Predictable Changes of Energy Stocks in High-Frequency Markets: The Example of the Period of the Russian-Ukrainian Conflict and COVID-19 in China

Course: "Machine Learning in Finance" **Instructor:** Wenyu Zhou **Authors:** Group 6

Name of members	Student ID	Name of members	Student ID
Yunxuan Lu	22371188	**Gloria Timone**	22371327
Yin He	22371221	**Mary Kilonzo**	22371315
Boyang Zhang	22371169	**Martina Martinova**	22371324
Alvaro Paredes	22371326	**Mohamed Abdelsamie**	22371320

Table of Contents

1. High-Frequency Trading Background

(1) Idealized Financial Hypotheses and Mathematical Models

The Efficient Market Hypothesis suggests that stock prices follow a random walk, implying no predictability in stock prices. However, a crucial assumption of this hypothesis is that "all market participants are perfectly rational," meaning they can make the most advantageous investment decisions at all times, which is clearly impractical, especially in markets like China with numerous retail investors.

In financial mathematics, the random walk model assumed to describe stock price changes is a simplified mathematical model. In reality, market trading conditions are more complex than what fixed mathematical formulas describe. Therefore, from a theoretical perspective, the assumptions of relevant stochastic theories are not entirely valid, indicating that there is some predictability in the stock market.

(2) High Returns in Investment Practice

In recent years, scholars have extensively researched stock market predictability at both low and high frequencies. In high-frequency studies, utilizing past trading and quote data alone has demonstrated an out-of-sample R^2 of 10.5% for future 5-second stock returns and a 64% accuracy in predicting trade direction (Ait-Sahalia et al., 2021). Low-frequency research used neural networks to predict the annualized out-of-sample Sharpe ratio for a monthly S&P 500 investment portfolio strategy, achieving a ratio of 0.77. This outperformed the 0.51 Sharpe ratio of a buy-and-hold strategy, resulting in notable investment returns (Gu et al., 2019).

2. Data Processing

2.1. Data Integration

(1) Focus on Millisecond-Level Executed Orders

The prediction task focuses explicitly on millisecond-level high-frequency order information traded for 'Shenzhen Energy (000027).' The analysis employs data from executed orders to gauge time series changes in stock returns. Within the tick data, each formally traded order (TYPE=T) documents specifics, including Execution Time, Execution Price, Execution Quantity, and Buy/Sell IDs.

(2) Supplementing Missing Information in Executed Orders

Critical information regarding the Buy/Sell Direction is missing from the traded information. The goal is to obtain the buying and selling information for these orders to uncover potentially advantageous trading insights in the market.

To accomplish this, the matching process involves pairing the "Order Number" with the corresponding "Buy/Sell IDs" (for buy orders, the order number aligns with the buyer's ID, while for sell orders, it corresponds to the seller's ID). Emphasis is placed on the earlier entry information (TYPE=E) associated with these traded orders in the tick data. Extracting the "Buy/Sell Direction" field from the order information, the transformation is applied, converting "Buy" to "+1.0" and "Sell" to "-1.0" as outlined by Lee and Ready (1991).

In the stock market, both limit orders and market orders coexist. The quote information of limit orders is recorded in the Snapshot Data, while the price of market orders necessitates reference to the best buying and selling prices in the market. Obtaining the nearest best bid and ask quote information for each traded order relative to its execution time is crucial for a better understanding of the dynamic trading process in the stock market. To achieve this, the extraction process involves obtaining the nearest highest bid price, lowest ask price, and their corresponding quantities (A1, B1, AQ1, BQ1) from the Snapshot Data.

At this stage, the gathered information includes "Execution Time" "Execution Price" "Execution Quantity" "Buy/Sell Direction" "Highest Bid Price" "Quantity corresponding to the Highest Bid Price" "Lowest Ask Price" and "Quantity corresponding to the Lowest Ask Price" for all traded orders. To integrate the information on the best bid & ask prices, a new value is computed as follows(Ait-Sahalia et al., 2021):

$$P_t = \frac{A1 + B1}{2}. \tag{1.1}$$

The renamed results are listed in Table 1.

