Trading, AI and Quantitative Investing in Domestic Market

交易实战, 人工智能与国内二级市场量化交易初探

`AI, Trading, Quantitative, Stocks, Multi-Factor`

221 Data LLC. Artificial Intelligence Group: Allen LI, Tracy LIU, Harvey LIU, Sophy LU

Section 1. Profile of Current Domestic Stock Market Participators

当下我国二级市场参与者的特征

1. 沪深A股市场参与者构成

我国沪深A股市场与西方股票市场从参与者构成上存在着本质的区别。百分之八十左右的参与者为散户构成。而西方股票市场的主要参与者为机构，Pension Fund, Public Fund, Hedge Fund etc.

此外，2016年股灾之后，我国A股市场又多了"国家队"这一独特而又举足轻重的参与者。“国家队”承担着维稳的作用，同时在其大量套牢盘的现实情况下，使得股市全面上扬面临着巨大的抛压。

游资，又是我国市场一道独特的风景线。关于游资众说纷纭，褒贬不一。其激进的操作手法，涨停敢死队的风格更是存在着巨大的争议。然而不可否定的是，游资在这众多散户心目中起着风向标作用。在我国市场被国家队任意践踏，政策无故干预的困难情形下，游资作为一股倔强的力量，敢于亮剑，能够引爆市场的风向，制造概念与龙头，从而也起到了二级市场的价值发现与资源合理配置的作用。

游资所参与的股票往往是老百姓喜闻乐见，熟悉的，股性较强的，国家队套牢盘较小的股票。而我国公募基金，券商自营部门，资管部门根据多因子模型选股而参与的股票，游资与散户很少涉及。可以说，我国沪深A股市场某种程度上是割裂的。这也是近两年来指数失真，二八，一九转换如此频繁的根本原因。那么，没有了游资的生存空间，也就没有了散户，这些市场百分之八十参与者的生存空间。

游资的操作风格与手法各不相同，通过对每日龙虎榜，买卖席位的数据挖掘，能够发现游资的蛛丝马迹，从而及时捕捉市场中的阶段性热点。这已经成为了我国市场举足轻重的因子。现阶段我国对市场起着举足轻重影响的游资有赵老哥，中信证券上海淮海中路，深圳金田路，益田路，成都帮双雄，佛山无影脚，山东帮，乔帮主以及孙哥等。

2.指数失真与二八转换

正如之前所提及的市场割裂现象。广大散户与机构参与的股票交集极小，偶尔才会出现交集，从而在当日龙虎榜中看到游资与机构之间的强博弈现象，制造出振幅巨大的个股。

我国市场的这种特殊构成，直接导致了近两年时常出现的权重股(银行，保险，券商)拉升而维稳指数的现象。实际情况是广大散户持有的个股暴跌，少数机构持有的权重，蓝筹，白马上涨。除了指数在数字上好看之外，市场情绪其实已经降到了冰点，亏钱效应十分严重。去年第四季度出现的一九分化，以及今年多次出现的二八分化，也是这种维稳操作情形下，游资，散户与国家队之间的博弈结果。

3.打新,新股与次新股

打新是我国市场特有的现象之一。中新股如同中彩票，因为我国新股上市往往是连续若干个集合竞价封板式的涨停板。由于之前的打新规则存在一定漏洞，再加上2015年股灾之后，大家发现只有打新股能够稳定的创造巨大的收益。于是为了打新，各募集产品持有大量的蓝筹股，再配合50ETF期权，不惜一切提高中新股的概率。这也是市场不可忽视的一个因素。

尽管今年打新规则进行了修改，新股与次新股在市场上的重要作用丝毫没有减轻。其关键因素之一便是新股与次新股没有套牢盘，没有国家队参与，向上阻力小。若新股与次新股能够结合当下概念与热点，更将成为全市场的龙头股票。比如近期的光威复材与五月份的杭州园林。

其关键因素之二: 目前资金存量博弈的情形下，新股发行速度过快，导致资金都被吸收至新股及次新股中。再无充裕的资金给其他股票以正确的估值。

此外，新股与次新股是市场情绪的晴雨表。新股开板是非常关键的博弈时刻，是否能立即反包，是否放量，反包的时间是上午还是下午收盘时，都存在着巨大的变数与重要的解读。另外，次新股的龙回头时机与第二波炒作也是市场博弈的关键因子。总之，我国的新股与次新股市场是值得单独提炼出来进行研究与博弈的。

