Research on Commodity Futures Pricing Efficiency: A Machine Learning Perspective

Elaine Huang*

Economics, Shenzhen College of International Education, Shenzhen, Guangdong Province, 518000, China

*e-mail:s22533.huang@stu.scie.com.cn

Abstract. The huge size of China's commodity market plays an important role in global asset pricing. It is significant to deeply analyse whether machine learning can extract effective information from futures market data and how effective artificial intelligence algorithms are. Based on the monthly data of index contracts of 71 varieties in China's commodity futures market, this paper uses OLS and machine learning algorithms to extract information from the price data. Also, based on trading strategies based on predictions, this paper discusses the impact of the application of machine learning (technological progress) on the pricing efficiency of the futures market. The results show that the machine learning algorithm can improve the strategy performance on the whole, but with the development of the market, the improvement effect decreases at a significant level of 5%, that is, the market effectiveness is improving.

Keywords: Commodity futures; Pricing efficiency; Machine learning; Information mining

1 Introduction

Analyzing the rate of return and volatility of assets is a core issue in the financial field and is of great significance. With its rapid economic development, China's gross domestic product has increased more than tenfold in the past 20 years, making it the world's largest futures market. Strong industrial demand has increased along with

P. Dou and K. Zhang (eds.), *Proceedings of the 2024 2nd International Conference on Economic Management, Financial Innovation and Public Service (EMFIPS 2024)*, Advances in Economics, Business and Management Research 327,
https://doi.org/10.2991/978-94-6463-706-9_7

production, and China now accounts for 30% of global gold demand, 53% of coal demand, 60% of nickel demand, 59% of aluminum and 59% of iron ore. The "2020 China Futures Market Yearbook" report said that the fuel oil trading volume, China's asphalt, steel, corn and other futures contract markets are among the world's top five. Capital market allocates resources to different assets and enterprises effectively through pricing and trading. Commodity futures are closely related to its related industrial economic activities. At the same time, the volatility of commodity prices is large and there is leverage effect, which has a significant impact on the global capital market and economic development. Is the main reason for the unusual volatility from real demand of new emerging economies such as in China or speculation from hedge funds and futures index traders? What is the transmission role of the real economy and financial markets? The scholars' understanding of this is not uniform. The available literature is attributed either to speculation, to real demand, or to the transmission role of financial markets and the real economy.

The rapid development of computer technology, especially big data, artificial intelligence and large-scale cluster technology, has injected strong scientific and technological force into the development of finance. A large amount of literature shows that machine learning can improve the accuracy of futures price prediction. However, there is no systematic evaluation of whether machine learning can extract effective information from futures market data and the improvement effect of artificial intelligence algorithms. Therefore, the research significance of this paper is as follows: (1) academic perspective. Artificial intelligence algorithms can capture information in the market that is difficult for humans to have a sharper response than humans in dealing with market effectiveness. On the other hand, artificial intelligence algorithm trading has a positive effect on improving market efficiency, increasing liquidity, reducing market volatility, and timely detecting and correcting misprizing of securities. Through empirical analysis, this paper shows the impact of the application of machine learning in futures market on the efficiency of futures pricing. (2) Practical perspective. The commodity futures market of the domestic derivatives market has now developed into a complete and huge market with more than 70 trading varieties. The intelligent derivatives investment model based on machine learning model can be used in the practice of commodity futures investment trading, and its practical effect can be analyzed.

The main contributions of the research are as follows: (1) Based on all commodity futures, this paper uses traditional algorithms and machine learning algorithms to compare and analyze the market effectiveness of futures markets and improves the research on the impact of machine learning on the pricing efficiency of commodity futures markets. (2) This paper reveals that machine learning promotes market efficiency while improving information mining. (3) The quantitative model in this paper can provide basis for investors to make decisions.

The follow-up papers are arranged as follows: The second part is a literature review, which analyzes the relevant literature from various perspectives. The third part is theoretical analysis, from the perspective of macroeconomic law, supply and demand relationship and quantitative model analysis and research theory. The fourth part is empirical analysis. Finally, the fifth part is the conclusion.

