**岭回归特征筛选、rolling测试**

1. **先放结论与分析**
   1. 对于特征筛选，我尝试了3种方式：cv筛特征；train筛特征；train的最后10%筛特征。不妨分别记为cv, tr和tr10
   2. 对于这三种特征筛选，我分别也测试了是否过滤掉train IC小于0的因子。过滤掉记为f，没过滤则无符号
   3. 对于这三种特征筛选，我也尝试了模型是否在筛选后的特征内进行rolling训练（不是rolling筛特征）。如果rolling，则记为r。
   4. 这样就可以给我的这些模型命名。例如：cv-f-r则代表在cv上筛出的特征上，过滤掉其中IC小于0的特征，并rolling训练；tr10-r则代表在train的最后10%筛特征，不做过滤，并rolling训练。以此类推。

* 1. **筛选特征的方式**
     1. 无f无r：cv > tr > tr10



* + 1. 有f无r：cv > tr10 > tr



* + 1. 无f有r：tr > cv > tr10



* + 1. 有f有r：cv > tr10 > tr

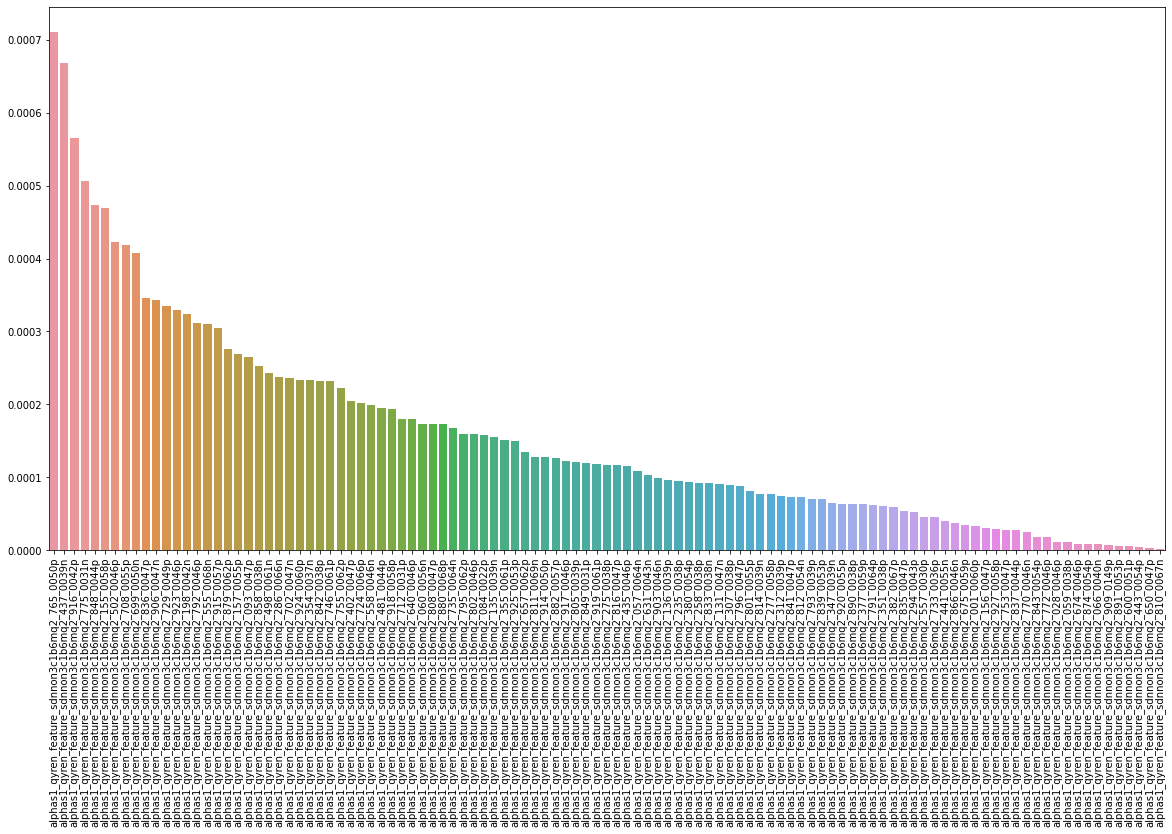


* + 1. 总结：cv ~ tr > tr10
  1. **是否过滤掉IC<0的因子？**

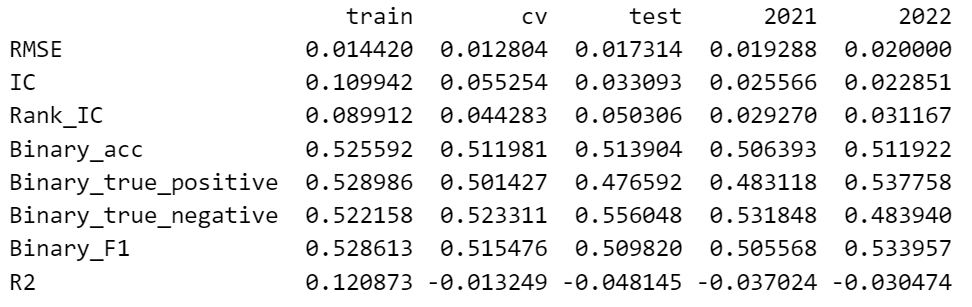
