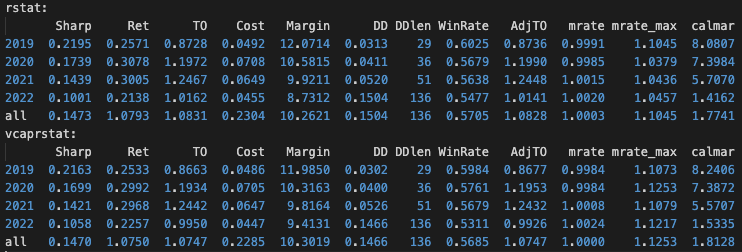
# 2024.01.08

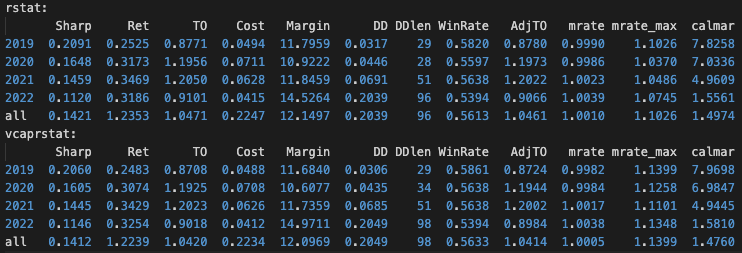
本周计划

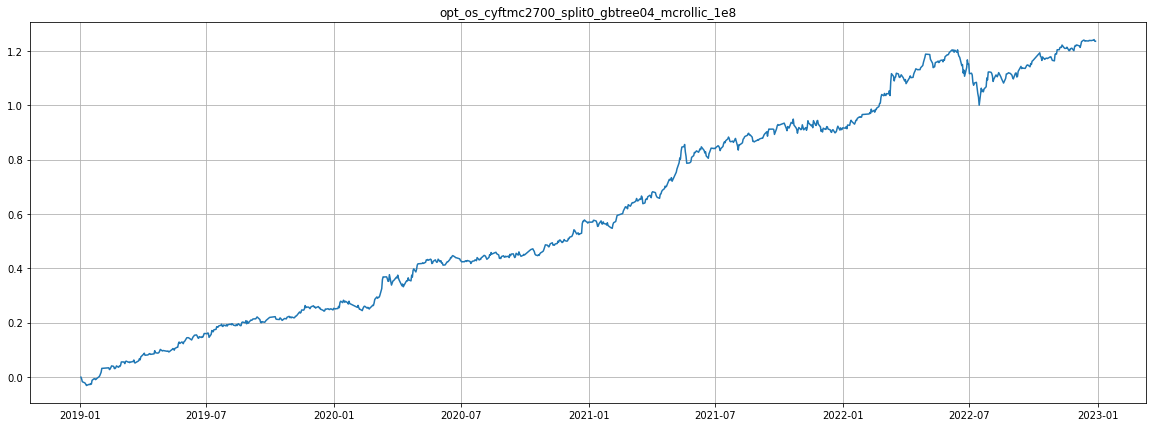
1. 2700因子的三个chunk全部筛完了，3个chunk也已经拼接起来又筛完了。目前获得了一个较优的因子集合，回测如下：
   1. 
2. 接下来是开始在一个较优的因子集合上进行模型的随机性训练并多模型融合以使结果稳定。
   1. 每棵树使用随机50%的特征
   2. 每棵树随机选取50%的数据
   3. 使用不同的随机种子迭代平均
3. 尝试加入品种信息
   1. 根据sigtot获得品种之间的相关性，将相关性当成余弦相似度，以此使用MDS方法获得每个品种的坐标编码。以此获得连续坐标
      1. 一维连续坐标
      2. 二维连续坐标
4. 尝试引入截面上的关联性以提升模型表现
   1. 因子的截面zscore
   2. 因子的截面rank

