QF604 Group Project Report

Team member:

Yu Lingfeng
Wang Wenjie
Wang Haotong
Chen Liangrui
Liu Jing

Contents

1

Abstract

01 Abstract

Original Paper –

Target: predict clean energy stock price under the impact of climate changes

Method: Random Forests, Extremely Randomized Trees, Gradient Boosting and Support Vector Machine outperform Lasso and Naïve Bayes

Result: prediction accuracy up to 85% under Random Forests, GBM methods

Improvement –

Method: Linear Regression (which gives best result and more stable)

Result: prediction accuracy around 60%

2

Replication

02 Replication

Study period: 1st, Jul 2022 -- 1st, Feb 2024

Train/Test Split: 70%/30%

Input feature

01

Original feature:

VIX (S&P500) VXN (NASDAQ100) OVX(Crude Oil)

Economic Policy Uncertainty(EPU) Economic Market Uncertainty(EMU)

Infectious Disease Tracker(IDT)

02

Technical Indicators:

Moving Averages(MA50 & MA200)

Williams Accumulation/Distribution(WAD)

03

Additional Feature for Improvement:

TNX (yield on U.S. Treasury bond index)

SPX (S&P 500 index) N225(Nikkei 225 index)

JPY=X (Japanese Yen to USD exchange rate)

EUR=X (Euro to USD exchange rate)

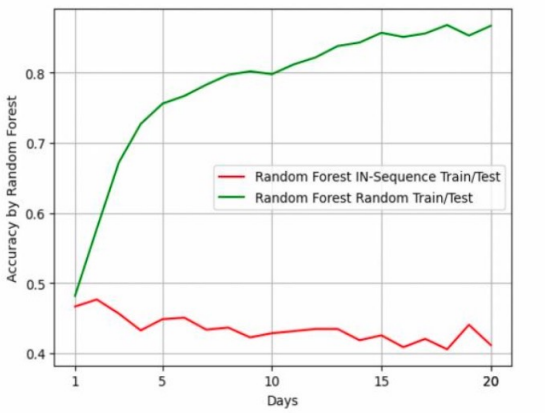
Hang Seng Index (HSI) Shanghai Composite Index (SCI)

