Crude Oil Price Prediction Based on Multiple Ensemble Learning Algorithms

Junhao Zhang[1 *]

[1]School of Economics and Management, Wuhan University, Wuhan, China

*Corresponding author: 2019301050249@whu.edu.cn

Abstract. The volatility of crude oil price affects people from all walks of life. Since the start of the 21st century, the frequent booms and busts in crude oil price have caused a greater emphasis on forecasting the price. In order to make prediction from a new perspective, this paper combined both financial and non-financial factors and applied both bagging and boosting algorithms in ensemble learning to construct models. According to the results, both bagging and boosting algorithms achieved impressive improvements over the benchmark model. The ensemble learning models can not only reach high goodness-of-fit, but also attain excellent accuracy in direction prediction, which indicates the models can well realize the long term trend prediction of crude oil price. In brief, this paper enriched the content in predicting crude oil price and offer a guideline for the application of ensemble learning algorithms in such a field.

Keywords: crude oil price; ensemble learning; bagging; boosting.

1. Introduction

The crude oil price forecasting has been a wide topic for decades. Even tiny Fluctuations in national crude oil price could cause exceeding variations in multiple areas, both for crude oil importer and exporter countries [1]. Crude oil, from another perspective, as a widely recognized tool to resist inflation and risk, its price changes correspond to changes in international situation. To be specific, the emergence of COVID-19 led to a drop in oil prices in early 2020, while the Russia-Ukraine conflict led to a surge in the first half of 2022.

A qualified crude oil price prediction model will undoubtedly play a significant instructive role in strategies or decisions for both people and government. Many studies have been performed to find factors that can be added to the prediction model. Barskey and Kilian figured out the correlation between oil and macroeconomic since 1970s. They introduced varied non-financial factors (e.g., CPI and real investment), which truly affect oil price and visualized the precise relevancy in 2004 [2]. In 2018, Mikhaylov discussed the impact of oil prices on world stock indices. He proposed a Probit model to analyze the major trends in oil market and describe the intrinsic correlation [3]. In 2019, An, Mikhaylov and Moiseev selected the best factors which market participants can use for machine learning approach. They also performed their prediction based on linear regression and modified linear regression algorithms [4]. Those work focuses on either financial or non-financial areas. Therefore, one of the marginal contributions of this paper is the combination of financial and non-financial factors affecting crude oil price.

Except for the choice of predictors, a number of papers have applied different kinds of algorithms to predict the price. In 2005, Moshiri and Foroutan concluded that crude oil futures prices were nonlinear and applied multiple nonlinear models (e.g., ARMA, GARCH and ANN) to predict the chaotic price fluctuation [5]. Xie et al. created a new prediction model based on Support Vector Machine, and achieved better results than ARIMA and BPNN in 2006 [6]. In 2017, Chen, He and Tso purposed a new hybrid forecasting model based on deep learning, while the empirical results showed improvement in forecasting accuracy [7]. Assaad and Fayek tried to make deeper exploration in deep learning and had upgraded LSTM by combining it with CNN in 2021 [8].

Recent papers on crude oil price prediction have focused on deep learning. However, one crucial prerequisite is that the size of the data needs to be large enough to meet the requirements of deep learning for the training dataset. Considering non-financial factors data is always monthly, which means the scale of datasets won't be large enough and deep learning method aren't going to work well as one expected. Therefore, another marginal contribution is, instead of deep learning, this paper concentrates on ensemble machine learning and applies both bagging and boosting methods. Ensemble learning can not only make the most of limited datasets, but also reduce the generalization error effectively. Gabralla, Jammazi and Abraham presented a model using ensemble machine learning algorithm. They had combined IBL, SMOreg and ensemble approach to improve the model and show the improvement by comparing the evaluation metrics in 2013 [9]. While in 2015, Aydogmus and Ekinci mainly investigated bagging ensemble models such as bagged artificial neural networks (BANN) and bagged regression trees (BRT). It is concluded that the bagging ensemble method could optimize the forecast accuracy [10]. Those papers have suggested ensemble learning algorithms have the potential to work well enough in forecasting areas.

