Predictability Changes of Chinese Energy Stocks in High-Frequency Trading Markets

— A Case Study on the Period of **Russia-Ukrain Conflict** & **COVID-19**

------- **Members of Group 6** -------

| 卢云轩 | 何尹 | 张伯杨 | Gloria Timone | Alvaro Paredes |

| Mary Kilonzo | Martina Martinova | Mohamed Abdelsamie |

Agenda

1. **Introduction to High-Frequency Trading**
   - Discussion on the predictability of stock returns.
2. **Data Processing Procedure**
   - Our process for handling raw data;
   - Construction of labels and variables.
3. **Model Prediction**
   - Model selection and Cross-validation techniques;
   - Model comparison and Temporal heterogeneity analysis.
4. **Model Explanation**
   - Financial explanation behind important factors and Advice on investment;
   - Explaining model predictive performance differences through factor selection.

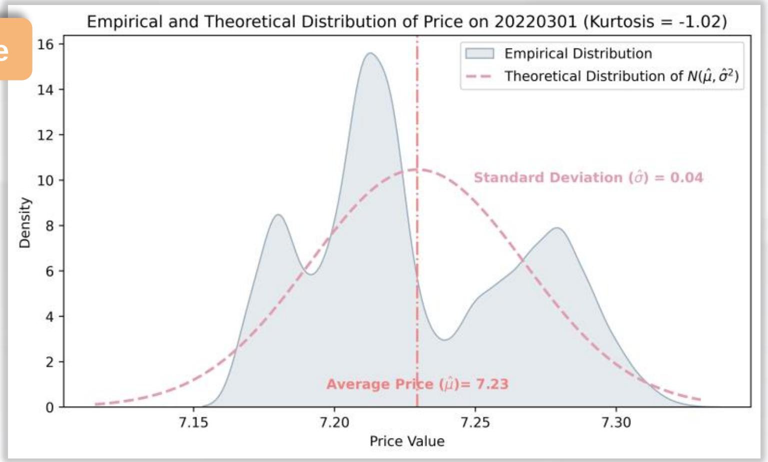
01 Background: Why can stock returns be predicted?

- Idealized financial hypotheses and mathematical models for stock price
  - Efficient Market Hypothesis : assume all investors in the stock market are perfectly rational;
  - Random walk model : simplify the complex trading markets.

* Both are not applicable in complex real stock markets

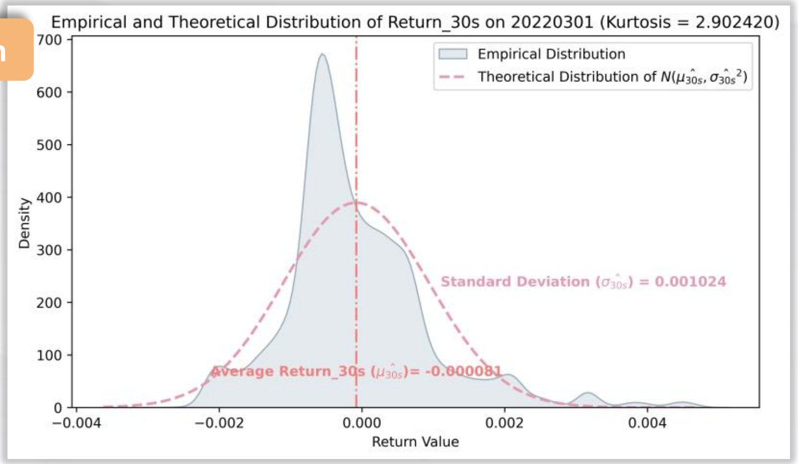
- The predictability of stock returns is stronger than that of stock prices
  - Stationarity : prices -- **non-stationary** ; returns – **stationary**;
  - Sample Distribution : the distribution of returns is closer to the assumption of **a normal distribution**.