02

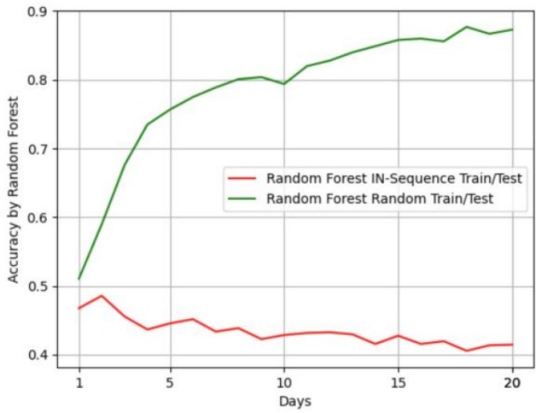
Replication– Machine Learning Method

Random Forest

Number of trees: 100

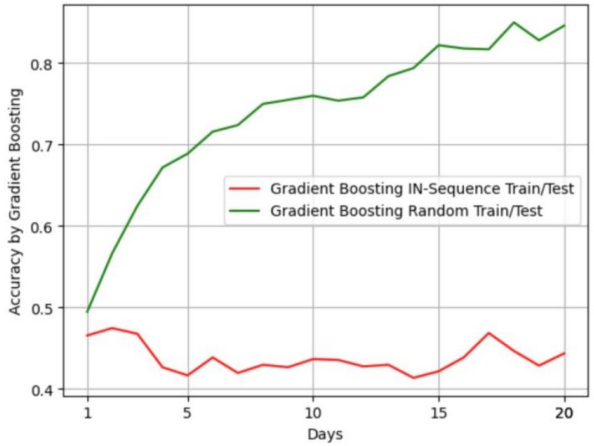


Number of trees: 500

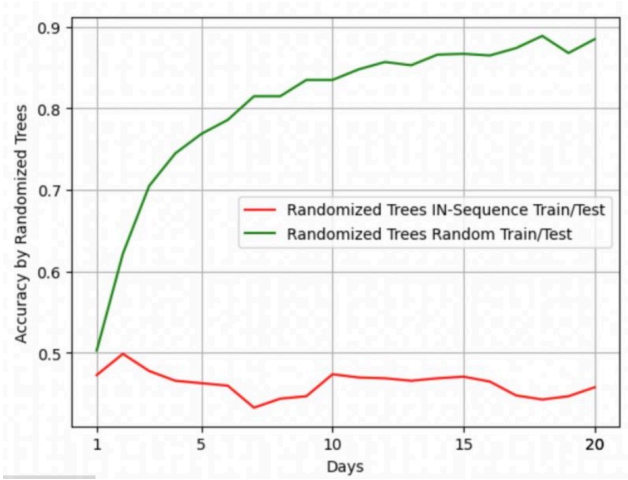


Gradient Boosting

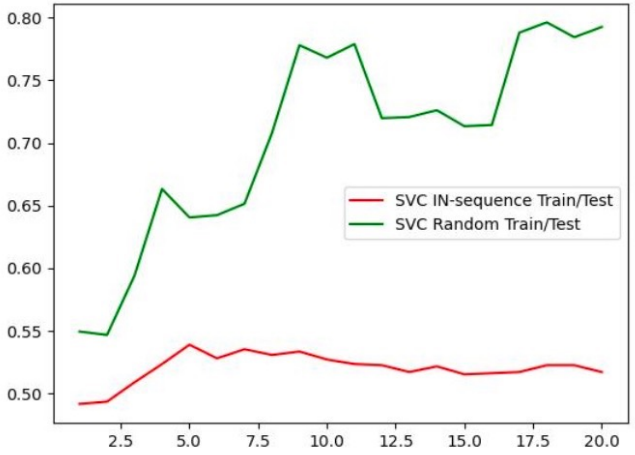
Number of boosting stages: 500 learning rate: 0.05



