A Hybrid Autoregressive LSTM Model-based Gasoline Price Predicting Method Using Optimal Time Window

Siyi Wei, Xin Zhang*, Chaoran Zhou, Wenwei Jiang

School of Computer Science and Technology, Changchun University of Science and Technology, Changchun, China

Abstract

Gasoline is the core energy of petroleum products. The accurate prediction of gasoline price can provide decision-making basis for urban economic construction, energy security and people's travel. At present, the oil price forecasting methods focus on single factor model, and lack of objective factors such as supply and demand. In the aspect of feature building, this paper presents a feature building method based on data window, and obtains the comprehensive feature data by introducing multiple factors and building feature. Aiming at the nonstationarity, nonlinearity and time variability of gasoline price data, this paper designs a gasoline price prediction model based on hybrid autoregressive long-term and short-term memory network (HARLSTM). In order to evaluate the research work of this paper, neural network and regression model are selected as baselines in the experiment. Experiments show that HARLSTM performs better in MSE, MAE and MAPE. In the aspect of optimal time window selection, when the window length reaches 12, this model performs best, MSE, MAE, MAPE index reduces at least 20%, 14%, 3%.

Keywords

Hybrid Autoregressive LSTM Model, Vector Autoregression, Gasoline Price Predicting, Deep Learning

1. INTRODUCTION

The price of gasoline is closely related to people's daily life. As a major part of people's daily consumption, gasoline price affects consumers' choice of car purchase and travel mode, and its change is the basis of making effective economic and environmental strategies. Such as [1] that gasoline prices on the enterprise's resource allocation, logistics and transportation and other aspects of the impact is very large, [2] that it significantly affects the frequency and time of bicycle travel. In addition, the retail price of gasoline is mainly affected by the price of crude oil and the level of supply and demand for gasoline, but also by refineries, gasoline taxes, environmental regulations and crude oil prices [3-4]. These factors lead to the highly nonlinear nature of gasoline prices, which makes it challenging to capture the volatility mechanism of oil prices in order to predict oil prices.

Aiming at the non-stationary, non-linear and time-varying characteristics of gasoline price, this paper proposes an auto-regressive LSTM model for gasoline price prediction. The main contributions of this approach are as follows:

In this paper, a feature analysis model is proposed to extract the key features that affect oil prices, and to determine the impact cycle of various factors on gasoline prices, so as to provide a basis for parameter setting of the prediction model.

A feature construction method based on data window is proposed and extended to multiple time steps. Explore the correlation between oil prices and other factors in time series, and timely capture the trend of oil prices in time.

ICCEIC2022@3rd International Conference on Computer Engineering and Intelligent Control
EMAIL: *Corresponding author: zhangxin@cust.edu.cn (Xin Zhang)

A prediction model is proposed. Based on the (recurrent neural network) RNN and LSTM, the prediction is decomposed into a separate time step, and the prediction of each step is taken as the feedback of the next step. Based on this model, the input time step of the model is determined by combining the calculated window length with the analysis model. The time step of model access under this scheme is the influence period of relevant factors on oil price, so the model can fully grasp the variation of characteristics with time. Experimental results show that the proposed model is better than the baseline model.

2. RELATED WORK

In the field of energy prices, oil price time series analysis and prediction is one of the most important research subjects. Capturing the underlying volatility mechanism of energy prices is challenging because of their significant nonlinear time-variability. Support Vector Machine Regression (SVR), Stochastic Forest Regression and Neural Network in Machine Learning can deal well with the series with nonlinear and volatility. For example, when predicting the future price level of gasoline products, Xu F et al. compared the performance of autoregressive integrated moving average (ARIMA) - generalized autoregressive conditional heteroscedasticity (GARCH), exponential smoothing, grey systems, neural networks and support vector machines models. Experimental results showed that support vector return (SVR) and feedforward neural network (FNN) were better than other models in predicting accuracy [5]. Baumeister et al. conducted an in-depth regression analysis of gasoline prices in the United States. On this basis, they suggested using a mixture of five forecasting models with equal weights [6]. Through the analysis of the above work, it is found that the oil price prediction methods based on machine learning have better accuracy of oil price time series prediction.

In view of the above problems, more and more deep learning methods are proposed and applied to solve the oil price time series analysis and forecasting problems. Nonlinear models based on neural networks exhibit higher accuracy in processing time series data [7]. The above research results have the advantage of learning the timing characteristics of oil price changes, but do not consider the objective factors that affect oil price changes. Yang said multivariate models can improve predictions [8].

To sum up, by analyzing the influence characteristics of various factors on gasoline prices, the author provides support for the parameter setting of subsequent models. Therefore, based on LSTM, this paper constructs a time series prediction model of autoregressive oil price, and establishes a nonlinear prediction model among multiple variables using multiple (multi-factor) data sets.