## 2700个因子的筛选结果

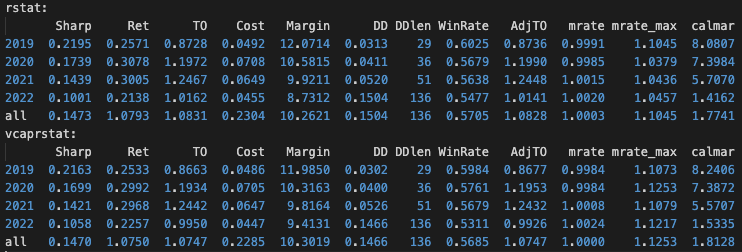
之前将2700个因子均匀分成三批，每批900个进行筛选。

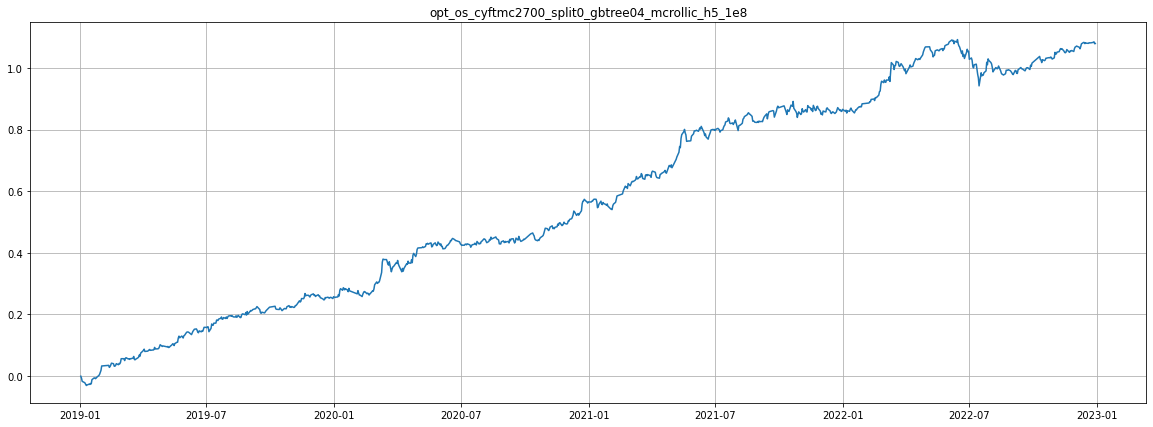
在chunk1上观察到模型和有不错的表现，夏普很高，每年模型表现很稳定，2022年夏普也能上0.1。下图是其样本外表现：