price



less heavy-tailed

return

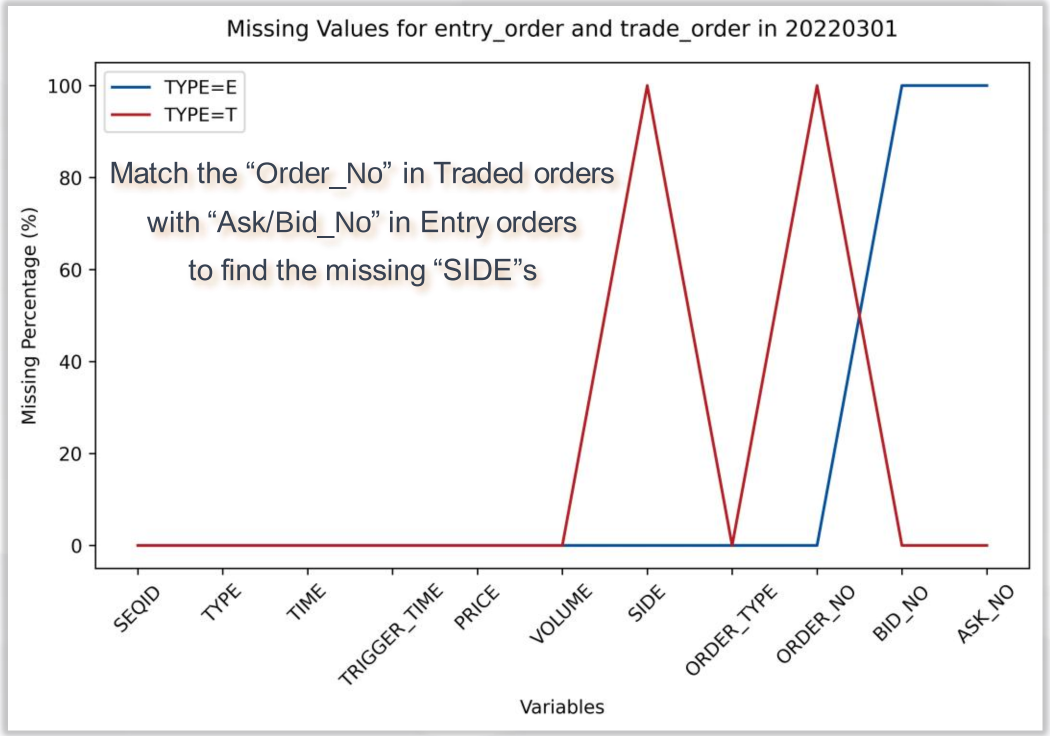


02 Data we used: Extracting useful info. from raw data

- Focus on traded orders of **stock 000027**
  - **Tick data**: records of executions

    Matching the **Entry** info. corresponding to **Traded** orders to obtain comprehensive data.
  - **Snapshot data**: records of quotations

    Matching the **optimal bid and ask prices** closest to the transaction time, along with the **corresponding traded volumes.**



- Handle with outliers and missing values
  - Remove orders outside of trading hours: **Trading hours: 9:30 - 11:30 & 13:00 - 15:00;**
  - Imputation operations: **Forward fill** the missing values (Delete orders with too many missing fields).

03 Predictors & Labels

- **Predictors:** consider 10 kinds of factors and 2 lookback windows (Calendar Clock)
  - Each factor is considered for **2 lookback windows**: (T-5s, T] & (T-30s, T-5s] → totally 20.

Category	Abbreviation	The specific meanings
CAT 1- Stock Trading Intensity	1.1. Breadth	**Number of trades** within the lookback window (T-Δ, T]
	1.2. Immediacy	**Average time interval** between adjacent trading orders within the lookback window (T-Δ, T]
	1.3. VolumnAll	**Total shares traded** within the lookback window (T-Δ, T]
	1.4. VolumnAvg	**Average number of shares traded** within the lookback window (T-Δ, T]
	1.5. VolumnMax	**Maximum single traded volume** within the lookback window (T-Δ, T]
CAT 2- Asymmetry of Trading	2.1. LobImbalanca	**Average imbalance indicator of the limit order book depth** within the lookback window (T-Δ, T]
	2.2. TxnImbalance	**Asymmetry of traded volume for buy and sell orders** within the lookback window (T-Δ, T]
	2.3. PastReturn	**Stock returns momentum** within the lookback window (T-Δ, T]
CAT 3- Inherent Speed and Cost in Trading	3.1. QuotedSpread	**Average nominal spread** (quoted spread) within the lookback window (T-Δ, T]
	3.2. EffectiveSpread	**Weighted percentage effective spread** within the lookback window (T-Δ, T]

