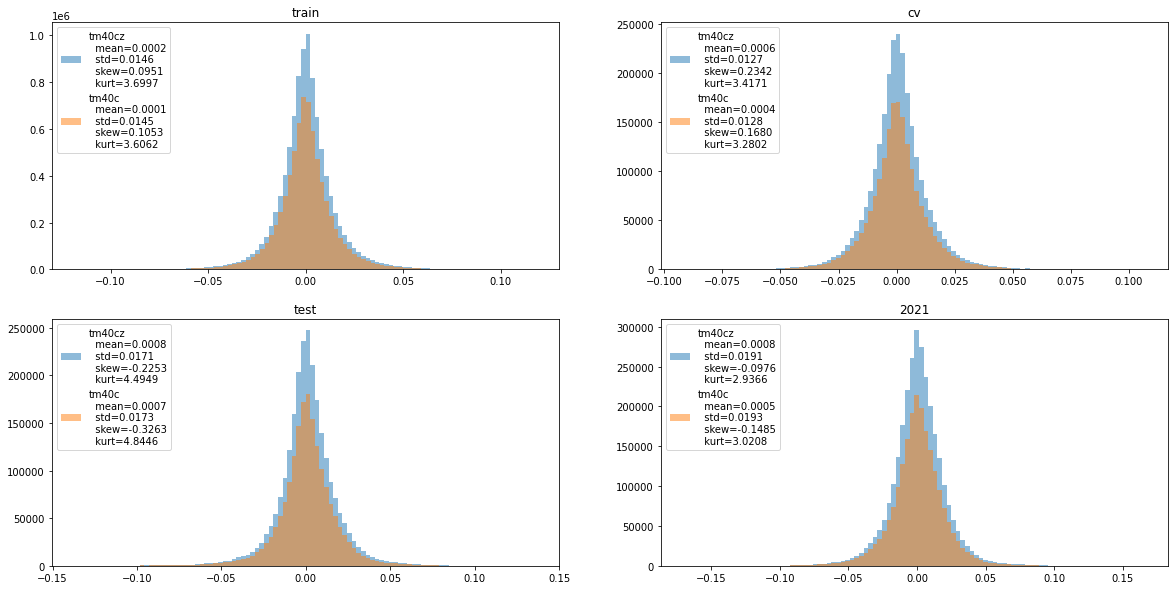
tm40c和tm40cz的差别

崔晏菲 2023.12.08

1. **数据总览**
   1. 数据量差异

|  |  |  |  |
| --- | --- | --- | --- |
|  | tm40cz | tm40c | 变化率 |
| train | 9652635 | 7209240 | -0.25313 |
| cv | 2572751 | 1850796 | -0.28062 |
| test | 2447595 | 1781010 | -0.27234 |
| 2021 | 2962710 | 2147340 | -0.27521 |