Table 1: Data fields used for Calculating Labels and Factors

Abbrev.	Meaning	Abbrev.	Meaning
Time	Order Trigger Time	**Pb**	Lowest Ask Price
TP	Execution Price	**Pa**	Highest Bid Price
Vt	Executed Quantity	**Sb**	Quantity corresponding to Lowest Ask Price
Dir	Buy/Sell Direction (1 for Buy, -1 for Sell)	**Sa**	Quantity corresponding to Highest Bid Price
		Pt	Simple average of Pa and Pb

2.2. Prediction Labels

(1) Predicting Returns Instead of Prices

Clearly, for technical analysis, stock prices are not suitable as direct objects for predictive exploration due to their significant trendiness, while stock returns are more appropriate.

Simultaneously, as illustrated in Figure 1, the notorious heavy-tailed nature of stock prices poses a challenge for many predictive models, whereas stock returns are closer to a normal distribution, aligning more closely with the general assumptions of the model regarding the sample's sampling distribution.

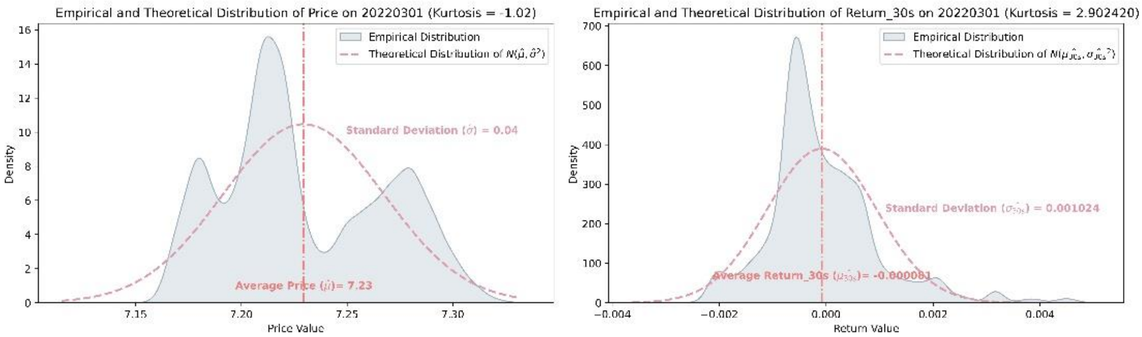

Figure 1 **(a)** Distribution of Stock Price (Not Stationary) **(b)** Distribution of Stock Return (Stationary)

(2) Prediction Labels: Average Returns within a Forward Window

For each execution moment T, the future stock return within a forward time window Δ is calculated as follows (Ait-Sahalia et al., 2021):

$$\mathrm{Re}turn(T,T+\Delta) = \frac{Average\ TP\ in\ (T,T+\Delta]}{Pt\ at\ T} - 1. \qquad (1.2)$$

Calculating average returns within a forward window, as opposed to single trades or fixed time points, reduces response variable volatility, enhancing stability in predictions. Studying aggregated trading behavior over a short span provides a more comprehensive reflection of participants' actions than focusing on a specific future time point. Predicting trade returns is more relevant for market makers, given the uncertainty in the execution time frame. Using average trade returns within a forward window better adapts to the current state of market activity and order book conditions.(Ait-Sahalia et al., 2021).

Two types of return horizons are considered, setting Δ to 5 seconds and 30 seconds. After completing the model predictions, an exploration of predictability duration involves comparing the predictive accuracy differences between the future 5-second returns and the future 30-second returns.

2.3. Predictive Factors

Ait-Sahalia et al. (2021) constructed 10 factors across three categories.

Table 2: Predictors Used

Abbrev.	Factor Meaning
1. Stock Trading Intensity	
1.1. Breadth	Number of trades within (T-Δ,T]
1.2. Immediacy	Average duration between adjacent executed orders within (T-Δ,T]
1.3. VolumeAll	Total number of shares traded within (T-Δ,T]
1.4. VolumeAvg	Average number of shares traded per transaction within (T-Δ,T]
1.5. VolumeMax	Maximum shares traded in a single transaction within (T-Δ,T]
2. Trading Asymmetry in Short Time Frames	
2.1. LobImbalance	Average imbalance indicator of LOB depth within (T-Δ,T]
2.2. TxnImbalance	Asymmetry in traded volumes within (T-Δ,T]
2.3. PastReturn	Stock return within (T-Δ,T]
3. Inherent Speed and Costs in Stock Trading	
3.1. QuotedSpread	Average quoted nominal spread within (T-Δ,T]
3.2. EffectiveSpread	Weighted percentage effective spread within (T-Δ,T]

To investigate the impact of information timeliness on predictions, each factor is calculated using historical information within two lookahead windows: (T-5s, T] and (T-30s, T-5s]. Consequently, 20 predictive factors are obtained in the end.

