Short-term Petroleum Price Prediction Based on LSTM and LightGBM

Yuefan Zhang [1, *], Yixiao Shi [1], Yadi Qiao [1, #], Weiwei Wang [2, #]

[1] Department of Economics and Management, Tiangong University, Tianjin, China, 300387

[2] Department of Business, Chengdu University of Technology, Sichuan, China, 610059

* Corresponding author: zkkdxbg@163.com

[#]These authors contributed equally.

Abstract. With the recent severe fluctuations in the international petroleum market, accurately predicting the direction of oil prices has enormous importance for multiple industries. This study employ the single-unit LSTM model and the multi-unit LSTM model based on word frequency to predict short-term petroleum prices. Results show that the multi-unit LSTM model provides better prediction performance. This study also utilized the LightGBM prediction model taking seven distinct variables into considerate. The result indicates that the model proposed is effective in forecasting crude oil price trends. In conclusion, LSTM model has better interpretability in the time dimension, while the LightGBM model has higher overall prediction accuracy than LSTM. And this study provides ideas for the construction of other oil price forecasting models.

Keywords: Crude Oil Price, LSTM, word frequency, LightGBM, Short-term Forecasting.

1. Introduction

International crude oil prices are critical for market economy and serves as a key performance indicator for businesses and investment entities. Recently, various methods have explored for oil price prediction. Two primary categories of methods commonly used-traditional econometric models and machine learning-based approaches.

Initially, econometric models utilized in crude oil price searching including GARCH, VAR, VEC, and so on [1-3]. Ma et al. improved the accuracy of predicting oil market volatility by incorporating timely reflected price jump information into the MIDAS model [4]. In addition, in hybrid econometric models, Aamir et al. finding that the ARIMA-Kalman mixed model performed best among three mixed time series models, and Mirmirani et al. finding that the BPN-GA model, which combines BP neural network with genetic algorithm, outperformed the VAR model [5-6]. Nevertheless, econometric models have limitations in capturing the nonlinear characteristics of crude oil price trends.

In recent years, more machine learning techniques are utilized in forecasting crude oil prices. Guo utilized VMD and FE to decompose and reconstruct crude oil price sequences, and obtained multiple high-accuracy prediction models [7]. Hybrid models have also been proposed by other scholars. Zhou proposed a new prediction model combined CEEMDAN and XGBoost [8]. Lin combined Bi-LSTM, attention mechanism, and CNN, greatly improved accuracy [9]. In conclusion, machine learning methods can predict short-term crude oil prices more efficiently and accurately compared to traditional econometric models.

However, deep learning algorithms and ensemble algorithms have not been widely researched for short-term crude oil price forecasting [10]. This study uses a combination of LSTM and LightGBM models to predict crude oil prices and compare their accuracy.

2. Data source

2.1. Oil price

Based on trading volume, market ranking, circulation of oil, this study selected WTI crude oil for

our research to represent the oil market. The daily trading data from October 12, 2017, to October 31, 2022, were collected from https://www.iea.org/. Existing research generally considers that oil price volatility is correlated to some macroeconomic variables such as stock index, exchange rate, and oil price and so on. This paper selected monthly central bank discount rate, monetary aggregates M2 and consumer price index, oil supply, oil demand and oil market size as the characteristics that could affect the price fluctuation of oil price[11-12]. The data is from https://tradingeconomics.com/.

The crude oil spot prices data in US dollars per barrel used in this study are sourced from https://research.stlouisfed.org/. This study first uses IQR outlier identification to handle outliers and implements cubic spline interpolation method to deal with missing data. The descriptive statistics results are presented in Table 1.

Table 1. The descriptive statistics of oil prices

		Central		Variability	Skewness and Kurtosis		
target	sample size	average	variance	range	Skewness	Kurtosis	CV
value	1827	61.938	190.072	125.0	0.375	0.247	0.223

2.2. Word frequency statistics

This study conducted a word frequency analysis of oil-related news text data. First, use Python web scraping technology to crawl all real-time news articles on https://oilprice.com from October 12, 2017 to October 31, 2022. 1060 articles were finally obtained. After data cleaning, a total of 85,562 pieces of text data were obtained. Then this study used these text data to perform daily word frequency statistics.

