Replication of TW Total Return Index

Yu-Ching Liao, Futures Prop. Trading Div., Prop. Trading Dept.

Problem Formulation

Problem Formulation:

- Totally 963 components → Difficulty of replication

```
1  TX_Closing
```
✓ 0.0s

	2330	2317	2454	2412	2308	6505	2881	2882	2303	2382	...
2018-08-15	213.17	77.85	195.99	88.38	92.48	105.46	35.12	41.63	13.15	37.75	...
2018-08-16	210.96	77.94	186.56	87.97	90.29	105.46	34.84	41.31	13.30	37.46	...
2018-08-17	211.40	77.85	186.19	88.38	92.04	105.90	34.98	41.55	13.34	38.04	...
2018-08-20	211.40	77.75	186.19	88.38	92.04	108.09	34.98	41.15	13.61	38.04	...
2018-08-21	212.73	77.75	185.06	89.62	92.48	109.40	35.05	41.63	13.96	38.26	...
...	...	...	...	...	...	...	...	...	...	...	...

Problem Formulation:

We are aiming to…

- Replicate the index with lesser components.
- Minimize the tracking error.

Problem Formulation:

Problem to be solved:

1. How much components?
2. How much training data?
3. How to achieve so?

$$N_{components} \times N_{days} \times Methodology$$

Dataset Overview

Adj. Close PX of Components

- Time Interval: Daily
- Ranged from 2018/11/30 to 2023/07/20

```
1  N_TX_Closing
```
✓ 0.0s

	2330	2317	2454	2412	2308	6505	2881	2882	2303	2382	...
2018-11-30	199.04	56.88	179.03	88.38	113.95	97.59	34.20	38.95	8.96	36.40	...
2018-12-03	207.43	58.55	185.06	87.55	113.95	101.09	34.45	39.39	9.07	36.66	...
2018-12-04	206.55	58.07	189.96	87.55	113.95	98.46	34.17	39.27	9.15	36.66	...
2018-12-05	199.49	56.65	184.30	87.97	113.08	98.90	33.89	38.82	8.96	36.25	...
2018-12-06	194.19	54.59	173.00	87.97	111.76	96.71	33.64	38.70	8.76	35.56	...
...	...	...	...	...	...	...	...	...	...	...	...

Weights of Components

- Time Interval: Daily
- Ranged from 2018/12/03 to 2023/07/20

```
1  N_TX_Weights
```

✓ 0.0s

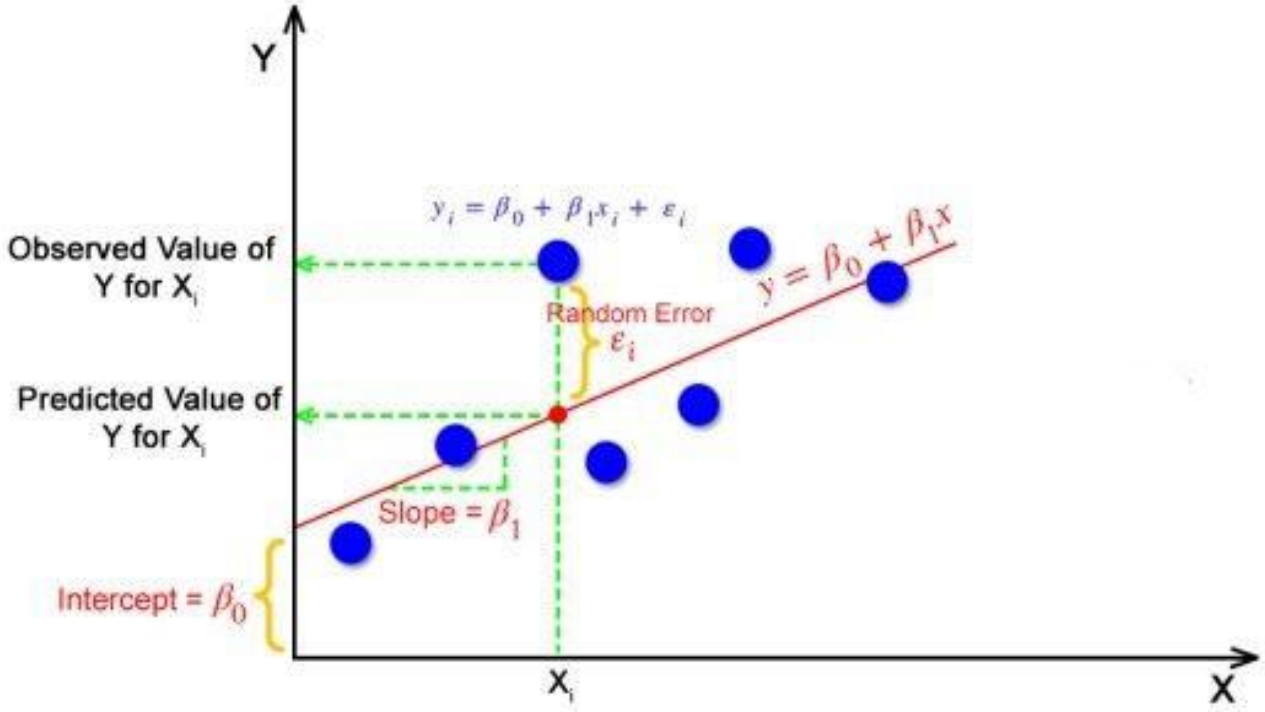
	2330	2317	2454	2412	2308	6505	2881	2882	2303	2382	...
2018-12-03	0.211463	0.000000	0.013670	0.030018	0.012212	0.038412	0.018061	0.022058	0.005189	0.006977	...
2018-12-04	0.211463	0.000000	0.013670	0.030018	0.012212	0.038412	0.018061	0.022058	0.005189	0.006977	...
2018-12-05	0.211463	0.000000	0.013670	0.030018	0.012212	0.038412	0.018061	0.022058	0.005189	0.006977	...
2018-12-06	0.211463	0.000000	0.013670	0.030018	0.012212	0.038412	0.018061	0.022058	0.005189	0.006977	...
2018-12-07	0.211463	0.000000	0.013670	0.030018	0.012212	0.038412	0.018061	0.022058	0.005189	0.006977	...
...	...	...	...	...	...	...	...	...	...	...	...

P.S. This should be adjusted to sum-up to 100%.

Methods Review

Linear Regression

$$min\frac{1}{n}\sum(\hat{y} - y)^2$$



Lasso Regression

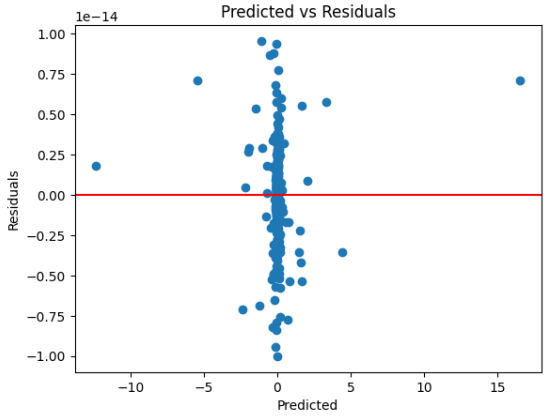
$$\min\{\frac{1}{n}\sum(\hat{y} - y)^2 + \lambda \sum \| \beta \|^2\}$$

Penalization to set features to 0
*→ Deal with **Multi-collinearity***

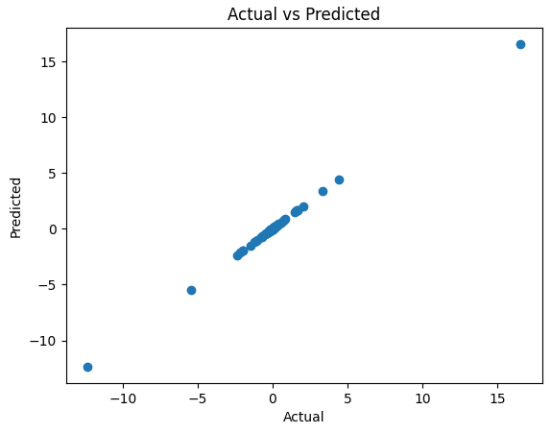
Linear Regr. Assumption Check:

```
Durbin Watson test statistic:  1.8111323743805 36
```

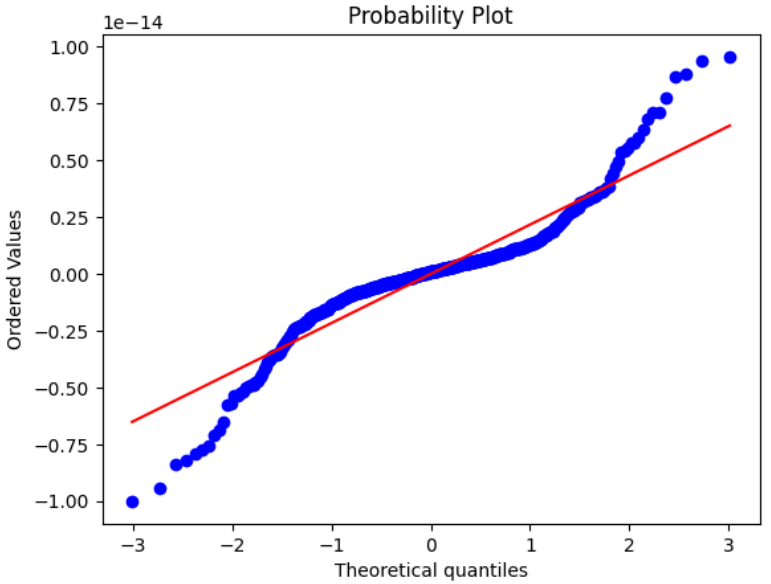
Independency: Passed



Homoscedasticity: Passed

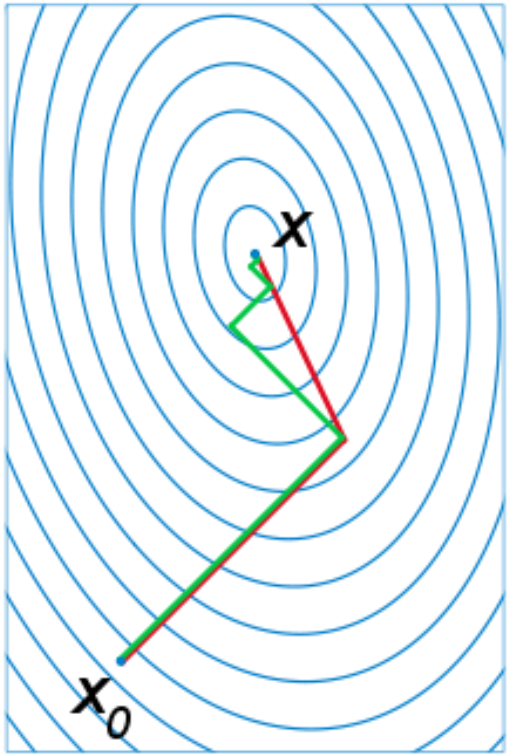


Linearity: Passed



Normality in residual: Unpassed !!!!

Conjugate Gradient Research



- Optimization Technique based on Slope.
- Risk of Local Optimum → Must have suitable Initial Condition.

Data Handling – Dealing with NaNs

let...

$$N_{components} = 700$$

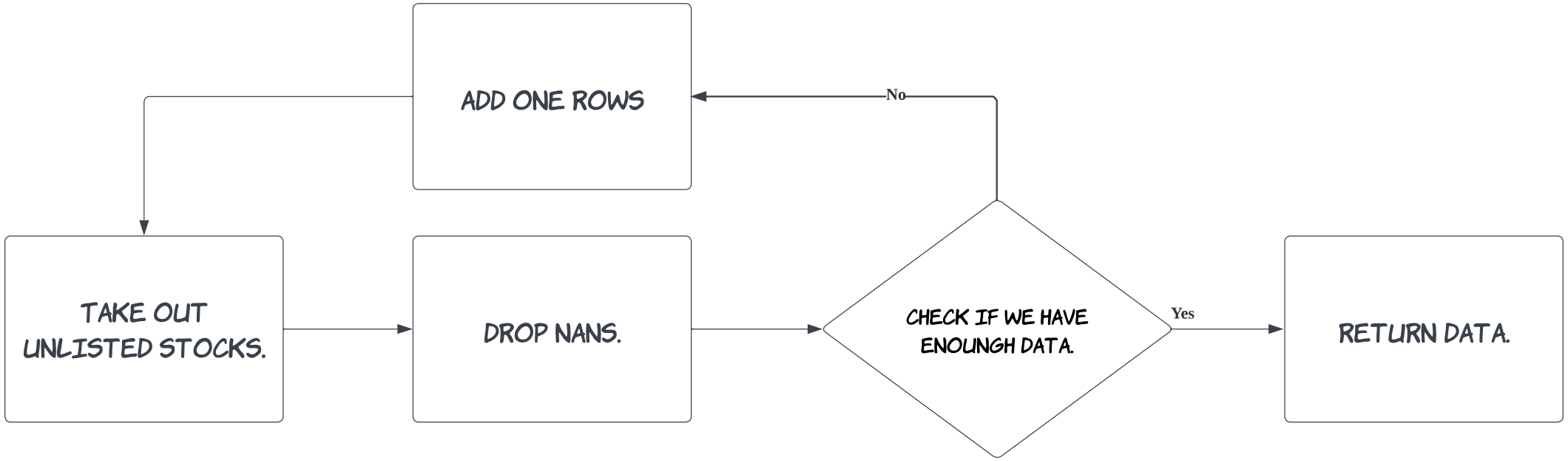
$$N_{days} = 800\ (days)$$

Step 1: Get Latest *700* x *800* daily data

	2330	2317	2454	2412	2308	6505	2881	2882	2303	2382	...
2019-12-20	304.54	75.80	345.89	95.50	132.55	89.12	34.25	35.57	13.49	49.50	...
2019-12-23	309.16	76.14	348.22	95.93	134.35	89.85	34.18	35.57	13.36	49.65	...
2019-12-24	307.31	75.64	345.11	95.07	132.55	89.57	34.07	35.45	13.36	49.65	...
2019-12-25	308.24	75.72	350.54	95.07	133.00	88.94	34.03	35.41	13.45	49.65	...
2019-12-26	308.24	75.55	348.22	95.50	132.09	88.39	34.07	35.45	13.36	49.65	...
...	...	...	...	...	...	...	...	...	...	...	...

P.S. Should have 800 rows after dealing with NaN values.

Tips: How to deal with NaNs?



We can get the data with no NaNs except unlisted stocks!!!

