**参数选择、数据清洗、模型工程**

韩昊阳

关键词：Sharpe指数；缺失值处理；rolling；特征工程；优化方法。

1. **背景**

承接上周，本周依旧选用21支因子作为自变量，用股票价格生成的的alpha作为因变量。由于模型表现不够突出，为在有限资源情况下生成更好模型结果，本研究选取了标准化（standardization）、缺失值处理、Pearson Correlation和sharpe指数选择因子等方法对输入数据进行处理，并通过IC/IR/Sharpe/Mean Return/MDD/DD period六个特征对选择后的模型进行评估。

1. **实验数据与方法**

实验选用21支因子作为自变量，形状为40000（分钟）\*1829（支）股票，其中有效股票数量为1082支。在数据集中存在一定数量的缺失值（nan）。自变量中8号因子（TSX[100018\_1]）含有nan值2055048个，9号因子含有nan值（TSX[100024\_1]）20851个，构建因子alpha(y)中nan值1269990个，其余因子均无。由于在上周的alpha因子预测中由于没有很好的处理数据和模型，最后各个模型的拟合效果均不够好（表1）。

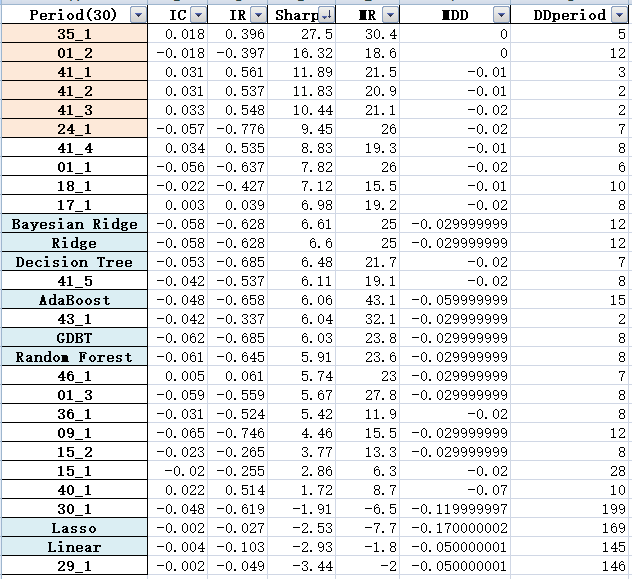


表1：21因子与模型回归alpha因子因子分析

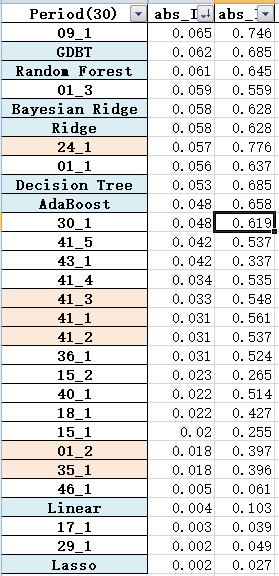
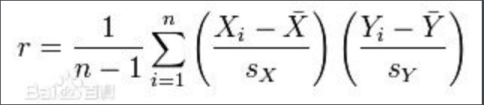


表2：全因子absolute IC/IR因子分析结果

由表1 可见，我们拟合的模型中，linear模型与Lasso模型的表现非常差，Sharpe、absolute IC、absolute IR表现均在因子尾部，而Mean Return value甚至为负值， DDperiod也在所有因子种处在最差位置。与之相对的是树形模型和ridge（L2 regularization）模型的表现相对较好，不仅在sharpe指数的评价中处在所有因子中部，更在日均收益（MR）、IR、IC评价中处在全因子的头部位置。我们针对这种情况，本周分别对整个数据集和linear/lasso/Decision Tree/Random Forest/GBDT等模型进行一定程度的筛选、调整和优化，以提升模型的评价指数。本研究按照实际操作中不同想法的实践顺序排列如下：Pearson Correlation；Standardization；nan-value；Sharpe Index；Rolling；Greedy Research for parameter selection。