03 Predictors & Labels

- **Future Labels:** consider the average return within the lookahead window (Not instantaneous price)
  - Consider 2 lookahead window (T, T+5s] & (T, T+30s] to estimate the duration of predictability ;

$$Return(T,T+\Delta) = \frac{Average\ \boldsymbol{Transaction\ Price}\ in\ (T,T+\Delta]}{Simple\ average\ of\ \boldsymbol{optimum\ ask}\ \&\ \boldsymbol{bid\ price}\ at\ T} - 1.$$

  - Strength of this calculation way of return compared to instantaneous price:
    - less volatility than instantaneous prices, with reduced data noise;
    - reflect more: aggregating trading behavior over a short time span;
    - Determining trading time intervals is easier to achieve than specific trading moment.
- **Final data cleaning**
  - Remove factors with missing information

delete orders for **the first and last 30 seconds of each day's records**, since calculating labels and predictors involves using lookahead and lookback windows of 30 seconds

04 Models & Cross-Validation

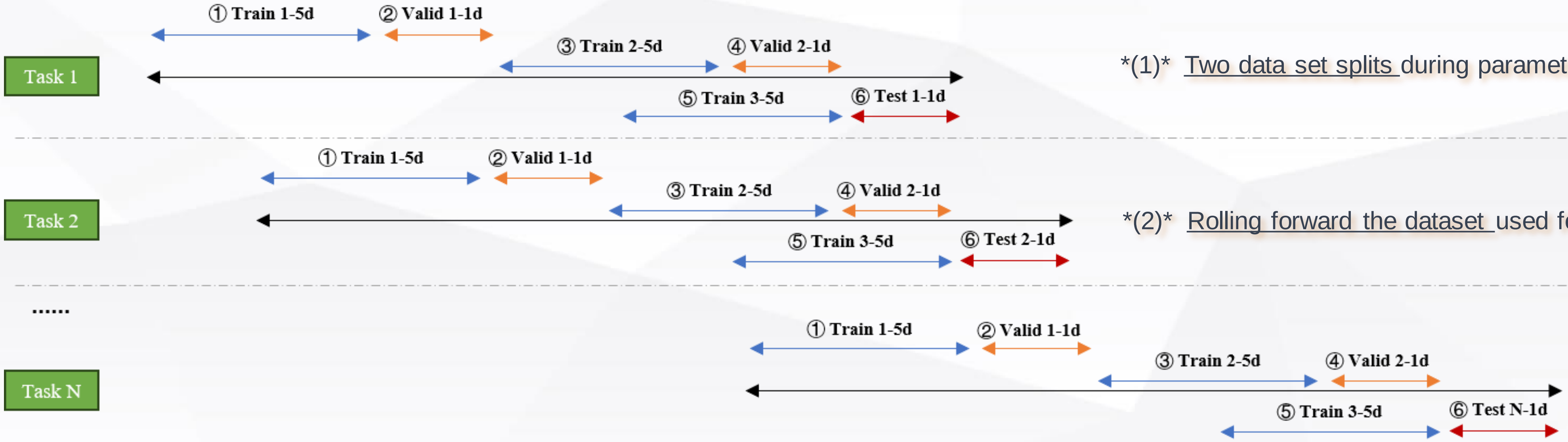
- **Machine Learning Models we chose:**

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

Flexible model structure of Machine Learning

- ■ **LASSO:** **Linear** + **Regularization** term to enhance sparsity;
- ■ **Ridge:** **Linear** + **Regularization** term to shrink the absolute values of all coefficients;
- ■ **Random Forest:** **Nonlinear** + Considering the complicated **interaction** among the predictors;

- **Rolling Prediction:** Synthesizing Ait-Sahalia et al. (2021) and Gu et al. (2019).