* 1. 分布差异
     1. 对y的分布差异查看

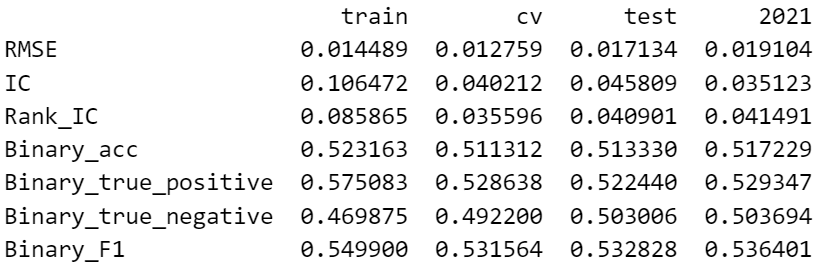


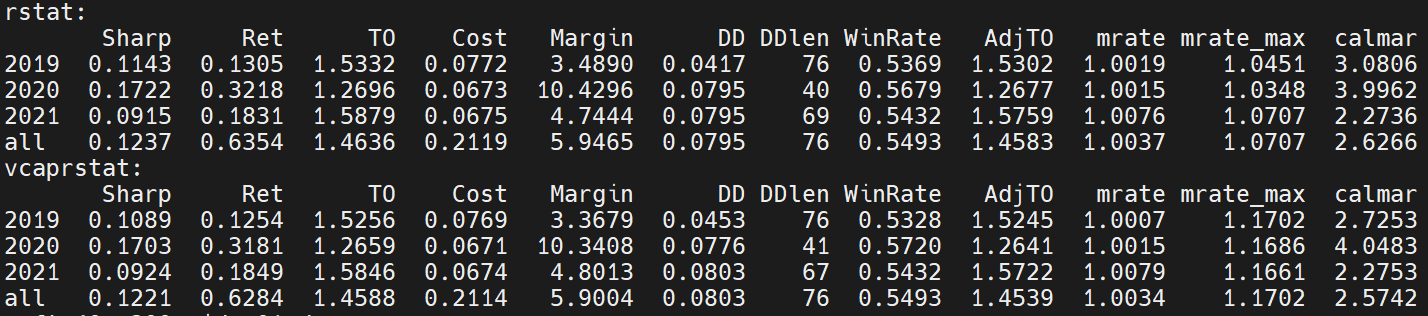
对均值变化的检验

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | tm40cz | tm40c | 变化率 | t统计量 | p值 | 结论 |
| train | 0.000193 | 0.000064 | -0.66839 | -18.0848 | 4.21E-73 | 均值变小 |
| cv | 0.000605 | 0.00042 | -0.30579 | -14.9997 | 7.40E-51 | 均值变小 |
| test | 0.000825 | 0.000659 | -0.20121 | -9.82131 | 9.12E-23 | 均值变小 |
| 2021 | 0.000792 | 0.000494 | -0.37626 | -17.3847 | 1.08E-67 | 均值变小 |

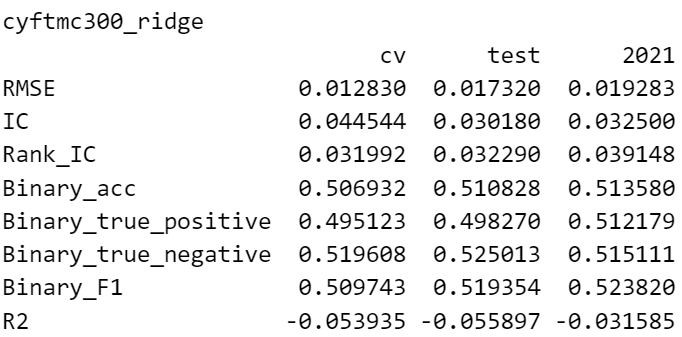
对方差进行检验，发现F值都几乎等于1，故认为方差不变。

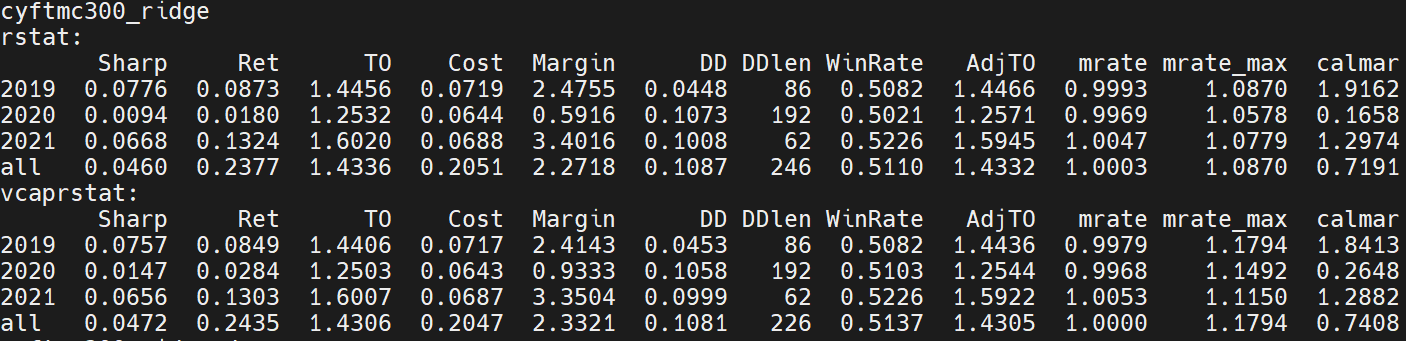
1. **模型表现差异**
2. **300特征**
   1. 岭回归表现差异
      1. lambda = 1
      2. tm40cz