Tips: How much data we need? :

$$Training\ Data = N_{days}$$

$$Testing\ Data = 50\ days$$

$$Cost\ of Calculating\ Return = 1$$

$$= Totally\ N + 50 + 1\ Data$$

Step 2: Get Daily Return

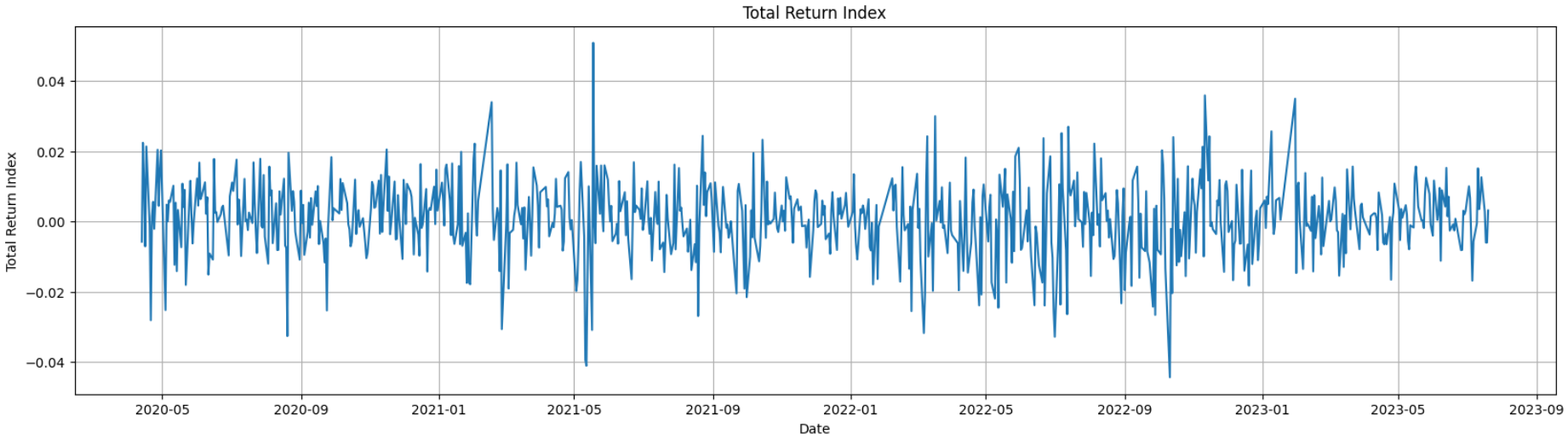
$$\text{Daily Return} = \ln\left(\frac{P_{t+1}}{P_t}\right)$$

```python
def calculate_log_returns(df: pd.DataFrame) -> pd.DataFrame:
    return df.apply(lambda x: np.log(x / x.shift(1)))
```

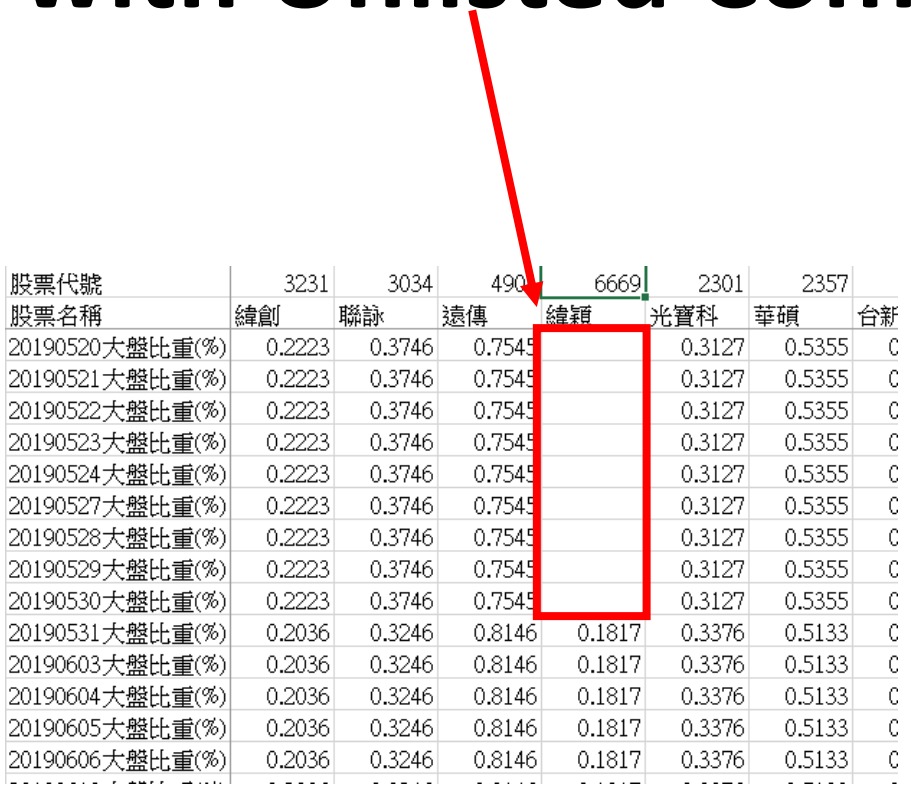
	2330	2317	2454	2412	2308	6505	2881	2882	2303	2382	...
2020-04-13	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
2020-04-14	0.023090	0.017599	0.039066	0.004617	0.054546	0.011023	0.028612	0.015638	0.023161	0.028623	...
2020-04-15	0.008709	-0.004046	0.000000	0.004596	0.033511	0.010903	0.010023	0.007885	-0.003276	-0.004680	...
2020-04-16	-0.003467	-0.001461	-0.010997	-0.009213	-0.018469	-0.009700	-0.021167	-0.012965	-0.013212	-0.004702	...
2020-04-17	0.067471	0.024070	0.017817	0.004617	0.003705	0.005991	0.010135	0.010448	0.009926	-0.004724	...
...	...	...	...	...	...	...	...	...	...	...	...

Step 3: Get Total Return Idx.

$$Total\ Return\ Index = \sum W_{stock\_id} \times R_{stock\_id}$$



Step 4: Deal with Unlisted Components

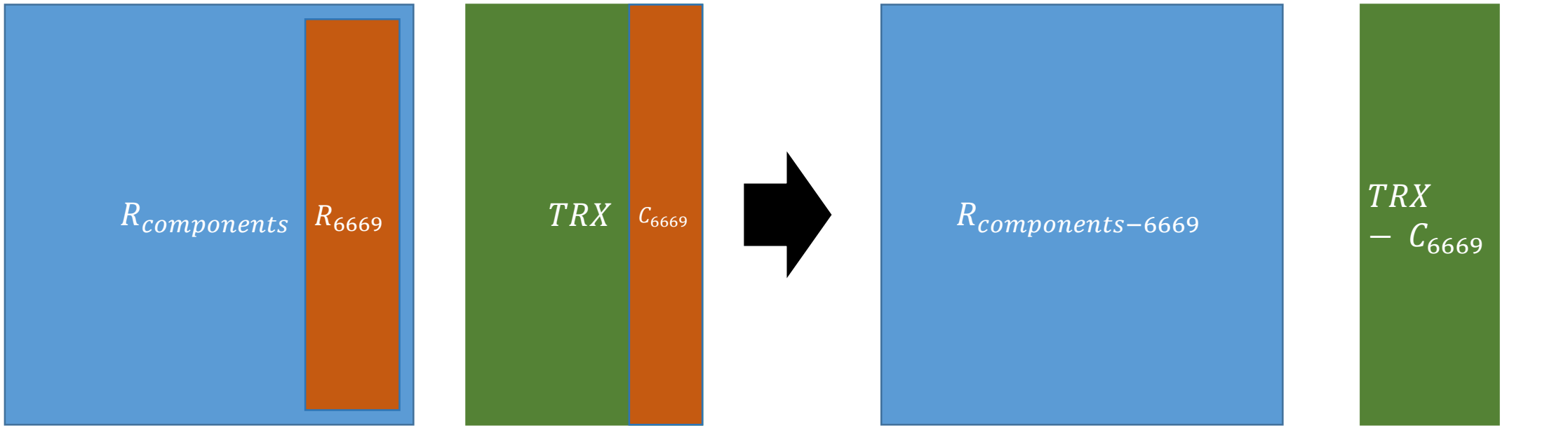


股票代號	3231	3034	490	6669	2301	2357	
股票名稱	緯創	聯詠	遠傳	緯穎	光寶科	華碩	台新
20190520大盤比重(%)	0.2223	0.3746	0.7545		0.3127	0.5355	0
20190521大盤比重(%)	0.2223	0.3746	0.7545		0.3127	0.5355	0
20190522大盤比重(%)	0.2223	0.3746	0.7545		0.3127	0.5355	0
20190523大盤比重(%)	0.2223	0.3746	0.7545		0.3127	0.5355	0
20190524大盤比重(%)	0.2223	0.3746	0.7545		0.3127	0.5355	0
20190527大盤比重(%)	0.2223	0.3746	0.7545		0.3127	0.5355	0
20190528大盤比重(%)	0.2223	0.3746	0.7545		0.3127	0.5355	0
20190529大盤比重(%)	0.2223	0.3746	0.7545		0.3127	0.5355	0
20190530大盤比重(%)	0.2223	0.3746	0.7545		0.3127	0.5355	0
20190531大盤比重(%)	0.2036	0.3246	0.8146	0.1817	0.3376	0.5133	0
20190603大盤比重(%)	0.2036	0.3246	0.8146	0.1817	0.3376	0.5133	0
20190604大盤比重(%)	0.2036	0.3246	0.8146	0.1817	0.3376	0.5133	0
20190605大盤比重(%)	0.2036	0.3246	0.8146	0.1817	0.3376	0.5133	0
20190606大盤比重(%)	0.2036	0.3246	0.8146	0.1817	0.3376	0.5133	0

Data Handling – Dealing with Unlisted

Step 1: Deal with Unlisted Components

※$C = Contribution$



Number of Features = 800

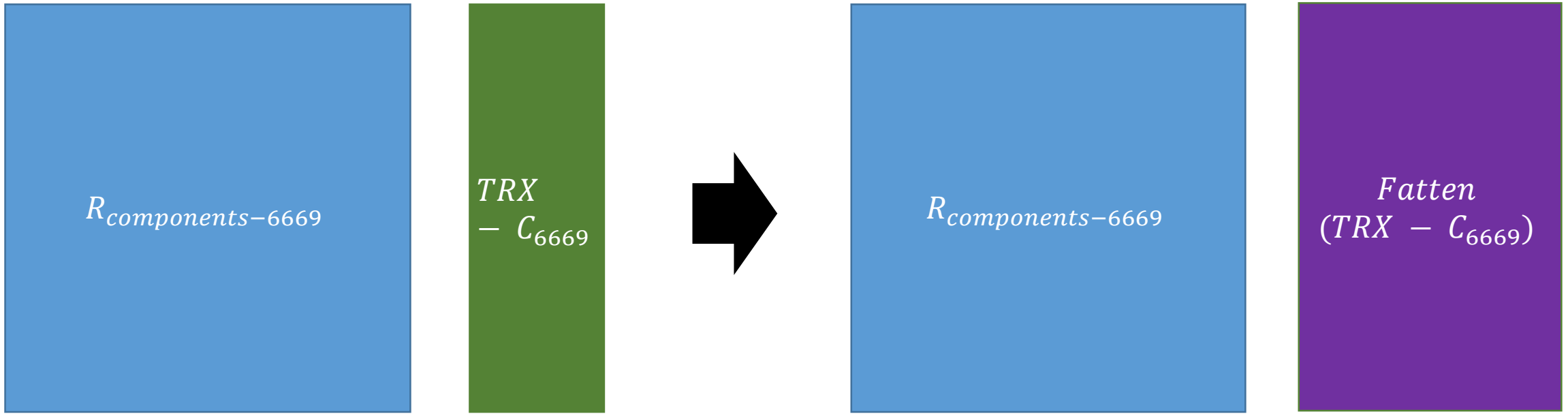
100% of Total Contribution

Number of Features = 799

(100 - W_{6669})% of Total Contribution

Step 1: Deal with Unlisted Components(cont.)

※$C = Contribution$



Number of Features = 799

(100 - W_{6669})% of Total Contribution

Number of Features = 799

100% of Total Contribution

Step 2: Fit and Adjust

$※C = Contribution$

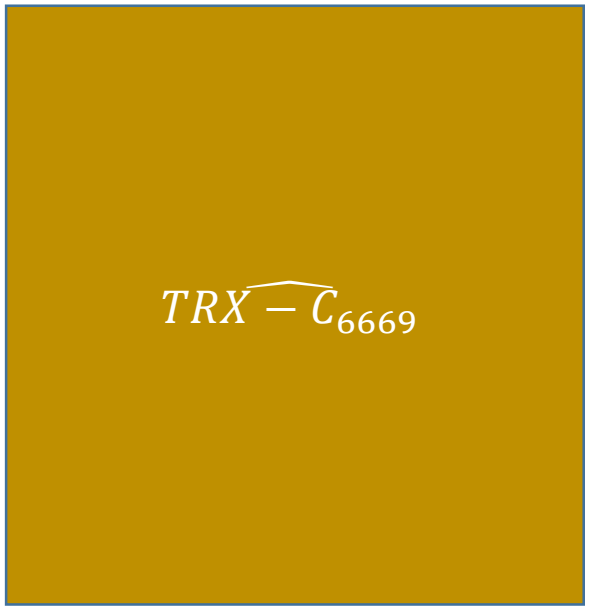
Fit $R_{components-6669}$ ***On*** $Fatten\ (TRX - C_{6669})$ ***To Get*** $\widehat{TRX - C_{6669}}$

Number of Features = 799

100% of Total Contribution

100% of Total Contribution

Step 3: Comparison

$(\mathbf{1} - \boldsymbol{W_{6669}}) \times$  $+\boldsymbol{W_{6669}} \times$  *VS* 

(100 - W_{6669})% of Total Contribution

W_{6669}% of Total Contribution

100% of Total Contribution

Experiment

$$N_{components} \times N_{days} \times Methodology$$

$$N_{components} = [100, 300, 500, 700, 900, 963]$$
$$N_{days} = [100, 200, 300, 400, 500]$$
$$Methodology = Linear\ Reg, Lasso, CGR$$

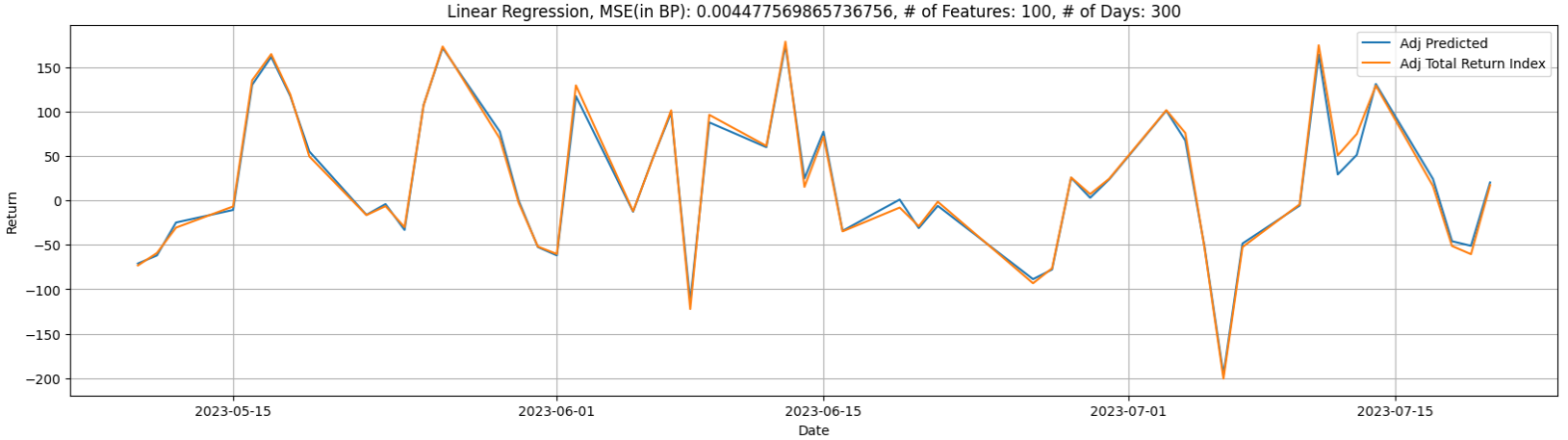
Method 1: *Linear Reg.*

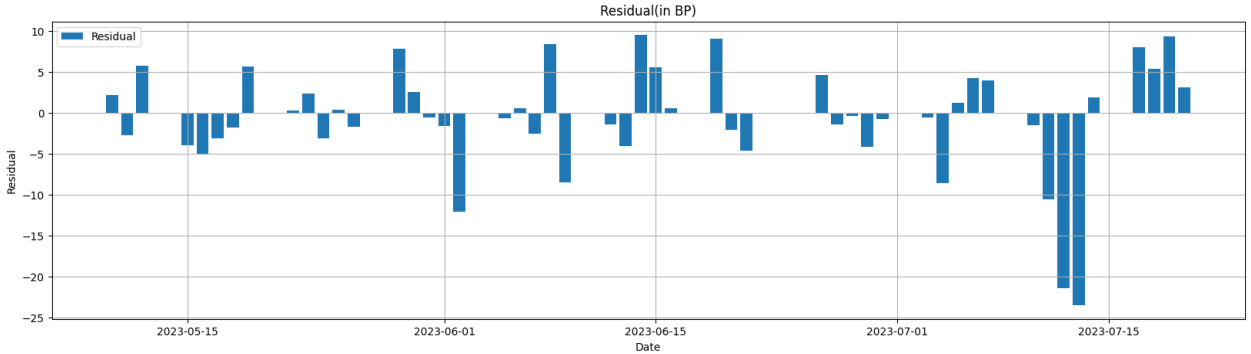
Numerical Results:

Mean Squared Error: $0.0045(basis\ points)$
Number of Features: 100
Number of Days: 300

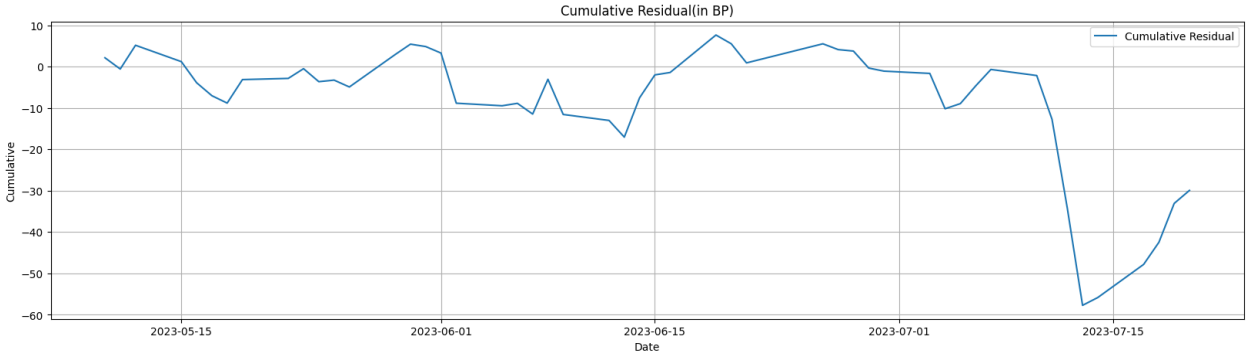
	Weights
2330	0.337531
2317	0.038400
2454	0.033472
2412	0.025185
2308	0.018970

Numerical Results:





Residuals.



Cum. Residuals.

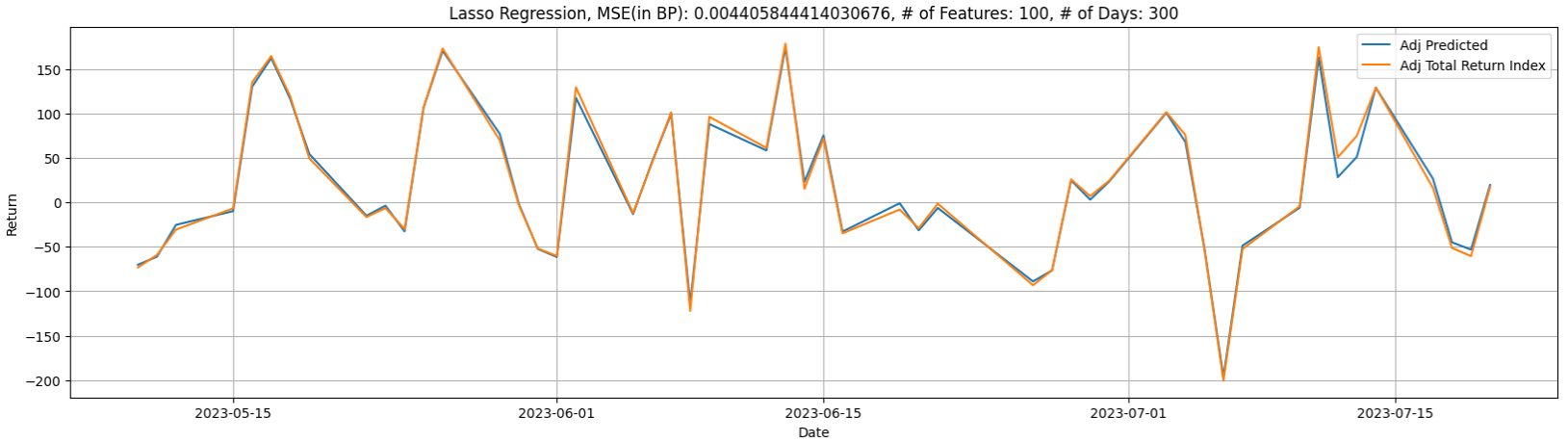
Method 2: *Lasso Reg.*

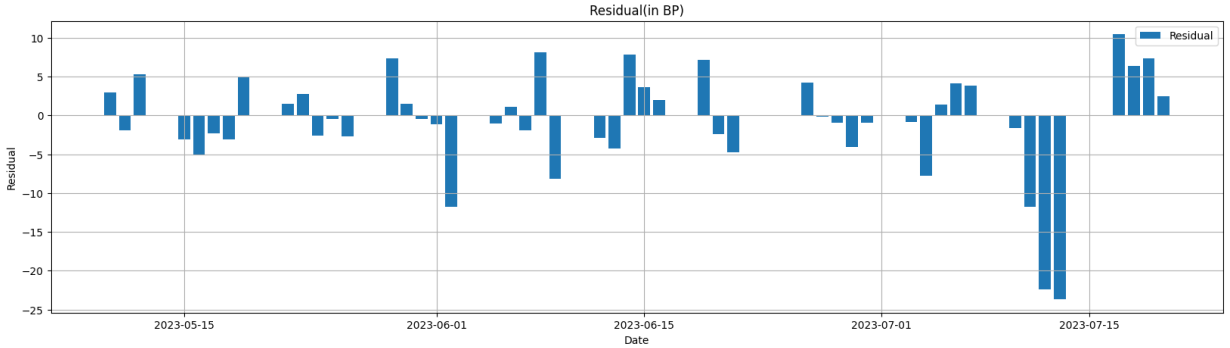
Numerical Results:

Best λ: 10^{-7}
Mean Squared Error: 0.0044 $(basis\ points)$
Number of Features: 100
Number of Days: 300

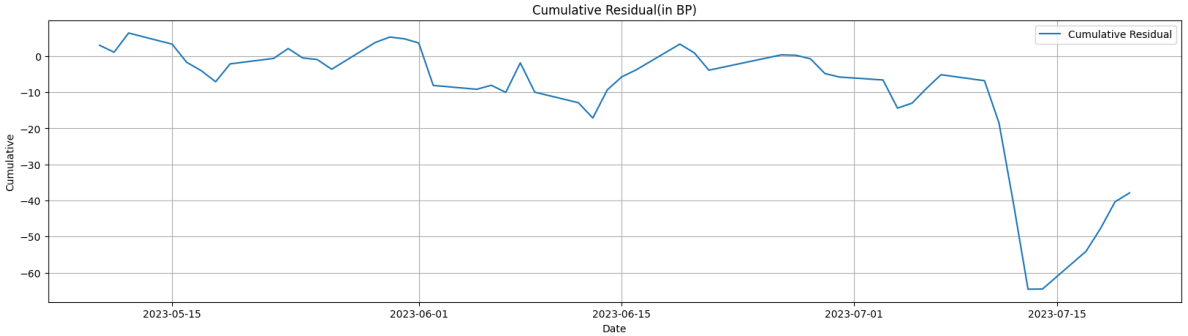
	Weights
2330	0.337202
2317	0.037894
2454	0.033400
2412	0.024280
2308	0.018881

Numerical Results:





Residuals.



Cum. Residuals.

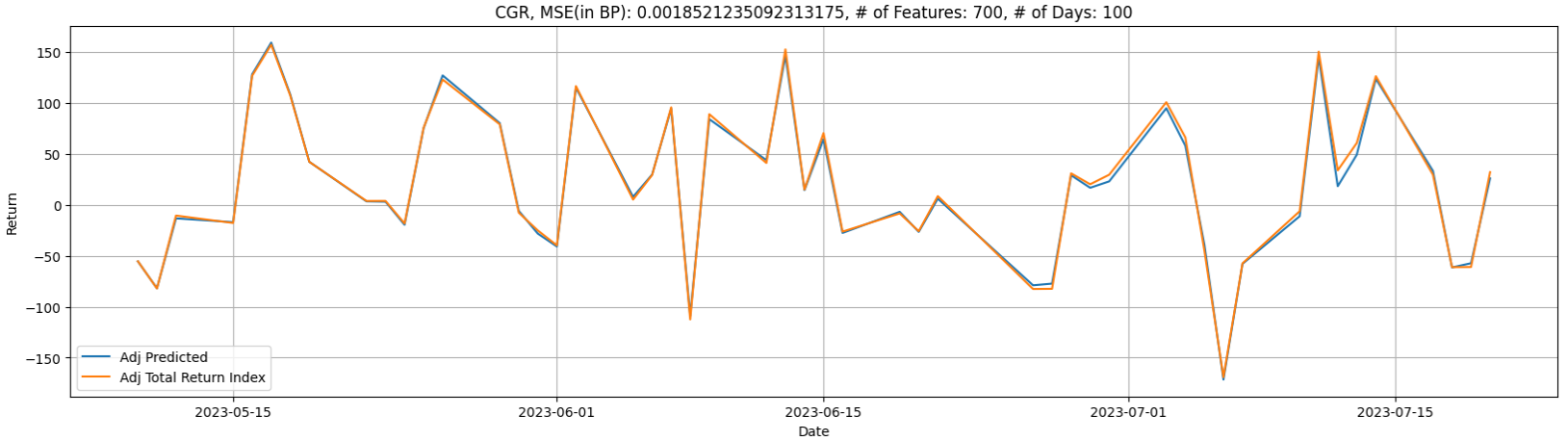
Method 3: Conjugate Gradient Research

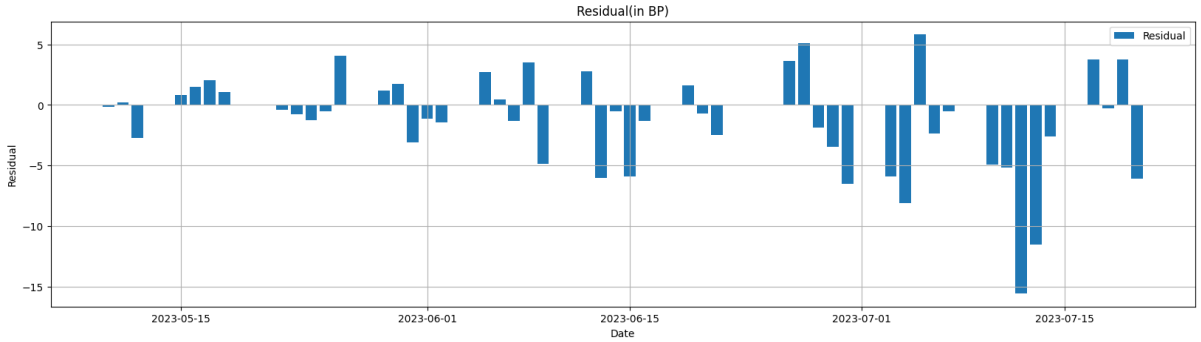
Numerical Results:

Mean Squared Error: $0.0018\ (basis\ points)$
Number of Features: 700
Number of Days: 100

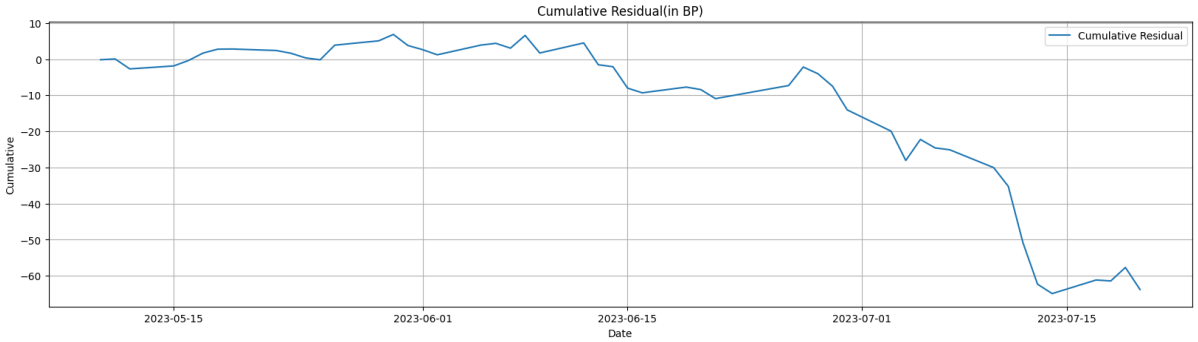
	Weights
2330	0.276540
2317	0.030873
2454	0.024497
2412	0.019524
2308	0.016451

Numerical Results:





Residuals.



Cum. Residuals.