(1) Two data set splits during parameter tuning

(2) Rolling forward the dataset used for each prediction

05 Data Range & Criterion for Accuracy

- **Time range of the test set:**
  - One stock: " *Shenzhen Energy* " (Stock Code: 000027);
  - Test set: 65 trading days containing **1,308,436 records** in **2022.01 - 2022.05** **(2021.12 is missing)**;
  - The specificity of the test set range for energy stock:

    Encompassing **the rebound of the COVID-19 pandemic in China** and **the outbreak of the conflict between Russia**(Major energy-supplying nation) **and Ukraine** in 2022.02.

- **Criterion for Accuracy:** out-of-sample R^2 refer to **Gu et al. (2019).**

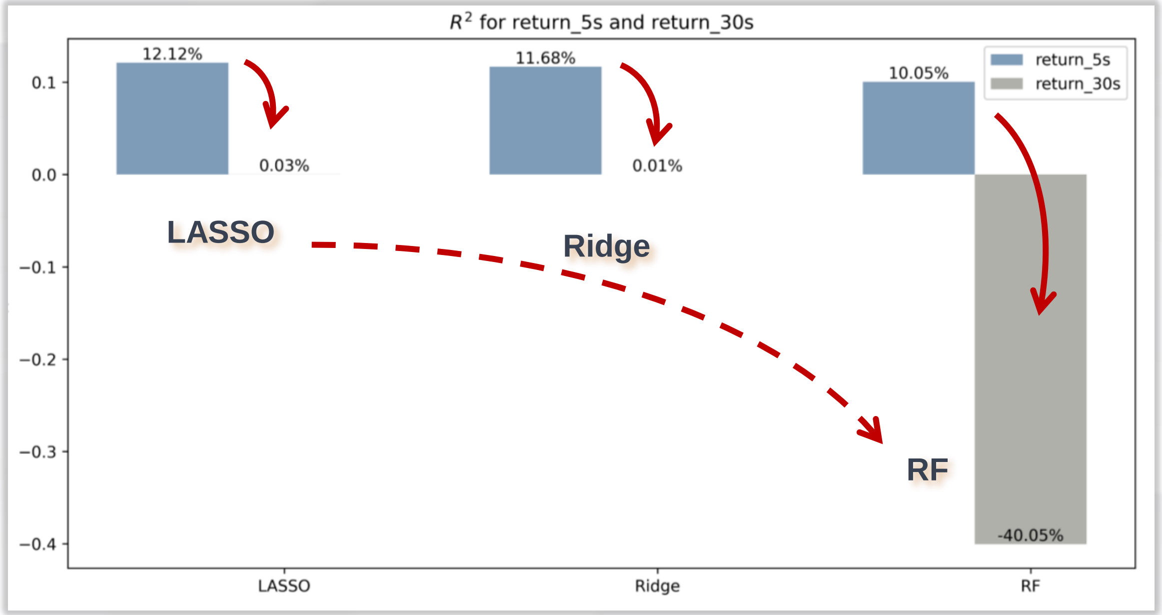
$$R^2_{oos} = 1 - \frac{\sum_{(i,t)} \left(r_{i,t+\Delta} - \hat{r}_{i,t+\Delta} \right)^2}{\sum_{(i,t)} \boldsymbol{r^2_{i,t+\Delta}}}$$

  - Exclude the mean of actual values from the denominator's squared term

    In financial forecasting, **mean predictions are less effective than zero predictions**. Calculating R^2 using the mean would artificially **lower the standards** for predictive evaluation

06 Model Comparison & Duration of predictability

- **Out-of-sample R^2 :** LASSO > Ridge > RF & return_5s > return_30s



	LASSO	Ridge	RF
R^2 for **return_5s**	**+ 12.12%**	+ 11.68%	+ 10.05%

	LASSO	Ridge	RF
R^2 for **return_30s**	+ 0.03%	+ 0.01%	**- 40.05%**

*need to explore why RF performed so terribly

- Advantage of the sparsity and sensitivity to redundant features of LASSO ;
- The advantage of LASSO & Ridge can reduce the risk of overfitting, adapting better to the noise;
- In the short term, the market tends to show a simpler, more linear behavior;
- The predictability duration of high-frequency returns is very short (rapid decay from 5s to 30s);

