Forecasting crude oil futures prices using Extreme Gradient Boosting

Yang, Qian; He, Kaijian; Zheng, Linyuan; Wu, Chiwai; Yu, Yi; Zou, Yingchao

Published in:
Procedia Computer Science

Published: 01/01/2023

Document Version:
Final Published version, also known as Publisher's PDF, Publisher's Final version or Version of Record

Published version (DOI):
10.1016/j.procs.2023.08.069

Publication details:
Yang, Q., He, K., Zheng, L., Wu, C., Yu, Y., & Zou, Y. (2023). Forecasting crude oil futures prices using Extreme Gradient Boosting. *Procedia Computer Science*, *221*, 920-926. https://doi.org/10.1016/j.procs.2023.08.069

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 221 (2023) 920–926

Procedia Computer Science

www.elsevier.com/locate/procedia

Information Technology and Quantitative Management (ITQM 2023)

Forecasting crude oil futures prices using Extreme Gradient Boosting

Qian Yang[a], Kaijian He[a], Linyuan Zheng[a], Chiwai Wu[b,c], Yi Yu[d], Yingchao Zou[a,*]

[a]*College of Tourism, Hunan Normal University, Changsha 410081, China*
[b]*School of Tourism Management, Macao Institute for Tourism Studies, Macao, China*
[c]*Department of Marketing, City University of Hong Kong, Hong Kong, China*
[d]*School of Foreign Languages, Hunan University, Changsha, 410082, China*

Abstract

Multi-source data is widely used in the field of energy future prices forecasting, the improvement of forecast ability and data screening are becoming the focus of current research. In this paper, two tree-based models (namely, Random Forest and XGBoost model) are employed to predict China's crude oil future prices. The empirical analysis confirms that Random Forest and XGBoost model have superior prediction performances than benchmark and the XGBoost model performs best. An important finding is that there is a time gap between investor information search and processing because the prediction performance within the time lags is obviously superior than that of the current period.

Peer-review under responsibility of the scientific committee of the Tenth International Conference on Information Technology and Quantitative Management

Keywords: Oil futures forecasting; search index; XGBoost; random forest

1. Introduction

Crude oil plays an indispensable role in economy development and production promotion around the world because its attributes of energy commodities, investment products and strategic resources [1, 2]. From the macro perspectives, the volatility of crude oil future prices concerns the stability and long-term development of global economy,society and politics [3, 4]. Accurate forecasting of crude oil futures has great implications for the risk management and investment portfolio for investors and decision makers from micro perspectives [5]. Many studies utilized different methods to improve the accuracy of crude oil prices prediction such as optimizing the structure of model, assembling new model based on the prediction advantage of different models, and adding additional influencing factors as the exogenous variables [6, 7, 8, 9, 10]. For example, [8] employed consumer price index (CPI), industrial price index (IPI), U.S. oil imports (USI), Baltic Dirty Tanker Index (BTI), capacity utilization (CU), LIBOR (LBR) and S& P500 Index (SP5), currency for Japanese yen to US dollar parity (JUS) and Chinese yuan to US dollar parity (CUS) as additional forecasting indicators to predict the price of Brent crude oil futures. [10]proposed a hybrid method combining k-means, KPCA and KELM by using the "divide and conquer" strategy, and indicated that the "divide and conquer" strategy can effectively improve the forecasting performance of model.

*Corresponding author. Tel.:+86-18511075982
Email address: evangeline1203@hunnu.edu.cn (Yingchao Zou)

Peer-review under responsibility of the scientific committee of the Tenth International Conference on Information Technology and Quantitative Management
10.1016/j.procs.2023.08.069

Extent research have shown that the inclusion of additional influencing factors as exogenous variable data can effectively improve the prediction performance of the model [11, 8]. With the progress of internet technique, the explosive spread of network information makes the search and transmission of information become the main way of information acquisition for network users in recent years. Especially in the investment market which requires access to the latest information quickly, investors rely on information obtained from internet website or social media platform as a reference to make investment divisions. These network footprints and search behavior was called investor attention and studied in the forecasting area of energy price,stock price, etc [12, 13, 14]. Extent studies indicated that investor attention about the multiple perspectives has impact on crude oil future prices volatility [13, 14, 15]. For example,[16] discussed the influence of investor attention on the crude oil futures volatility, found out the attention index measures the retail investors attention perfectly and the using of investor attention effectively improved the predictive performance of model.[3] revealed the mechanism of investor attention on crude oil futures prices and analyzed the reaction of crude oil prices to the investor attention shock.