2.4. Outlier and Missing Value Handling

Due to the Shenzhen Stock Exchange's trading hours (9:30-11:30 & 13:00-15:00), orders outside these times are excluded. Labels and factors are calculated within a 30-second window before and after each traded moment, so orders in the first 30s after the first order and the last 30s before the last order are removed. In the initial data analysis, no missing values were observed in label and factor calculations. For robustness, records with a significant number of missing fields will be directly removed in subsequent calculations for other dates. Individual missing values will be addressed using forward filling.

3. Model Prediction

3.1. Machine Learning Model Selection

Machine learning models, with their flexible structures, offer an advantage over econometric models with strict assumptions when predicting complex problems. In the exploration of various models for stock return prediction, the focus is on three: LASSO, Ridge regression, and Random Forest. While the first two are linear models with regularization terms, each exhibiting unique characteristics, Random Forest stands out as an algorithm capable of capturing non-linear relationships and complex interactions between factors.

Assuming the model i aims to predict the return $r_{i,t+\Delta}$ within the future Δ interval at time t, and the vector consisting of the 20 factors at time t is $X_{i,t}$. Then there is

$$E(r_{i,t+\Delta}) = g(X_{i,t}), \tag{2.1}$$

where $g(\cdot)$ represents the flexible intrinsic structure of machine learning. The predictive task involves using three models and 20 factors to forecast the future 5-s and 30-s returns of "Shenzhen Energy" at every traded moment at the millisecond level in the past.

3.2. Rolling Cross-Validation

During hyperparameter tuning (see Figure 2), the model is initially trained on a 5-day test set. Subsequently, MSE1 is computed using the first validation set for the next trading day, with similar validations over the next 6 days resulting in MSE2. The optimal hyperparameter combination minimizes MSE1 + MSE2. Post-selection, the model is retrained on the past 5 days' data closest to the test day, predicting target values for the next day (highlighted in red). After one day of test set prediction, hyperparameters are readjusted. The time ranges of the training, validation, and test sets are shifted forward by 1 day for closer alignment with the test set during subsequent training and tuning.

It's crucial to highlight our higher frequency of hyperparameter adjustment and models' generalization compared to the literature by Ait-Sahalia et al. (2021) and Gu et al. (2019).

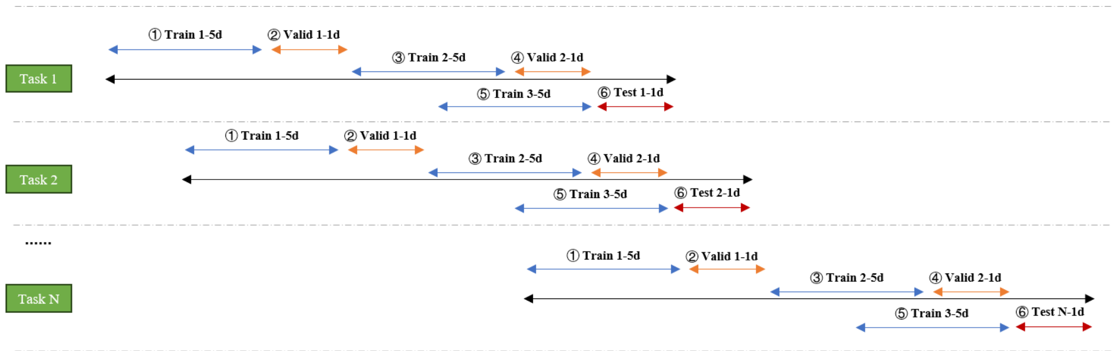

Figure 2: Rolling Prediction

Due to the large datasets and high computational resource requirements, the focus is on optimizing only one most crucial hyperparameter for each machine learning model as follows:

Table 3: Hyperparameter to Tune

Model	Hyperparameter to tune	Range of tuning
LASSO	regularization coefficient alpha	{0.01, 0.1, 1.0, 10}
Ridge	regularization coefficient alpha	{0.01, 0.1, 1.0, 10}
RF	max_depth (n_estimators=100)	{3, 4, 5, 6, 7}

A continuous period of 65 trading days in 2022.01-2022.05 is chosen as the test set. It is noteworthy that this test set interval encompasses two significant international events in China: the resurgence of the COVID-19 and the outbreak of the Russia-Ukraine conflict in 2022.02.