07 Cautions for Extreme return

- **The accuracy of the model in identifying extreme returns is not very high**
  - Extreme real return: **-100%**

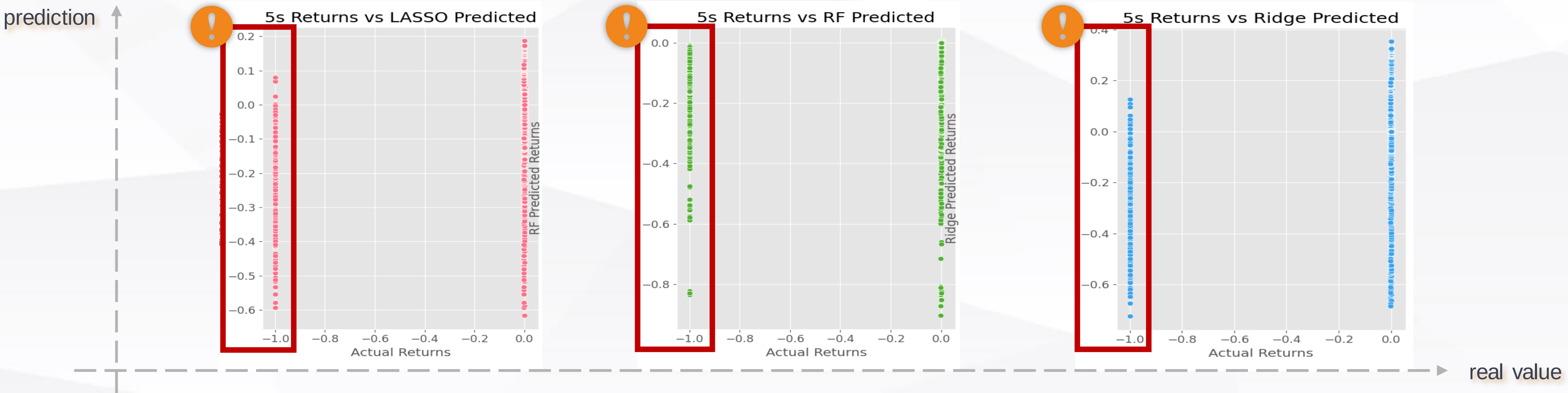
    The “**return = -1**” means there are no traded orders within the lookahead window.

    ➡ It does **happen in the real trading market**, but the models we used couldn’t identify it well.

    ➡ **The overall prediction accuracy** of high-frequency trading is impacted.

    ➡ **High-frequency investors** need to pay attention to the impact of this situation.

Real return = -1 can’t be predicted well



08 Specialty around opening & closing time

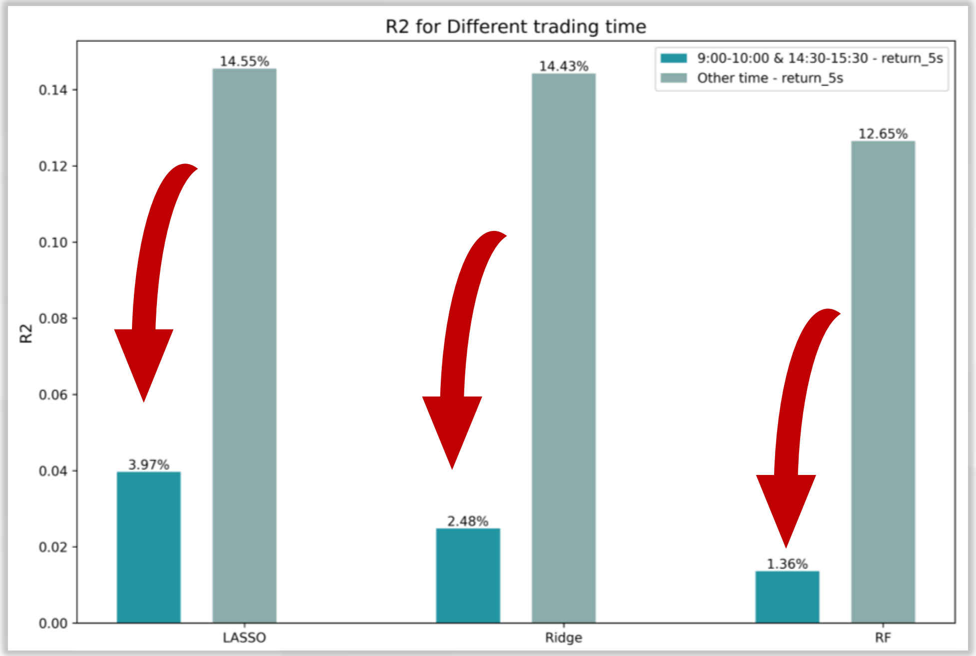
● 9:30-10:00 & 14:30-15:00 V.S Other trading time