4.高送转

高送转行情是中国市场特有的，最具人气之一的行情。分为预期高送转炒作期，高送转落地炒作期以及填权期. 一支被市场认可的高送转标的价格翻倍是至少的。每个季度，尤其年末是高送转行情的爆发期，且该板块具备记忆性，会被拿来反复炒作。(年末还有ST摘帽炒作)

5.股指期货市场的不完善

自从股灾之后对股指期货的开平仓，手续费的诸多限制，沪深A股市场的对冲工具已经名存实亡。极差的流动性，负基差的真实存在，限仓等诸多因素，扭曲了市场。西方多因子量化对冲策略无法直接适用。

6.我国商品期货市场情况 (篇幅限制略)

Section 2. Our Unique 221 Factors

基于交易实战衍生出的因子构建

随着近年来人工智能的兴起，海外广大对冲基金招兵买马，因子库由一百个上升到三百，四百个。同时不断尝试新的算法，较为流行的有Deep Random Forest, Deep Neural Networks combined with Gradient Boosting等等。

本篇我们依然适用数量为一百的因子库，同时加上221公司特有的因子，也就是Section 1中提及的，从交易实战中总结出来的因子。

Subsection 1. Looking into the Data

- 让我们从数据出发。以万得导出的wsd数据为例

```
rm(list=ls())

### set working directory
#setwd("D:/R")

library(quantmod)
library(dtw)
library(sm)
library(fBasics)
library(data.table)
library(ggplot2)
library(WindR);

w.start();
stockCodes=w.wset("sectorconstituent","date=20171108;sectorid=a001010100000000;field=wind_code")

start.date='20090101'
end.date='20171108'
codes <- stockCodes$Data
code <- codes$wind_code
code_length<-length(code);
for (i in 1:code_length) {
  wsd_data<- w.wsd(code[i],"trade_code,open,high,low,close,pre_close,volume,amt,dealnum,chg,pct_chg,vwap, adjfactor,close2,turn,free_
  data_df<-data.frame(wsd_data$Data);
  filepath<-paste('D:/winddata/',code[i],'.csv',sep="");
  write.csv(data_df,file = filepath)
}
```

- 目前市场3149支股票, 不到九年的时间，一年252个交易日左右，这些数据远远不够算法进行训练与学习。且我国是发展中国家，金融市场也有待完善，五年前的市场特性与当下相比差距很大，过长的历史数据会对我们发现市场的alpha造成极大的困难。既然如此，那么我们尝试使用221大数据库提取股票Tick级数据

```
### specify the contracts
contract <- code

for (i in 1:code_length) {
    #all.files <- list.files(pattern='*\\.csv',recursive=TRUE)
    the.files <- grep(pattern=contract[i],all.files,value=TRUE)

    if(any(file.info(the.files)$size == 0)) {
        cat("zero bytes\n")
        zero.files <- which(file.info(the.files)$size == 0)
        cat(zero.files,"\n")
        the.files <- the.files[-zero.files]
    }

    ### read csv into a list temp, using Encoding GBK to solve the Chinese issue
    temp <- lapply(the.files,function(x) read.csv(x,header=T,stringsAsFactors=F,encoding="GBK"))
    #x<-do.call(rbind,temp)
    x<-rbindlist(temp)
    x <- data.frame(x)
    ### Save the data
    colnames(x) <- c('exchange','symbol','time','price','opi','deltaopi','volamount',
    'Volume','openopi','closeopi','opitype','opidirection','bid1','ask1','bidsize1','asksize1')
    # eliminate exchange,symbol,volamount
    x <- x[,c(-1,-2,-7,-11,-12)]
    # create xts
    xtsindex <- as.POSIXct(as.character(x$time),tz="Asia/Shanghai",origin=ISOdatetime(1970,1,1,0,0,0))
    x.xts <- as.xts(x[,c(-1)],order.by=xtsindex)
}
```

- 从2015年至今,沪深A股Tick级别数据共51G大小. 然而股票并不是期货，无法在Tick级别从事交易(日内回转交易另当别论). 若将Tick级别数据转换成2分钟，15分钟，半小时，一小时K线组合，依然无法在实践中产生意义，毕竟我国是T+1交易制度的。

