Price Prediction of Cude oil, Gold, and Cotton Based on OLS, Random Forest, and Lightgbm

Yutong Chen *

Rutgers Business School, Rutgers University, New Brunswick, The United States

*Corresponding author: yc1124@rutgers.edu

Abstract. Contemporarily, the predicted price becomes a necessary step for investors in order to gain extra returns. This study applies three regression models (i.e., OLS, Random Forest, and Lightgbm) to forecast the future prices of three products (i.e., crude oil, gold, and cotton) The 0.3 rates is used to split the train and test set. Comparing each graph of errors, two obvious lines representing the original data and predicted data tell the direct result of the advantages and disadvantages of the three models. In the experiment, it is found that the Random Forest model takes part in an important role to predict these three prices. The Random Forest model has less error between the three metrics and displays a closer graph between the original data and predicted data. According to the results, Random Forest model predicts the futures prices accurately and other models can also play the same role for specific futures to some extent. These results shed light on guiding further exploration of future price prediction.

Keywords: Future Price prediction; Prediction model; Regression.

1. Introduction

Futures first appeared on the Grain Exchange in Chicago as a tool for people to agree on a specific price in the contract and trade at a specified time in the future [1-3]. The specific price are required to be lower than the seller's total assets, otherwise, a seller should pay the liquidated damages if not keeping a promise in the continuous time series [3]. The earliest types of futures were metal futures such as gold, followed by energy futures including crude oil, financial futures, etc. In the early stage of development, because of the low transaction cost and fast transaction speed of futures, futures occupied a short-term main position in the trading market. Additionally, futures prices can be used as a weather vane for the future economic market from a macro perspective [4]. From a micro perspective, the transaction cost of futures can measure the efficiency of the transaction market including buyers and sellers [5]. However, futures cannot exist as the main factor determining the long-term trend of market prices because it does not essentially have a decisive impact on long-term market volatility [6]. The leverage in futures trading using the ratio of the initial margin to leverage the return exponentially expands the risk rate of futures trading. French and Roll raised the explanation for the futures price volatility. The reason is that public information about the price is released widely in normal hours, private information is also released after trading, and the existing price difference in the trading [7]. Because the risk of futures trading requires high capital, investors must pay attention to the daily futures to predict the price of the transaction to avoid risks when conducting futures trading. According to the investigation of Louis Ederington and Jae Ha Lee about most traders, it's clear that all different type of traders had analyzed their own assets and the futures market before they start to invest in futures. But due to the underestimated or overestimated prices brought by changing futures, the traditional statistical method cannot accurately predict the futures price [2].

Until now, there are many studies predicting the futures price by using the existing prices to establish some models that have been completed by other scholars. Goss used the OLS model to test the prices of copper, tin, lead, and zinc and found that there existed a risk premium for the them with the passing of time [2]. Other scholar investigated the price of crude oil to test whether the spot prices will be cointegrated provably and then summarized that spot price can be predicted by posted price and futures price with a higher success rate [8]. Moreover, Zou uses three Commodity Exchange data: Shanghai, Zhengzhou, and Dalian to make a Random Forest prediction in which the error range is in

the appropriate range, so the Random Forest model plays an accurate role in the prediction [9]. As for Lightgbm, Sun et al. use the model to forecast the trend of novel cryptocurrency price and finds that it's advantage will be greater when the data is increasing [10].

Thus, this paper seeks to explore the practicality and application of three models including OLS, Random Forest, and Lightgbm in predicting the price of futures (crude oil, gold, and cotton), examining whether a new model can predict the price specifically and if so, whether there are some limitations should be adapted. To illustrate this research motivation clearly, this study uses the following framework to develop. Sec. 2 will be Data and Method. Sec. 3 will be Results & Discussion. Sec. 4 will be Limitations & Prospects. Eventually, a brief summary will be given in Sec. 5.

2. Data & Method

2.1 Data

All the data in this study comes from the website Investing.com. The main research objects are three typical futures, which are metal futures gold, energy futures crude oil, and crop futures cotton. Because the three futures belong to different production fields, such a mixed research object can be more represented as the overall futures trading research. On the website Investing.com, this study chooses the daily prices of the futures Barrick Gold Corp, Crude Oil WTI Futures, and US Cotton #2 Futures as the basis for data analysis. The whole-time range in the research is from Nov.15th 2021 to Nov.13rd 2022. As the research model, it uses the open price, high price on that day, and low price on that day as the input data and the closed price as the output data. In addition, in the train test split, the research uses the ratio of 7:3 to split the train set and test set.