3. THE STUDY.

In this section, we describe the research work on HARLSTM as a prediction model. First, we analyzed the relevant factors influencing the price fluctuation of gasoline and the data processing process. Then we build features with analytical models. Finally, we describe how to build the model and its internal structure.

3.1. Data Preprocessing

Because of the complexity and diversity of data, the time of data acquisition is not the same. In order to keep the loss of data features, this paper fills in the feature factors. In addition, data loss is inevitable, and according to our observations, the percentage of missing values is small. The data missing in this paper can be divided into two cases: data feature missing and time stamp missing. For data feature loss, we use linear interpolation to fill them based on their adjacent data points. Another problem is the missing timestamp problem, which is supplemented by formula (1). T_{insert}represents the current insertion position, T_{back} the position after the vacancy, and $Timestep$ the time granularity. Secondly, there is still inconsistency of timestamp. For example, the date of collecting gasoline price is January 24, 2000, but some characteristic factors are collected later than the time of collecting gasoline price, so there is no corresponding gasoline price on January 24. The price of gasoline on Jan. 28 is perfectly fine, according to observations.

$$T_{MISS} = T_{insert} + Timestep * (T_{back} - T_{insert}) \quad (1)$$

According to the analysis, oil price data is cyclical, especially when seasonal changes, oil price data will also change. In order to obtain a year's worth of available signals, this article creates them using the following formula:

$$\mathrm{Ys} = \sin\left(\mathrm{timestamp_s} * \left(2 * \frac{\pi}{\mathrm{year}}\right)\right) \quad (2)$$

$$\mathrm{Yc} = \cos\left(\mathrm{timestamp_s} * \left(2 * \frac{\pi}{\mathrm{year}}\right)\right) \quad (3)$$

3.2. Feature Construction

Through the above analysis, this section calculates the impulse response function of each variable through vector autoregressive model, and judges the influence on other variables when a variable changes one standard deviation. It can help to select the factors that affect gasoline prices, and analyze each factor's specific impact on gasoline prices, such as the size and duration of the impact of characteristic factors on gasoline prices, to obtain key factors.

Taking Florida as an example, some factors are analyzed as follows:

As you can see from **Figure 1a**, Gasoline demand also has a positive impact on gasoline prices, which is in line with demand growth and price increase. **Figure 1b** Gasoline inventory has a negative impact on the price of gasoline, which lasts for four months and then closes to zero. This also indicates that inventory serves as a buffer between short-term supply and demand, suppressing gasoline prices when supply problems arise. The other factors were analyzed according to this method. For example, gasoline output has a short negative impact on gasoline price, and then shows a continuous positive impact until approaching zero. This means that as gasoline production increases and supply increases, prices fall. The subsequent positive shocks show that stocks are low, it is not enough to cushion supply and demand, and gasoline production needs to increase. Gasoline imports on gasoline prices showed a positive impact for three months gradually towards zero. Inventories are falling rapidly, leading to higher oil prices as imports rise, indicating a production shortfall.

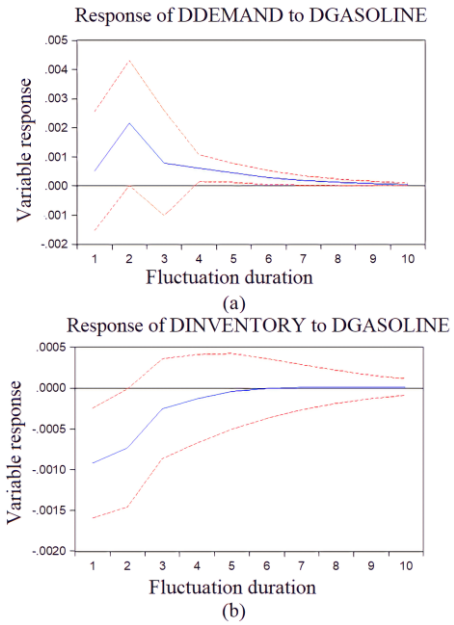

Figure 1 Impulse Response Results of a Standard Deviation to Gasoline Price for Gasoline Demand, Gasoline Stock Changes

3.3. Prediction model modeling

The first method in the model constructed in this paper is to use a preparation structure based on a combination of RNN and Dense layers, using oil price series {*X_1, ..., X_N*} tensors containing multiple features involved in feature construction as input. Extracting time series features from each input tensor by the Lstm_rnn layer shown in **Figure 2**. Specifically, each of the tensors in this paper is a 3-D tensor of 13 channels, via Lstm_rnn gets 2D tensor $I_t^w = \text{rnn} \ldots \text{dense}(X_t)$. This structure is a state that initializes LSTM based on the input and, once trained, captures the relevant portion of the historical data, which is equivalent to a one-step prediction model. The first dimension of this 2D tensor is batchsize, and the second dimension is the eigenvalue.