According to the result of the word frequency statistics, this paper removed low-frequency words and invalid words. Then screening out text vocabulary that is highly associated with oil price changes based on the Random Forest model.

The core formula of the Random Forest regression algorithm is:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} f_b(x) \tag{1}$$

Where $\hat{y}$ represents the predicted value of the predicted oil value, B represents the number of all decision trees,$f_b(x)$ represents the predicted value of the bth decision tree to sample x.

According to the ranking results of feature importance (FI), the top 4 variables with high importance were selected as highly correlated vocabulary was shown in Table 2.

Table 2. Top 4 variables with high importance

Words	ukraine	coronavirus	cuts	variant
Feature importance	0.2605	0.1126	0.0547	0.0337

3. The establishment of the model

3.1. Long Short-Term Memory Model

LSTM is kind of recurrent neural network, which contains "long-term memory" and "short-term memory". The connections and weights undergo dynamic changes at each training step, forming long-term memory using short-term memory of thousands of steps. The problem of gradient disappearance generated by classical RNNs in long time sequences is basically solved.

LSTM regulate the flow of information into and out of the cell through an input gate, an output gate and a forget gate. LSTM network topology is shown in Figure 1.

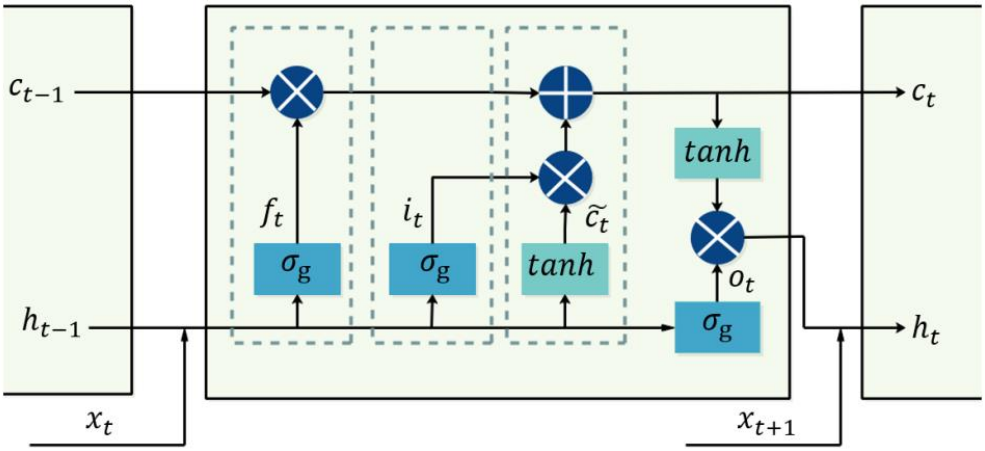

Figure 1. LSTM network topology

Input gate decide which pieces of information to store in the current state. Assume i_t is input gate's activation vector, $\widetilde{c_t}$ is cell input activation vector, c_t represents the cell state vector, The subscript t indexes the time step.

The input calculation formula can be expressed as

$$i_t = \sigma_g(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$\widetilde{c_t} = tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{3}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \widetilde{c_t} \tag{4}$$

Where, σ_g represents the sigmoid function, $tanh$ represents the tanh function, $W \in R^{h \times d}$ and $b \in R^h$ represents the weight matrices and bias vector parameters which need to be learned during training, x_t is the input vector to the LSTM unit, h_{t-1} represents the last output vector of the LSTM unit, $\odot$denotes the Hadamard product.

Output gates assign a value from 0 to 1 to the information control its output in the current state by considering the previous and current states. Assume that $o_t$represents the output gate's activation vector, the output can be computed as:

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

$$h_t = o_t \odot tanh(c_t) \tag{6}$$

Forget gate take the same system as input gate, according to the calculation result is 1 or 0 decide to keep or discard the information. Assume forget gate is f_t, The calculation method for the forget gate can be defined as follows:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{7}$$

In this study, we used historical oil prices the as single-factor for LSTM input, predict the value of next step by training the historical 30 step according to the obtained 1826 data.