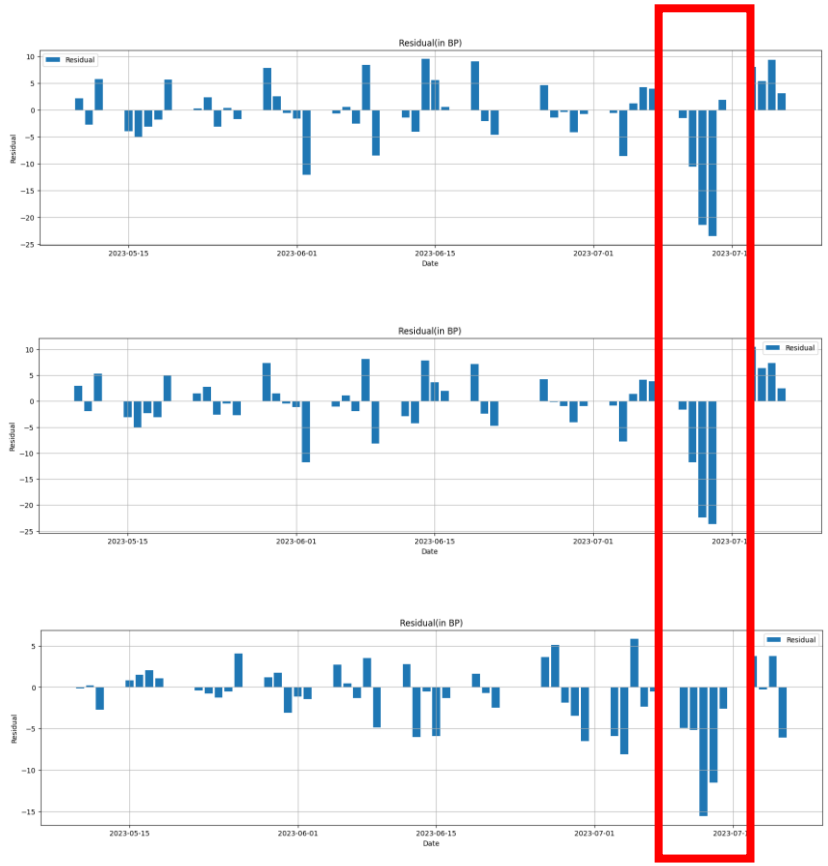
Inspection

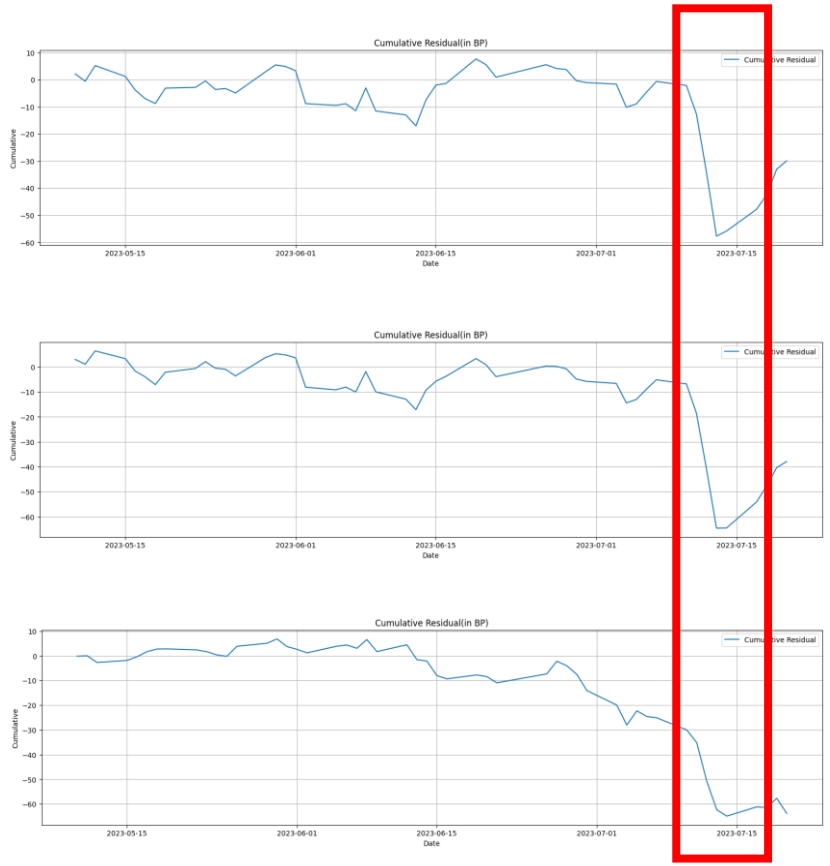
Anomalies in Residuals:

Linear Regr.:

Lasso:

CGR:



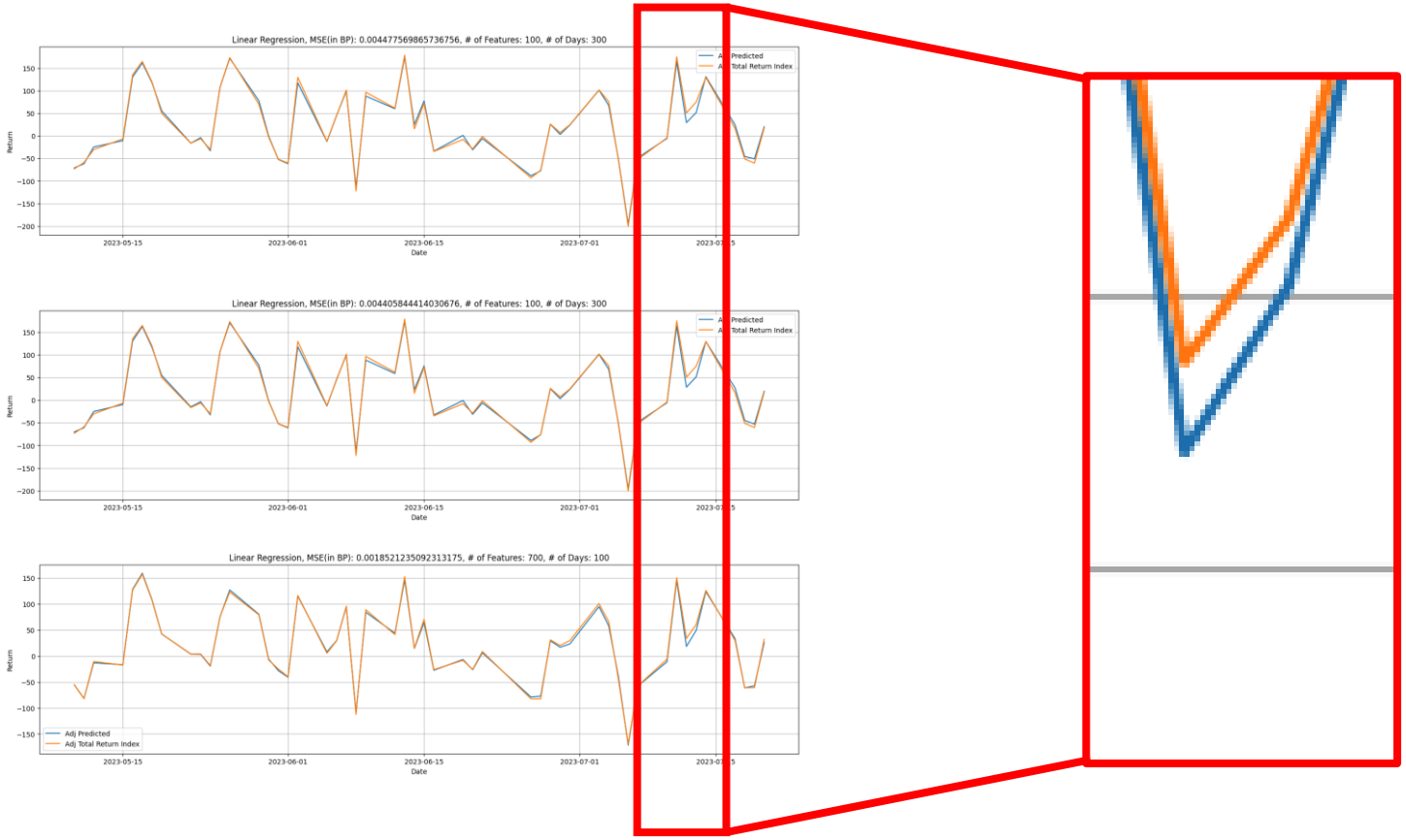


Anomalies in Residuals:

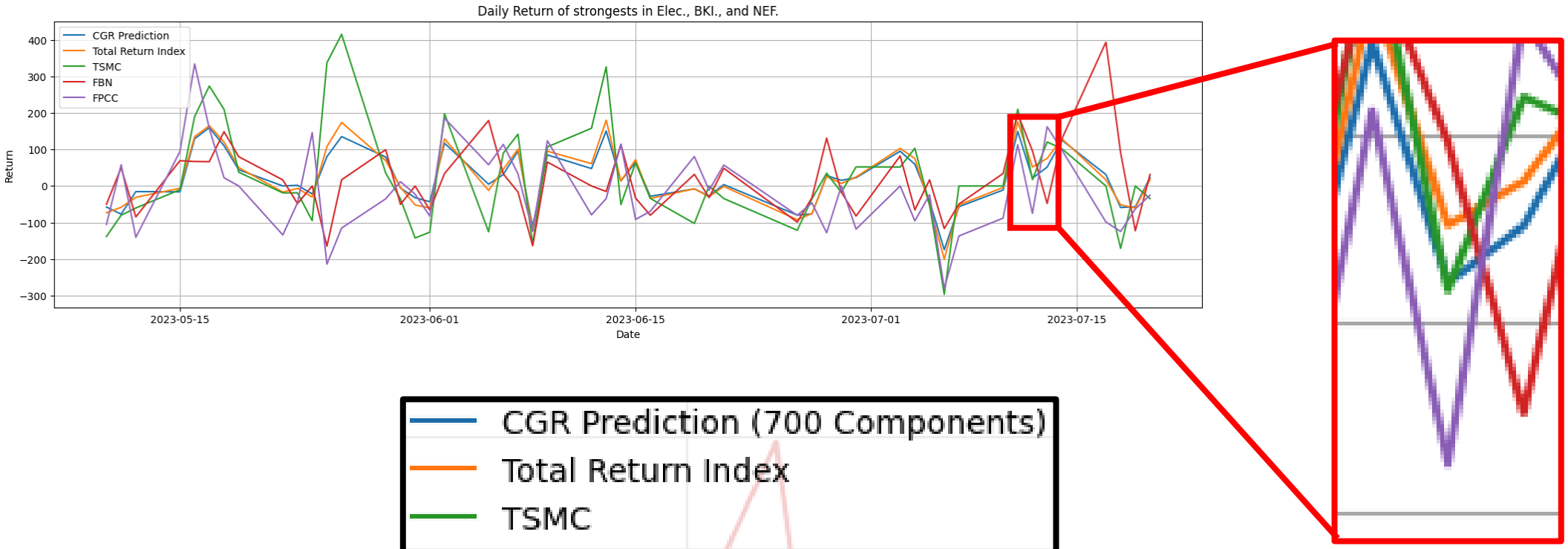
Linear Regr.:

Lasso:

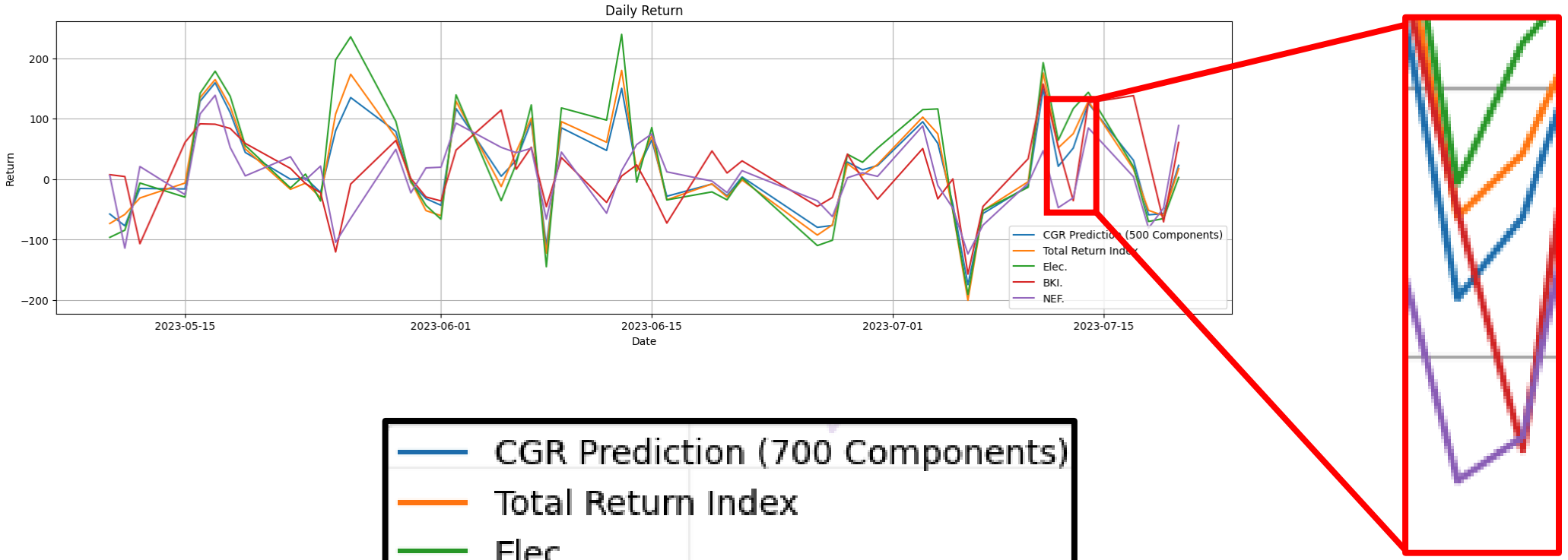
CGR:



Look into TSMC, FBN and FPCC:



Look into Elec, BKI and NEF idx:



Big Mistake!!!

Our model: Heavily depends on weighted stocks (e.g. Elec.).

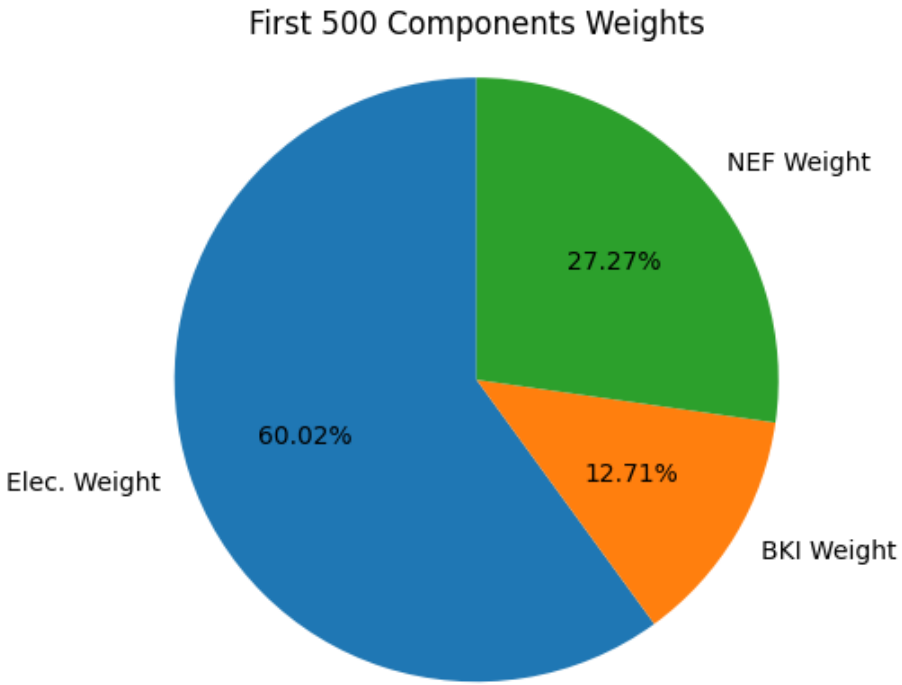
Reality: Stocks other than Elec. have majority in amount.

Choose different way of feature selection!!!!