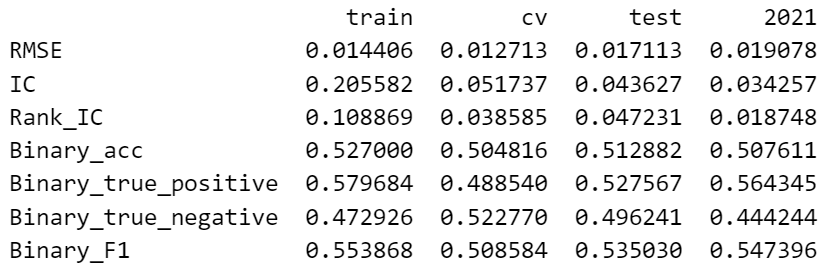
* + 1. tm40c

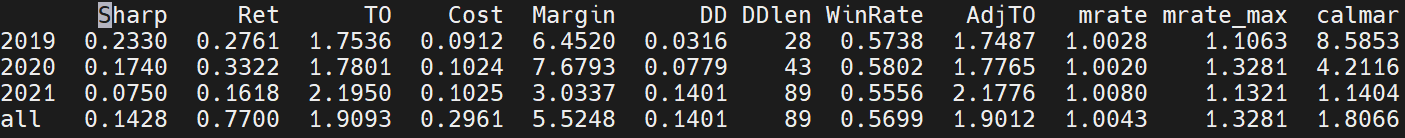


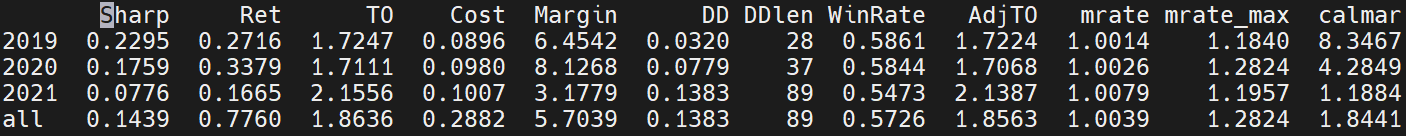


可见，300特征的ridge在tm40c上的表现很差。夏普不如，年化收益也不如。换手率稍低，DD也不如。并且tm40cz偏正，tm40c偏负。

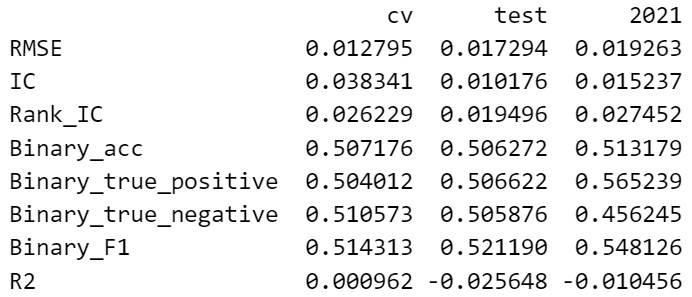
* 1. gbtree表现差异
     1. n\_eatimators = 30, max\_depth = 3, learning\_rate = 0.1
     2. tm40cz