3.3. Prediction Accuracy

(1) Comparison of Out-of-Sample R^2 for Different Models

Enhanced R^2 is employed to assess the prediction accuracy proposed by Gu et al. (2019), which replaced the mean prediction benchmark in the denominator with zero prediction, a more scientifically sound approach in financial forecasting:

$$R_{oos}^2 = 1 - \frac{\sum_{(i,t)} (r_{i,t+\Delta} - \hat{r}_{i,t+\Delta})^2}{\sum_{(i,t)} r_{i,t+\Delta}^2}. \quad (2.2)$$

The out-of-sample R^2 of the three models across the entire test set is shown in Figure 3.

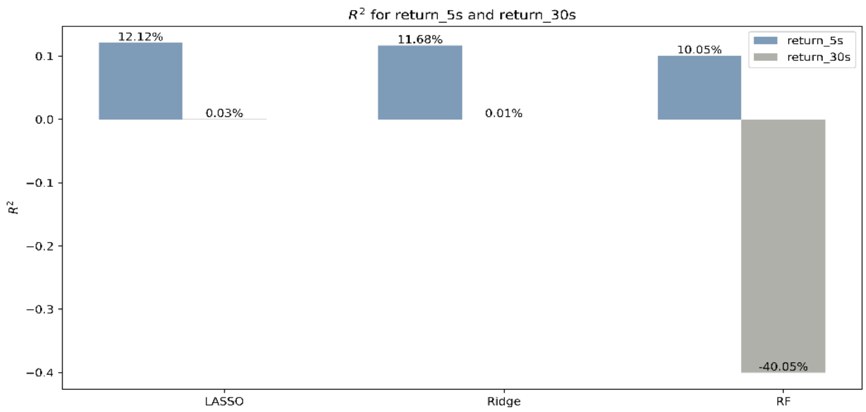

Figure 3: R^2 of Each Model for Return_5s & Return_30s

The results show that R2 follows the order: **LASSO > Ridge > RF**. This suggests that the sparsity and sensitivity to redundant features in LASSO may offer an advantage in high-frequency return prediction. LASSO's sparsity helps mitigate the risk of overfitting, adapting better to noise and redundant information in high-frequency data. Its exceptional performance indicates a more linear relationship between the studied factors and high-frequency returns. Conversely, Random Forest, especially for longer-term forecasts in this scenario, may encounter issues such as overfitting, presenting challenges for accurate predictions.

(2) Duration of High-Frequency Return Predictability

As depicted in Figure 3, predictability in future 5-second returns exceeds that in future 30-second returns, with a notably low R^2 for return_30s. The models encounter difficulty in capturing variance in real returns, highlighting short-term patterns. This implies that predictability for high-frequency returns in actual stock markets is transient. For technical investors, seizing profitable opportunities within these brief trading intervals is crucial. Moreover, the findings suggest that a quicker market reaction leads to higher prediction accuracy. The short-term market may lack complexity, limiting the efficacy of Random Forest.

(3) Impact of Extreme Real Returns on Predictions

The occurrence of return=-1, signifying no traded orders within the lookahead window, is plausible in real market. Figure 4 illustrates the models facing challenges in identifying such abnormal returns. In practical applications, high-frequency investors should exercise additional caution regarding these uncommon but possible scenarios that can impact investment outcomes.

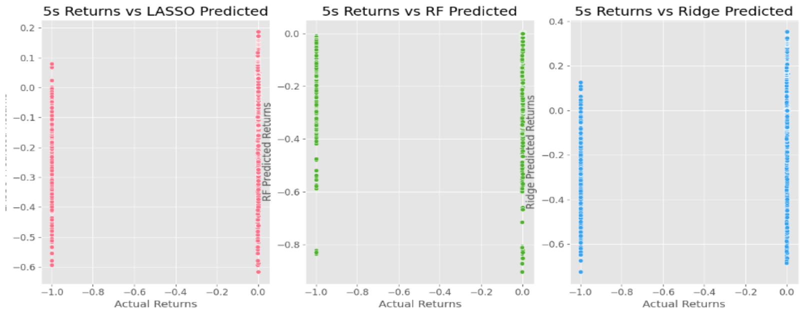

Figure 4: Prediction for Extreme Return

3.4. Temporal Heterogeneity

(1) Impact of the Russo-Ukrainian Conflict and COVID-19 on the Stock Market

Volatility of energy price is a complex phenomenon driven by the interplay of supply and demand dynamics, along with factors like seasons, geopolitical tensions and so on.[1]

The daily R2 variations for the test set were computed, as shown in Figure 5. It is observed that towards the end of 2022.02, the predictability of returns suddenly declined significantly. R^2 dropped below 0 for all models, with R^2 for predicting return_5s reaching close to -600,000%. This could indicate that the models' assumptions were particularly misaligned with market behavior. Lastly, the poor R^2 of RF in predicting return_30s also exposes its instability.