Studies have indicated that big data representing investor attention on the internet contributes to better performance of the predictive model [14, 17],but not all keywords data have the unique ability. Keywords, searched by investors, represent investors attention and contain some potential information about the changes or trends of crude oil prices which may influence the decision-making of investors and in turn affect the changes of crude oil future prices. It's necessary to select the most effective keywords from a large amount of keywords data which is provided by internet platform. Current research using keywords data filter the effective keywords by the method of correlation coefficient, principal component analysis or manually limited keywords. For example,[18] calculated the maximum correlation coefficient to select the most relevant search keywords as Google search volume index to measure the investor attention. [19] grouped the keywords into 3 groups, extracting the first three principal components of each group of keywords and the first three principal components of all keywords to screen the core terms for crude oil price forecasting. The selection of effective keywords is still an issue discussed by relevant scholars.

In this study, the main purpose of this paper is to find out whether the forecasting performance of tree-based models in crude oil futures prediction is significantly different from that of time series model, using the advantages of tree-based models in the features selection mechanism.

The rest of this paper is organized as follows. Section 2 mainly introduces the background of Random Forest and XGBoost algorithm. The experiments results analysis are presented in Section 3. In Section 4,we summarized the main conclusion of this paper.

2. Methodology

2.1. Random forest

Random forest is an ensemble learning method with decision tree framework, which constructs a series of random decision trees and averages the output of results for each tree to obtain better forecasting accuracy value and control over-fitting problem [20, 21, 22]. In random forest algorithm, it's a tree-based classification method when the dependent variable is classified, while it is a tree-based regression method when the dependent variable is continuous [23]. Specifically, random forest algorithm returns the importance value to examine the casual relationship between input variables and target variables and to measure the contribution of each variable for model's forecasting accuracy [24]. Therefore, variables with high contribution for predictive accuracy can be screened by the importance value of keywords and the core keywords that affects the forecasting performance of model can be determined. Random forest is composed of bagging algorithm and CART trees [20]. The basic procedure is as follows:

Firstly, randomly select n samples from the training set to generate several new sub-training sets. Then several classification and regression trees are trained with the sub-training sets. K features are randomly selected from all features, and the optimal split points are determined for subtrees splitting.

Secondly, a classification and regression tree is generated on the assumption that the input space is split into N units $(R_1, R_2, \ldots, R_n)$, using the square error minimization to find the optimal segmentation feature and the optimal split point. Assume that $\widehat{x^{(j)}}$ variables and s value as the segmentation feature and split point, and defining two space regions, the formula is represented by Eq. (1):

$$R_{1(j,s)} = x|x^{(j)} \leq s, R_{2(j,s)} = x|x^{(j)} > s \tag{1}$$

The two sub-regions are recursively divided to find the optimal score pair (j, s) until the stop condition is satisfied. A least squares regression tree is generated by averaging and outputting the final partition region values, which can be expressed as in Eq. (2).

$$f(x) = \sum_{n=1}^{N} C_n I \tag{2}$$

Where C_n represents the output value and $I(x \in R_n)$ is an indicative function.

Thirdly, several classification and regression trees are trained by sub-training sets, selecting k features randomly and determining the optimal split point for the further splitting.

Fourthly, the final predicted result of the random forest is the average of all outputs of classification and regression trees.

2.2. XGBoost model

XGBoost is an advanced supervised algorithm within the ensemble learning framework and has been applied in many fields [25]. There are n decision trees in XGBoost model, new tree will be iteratively updated by the gradient algorithm and the residual of previous tree until the most accurate result is produced by reducing the residuals [25, 26]. The objective function of XGBoost consists of training loss and regularization term as in Eq. (3):

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t) \tag{3}$$

Where $\sum_{i=1}^{n} l(y_i, \hat{y}_i)$ refers to the training loss, $\sum_{t=1}^{T} \Omega(f_t)$ refers to the regularization term for punishing the complexity of trees, preventing over-fitting and increasing the generalization ability of tree model. l represents the loss function, $\hat{y}_i$ shows the forecast value of the ith sample x_i.