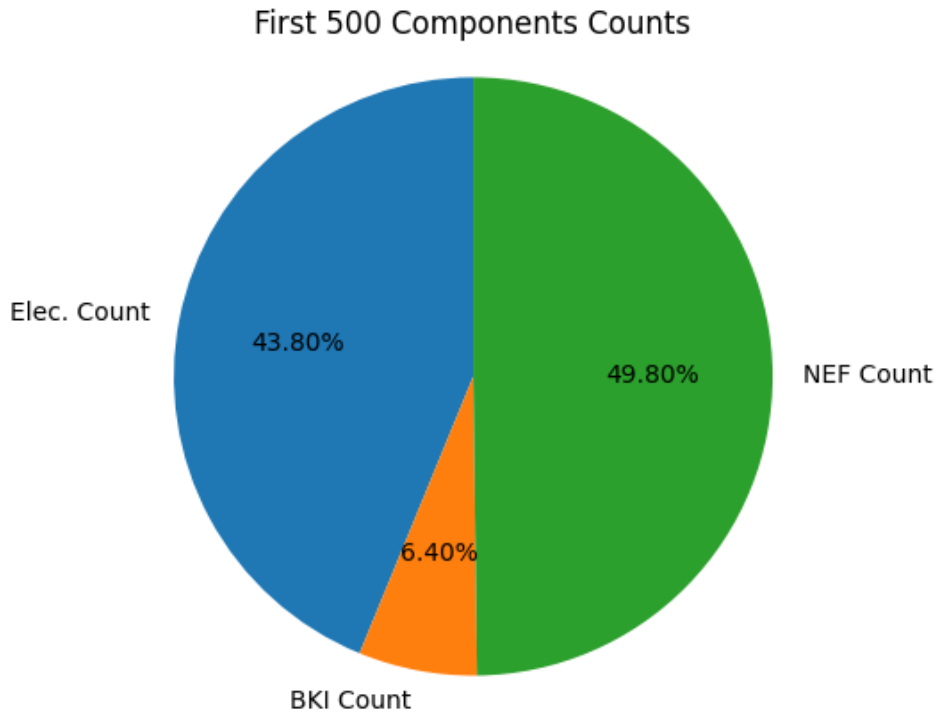
Selection of Components

Selection Method 1: Replicate the Shape

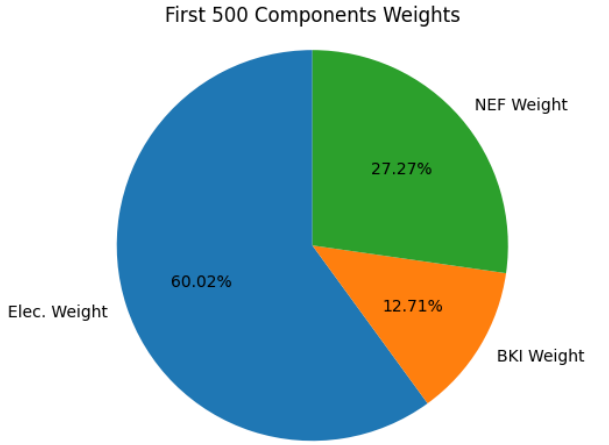
Chosen by Weight VS Chosen by Amount:

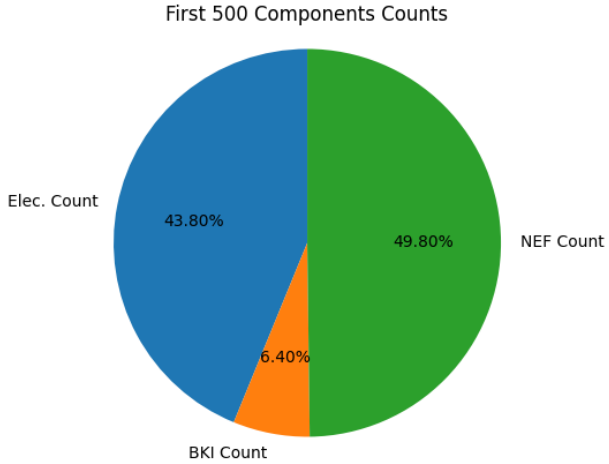


≠

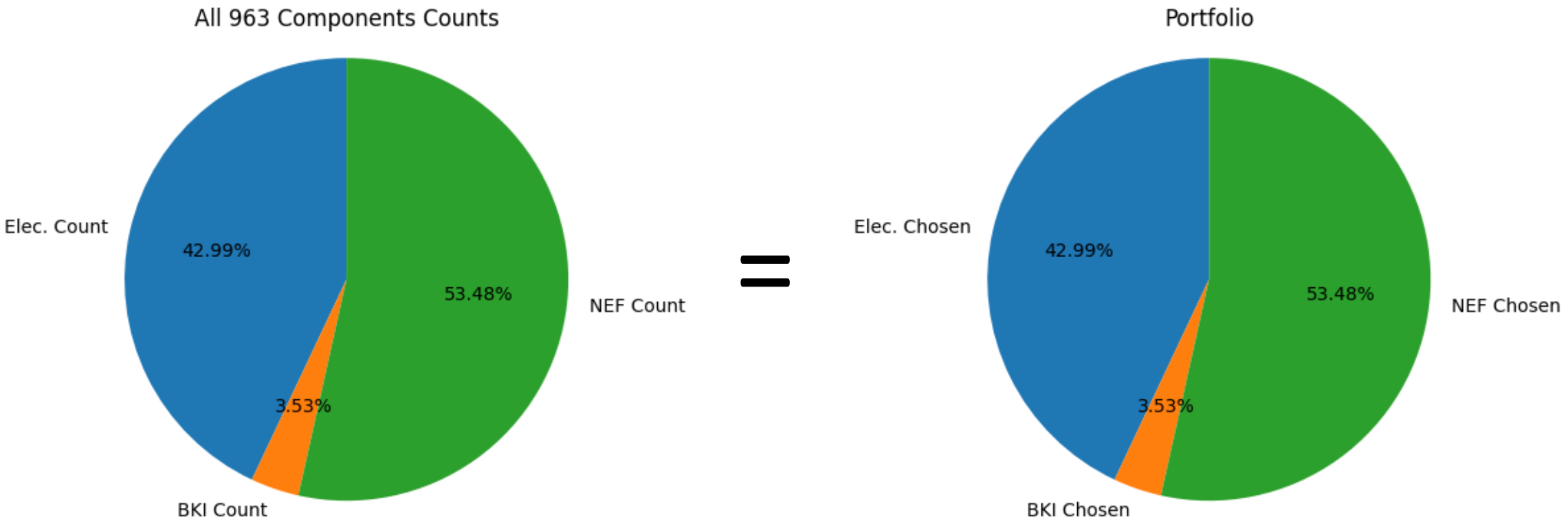


Chosen by Weight VS Chosen by Amount:





New replication approaches!

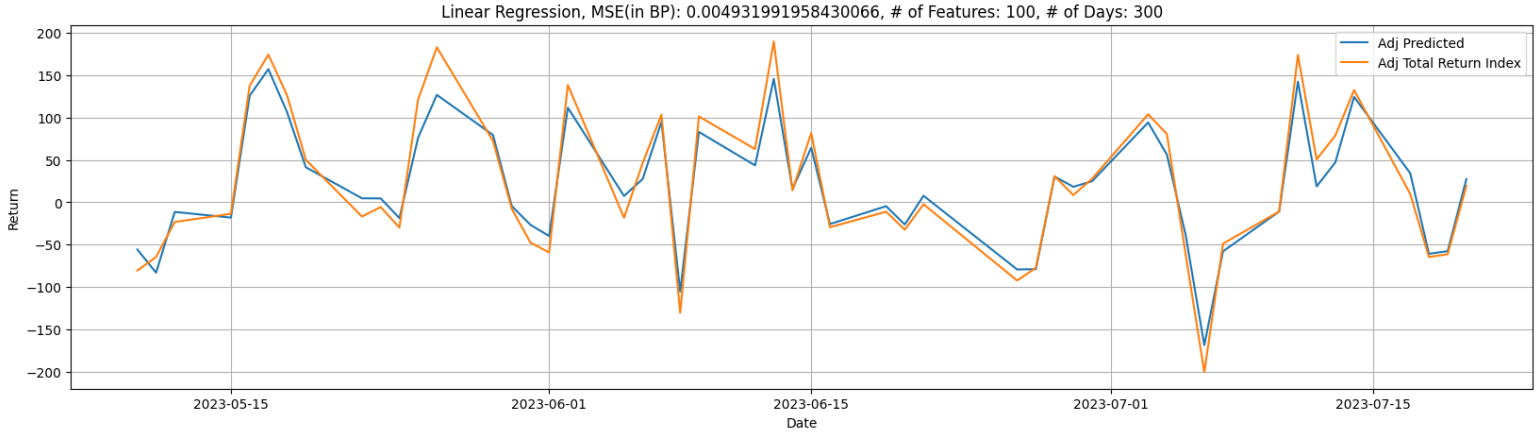


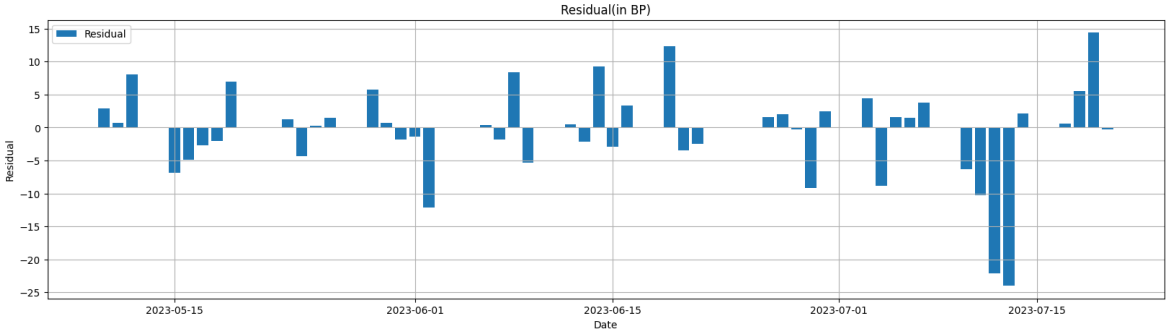
Numerical Results (Linear Regr.):

Mean Squared Error: 0.0049 $(basis\ points)$
Number of Features: 100
Number of Days: 300

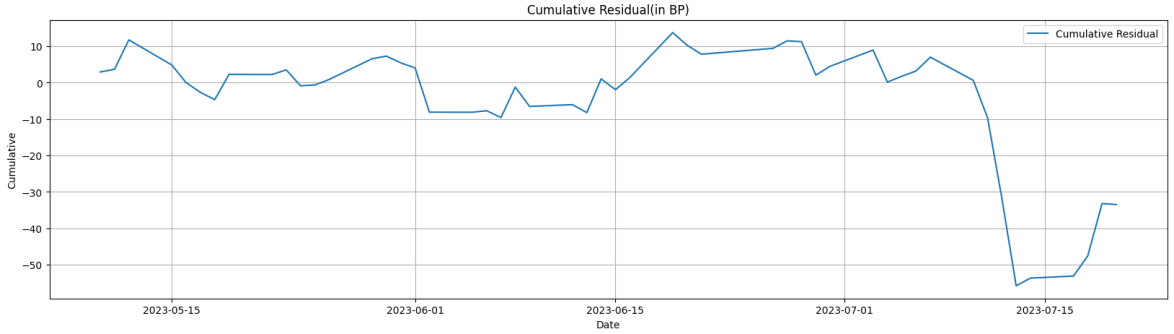
	Weights
2330	0.363868
2317	0.043070
2454	0.036380
2412	0.028987
2308	0.021329

Numerical Results (Linear Regr.):





Residuals.



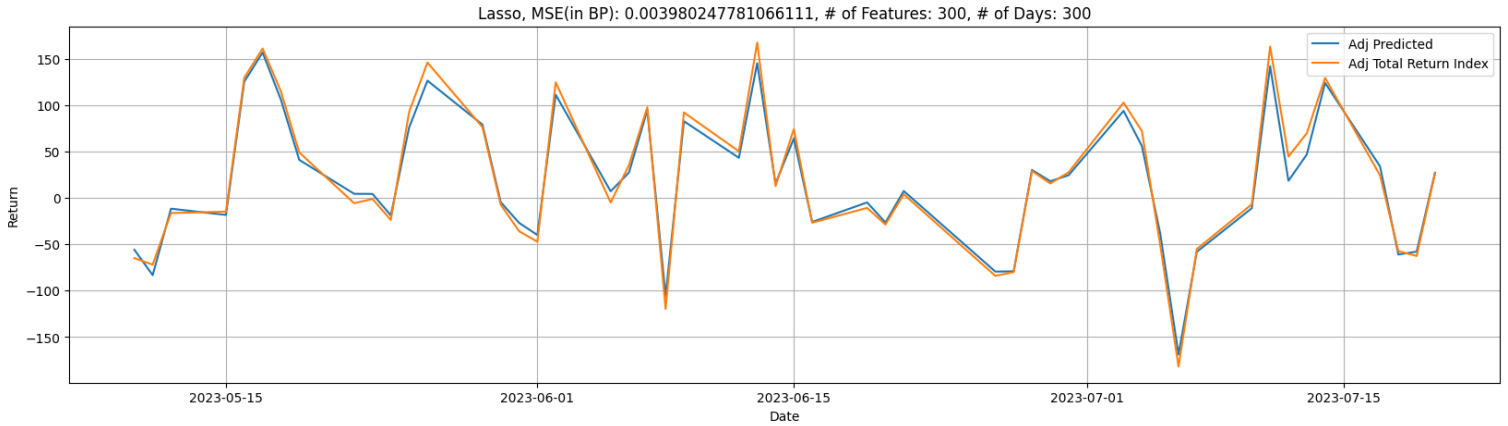
Cum. Residuals.

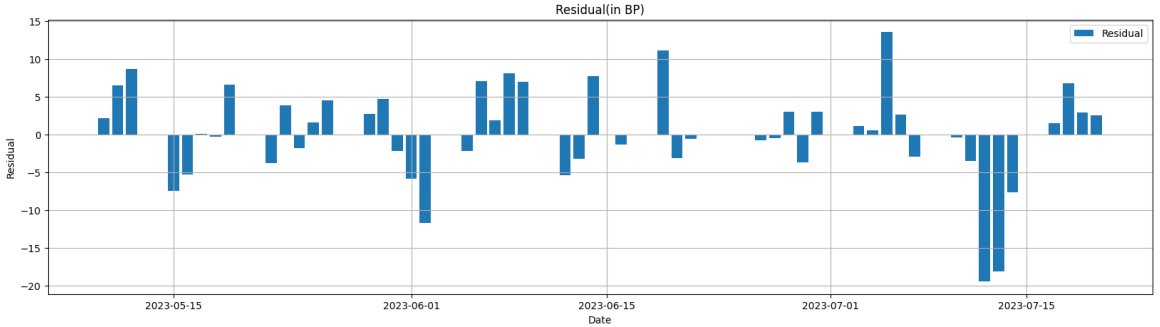
Numerical Results (Lasso):

Mean Squared Error: $0.004(basis\ points)$
Number of Features: 300
Number of Days: 300

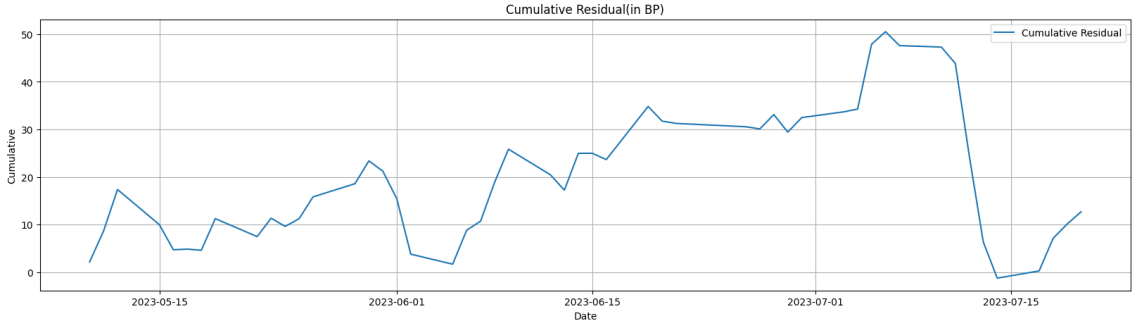
	Weights
2330	0.297671
2317	0.029504
2454	0.027582
2412	0.023294
2308	0.017023

Numerical Results (Lasso):





Residuals.