1. Pearson correlation.

皮尔森相关系数（Pearson correlation coefficient）也称皮尔森积矩相关系数(Pearson product-moment correlation coefficient) ，是一种线性相关系数。皮尔森相关系数是用来反映两个变量线性相关程度的统计量。相关系数用r表示，其中n为样本量，分别为两个变量的观测值和均值。r描述的是两个变量间线性相关强弱的程度。r的绝对值越大表明相关性越强。其公式如下：

 （1）

在思考模型为何会表现不佳时，首先考虑到的就是自变量不同因子之间可能存在强相关性。倘若模型同时考虑相关性极强的因子，那么整个模型的权重和回归倾向就会趋向于具有强相关性的因子。因此我们首先通过numpy中corrcoef()函数迭代运算得到所有因子之间的相关性，并以0.6作为两两因子是否强相关的凭依。随后使用可视化库seaborn中heatmap()函数生成pearson系数热点图，得到：

3号因子和8、18号因子强相关；

8号因子和3、7、10号因子强相关；

17号因子和16/18号因子强相关。

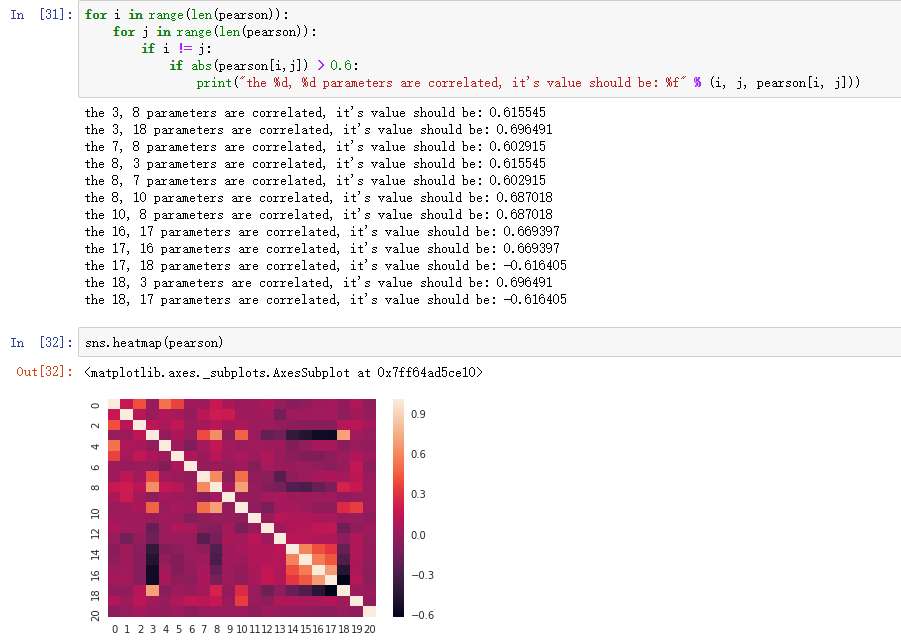


图1.Pearson Correlation Results and Heatmap Visualization.

由此，我们删去第3、8、17号因子，并以新组成的数据作为新数据集继续做数据集优化。

2.Standardization(Feature Scaling via standard normalization)