可以发现，在岭回归上，过滤掉IC<0的会导致表现均不如原来。

* 1. **Rolling前后对比**
     1. 对于没有f的来说
        1. Rolling会让结果略微变好，但幅度不大
     2. 对于有f的来说
        1. Rolling对结果的影响很难界定好坏
  2. **回归的特征重要度分析**

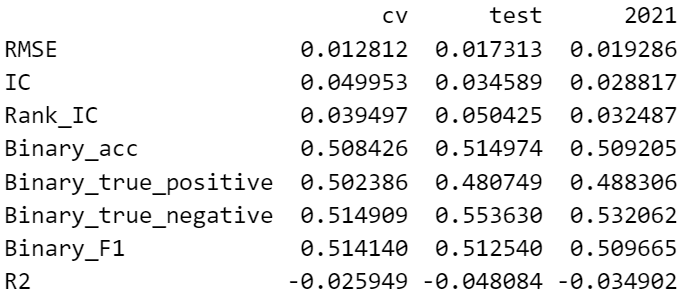
对tr的岭回归做了特征重要性分析，如下图所示：



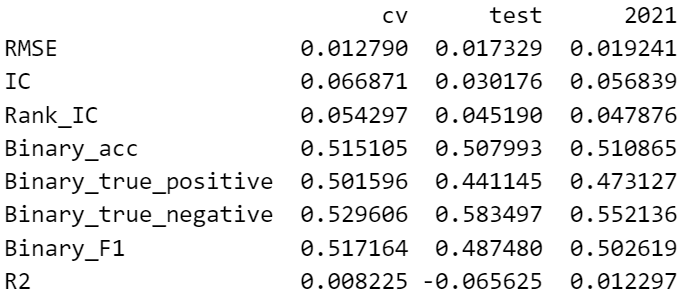