The tensor I_t^w obtained by preparing the structure passes through the pseudo-cyclic neural network structure of the model. This structure is a rnn composed of LSTM cells. Its purpose is to use the state of the preparation structure and the initial prediction as input, continue iterating the model, use the prediction of each step as input feedback, and set the number of iterations to the output step.

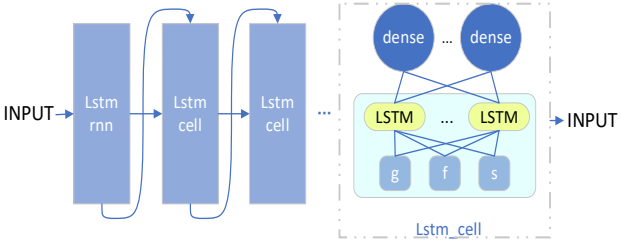

Figure 2 Structure of HARLSTM

3.4. Internal architecture of prediction model

The above research results have the performance advantage of learning the time series characteristics of oil price changes, but the impact of objective factors of oil price changes on oil price time series prediction is not considered. **Figure 3** shows the Architecture of lstm_cell. The output is *(t+12) th* week gasoline price, and the input is *(t-timestep+1) th* to *t th* week data. This time step (TS) represents the time step of this structure, which is calculated by the method discussed above. At *TSth* time step (*week (t-timestep+1)*), the data of gasoline price on the day, related factors and statistical characteristics are connected to multiple lstm_cell Modules, each module outputs a cell state based on input. This unit state captures the contribution of factor data to the characteristics of oil price fluctuations. Each cell state carries some useful information and predictions of each step to its subsequent module, on which each lstm_cell module's cell status will be updated at week *(t-ts + 2)*. Loop through this process until you are connected to a 12-week goal (label).

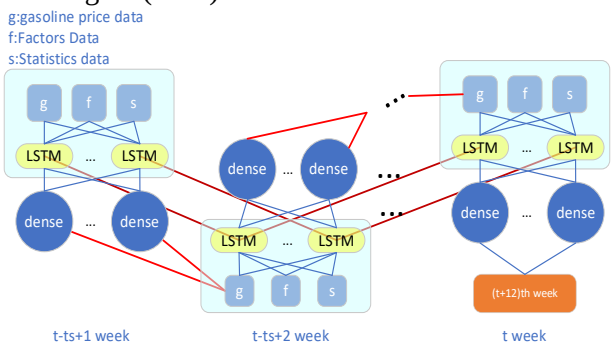

Figure 3 The Lstm_cell Structure

4. Experimental result

The petrol price data set is from the official website of the U.S. Energy Information Administration (wwws://eia.gov/) and contains retail petrol price data for cities in nine states, including Florida, from

January 3, 2000 to June 28, 2021. Each sample of the data set contains timestamp and oil price information data. The Crude Oil Price Data Set is from Cushing, Oklahoma, Crude Oil Price (wwws://eia.gov/), Fuel Oil Price is from New York Port No. 2 Fuel Oil (wwws://eia.gov/), and National Car Volume is from the Trade Economics Network. Ref = ieconomics.com/& iis). The remaining variables (crude oil stocks, crude oil acquisition costs of refineries, gasoline imports, gasoline stocks, gasoline production, gasoline demand) are data for the Petroleum Authority area of the Defence Zone.

Figure 4 shows the result of a multi-time step forecast, which, unlike a single-step forecast, is a sequence of predicted future values. The HARLSTM model predicts gasoline prices for the next 12 weeks. The blue line indicates the 12-week price of gasoline for the model input and the light blue cross indicates the forecast for the next 12 weeks. It is found that the trend of the predicted curve is basically consistent with the actual value, which proves that the model has good fitting performance.

In order to quantitatively measure the prediction results of the model, Table 1 lists all the model results under the three evaluation indicators. The MSE, MAE and MAPE of LGBM model and SVR model are less than those of MLP and CONV _ DENSE model, so the relationship between gasoline price and the factors affecting gasoline price is nonlinear. Because LSTM is a large depth neural network constructed as a complex nonlinear element, it has the ability to deal with high dimensional large data and adapt to time variability. Therefore, it can better capture the volatility and trends in gasoline prices, to tap its potential characteristics, showing the best precision and better indicators. The HARLSTM produced a mean square error of 0.008 (MSE), an average absolute error of 0.06 (mae) and an average absolute percentage of 0.19 (mape); it performed well in multi-step forecasts, with MSE reduced from 0.13 to 0.09. Through comparative analysis, it is proved that the model can predict gasoline price effectively.