The result of XGBoost model is the sum of output values of all trees, which can be expressed as Eq. (4).

$$\hat{y} = \sum_{t=1}^{T} f_t(x_i) \tag{4}$$

x_i represents the sample i, $\hat{y}$ means the output of XGBoost model, f_t shows the tree t in the model.

XGBoost model can also output corresponding results to explain the importance of each keywords and analyze the most relevant feature variables. This model decides variables' importance by three indicators, i.e. Gain, Cover, Frequency [27]. Among them,Gain represents the importance of tree node features, Cover means the number of observations, Frequency shows the relative percentage of time that corresponding feature appears in the model tree. The importance of each keyword can be acquired by the feature of Gain,the higher the Gain value, the more important it is for model forecasting [28]. Therefore,XGBoost model is a boosting algorithm that can not only carry out the feature selection but improve the forecasting accuracy in terms of predictive ability.

3. Experiment results

3.1. Data

China's Shanghai crude oil future prices and the search index related to the crude oil price are chosen as the data variables. The daily closing price of principal continuous contract of Shanghai crude oil future is determined for crude oil future price forecasting as the dependent variable and the daily search index data are chosen as the exogenous variables. We acquire the daily crude oil future closing price from Flush database and collect the search query data from Baidu Index platform (https://index.baidu.com/). The data length is determined by the listing time of China's Shanghai crude oil future, from March 26,2018 to February 21,2023. Random forest algorithm and

XGBoost algorithm have advantages in processing a large number of observations, acquiring a smaller set of variables and assessing the importance of feature variables [20, 28]. In this study,we choose random forest and XGBoost algorithm as main methods to construct tree-based model for crude oil futures forecasting. Meanwhile, ARIMA model and ARIMA model with exogenous variables are utilized as benchmark model for comparison. Additionally, Root Mean Square Error (RMSE) is selected as evaluation criteria to judge the forecast performance of the model.

Keywords acquisition is considered as a critical step before the crude oil future prediction. In terms of keyword selection, many studies drew up the seed keywords based on the perspective of crude oil, alternative energy, relative market, emergency, geopolitics and economics [1, 29, 10, 30]. For example,[1] acquired 25 keywords and constructed 5 investor concern indexes according to the categories of crude oil, related market, alternative energy, geopolitics and macroeconomy. This study considered the key factors of crude oil price, alternative energy, related market, related conference and China's oil price pricing mechanism determined 5 categories and acquired 34 keywords from Baidu index.

Descriptive statistical analysis of crude oil futures data and Baidu search index data are carried out in this paper (shown as Table 1). The P value of ADF test of Baidu search index related to crude oil future were mostly 0.01, while the original data of crude oil futures did not pass the ADF test. Therefore, differential processing was carried out for it.

Table 1. Descriptive statistical analysis of crude oil futures and Baidu search index data

Variable	Mininum	Median	Mean	Maximum	Standard Deviation	Skewness	Kurtosis	ADF	Dickey-Fuller
Data$_{OilPrice}$	206.6	456.1	464.7	806.6	125.56	0.27	-0.26	0.77	-1.55

Considering that there is a time gap in search behavior, which means that the search behavior with several days in advance and search behavior in the current period may have an impact on the crude oil future prices fluctuations. Therefore, we calculated the search query data with a lag of 10 days to explain the difference of influence on the crude oil futures volatility between search behavior within 10 days in advance (represented by lag10 in this paper) and the influence of search behavior in the current period (showed by lag0 in this paper).

In this paper, the optimal sub-node of trees is determined for out-of-sample forecast based on the fitting results of in-sample. Specifically,in Random Forest model, the main reference parameters for determining the mtry parameter are the minimum RMSE and Mean of squared residuals. The reference parameter for selecting the optimal nsplit is RMSE$_{XGBoost}$ in the XGBoost model. The specific results of each time lag are listed in Table 2.