2.2 Models

This discusses select three research models, i.e., OLS, Random Forest, and Lightgbm. OLS means ordinary least square, which applies to the linear regression model to interpret the relationship between one or multiple independent variables and one dependent variable. In the OLS model discussed in the study, the input parameters are used as the independent variables, and the output parameter is used as the dependent variables. So, such OLS is the multiple linear regression model. Random Forest means the model uses many decision trees to construct the regression model. This model splits the whole data into various data sets to combine as small nodes by using the criterion of decision trees to separate the data. In each divided small tree, also starts to divide the data into small trees. Until the model fails to divide, the Random Forest model uses the result as the decision in every decision tree to make appropriate predictions for the output data. Lightgbm model also uses the decision trees to construct, but it uses a more accurate decision tree to decrease the level of decision trees to avoid the predictions overlapping the existing data.

2.3 Metrics

This study uses four metrics to measure the accuracy of three models. The first metric is R-square which represents the probability that the predicted value y and the real value y are predicted to coincide, and the interval is between [0, 1]:

$$R^2 = 1 - \frac{RSS}{TSS} \tag{1}$$

Here, RSS is the sum of squares of residuals and TSS is the total sum of squares. The second metric is MSE, which represents the value of error in the model prediction:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{2}$$

Here, n is number of data points, Y_i is observed values, $\hat{Y}_i$ is predicted values. The less the value of MSE is, the more accurate the model is. The third metric is MAE, which is used to calculate the absolute error:

$$MAE = \frac{\sum_{i=1}^{n}|Y_i - \hat{Y}_i|}{n} \tag{3}$$

MAE is useful to measure the accuracy of the model in the time series. Also, the less the value of MAE is, the more accurate the model is. The fourth metric is RMSD which is to calculate the square root of residuals(errors):

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{N}} \quad (4)$$

RMSD is affected by the outliers easily and is always nonnegative which means the model is more accurate when RMSD is closer to zero.

3. Results & Discussion

3.1 Crude Oil

In the OLS model (seen from Table 1), x1, x2, and x3 correspond to open, high, and low prices respectively, and y corresponds to closed prices. The value of R-squared is 0.069 shows that this OLS presents a prediction coincidence probability of 0.069. The value that is very far from 1 illustrates that the prediction of the model cannot achieve the same as the original data. According to the coefficients of OLS statistics, the coefficient of variable x1 is negative, so it can be obtained that x1 generates a negative relationship to the dependent variable closed prices. In other words, for the increasing open prices, the closed prices show a decreasing trend. The coefficient of x2 and x3 are positive, which means that if the high and low prices show an upward trend, the dependent variable closed prices will still be affected by growth. For the results with p-values of three variables that are greater than 0.05 (confidence level), this proves that the probability that their coefficients are zero is not small. However, if the coefficient is zero, it means that there is no influencing relationship between the independent variable and the dependent variable, which is meaningless to construct the relationship between the given independent variables and dependent variables. Jarque -Bera statistics indicate whether the residual presents a normal distribution, and its probability is 2.96e-12, which is less than the confidence level of 0.05. Therefore, in OLS, the difference between the predicted closed price and the actual closed prices is normally distributed. The Durbin-Watson value is 0.512 when there are three independent variables and the degree of freedom is 75, which is much smaller than dL. This means that there is a positive autocorrelation in the residual relationship. It's a violation the of independence-error assumption.

Table 1. OLS Regression Results (Crude Oil)

R-squared	0.069	AIC:	896.9	x1(coef):	-71.806	x2(coef):	57.223	x3(coef):	24.483
Adj. R-squared	0.030	BIC:	903.8	t-statistic(x1):	-1.665	t-statistic(x2):	0.968	t-statistic(x3):	0.522
F-statistic:	1.786	Jarque-Bera(JB):	53.090	P>|t|(x1):	0.100	P>|t|(x2):	0.336	P>|t|(x3):	0.603
Prob (F-statistic):	0.157	Prob(JB):	2.96e-12	Std err(x1):	43.132	Std err(x2):	59.140	Std err(x3):	46.909
Durbin-Watson:	0.084	Log-Likelihood:	-445.44	MSE:	8437.477	RMSD:	91.856	MAE:	89.817

