# 数据

Fn1表示5月期货当月合约；fn2表示6月期货当月合约。

# 思路历程

## 最开始

首先做简单Probit，大概梳理思路，发现准确率不好，结果有偏，五个状态划分时没有几个数据在1和5全都在中间了，fn1内部前80%做训练，后20%做测试时，发现训练集内平均日信号600次，测试集内平均日信号100次。以为是过拟合，后来发现真实信号在训练和测试集内大概是6000和1000，因此认为是期货交易特征（当月合约在临近月末时交易下降）。考虑到数据多，因此之后采用fn1做训练集（或训练集+验证集），fn2做测试集。

## 之后于老师

之后异方差Probit，估计出来结果并不理想，体现为门限值很大，大部分都落在状态3（价格不变），对其他状态预测效果也不好；

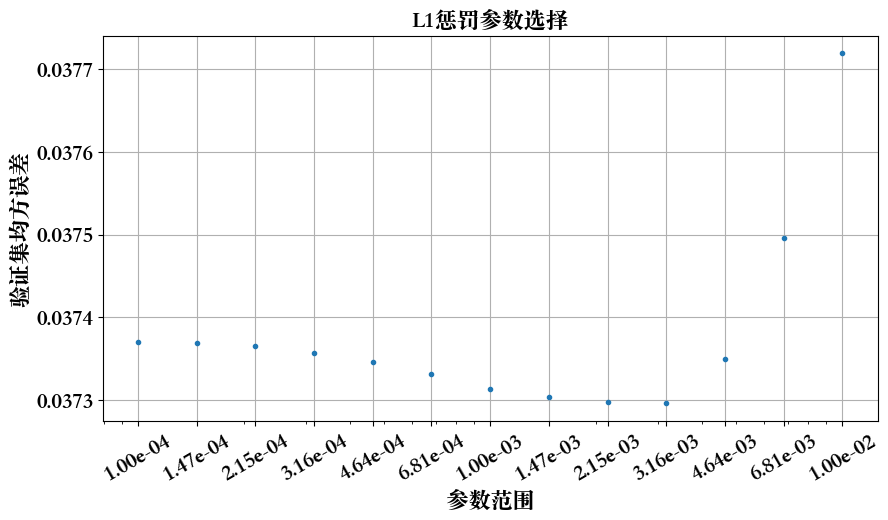
## 再之后补齐时间戳

再之后补齐时间戳以解决非等距时间带来的异方差效应，具体方式是对处理后数据dropna，然后join原来半秒的dataframe，用前填充方式补全数据，开始接下来的模型建立。

# 运用Lasso进行特征选取

首先认为特征选取不合适，这里加入多个特征后开始用Lasso选取特征。

一个很自然的预测思路是，用T时刻数据预测T+1时刻价格。根据参数表现规律，选取Lasso惩罚参数在e-4到e-2之间，按log等分13个点。划分fn1前80%为训练集，后20%为验证集；fn2为测试集。得到验证集误差随L1参数变化如图所示。



对应结果和变量为：

==============================

result for fn\_lasso\_shift1

训练集的特征大小为 (410024, 43)

训练集的标签大小为 (410024,)

验证集的特征大小为 (102506, 43)

验证集的标签大小为 (102506,)

测试集的特征大小为 (551260, 43)

测试集的标签大小为 (551260,)

best validation error: 0.03730

corresponding test error:0.03430

best lambda: 0.00316

number of the features: 18.00000

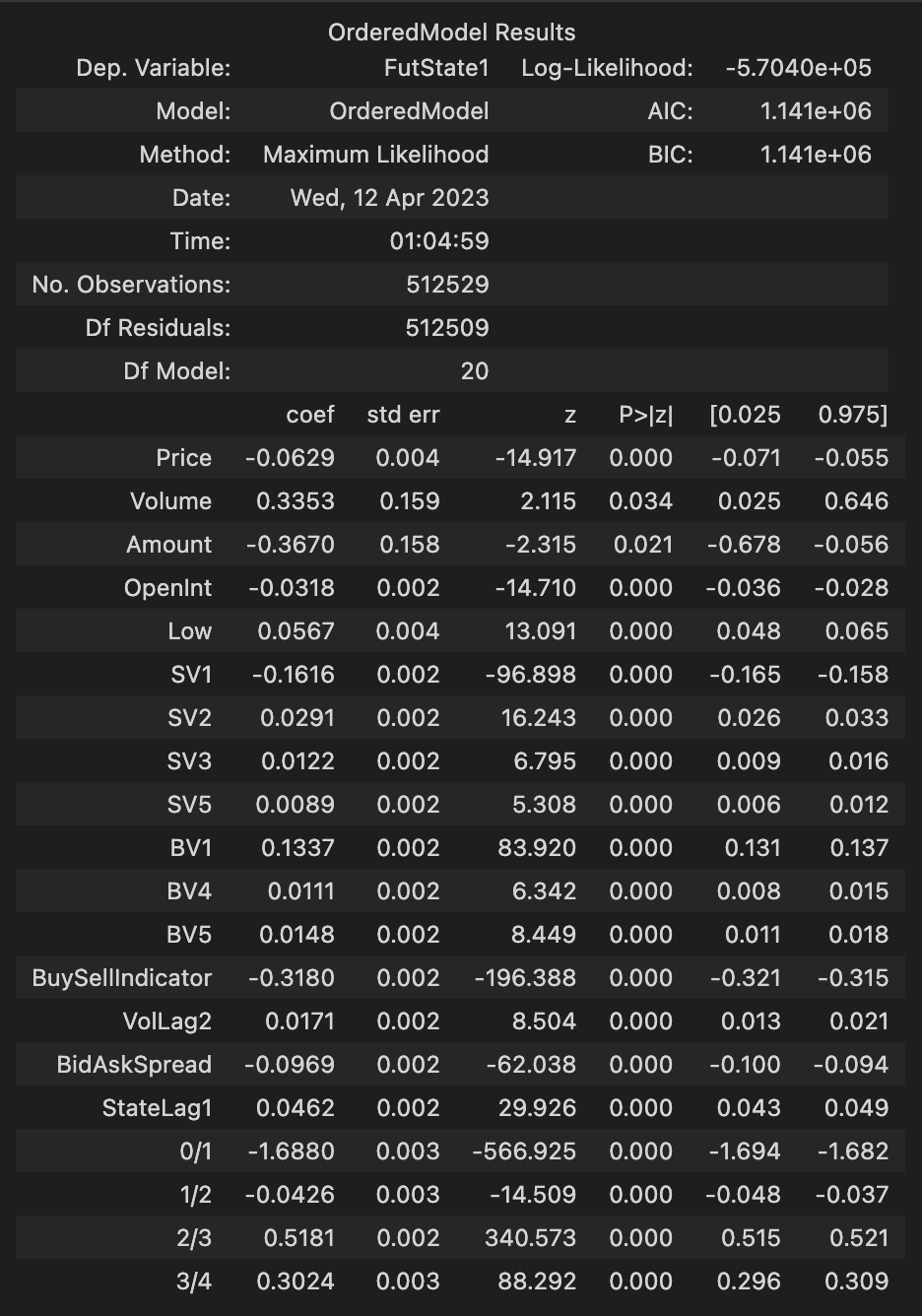
['Price' 'Volume' 'Amount' 'OpenInt' 'Open' 'Low' 'SV1' 'SV2' 'SV3' 'SV5'

'BV1' 'BV4' 'BV5' 'BuySellIndicator' 'VolLag1' 'VolLag2' 'BidAskSpread'

'StateLag1']