2 Literature Review

This paper mainly involves three aspects: the fundamentals and economic laws related to futures, the application of machine learning in finance, and digital economy and financial technology. Based on the literature, this paper combines the characteristics of China's futures market and relevant theoretical analysis to put forward a research hypothesis.

(1) Futures related fundamentals and economic laws

Shamim Ahmed et al. (2016) used two risk factors to build a factor model to study the temporal predictability of excess returns of commodity futures. Meanwhile, Bakshi (2019) proves that a model containing average commodity factor, carry factor, and momentum factor can describe cross-sectional changes in commodity returns. Chinn (2014) studied the prediction of price changes and fairness in energy, agricultural, precious and base metal futures. The results showed that metals failed most unbiased tests and failed to predict subsequent price changes, but energy and agricultural products did much better. Chauvet et al. (1998) proposed a dynamic factor model with institutional switching as an empirical representation of business cycles. Hong (2012) argues that open interest may be more informative than futures prices when there is hedging demand and limited risk absorption capacity in futures markets. Inoue (2005) shows that neither traditional data mining nor parametric instability can reasonably explain the tendency that in-sample tests lead to unpredictability more often than out-of-sample tests.

(2) The application of machine learning in finance

Zhang Ning et al. (2022) use Transformer deep learning model to forecast the return rate of alternative assets to improve the accuracy of expected return rate. John Cotter et al. (2020) used a forecast combination of 28 potential predictor variables to find the out-of-sample predictability of the excess return of commodity futures. Overall, portfolio forecasts act as state variables in ICAPM, thus restoring the central role of macroeconomic risk in determining expected returns. Juan.D. Diaz et al. (2023) identified the accuracy of several machine learning models in predicting the gold risk premium using 186 predictors by performing out-of-sample assessments and portfolio measures. The results show that the combination of machine learning methods and forecasts has limited ability to outperform historical averages in predicting the gold risk premium, but slightly better results are obtained when the predictors are used alone. Xinjie Lu et al. (2022) comprehensively investigated the relationship between oil futures volatility and financial markets in a rich model environment. Baris Kocaarslan (2024) investigates the use of the CatBoost algorithm and the Shapley additive interpretation method to explore the link between the dollar and oil market uncertainty. The dollar, by contrast, was the most influential factor.

(3) Digital economy and financial technology

The impact of big data, artificial intelligence and other technologies on society in the era of digital economy is comprehensive, and the pricing of the capital market interacts with it. Du Li et al. (2022) calculated through digital finance that strengthening the risk constraints of commercial banks would affect their operational efficiency and improve them. Jiang Hai et al. (2023) judged that digital transformation could reduce the management costs and risks of banks and improve their operational efficiency through theoretical calculations and experiments. Sun Huijun et al. (2023) set up a government-led market resource allocation framework with the aim of balancing the interests of multiple entities as the main body and analyzed Internet technology and service mode of shared travel. Zhang Wei et al. (2023) compared the help brought by "traditional finance" and the newly developed "computational experimental financial engineering" tool for management decision-making through a new formal conceptual model. Tang Song et al. (2020), based on the data of A-share listed companies in Shanghai and Shenzhen from 2011 to 2017, deduced that the development of digital finance has A "structural" driving effect on enterprise technological innovation, and provided a complete plan.

In his 1970 book Efficient Capital Markets: A Review of Theoretical and Empirical Work, Fama refined the hypothesis of market efficiency and argued that markets should be divided into three situations, "weak form", "semi-strong form" and "strong form". These three market conditions respectively represent the efficiency of market prices to reflect potential factors1. The "weak form" is effective and cannot be arbitraged by building a model based on historical data. "Semi-strong" is effective and cannot be arbitraged through historical data, nor can it be arbitraged through the current public data building model. "Strong" efficiency, even using private data to build a model, can not achieve arbitrage (to make money).

Hypothesis 1: The application of machine learning (the development of technology) affects the measurement of market effectiveness.

Expected result 1: The machine learning algorithm can make money, indicating that the market "weak formula" efficiency does not hold. The discussion of market efficiency needs to take into account technological developments.