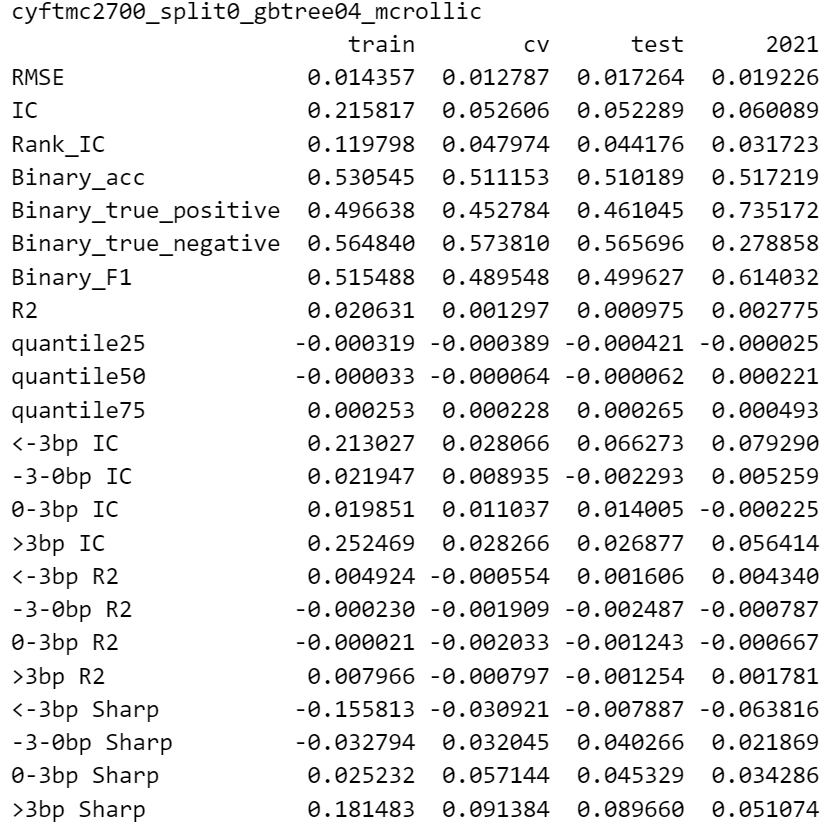
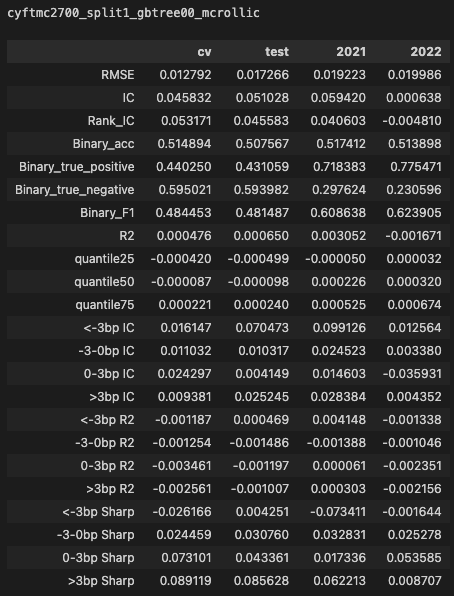
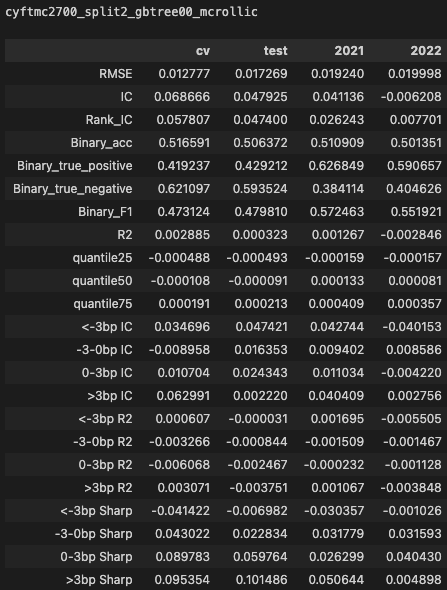
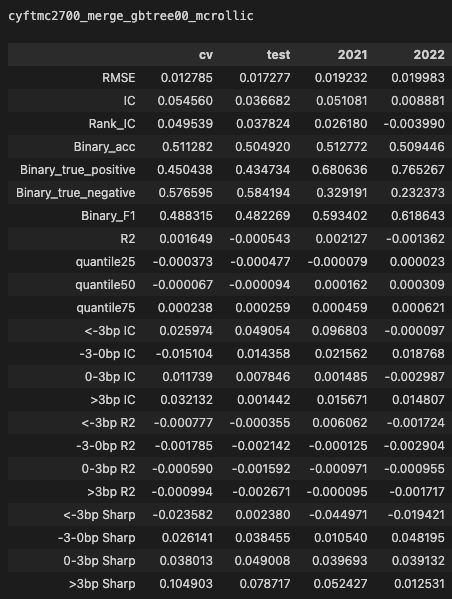
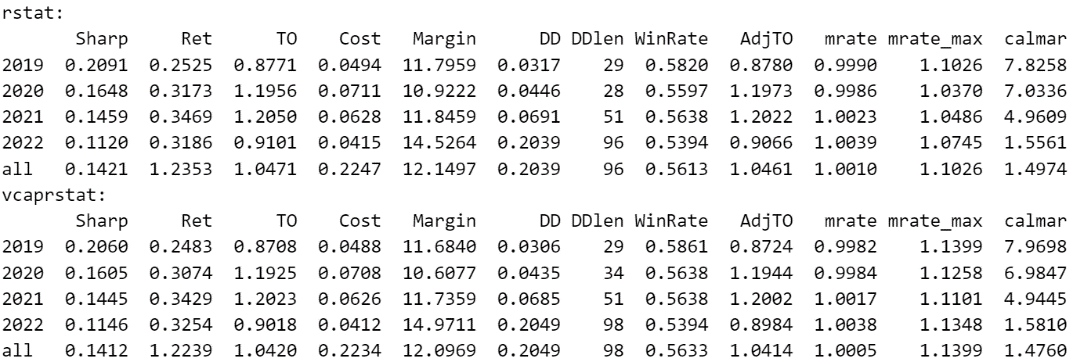
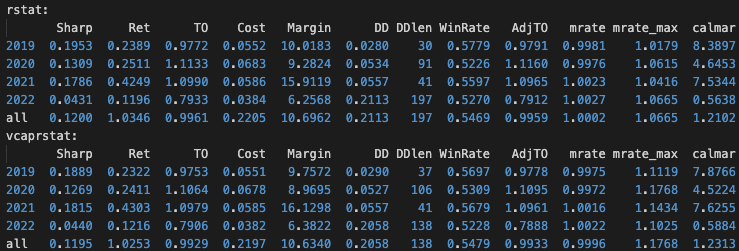
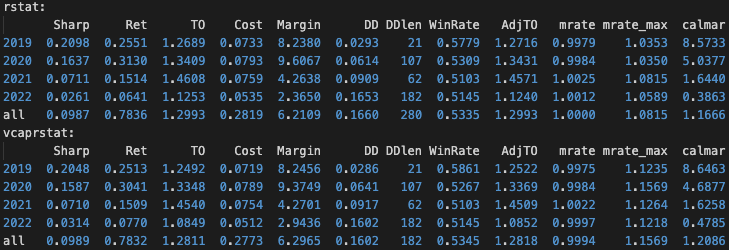
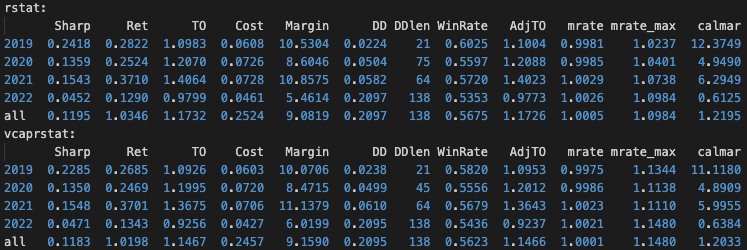
在调整了hconstrain=0.5之后，模型的回撤降低到了0.15以下，夏普仍在0.1以上。