■ Comparison Result:

R^2 around opening & closing time is much worse.

■ Potential Reasons

- Poor market **liquidity** around opening & closing time;
- **Trading volume** may experience significant increases or decreases;
- **New information** may be released before the opening and after the closing, leading to **information asymmetry** in the market and causing significant volatility;
- **Investor sentiment** may become more volatile, leading to increased uncertainty in market behavior;
- **Intrinsic characteristics** like the procedure of trading.



9:30-10:00 & 14:30-15:00 worse than Other trading time

09 Variation of R^2 by date

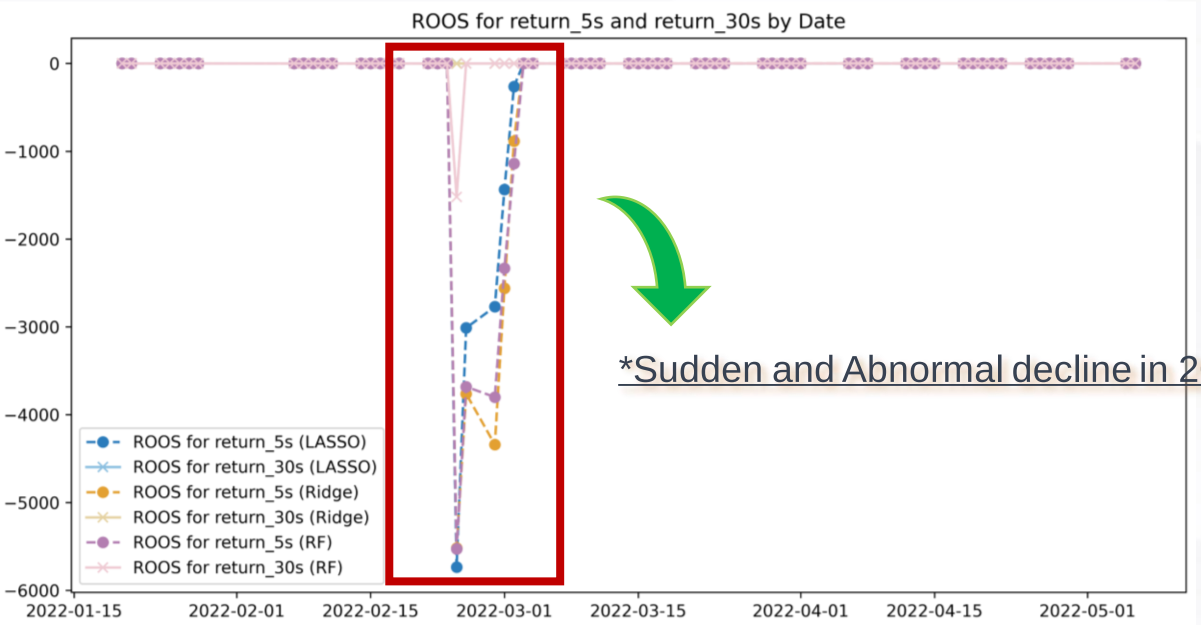
● Sudden Decline in the end of 2022.02 :

■ the rebound of the pandemic in China

Negatively impacting the **production chain**, **energy demand**, **investor confidence**, and **overall market sentiment.**

■ the conflict between Russia and Ukraine

Russia, a major global **energy supplier** and **reserve holder**, will influence **the global energy market**, thereby impacting the energy stock(code 000027) under our study.