Hypothesis 2: Machine learning algorithms can improve the information mining ability of futures markets.

Expected Result 2: Machine learning algorithms make money while linear regression does not, indicating that the development of algorithms such as big data and artificial intelligence has an impact on the evaluation of market effectiveness. That is, machine learning algorithms can improve information mining capabilities, which poses challenges to market effectiveness.

Hypothesis 3: Even taking into account the application of machine learning (technology development), market efficiency is improving.

Expected Result 3: The earning power of machine learning algorithms is declining, indicating that the development of algorithms such as big data and artificial intelligence is improving market effectiveness.

3 Theoretical Analysis

3.1 Analysis of Economic Laws

(1) The price of aluminium, whether spot or futures, from the perspective of economic laws, is first determined by supply and demand: when supply is greater than demand, the price of aluminium decreases; When supply is less than demand, the price of

aluminium rises. For example, eggs provide nutrition for human beings, when the demand for food increases, the price of eggs rises, and vice versa.

(2) The specific demand comes from the downstream industrial chain, which needs to supply bauxite and electrolysis from the upstream from real estate development, machinery and equipment, transportation, etc.

(3) From the macro perspective, environmental protection policies such as carbon peaking and electricity price policies will affect output, while fiscal policies, import and export policies, interest rates, etc. will affect export and demand, and then have an impact on commodity prices.

(4) Behavioural game. Price and transaction data are used to construct indicators. This paper uses an average of monthly returns.

3.2 Description of Quantitative Model

(1) OLS regression

$$y_{it} = \alpha_0 + \sum_{k=1}^{n} \beta_k x_{kit} + u_{it} \tag{1}$$

All data in the sample were treated equally and mixed OLS regression was performed. Where y_{it} is the yield of futures variety i in t month; ${}_0$ is the intercept term, the constant; ${}_{kit}$ is the independent variable; ${}_{it}$ is the interference item.

Xgboost regression2

XGBoost (Extreme Gradient Boosting) is an optimized distributed gradient enhancement library designed to enable efficient, flexible and portable machine learning algorithms. It is based on the Gradient Boosting framework and solves many data science problems quickly and accurately through parallel tree lifting (also known as GBDT, GBM). The core principles of XGBoost include loss function optimization, regularization, Shrinkage, column sampling, parallel processing, etc.

4 Empirical analysis

(1) Variable description

The calculation and description of variables used in this paper are shown in Table 1.

Table 1. Variable definition and explanation

Variable	sign	Variable name	Definition	Theory
Explained variable	Re	The monthly yield of a futures index contract	$\mathrm{close}_{t+1}/\mathrm{close}_t - 1$	
Explanatory variable	Ma1	Average rate of return	MA1: Lag 1 period of the average of the last 5 periods of the futures yield	behavioural game
	Ma2	Average rate of return	MA2: Lag 2 period of the average of the last 5 periods of the futures yield	behavioural game
	Ma3	Average rate of return	MA3: Lag 3 period of the average of the last 5 periods of the futures yield	behavioural game
	Ma4	Average rate of return	MA4: Lag 4 period of the average of the last 5 periods of the futures yield	behavioural game
	Ma5	Average rate of return	MA5: Lag 5 period of the average of the last 5 periods of the futures yield	behavioural game

(2) Data description and descriptive statistics

The futures price data is derived from the Broad-based Quantisation platform database, which contains monthly index contract data for all active varieties in the commodity futures market from 2016 to 2024. After data processing, the study sample contains 5317 data. As shown in Table 2, the mean value of re is 0.0075, the minimum value is -0.3909, the maximum is 13.8752, and the standard deviation is 0.2022. From ma1 to ma5, the mean value increases from 0.0082 to 0.0096, and the standard deviation value also increases slightly. Nevertheless, the minimum value of -0.1246 and the maximum value of 2.8877 are constant.