In order to fully reflect the above marginal contribution, the paper is further organized into the following parts. The Sec. 2 mainly introduces the methodology including data, exploratory data analysis, models and metrics. The Sec. 3 discusses the results obtained from the study and current limitation. Eventually, a brief summary will be given in Sec. 4.

2. Methodology

2.1 Data Collection and Descriptive Statistics

In this paper, a lot of monthly datasets have been collected. All of the datasets are monthly from 2001-01 to 2019-12 and crawled from U.S. Energy Information Administration (EIA) or from Kaggle. The selected dependent variable(y) is the price of WTI crude oil(crude), the unit used is dollars per barrel. As for the independent variables (x), this study chooses: (1) the OPEC crude oil production(production) which is measured in million barrels per month; (2) the sp500 index(sp500); (3) the US dollar index(udi); (4) the US Consumer Price Index(ucpi); (5) the price of gold(gold) which is measured in dollars per ounce; (6) the brent oil future price(brent); (7) the inventory of WTI crude oil(inventory) which is measured in million barrels. The descriptive statistics before data scaling is given in Table. 1.

Table 1. Descriptive statistics before scaling.

variables	count	mean	std	min	25%	50%	75%	max
crude	228	63.41	25.77	19.39	45.11	59.35	84.68	133.88
udi	228	89.66	11.14	72.37	80.85	87.58	96.56	119.79
ucpi	228	217.63	24.07	175.10	197.40	218.75	237.53	257.35
production	228	33.98	2.35	27.25	33.30	34.34	35.52	37.82
brent	228	67.08	29.69	19.62	45.74	62.78	89.51	140.67
gold	228	969.26	451.20	260.75	466.56	1125.51	1298.83	1780.65
sp500	228	1579.88	595.39	700.82	1134.69	1350.45	2013.95	3113.87
inventory	228	1762850	143325	1460452	1658453	1757518	1862183	2064123

2.2 Data Preprocessing

Before developing and training models, the selected dataset should be appropriately pre-processed. On this basis, this work finished data scaling and data splitting. One can obviously find that the order of magnitude difference between the values of the variables is quite large. To reduce the negative impact of the magnitude difference and improve the accuracy of models, it's imperative to do suitable data scaling. In this work, those time series datasets were scaled to a range of [0,1] using the log scaling method since the data fits better after log scaling than using MinMaxScaler:

$$z = \frac{\ln(x)}{\ln(\max(x))} \tag{1}$$

where z is the scaled version of the unscaled variable x. The descriptive statistics table after scaling is as follow. With regard to data splitting, this paper used the former 80% data as the training dataset, which is in the period from 2001-01 to 2016-03. The test dataset used the latter 20% data which is in the period from 2016-04 to 2019-12.

Table 2. Descriptive Statistics after scaling.

variables	count	mean	std	min	25%	50%	75%	max
crude	228	0.829	0.091	0.605	0.778	0.834	0.906	1.000
udi	228	0.938	0.025	0.895	0.918	0.935	0.955	1.000
ucpi	228	0.969	0.020	0.931	0.952	0.971	0.986	1.000
production	228	0.970	0.020	0.910	0.965	0.973	0.983	1.000
brent	228	0.828	0.099	0.602	0.773	0.837	0.909	1.000
gold	228	0.899	0.078	0.743	0.821	0.939	0.958	1.000
sp500	228	0.908	0.044	0.815	0.874	0.896	0.946	1.000
inventory	228	0.989	0.006	0.976	0.985	0.989	0.993	1.000

2.3 Exploratory Data Analysis

Before the main project begin, some essential exploratory data analysis may help to thoroughly understand the whole dataset. Fig. 1 shows the variance of crude oil price, the brent oil future price and the difference (price minus futures price) between them. It's widely recognized that the relationship between commodity prices and its futures prices could reflect the market trend. The position of difference line and basic zero line show how enthusiastic investors were about the crude oil market before 2008. After the Financial crisis of 2007-2008, the frenzy gradually returned to normal. One can also observe the astonishing collapse of crude oil prices in 2014-2016 due to oversupply and underdemand. One important approach that could help us find out if there exists enough relevance between dependent variable and independent variables is the correlation analysis. As shown in Figs. 2 and 3, if |corr(crude,x)| < 0.5, the varible will be drop to avoid the interference brought by the factors of weak correlation. Hence, the variables sp500 and inventory are discarded and selected the other five independent variables as predictors to further build forecasting models.