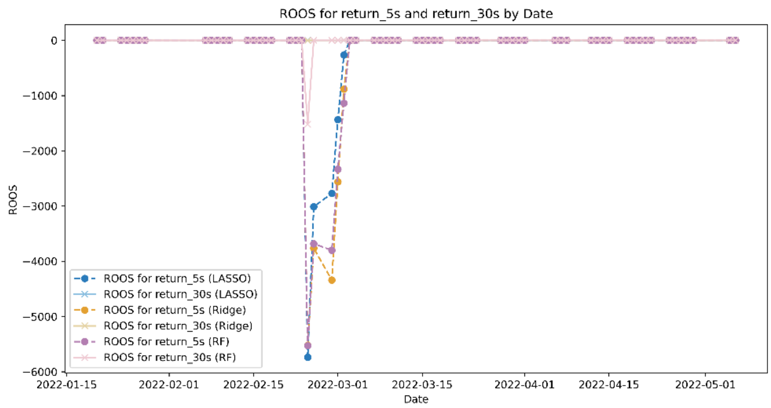

Figure 5: Variation of R^2 by Date

[1] Refer to U.S. Energy Information Adiministration

During this period, coinciding with the onset of the Russia-Ukraine conflict, the study's focus, "Shenzhen Energy," being an energy stock, felt the impact of Russia's significant role as a global energy supplier (with 15% of global exports and 10% of reserves in 2020[2]). The conflict led to a global energy supply shortage, resulting in a substantial surge in global oil prices. For instance, Brent crude prices surged by 56% between January 3 and March 7, 2023, maintaining significantly higher levels than before the conflict. Consequently, China's energy market was affected, causing a notable increase in energy stock volatility during this period and a sharp decline in the predictability of high-frequency returns.

Additionally, this period coincided with a critical juncture in the resurgence of the COVID-19 pandemic in China. The pandemic could lead to a slowdown in economic activities, disruptions in the production chain, constrained capacities, and a decrease in demand. Simultaneously, it could directly impact overall market sentiment, ultimately influencing stock market volatility. In March 2022, China encountered economic challenges as key indicators like net exports and consumption slowed down, as illustrated in Figure 6.

Figure 6: Sudden Increase of Macroeconomic index when COVID-19[3]

[2] Data source: The systemic impacts of the war in Ukraine: a triple shock

[3] Data source: China National Bureau of Statistics

(2) Predictability during the Near Opening and Closing Periods

As widely recognized, near the opening and closing periods exhibit lower market liquidity and wider bid-ask spreads due to fewer investors. These times are also vulnerable to the impact of overnight news or shifts in global energy markets. To validate, the test set is divided into two segments: the half-hour near the opening and closing (9:30-10:00 and 14:30-15:00) and other intervals. Figure 7 illustrates the R^2 performance of the three models in predicting high-frequency returns during these periods.

Results in Figure 7 confirm the hypothesis: predictability of high-frequency returns during the near opening and closing half-hour is significantly lower than during other intervals. Accurate prediction of high-frequency returns is particularly challenging in this period, as shown by the notably low predictability of return_30s. External factors such as the Russia-Ukraine conflict and COVID-19 in China directly influence this lower predictability.