Moreover, MSE in the OLS model is 8437.477, RMSD is 91.856, and MAE is 89.817, which means there are lots of errors in the prediction. In the comparison between the predicted and the original data (as shown in Fig.1), it's easy to find that some outliers exist the clear difference between the predicted and the original such as the data when x = 4 and 37 approximately. Therefore, OLS is inappropriate as a regression analysis for the data of crude oil. In the Random Forest model (seen from Fig 2), its MSE is 2.684, which shows that its prediction loss is less in this model. RMSE is 1.638 and MAE is 1.257, which means that the three metrics to measure the error are very small. So Random Forest model fits the prediction successfully. The Lightgbm model (as illustrated in Fig.3) shows the error between the predicted value and the real value in Lightgbm. It illustrates the predicted value of Lightgbm almost not coincides with the real value in many outliers. Because MSE is 9020.757, MAE is 1.257, and RMSD is 94.978, three big value metrics make the error also big clearly. In the displayed feature importance (as depicted in Fig.4), column_0, column_1, and column_2

respectively represent open, high, and low prices. It is not hard to see that open price and low price have a greater impact on the closed price.

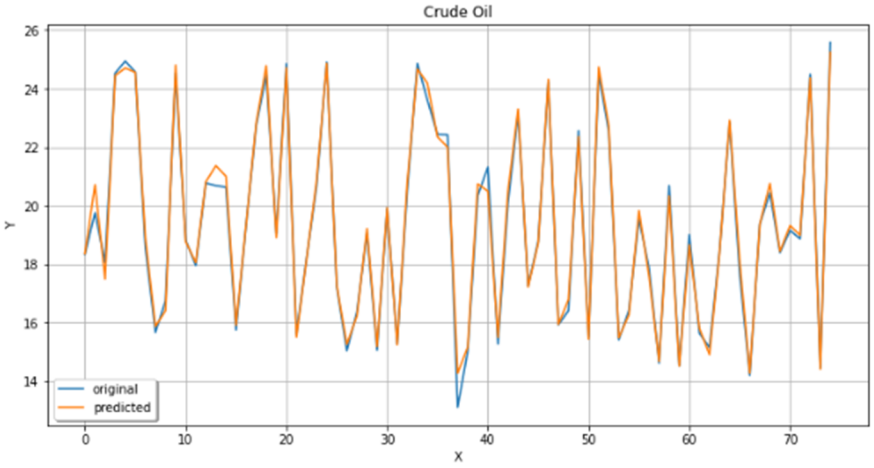


Fig. 1 Difference between Original and Predicted data (Crude Oil in OLS Model)

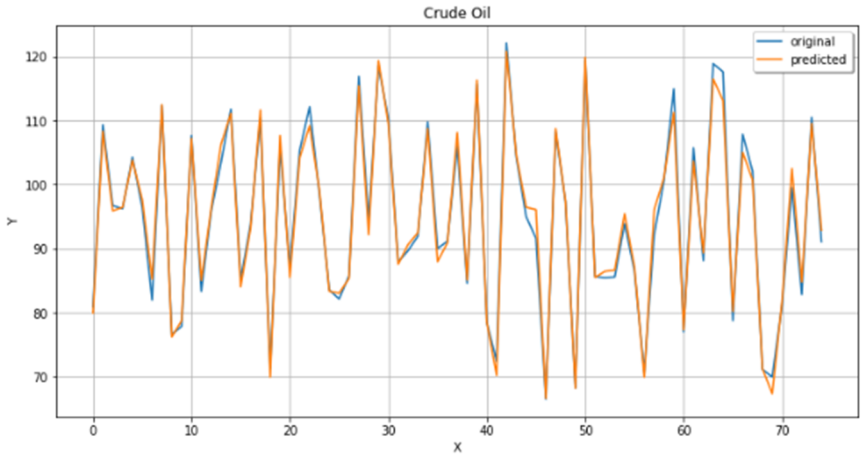


Fig. 2 Difference between Original and Predicted data (Crude Oil in Random Forest Model)