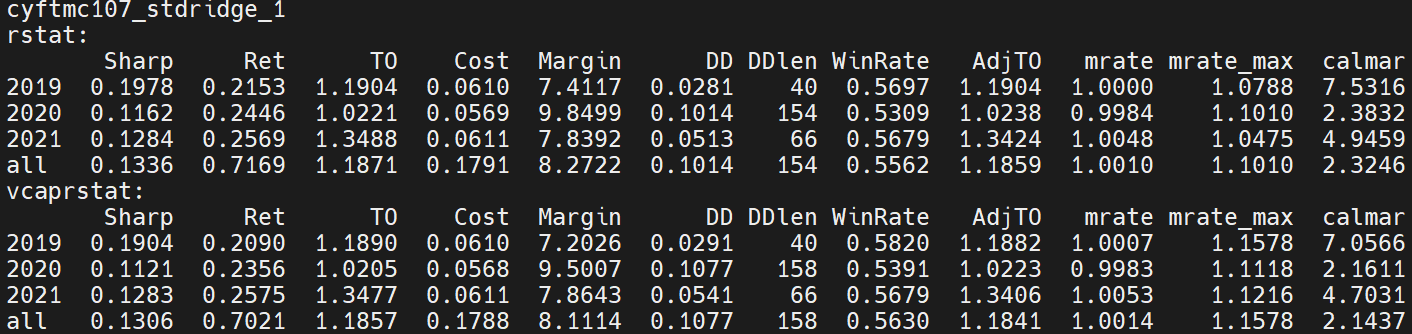
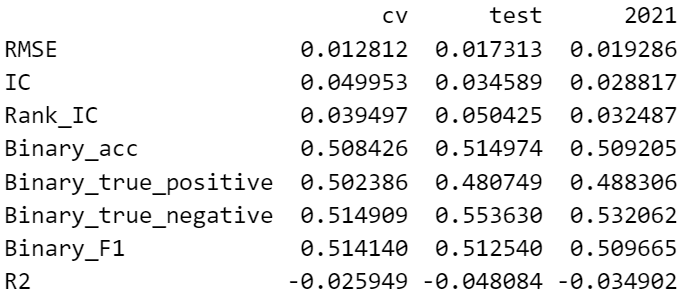
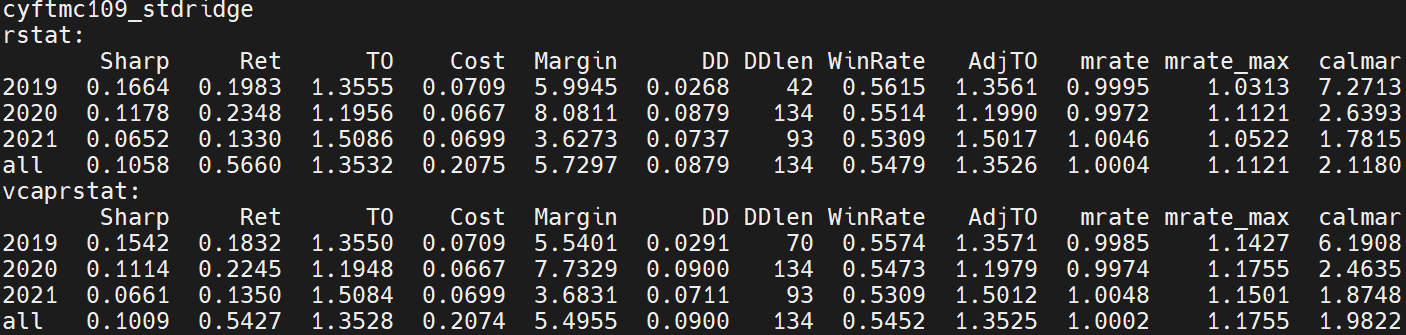
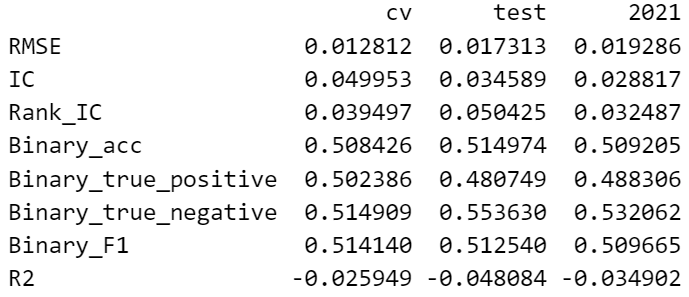
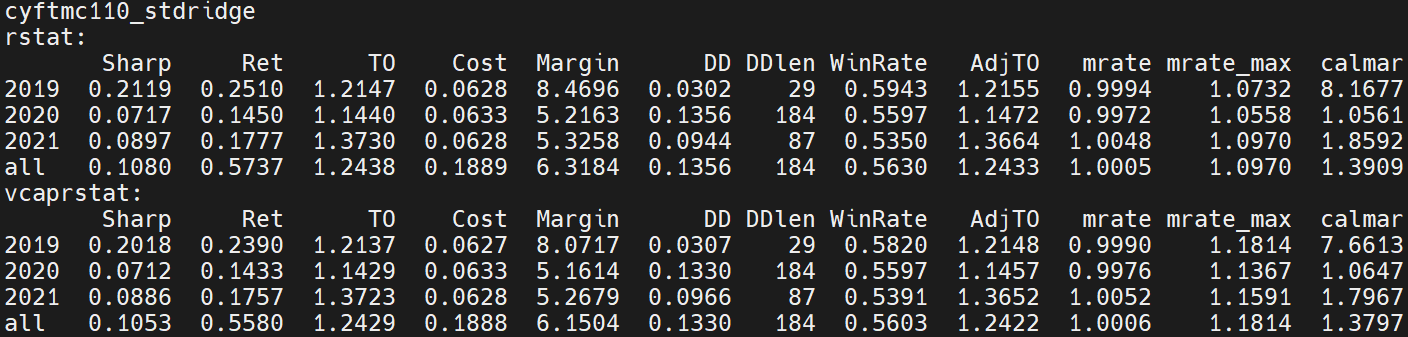
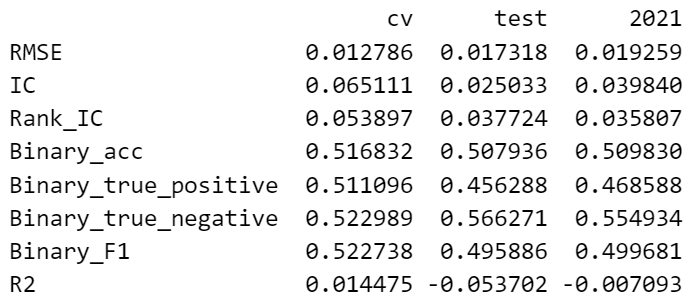
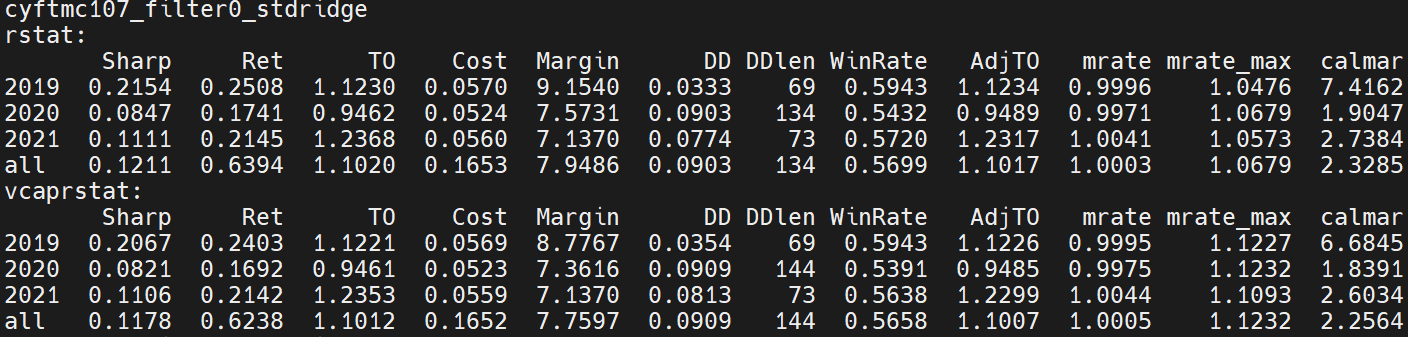