Subsection 2.Thinking Out of the Box

我们面临的困境是 1. 数据量极度短缺，无法进行充分有效的机器学习。2.没有合适的训练集进行训练

二级市场并不是围棋，图片，声音，可以直接应用Neural Networks 或 Deep Neural Networks。我们很难定义二级市场的优与劣，胜与负，因为参与者是各向异性的，个人效用函数和交易周期都不相同。可以说是大家在不同维度上面的一种博弈，何况如Section 1描述，我国市场又具有如此多的特性。

问题的本质是一种博弈，是策略间的对抗。市场结构瞬息万变，每个时点适用的策略不尽相同。试图完全解构市场是不切实际的，我们只能坚持自己的策略，控制下行风险，享受**概率上的不确定性**。

那么问题很显然了，我们用来学习的数据应该是市场上各个主体使用的策略，每个策略本质上相当于一个弱分类器。市场行情本身是没有意义的，其背后各个主流策略的表现是有意义的。如果直接从市场的价格，因子构建策略，也就是构建弱分类器，会出现大量的噪音，因为市场的维度太高了

Subsection 3. Factor and Strategy Construction

- 接下来让我们构建新的因子及策略，让这些策略能够广泛的代表当下市场的主要参与者。这里除了221大数据股票Tick级数据，也用到了221爬虫组获取的每日龙虎榜数据

1.市场温度

- 根据当前时点的涨停数量，非一字板数量及封板时间，开板数量及开板时间，跌停数量综合计算得到的当前时点的市场温度因子。度量了当前市场的热度。(过滤掉新股与次新股板块)

2.赚钱效应

- 根据昨日打板成功率及盈利率得到的赚钱效应。体现了市场上最激进的游资与跟风盘的盈利情况。(过滤掉新股与次新股板块)

3.一线游资关注个股，同城营业部协作

- 根据龙虎榜推算出的一线游资关注个股，同城市营业部协作，对倒的行为。根据游资的实力与关注度加权。体现了市场风口的动量效应。(过滤掉新股与次新股板块)

4. 高送转异动

- 如前所述，高送转板块值得单独作为一个重要的因子与策略

5. 近五日内盘中出现托单，压单，对倒单的总数

- 日内主力会进行托单，压单，对倒的行为。五日的累积数量体现了该股被主力关注的程度。(过滤掉新股与次新股板块)