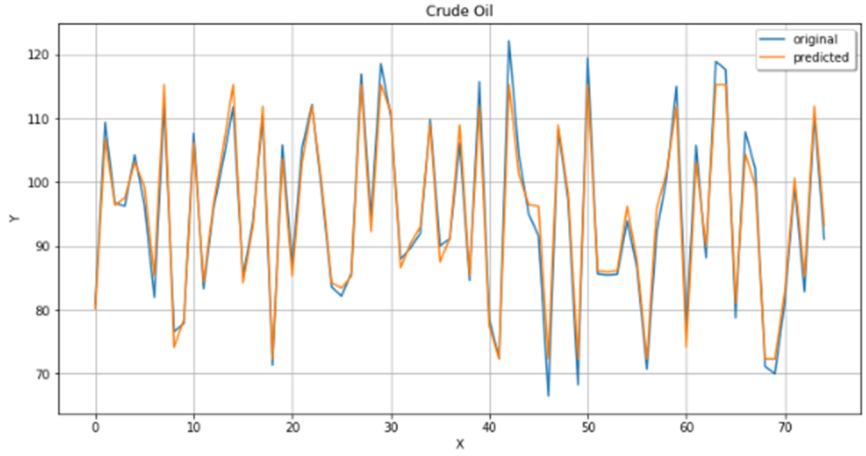


Fig. 3 Difference between Original and Predicted data (Crude Oil in Lightgbm Model)

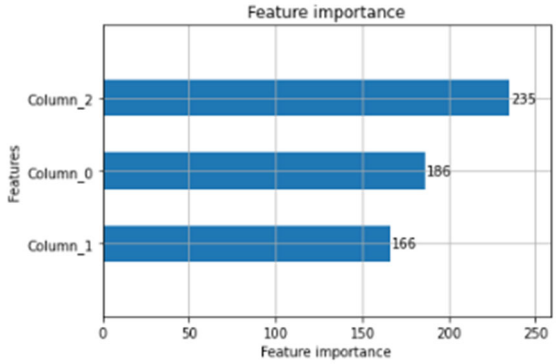


Fig. 4 Feature Importance (Crude Oil in Lightgbm Model)

3.2 Gold

In the OLS model (as shown in Table.2), x1, x2, x3, and y are the same as before. The R-squared shown in the table is 0.102, which means that OLS has an advantage in the coincidence rate of the predicted value and the actual value. According to the coefficients of OLS statistics, the coefficients of x1 and x3 are positive, which means that the impact of x1 and x3 on y is positively correlated. On the contrary, x2 is negatively correlated. The p-values of the three variables are greater than 0.05, which means the probability of the coefficient of the three variables being 0 is high. In this OLS model, the probability of Jarque-Bera statistics is 0673. The probability of close to 1 makes the residual of this model present a less normal distribution. However, the Durbin-Watson value of 0.065 exists in the condition of three variables, and the degree of freedom is 75, which is much smaller than Dl, which shows that it still has positive autocorrelation. The comparison (exhibited in Fig. 5) between the predicted and original values fits some extent.

Table 2. OLS Regression Results (Gold)

R-squared	0.102	AIC:	657.8	x1(coef):	5.123	x2(coef):	-12.010	x3(coef):	13.485
Adj. R-squared	0.065	BIC:	664.7	t-statistic(x1):	0.353	t-statistic(x2):	-0.408	t-statistic(x3):	0.465
F-statistic:	2.738	Jarque-Bera(JB):	0.793	P>|t|(x1):	0.725	P>|t|(x2):	0.684	P>|t|(x3):	0.643
Prob (F-statistic):	0.050	Prob(JB):	0.673	Std err(x1):	14.518	Std err(x2):	29.436	Std err(x3):	28.988
Durbin-Watson:	0.065	Log-Likelihood:	-325.88	MSE:	347.983	RMSD:	18.654	MAE:	18.405

