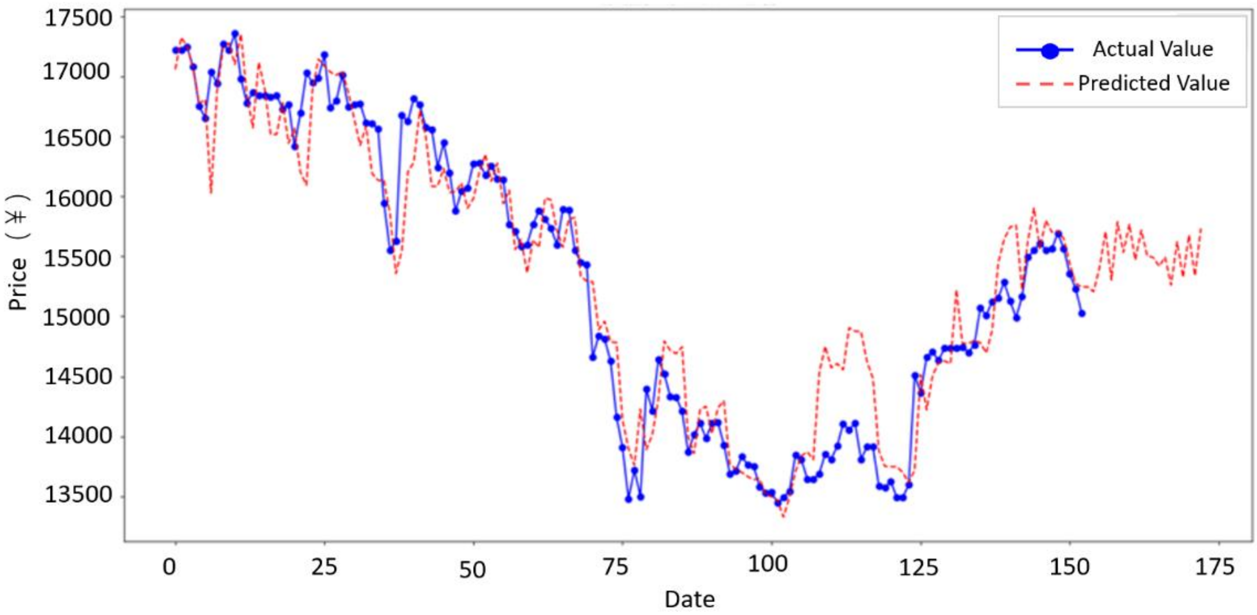

Figure 5. Chart of Market Trend for the Next 20 Days

The forecast results indicate that the current live hog futures prices are at the bottom. Based on historical market data, the downside potential for live hog futures prices is limited, and there is an upward trend. Previously, the rise in grain prices and the increased costs for hog farmers resonated, causing live hog prices to experience a period of volatility, reaching an unprecedented level of extreme market fluctuations. Given the current relatively stable live hog market, it is predicted that although there will be an upward trend in live hog futures prices for a period after March 29, 2024, it is unlikely to see a repeat of the extreme conditions in 2021 and only a modest recovery is expected.

4. Conclusions

This study uses the XGBoost model to analyze and predict live hog futures price trends, aiming to explore price fluctuation characteristics and forecasting methods in the current complex market environment. Results show that the regularity of the live hog price cycle has been disrupted, with a shortened upward cycle and an extended downward cycle. Feature importance analysis reveals that prices from 1–3 days, 5 days, and 22 days prior are highly important for prediction, aligning with the use of the "5-day moving average" and "20-day moving average" in Turtle Trading. Optimizing the model with these features, the study achieves a MAPE of 0.11% on the training set and 1.71% on the testing set, indicating good predictive effectiveness and generalization ability. The next 20 days are forecasted to see an upward trend with fluctuations, with limited downside potential and a modest recovery expected.

However, the study has limitations. It mainly relies on historical price data and simple macroeconomic indicators, failing to fully account for policy changes, public health events, and international trade impacts. The limited time span and sample size may also affect the model's generalization ability. Future research can expand the data range, incorporate more factors, and combine other advanced methods to enhance predictive accuracy and adaptability. This study provides a reliable decision-support tool for live hog futures market participants, with significant implications for improving market forecasting and risk management.

References

[1] Ling Liwen, Huang Rongxi, Huang Chuxin. Construction and Analysis of Knowledge Graph on Pork Price Research Based on CiteSpace [J]. Chinese Agricultural Science Bulletin, 2023, 39 (25): 155-164.

[2] Zhu Zengyong, Zhou Liang. Causes and Operating Mechanism of Cyclical Fluctuations in Pork Prices [J]. Price Theory and Practice, 2023, (10): 158-164.

[3] Hou Senyang, Shi Jing. Research on the Characteristics of Pork Price Fluctuations and Regulatory Countermeasures in China [J]. Journal of Anhui Agricultural Sciences, 2024, 52 (4): 189-192.

[4] Liu Yanhong, Liu Hebing, Shang Junping. Research on the Prediction of Agricultural Product Futures Prices Based on Deep Learning [J]. Henan Science, 2024, 42 (3): 430-439.

[5] Zhang Yingying. Prediction of China's Pork Prices Based on the ARIMA Model [J]. Business Exhibition Economy, 2020 (5): 21-23.

[6] Jiang Baichen, Feng Kaijie, Peng Sixi. Research on Pork Price Prediction Based on Improved Support Vector Machine [J]. Guangdong Agricultural Sciences, 2018, 45 (12): 158-164.

[7] Cao Huiru, Xu Peixin, Chen Junjiong, Zhong Lin, Li Xiaomin. Prediction and Experiment of Domestic Live Hog Prices Based on the Bi-RNN-LSTM Model [J]. Mechanical and Electrical Engineering Technology, 2023, 52 (10): 260-263+289.

[8] Hu Chun'an, Jiang Wei. Pork Price Prediction Model Based on VMD-BO-BiLSTM [J]. Journal of Applied Science, 2023, 41 (04): 692-704.

[9] Yang Hanyu, Zhao Xiaoyong, Wang Lei. A Review of Data Normalization Methods [J]. Computer Engineering and Applications, 2023, 59 (3): 13-22.

[10] Miao Zhichun, Liu Xiuqin. Analysis of Factors Affecting Pork Price Fluctuations in Fujian Province [J]. Hebei Enterprise, 2021, (12): 9-12.