可见，hconstrain主要是通过降低收益来减小回撤。

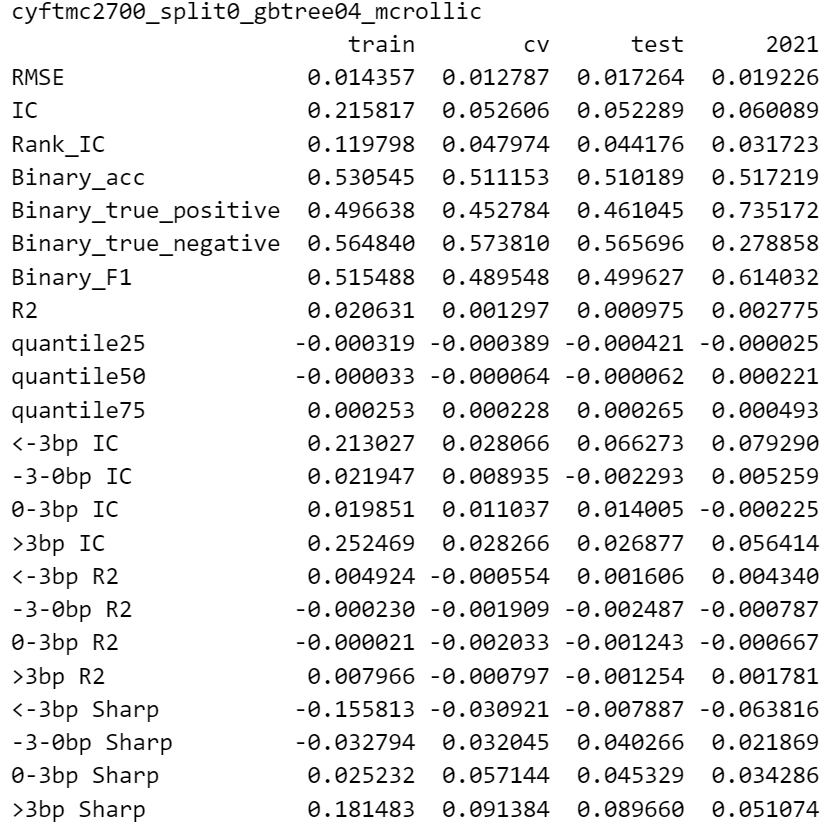
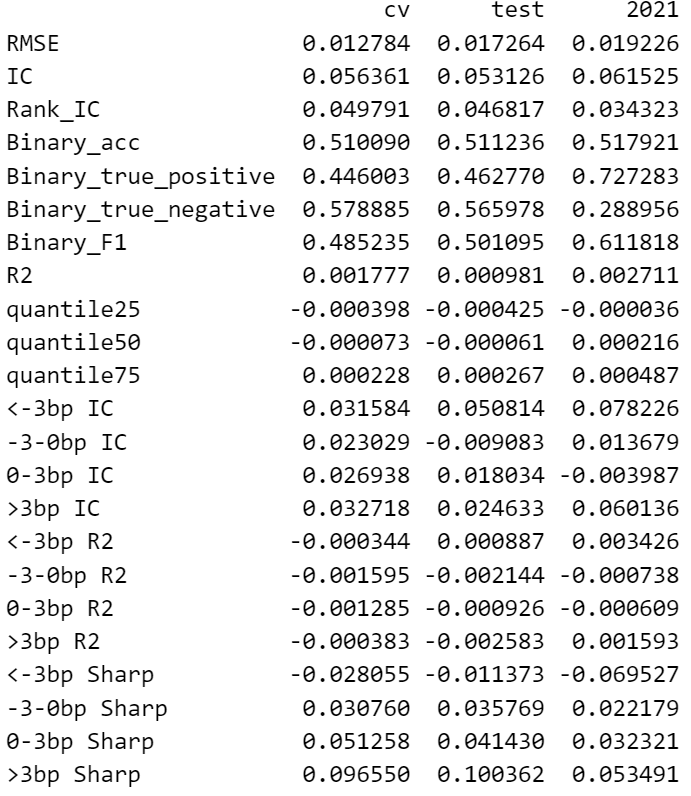
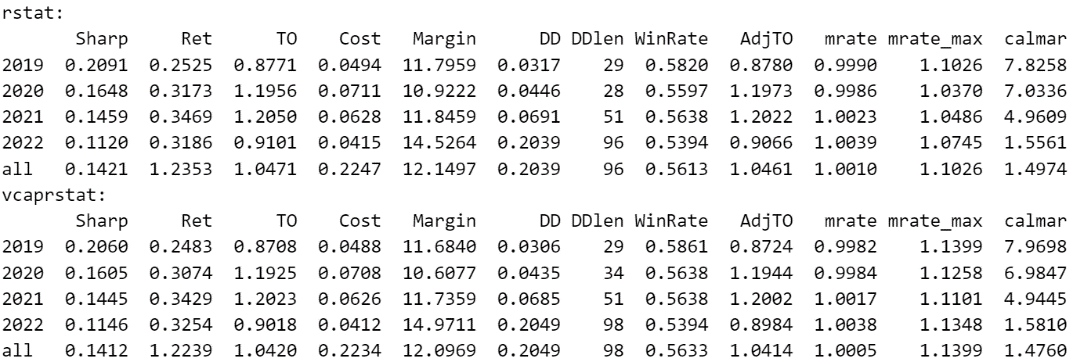
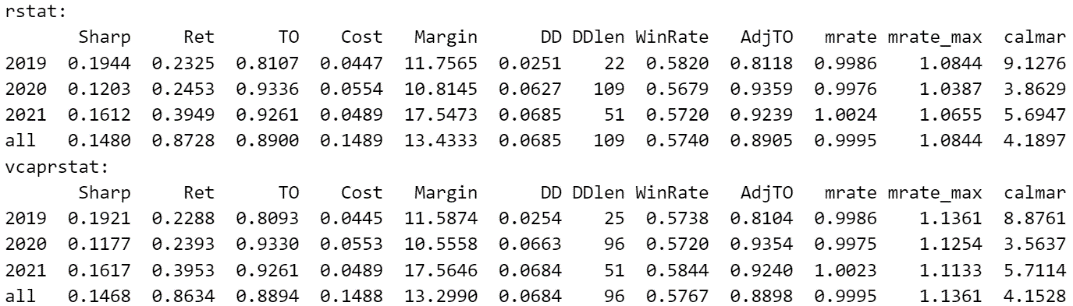
其余的chunk以及merge的表现如下