Fig. 1 Price and futures price

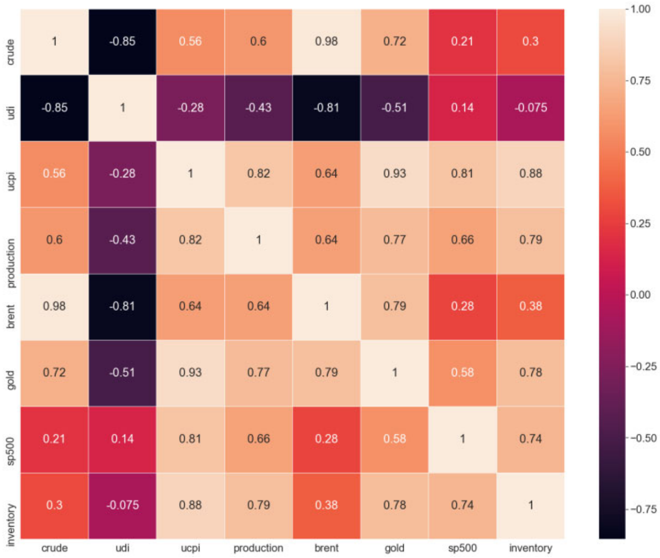

Fig. 2 Heatmap analysis

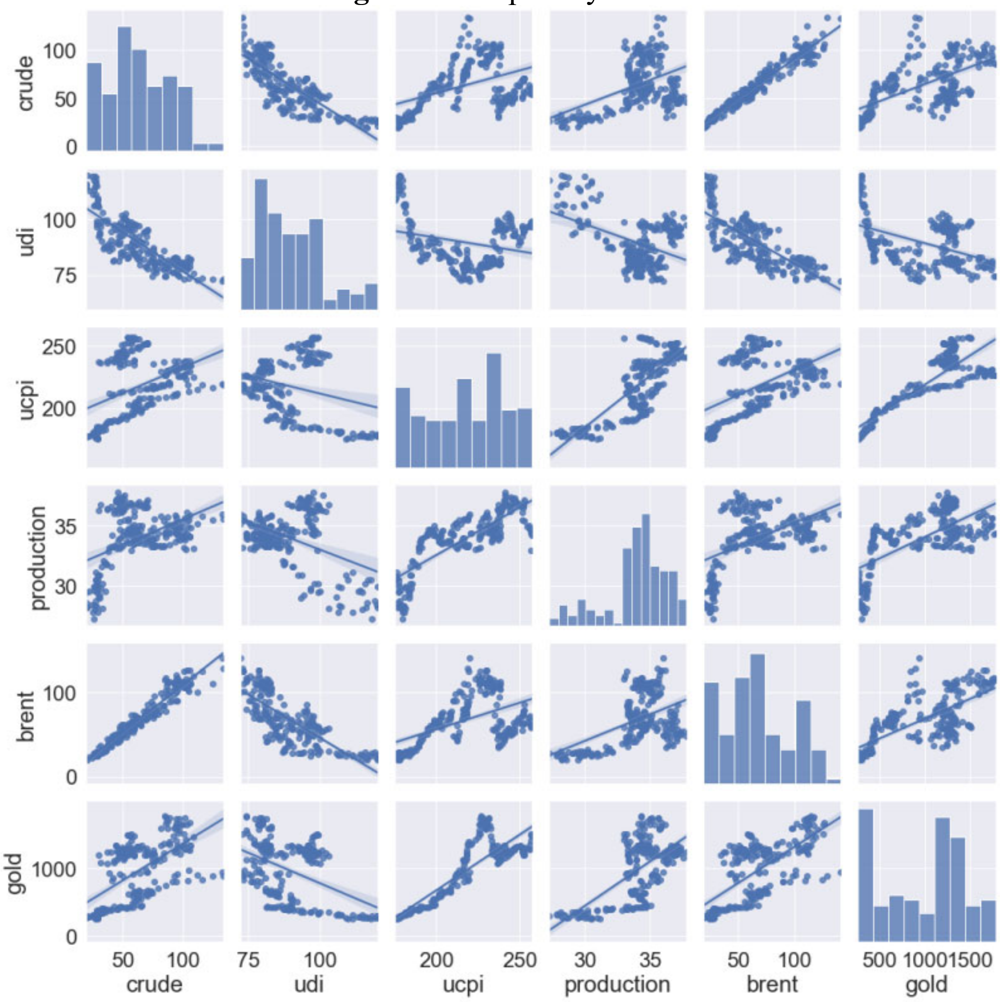

Fig. 3 Pairplot analysis

2.4 Models

To better present the benefits and advantages of ensemble learning, this paper chooses two basic models. One is the most basic model used in econometrics for forecasting, the Linear Regression model, also known as OLS. Another is one supervised learning algorithm, the Decision Tree Regressor, cause the ensemble learning models this paper used are based on tree operations (also known as the computation tree logic mostly). The Linear Regression model can be shown as follow:

$$crude_t = \alpha_0 + \alpha_1 udi_t + \alpha_2 production_t + \alpha_3 ucpi_t + \alpha_4 brent_t + \alpha_5 gold_t + \varepsilon_t \quad (2)$$

The Decision Tree Regressor (DTR) is based on the module sklearn.tree from package sklearn. The ensemble learning methods this research are going to talk about includes Random Forest Regressor (RFR) in Bagging algorithms and Gradient Boosting Regressor(GBR), eXtreme Gradient Boosting Regressor(XGBR) and Light Gradient Boosting Machine Regressor(LGBMR) in Boosting algorithms.

2.5 Metrics

This paper chose four main metrics to evaluate the models' performance. First is R-squared(***R²***), measures how well the data fit the regression model, or we say the goodness of fit.

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{\sum_{t=1}^{n}(\bar{y}_t - y_t)^2} \quad (3)$$

The second is Root Mean Squared Error (***RMSE***), which measures the sample standard deviation of the difference (called the residual) between the predicted and observed values. RMSE can indicate the degree of sample dispersion.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y_t})^2} \quad (4)$$