Extremely Randomized Tree



Support Vector Machine

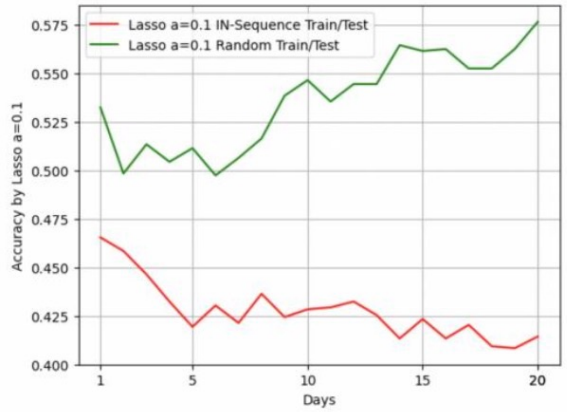


02

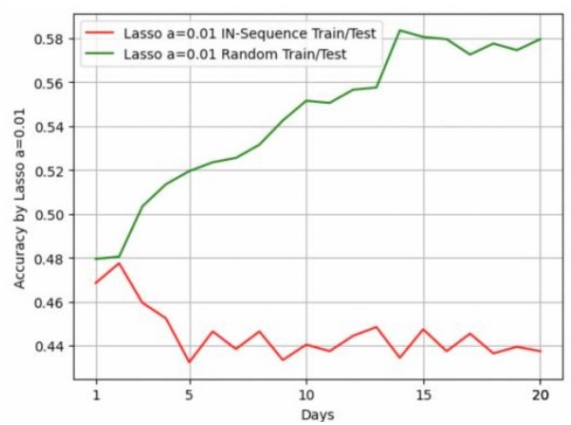
Replication– Machine Learning Method

Lasso

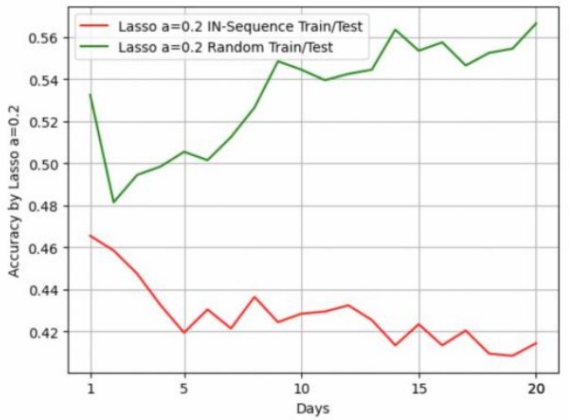
Alpha = 0.01



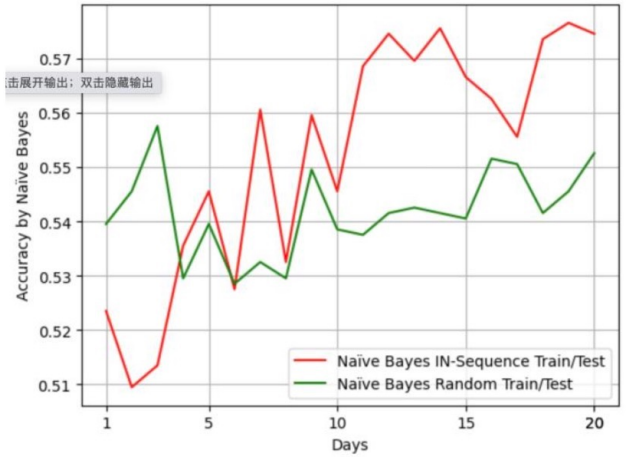
Alpha = 0.1



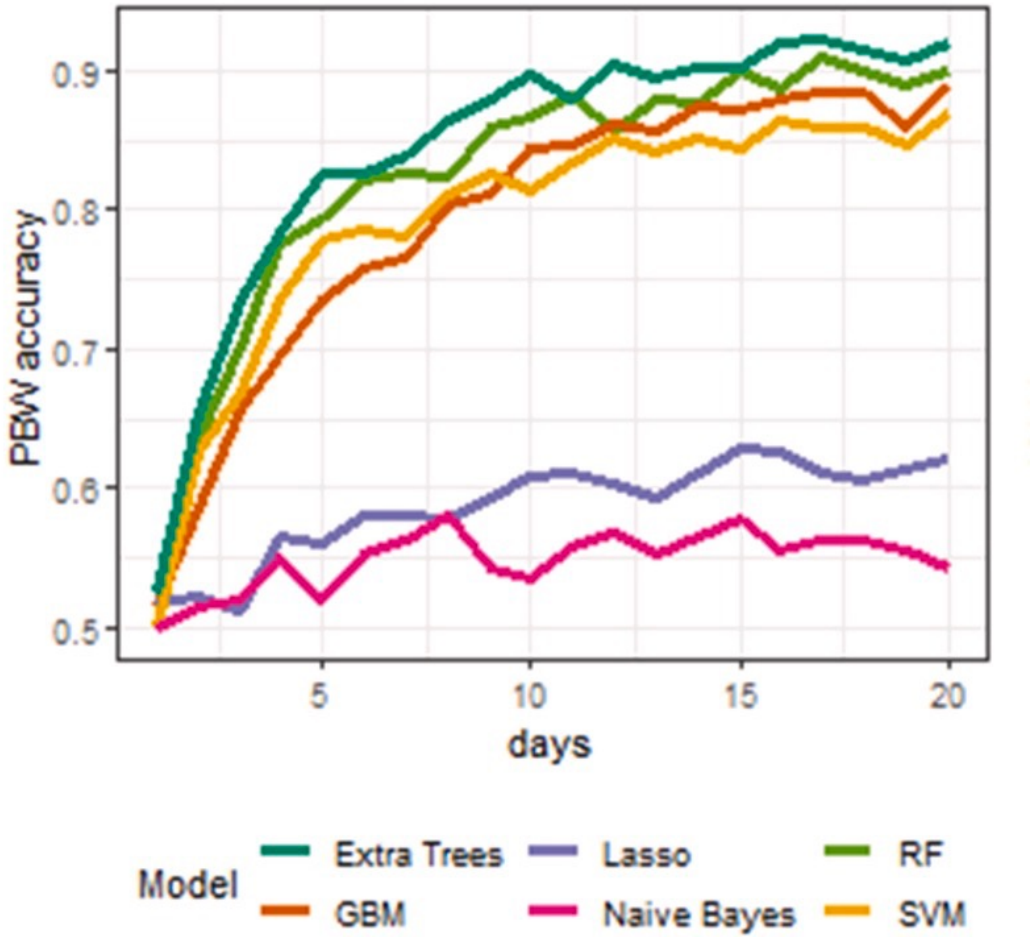
Alpha = 0.2



Naïve Bayes



02 Replication – Compare results



Compare

The replicated results are **similar**, which is **as high as 90%**, just be doing a random partition on training and test set.