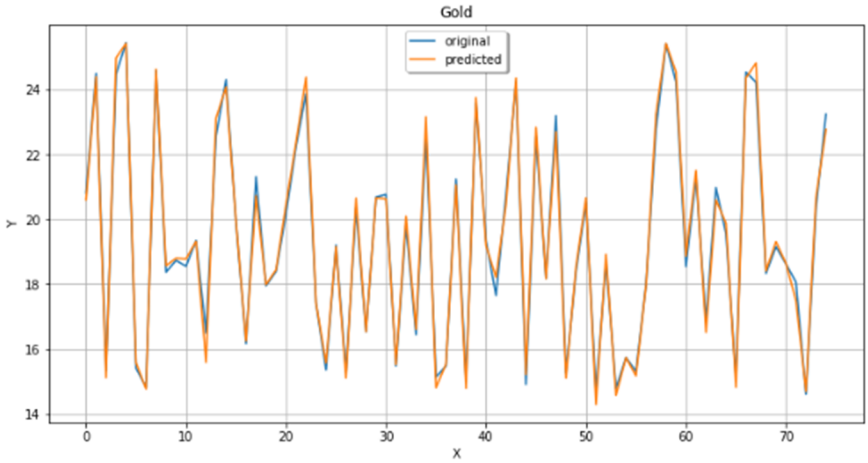


Fig. 5 Difference between Original and Predicted data (Gold in OLS Model)

Besides, its MSE is 347.983, RMSD is 18.654, and MAE is 18.405, which means there are also some errors here. In the Random Forest model (seen from Fig. 6), the MSE value is only 0.0916, RMSD is 0.303, and MAE is 0.236, which represents the prediction loss of this model is very small. In the comparison between the predicted value and the real value in Lightgbm (as shown in Fig.7),

when x = 5 and 57 respectively, there exists an obvious difference between the predicted and original. The three metrics are MSE = 8922.726, RMSD = 94.460, and MAE = 90.586. Moreover, the open price and low price (as shown in Fig.8) are more important to the closed price.

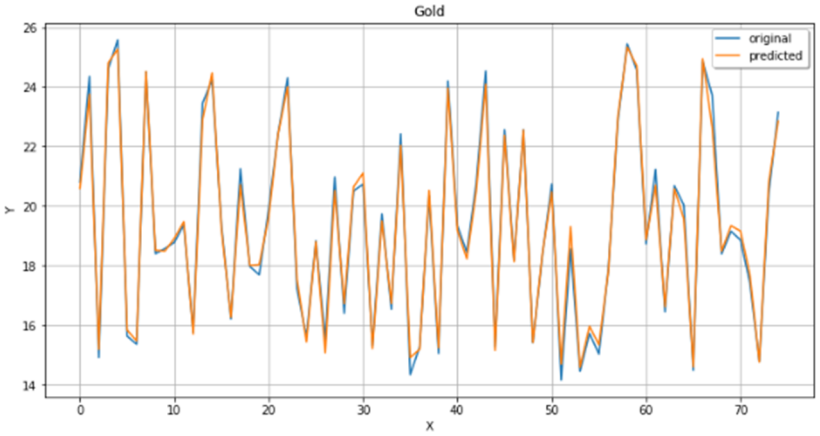


Fig. 6 Difference between Original and Predicted data (Gold in Random Forest Model)

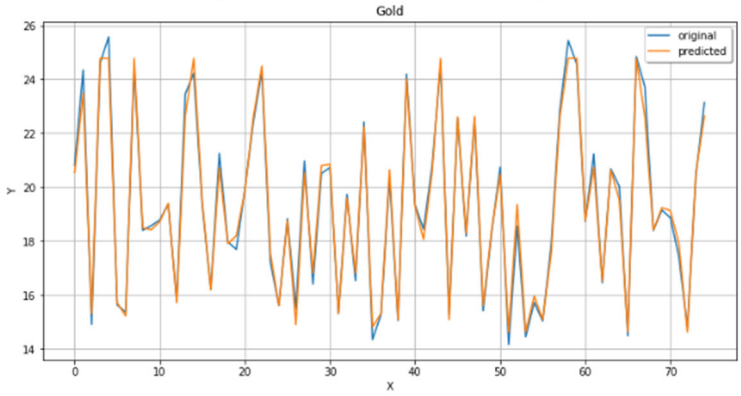


Fig. 7 Difference between Original and Predicted data (Gold in Lightgbm Model)

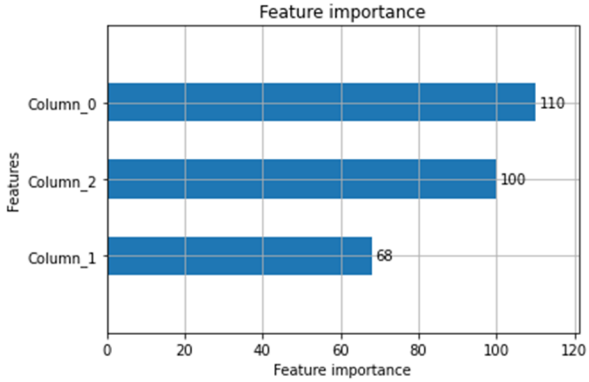


Fig. 8 Feature Importance (Gold in Lightgbm Model)