# Probit

根据上述变量，开个Probit模型，剔除显著性大于0.05的特征，得到



|  |  |  |  |
| --- | --- | --- | --- |
| OrderedModel Results | | | |
| Dep. Variable: | FutState1 | Log-Likelihood: | -5.7040e+05 |
| Model: | OrderedModel | AIC: | 1.141e+06 |
| Method: | Maximum Likelihood | BIC: | 1.141e+06 |
| Date: | Wed, 12 Apr 2023 |  |  |
| Time: | 01:04:59 |  |  |
| No. Observations: | 512529 |  |  |
| Df Residuals: | 512509 |  |  |
| Df Model: | 20 |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | coef | std err | z | P>|z| | [0.025 | 0.975] |
| Price | -0.0629 | 0.004 | -14.917 | 0.000 | -0.071 | -0.055 |
| Volume | 0.3353 | 0.159 | 2.115 | 0.034 | 0.025 | 0.646 |
| Amount | -0.3670 | 0.158 | -2.315 | 0.021 | -0.678 | -0.056 |
| OpenInt | -0.0318 | 0.002 | -14.710 | 0.000 | -0.036 | -0.028 |
| Low | 0.0567 | 0.004 | 13.091 | 0.000 | 0.048 | 0.065 |
| SV1 | -0.1616 | 0.002 | -96.898 | 0.000 | -0.165 | -0.158 |
| SV2 | 0.0291 | 0.002 | 16.243 | 0.000 | 0.026 | 0.033 |
| SV3 | 0.0122 | 0.002 | 6.795 | 0.000 | 0.009 | 0.016 |
| SV5 | 0.0089 | 0.002 | 5.308 | 0.000 | 0.006 | 0.012 |
| BV1 | 0.1337 | 0.002 | 83.920 | 0.000 | 0.131 | 0.137 |
| BV4 | 0.0111 | 0.002 | 6.342 | 0.000 | 0.008 | 0.015 |
| BV5 | 0.0148 | 0.002 | 8.449 | 0.000 | 0.011 | 0.018 |
| BuySellIndicator | -0.3180 | 0.002 | -196.388 | 0.000 | -0.321 | -0.315 |
| VolLag2 | 0.0171 | 0.002 | 8.504 | 0.000 | 0.013 | 0.021 |
| BidAskSpread | -0.0969 | 0.002 | -62.038 | 0.000 | -0.100 | -0.094 |
| StateLag1 | 0.0462 | 0.002 | 29.926 | 0.000 | 0.043 | 0.049 |
| 0/1 | -1.6880 | 0.003 | -566.925 | 0.000 | -1.694 | -1.682 |
| 1/2 | -0.0426 | 0.003 | -14.509 | 0.000 | -0.048 | -0.037 |
| 2/3 | 0.5181 | 0.002 | 340.573 | 0.000 | 0.515 | 0.521 |
| 3/4 | 0.3024 | 0.003 | 88.292 | 0.000 | 0.296 | 0.309 |

Word粘贴牛逼

预测结果为

其中状态预测正确率为与原数据相同的比例；涨跌方向正确率为0,1状态为开空仓，3,4状态为开多仓，其他信号不操作下操作对的比率；买卖方向正确率为0状态为开空仓，4状态为开多仓，其他信号不操作下操作对的比率