The third is Mean Absolute Error (***MAE***), which measures the average of absolute errors between predicted and observed values. The two have the similar function, but are not as sensitive to outliers. Considering both together can evaluate the model more comprehensively.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|(y_t - \hat{y_t})| \quad (5)$$

The last metric is Accuracy of Direction Prediction (***DA***), measures whether the model is successful in predicting the direction of price change from this period to next period. Hence, if the model correctly predicts the changing direction, this paper gives it the value **1**; else, this paper gives it **0**. Therefore, DA is the proportion of correct direction prediction.

$$DA = \sum_{t=0}^{n-1} A_t/n \quad (6)$$

Here,

$$(y_{t+1} - y_t)(y_{t+1} - y_t) > 0, A_t = 1; (y_{t+1} - y_t)(y_{t+1} - y_t) < 0, A_t = 0 \quad (7)$$

For all the formulae above, $\boldsymbol{y_t}$ is the actual value of the price of crude oil at time $\boldsymbol{t}$, $\boldsymbol{\hat{y}_t}$ is the predicted value, and $\boldsymbol{\bar{y}_t}$ is the mean value of crude oil price.

3. Results & Discussion

This section aims to present and analyze the results for basic models LR, DTR and ensemble learning models RFR, GBR, XGBR, LGBMR.

3.1 Basic Models Results

As for the LR, this paper just uses the LinearRegression Module in sklearn package. LR model doesn't have additional parameters to adjust. This work used DecisionTreeRegressor Module in sklearn package for constructing the DTR model. The best parameters for DTR are: max_depth = 9, min_samples_split = 4. The parameters which are not mentioned here just use the default value. The actual versus predicted crude oil prices using basic models are shown in Fig.4.