3.3 Cotton

In the OLS model (seen from Table.3), R-squared is 0.994, which means that the predicted value and the actual value are highly coincident. At the same time, the coefficients of three variables are both positive. Additionally, the p-value of three variables is only low (less than 0.05). Jarque-Bera statistics is 171.837, and its probability is 4.85e-38, which means that the residual presents a more standard normal distribution. In the Durbin-Watson test, when there exist the same three variables and 75 degrees of freedom, the statistic is 1.722 that greater than the proper Dl value, so there is not a positive autocorrelation. And metrics are MSE is 2.703, RMSE is 1.644, and MAE is 1.152. The smaller value of metrics means less errors in the prediction (as shown in Fig. 9). In the Random Forest model (as presented in Fig. 10), the MSE value is 4.355. The closer to 0 the MSE is, the smaller the

expected loss is. A larger MSE means the Random Forest model does not have an advantage in regression prediction. Also, RMSE is 2.087 and MAE is 1.435. From the figure shown, there are some different places in outliers between the orange line and the blue line. At the same time, in the Lightgbm model (as illustrated in Fig.11), it is obvious that the original value represented by the blue line and the predicted value represented by the orange line are largely non-overlapping. When x=8, the error of y has reached about 10 approximately. This means that Lightgbm has a large error in the prediction. According to the three metrics, MSE is 9.353, RMSE is 3.058, and MAE is 1.982. In Feature Importance analysis (as depicted in Fig.12), the variable high price has become a variable that has the greatest influence on closed prices.

Table 3. OLS Regression Results (Cotton)

R-squared	0.994	AIC:	299.2	x1(coef):	3.6773	x2(coef):	7.977	x3(coef):	7.948
Adj. R-squared	0.993	BIC:	308.6	t-statistic(x1):	2.524	t-statistic(x2):	3.243	t-statistic(x3):	2.930
F-statistic:	3820.0	Jarque-Bera(JB):	171.837	P>|t|(x1):	0.014	P>|t|(x2):	0.002	P>|t|(x3):	0.005
Prob (F-statistic):	2.94e-79	Prob(JB):	4.85e-38	Std err(x1):	1.457	Std err(x2):	2.459	Std err(x3):	2.713
Durbin-Watson:	1.722	Log-Likelihood:	-145.62	MSE:	2.703	RMSE:	1.644	MAE:	1.152

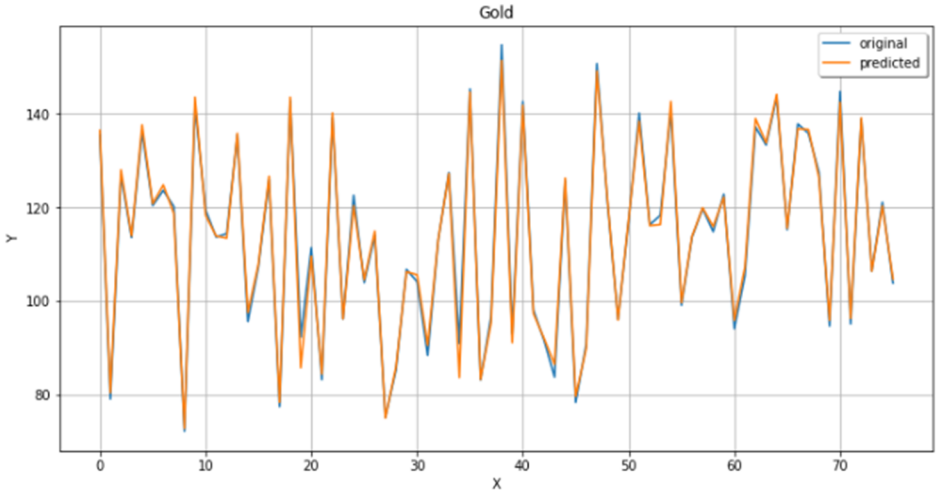


Fig. 9 Difference between Original and Predicted data (Cotton in OLS Model)