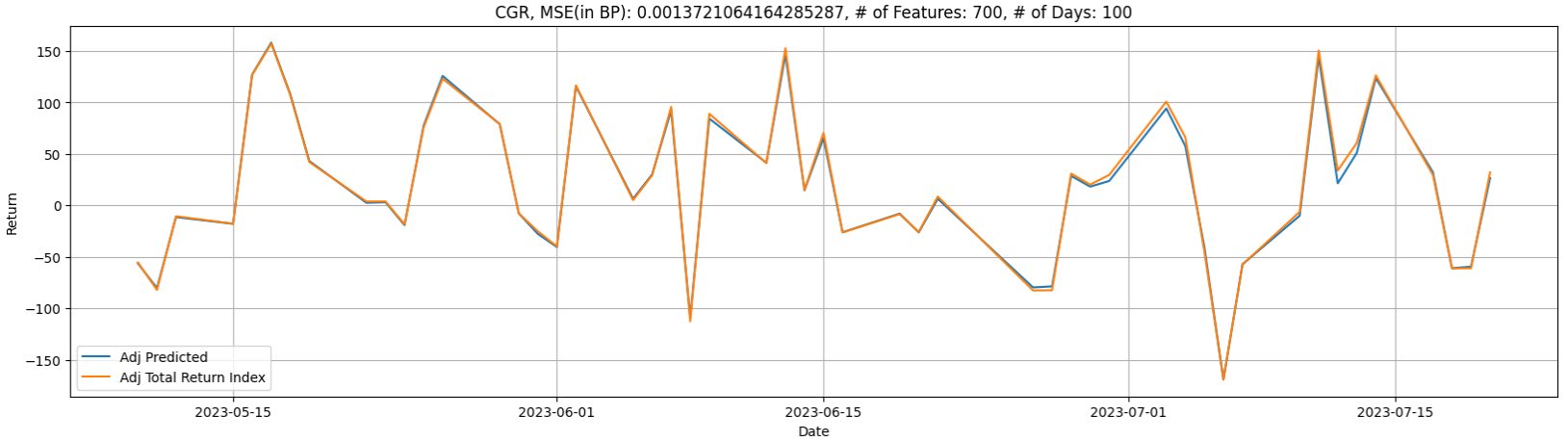
Cum. Residuals.

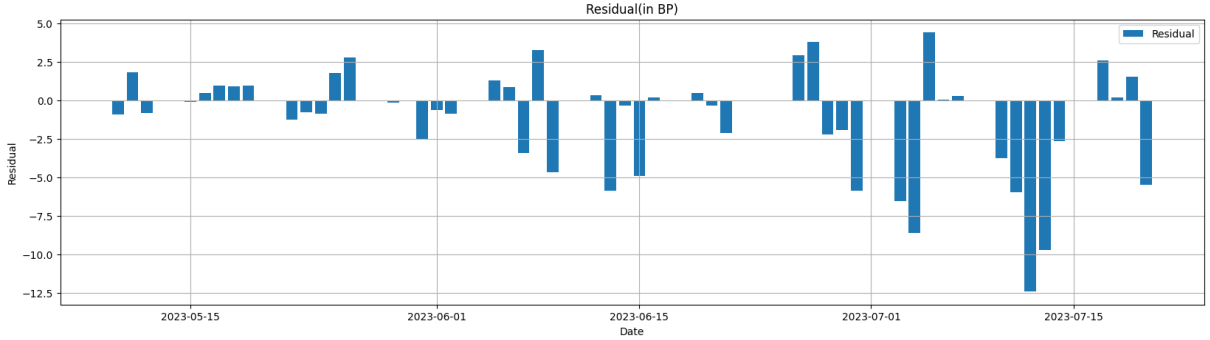
Numerical Results (CGR):

Mean Squared Error: $0.0014\ (basis\ points)$
Number of Features: 700
Number of Days: 100

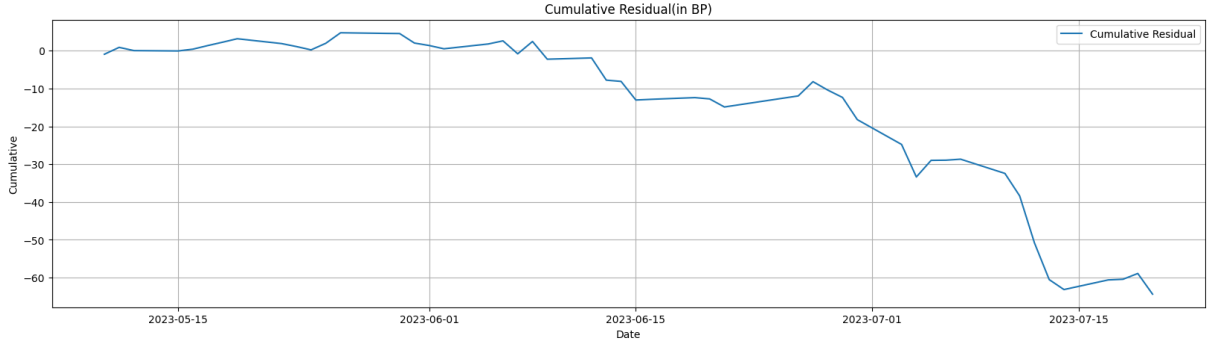
	Weights
2330	0.272726
2317	0.030350
2454	0.022286
2412	0.020639
2308	0.016325

Numerical Results (CGR):





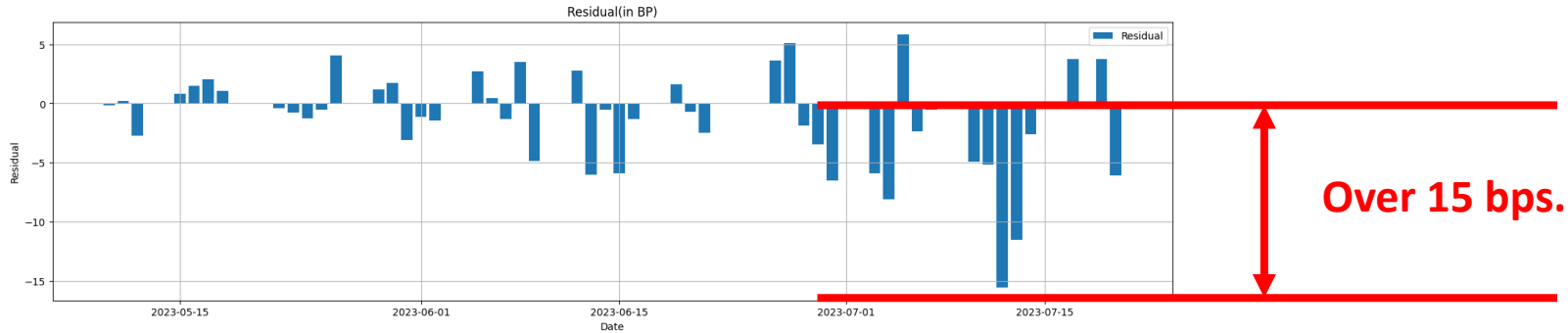
Residuals.



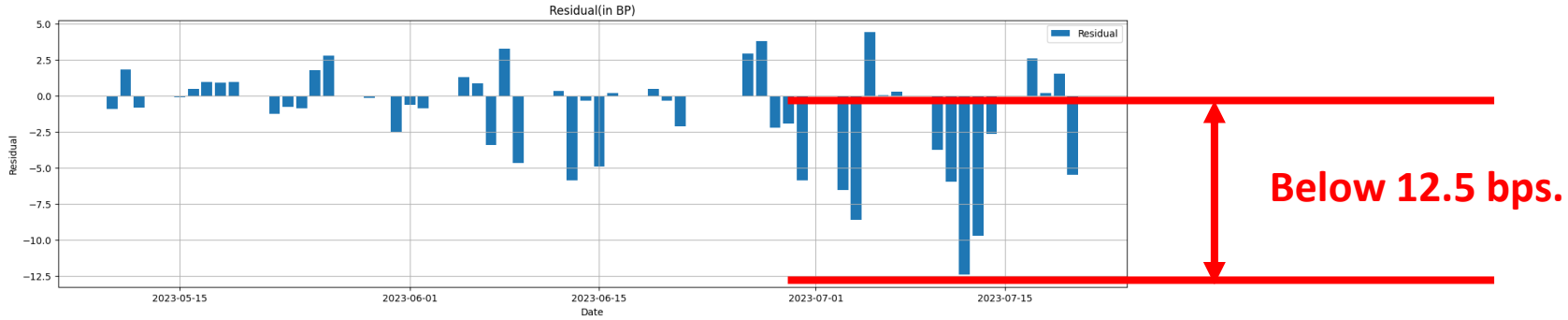
Cum. Residuals.

Comparison:

Original CGR:

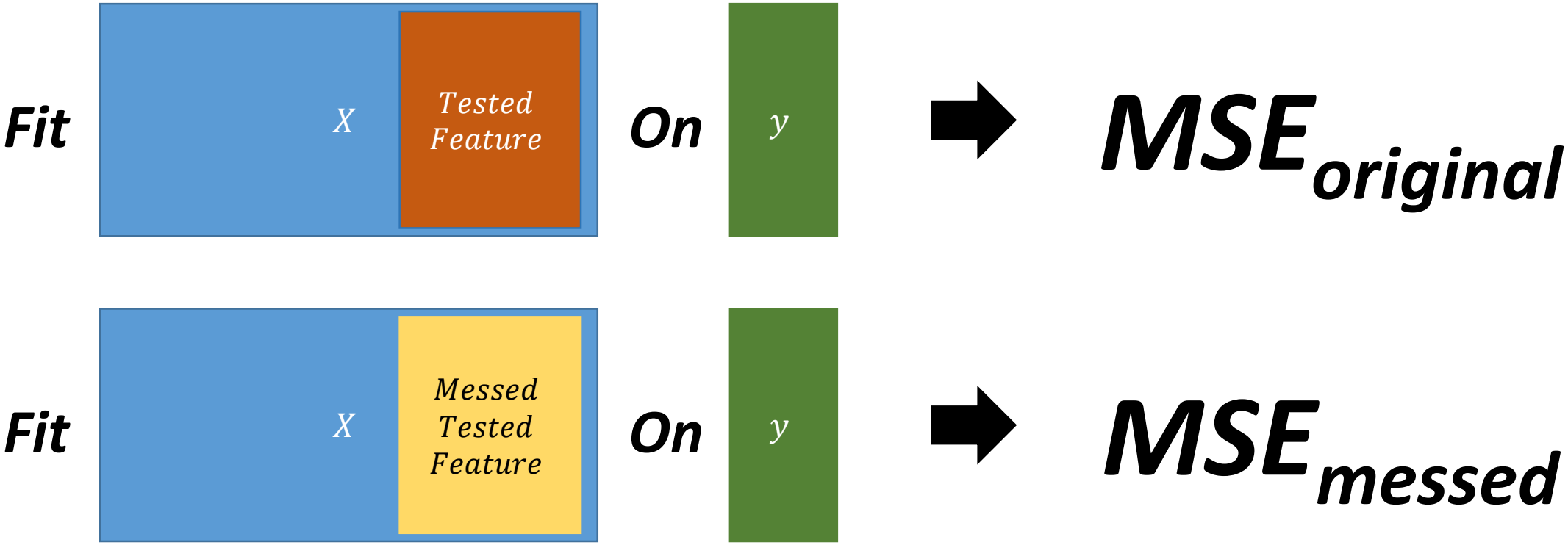


New approach CGR:



Selection Method 2:
Random Forest Feature Importance

Random Forest Feature Importance:



$$Importance = MSE_{messed} - MSE_{original}$$

Importance of Components:

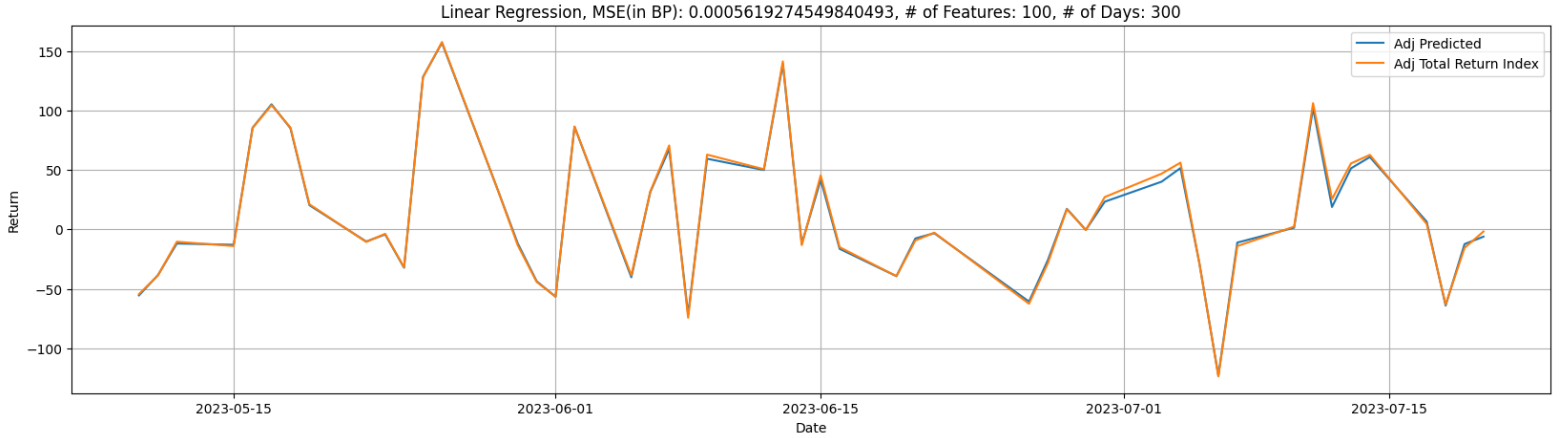
	feature	name	importance
0	2330	台積電	0.497079
95	3653	健策	0.020009
300	3413	京鼎	0.019690
73	6770	力積電	0.017479
263	2338	光罩	0.017137
19	5880	合庫金	0.012226
530	8110	華東	0.011321
624	4952	凌通	0.009434

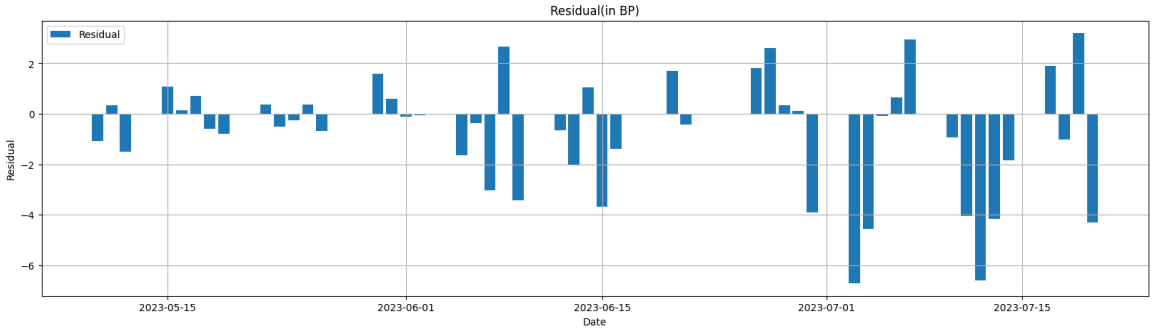
Numerical Results (Linear Regr.):

Mean Squared Error: 0.00056 (*basis points*)
Number of Features: 100
Number of Days: 300

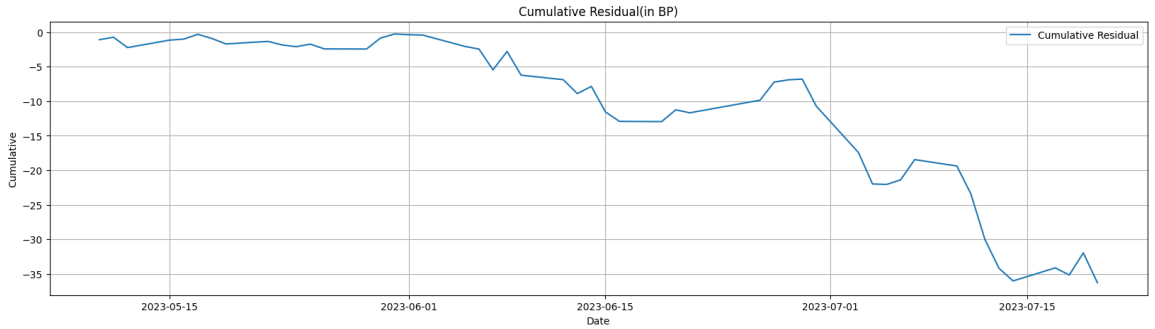
	Weights
2330	0.363455
3653	0.001457
3413	0.001072
2338	0.000688
5880	0.010240

Numerical Results (Linear Regr.):





Residuals.