Table 2. The optimal feature division of crude oil future prices forecast based on Random Forest and XGBoost model

Parameter	mtry$_{RF}$	RMSE$_{RF}$	RMSE$_{XGBoost}$	iter$_{XGBoost}$
lag0	11	3.3753	0.0010	789
lag1	56	3.2558	0.0008	652
lag2	38	3.2784	0.0007	635
lag3	89	3.2440	0.0007	609
lag4	147	3.2495	0.0006	573
lag5	161	3.2263	0.0006	991
lag6	140	3.2396	0.0006	597
lag7	178	3.2217	0.0006	564
lag8	177	8.1502	0.0006	563
lag9	182	8.1502	0.0006	995
lag10	221	8.1502	0.0006	584

Results from Table 2 indicated that the characteristic of Random Forest model is the change for RMSE$_{RF}$

shows a trend from decline to rise with the increase of time lags (namely, the extension of the advance search time). The minimum $RMSE_{RF}$ appears at the 7th order lag (i.e. lag7),with the specific value of 3.2217, and the optimal number of mtry for model tree is 178. In the XGBoost model, the most prominent feature is the value of $RMSE_{XGBoost}$ decreases with the increase of time lag order, in which the minimum value of $RMSE_{XGBoost}$ is 0.000553 and the corresponding lag order is 10. The optimal iteration of the model tree is 584.

In summary, the mtry of Random Forest model tree is 178 at 7th lag order when the RMSE of model is the minimum and the optimal prediction accuracy is acquired. The $RMSE_{XGBoost}$ of XGBoost model is 0.000553 at 10th lag order while the best number of iteration is 584. This paper further makes the out-of-sample prediction in terms of results in Table 2 and the forecast results are shown in Table 3.

Table 3. Out-of-sample RMSE value of crude oil futures prices forecast based on Random Forest and XGBoost model

Model	ARIMA	ARIMAX	Random Forest	XGBoost
$\mathbf{RMSE}_{lag0}$	17.5073	17.7484	16.8198	16.8268
$\mathbf{RMSE}_{lag7}$	N/A	17.6072	16.8058	16.7813
$\mathbf{RMSE}_{lag10}$	N/A	17.6442	16.6480	16.6583

RMSE is selected as the evaluation criterion for the forecasting performance of models. Three obvious findings can be obtained from Table3. Firstly, compared with the benchmark models ARIMA and ARIMAX, Random Forest model outperformed based on the search query data of current period. Secondly, when all search query data (namely 34 keywords data) were added into the three models, the forecasting accuracy of ARIMA model with exogenous variables performed worse than that of the model without exogenous variables. This means that putting all unfiltered search data into the model does not necessarily make the model perform better, although the prediction performance of the model gradually improves with the increase of the lag order. Additionally, XGBoost and Random Forest model with different time lags have higher forecasting accuracy, which indicates that search query data characterizing information search behavior in advance has a significant improvement for the model forecasting accuracy. The prediction performance of Random Forest model appears at the 7th lag order with the RMSE value of 16.8058, which was further verified according to the fitting result of in-sample. So it's obvious that the search behavior with one week in advance influences greater on the prices volatility of crude oil futures than that in the current period. For XGBoost model, the feature of fluctuation for prediction accuracy is clearly with the forecasting performance with the 10th lag and its RMSE value is 16.6583. It can be seen that XGBoost model predicts with lower RMSE value.

4. Conclusions

This paper applied two tree-based models (namely, Random Forest and XGBoost model) to predict China crude oil future prices and used ARIMA model and ARIMA model with exogenous variables as benchmarks for comparison. We aim to screen out the key variables which affects the investor decision-making and model forecasting accuracy, and discuss the prediction performance with different time lags. Work in this paper confirms that XGBoost model performs superior, and the search behavior with lag of more than one week influences the model significantly, which shows a better forecasting performance.

Acknowledgments

The work described in this paper was supported by a grant from National Natural Science Foundation of China (Grant no. 72271089),Hunan Provincial Natural Science Foundation of China (Grant no. 2022JJ30401).