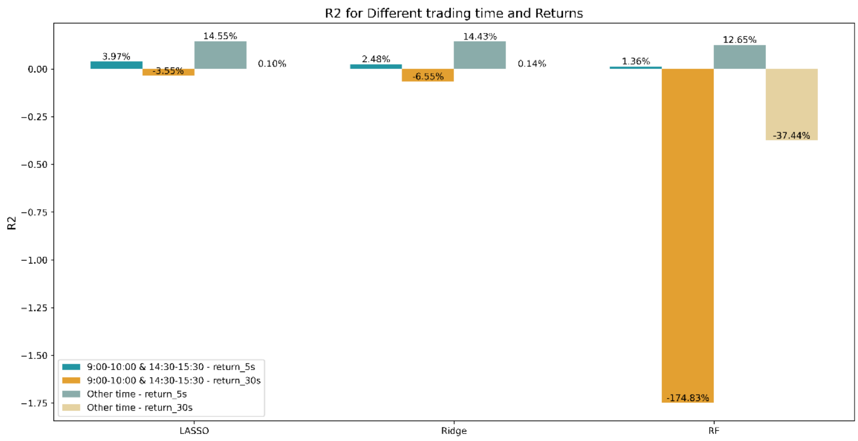

Figure 7: R^2 comparision for different trading time

4. Model Interpretation

4.1. Factor Importance

(1) Financial Explanation Behind Important Factors

While Random Forest demonstrates good accuracy in predicting return_5s, it lags behind linear models. To investigate this, the inherent feature importance method of tree models is used, calculating variations in importance metrics for the 20 factors in predicting each training model of the test set (see Figures 8). Notably, when predicting return_5s, Breadth, Immediacy, and Pastreturn emerge as the top three factors, with a stronger influence observed for order placements closer to the prediction time. These findings highlight the primary sources of predictability for high-frequency stock returns, shedding light on the less satisfactory performance of Random Forest in variable selection.

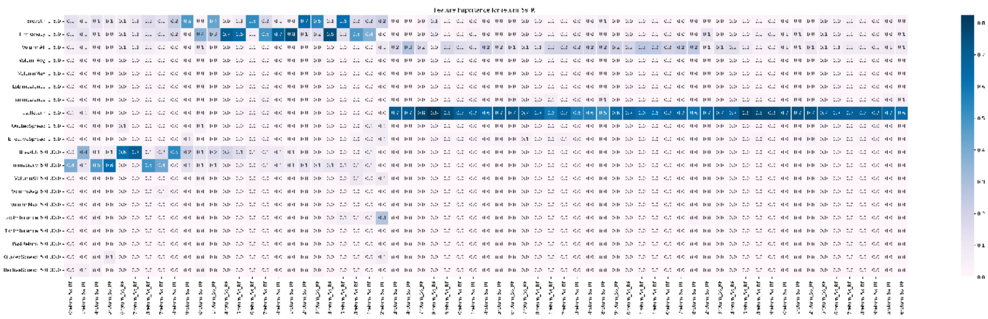

Figure 8: Importance of predictor when predicting return_5s

(2) Sudden Shifts in Factor Importance Rankings

Observing Figures 8, we notice on March 3, 2022, variable importance shifts coincide with the Russia-Ukraine conflict. Post this date, PastReturn dominates, indicating its robust predictive nature during market turmoil. This is crucial for investors navigating uncertainty.

(3) Singular Focus in Random Forest Variable Selection

Post-March 1, 2022, Random Forest leans heavily towards PastReturn, potentially hindering its adaptability to non-linear relationships. This bias contributes to weaker predictive performance compared to LASSO, raising concerns about overfitting. The heatmap decoding reveals a preference for simplicity in feature selection, resembling LASSO. Consistent feature importance suggests reliability, while variability may indicate sensitivity to specific data subsets or model configurations.

(4) Key Variables in Energy Stock Prediction

Understanding variables like Market Breadth, Immediacy, Trading Volumes, Past Return, LobImbalance, and Effective Spread is essential for interpreting and predicting stock performance in the energy sector. These variables play pivotal roles in reflecting sentiment, liquidity, and historical trends, capturing both micro-level trading mechanics and broader market trends in the dynamic energy sector.

4.2. Advice on Investment

Investing in Chinese energy stocks requires diversification due to market volatility. Market sentiment indicators should be approached cautiously, considering potential swift reversals. Liquidity, reflected in immediacy, offers advantages in responding to market changes, while trading volumes indicate interest but may trigger volatility. Historical performance, while tempting, should be viewed skeptically in the dynamic energy sector. A diversified portfolio and careful risk management are crucial for navigating market changes.

5. References

[1] Ait-Sahalia, Yacine, et al. "How and When Are High-Frequency Stock Returns Predictable?" *SSRN Electronic Journal*, 2022.

[2] Ding, Yi, et al. "Forecasting Realised Volatility: Does the LASSO Approach Outperform HAR?" *Journal of International Financial Markets, Institutions and Money*, vol. 74, Sept. 2021, p. 101386, https://doi.org/10.1016/j.intfin.2021.101386.