3

Improvement

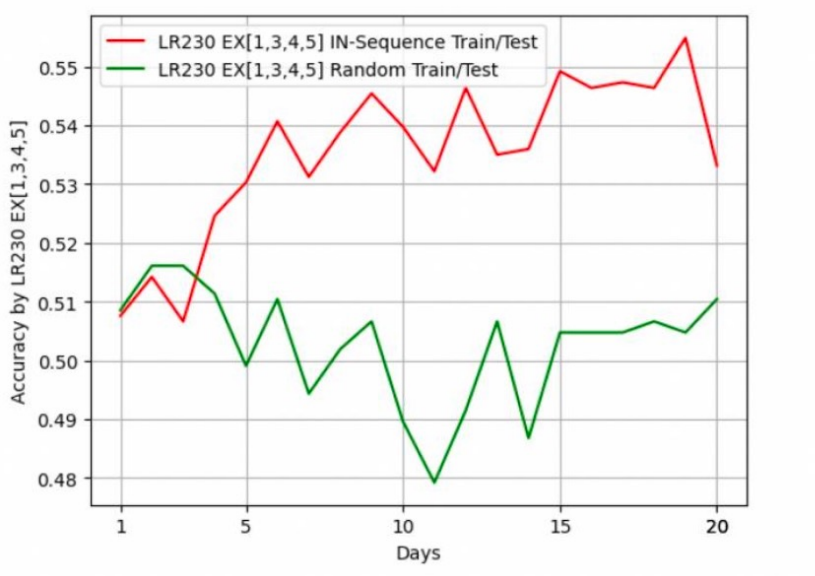
03 Improvement – Linear Regression

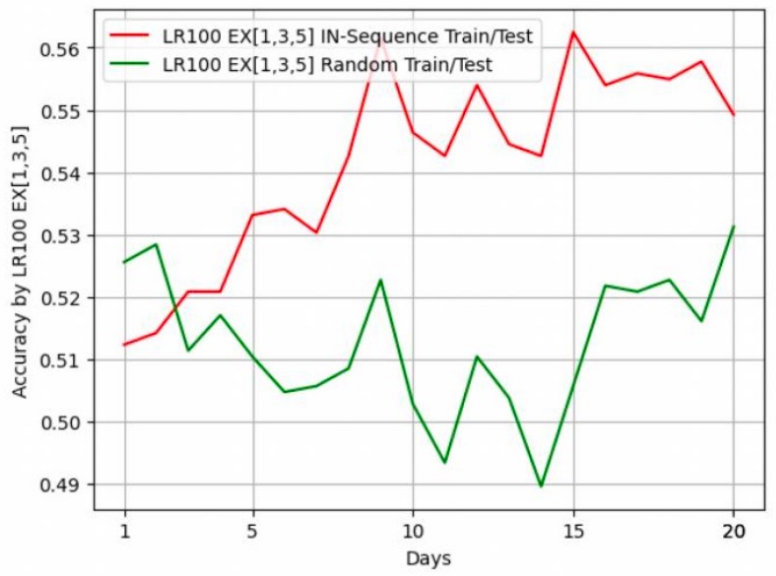
Input features: **9 features** (exclude EPU, EMU, IDT)

Do **correlation analysis** to identify effective predictors.

Do **mean correlation cut off** to select effective predictors.

Generate **511 combinations** and select **best two** performances.