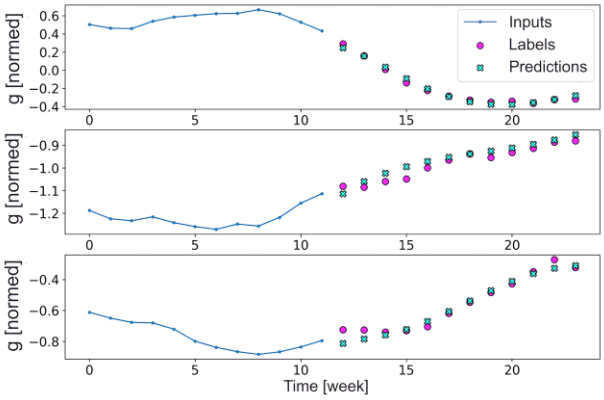

Figure 4 Line Plot of Actual Values and Prediction Values

TABLE 1 COMPARSION AMONG MODELS

Model	MSE	MAE	MAPE
MLP Single step	0.1	0.25	0.3
MLP multi-step	0.1	0.23	0.22
Conv_dense Single step	0.08	0.16	0.34
Conv_dense multi-step	0.13	0.23	0.45
residual_lstm Single step	0.01	0.09	0.34
residual_lstm multi-step	0.12	0.3	0.39
LGBM Single step	0.01	0.08	0.24
RFR Single step	0.09	0.06	0.23
SVRSinglestep	0.01	0.07	0.26
HARlstm Single step	0.008	0.06	0.19
HARlstm multi-step	0.09	0.23	0.26

To verify the validity of the window length calculation, this article sets different window lengths for comparison, as shown in Table 2. The experimental results show that HARLSTM performs best when

the window length is 12, and the influence period of each factor on gasoline price fluctuation is 12 weeks. Length is 8 when next. But when the length is 24, the error is the biggest, so the impact on gasoline prices is more on the data within the cycle than the impact of long-term historical data. In this paper, the window length, the three indicators are reduced by 20%, 14%, 3%, significantly improve the performance of the model, confirming the advantages of this model and analysis model.

TABLE 2 TABLECOMPARSION AMONG DATAWINDOW

Window Length	MSE	MAE	MAPE
24	0.23	0.44	0.27
12	0.08	0.22	0.26
8	0.1	0.25	0.35

5. Conclusions

This paper describes a model called HARLSTM to predict urban oil prices. HARLSTM determines the optimal time window based on vector autoregressive model. In the aspect of feature extraction, multi-step time windows are used to improve the performance of HARLSTM mining and exploring time-dependent high-dimensional and statistical features. Experimental results show that the HARLSTM is superior to other baseline models. In the oil price market, the price of gasoline varies from place to place. Therefore, the authors will construct the geographical feature information and introduce HARLSTM to achieve more accurate and objective oil price prediction.

6. Acknowledgment

This work is supported by the Jilin Scientific and Technological Development Program (No. 20200201182JC).

7. References

[1] Busse M R, Knittel C R, Silva-Risso J, et al. Who is exposed to gas prices? How gasoline prices affect automobile manufacturers and dealerships[J]. Quantitative Marketing and Economics, 2016, 14(1): 41-95. [J]. Quantitative Marketing and Economics, 2016, 14(1): 41-95.

[2] He P, Zou Z, Zhang Y, et al. Boosting the eco-friendly sharing economy: the effect of gasoline prices on bikeshare ridership in three US metropolises[J]. Environmental Research Letters, 2020, 15(11): 114021.

[3] Kilian L. Explaining fluctuations in gasoline prices: a joint model of the global crude oil market and the US retail gasoline market[J]. The Energy Journal, 2010, 31(2).

[4] Borenstein S, Kellogg R. The incidence of an oil glut: who benefits from cheap crude oil in the Midwest?[J]. The Energy Journal, 2014, 35(1).

[5] Xu F, Sepehri M, Hua J, et al. Time-series forecasting models for gasoline prices in China[J]. International Journal of Economics and Finance, 2018, 10(12): 43-53.

[6] Baumeister C, Kilian L, Lee T K. Inside the crystal ball: new approaches to predicting the gasoline price at the pump[J]. Journal of Applied Econometrics, 2017, 32(2): 275-295.

[7] Firouzjaee J T, Khaliliyan P. LSTM Architecture for Oil Stocks Prices Prediction[J]. arXiv preprint arXiv:2201.00350, 2022.

[8] Yang Y, Guo J, Sun S, et al. Forecasting crude oil price with a new hybrid approach and multi-source data[J]. Engineering Applications of Artificial Intelligence, 2021, 101: 104217.