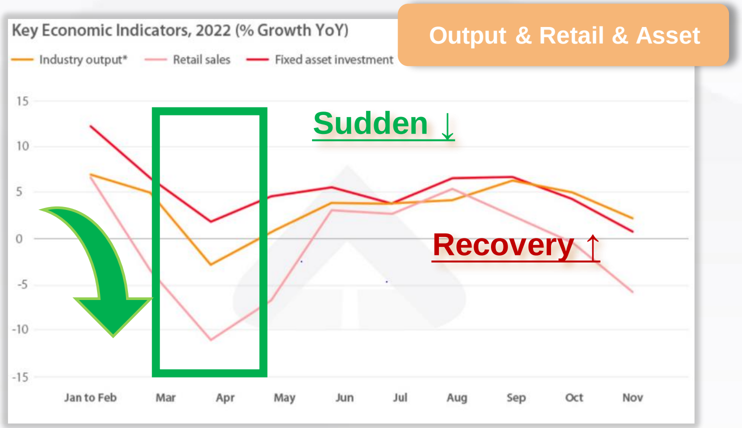


*Sudden and Abnormal decline in 2022.02

R^2 of only the RF decreases when predicting return_30s. This could be one of the reasons for the low accuracy of RF in predicting return_30s

09 Variation of R^2 by date

- **Influence of the Rebound of the COVID-19 pandemic in China:**
  - A decrease in macroeconomic index
  - Investors' concerns about the capital markets

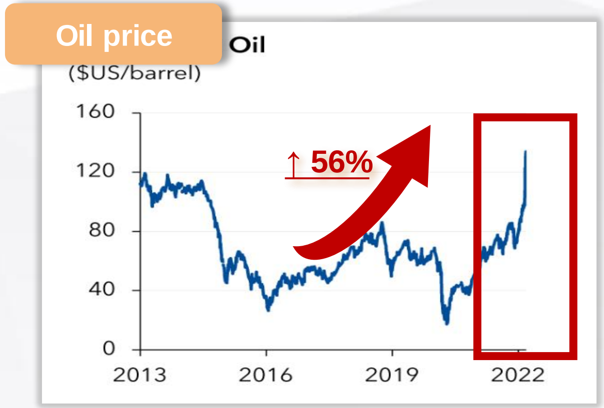


- **Influence of the Conflict between Russia and Ukraine**
  - the Surge in **global crude oil prices**

    In the past year, **the US dollar expenditure** on China's crude oil imports increased ↑ **44%.**

    ➡ Lead to the **supply pressure in the renewable energy** industry continues to rise.(**main business of 000027**)
  - the fluctuations in **global energy supply chain** & **demand**

    Russia: major supplier of crude oil (**15%** of **global exports** & **10%** of **preparations**);
  - **Market sentiment** is cautious



10 Importance Measurement for predictors

● Utilizing the variable importance method provided by tree model

■ Why choose this measurement?

RF shows **decent accuracy in predicting return_5s** but **lags behind linear models**(worth investigating).We aim to explore potential reasons for the suboptimal performance of RF **focusing on variable selection**.

■ Best Predictors and Best lookback window

- **Breadth**, **Immediacy** and **PastReturn** are the most important predictors;
- **Closer lookback window** performed better.