4

Conclusion

04 Conclusion 1 – Train/Test Effects

Train/Test Effects

Factor

The **choice of train/test split** effect performance of ML methods

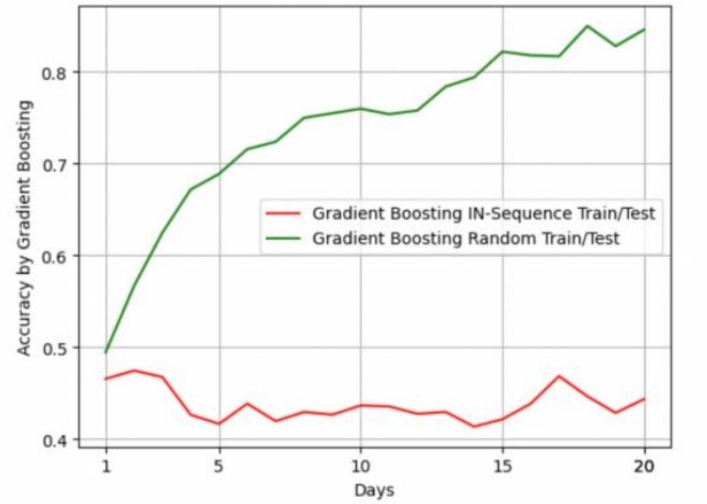
The **sequential arrangement** of data effect prediction accuracy

Result

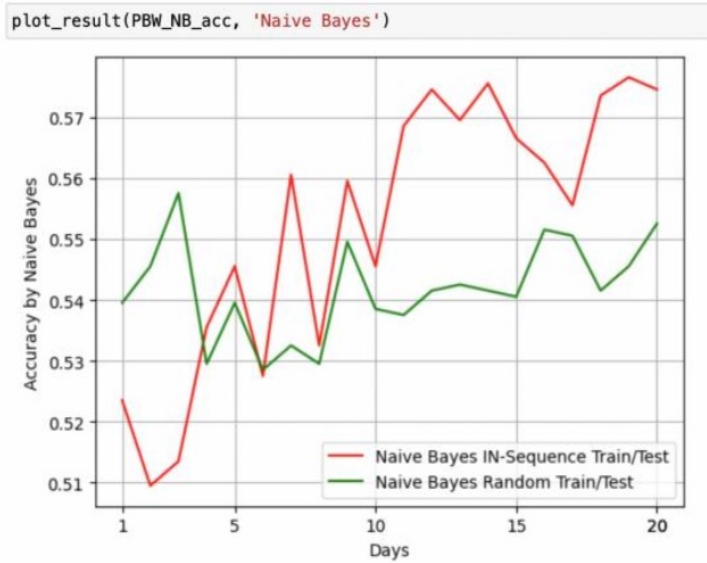
Gradient Boosting preforms better than Naïve Bayes under the way we split

Take Naïve Bayes as benchmark (as author), GB demonstrate **instability**.

```
plot_result(PBW_GB_acc, 'Gradient Boosting')
```




```
plot_result(PBW_NB_acc, 'Naive Bayes')
```



04 Conclusion 2 – Features Set Effects

Features set Effects

Replication Part

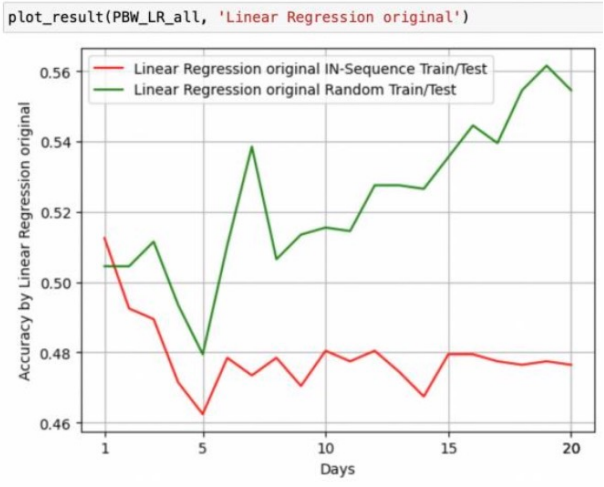
High performance due to **improper partitioning train/test set.**
The assessment of feature significance remains **inconclusive.**