3.2. Word Frequency-Guided Multivariate LSTM Model

BP neural network is back propagating, mainly composed of three parts: input layer, middle layer and output layer. The number of nodes in the input and output layers is relatively easy to determine, but the determination of the number of nodes in the hidden layer is a very important and complex problem.

3.3. The GBDT model under the LIGHTGBM framework

Gradient Boosting Decision Tree is kind of ensemble learning technique using Boosting Decision Tree, which is the weak prediction models based on Boosting algorithm. GBDT is established through the phased approaches, which allows it to be more flexible to differentiable loss function to approach optimization.

On the basis of GBDT, LightGBM utilizes two novel techniques Gradient-Based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB),, which allow the algorithm to run faster while maintaining a high level of accuracy.

This study uses the GBDT model in the LightGBM framework for short-term oil price forecasting. The input variable is a seven-dimensional matrix consists of seven searched factors: history oil price, monthly central bank discount rate, monetary aggregates M2 and consumer price index, oil supply, oil demand and oil market size on the given date and its adjacent dates. It is labeled as $x = (x_1,\ x_2,\ldots,\ x_n)$. The output is the price of oil for the given date, which is labeled as$y = (y_1, y_2, \ldots, y_n)$, and the dataset is set as $X = (x_i, y_i)_{i=1}^{n}$.

As the future oil price forecasts are continuous response variables, this study use squared-error L2 loss as the loss function to measure the model forecasting effect.

Assume that the estimation function is $f(x)$,The forecast value of the oil price for the specified i date is $f(x_i)$, the true value is y_i,N is the total number, the fitted error $L(f)_{L_2}$ can be computed as:

$$L(f)_{L_2} = \sum_{i=1}^{N} \frac{1}{2}[y_i - f(x_i)]^2 \tag{8}$$

The optimization goal of the LightGBM model is to optimize the predictive function $f(x)$, and minimize the expected value of the loss function $L\big(y, f(x)\big)$ over the entire distribution of y and X, denoted as $E_{y,X}$:

$$\hat{f} = \arg\min_{f} E_{y,X} L\big(y, f(x)\big) \tag{9}$$

Assume $\widehat{f_0}(x)$ is the initial oil price, $\widehat{f_t}(x)(i = 1,2 \ldots M)$ is the function increments of per iteration. Assume M is the number of regression tree. The overall ensemble function estimate carrying Mth iterations can be expressed:

$$\widehat{f_{\mathrm{M}}}(\mathrm{X}) = \sum_{i=0}^{M} \widehat{f_{\mathrm{t}}}(\mathrm{X}) \tag{10}$$

Assuming that the model at the end of the $t-1$step iteration is $F_{t-1}(x)$, the decision tree model added at the m step iteration is $f_t(x)$, at step m, LightGBM would be trained as:

$$\Gamma_{\mathrm{m}} = \sum_{i=1}^{n} l\left(y_i, \widehat{f_{\mathrm{m-1}}}(x_i) + f_{\mathrm{m}}(x_i)\right) \tag{11}$$

Where, the negative gradient for mth iterations is defined as:

$$g_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x_i)=f_{m-1}(x_i)} \tag{12}$$

Where, $g_{\mathrm{m}}(\mathrm{x})$is the direction in which the loss function fell fastest in the last iteration. Assuming the second-order gradient statistics is h_{im}, the transformed formula is obtained as:

$$\Gamma_m \cong \sum_{i=1}^{n} (g_{im} f_m(x_i) + \frac{1}{2} h_{im} f_m^2(x_i)) \tag{13}$$