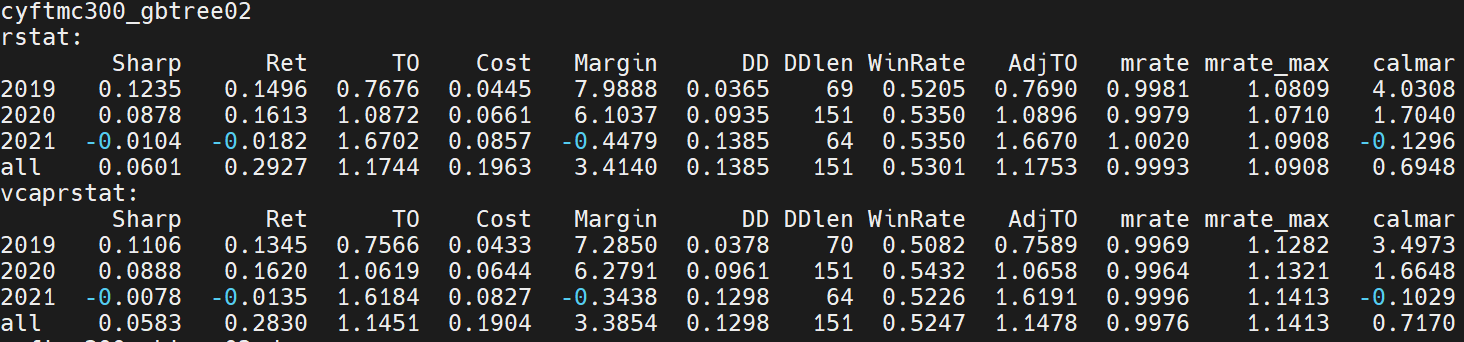






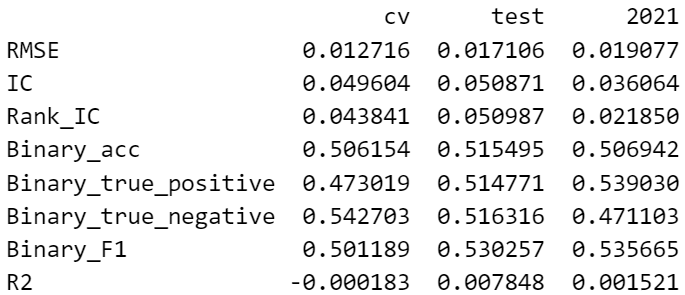
* + 1. tm40c

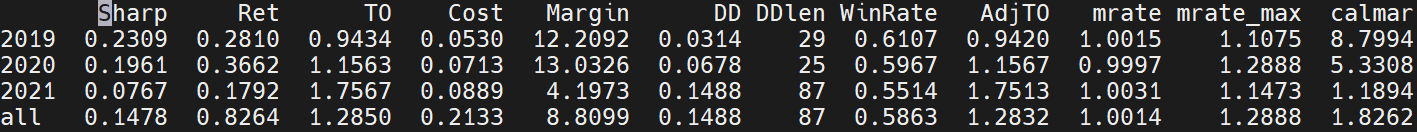


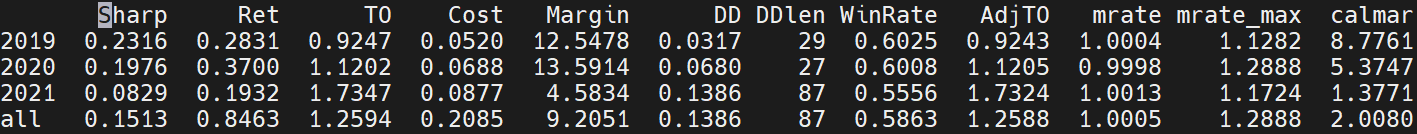


可见，300特征的gbtree在tm40c上的表现非常差。相比于tm40cz，tm40c的夏普严重降低，年化收益严重降低，TO降低，DD几乎不变。并且二者均偏正。

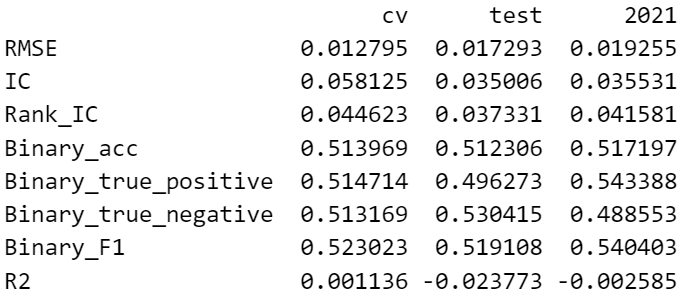
1. **300特征+固定方法筛选因子**
   1. 在tm40cz上筛出118个特征，在tm40c上筛出107个特征
   2. 岭回归表现差异
      1. tm40cz