==================================================

最大状态数为 5.0

==================================================

训练集内状态预测正确率为 0.5715

训练集内涨跌方向预测正确率为 0.5739

训练集内买卖方向预测正确率为 0.8176

模型每日去重开仓信号次数 1.2000

==================================================

==================================================

最大状态数为 5.0

==================================================

测试集内状态预测正确率为 0.6097

测试集内涨跌方向预测正确率为 0.6116

测试集内买卖方向预测正确率为 0.8352

模型每日去重开仓信号次数 1.0952

==================================================

大概和真实的比一下

fn1真实状态计数 fn1预测状态计数 fn2真实状态计数 fn2预测状态计数

0 27940 449 22889 633

1 98690 1139 100988 1550

2 290651 501830 334663 541865

3 87034 9109 86466 7203

4 8214 2 6254 9

# 怀疑模型不行，上个adaboost

其实adaboost早就上了，那时候probit还没站起来。不过现在也没站起来，因为去重后每天就一手，没啥动静。

Adaboost选取最大深度为3，100个树，得到结果

==================================================

最大状态数为 5.0

==================================================

训练集内状态预测正确率为 0.5788

训练集内涨跌方向预测正确率为 0.5963

训练集内买卖方向预测正确率为 0.8137

模型每日去重开仓信号次数 163.5500

==================================================

==================================================

最大状态数为 5.0

==================================================

测试集内状态预测正确率为 0.6015

测试集内涨跌方向预测正确率为 0.6169

测试集内买卖方向预测正确率为 0.8173

模型每日去重开仓信号次数 131.5714

==================================================

再看看和真实的状态差异

fn1真实状态计数 fn1预测状态计数 fn2真实状态计数 fn2预测状态计数

0 27940 11202 22889 9565

1 98690 11726 100988 10407

2 290651 447468 334663 484328

3 87034 36278 86466 35847

4 8214 5855 6254 11113

已经很能打了

下面是变量的权重，权重越高越重要

变量名 权重

Price 0.04921115644884169

Volume 0.007715045024669852

Amount 0.10653425274440345

OpenInt 0.10863767289922106

Open 0.004044411213355352

Low 0.02401601564712642

SV1 0.0542891420162907

SV2 0.08872062860604231

SV3 0.08351528958597251

SV5 0.07897021759083893

BV1 0.07869697245204216

BV4 0.06300160241132126

BV5 0.08497162311503698

BuySellIndicator 0.03476843737489511

VolLag1 0.055810430920015926

VolLag2 0.0489115097292317

BidAskSpread 0.012769149821673747

StateLag1 0.015416442399020758

# 回测看看

## Probit

Probit，双边费率万分之0.23，结果如下：

==============================

result for probit\_shift1 with commission 2.3e-05