6. 超跌妖股

- 关注超跌的曾经炒作过的龙头股票。当盘中无热点的时候，主力会选择拉升这些有着妖股记忆的股票

7. 二八转换

- 关注市场出现二八转换的情况，发现国家队的护盘行为，本质是选择出的一篮子权重股票与一篮子热点股票的比价。

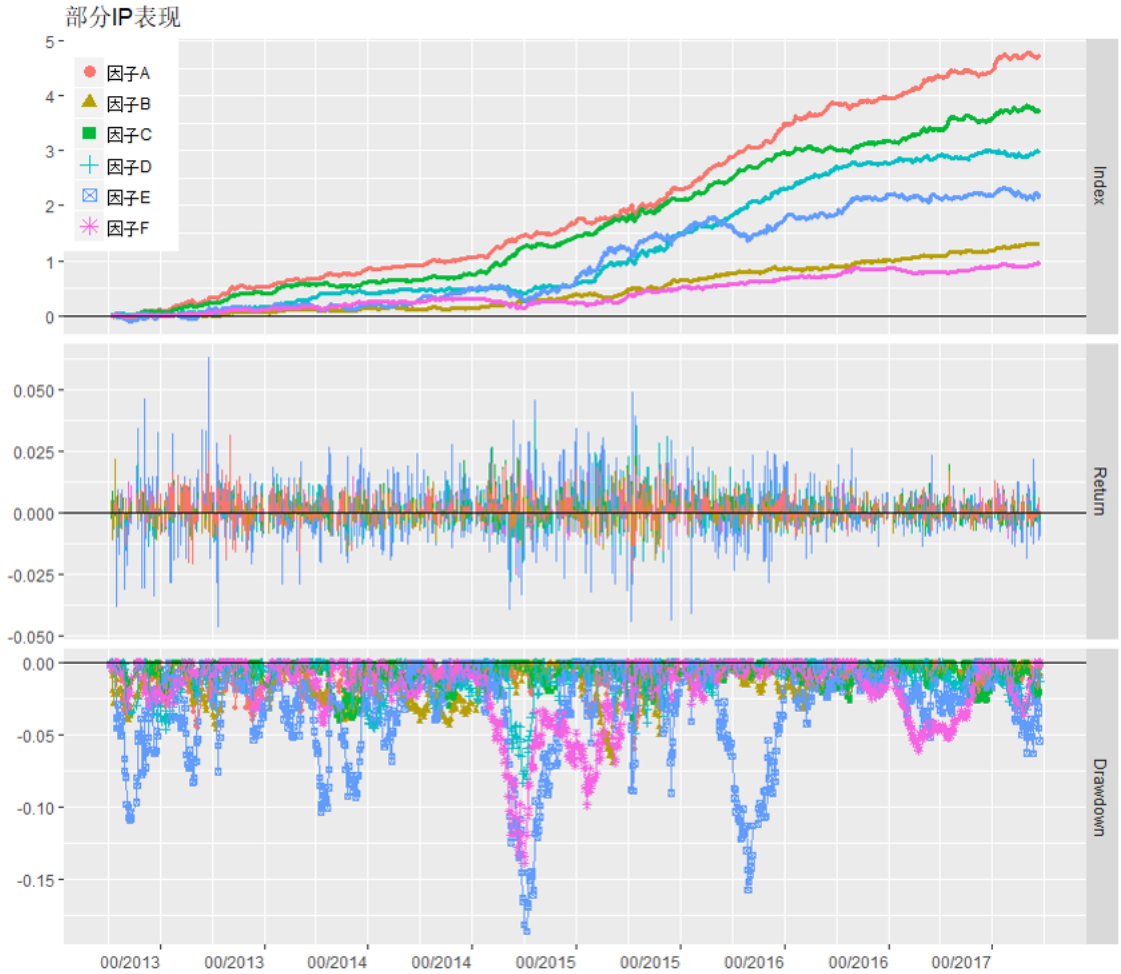
8. 热点轮动

- 近一年出现的热点股票会反复炒作，呈现出一种轮动特性。本质是选择出的各个热点的龙一，龙二，龙三标的。

9. 新股与次新股

- 只关注新股与次新股板块，观察这里的强博弈情况。

- 通过以上因子构建新的交易策略，共六大类用来模拟游资的交易模式，表现如下。纵坐标 1,2,3,4,5 分别表示年化100% 200% 300% 400% 500%.



- 曲线的特征也反应了游资激进的风格.且该策略回测需要Tick级别的数据来模拟其强博弈的特性。比如我们某策略可以设定规则今天买入，第二天冲级涨停失败回调至7%时即抛出。

Section 3. 人工智能算法尝试

Deep Neural Network

我们采用Deep Neural Network combined with Gradient Boosting的方法，对策略(因子)进行机器学习，从而进行预测，即动态的选择因子。

我们的输入是传统因子库中100个有代表性的因子与221自行构建的几十个因子，还有上述若干策略及经典多因子策略(用来代表机构当期的收益表现)。

如果我们把单因子也当做一个策略，其实我们的输入是能够代表市场参与人群的众多策略。本质上这是一种Bayesian的，heuristic的研究方法。先验策略的构建体现了工程师的核心价值。

之后用来训练的输出是策略的超额收益，我们同时进行Rolling Training.

尽管我们输入的是策略而非因子，数据量依旧略显单薄，所以我们需要做Data Augmentation的工作来扩展数据量。本篇采用PCA Jittering的方法，对沪深A股行情进行处理。优秀的策略，能够抵抗这种概率事件的发生，体现一种Robustness. 这是因为策略的最大回撤的限定，能够抵御行情一定程度的随机性。相对的，简单的单因子策略就会因为这样的操作而给出极差的表现，使得优秀的策略得到奖励。这就是输入为策略而非因子的优势。

Gradient Boosting Machine

Deep Neural Network(DNN)的在学习控制着向量变换方式的权重矩阵后，接下来的问题就是如何学习每一层的权重矩阵。

本篇采用Gradient Boosting的方法.