References

[1] H. Qu, G. Li, Multi-perspective investor attention and oil futures volatility forecasting, Energy Economics 119 (2023) 106531. doi:https://doi.org/10.1016/j.eneco.2023.106531.

[2] K. Xu, H. Niu, Do eemd based decomposition-ensemble models indeed improve prediction for crude oil futures prices?, Technological Forecasting and Social Change 184 (2022) 121967. doi:https://doi.org/10.1016/j.techfore.2022.121967.

[3] T. Yao, Y.-J. Zhang, C.-Q. Ma, How does investor attention affect international crude oil prices?, Applied Energy 205 (2017) 336–344. doi:https://doi.org/10.1016/j.apenergy.2017.07.131.

[4] X. Feng, S. Wang, S. Wang, X. Zhang, Exploring the relationship between cude oil spot and futures prices: New perspective from multi-scale decomposition, International Journal of Energy and Statistics 02 (2014) 269–286. doi:10.1142/S2335680414500185.

[5] J. J. Guo, Z. L. Zhao, J. Y. Sun, S. L. Sun, Multi-perspective crude oil price forecasting with a new decomposition-ensemble framework, RESOURCES POLICY 77, times Cited in Web of Science Core Collection: 0 Total Times Cited: 0 Cited Reference Count: 59 ER -. doi:10.1016/j.resourpol.2022.102737.

[6] L. Yu, S. Wang, K. K. Lai, Forecasting crude oil price with an emd-based neural network ensemble learning paradigm, Energy Economics 30 (5) (2008) 2623–2635. doi:https://doi.org/10.1016/j.eneco.2008.05.003.

[7] X. Feng, S. Wang, S. Wang, X. Zhang, Exploring the relationship between crude oil spot and futures prices: New perspective from multi-scale decomposition, International Journal of Energy & Statistics 2 (04) (2014) 269–286.

[8] K. Bekiroglu, O. Duru, E. Gulay, R. Su, C. Lagoa, Predictive analytics of crude oil prices by utilizing the intelligent model search engine, Applied Energy 228 (2018) 2387–2397. doi:https://doi.org/10.1016/j.apenergy.2018.07.071.

[9] Y. Ding, A novel decompose-ensemble methodology with aic-ann approach for crude oil forecasting, Energy 154 (JUL.1) (2018) 328–336.

[10] Y. Yang, J. E. Guo, S. Sun, Y. Li, Forecasting crude oil price with a new hybrid approach and multi-source data, Engineering Applications of Artificial Intelligence 101 (2021) 104217. doi:https://doi.org/10.1016/j.engappai.2021.104217.

[11] S. Birim, I. Kazancoglu, S. K. Mangla, A. Kahraman, Y. Kazancoglu, The derived demand for advertising expenses and implications on sustainability: a comparative study using deep learning and traditional machine learning methods, Annals of Operations Research (1) (2022) 1–31.

[12] C. Deng, L. Ma, T. S. Zeng, Crude oil price forecast based on deep transfer learning: Shanghai crude oil as an example, SUSTAINABILITY 13 (24), times Cited in Web of Science Core Collection: 1 Total Times Cited: 1 Cited Reference Count: 19 ER -. doi:10.3390/su132413770.

[13] Y. Gao, K. Zhao, C. Wang, C. Liu, The dynamic relationship between internet attention and stock market liquidity: A thermal optimal path method, Physica A: Statistical Mechanics and its Applications 550 (2020) 124180. doi:https://doi.org/10.1016/j.physa.2020.124180.

[14] J. Xiao, Y. Wang, Investor attention and oil market volatility: Does economic policy uncertainty matter?, Energy Economics 97 (2021) 105180. doi:https://doi.org/10.1016/j.eneco.2021.105180.

[15] Y. Gao, Y. Wang, C. Wang, C. Liu, Internet attention and information asymmetry: Evidence from qihoo 360 search data on the chinese stock market, Physica A: Statistical Mechanics and its Applications 510 (2018) 802–811. doi:https://doi.org/10.1016/j.physa.2018.07.016.