Assume $I_j(j = 1,2, \dots, J)$is represents all the samples that end up in leaf node j after traversing the regression tree. Set $q(x)$ to be the decision rules of the tree and $\mathrm{w_j}$ is the vector that denotes the sample weight of leaf nodes. So the calculation could be transformed as follows:

$$\Gamma_m = \sum_{i=1}^{j}\left(\left(\sum_{i\in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i\in I_j}(h_i + \lambda)\, w_j^2\right)\right) \tag{14}$$

For a certain tree structure $q(x)$, can calculate the optimal leaf weight scores of each leaf node w_j^* as:

$$w_j^* = -\frac{\sum_{i\in I_j} g_{im}}{\sum_{i\in I_j}(h_{im}+\lambda)} \tag{15}$$

The extreme value of Γ_k could be solved as:

$$\Gamma_M^* = -\frac{1}{2}\sum_{j=1}^{J}\frac{\left(\sum_{i\in I_j} g_{im}\right)^2}{\sum_{i\in I_j}(h_{im}+\lambda)} \tag{16}$$

Where Γ_T^* could be viewed as the scoring function that measures the quality of the tree structure q. Finally, the objective function after adding the split is:

$$G = \frac{1}{2}\left(\frac{\left(\sum_{i\in I_L} g_{im}\right)^2}{\sum_{i\in I_L} h_{im}+\lambda} + \frac{\left(\sum_{i\in I_R} g_{im}\right)^2}{\sum_{i\in I_R} h_{im}+\lambda} + \frac{(\sum_{i\in I} g_{im})^2}{\sum_{i\in I} h_{im}+\lambda}\right) \tag{17}$$

Where I_L and I_R are the sample sets of the left and right branches.

The calculation results of LightGBM are sensitive to the values of its parameters, so this study uses the particle swarm algorithm to find the optimal values of the parameters feature_fraction, learning_rate, and the other values are taken as default. The values of the parameters are shown in Table 3.

Table 3. Parameters of LightGBM

Parameters	Value
Num_GBDT	200
learning_rate	0.1
max_depth	10
min_data	10
feature_fraction	1.0
bagging_fraction	0.7

3.4. Model evaluation indicators

This paper choose five evaluation criteria, MSE, RMSE, MAE, MAPE, and R^2, to judge and evaluate the predictive results of the model.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{18}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{19}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{21}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right| \times 100\% \quad (22)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \quad (23)$$

Where, n is the sample size, y_i is the true value, $\hat{y}_i$ is the predicted value, $\bar{y}$ indicates the average of the actual values.

4. Experimental results and analysis

4.1. Prediction results of the one-factor LSTM

The prediction results of the LSTM using the individual feature dimensions for analysis are obtained in Figure 2.