[3] Gu, Shihao, et al. "Empirical Asset Pricing via Machine Learning." *The Review of Financial Studies*, vol. 33, no. 5, 26 Feb. 2020, https://doi.org/10.1093/rfs/hhaa009.

[4] LEE, CHARLES M. C., and MARK J. READY. "Inferring Trade Direction from Intraday Data." *The Journal of Finance*, vol. 46, no. 2, June 1991, pp. 733–746, https://doi.org/10.1111/j.1540-6261.1991.tb02683.x.

[5] "The Systemic Impacts of the War in Ukraine: A Triple Shock." ODI: Think Change, odi.org/en/insights/the-systemic-impacts-of-the-war-in-ukraine-a-triple-shock/.

[6] "How War in Ukraine Is Reverberating across World's Regions." IMF, 15 Mar. 2022, meetings.imf.org/en/IMF/Home/Blogs/Articles/2022/03/15/blog-how-war-in-ukraine-is-reverberating-across-worlds-regions-031522. Accessed 6 Jan. 2024.

[7] Brian Michael Jenkins, Brian. "Consequences of the War in Ukraine: The Economic Fallout." The RAND Corporation, 7 Mar. 2023, www.rand.org/pubs/commentary/2023/03/consequences-of-the-war-in-ukraine-the-economic-fallout.html. Accessed 2 Jan. 2024.

[8] Harris, Larry. Trading and Exchanges : Market Microstructure for Practitioners. Oxford ; New York, Oxford University Press, 2003.

[9] Hasbrouck, Joel. Empirical Market Microstructure. Oxford University Press, 4 Jan. 2007.

[10] L, Gao, et al. "Market Intraday Momentum." Journal of Financial Economics, 2018, pp. 394–414.

[11] Li, Yi, et al. "Does Intraday Time-Series Momentum Exist in Chinese Stock Index Futures Market?" Finance Research Letters, Sept. 2019, https://doi.org/10.1016/j.frl.2019.09.007.

[12] M, Jin, et al. Intraday Time-Series Momentum: Evidence from China. 2020.

[13] Chu, Xiaojun, et al. "Intraday Momentum and Reversal in Chinese Stock Market." Finance Research Letters, vol. 30, Sept. 2019, pp. 83–88, https://doi.org/10.1016/j.frl.2019.04.002.

[14] Zheng, Zunxin, et al. "High-Speed Rail and Stock Return Comovement in China." Research in International Business and Finance, vol. 67, 1 Jan. 2024, pp. 102107–102107, https://doi.org/10.1016/j.ribaf.2023.102107. Accessed 6 Jan. 2024.

[15] Xingyue, et al. "Network Momentum across Asset Classes." ArXiv (Cornell University), 22 Aug. 2023, https://doi.org/10.48550/arxiv.2308.11294. Accessed 6 Jan. 2024.

[16] Wang, Xin, and Haofei Zhang. "The Cross-Predictability of Industry Returns in International Financial Markets." International Review of Finance.

[17] Huld, Arendse. "China's Economy in 2022: A Tumultuous Year Ending with an Optimistic Outlook." China Briefing, 22 Dec. 2022, www.china-briefing.com/news/chinas-economy-in-2022-a-tumultuous-year-ending-with-an-optimistic-outlook/.

[18] Xing, Xiaoyun, et al. "The Impact of the Russia–Ukraine Conflict on the Energy Subsector Stocks in China: A Network-Based Approach." Finance Research Letters, vol. 53, 1 Jan. 2023, pp. 103645–103645, https://doi.org/10.1016/j.frl.2023.103645.

[19] Davis, Marc. "The 8 Most Volatile Sectors." Investopedia, 13 July 2022, www.investopedia.com/financial-edge/0712/the-8-most-volatile-sectors.aspx.

[20] Uberti, David, and Bob Henderson. "Oil Prices and Energy Stocks Are Headed in Different Directions. What Gives?" Wall Street Journal, 6 Dec. 2022, www.wsj.com/articles/oil-prices-and-energy-stocks-are-headed-in-different-directions-what-gives-11670294602. Accessed 6 Jan. 2024.

[21] "With Energy Stocks, It's the Volatility That Kills You." Bloomberg.com, 16 Feb. 2022, www.bloomberg.com/opinion/articles/2022-02-16/with-energy-stocks-it-s-the-volatility-that-kills-you. Accessed 6 Jan. 2024.