LSTM-Based Gasoline Price Prediction - An Example of Gasoline 95# in Beijing

Lingyue Zeng *

Beijing-Dublin International College, Beijing University of Technology, Beijing, 100124, China

* Corresponding Author Email: lingyue.zeng@ucdconnect.ie

Abstract. Gasoline, as one of the indispensable energy sources in the global economy, the stability of its price has an important impact on national energy security and economic development. This study aims to improve the accuracy of gasoline price prediction through the use of Long Short-Term Memory Network (LSTM). By collecting and analyzing the historical price data of No. 95 gasoline in Beijing, the author constructed a prediction model based on LSTM, which is particularly suitable for dealing with long-term dependence in time series data. By comparing their performance to the traditional Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN), LSTM demonstrates its advantages in processing complicated time series data. The validity of the model is verified by several evaluation metrics, including mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). The results of the study indicate that the LSTM model shows high accuracy and reliability in predicting gasoline prices. In addition, future work will explore the introduction of multivariate analysis and hybrid network models to further the accuracy of the predictions and its generalization ability.

Keywords: LSTM; Gasoline Price; Prediction Accuracy; Time Series Analysis.

1. Introduction

In the international world, gasoline is a very important resource, and the fluctuation of its price will have a great impact on the global economy and regional economy [1]. Especially for a large developing country like China, the stability of gasoline price is of great significance to guarantee national energy security and promote sustainable and healthy economic development [2]. As the capital of China, the fluctuation of gasoline price in Beijing not only affects the city's daily operation, but also reflects the change of supply and demand in the energy market and the adjustment of macroeconomic policies. Therefore, an accurate prediction of gasoline prices in Beijing is an important reference value for policy makers, enterprises and ordinary consumers.

In this study, the author used the Long Short-Term Memory Network (LSTM) model, which is an efficient method for analyzing time series data and is particularly suitable for dealing with and predicting long-term dependencies in serial data. By collecting and analyzing the historical price data of No. 95 gasoline in Beijing, the author constructed a prediction model using the LSTM model. The model not only helps people understand the price trends in historical data, but also predicts the movement of gasoline prices in the future period.

The data processing process of the article includes data normalization, division of training and test sets, construction and training of LSTM models, and evaluation of model performance. The goal of this study is to verify the LSTM's effectiveness in time series analysis through the accuracy of the model prediction, which provides a scientific basis for subsequent economic decisions and market analysis. By integrating data science and machine learning techniques, this study aims to provide a new perspective on gasoline price forecasting and explores the use of machine learning techniques in the energy sector.

2. Theoretical Foundations and Methodology

2.1. Characteristics of gasoline prices

This study concentrates on analyzing the fluctuation of the price of gasoline 95 in Beijing. By analyzing 5,425 pieces of data from January 15, 2009, to November 22, 2023, it can be concluded that the price of No. 95 gasoline has experienced different degrees of fluctuation. The price data shows that the lowest price of gasoline 95 was 5.68RMB, the highest price reached 9.93 RMB, and the average price was 7.44 RMB. This indicates that during the observation period, the price of No. 95 gasoline showed obvious volatility, reflecting the influence of market supply and demand, changes in crude oil prices and national policies. Therefore, in-depth analysis and forecasting of gasoline price changes are critically important for the development of related economic policies and individual consumption plans. The fluctuation in the price of gasoline in shown in figure 1.