Table 2. Descriptive statistics

	count	mean	std	min	25%	50%	75%	max
ma1	5317	0.0082	0.0909	-0.1246	-0.0098	0.0028	0.0192	2.8877
ma2	5317	0.0086	0.091	-0.1246	-0.0095	0.0031	0.0198	2.8877
ma3	5317	0.0091	0.091	-0.1246	-0.0091	0.0034	0.0204	2.8877
ma4	5317	0.0094	0.091	-0.1246	-0.0088	0.0037	0.0207	2.8877
ma5	5317	0.0096	0.0911	-0.1246	-0.0088	0.0038	0.0212	2.8877
re	5317	0.0075	0.2022	-0.3909	-0.029	0.0006	0.0365	13.8752

Note: All data range from 2009 to 2021, the unit of data in the table is %.

In Table 3, the mean value, standard deviation, minimum value of -1.0398 and maximum value of 14.0022 of the strategy yield constructed by the prediction of the XGBoost model are 0.4134, 1.7257, and 14.0022. Compared with OLS, XGBoost model increases by 0.3375 on average, 0.0117 on standard deviation, 0.5195 on minimum value and 0.4636 on maximum value. In summary, XGBoost is more efficient than OLS in all respects.

Table3. Comparing the performance of models

index	XGBoost	OLS	improve
count	73	71	
mean	0.4134	0.0759	0.3375
std	1.7257	1.714	0.0117

min	-1.0398	-1.5592	0.5195
max	14.0022	13.5386	0.4636

Table 4. .OLS linear model and machine learning algorithm XGBost performance

year	xgb_mean	ols_mean	did	meanDelta1	meanDelta2
2016	0.0166	-0.0185	0.0351	0.0106	0.004
2017	-0.0014	-0.006	0.0046	0.0106	0.004
2018	-0.0007	-0.0014	0.0007	0.0106	0.004
2019	0.0252	0.0231	0.0022	0.0106	0.004
2020	0.0005	-0.0034	0.0039	0.0106	0.004
2021	0.0073	0.0013	0.006	0.0106	0.004
2022	0.0002	0.0061	-0.0059	0.0106	0.004
2023	0.0057	-0.0065	0.0122	0.0106	0.004

As shown in Table 4, xgb_mean is listed as the average yield of all futures analyzed by the machine learning algorithm, which is 0.0166 in 2016 and 0.0057 in 2023. Similarly, ols_mean is listed as the average yield of all futures analyzed by the linear regression algorithm, which is -0.0185 in 2016 and -0.0065 in 2023. did is listed as the mean return from machine learning algorithm minus the mean return from linear regression algorithm, which is 0.0351 in 2016 and 0.0122 in 2023. meanDelta1 is the average value of did in the first four years from 2016 to 2019, which is 0.0106; meanDelta2 is the mean value of did in the next four years from 2020 to 2023, which is 0.004, and the difference between them is -0.0066. The sample difference of this difference is tested in this paper, and the statistical value is 1.96, and the p-value is 0.05. This shows that the machine learning algorithm can improve the performance of the strategy as a whole, but with the development of the market, this improvement effect decreases at a significant level of 5%, that is, the market effectiveness is improving.

(3) Robustness test

the paper analyzes the parameters of XGBoost algorithm, the period of variable consideration, and the sample time span of the two algorithms from multiple perspectives, and the conclusions are consistent.

5 Conclusions

Based on the monthly data of index contracts of 71 varieties in China's commodity futures market, this paper uses OLS and machine learning algorithms to extract information from the price data, constructs trading strategies based on predictions, and then discusses the impact of the application of machine learning (technological progress) on the pricing efficiency of the futures market. The results show that, first of all, the mean value, standard deviation, minimum value of -1.0398 and maximum value of 14.0022 of the strategy yield constructed by the prediction of the XGBoost model are 0.4134, 1.7257, and 14.0022. In summary, XGBoost is more efficient than OLS in all respects.

Secondly, this paper discusses the difference between machine learning algorithm XGBoost and OLS algorithm on the average yield of all futures varieties. meanDelta1 is the average of did in the first four years from 2016 to 2019, which is 0.0106; meanDelta2 is the mean value of did in the next four years from 2020 to 2023, which is 0.004, and the difference between them is -0.0066. The sample difference of this difference is tested in this paper, and the statistical value is 1.96, and the p-value is 0.05. This shows that the machine learning algorithm can improve the performance of the strategy as a whole, but with the development of the market, this improvement effect decreases at a significant level of 5%, that is, the market effectiveness is improving.