DNN每一步需要比较预测值与目标值，根据两者的差异来更新每一层的权重矩阵。因此需要设定目标函数，也称为损失函数(Loss Function). 如果用的是梯度下降的方法(Gradient Descent)，通过使当前点对应梯度下降的方向移动，来降低损失。这里有个小trick，基于二级市场本身的随机性，梯度下降的速率采用随机梯度下降Stochastic Gradient Descent (SGD)的方式，赋予整个网络以一定程度的随机性。

Boosting

Boosting是一种思想。早在2006年笔者在燕园撰写论文时便已兴起，如Adaboost等等。

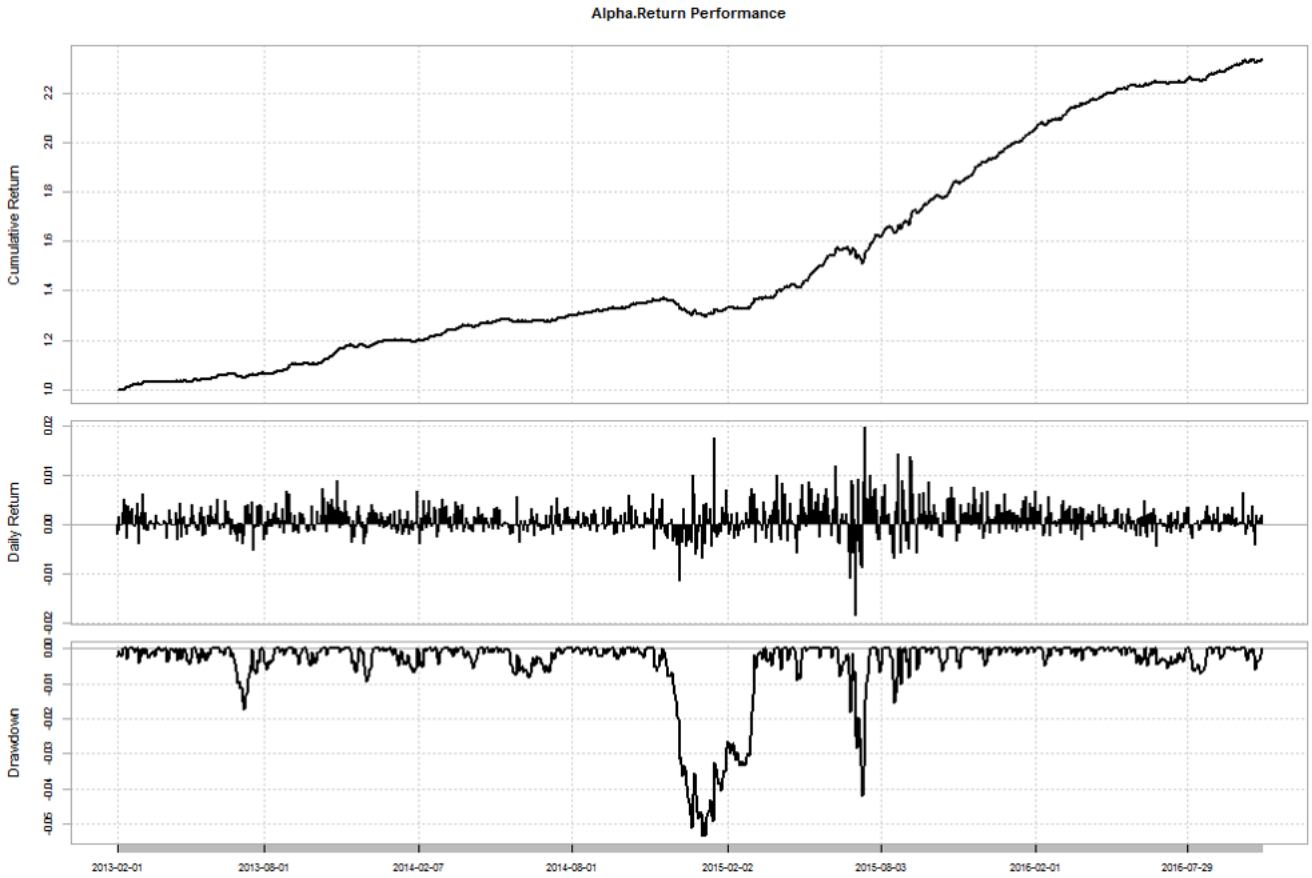
Boosting本质是惩罚错误，每次学习要对错误加权。之后得到若干弱分类器的线性组合。就像是一个人学习的过程，开始学一样东西的时候，会去做一些习题，但是常常连一些简单的题目都会弄错，但是越到后面，简单的题目已经难不倒他了，就会去做更复杂的题目，等到他做了很多的题目后，不管是难题还是简单的题都可以解决掉了。

具体到我们的问题，其实每个因子或策略本身是一个弱分类器(weak learner), boosting的思想用在这里非常合适。

那么Boosting的思想配合Gradient的惩罚方法，就是Gradient Boosting Machine.