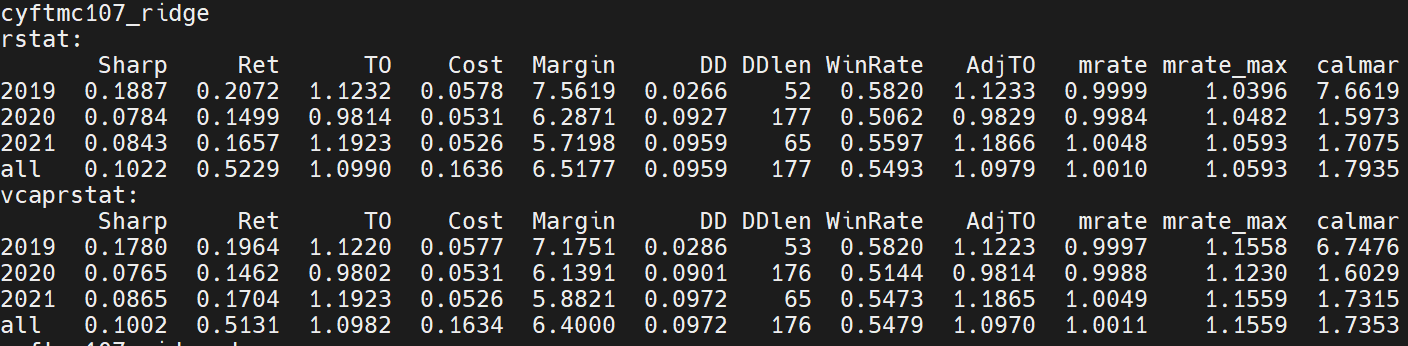






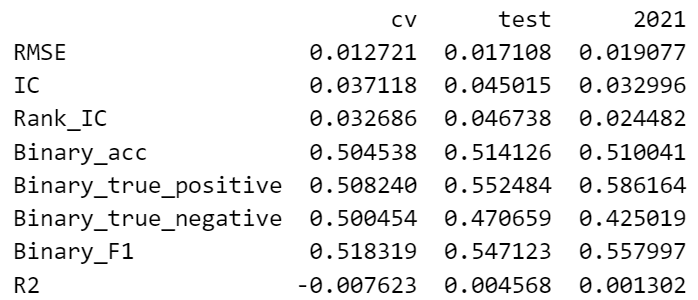
* + 1. tm40c

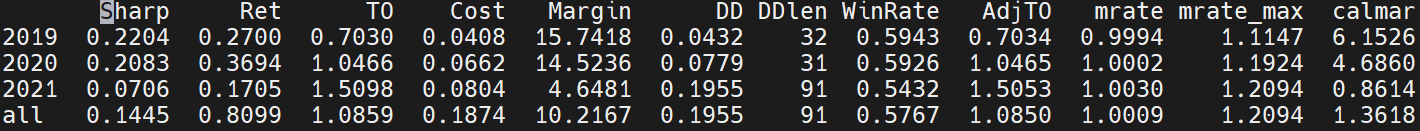


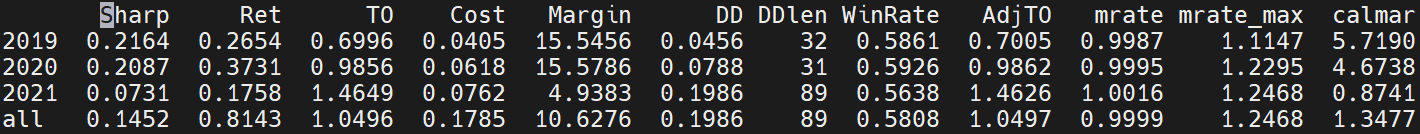


可见，特征筛选的岭回归在tm40c上的表现略弱于tm40cz，夏普降低，年化收益降低，TO降低，DD减小。并且二者都没有偏正或偏负。

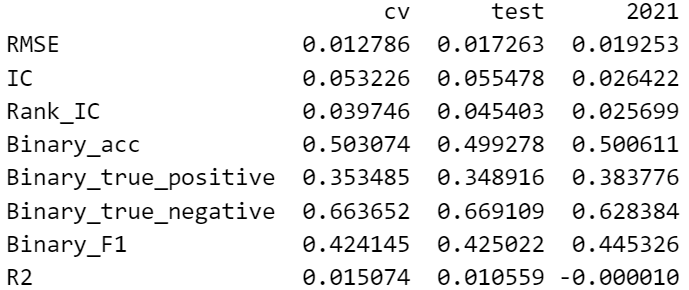
* 1. gbtree表现差异
     1. n\_eatimators = 30, max\_depth = 3, learning\_rate = 0.05
     2. tm40cz