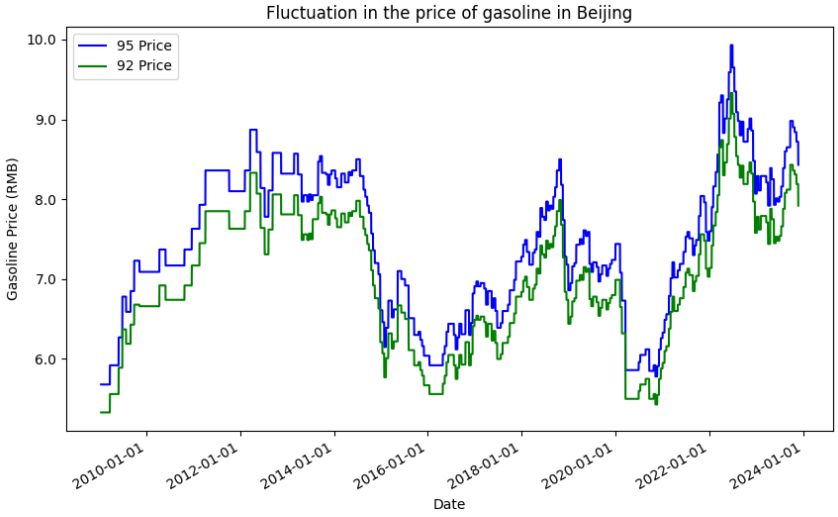

Fig. 1 Fluctuation in the price of gasoline in Beijing (Photo/Picture credit: Original)

2.2. LSTM algorithm and its application to Gasoline price prediction

LSTM is a specialized type of Recurrent Neural Network (RNN) designed specifically to tackle the problems of vanishing gradients that traditional RNNs face when processing long data sequences [3]. The LSTM addresses this issue by incorporating three gates: an input gate, a forgetting gate, and an output gate. The input gate manages the intake of new information, the forgetting gate determines which past information to eliminate, and the output gate oversees the release of information. The architecture of the LSTM model is shown in figure 2.

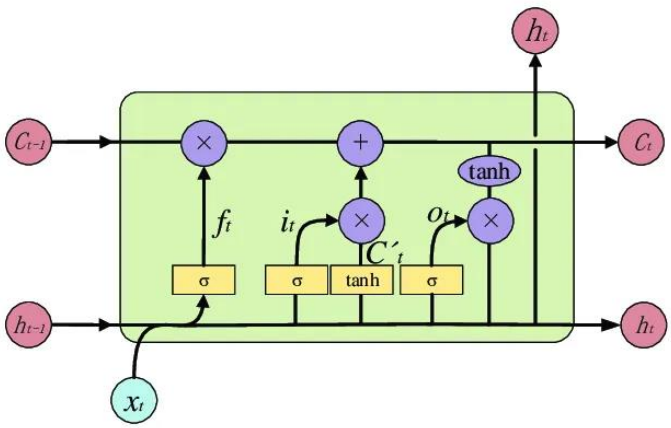

Fig. 2 The LSTM model's architecture [3]

The core of LSTM lies in its ability to learn important features of data at different points in time in a sequence and predict future values based on these features [4]. This structural design makes LSTM particularly suitable for processing and forecasting time series data, especially in scenarios with long term dependencies [5]. Gasoline price time series data is highly nonlinear and has complex time dependencies, and LSTM's powerful time series data processing capabilities can provide more accurate prediction of gasoline prices.

3. Application of LSTM Networks in Predicting Gasoline Prices

3.1. Construction of Gasoline Price Prediction Model Based on LSTM

3.1.1 Data pre-processing and preparation

Before conducting a study on gasoline price prediction based on LSTM networks, proper pre-processing of the raw dataset is a necessary antecedent step. The process mainly involves key aspects such as data cleaning, feature selection and data normalization, aiming to provide clean and standardized data for the model.

First, the data cleaning phase mainly focuses on resolving missing values, outliers, and non-numeric data in the dataset. For the 95 gasoline price data in this study, ensuring that each record has accurate date and corresponding price information is the first task of the cleaning process. At the same time, missing values in the data must be removed or filled in to maintain data continuity and improve model training.

Secondly, the feature selection phase aims to identify the data dimensions that will be used in model training. Considering the input requirements of the model, the price data in the time series is selected as the main feature, a decision based on the objective of studying the price movements of gasoline and the advantages of LSTM in handling time series data.

Finally, data normalization is performed by applying the Minimum-Maximum Normalization (MinMaxScaler) method to scale all feature values to a range between 0 and 1. This measure is crucial for enhancing the model's convergence rate and predictive accuracy, as it aids in mitigating the issue of vanishing that the model faces throughout the training phase [6].