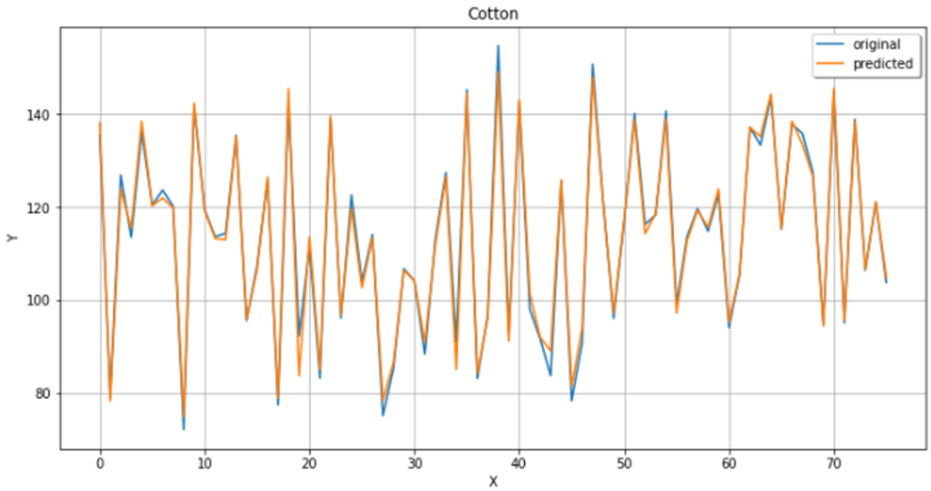


Fig. 10 Difference between Original and Predicted data (Cotton in Random Forest Model)

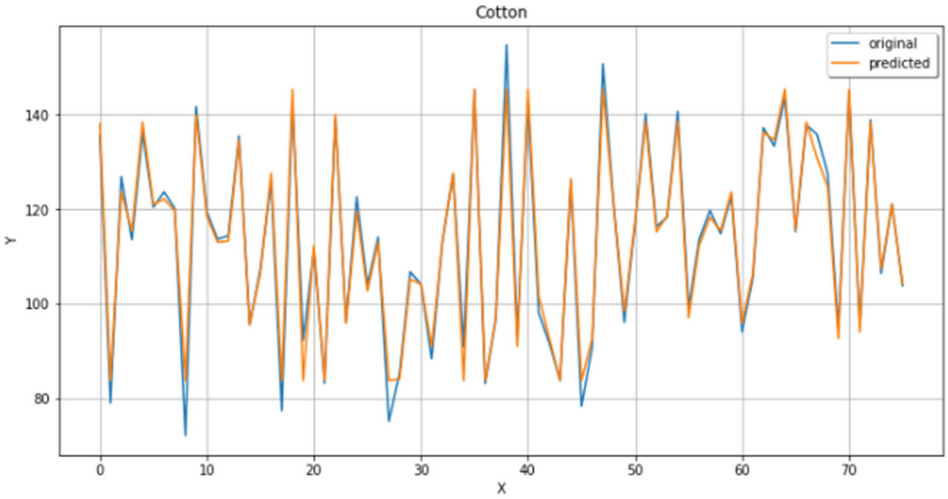


Fig. 11 Difference between Original and Predicted data (Cotton in Lightgbm Model)

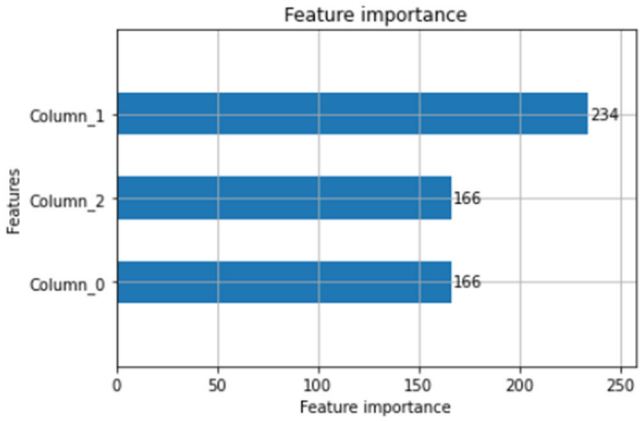


Fig. 12 Feature Importance (Cotton in Lightgbm Model)