1. 统计指标
   1. chunk1
      1. 
   2. chunk2
      1. 
   3. chunk3
      1. 
   4. merge
      1. 
   5. 可见，根据统计指标，merge之后的结果均不如三个chunk单独的表现。尤其是2020年表现略差
2. 回测
   1. chunk1
      1. 
   2. chunk2
      1. 
   3. chunk3
      1. 
   4. merge
      1. 
   5. 根据回测结果，chunk3表现最差，chunk次之，但二者都在2022年烂掉了。将三者拼在一起按相同规则筛选，2022年表现也较差。
      1. 说明chunk 1 2 3的因子本身有差异，应当寻找到这种差异到底在哪
         1. 可能是在21-22年，存在一些因子在21年表现很好但是22年表现很差，导致在筛因子的过程中留下来21年好的但22年不好
            1. 检验方法：查看每个因子在21和22年的夏普，并查看chunk 1 2 3有何差异
            2. 解决办法：这不是筛因子的能解决的问题，需要用因子融合的方法。

一个合理的方式是按照名字排序，每10个因子训练一个线性模型，以线性模型的输出作为特征再用树模型训练

## 增加随机性取平均

在chunk1上取sub\_sample = 0.5, colsample\_bytree = 0.5，以不同的随机种子训练是个模型取平均，和原本结果对比

1. 统计指标
   1. 不随机化
      1. 
   2. 随机取平均
      1. 
      2. 可以看到统计指标都变好了
2. 回测表现
   1. 不随机化
      1. 
   2. 随机取平均
      1. 
      2. 但是回测结果却略差