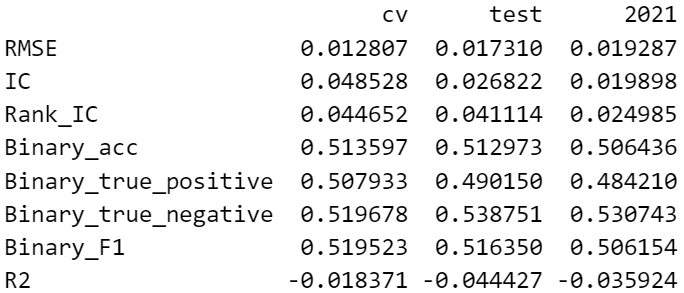
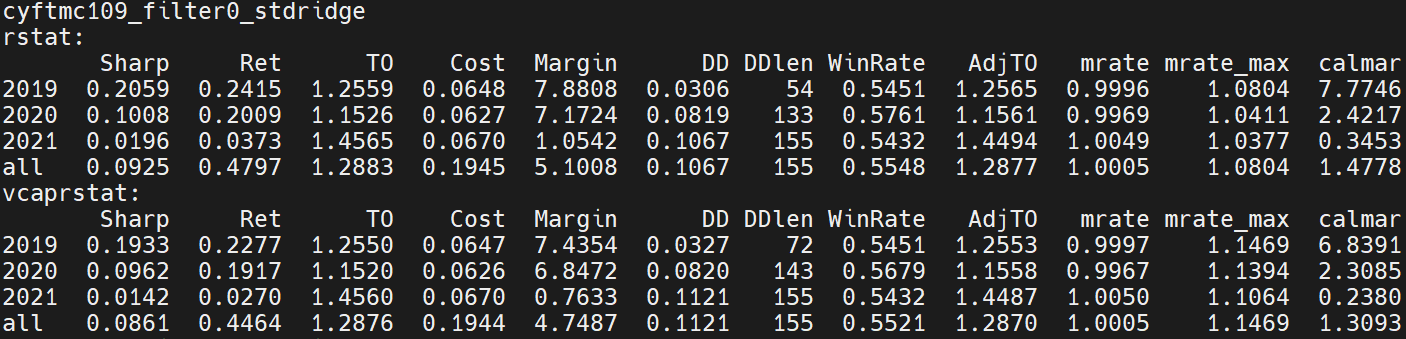
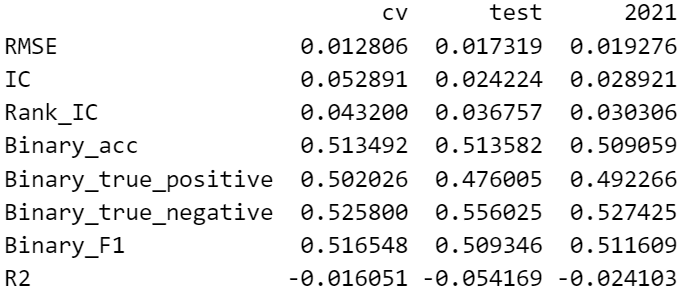
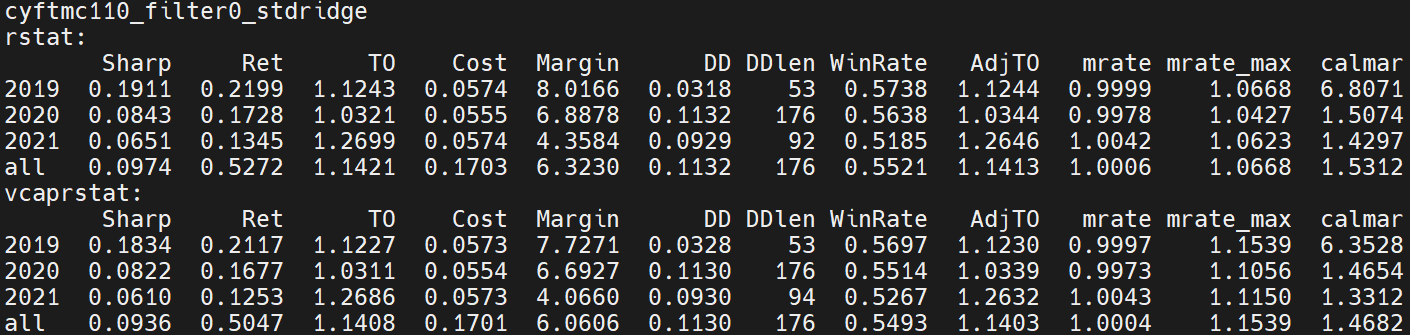
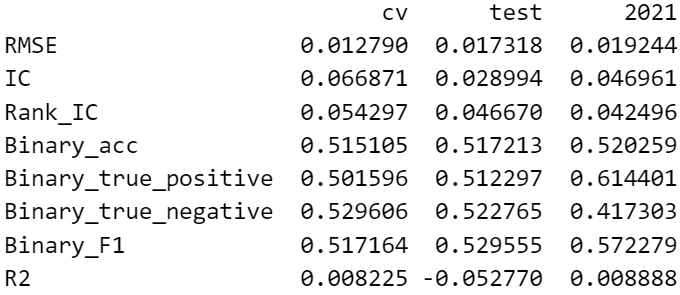
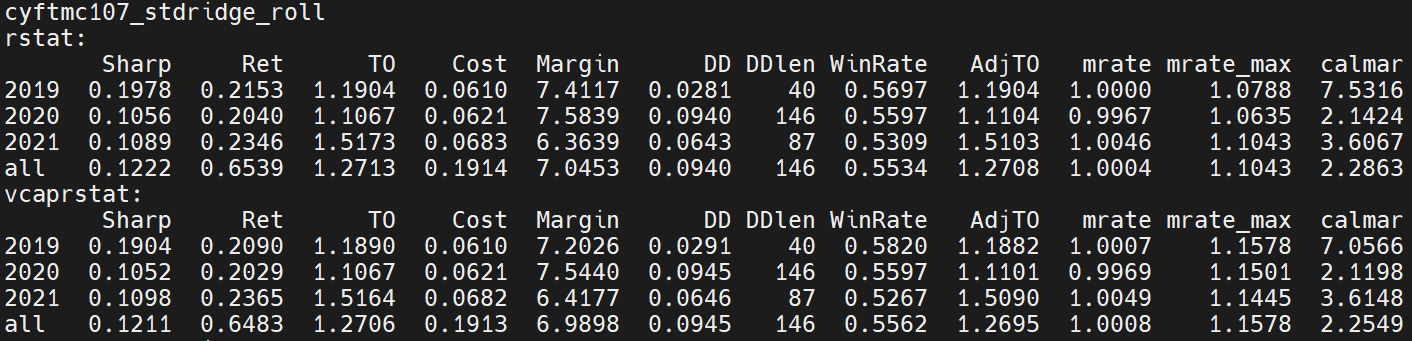
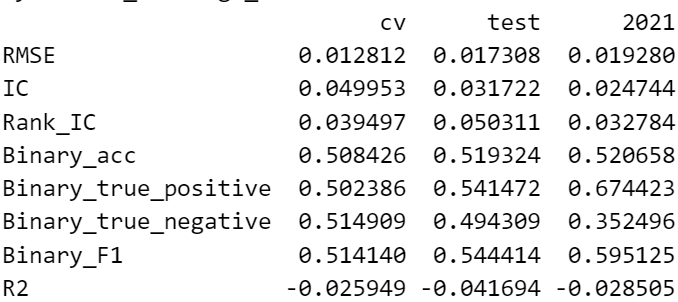
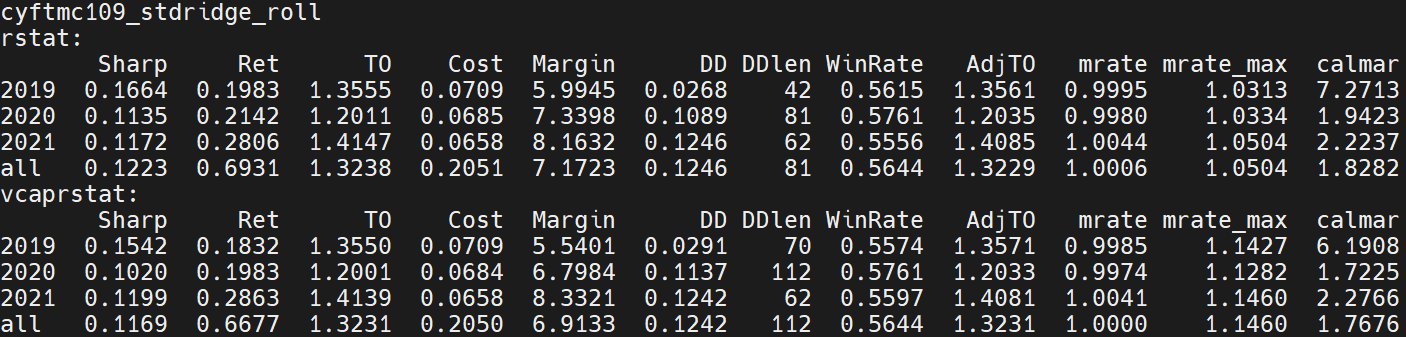