These refined data preprocessing steps provide a solid foundation for constructing and training the LSTM-based gasoline price prediction model, which ensures t data quality and input consistency, and thus supports the realization of highly accurate price prediction.

3.1.2 The construction of training and test sets

Constructing training and test sets is a key step in the development process of gasoline price prediction model, which directly affects the training effect of the model. In this study, the author used the sliding window method to generate training and test samples from time series data, which is particularly suitable for dealing with data with time-dependent characteristics, such as gasoline price movement data. The sliding window method creates data samples by sliding a fixed-length window over the original time series. For each window, the method uses the historical price within the window as a feature (X) and the next price after the window as a label (Y). In this way, the model learns the ability to predict future prices given past price information. In the code implementation, through the create_dataset() function, the author set the time step for each window, and here the time step chosen is 100 days. This means that the model will use the first 100 days of price data to predict the price on the 101st day. Choosing the right time step is the key to achieving efficient prediction. A time step that is too short may not be sufficient to capture enough historical information, while a step that is too long may make it difficult for the model to learn effective prediction patterns from too much historical information [7].

To assess the model's performance with unknown data, the authors partitioned the dataset into training and the test sets. Specifically, 80% of the data is used for training, while the remaining 20% is designated for testing. This division ensures that the model has enough data for learning, while also leaving a portion of the data for testing the model’s generalization ability. Importantly, to preserve

the sequential integrity of the time series, the author divides the dataset in a chronological rather than a randomized way.

3.1.3 Architectural design of LSTM model

In this study, the author designed a LSTM based architecture aimed at efficiently predicting the movement of gasoline prices. This model architecture encompasses the configuration of the input and hidden layers, the design of the output layer, and also includes the choice of activation function, loss function, and optimizer. These elements are integrated to ensure that the model effectively captures the complex dependencies present in the time-series data and provides precise predictive outcomes. The description of the model's training process is shown in Figure 3.