The primary sources of predictability

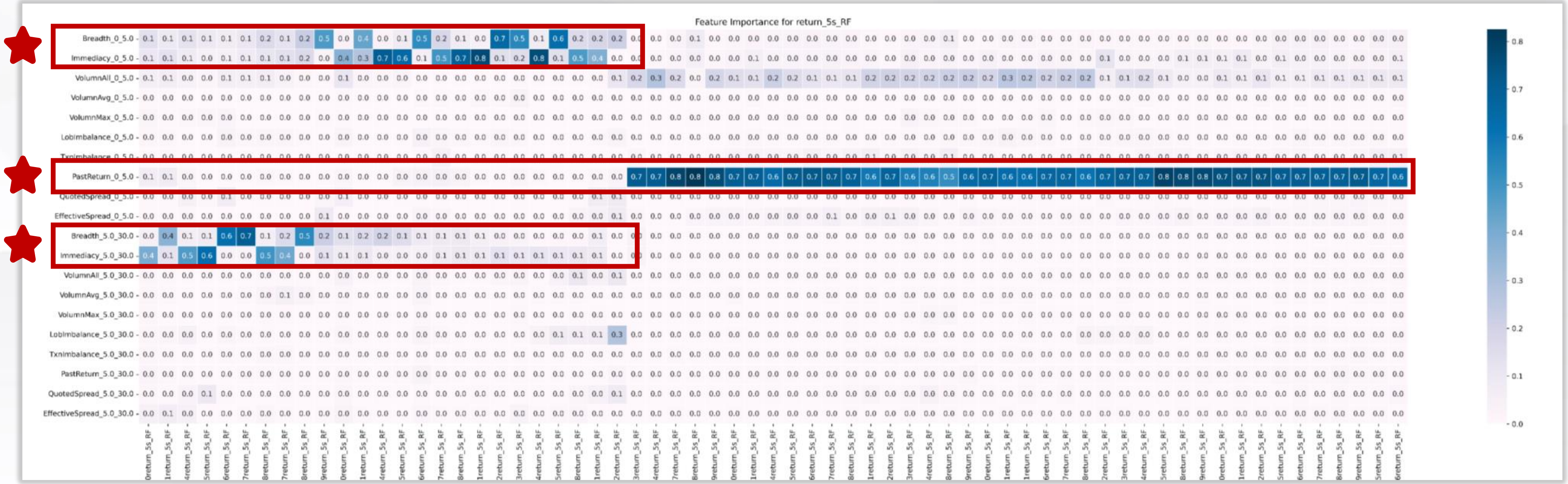
Breadth_0_5s

Immediacy_0_5s

PastReturn_0_5s

Breadth_5_30s

Immediacy_5_30s



11 Sudden shift in ranking patterns

● Ranking pattern shift at the beginning of 2022.03

■ What is the specialty of the beginning of 2022.03 ?

2022.03.03 is **the corresponding first test** set after **including the period of the Russia-Ukraine conflict and the resurgence of the COVID-19 pandemic in China in the training set.**

■ The importance of Past Return

In periods of high market volatility, only **the momentum factor** proves to be a **robust source of predictability**;

■ The variable selection of RF is very limited

After 2022.03, RF tends to favor 1/2 variables, limiting its interaction ability(potential reason for worse performance);

12 Financial Explanation for important predictors

- **Financial meaning behind important predictors (when predicting return_5s/30s)**
  - **Breadth:** an indicator for the general **market sentiment** and its impact on energy stocks;
  - **Immediacy:** **liquidity** & **immediacy**;
  - **VolumeAll:** **overall market activity** and interest;
  - **VolumeAvg:** a more stable perspective by smoothing out daily fluctuations;
  - **VolumeMax:** the **extremes** of market activity;
  - **PastReturn:** the **momentum** effect;
  - **LobImbalance:** the **supply and demand dynamics** within the market;
  - **EffectiveSpread:** the significance of **transaction costs** and **market liquidity**;

13 Advice on Risk and Investment

● Cautious Investing Approach

Approach stock market investing cautiously, considering varied predictive abilities and evolving factors, and choose **distinct models** to predict.

● Cautious Interpretation of Market Sentiment

While **market sentiment**, reflected in market breadth, provides insights, caution is advised, especially in the **dynamic energy sector** where trends can swiftly reverse.

● Diversification for Volatility

Given the inherent volatility of energy equities influenced by factors like market breadth and trading volumes, **a diversified portfolio** is crucial to mitigate risk.

Thanks for Listening

------- Members of Group 6 -------

| 卢云轩 | 何尹 | 张伯杨 | Gloria Timone | Alvaro Paredes |

| Mary Kilonzo | Martina Martinova | Mohamed Abdelsamie |