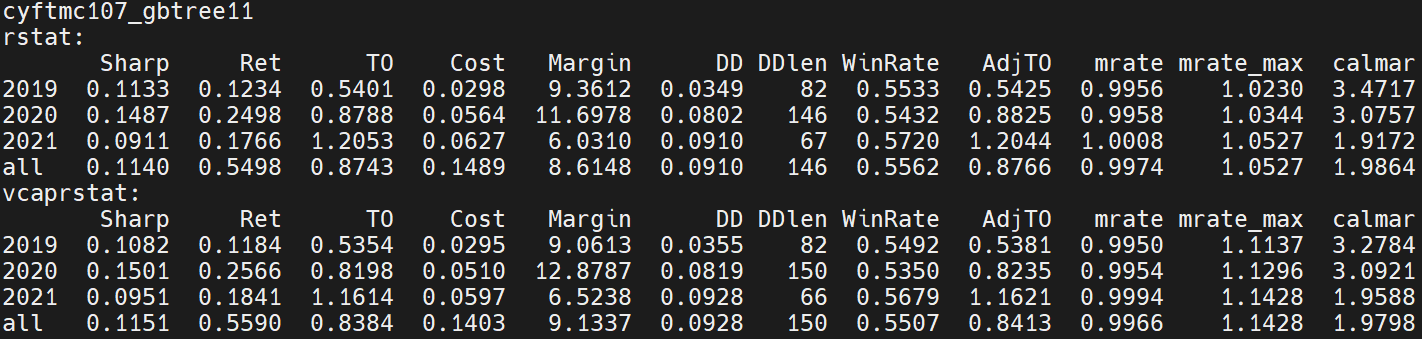






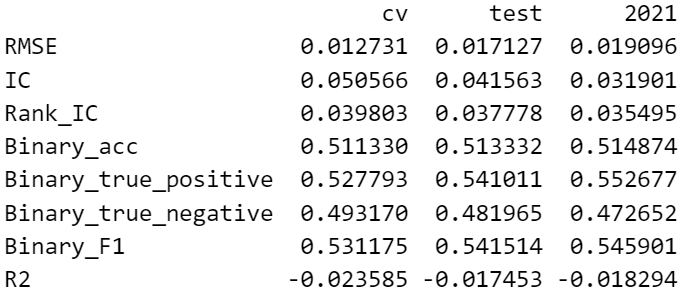
* + 1. tm40c

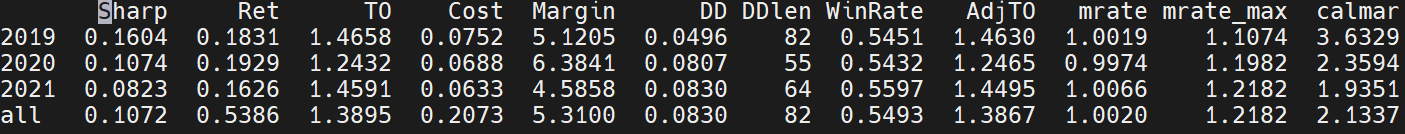


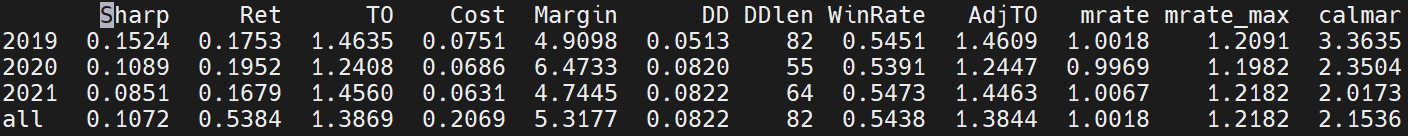


可见，特征筛选的gbtree在tm40c上的表现略弱于tm40cz。夏普降低，年化收益降低，TO降低，DD减小。tm40cz偏正，tm40c偏负。

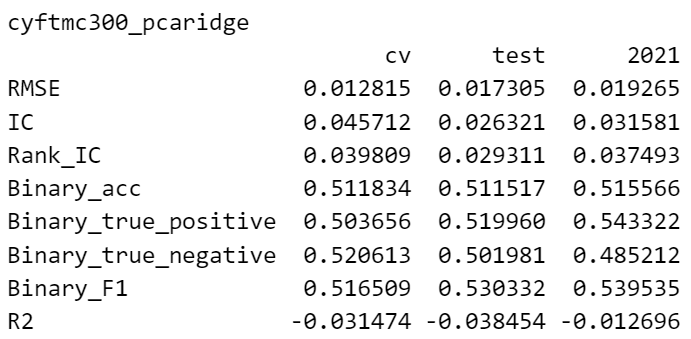
1. **300特征+固定参数PCA**
   1. 在tm40cz上降维为112，在tm40c上降维为86
   2. 岭回归表现差异
      1. tm40cz