[16] L. Han, Q. Lv, L. Yin, Can investor attention predict oil prices?, Energy Economics 66 (2017) 547–558. doi:https://doi.org/10.1016/j.eneco.2017.04.018.

[17] M. Qadan, H. Nama, Investor sentiment and the price of oil, Energy Economics 69 (2018) 42–58. doi:https://doi.org/10.1016/j.eneco.2017.10.035.

[18] X. Li, J. Ma, S. Wang, X. Zhang, How does google search affect trader positions and crude oil prices?, Economic Modelling 49 (2015) 162–171. doi:https://doi.org/10.1016/j.econmod.2015.04.005.

[19] D. Fantazzini, Forecasting the real price of oil using online search data, Int. J. of Computational Economics and Econometrics 4 (2014) 4 – 31. doi:10.1504/IJCEE.2014.060284.

[20] L. Peng, L. Wang, X.-Y. Ai, Y.-R. Zeng, Forecasting tourist arrivals via random forest and long short-term memory, COGNITIVE COMPUTATION 13 (1) (2021) 125–138, article. doi:10.1007/s12559-020-09747-z.

[21] D. Shrestha, W. Tan, B. Gaudel, N. Rajkarnikar, S. R. Jeong, Multi-dimensional analysis and prediction model for tourist satisfaction, KSII Transactions on Internet and Information Systems 16 (2022) 480–501. doi:10.3837/tiis.2022.02.007.

[22] R. K. Mishra, S. Urolagin, J. Angel Arul Jothi, N. Nawaz, H. Ramkissoon, Machine learning based forecasting systems for worldwide international tourists arrival, International Journal of Advanced Computer Science and Applications 12 (2021) 2021. doi:10.14569/IJACSA.2021.0121107.

[23] C. Kuzey, A. S. Karaman, E. Akman, Elucidating the impact of visa regimes: A decision tree analysis, Tourism Management Perspectives 29 (2019) 148–156.

[24] N. Leelawat, S. Jariyapongpaiboon, A. Promjun, S. Boonyarak, K. Saengtabtim, A. Laosunthara, A. Yudha, J. Tang, Twitter data sentiment analysis of tourism in thailand during the covid-19 pandemic using machine learning, Heliyon 8 (2022) e10894. doi:10.1016/j.heliyon.2022.e10894.

[25] A. Shehadeh, O. Alshboul, R. E. Al Mamlook, O. Hamedat, Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, lightgbm, and xgboost regression, Automation in Construction 129 (2021) 103827. doi:https://doi.org/10.1016/j.autcon.2021.103827.

[26] W. Zhang, C. Wu, L. Tang, X. Gu, L. Wang, Efficient time-variant reliability analysis of bazimen landslide in the three gorges reservoir area using xgboost and lightgbm algorithms, Gondwana Researchdoi:10.1016/j.gr.2022.10.004.

[27] S. K. Bhagat, T. Tiyasha, S. M. Awadh, M. T. Tran, A. H. Jawad, Z. M. Yaseen, Prediction of sediment heavy metal at the australian bays using newly developed hybrid artificial intelligence models, ENVIRONMENTAL POLLUTION 268, article. doi:10.1016/j.envpol.2020.115663.

[28] J. Dong, W. Zeng, L. Wu, J. Huang, T. Gaiser, A. K. Srivastava, Enhancing short-term forecasting of daily precipitation using numerical weather prediction bias correcting with xgboost in different regions of china, Engineering Applications of Artificial Intelligence 117 (2023) 105579. doi:https://doi.org/10.1016/j.engappai.2022.105579.

[29] J. Wang, G. Athanasopoulos, R. J. Hyndman, S. Wang, Crude oil price forecasting based on internet concern using an extreme learning machine, International Journal of Forecasting 34 (4) (2018) 665–677. doi:https://doi.org/10.1016/j.ijforecast.2018.03.009.

[30] L. Yu, Y. Zhao, L. Tang, Z. Yang, Online big data-driven oil consumption forecasting with google trends, International Journal of Forecasting 35 (1) (2019) 213–223. doi:https://doi.org/10.1016/j.ijforecast.2017.11.005.