3.2 Bagging Algorithm Model Result

There are two main algorithms in ensemble learning: Bagging and Boosting. To entirely show the advantage of ensemble learning, this paper applied both algorithms. In the aspect of Bagging method, this work chose RandomForestRegressor module in sklearn.ensemble. The best parameters of RFR are:n_estimators=13, max_depth=6, criterion='mse'. The actual versus predicted crude oil price using Bagging algorithm model is shown in Fig. 4.

3.3 Boosting Algorithm Models Results

As for Boosting methods, this work chose GBR, XGBR and LGBMR. For GBR, this paper used GradientBoostingRegressor module in sklearn.ensemble, and the best parameters of GBR are : n_estimators=18, max_depth=4, min_samples_split=7, learning_rate=0.09. For XGBR, this work used XGBRegressor module in xgboost package, and the best parameters are: eta=0.15, n_estimators=43, max_depth=15, min_child_weight=12, gamma=0, reg_lambda=6. For LGBMR, this paper applied LGBMRegressor in lightgbm package, while the best parameters for LGBMR are: boosting_type='gbdt', learning_rate=0.08, n_estimators=40, max_depth:2, num_leaves=10, colsample_bytree=0.7, min_child_samples=19. The actual versus predicted crude oil prices using three Boosting algorithm models are shown in Fig. 4.

3.4 Evaluation & Results Review

To facilitate the selection of the best model, this paper aggregated all the results in one image and gathered all the evaluation metrics in Table. 3. Based on the results, the four ensemble learning methods obviously outperformed the two benchmark methods in both goodness of fitting and variation. In contrast to the evident variation between basic models' prediction and true price, the figure indicates that the overall price prediction results of the ensemble learning models are relatively excellent, which are basically consistent with the real value. Table. 3 includes R-squared for both training dataset and testing dataset, RMSE, MAE, DA for each model. The yellow area means that the model outperforms other models and the green area means the model is the second best. In the last four columns this research focuses on, the GBR model performs the best and after GBR is the RFR model. Comparing with the benchmark models, the ensemble learning models achieve improvements in almost every aspect, especially the GBR model, RFR model and LGBMR model, which also signifies both Bagging and Boosting algorithms can be used as an optimizer. The three models boost the goodness of fit and control the variance, and at the same time, their direction prediction accuracy rates are more than 75%, even reach 82.2%. The high prediction accuracy means the models can well realize the long-term trend prediction of crude oil price.

After entirely results review, this paper draws the conclusion that the ensemble learning models bring value. They heavily outperform the simple basic models and there's enough reasons to believe they are able to realize high quality forecasts.

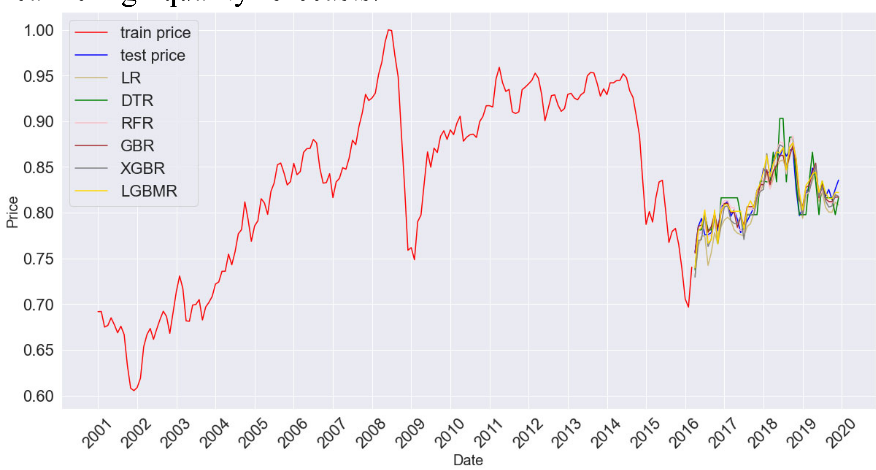

Fig. 4 Whole result visualization

Table 3. Metrics Summary

Metrics	R2_train	R2_test	RMSE	MAE	DA
lr	0.9803	0.7197	0.0155	0.0133	55.6%
dtr	0.9985	0.6903	0.0163	0.0130	66.7%
rfr	0.9929	0.8745	0.0104	0.0081	77.8%
gbr	0.9575	0.8798	0.0102	0.0077	82.2%
xgbr	0.9934	0.7922	0.0134	0.0101	57.8%
lgbmr	0.9708	0.8509	0.0113	0.0083	75.6%