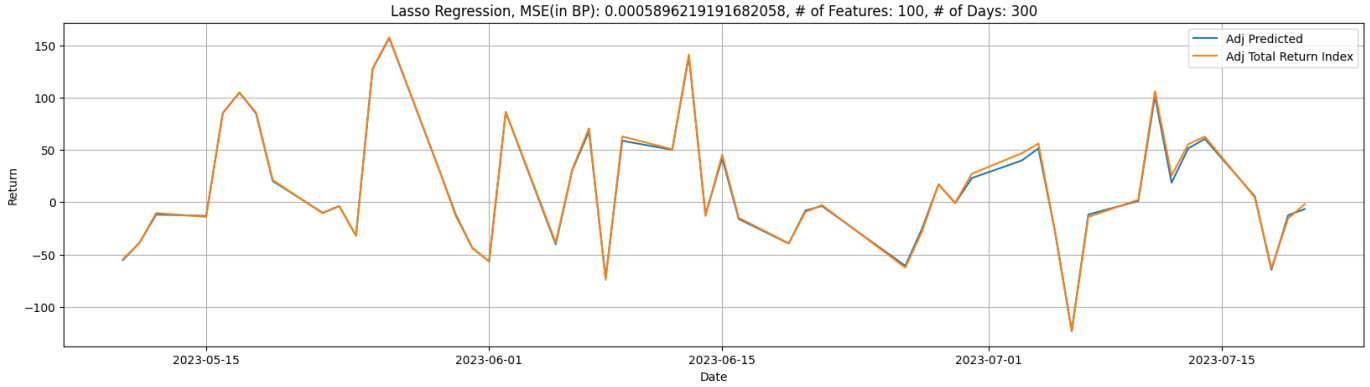
Cum. Residuals.

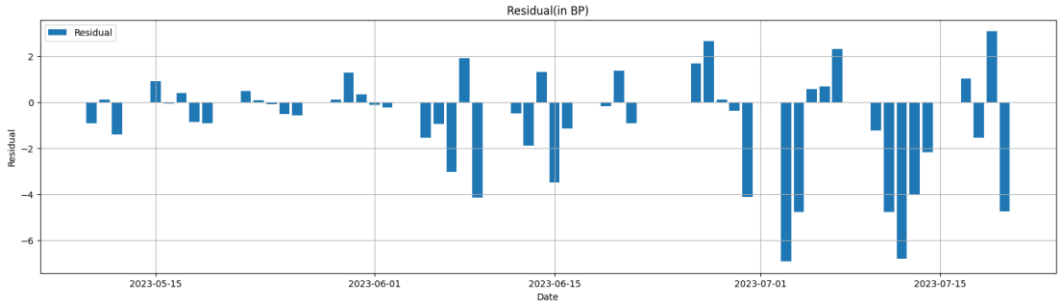
Numerical Results (Lasso Regr.):

Mean Squared Error: $0.00059\ (basis\ points)$
Number of Features: 100
Number of Days: 300

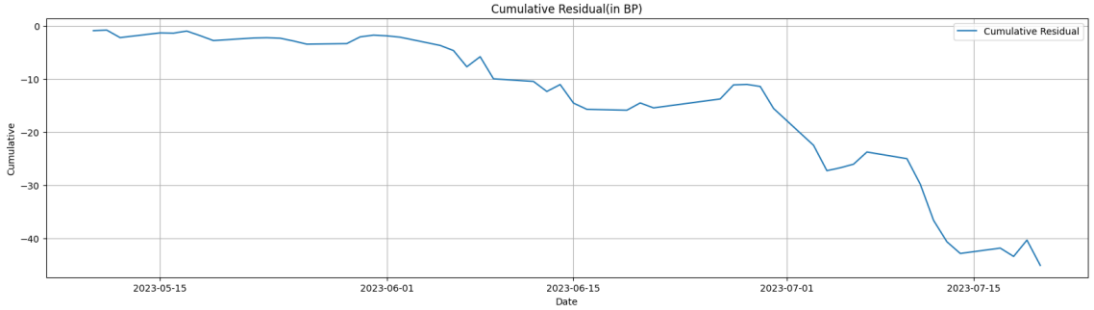
	Weights
2330	0.363411
3653	0.001474
3413	0.001028
2338	0.000656
5880	0.010111

Numerical Results (Lasso Regr.):



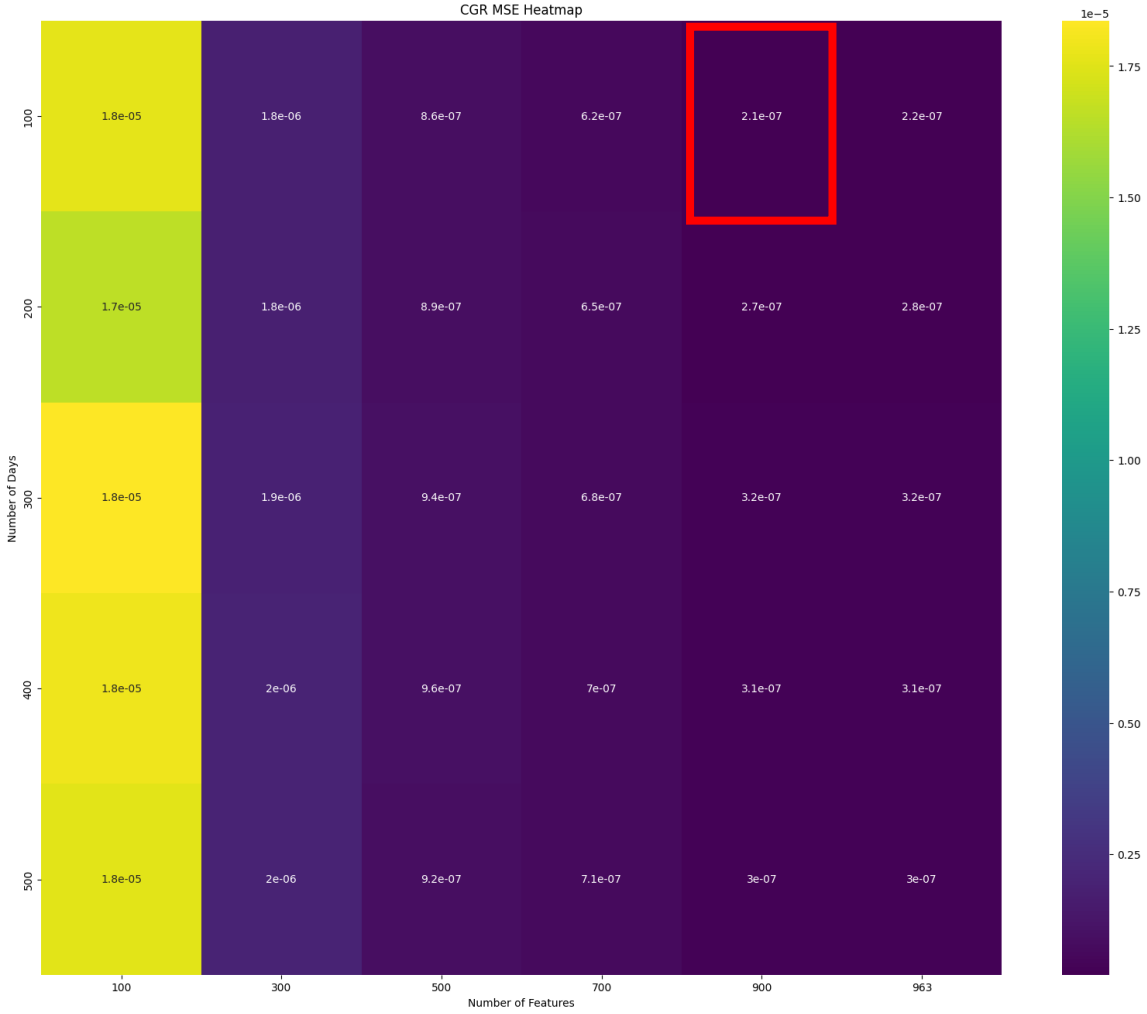


Residuals.



Cum. Residuals.

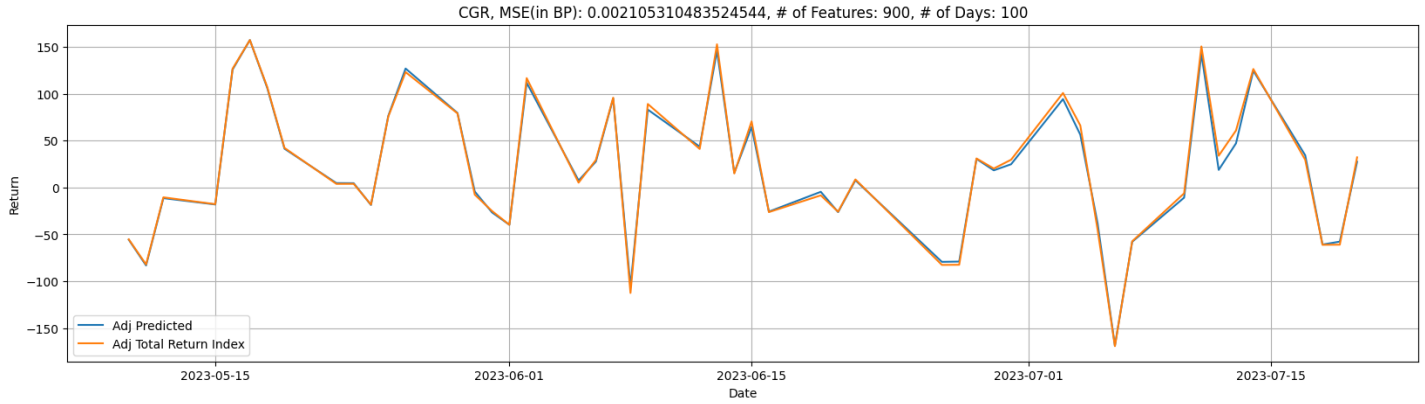
Numerical Results (CGR):

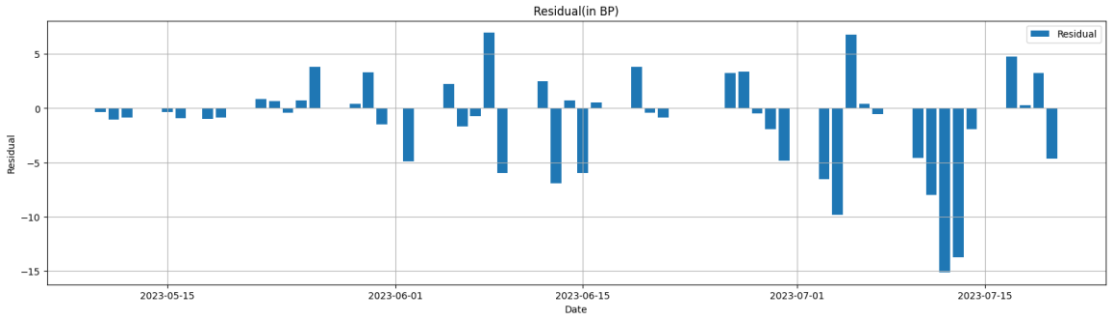


Mean Squared Error: $0.002(basis\ points)$
Number of Features: 900
Number of Days: 100

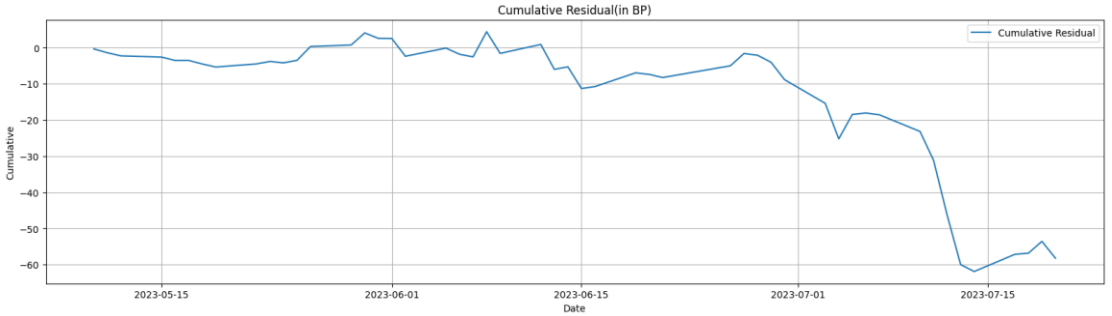
	Weights
2330	0.274101
3653	0.001292
3413	0.000464
2338	0.000597
6770	0.002949

Numerical Results (CGR):





Residuals.



Cum. Residuals.

Selection Method 3:
Feature Importance with Shape

Importance of Components of each Division:

	Feature	Name	Importance
0	2330	台積電	0.717298
313	6695	芯鼎	0.014532
105	6213	聯茂	0.013746
282	3356	奇偶	0.007593
334	8104	錸寶	0.006826

Elec. Importance

	Feature	Name	Importance
0	2881	富邦金	0.259498
7	2880	華南金	0.154056
4	5880	合庫金	0.134619
1	2882	國泰金	0.120053
3	2891	中信金	0.083562

BKI. Importance

	Feature	Name	Importance
98	6806	森崴能源	0.234574
443	2032	新鋼	0.086002
162	1718	中纖	0.052010
6	1326	台化	0.035379
383	2038	海光	0.031689