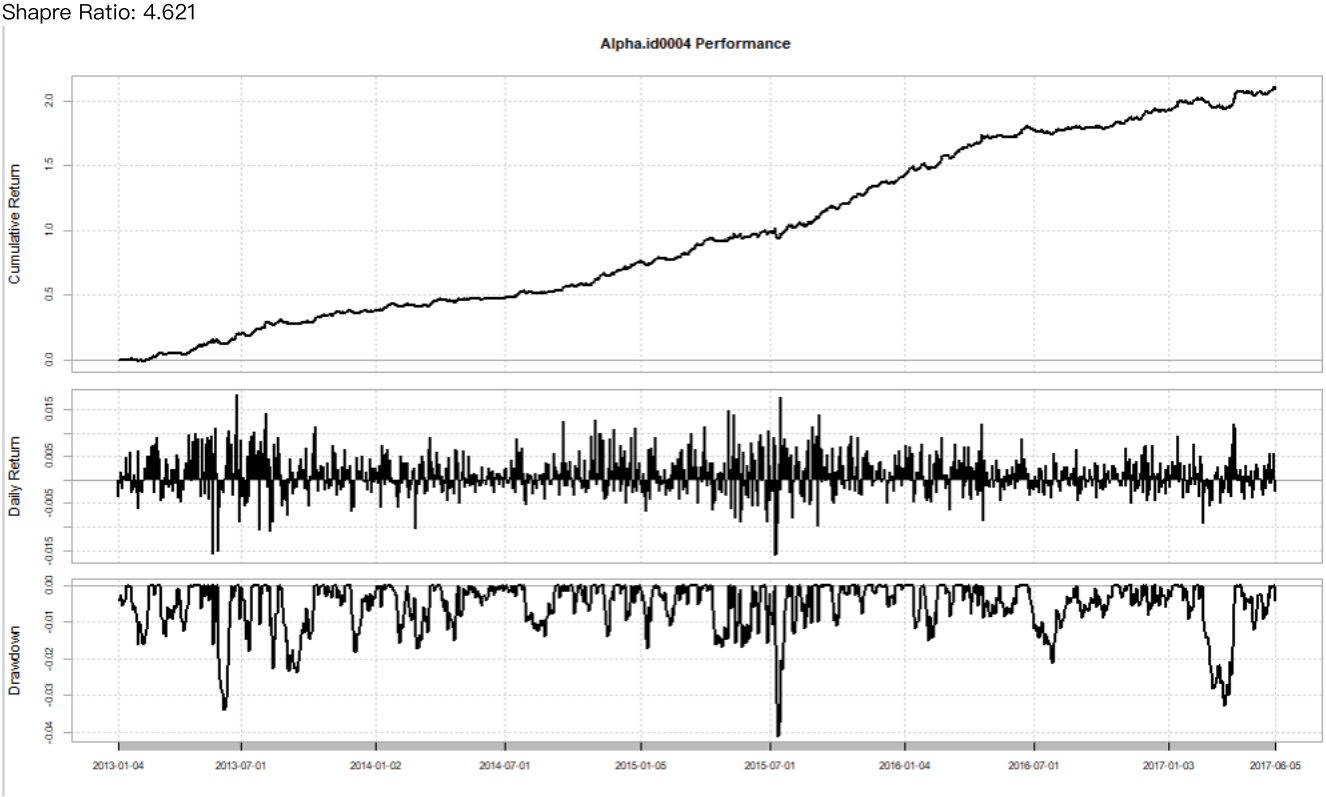
Section 4. Results

- 策略1 笔者在2016年研究的策略。
- Shapre Ratio: 3.417
- Calmar Ratio:2.389
- Sortino Ratio: 0.572