Annualized\_Mean 1.324368

Annualized\_Std 0.126008

Sharpe 10.510211

Excess\_Annualized\_Mean 1.334613

Excess\_Annualized\_Std 0.099956

Excess\_sharpe 13.351996

MaxDrawDown 0.024670

LongCounts 265.000000

MeanLongTime 928.618868

PerLongReturn 0.000170

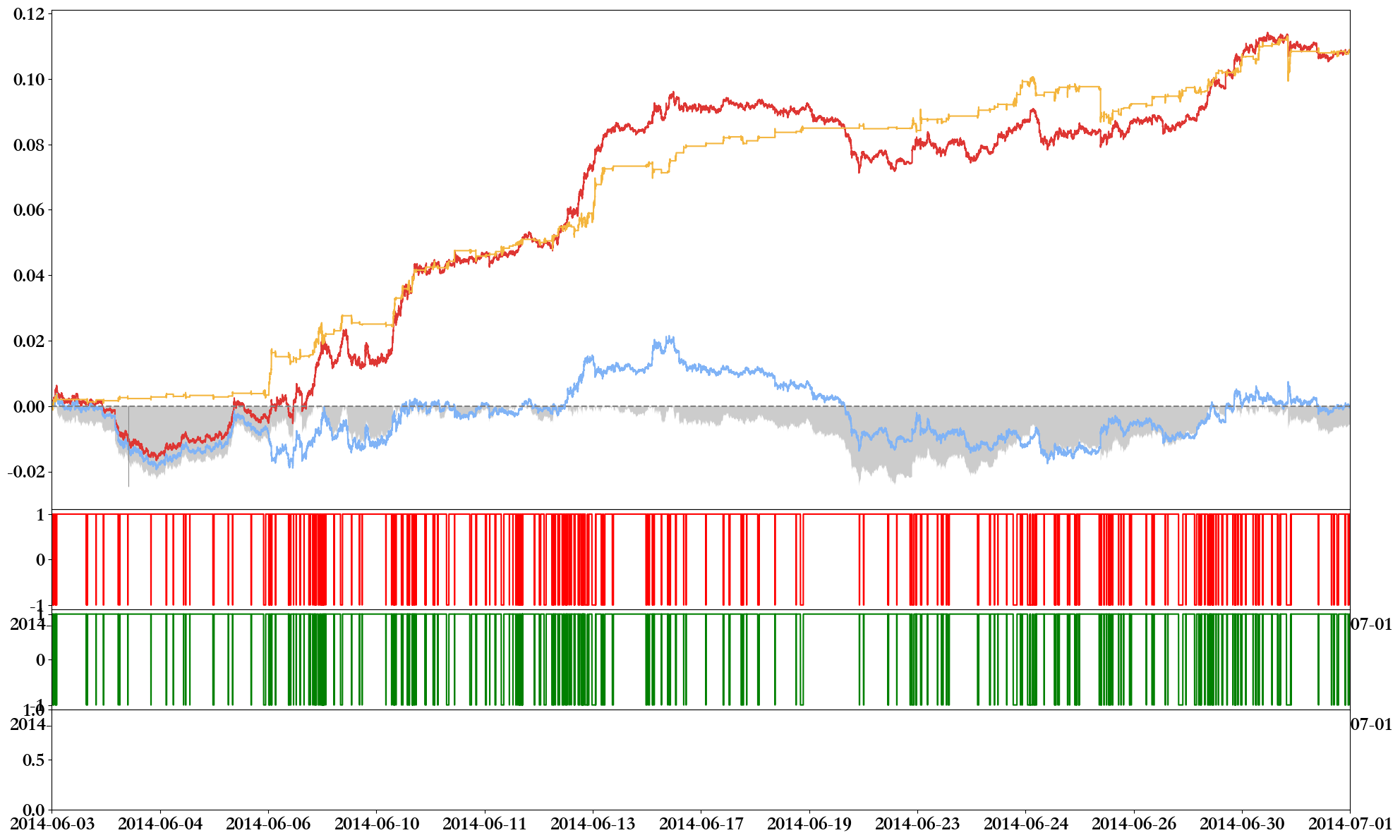
ShortCounts 264.000000

MeanShortTime 107.681818

PerShortReturn 0.000238

PnL 0.834517

WinRate 0.650284



挺牛的

## Adaboost

==============================

result for adaboost\_shift1 with commission 2.3e-05

Annualized\_Mean 7.155580

Annualized\_Std 0.125877

Sharpe 56.845884

Excess\_Annualized\_Mean 7.165826

Excess\_Annualized\_Std 0.125295

Excess\_sharpe 57.191842

MaxDrawDown 0.006713

LongCounts 3051.000000

MeanLongTime 71.659456

PerLongReturn 0.000085

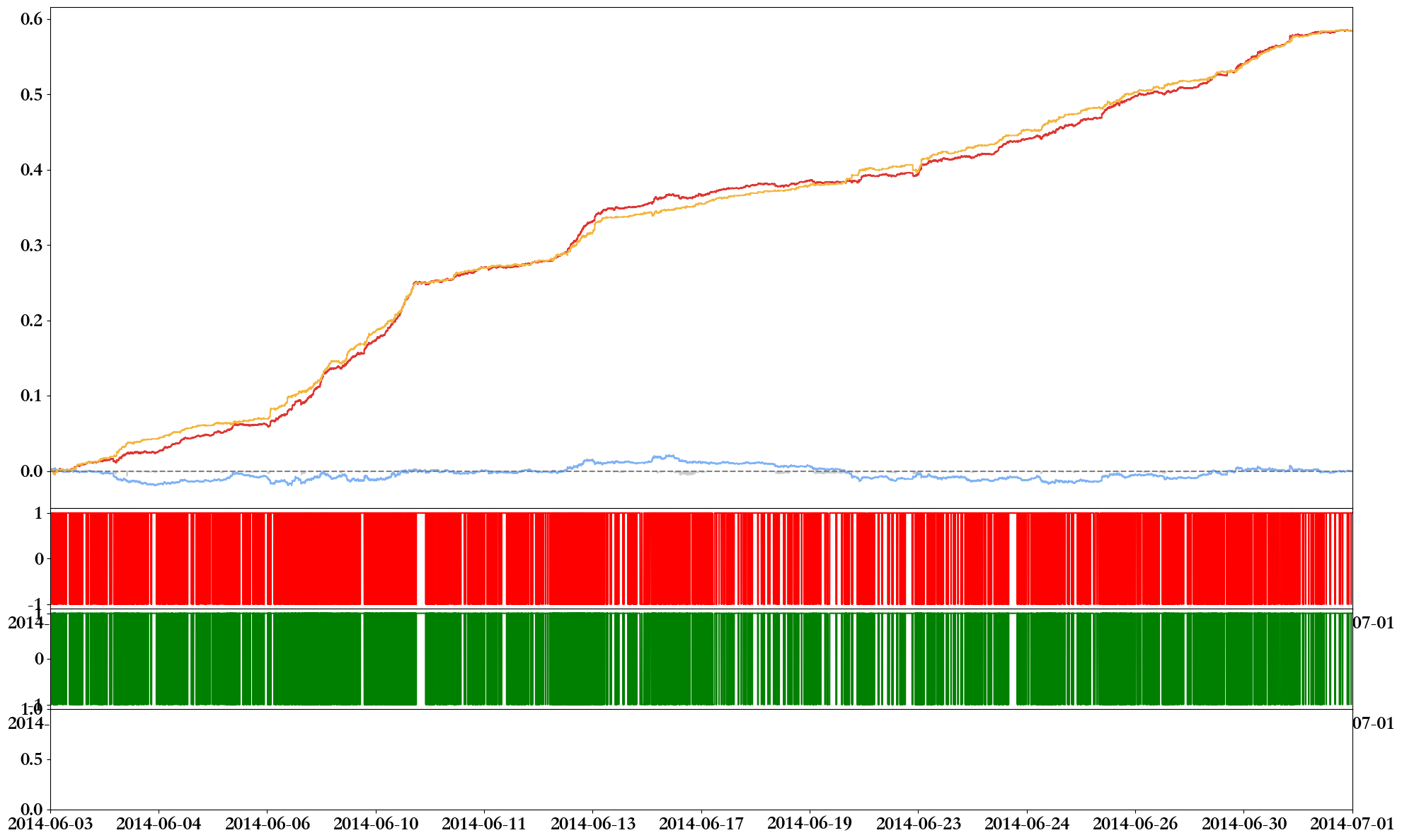
ShortCounts 3050.000000

MeanShortTime 18.361311

PerShortReturn 0.000106

PnL 1.205437

WinRate 0.620062



太可怕了，年化七倍。找找原因。

## 原因

* + - 1. 可能有未来信息
      2. 重新考虑费率

# 首先考虑未来信息，重新lasso

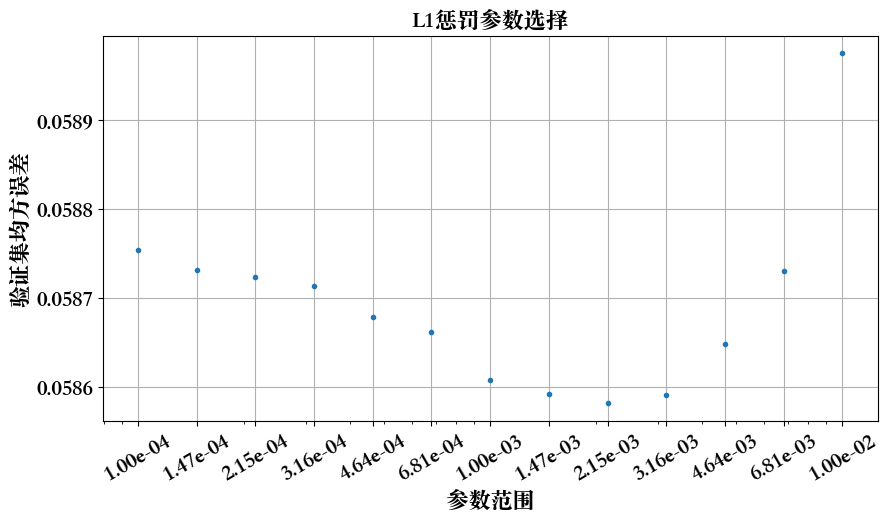
考虑到可能有未来信息，因为之前是思路一：

思路一：在训练集5月期货内用T时刻预测T+1的状态，得到模型运用于测试集6月期货得到因子值，T时刻因子值作为T时刻持仓进行投资；

接下来采取思路二，即

思路二：在训练集5月期货内用T时刻预测T+2的状态，得到模型运用于测试集6月期货得到因子值，T时刻因子值作为T+1时刻持仓进行投资；

重新估计lasso，用T期特征T+2期价格回归，对参数采取同样插值方法，得到



==============================

result for fn\_lasso\_shift2

训练集的特征大小为 (410024, 43)

训练集的标签大小为 (410024,)

验证集的特征大小为 (102506, 43)

验证集的标签大小为 (102506,)

测试集的特征大小为 (551260, 43)

测试集的标签大小为 (551260,)

best validation error: 0.05858

corresponding test error:0.05462

best lambda: 0.00215

number of the features: 24.00000

['Price' 'Volume' 'Amount' 'OpenInt' 'TotalVol' 'TotalAmount' 'Open' 'Low'

'SV1' 'SV2' 'SV3' 'SV5' 'BV1' 'BV4' 'BV5' 'DeltaTick' 'BuySellIndicator'

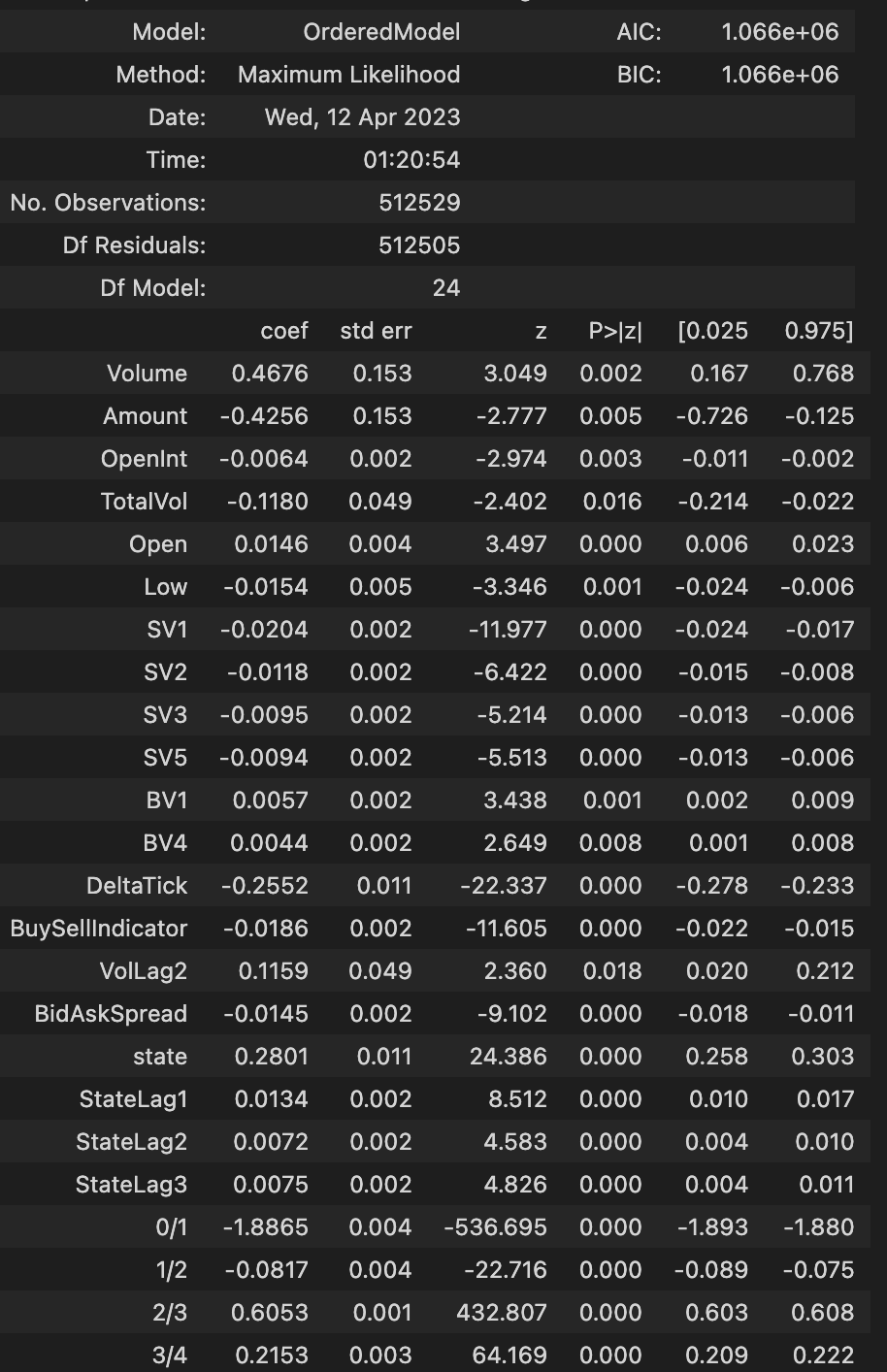
'VolLag1' 'VolLag2' 'BidAskSpread' 'state' 'StateLag1' 'StateLag2'

'StateLag3']

误差高了很多。

# 重新Probit

重新Probit，得到



OrderedModel Results

Dep. Variable: FutState2 Log-Likelihood: -5.3300e+05

Model: OrderedModel AIC: 1.066e+06

Method: Maximum Likelihood BIC: 1.066e+06

Date: Wed, 12 Apr 2023

Time: 01:20:54

No. Observations: 512529

Df Residuals: 512505

Df Model: 24

coef std err z P>|z| [0.025 0.975]