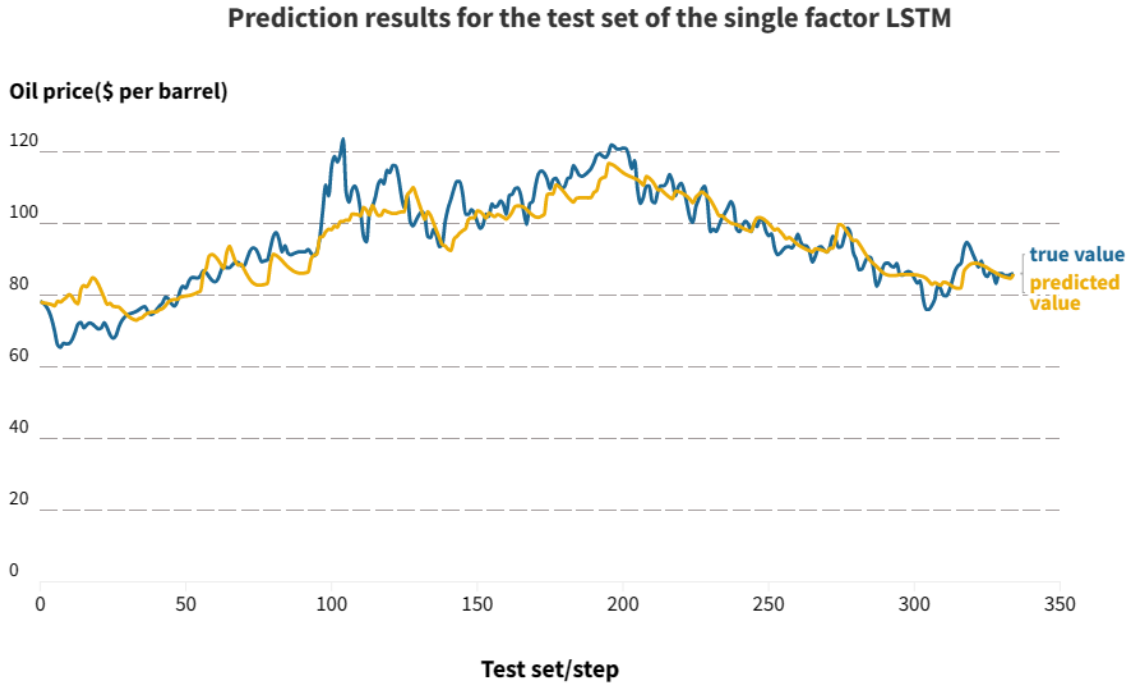

Figure 2. Single-factor LSTM prediction results

Figure 2 indicates that the univariate LSTM prediction model performs remarkably well in predicting the time series data. Through calculation, the minor discrepancy between the predicted and observed values indicates a high degree of accuracy in the forecasting model.

According to the model evaluation indicator in section 3.4, the evaluation results were showed in Table 4.

Table 4. Unitary LSTM prediction evaluation results

	MSE	RMSE	MAE	MAPE	R^2
Test set	36.6794	6.0564	4.5501	0.0484	0.8053

Table 4 shows that the single factor LSTM model performs relatively well, with small mean absolute error and mean absolute percentage error, indicating errors in prediction results is relatively small. And the coefficient of determination is large, meanting that the prediction results correlate relatively well with the true values. However, the mean squared error values are large and there may be some fluctuation or dispersion.

4.2. Prediction results of a multivariate LSTM model

The prediction results of the multivariate LSTM model is presented in Figure 3.