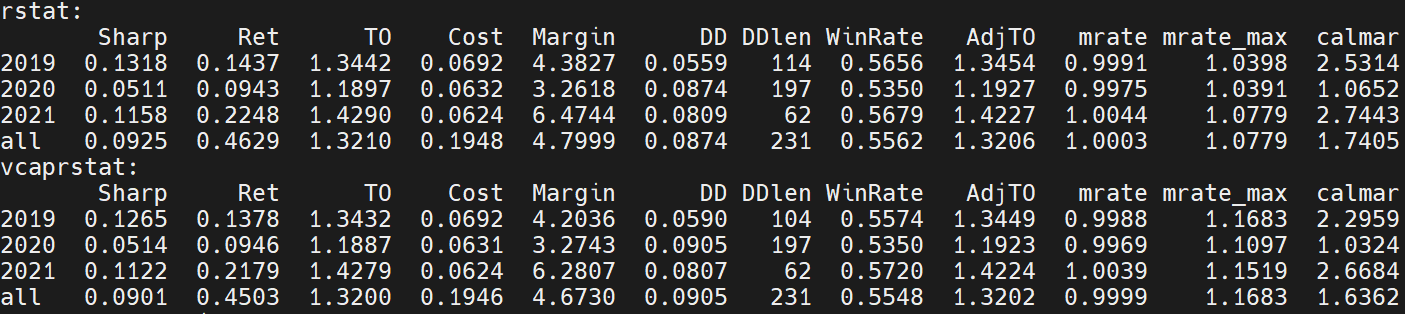






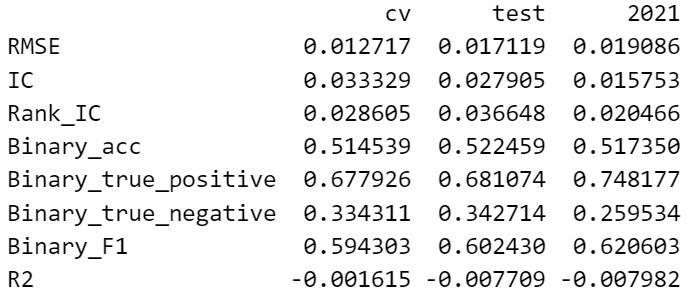
* + 1. tm40c

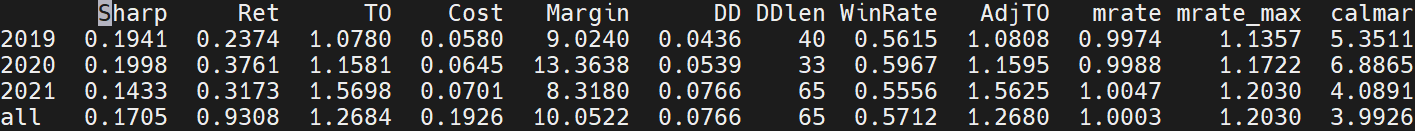


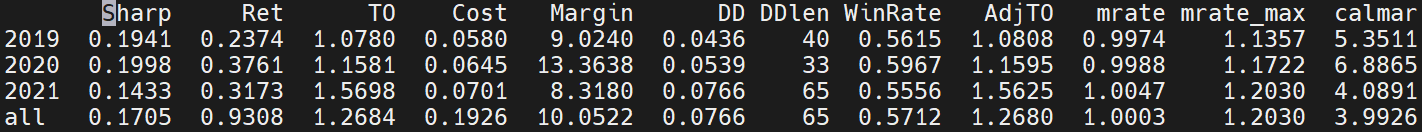


可以发现，pca+ridge在tm40c上的表现略弱于tm40cz，TO略微变小，年化收益减小，DD增大，夏普降低。且在tm40cz上倾向于预测偏正，而在tm40c上没有这种倾向。

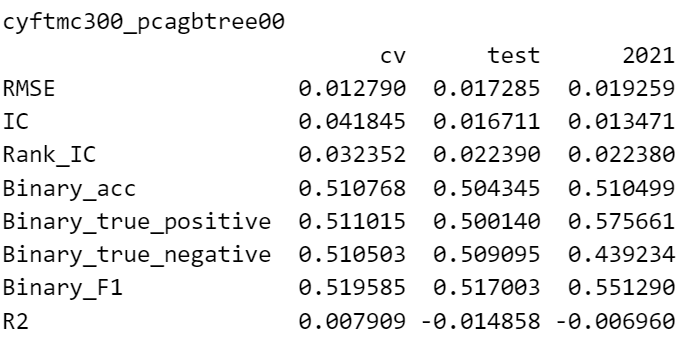
* 1. gbtree表现差异
     1. n\_eatimators = 200, max\_depth = 3, learning\_rate = 0.01