Volume 0.4676 0.153 3.049 0.002 0.167 0.768

Amount -0.4256 0.153 -2.777 0.005 -0.726 -0.125

OpenInt -0.0064 0.002 -2.974 0.003 -0.011 -0.002

TotalVol -0.1180 0.049 -2.402 0.016 -0.214 -0.022

Open 0.0146 0.004 3.497 0.000 0.006 0.023

Low -0.0154 0.005 -3.346 0.001 -0.024 -0.006

SV1 -0.0204 0.002 -11.977 0.000 -0.024 -0.017

SV2 -0.0118 0.002 -6.422 0.000 -0.015 -0.008

SV3 -0.0095 0.002 -5.214 0.000 -0.013 -0.006

SV5 -0.0094 0.002 -5.513 0.000 -0.013 -0.006

BV1 0.0057 0.002 3.438 0.001 0.002 0.009

BV4 0.0044 0.002 2.649 0.008 0.001 0.008

DeltaTick -0.2552 0.011 -22.337 0.000 -0.278 -0.233

BuySellIndicator -0.0186 0.002 -11.605 0.000 -0.022 -0.015

VolLag2 0.1159 0.049 2.360 0.018 0.020 0.212

BidAskSpread -0.0145 0.002 -9.102 0.000 -0.018 -0.011

state 0.2801 0.011 24.386 0.000 0.258 0.303

StateLag1 0.0134 0.002 8.512 0.000 0.010 0.017

StateLag2 0.0072 0.002 4.583 0.000 0.004 0.010

StateLag3 0.0075 0.002 4.826 0.000 0.004 0.011

0/1 -1.8865 0.004 -536.695 0.000 -1.893 -1.880

1/2 -0.0817 0.004 -22.716 0.000 -0.089 -0.075

2/3 0.6053 0.001 432.807 0.000 0.603 0.608

3/4 0.2153 0.003 64.169 0.000 0.209 0.222

准确率

==================================================

最大状态数为 5.0

==================================================

训练集内状态预测正确率为 0.6386

训练集内涨跌方向预测正确率为 0.6386

训练集内买卖方向预测正确率为 0.8065

模型每日去重开仓信号次数 1.1000

==================================================

==================================================

最大状态数为 5.0

==================================================

测试集内状态预测正确率为 0.6775

测试集内涨跌方向预测正确率为 0.6775

测试集内买卖方向预测正确率为 0.8280

模型每日去重开仓信号次数 1.0000

==================================================

分布

fn1真实状态计数 fn1预测状态计数 fn2真实状态计数 fn2预测状态计数

0 15101 12072

1 70964 70856

2 327324 512508.0 373500 551251.0

3 89826 19.0 88223 9.0

4 9314 2.0 6609

离谱，这里直接没有预测出来状态0和1,但是矩阵收敛成功了

# 重新adaboost

又来到喜闻乐见的adaboost环节

预测准确率

==================================================

最大状态数为 5.0

==================================================

训练集内状态预测正确率为 0.6385

训练集内涨跌方向预测正确率为 0.6415

训练集内买卖方向预测正确率为 0.8030

模型每日去重开仓信号次数 112.9500

==================================================

==================================================

最大状态数为 5.0

==================================================

测试集内状态预测正确率为 0.6454

测试集内涨跌方向预测正确率为 0.6523

测试集内买卖方向预测正确率为 0.8183

模型每日去重开仓信号次数 97.7619

==================================================

看看分布

fn1真实状态计数 fn1预测状态计数 fn2真实状态计数 fn2预测状态计数

0 15101 1225 12072 16154