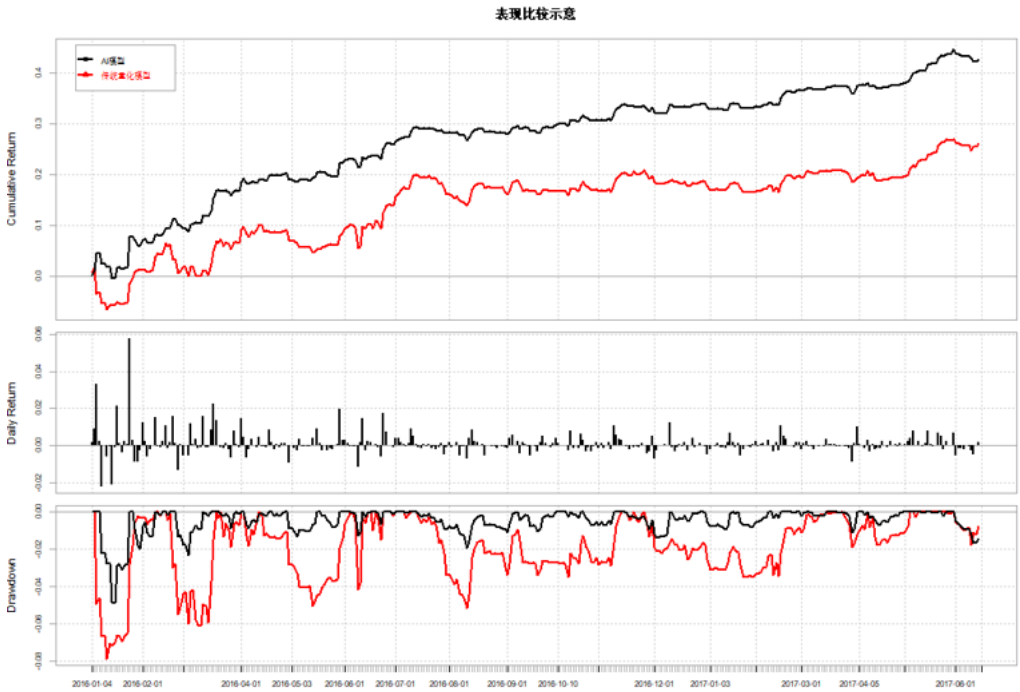


- 策略2 本次研究的结果 曲线更加平滑

- ○ Shapre Ratio: 4.621



- CTA策略 篇幅限制本文未提及我国商品期货市场。同样算法用于CTA市场的回测曲线



-

	..1	DL1	DL2	DL3	GBM1	RF1	DL5	DL6
2016-01-04	-0.63	0.9857929	-0.9423207	0.8846074	0.7411316	0.1950323	0.8864449	-0.7701291
2016-01-05	0.25	0.4777664	-0.5301601	0.9275566	0.6777024	0.3120125	0.9399622	0.7277877
2016-01-06	0.58	-0.9949881	-0.8796300	-0.9814105	-0.5927105	-0.3846154	-0.3508014	-0.8910408
2016-01-07	0.22	0.8124298	0.1985840	-0.2607375	-0.4069722	0.4538578	0.8987049	0.4404571
2016-01-08	0.32	0.8088368	0.1673667	0.8943529	0.2794076	0.0798722	0.7768374	0.9856895
2016-01-11	0.40	0.9898540	-0.1003699	0.9076746	0.4975551	0.4538578	-0.5649340	-0.8192966

	DL7	DL3.1	DL1.1	DL4	DL5.1	GBM1.1
2016-01-04	-0.94780338	-0.9999948	0.74398264	-0.5859892	-0.6434350	0.4872827
2016-01-05	-0.95559698	0.9981320	0.71643409	0.3253182	-0.3488607	0.6571112
2016-01-06	-0.98836421	-0.9995106	0.09634899	0.4836077	-0.6446489	-0.9850307

Section 5. Conclusion

主观投资者需要同时面对强大的市场机制和人性的弱点，这样很难持续高胜率的交易。

传统量化手段的开发时间至少是3个月，后期维护升级成本非常高，任何的‘升级’
都可能有策略漂移的风险。

策略漂移是行业内的禁地，所以几乎没有听过有持续升级的对冲基金。

AI开发平台对数据的前端处理要求很高，因为包含的策略多，策略漂移的风险就小，后期的开发/维护/升级效率很高。

不断地升级也就意味着AI模型更有可能在瞬息万变的金融市场持续的捕捉规律