3.5 Limitation

Although the ensemble learning methods got qualified long-term crude oil price prediction models in this paper, there's still some limitation can't be ignored. First is about the lack of latest datasets. It's well known that in the beginning of 2020, the crude oil price drop down dramatically because of the impact of COVID-19. Until now, the price goes up again even higher than before. Therefore, if the models in this paper want to fit the impact of the entire COVID-19 shock, the datasets that can cover the whole period would be a prerequisite. Unfortunately, this work had to use the data before COVID-19 emerge because the lack of latest data for some non-financial factors, which made it impossible to quantify the impact of COVID-19. Second shotroming is the project had dropped some factors that maybe important to further improve the model. It's shown before that the variables sp500 and inventory are dropped because of the low correlation coefficients. It's worth mentioning that the abandonment of variables doesn't mean that the factors they represent are unimportant. Hence, how to incorporate the effects of the stock market and crude oil inventory into the model is a problem needed to be addressed. The third one is datasets used are all monthly so the model is not sophisticated enough which means the model may perform poorly in short term. At present, the model is more suitable for long-term price forecasting.

4. Summary

In summary, this paper investigates crude oil price prediction based on multiple ensemble learning algorithms. It's unnecessary to reiterate the significant role of crude oil and its price forecasting. Considering accurate crude oil price prediction has prominent benefits for the economic conditions of a country, large quantities researchers made full use of various predictors and applied multiple algorithms to realize this mission. Different models may have different advantages. This work applied some popular ensemble learning algorithms that have achieved good results in other fields (e.g., XGB and LGBMR) and enriched the content in predicting crude oil price. The final models realized marked improvement and performed well in long-term trend prediction. To be specific, the goodness-of-fit of the two best models, RFR model and GBR model, reached 0.87, while the accuracy of direction prediction reached 77.8% and 82.2%, respectively. Nevertheless, this work is still a rudiment. In order to further improve and enrich the model, it's essential to address the limitation mentioned before. Future work should figure out the way to quantify short-term shocks such as COVID-19 and incorporate the effects of the stock market and crude oil inventory into the model. Overall, these results offer a guideline for the application of ensemble learning in crude oil price forecasting.

References

[1] Gupta V, Pandey A. Crude oil price prediction using LSTM networks. International Journal of Computer and Information Engineering, 2018, 12(3): 226-230.

[2] Barsky R B, Kilian L. Oil and the macroeconomy since the 1970s. Journal of Economic Perspectives, 2004, 18(4): 115-134.

[3] Mikhaylov A Y. Pricing in oil market and using probit model for analysis of stock market effects. 2018.

[4] An J. Oil price predictors: Machine learning approach. 670216917, 2019.

[5] Moshiri S, Foroutan F. Forecasting nonlinear crude oil futures prices. The energy journal, 2006, 27(4).

[6] Xie W, Yu L, Xu S, et al. A new method for crude oil price forecasting based on support vector machines。International conference on computational science. Springer, Berlin, Heidelberg, 2006: 444-451.

[7] Chen Y, He K, Tso G K F. Forecasting crude oil prices: a deep learning based model. Procedia computer science, 2017, 122: 300-307.

[8] Assaad R H, Fayek S. Predicting the price of crude oil and its fluctuations using computational econometrics: deep learning, LSTM, and convolutional neural networks. Econometric Research in Finance, 2021, 6(2): 119-137.

[9] Gabralla L A, Jammazi R, Abraham A. Oil price prediction using ensemble machine learning. 2013 International Conference on Computing, Electrical and Electronic Engineering (ICCEEE). IEEE, 2013: 674-679.

[10] Hacer Y A, Aykut E, Halil E, et al. Optimizing the monthly crude oil price forecasting accuracy via bagging ensemble models. Journal of Economics and International Finance, 2015, 7(5): 127-136.