1 70964 216 70856 7524

2 327324 503722 373500 518663

3 89826 3310 88223 3159

4 9314 4056 6609 5760

真不错

看看权重

变量名 权重

Price 0.048180481834215014

Volume 0.015406939672658987

Amount 0.10865077001280145

OpenInt 0.08818848455745627

TotalVol 0.02906120223260613

TotalAmount 0.03466930693283725

Open 0.012327710019193678

Low 0.027459706802000716

SV1 0.04018529638991711

SV2 0.08140076551266814

SV3 0.09954295268319488

SV5 0.07016164885921732

BV1 0.08481838469103385

BV4 0.06542340933746094

BV5 0.07272451355347065

DeltaTick 0.012903785935197865

BuySellIndicator 0.003507243568388182

VolLag1 0.013467325792178846

VolLag2 0.03740648596330789

BidAskSpread 0.005332664693473314

state 0.012300202022004022

StateLag1 0.012764582958352932

StateLag2 0.015141782952472704

StateLag3 0.008974353023891822

# 重新回测

## Probit

==============================

result for probit\_shift2 with commission 2.3e-05

Annualized\_Mean 0.056792

Annualized\_Std 0.055308

Sharpe 1.026839

Excess\_Annualized\_Mean 0.067037

Excess\_Annualized\_Std 0.113279

Excess\_sharpe 0.591789

MaxDrawDown 0.010145

LongCounts 1.000000

MeanLongTime 51137.000000

PerLongReturn 0.004635

ShortCounts 0.000000

MeanShortTime 0.000000

PerShortReturn 0.000000

PnL NaN

WinRate 1.000000

去重后就没啥信号了。。

## 看看Adaboost

==============================

result for adaboost\_shift2 with commission 2.3e-05

Annualized\_Mean -0.009579

Annualized\_Std 0.125950

Sharpe -0.076057

Excess\_Annualized\_Mean 0.000666

Excess\_Annualized\_Std 0.162371

Excess\_sharpe 0.004101

MaxDrawDown 0.062283

LongCounts 84.000000

MeanLongTime 2014.071429

PerLongReturn -0.000206

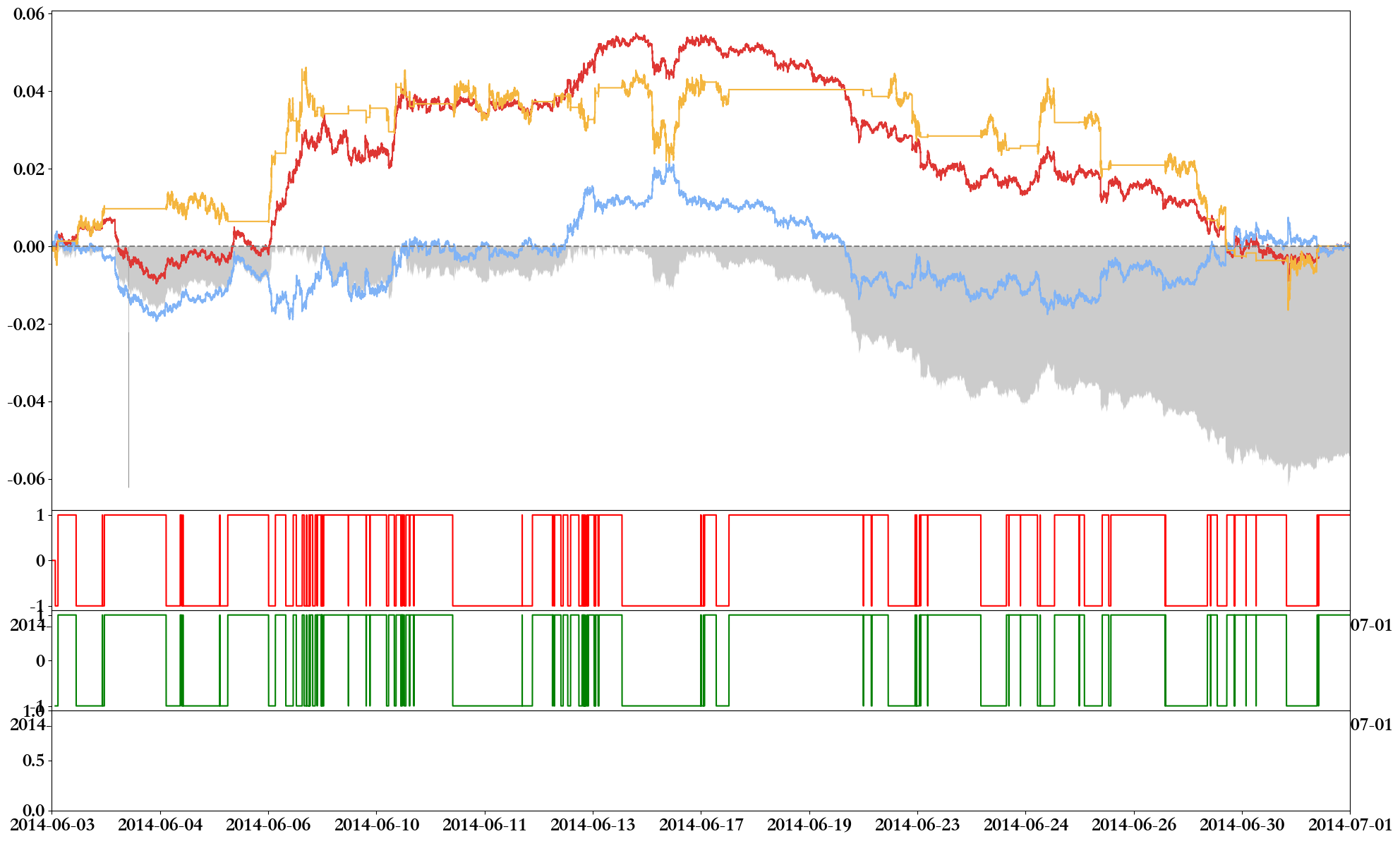
ShortCounts 84.000000

MeanShortTime 1246.928571

PerShortReturn 0.000197

PnL 0.693005

WinRate 0.589286



变拉了，小亏

如果不考虑交易费率呢？

==============================

result for adaboost\_shift2 with commission 0

Annualized\_Mean 0.084838

Annualized\_Std 0.125931

Sharpe 0.673688

Excess\_Annualized\_Mean 0.095083

Excess\_Annualized\_Std 0.162303

Excess\_sharpe 0.585836

MaxDrawDown 0.059778

LongCounts 84.000000

MeanLongTime 2014.071429

PerLongReturn -0.000160

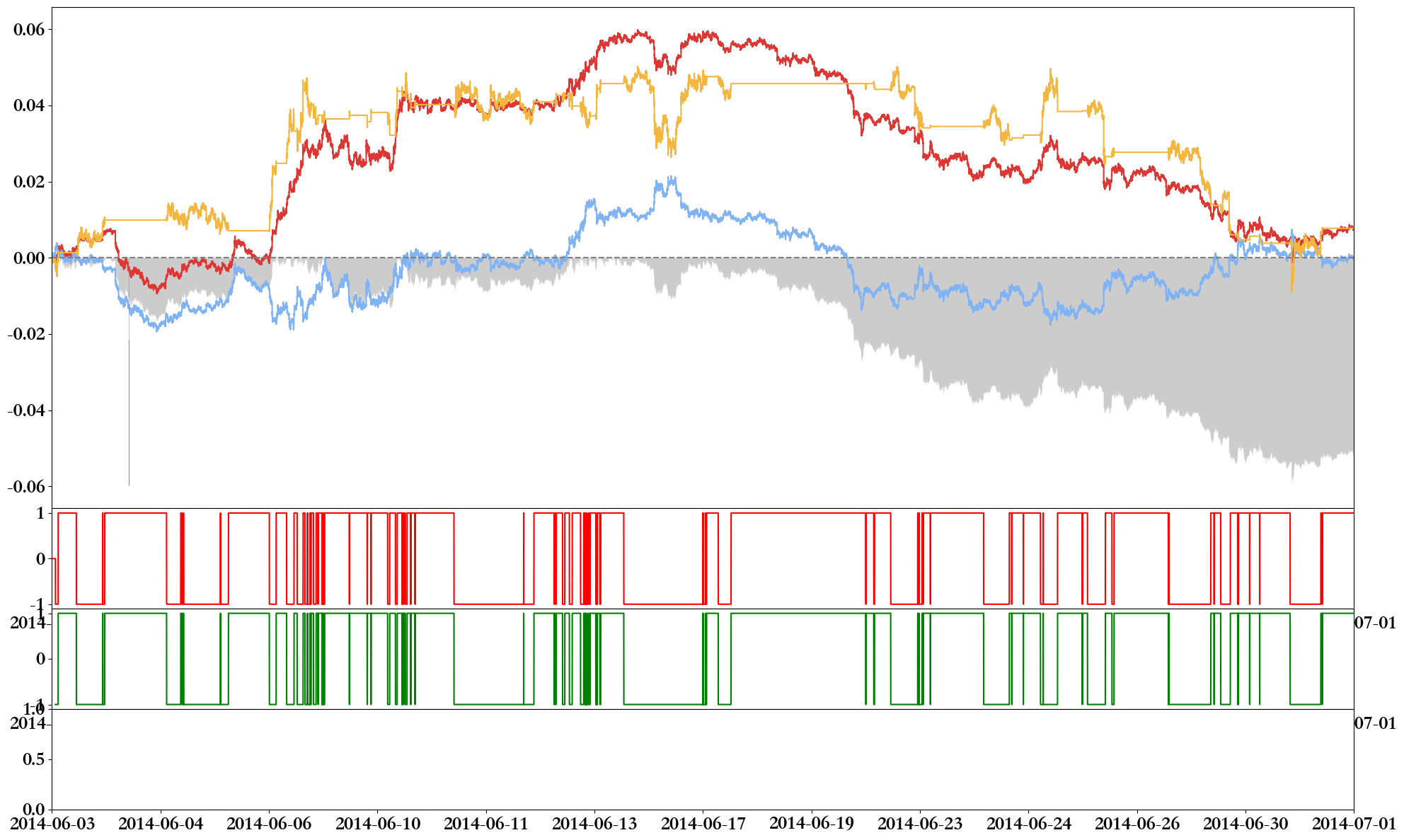
ShortCounts 84.000000

MeanShortTime 1246.928571

PerShortReturn 0.000243

PnL 0.687166

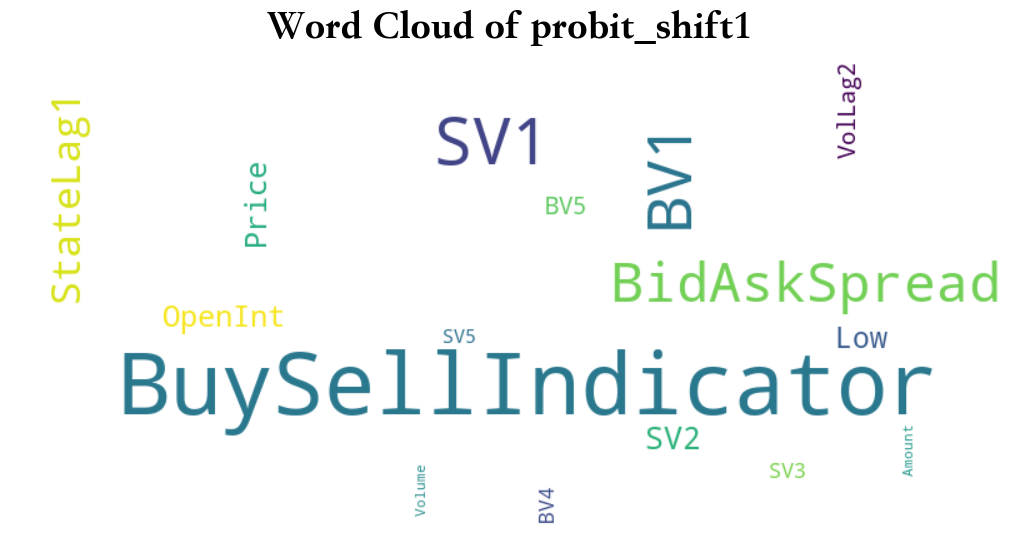
WinRate 0.601190



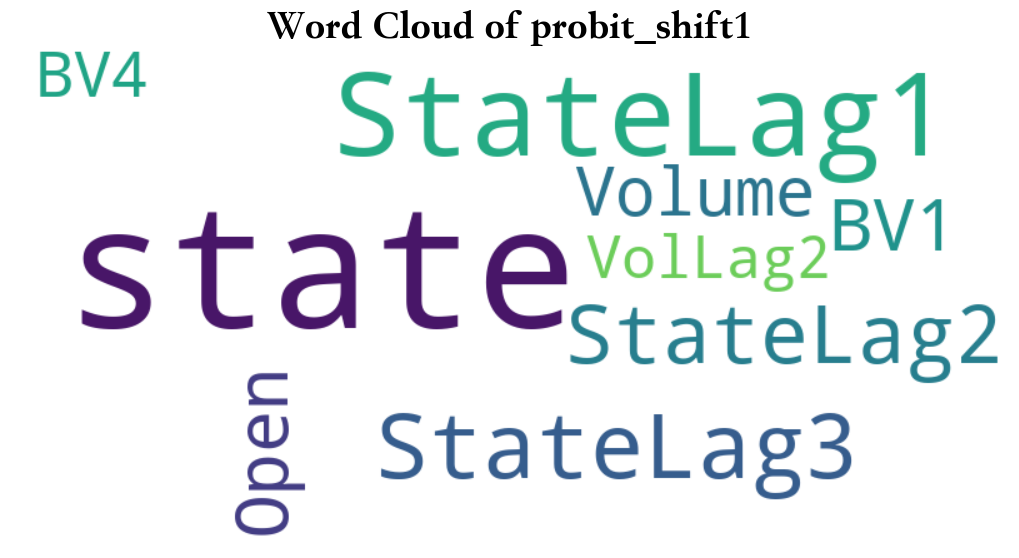
稍微赚点，但后来还是又下去了。毕竟是五月训练好模型在6月一整个月跑，所以能维持半个月收益也还不错了。