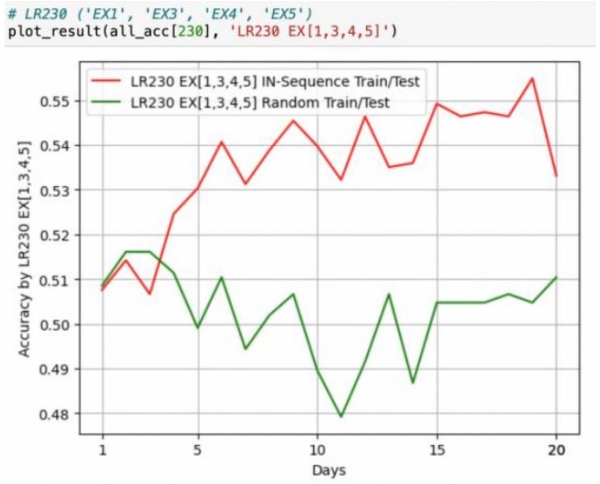
Some features exhibit **improper significance**. IDT is susceptible to **look-forward bias**.
Most technical Indicators have **high degree of freedom** and is subject to **overfitting**.

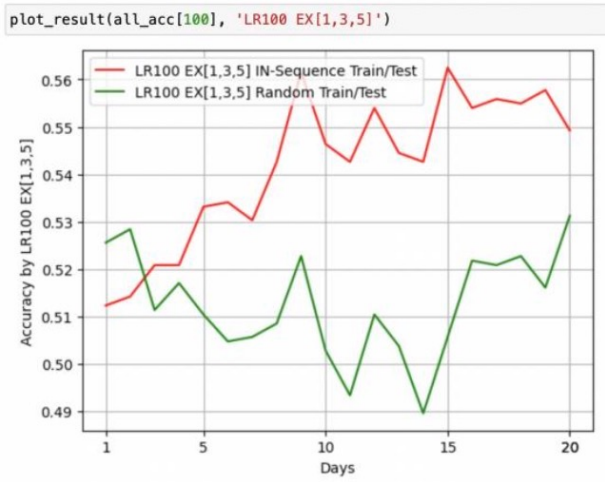
Improvement Part

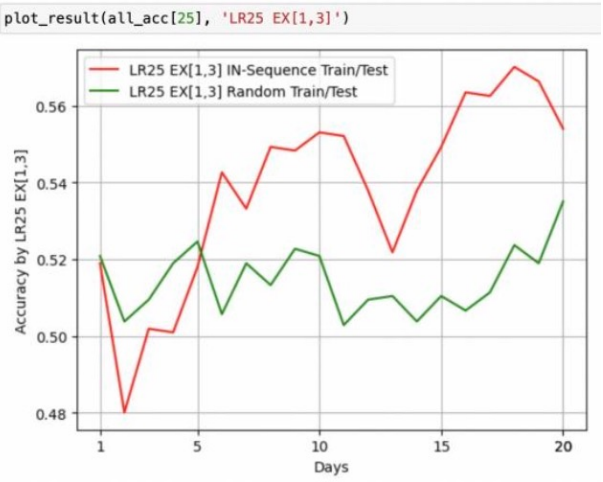
Feature selection and tuning can improve prediction accuracy.
The testing window we introduced, with a downward trend, limit the testing results.

The best linear model is feature combinations EX1, EX3, EX4,EX5.
The features selected by the author have **lower prediction accuracy** than we added.