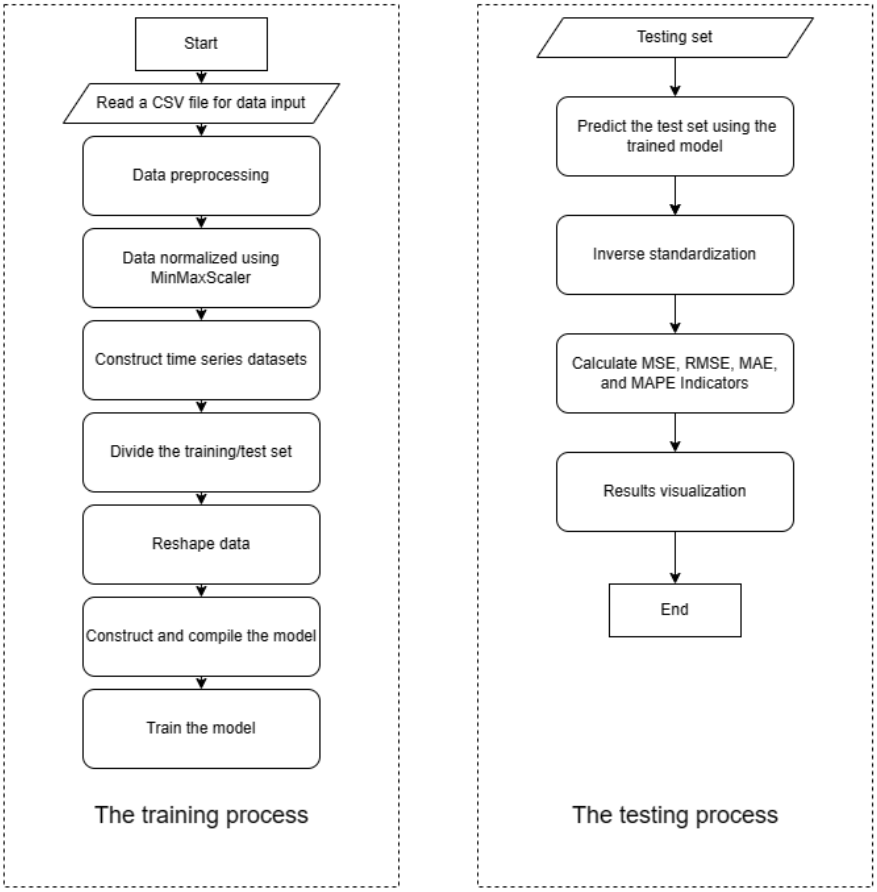

Fig. 3 The model's training process (Photo/Picture credit: Original)

The model's input layer receives data preprocessed and reshaped by the previous steps in the shape of [samples, time steps, features], corresponding to the quantity of samples in the training set, the time step for each sample (100 days in this study), and the number of features in each time step (1 in this case, i.e., the price of 95 gasoline). This structure takes full advantage of the LSTM's ability to process sequential data, providing the model with the necessary time-series information.

The hidden layer of the model consists of two LSTM layers; the first LSTM contains 128 neurons and returns the sequence in order to provide the complete time series input to the next LSTM layer; the second LSTM contains 64 neurons and does not return the sequence, but only outputs the result at the end of the sequence. This layered LSTM structure allows the model to learn long-term dependencies in time series data at a deeper level. Next, the model further processes the information through two densely connected layers, where the first dense layer consists of 32 neurons and the second dense layer includes16 neurons and uses the ReLU activation function to introduce nonlinear properties and improve the expressive power of the model.

Finally, the output layer of the model consists of a single neuron dense layer that directly outputs the predicted values of future prices. In the compilation stage of the model, the author chooses the

mean squared error (MSE), root mean square error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) to quantify the difference between the model's predicted value and the actual value. Meanwhile, the Adam optimizer is used to optimize the learning process.

With this architectural design, the model is able to effectively learn features extracted from historical price data and use these learned features to predict future Gasoline prices, demonstrating the power of LSTM in handling complex time series prediction problems.

3.1.4 Model training

During the stage of model training in this study, the author used carefully selected batch sizes and epochs to optimize the performance of the LSTM network model. Specifically, 32 was used as the batch size for the training process of the model, which means that 32 samples were fed into the network at a time for training. This batch size takes into account the efficient utilization of computational resources and also ensures the stability and efficiency of the training process. At the same time, the model is set to be trained iteratively for 10 epochs, in which the model will completely traverse the training set 10 times for a sufficient number of iterations to ensure that complex features and dependencies in the data are learned. This phase of the work provides the reliability and accuracy of the model and provides a solid foundation for achieving high quality gasoline price predictions.

3.1.5 Performance Evaluation

After completing the training of the LSTM network-based model, the evaluation of its prediction performance is a crucial step. In this study, several key statistical metrics are used to comprehensively assess the performance of the model, including MSE, RMSE, MAE and MAPE.

Specifically, MSE quantifies the average of the squares of the discrepancies between the predicted and actual values, and is a commonly used metric for assessing model performance. RMSE, on the other hand, as the square root of MSE, calculates the prediction error using the same units as the original data, providing a more intuitive assessment of the magnitude of the error. MAE computes the average of the absolute discrepancies between the predicted and the actual values, providing an intuitive measure for evaluating the accuracy of predictions. Finally, MAPE, which represents the percentage of prediction error relative to the actual value, is a common measure of prediction accuracy, and is particularly useful for comparing data of different sizes.

To further visually assess the predictive effectiveness of the model, the authors visualize the predictive ability of the model by plotting actual prices against predicted price, which is shown in Figure 4. In this chart, the actual price is depicted by a blue line, and the predicted price is shown by a red line. By comparing the trend of the two, the accuracy of the model's prediction and trend-capturing ability can be intuitively assessed.