The results of this paper show that investors and regulators can use machine learning algorithms to effectively track futures market pricing. The discussion of market efficiency should take into account the technological advances represented by big data and artificial intelligence.

References

1. Ruobing Bai. 2019. Quantitative trading of commodity futures based on the analysis of market capital flow. Modern commerce 14, 2.

2. Mingtan Lai. 2019. Discussion and analysis of copper option volatility. China securities and futures 2, 5.
3. Bin Liu, Xinyue Shao, Yueyang Li. Fundamental quantitative investment research driven by machine learning [J]. Chinese industrial economy,201908:61-79.DOI:10.19581/j.cnki.ciejournal.2019.08.004.
4. Hannan Liu, Yuping Lan2020. The application of intertemporal arbitrage method in commodity futures trading [J]. Finance, 2020, 104: 267-277.
5. Fei Lu. 2018. Analysis of the influence of commodity attributes on the layout of futures delivery warehouse. Modern commerce 14, 2.
6. Xuyang Lu. 2020. Optimization of commodity futures arbitrage strategy based on harmony search algorithm. China's collective economy 35, 3.
7. Qiuhui Xu. 2005. On price risk and control of farmers' production and management. Economics and management, 191, 4.
8. Xiaojian Yu, & Qianqian Zou. 2018. Procedure-based pair trading strategy in commodity futures market. Southern Finance 3, 9.
9. Xianbing Zeng, Can Yin, Bo Zhang. 2017. Research on the possibility of arbitrage in commodity futures market based on high-frequency data. Gansu finance 2, 6.
10. ZuoBing Zhao, & Shangxiang Yang. 1994. Analysis of commodity futures price composition. Contemporary finance and economics 6, 48-51.
11. Jianing Zhang. 2018. Asset pricing, behavioral finance. F832 [A] 1002—6487201821—0160~03
12. Ning Zhang, Shaobin Yan, Dan Fan. Yield prediction and portfolio model based on deep learning. Statistics and Decision-making Journal 23, 2022 - Total Journal 611. DOI: 10.13546/j.cnki.tjyjc.2022.23.009
13. Baris Kocaarslan,US dollar and oil market uncertainty: New evidence from explainable machine learning,Finance Research Letters,Volume 64,2024,105375, ISSN 1544-6123, https://doi.org/10.1016/j.frl.2024.105375.
14. Benmenzer, G. , Gobet, E. , & Jrusalem, C. . 2007. Arbitrage free cointegrated models in gas and oil future markets. arXiv.org.
15. Cotter J, Eyiah-Donkor E ,Potì, Valerio.Commodity Futures Return Predictability and Intertemporal Asset Pricing[J]. Social Science Electronic Publishing [2024-07-23].DOI:10.2139/ssrn.3710435.
16. Juan D. Diaz, Erwin Hansen, Gabriel Cabrera. Gold risk premium estimation with machine learning methods, Journal of Commodity Markets, Volume 31,2023,100293, ISSN 2405-8513 https://doi.org/10.1016/j.jcomm.2022.100293

17. Kuhn Guo, Yichong Li, Yunhan Zhang, Qiang Ji, Wanli Zhan. How are climate risk shocks connected to agricultural markets? Journal of Commodity Markets, Volume 32,2023,100367, ISSN2405-8513 https://doi.org/10.1016/j.jcomm.2023.100367
18. Kanamura, T., Rachev, S. T., & Fabozzi, F. J.. 2010. A profit model for spread trading with an application to energy futures. Journal of Trading, 5(1), 48-62.
19. Michael T. Chng. 2009. Economic linkages across commodity futures: hedging and trading implications. JournChngal of Banking & Finance, 33(5), 958-970.
20. Xinjie Lu, Feng Ma, Jin Xu, Zehui Zhang. Oil future volatility predictability: New evidence based on machine learning models, International review of Financial Analysis, 83, 2022, 102299, issn 1057-5219 https://doi.org/10.1016/j.irfa.2022.102299