04 Conclusion 3 – Final Model Testing with Out-of-Sample testing

Add a final test on out-of-sample data set from **07/01/2022 – 01/02/2024**

04 Conclusion 3 – Final Model Testing

Serve as a **benchmark** for comparison with other models

Naïve Bayes

Gradient Boosting

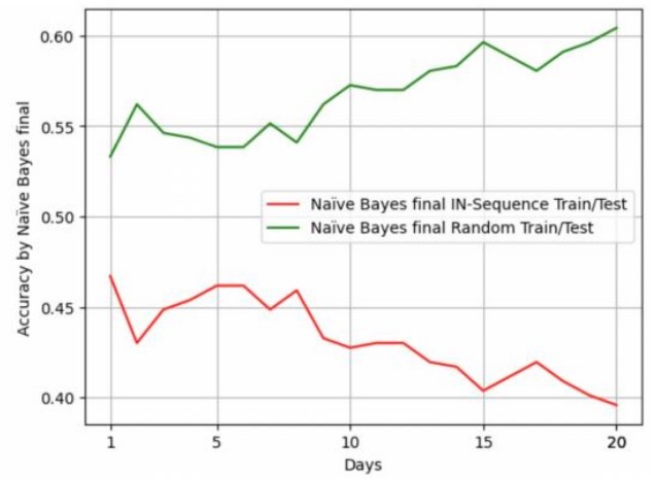
Stand out as top performing ML methods

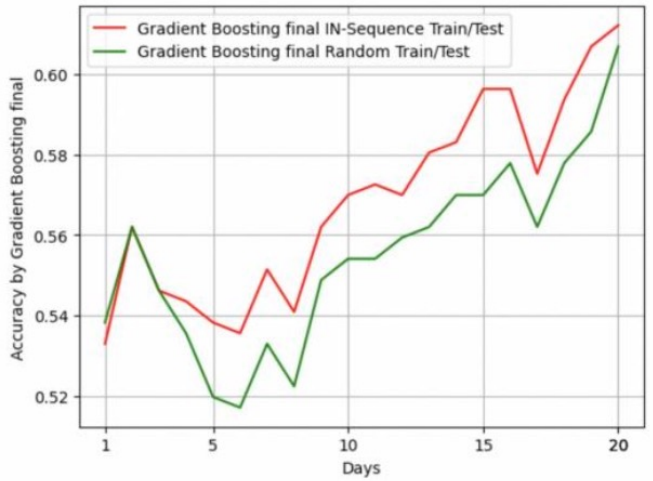
The In-Sample test with random partition gives an 80% accuracy; in-sequence partition gives around 45%.

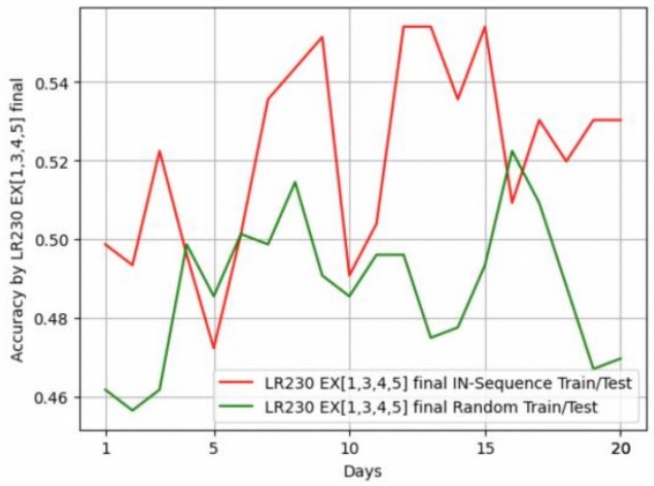
The Out-of-Sample test with random partition gives an 60% accuracy; in-sequence partition gives around 60%.

The performance **dropped significantly**, and implied **non-consistency**.

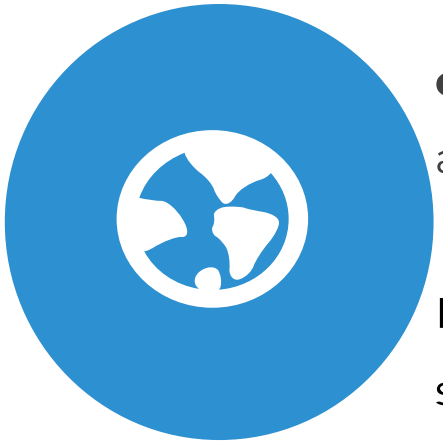
The result is better than final model testing, yet this could be due to **market regime change.**







04 Conclusion - Summary



The author has a problem with division of dataset, resulting in **test set containing large information in train set**, which leads to high prediction accuracy results.

Dividing the dataset **in chronological order** and selecting **stable model** such as linear regression and SVM, may be a more appreciate way for prediction.

THANK YOU!