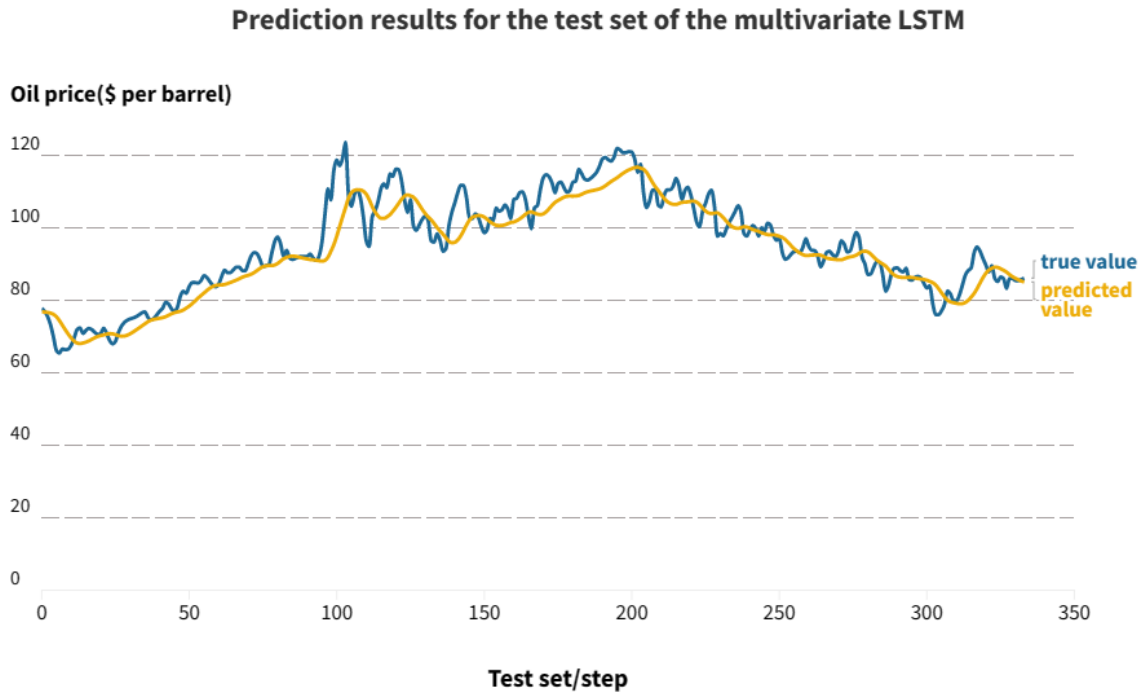

Figure 3. Prediction results of a multivariate LSTM model

Figure 3 indicates that multivariate LSTM model based on word frequency predicts well in the severe fluctuation phase, with stronger time explanations that can more accurately portray price inflection points.

According to the model evaluation indicator in section 3.4, the evaluation results of the training set are showed in Table 4.

Table 5. Evaluation results of multivariate LSTM prediction

	MSE	RMSE	MAE	MAPE	R^2
Test set	27.6917	5.2623	3.8972	0.0399	0.8583

Table 5 indicates that multivariate LSTM model based on word frequency performed relatively well in terms of prediction results. Compared to the unitary LSTM, MSE, RMSE, MAE and MAPE decreased, indicating that the error between the prediction model and the true value was reduced, while R^2 increased, indicating that the correlation between the model prediction results and the true value was enhanced.

4.3. LightGBM prediction results

The predictions obtained using LightGBM are obtained in Figure 3.

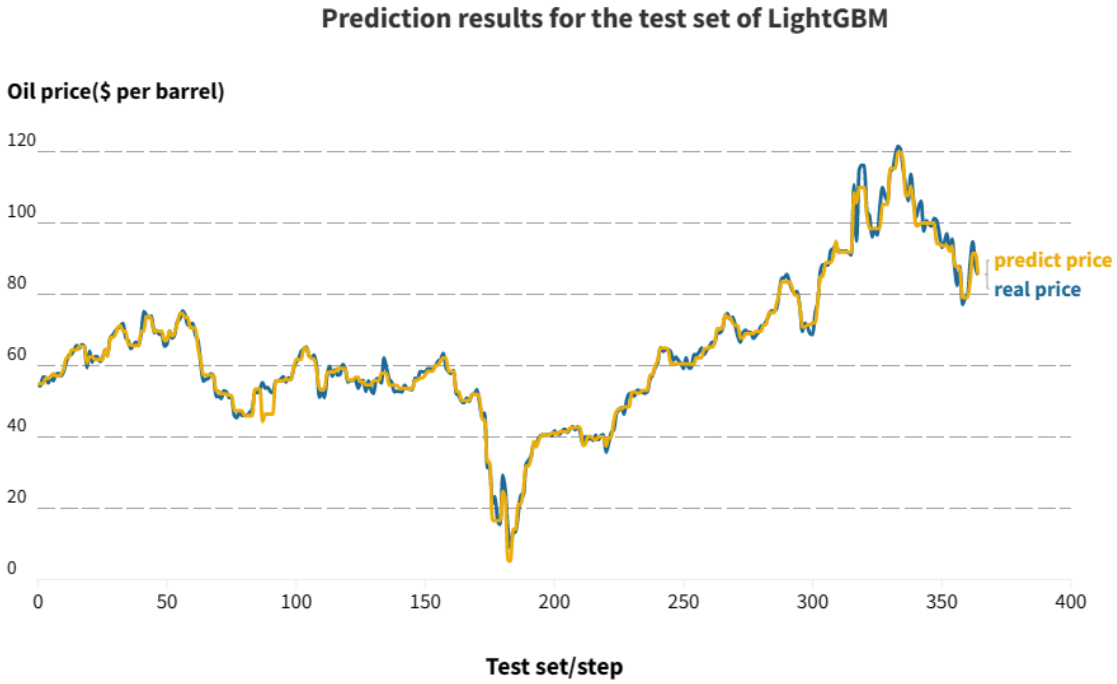

Figure 4. LightGBM prediction results

In Figure 3, during periods of sustained steady growth or decline, the predicted data from the test set fit the real data relatively well, indicating that the model was able to capture these trends effectively. However, during periods when oil price changes are highly volatile and unstable, the accuracy of the forecasts is reduced and there is some error.