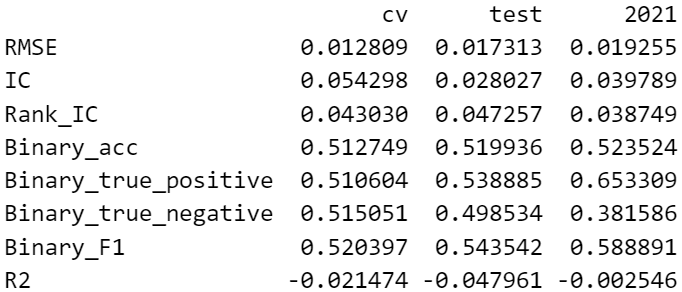
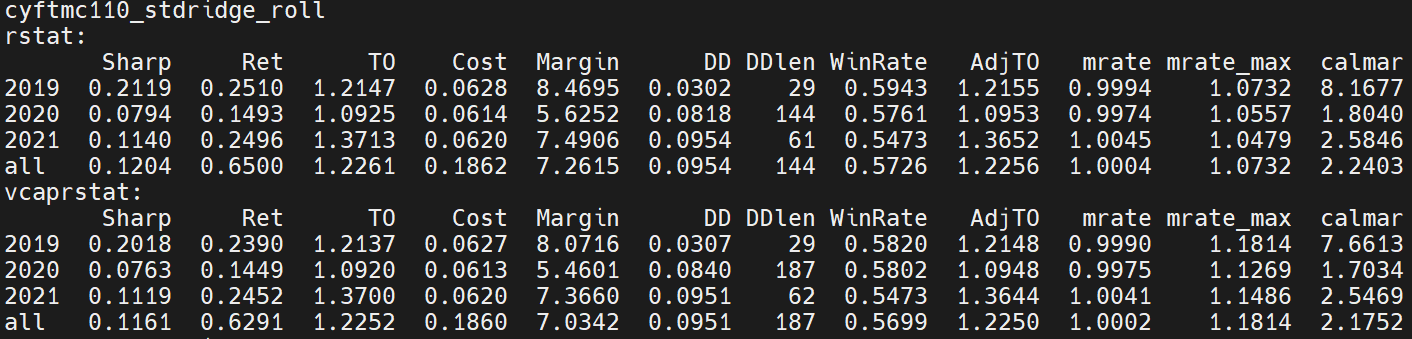
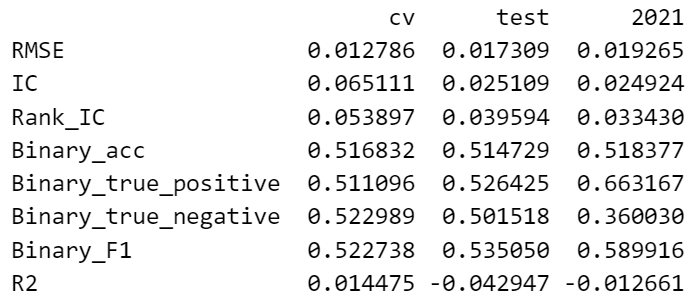
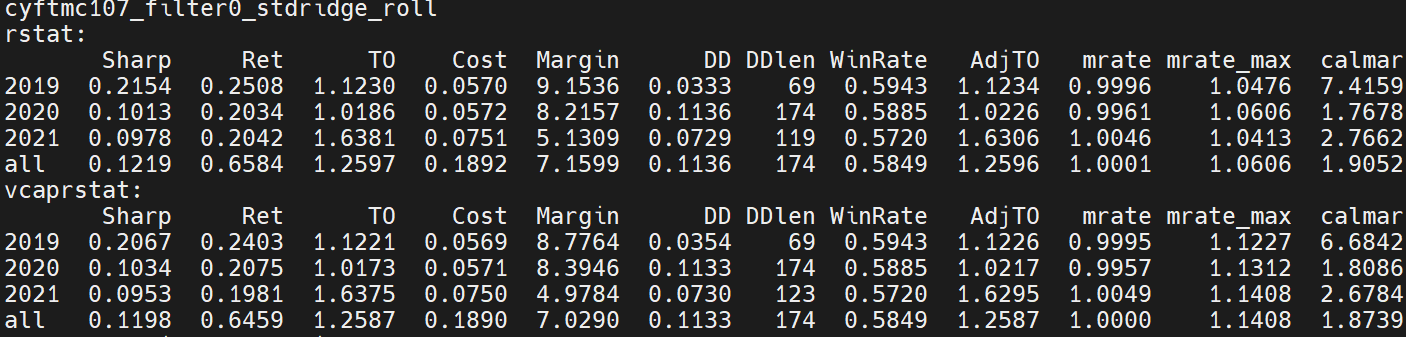
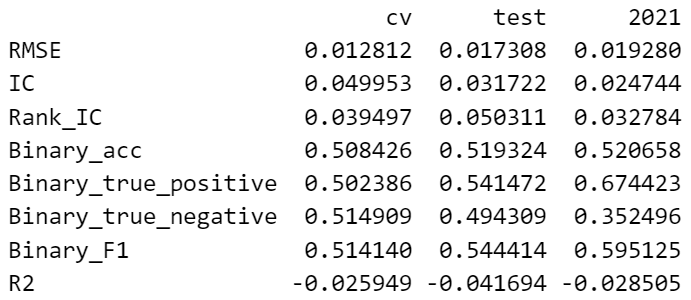
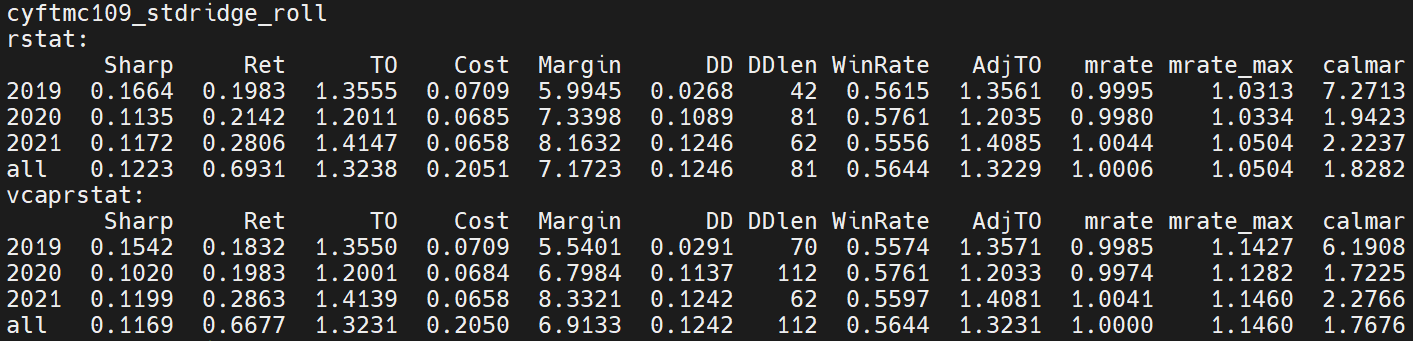
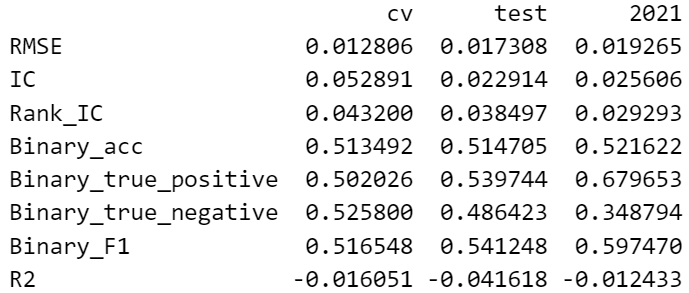