     2. tm40cz

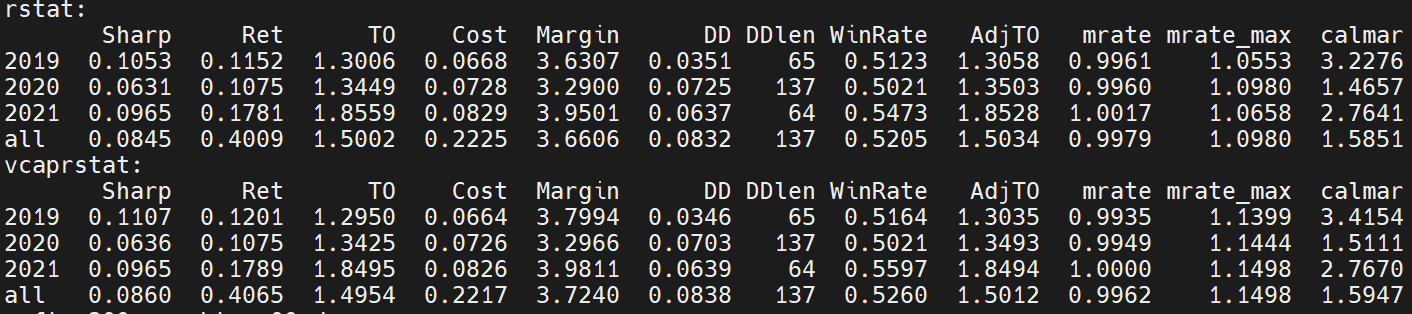






* + 1. tm40c





可以发现，pca+gbtree在tm40c上的表现弱于tm40cz。夏普降低，年化收益砍半，TO升高，回撤增大，margin变低。但在tm40c上的预测结果没有很强的偏正或偏负，而tm40cz是偏正的。

**总结：**

tm40c的模型表现均不如tm40cz，夏普和年化收益均降低。但TO总体上是下降的，DD变化不明显。并且模型的预测值的正负倾向更依赖于参数的设定而非universe的区别。

岭回归在tm40cz上的表现很好，但在tm40c上稍差。但岭回归的表现严重依赖于特征的筛选，因此在新的universe上仍有优化空间与尝试的价值。