# 汇总一下显著度或权重

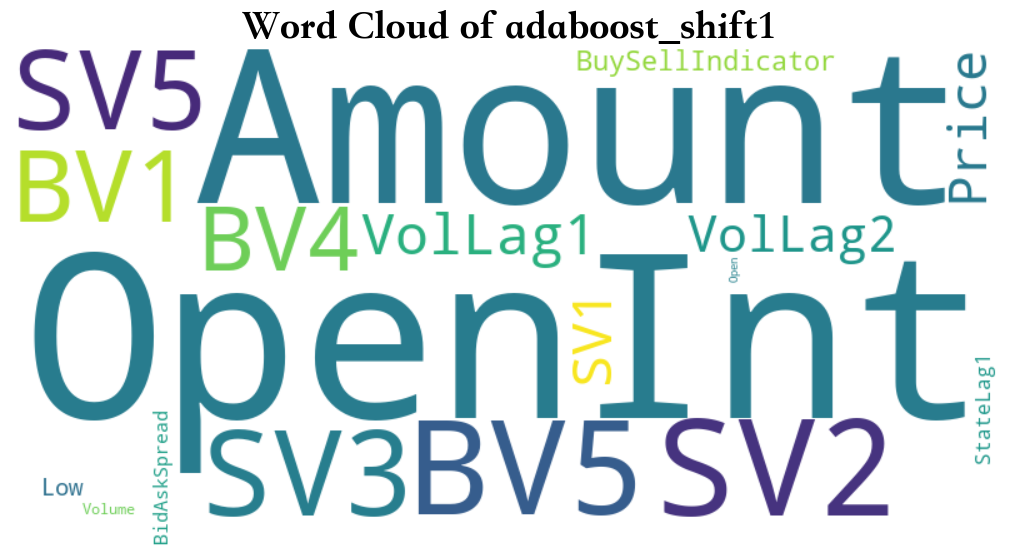
Probit预测T+1得到的z统计量



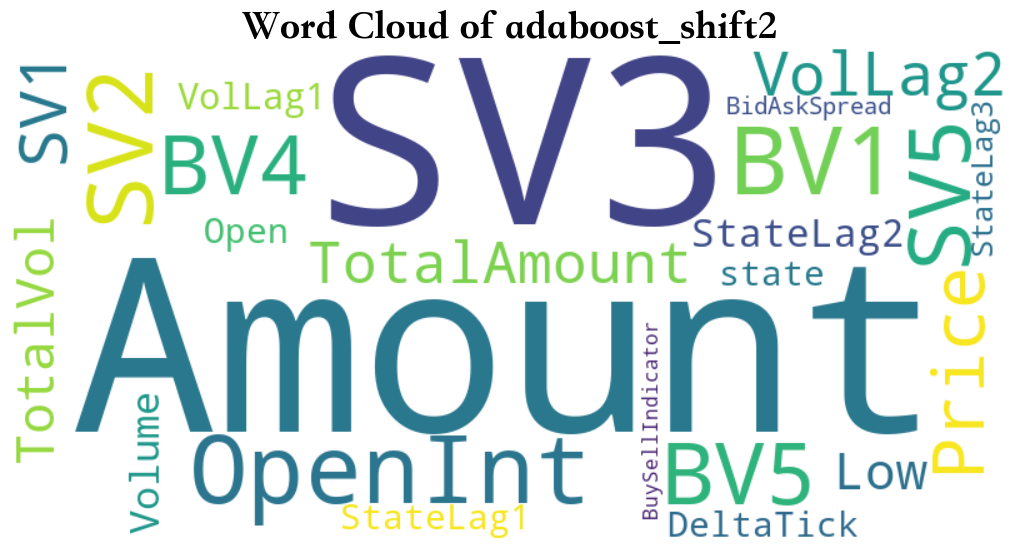
Probit预测T+2得到的z统计量



Adaboost预测T+1得到的特征权重



Adaboost预测T+2得到的特征权重



# 之前回测的操作逻辑是涨跌方向，调整成买卖方向

这里双边费率相同，取成万2.3和万0.23的中位数

## Probit对T+1

调整成买卖方向后

==============================

result for probit\_shift1 with commission 0.0001265

Annualized\_Mean 0.043918

Annualized\_Std 0.125932

Sharpe 0.348748

Excess\_Annualized\_Mean 0.054164

Excess\_Annualized\_Std 0.251927

Excess\_sharpe 0.214997

MaxDrawDown 0.039818

LongCounts 0.000000

MeanLongTime 0.000000

PerLongReturn 0.000000

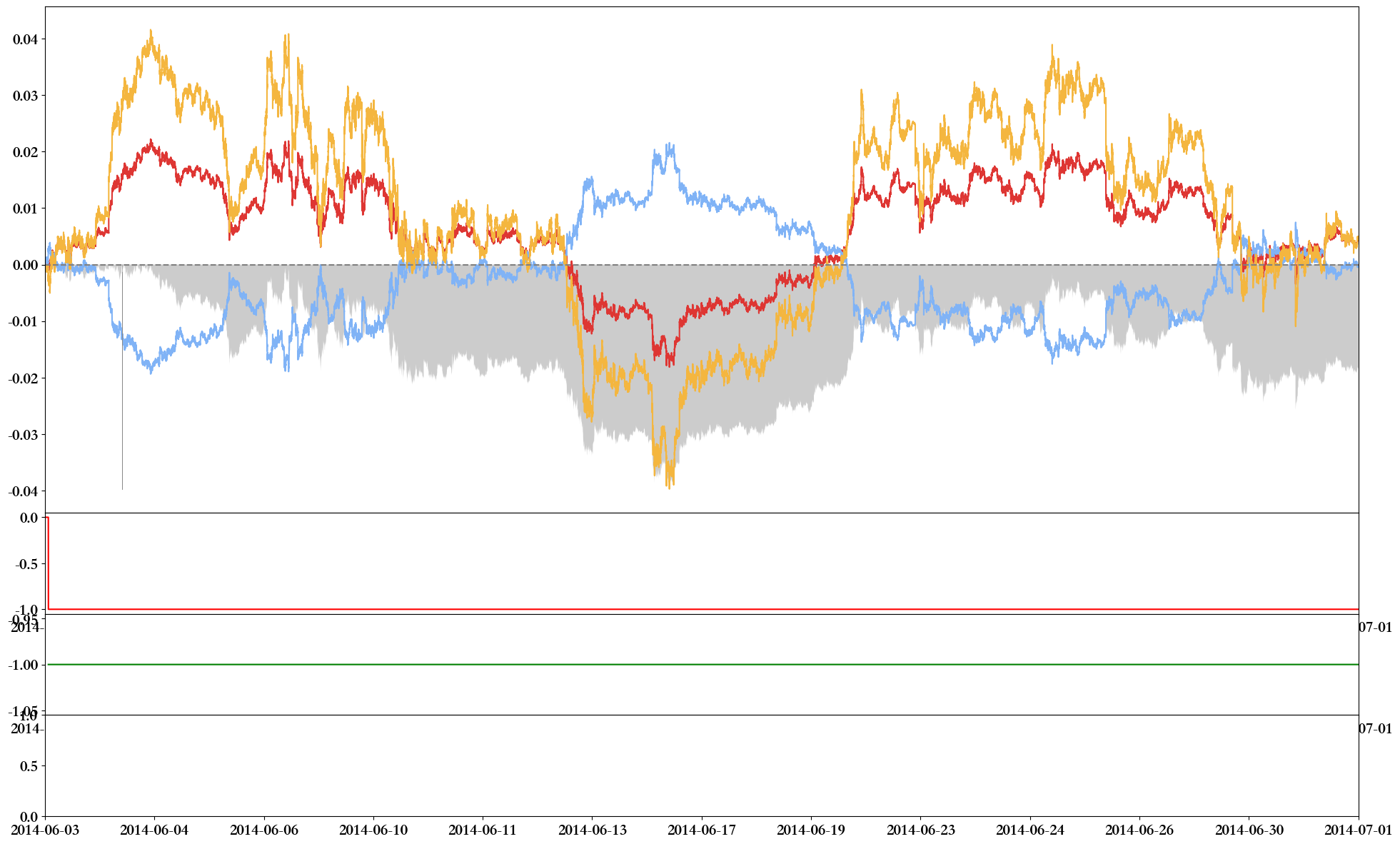
ShortCounts 1.000000

MeanShortTime 273924.000000

PerShortReturn 0.003584

PnL NaN

WinRate 1.000000



就开了一手，没啥信号

## Probit对T+2

直接没信号，nan

## Adaboost对T+1

==============================

result for adaboost\_shift1 with commission 0.0001265