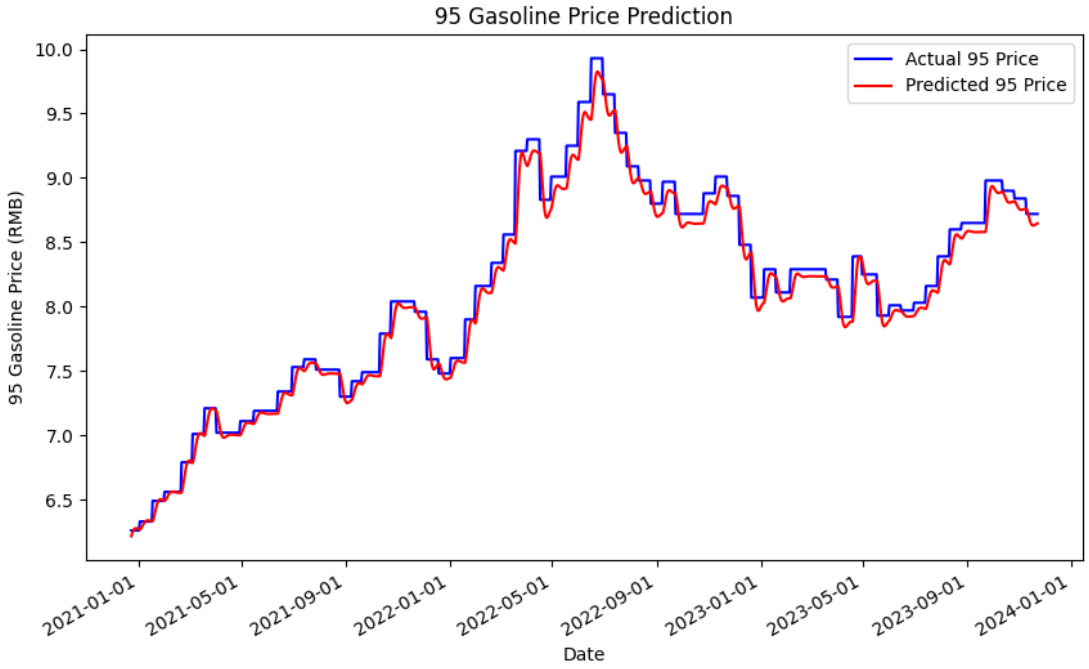

Fig. 4 Comparison Chart of Actual and Predicted Price Curves by the Model (Photo/Picture credit: Original)

In this study, through these comprehensive performance evaluation methods, the effectiveness of the LSTM model in the Gasoline price prediction task can be relatively objectively analyzed and judged.

3.2. Analysis and Comparison of LSTM Model Prediction Results

3.2.1 Model prediction results analysis

In this study, the LSTM based model shows good performance on the gasoline price prediction task. The results of the performance evaluations are shown in Table 1.

Table 1. The result of the performance evaluations of the model

MSE	RMSE	MAE	MAPE
0.0076	0.0872	0.0463	0.5610

In particular, the lower values of MAE and RMSE indicate a smaller average error between the model's predicted and actual values, suggesting that the model has a high prediction accuracy [8]. The value of MAPE, on the other hand, indicates that the model-predicted prices only have an average deviation of about 0.56% from the actual prices in the vast majority of cases, which further confirms the effectiveness of the model in capturing price trends and volatility.

By visually comparing the actual prices with the predicted prices, it can be noted that the model successfully captures the major trends and cyclical characteristics of the data, and despite some errors, overall, the model demonstrates a better ability to predict future price changes. The analysis of these prediction results not only demonstrates the potential of LSTM-based models for time series forecasting tasks, but also provides valuable information for further optimization of the model structure and parameters, with a view to achieving higher prediction accuracy in future work.

In summary, the analysis of the results in this study shows that the LSTM-based model has considerable application value in gasoline price prediction, which provides useful references and insights for future research and applications in related fields.

3.2.2 The comparison of LSTM model with CNN, RNN models

In the task of price prediction, two models, CNNs and RNNs, besides the LSTM model, are also often seen in price prediction. However, the LSTM model show superior prediction accuracy

compared to CNNs and traditional RNNs due to their ability to manage long-term dependencies in time series data [9]. Specifically, the CNN model, while excelling in image recognition and processing, slightly underperforms the LSTM model designed for sequential data when dealing with continuous time series prediction tasks. This is mainly due to the fact that CNNs lack the intrinsic structure to deal with the long-term dependence of time series.