The evaluation metrics for prediction results of the training set, test set and tri-fold cross-validation set is showed in Table 6.

Table 6. LightGBM model evaluation results

	MSE	RMSE	MAE	MAPE	R^2
Training set	5.538	2.353	1.368	2.524	0.987
Cross-validation set	7.861	2.769	1.672	2.999	0.981
Test set	4.812	2.194	1.452	2.658	0.987

Table 6 demonstrated that the model produced highly accurate and reliable predictions. MAPE and R^2 also demonstrated the reliable prediction ability of the model, highlighting its usefulness for practical applications.

5. Conclusion

This paper set up single factor LSTM model and multivariate LSTM based on word frequency. We concluded that multivariate LSTM based on word frequency presents better oil price forecasting capability. In addition, this study employed the LightGBM prediction model that considers multiple factors, and evaluated the predictive results of the two types of models using evaluation criteria such as MSE. In conclusion, the prediction accuracy of the LightGBM model is relatively better, while the LSTM model has a higher temporal interpretation.

In order to identify the key themes and trends in the alternative data sources, the incorporation of alternative data sources such as internet search trends and social media sentiment can provide a more complete picture of the various sources of information that influence oil prices.

References

[1] Caldara D, Cavallo M, Iacoviello M. Oil price elasticities and oil price fluctuations [J]. Journal of Monetary Economics, 2019, 103 (1): 1 – 20.

[2] Kuhe D A. The Dynamic Relationship between Crude Oil Prices and Stock Market Price Volatility in Nigeria: A Cointegrated VAR-GARCH Model[J]. Current Journal of Applied Science and Technology, 2019, 59 (1): 1 – 12.

[3] Cheng F, Li T, Wei Y, et al. The VEC-NAR model for short-term forecasting of oil prices [J]. Energy Economics, 2019, 78 (1): 656 – 667.

[4] Ma F, Liao Y, Zhang Y, et al. Harnessing jump component for crude oil volatility forecasting in the presence of extreme shocks [J]. Journal of Empirical Finance, 2019, 52: 40 - 55.

[5] Aamir M, Shabri A. Modelling and forecasting monthly crude oil price of Pakistan: A comparative study of ARIMA, GARCH and ARIMA Kalman model [C]//AIP Conference Proceedings. AIP Publishing LLC, 2016, 1750 (1): 060015.

[6] Mirmirani S, Cheng Li H. A comparison of VAR and neural networks with genetic algorithm in forecasting price of oil[M]//Applications of artificial intelligence in finance and economics. Emerald Group Publishing Limited, 2004: 203 - 223.

[7] Guo J, Zhao Z, Sun J, et al. multi-perspective crude oil price forecasting with a new decomposition-ensemble framework [J]. Resources Policy, 2022, 77: 102737.

[8] Zhou Y, Li T, Shi J, et al. A CEEMDAN and XGBOOST-based approach to forecast crude oil prices [J]. Complexity, 2019, 2019: 1 - 15.

[9] Lin Y, Chen K, Zhang X, et al. Forecasting crude oil futures prices using BiLSTM-Attention-CNN model with Wavelet transform [J]. Applied Soft Computing, 2022, 130: 109723.

[10] Gao H, Hu Y, Yu L. Forecasting Crude Oil Price with an Emd-based LSTM Neural Network Ensemble Learning Paradigm. [J]. Computer Applications and Software, 2021, 38 (10): 78 - 83.

[11] RATTI R A, VESPIGNANI J L. Oil prices and global factor macroeconomic variables [J/OL]. Energy Economics, 2016, 59 (1): 198 - 212.

[12] PIRYONESI S M, EL-DIRABY T E. Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index[J/OL]. Journal of Infrastructure Systems, 2020, 26 (1): 04019036.