Annualized\_Mean 0.742746

Annualized\_Std 0.127491

Sharpe 5.825879

Excess\_Annualized\_Mean 0.752991

Excess\_Annualized\_Std 0.203745

Excess\_sharpe 3.695761

MaxDrawDown 0.022727

LongCounts 319.000000

MeanLongTime 297.821317

PerLongReturn 0.000043

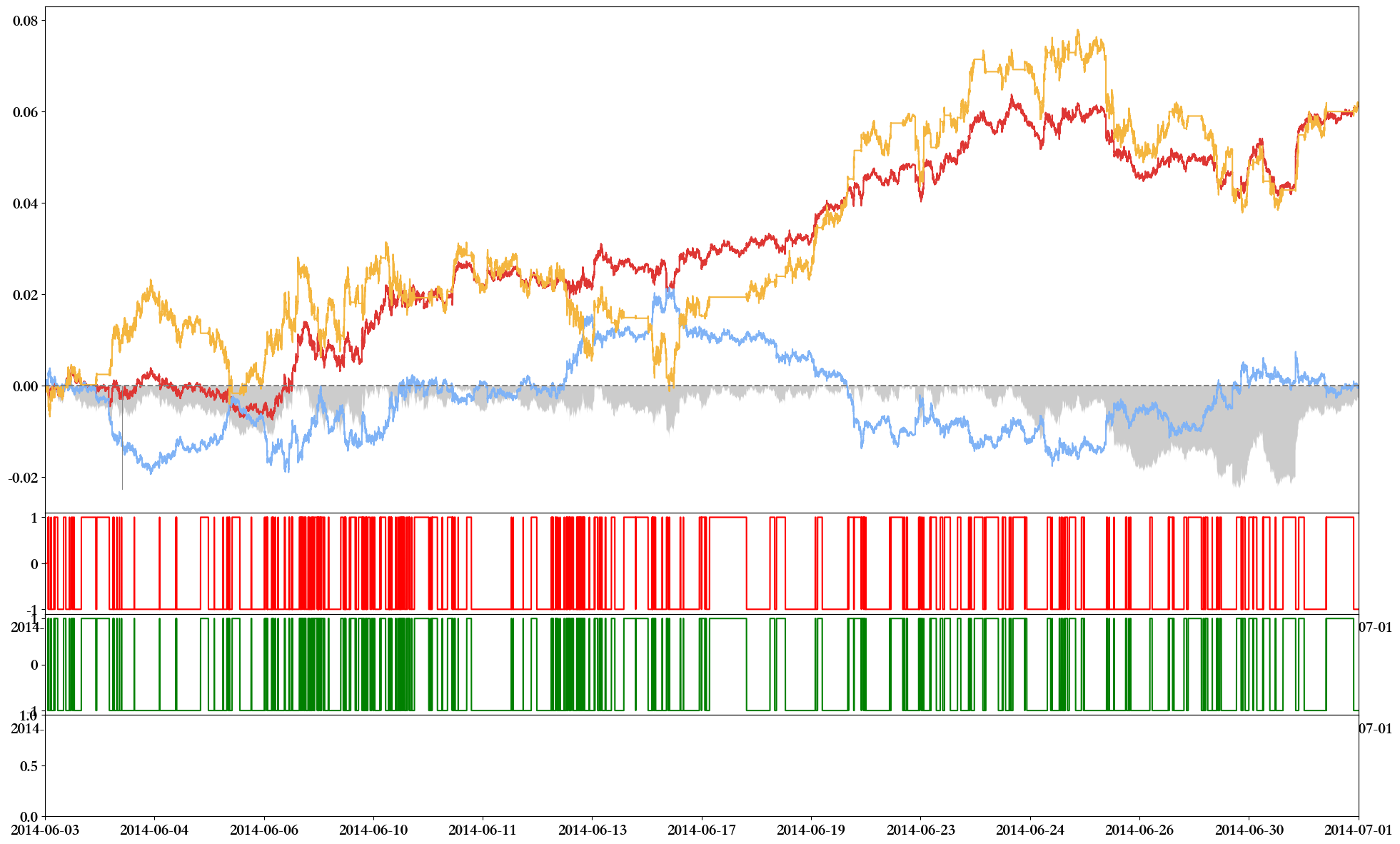
ShortCounts 320.000000

MeanShortTime 559.218750

PerShortReturn 0.000147

PnL 1.338334

WinRate 0.483568



居然赚钱了

## Adaboost对T+2

==============================

result for adaboost\_shift2 with commission 0.0001265

Annualized\_Mean 0.120488

Annualized\_Std 0.126111

Sharpe 0.955415

Excess\_Annualized\_Mean 0.130733

Excess\_Annualized\_Std 0.185942

Excess\_sharpe 0.703087

MaxDrawDown 0.037982

LongCounts 24.000000

MeanLongTime 5435.000000

PerLongReturn -0.000084

ShortCounts 24.000000

MeanShortTime 5979.458333

PerShortReturn 0.000493

PnL 1.147059

WinRate 0.500000