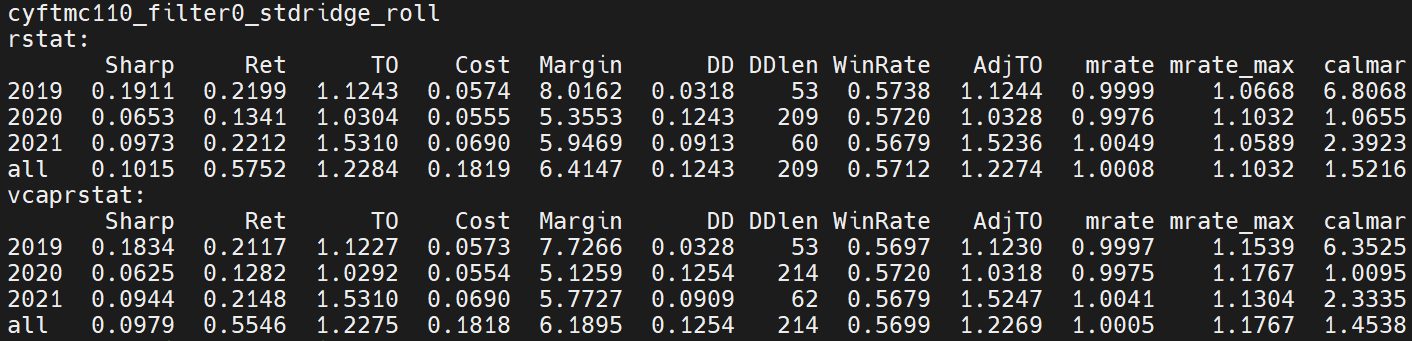
只保留其中大于0.0001的特征，其余略去，重新训练模型得到：



和原结果相比，变化不大（下为原结果）



这说明通过特征重要度删除不重要的特征，对线性模型没有较大影响。

1. **再放数据**
   1. 以下是结果罗列：
   2. **cv**
      1. 
      2. 
   3. **tr**
      1. 
      2. 
   4. **tr10**
      1. 
      2. 
   5. **cv-f**
      1. 
      2. 
   6. **tr-f**
      1. 
      2. 
   7. **tr10-f**
      1. 
      2. 
   8. **cv-r**
      1. 
      2. 
   9. **tr-r**
      1. 
      2. 
   10. **tr10-r**
       1. 
       2. 
   11. **cv-f-r**
       1. 
       2. 
   12. **tr-f-r**
       1. 
       2. 
   13. **tr10-f-r**
       1. 
       2. 
   14. **总结**
   15. 