Meanwhile, compared to the traditional RNN model, LSTM also demonstrates significant advantages in prediction accuracy. Although RNN is theoretically able to capture the dependencies in time series, it often faces the problem of gradient vanishing in practical applications, which limits its effectiveness on long sequence data [10]. On the contrary, LSTM effectively solves these problems by introducing a gating mechanism, which improves the model's capability to capture long-term dependencies within time-series data. This improvement leads to more accurate prediction outcomes in applications such as gasoline price forecasting.

By analyzing the application and results of these three models, it can be clearly seen that LSTM has advantages in dealing with complex time series prediction problems. This comparison not only proves the applicability and effectiveness of LSTM on the task of gasoline price prediction, but also provides a valuable reference for the selection of suitable models in the future, especially in the task of sequence prediction that needs to deal with long-term time dependencies.

4. Conclusion

In this study, gasoline price prediction was successfully realized by constructing and training a model based on LSTM. Through comparative analysis, the LSTM model demonstrated the ability to outperform traditional RNN and CNN models in processing time series data. The prediction results of the model are validated by MSE, RMSE, MAE, MAPE metrics and show high accuracy and reliability, proving the applicability and effectiveness of the LSTM model in such prediction tasks. However, despite the results achieved, there are still many challenges in the practical application of gasoline price prediction. First, the high dependence of the model on data quality may limit its performance in the presence of missing or abnormal data. Second, although LSTM can effectively deal with the time-dependent problem, the adaptability and flexibility of the model in the face of extreme market movements still needs to be improved. In terms of limitations, the main limitation of the current study is that it relies on a single variable (gasoline price) for forecasting, ignoring other multivariate factors that may affect price movements, such as policy changes, market demand fluctuations. In addition, both the training and testing of the model were conducted using datasets from the same geographic region, which may have geographical limitations.

For future research, the introduction of multivariate models can be considered to improve the accuracy of prediction. Additionally, expanding data sources to incorporate datasets from a wider range of regions and time horizons will also be a critical step in improving the generalization capabilities and utility of the models. Ultimately, by continuously optimizing the model structure and algorithms, it is expected that more intelligent and efficient forecasting tools can be developed to respond to the diverse needs of future markets.

References

[1] Wang M, Chen Y, Tian L, et al. Fluctuation behavior analysis of international crude oil and gasoline price based on complex network perspective. Applied Energy, 2016, 175: 109-127.

[2] Rahman Z U, Khattak S I, Ahmad M, et al. A disaggregated-level analysis of the relationship among energy production, energy consumption and economic growth: Evidence from China. Energy, 2020, 194: 116836.

[3] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation, 1997, 9(8): 1735-1780.

[4] Yu Y, Si X, Hu C, et al. A review of recurrent neural networks: LSTM cells and network architectures. Neural computation, 2019, 31(7): 1235-1270.

[5] Burgueño L, Cabot J, Gérard S. An LSTM-based neural network architecture for model transformations. 2019 ACM/IEEE 22nd International Conference on Model Driven Engineering Languages and Systems (MODELS). IEEE, 2019: 294-299.

[6] Dai Z, Heckel R. Channel normalization in convolutional neural network avoids vanishing gradients. arxiv preprint arxiv:1907.09539, 2019.

[7] Nar K, Sastry S. Step size matters in deep learning. Advances in Neural Information Processing Systems, 2018, 31.

[8] Wang W, Lu Y. Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model. IOP conference series: materials science and engineering. IOP Publishing, 2018, 324: 012049.

[9] Mehtab S, Sen J. Stock price prediction using CNN and LSTM-based deep learning models. 2020 International Conference on Decision Aid Sciences and Application (DASA). IEEE, 2020: 447-453.

[10] Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 2020, 404: 132306.