3.4 Comparison & Suggestion

For the results of the three models of crude oil, OLS cannot be used as a suitable linear regression model. The reason is that the Durbin-Watson statistics are too small, resulting in the existence of positive autocorrelation. The Random Forest model, whose MSE is only 2.684, is a very good prediction model. Similarly, the error picture of the predicted value and the original value directly displayed in the Lightgbm model shows that the Lightgbm model cannot play a very important role in the prediction of crude oil. For the three model results of gold, OLS is due to the R-squared value being only 0.102, which means that it deviates greatly from the predicted value. Therefore, it cannot be used as a model with a high success rate. The MSE in the Random Forest model is only 0.0916, which is more accurate than the Random Forest in crude oil. It can become a successful prediction model. Meanwhile, in the error picture displayed by Lightgbm, the two separate lines in some given outliers indicate that Lightgbm has not successfully predicted the price of gold (e.g., Random Forest model). For Cotton, due to the larger R-squared value of 0.994 (closer to 1), OLS can be used as a predictive model with a large effect. Random Forest with an MSE of 4.355 belongs to a model with a relatively large error because OLS model only has 2.703 MSE. Similarly, in the error pictures displayed by Lightgbm, the predicted values seldom match the original values in extreme data. In summary, Lightgbm shows the weakest predictive ability in predicting given futures prices in this study, and Random Forest almost produces the smallest MSE, which is an appropriate forecasting model. In the meantime, OLS for the prediction of cotton can produce sufficiently overlapping error graphs, which can also be used as a suitable forecasting model. Moreover, among the three futures,

cotton, as the three models, can be predicted accurately enough, maybe because its change and fluctuation produce a sufficient forecasting rule.

4. Limitations & Prospects

Nevertheless, this study has some shortcomings and defects. Regarding to choosing the independent variable, it used open prices, high prices, and low prices. The method of choice is lack normal samples (e.g., average prices and median prices). High and low prices mean the influence of extreme data on closed prices. In addition, after splitting the train set and test set, the remaining data only has 76 data which is less representative of the whole study. For further study, it should add more representative data and independent variables. Besides, after constructing three models, it's necessary to explore the reason for the disparity of the models of different futures.

5. Conclusion

In summary, this paper investigates the prediction of futures prices including futures crude oil, gold, and cotton based on three models (i.e., OLS, Random Forest, and Lightgbm). Putting the given independent variables and dependent variables into three models to construct the predicted prices, this study compares the predicted prices and the original prices. According to the analysis, Random Forest model predicts the prices with less error. Moreover, OLS model sometimes can be the model with relatively less error. Besides, Lightgbm produces the clearer difference between the predicted and the original in some outliers. In the future, different impacts posed by different models on different futures prices should take more into consideration. Overall, these results offer a guideline for predictions of futures prices or other prices.

References

[1] Lingling Z, Renxiang W, Ying F. The comparative analysis of stock index futures trading effect on the trend of spot market[C]. 2010 2nd IEEE International Conference on Information and Financial Engineering. IEEE, 2010: 659-661.

[2] Chowdhury A R. Futures market efficiency: evidence from cointegration tests[J]. Journal of Futures markets, 1991, 11(5): 577-589.

[3] Sutherland I E. A futures market in computer time[J]. Communications of the ACM, 1968, 11(6): 449-451.

[4] Hong H, Yogo M. What does futures market interest tell us about the macroeconomy and asset prices?[J]. Journal of Financial Economics, 2012, 105(3): 473-490.

[5] Locke P R, Venkatesh P C. Futures market transaction costs[J]. The Journal of Futures Markets (1986-1998), 1997, 17(2): 229.

[6] Gülen S G. Efficiency in the crude oil futures market[J]. Journal of Energy Finance & Development, 1998, 3(1): 13-21.

[7] Harvey C R, Huang R D. Volatility in the foreign currency futures market[J]. The Review of Financial Studies, 1991, 4(3): 543-569.

[8] Ederington L, Lee J H. Who trades futures and how: Evidence from the heating oil futures market[J]. The Journal of Business, 2002, 75(2): 353-373.

[9] Zou Z B, Peng H, Luo L K. The application of random forest in finance. Applied Mechanics and Materials. Trans Tech Publications Ltd, 2015, 740: 947-951.

[10] Sun X, Liu M, Sima Z. A novel cryptocurrency price trend forecasting model based on LightGBM[J]. Finance Research Letters, 2020, 32: 101084.