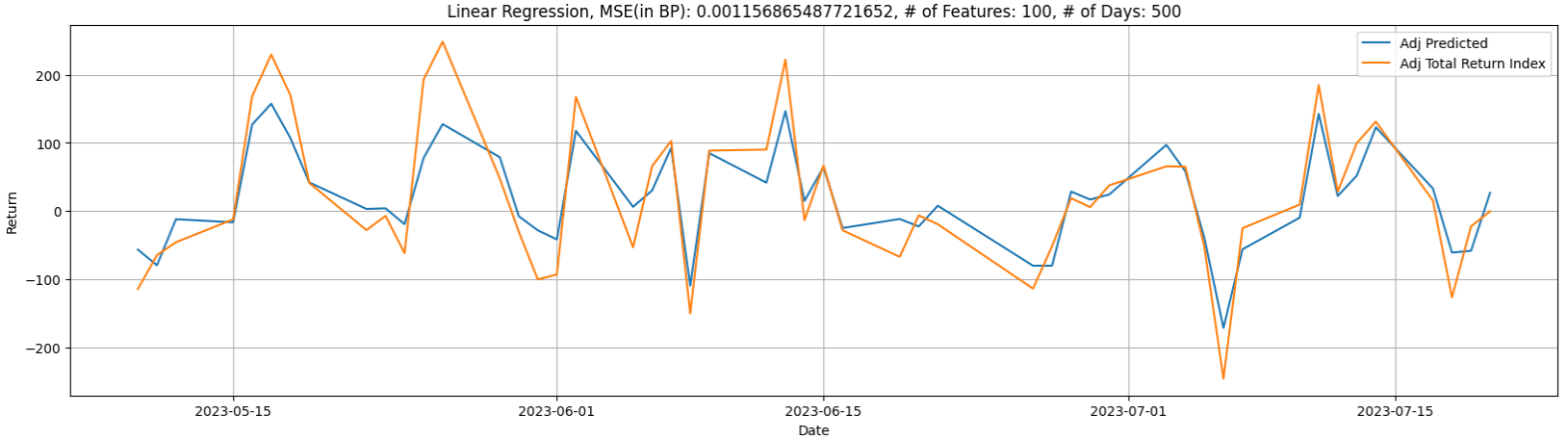
NEF. Importance

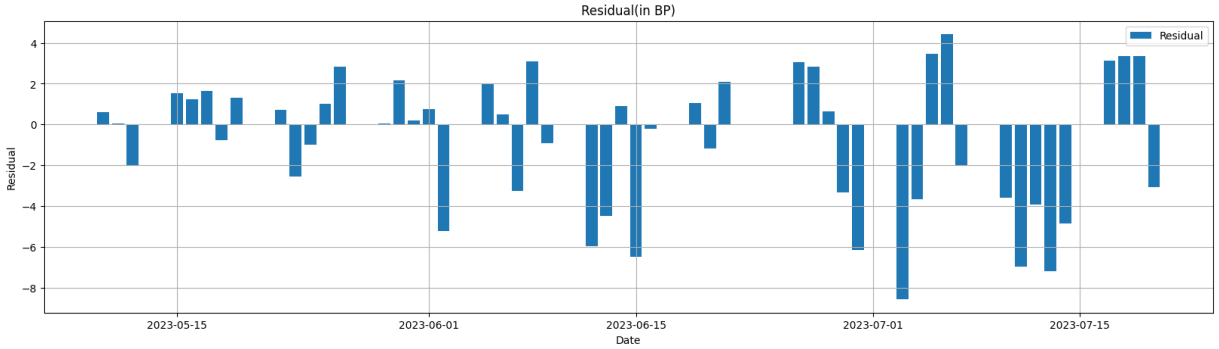
Numerical Results (Linear Regr.):

Mean Squared Error: $0.0012 (basis\ points)$
Number of Features: 100
Number of Days: 500

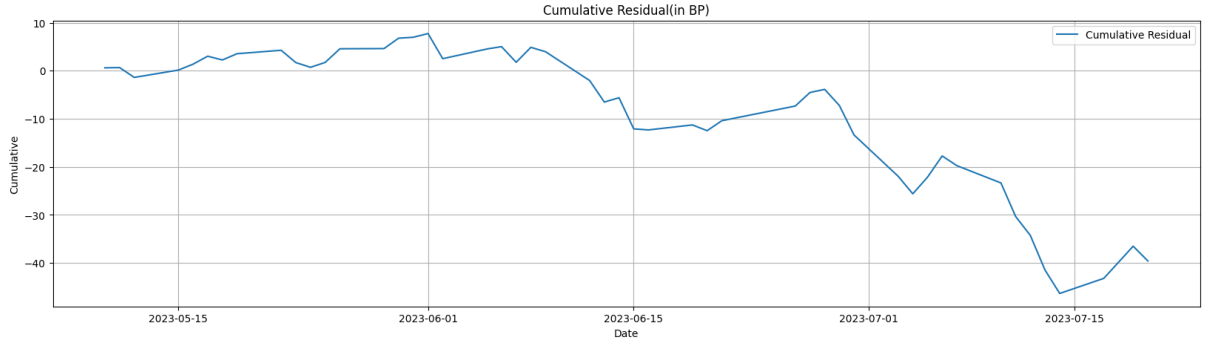
	Weights
2330	0.635407
6213	0.000752
3356	-0.000684
8104	0.001446
6533	0.001151

Numerical Results (Linear Regr.):





Residuals.



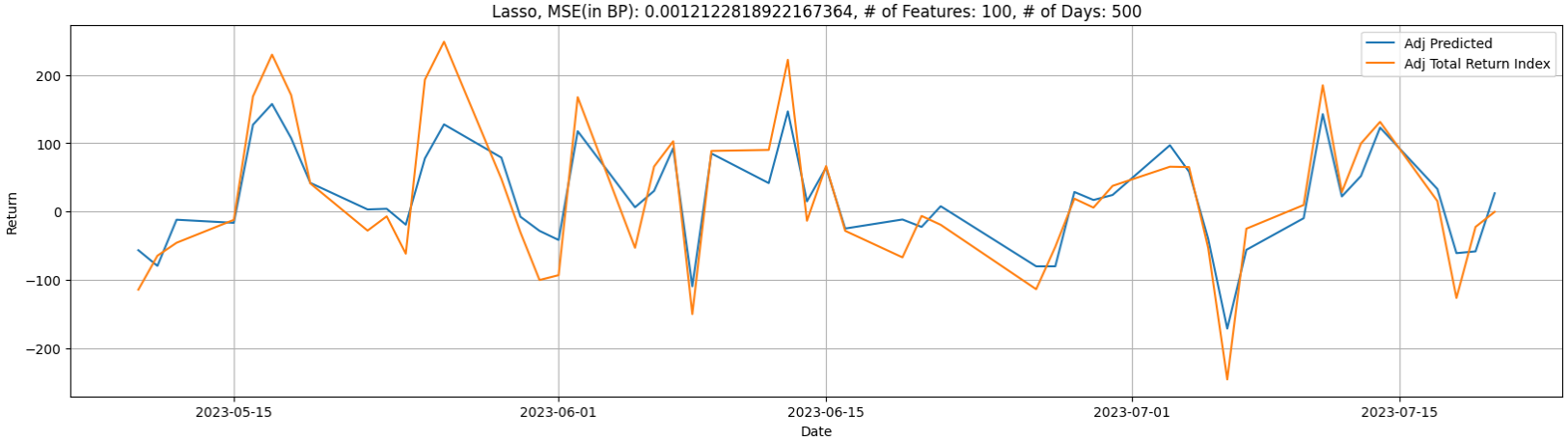
Cum. Residuals.

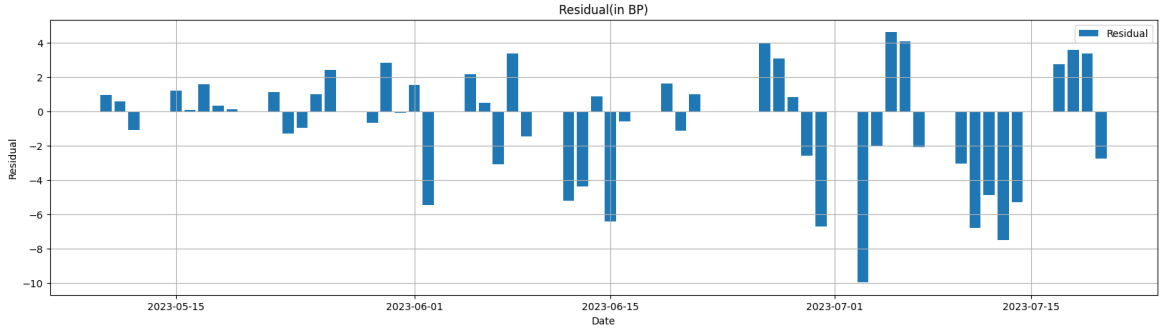
Numerical Results (Lasso Regr.):

Mean Squared Error: $0.0012\ (basis\ points)$
Number of Features: 100
Number of Days: 500

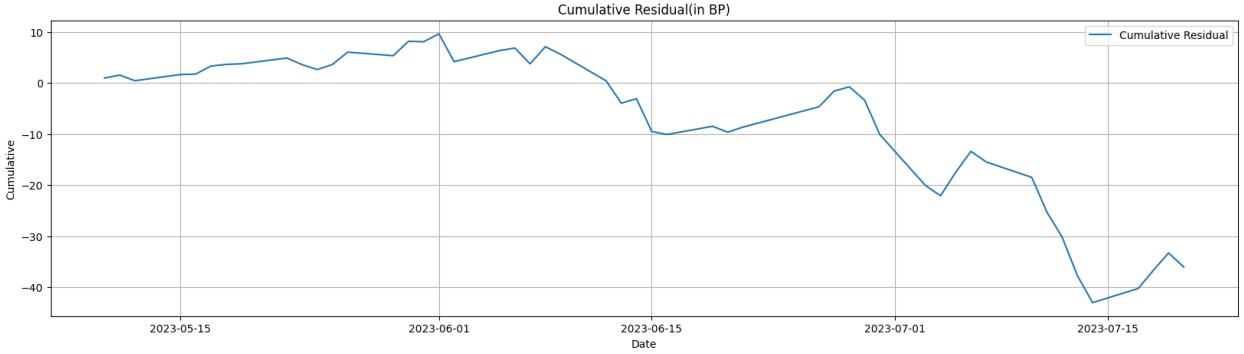
	Weights
2330	0.635450
6213	0.000534
3356	-0.000246
8104	0.001478
6533	0.001272

Numerical Results (Lasso Regr.):





Residuals.



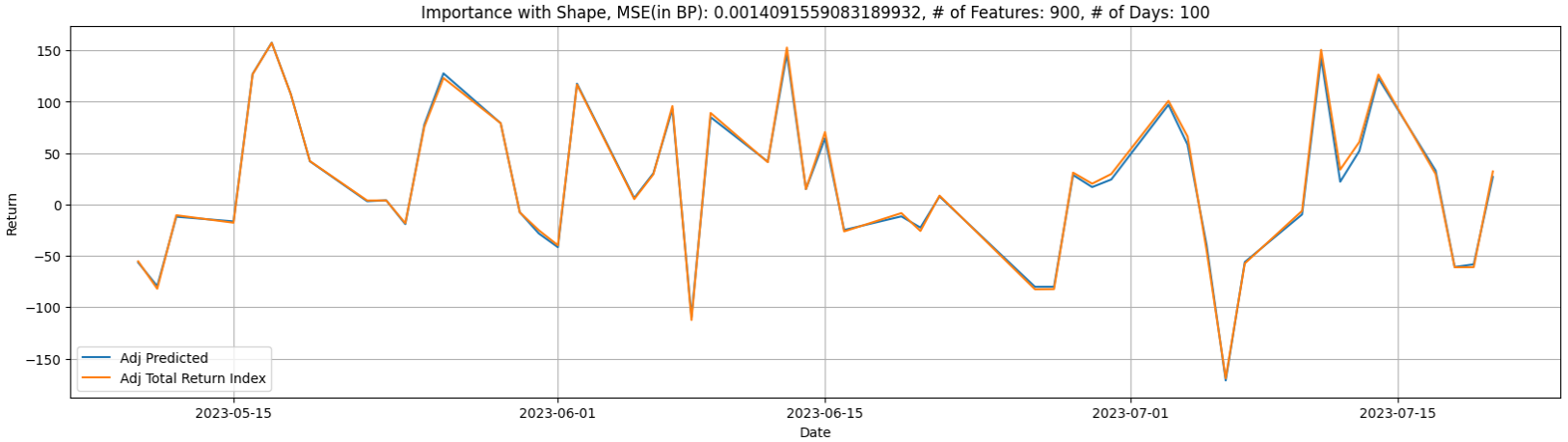
Cum. Residuals.

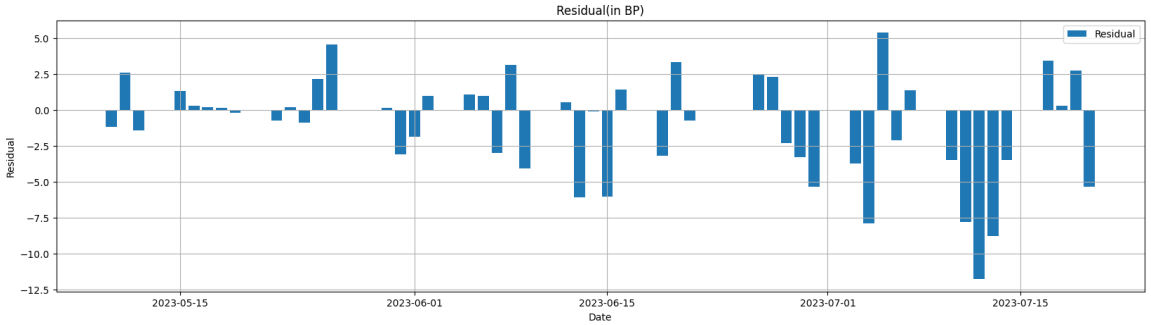
Numerical Results (CGR):

Mean Squared Error: $0.0014(basis\ points)$
Number of Features: 900
Number of Days: 100

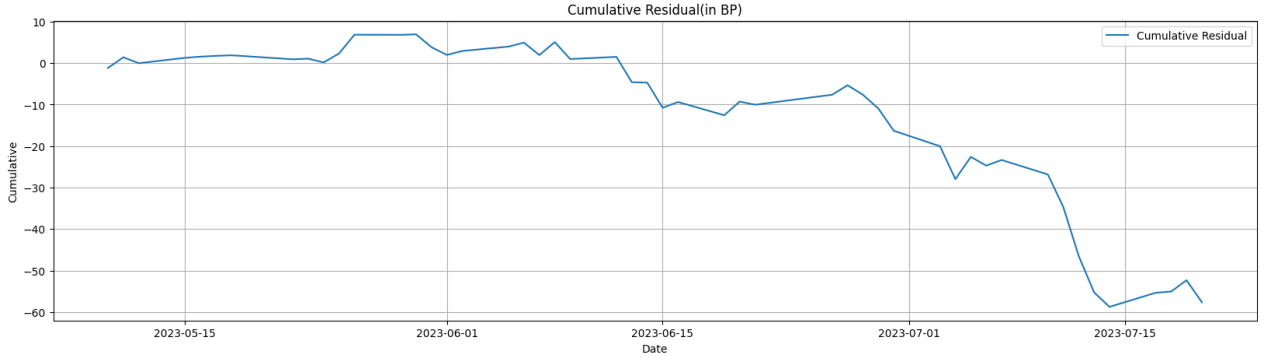
	Weights
2330	0.276477
6695	0.000077
6213	0.000539
3356	0.000071
8104	0.000065

Numerical Results (CGR):





Residuals.



Cum. Residuals.

Conclusion and Possible Amelioration

Possible Amelioration:

Some obstacles……

- NaNs. (Need to find the better way to deal with it!!)
- Anomaly in Weights.
- Calculation Capacity. (2^{963}-1 iterations)
- Some stocks are dominating the market.

Possible Amelioration……

- Divided by more sections. (not just Elec, BKI…)
- Try other optimization technique (not just CGR).
- Length of Testing Dates.
- Include the way of choosing the components into account.

$$N_{components} \times N_{days} \times Model \times Feature\ Selection$$

Conclusion (in basis points):

With best parameters……

MSE	Select by Weights	Select by Shape	Select by Importance	Importance and Shape
Benchmark	0.0013	0.0014	0.0027	0.0014
Linear Regr.	0.0045	0.0049	0.00056	0.0012
Lasso Regr.	0.0044	0.004	0.00059	0.0012
CGR	0.0018	0.0013	0.002	0.0014

Max |Residual|	Select by Weights	Select by Shape	Select by Importance	Importance and Shape
Benchmark	12.4	12.4	16.3	11
Linear Regr.	23.5	24	6.7	8.5
Lasso Regr.	23.7	19.4	6.9	9.9
CGR	15.6	12.4	15	12

Conclusion:

- Have to find the way to deal with NaNs better!!
- Results of regression is not reasonable due to our approaches of dealing with NaNs.
- CGR based on initial weighting might be the best and the most reasonable way for now.
- Choosing the features based on the Importance rather than weights has better performance.
- To divide into sections can as well perform better than not to.

CAPITAL