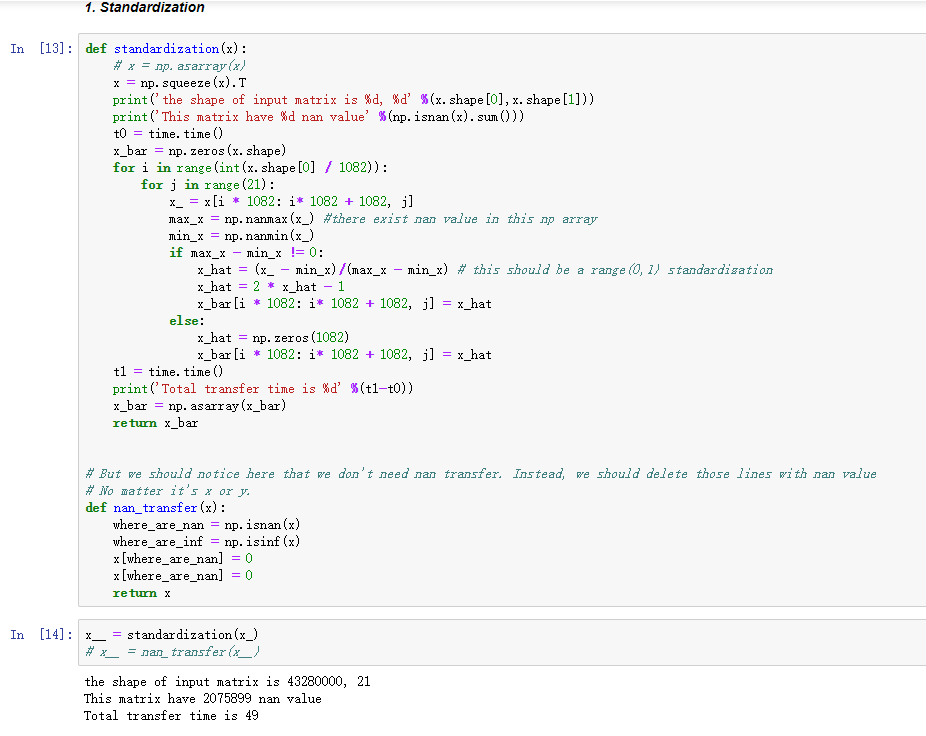
数据标准化指的是通过人工设定最大值和最小值的方法，将不同值域大小的自变量人为整理到同样的值域中（本次试验选择为[-1,1]）。显而易见，倘若某一自变量的最值仅为10^(-1)数量级，其对模型的影响会远逊于最值为10^(9)数量级的自变量因子（而这种情况确实发生在了本数据集中）。由此，我们应当对数据集进行数据标准化。标准化相关的代码如下：

图2：Standardization method used for feature engineering.

标准化的逻辑为：输入（40000\*1082,21）的x，对每个时间点的每一个因子的1802支股票，我们对其进行标准化，标准化的方法为：

X\_standardized = 2\*(x-x\_min)/(x\_max-x\_min) – 1 (2)

通过本方法将所有数据集按照时间不同全部归类到[-1, 1]的值域中。

注意：此时有三种其他标准化方法：对同一个因子所有时间点所有数据找x\_min/x\_max，进行标准化；对同一个因子的连续N个时间点（eg. N=300）进行标准化；通过排序而不是标准化的方法处理数据。不同的方法处理结果不同，但也要考虑到由于未来的时间点的数据值大小信息是缺失的，因而在使用过去的标准化值估计未来标准化值时会遇到二次估计从而加大误差的问题。

3. nan-value processing

在上周对数据集进行预处理时，有非常大的错误：我们机械的将所有nan值通过numpy的nan\_to\_num()函数全部转换为了0值，但事后思考了一下这是非常非常错误的。首先，y值缺失一般是遇到了股市停牌等实际情况，如果直接将其变为0不仅不符合实际，也会对后续数据处理产生巨大影响。而y值的缺失值较多，为1269990（约2.93% ），如果全部机械的变为0对模型的影响应当是巨大的。我们将y值为nan的整个时间点抛弃（抛弃21自变量+1因变量个值）而对于x值的取舍，考虑到自变量仅8，9号存在nan值且nan值数量较多，我们将这两个自变量直接抛弃。­­­­­

4. Sharpe Index

在进行完上示的操作后，我们对linear模型进行了一些拟合，发现结果仍不理想。拟合结果如下所示（图3）：



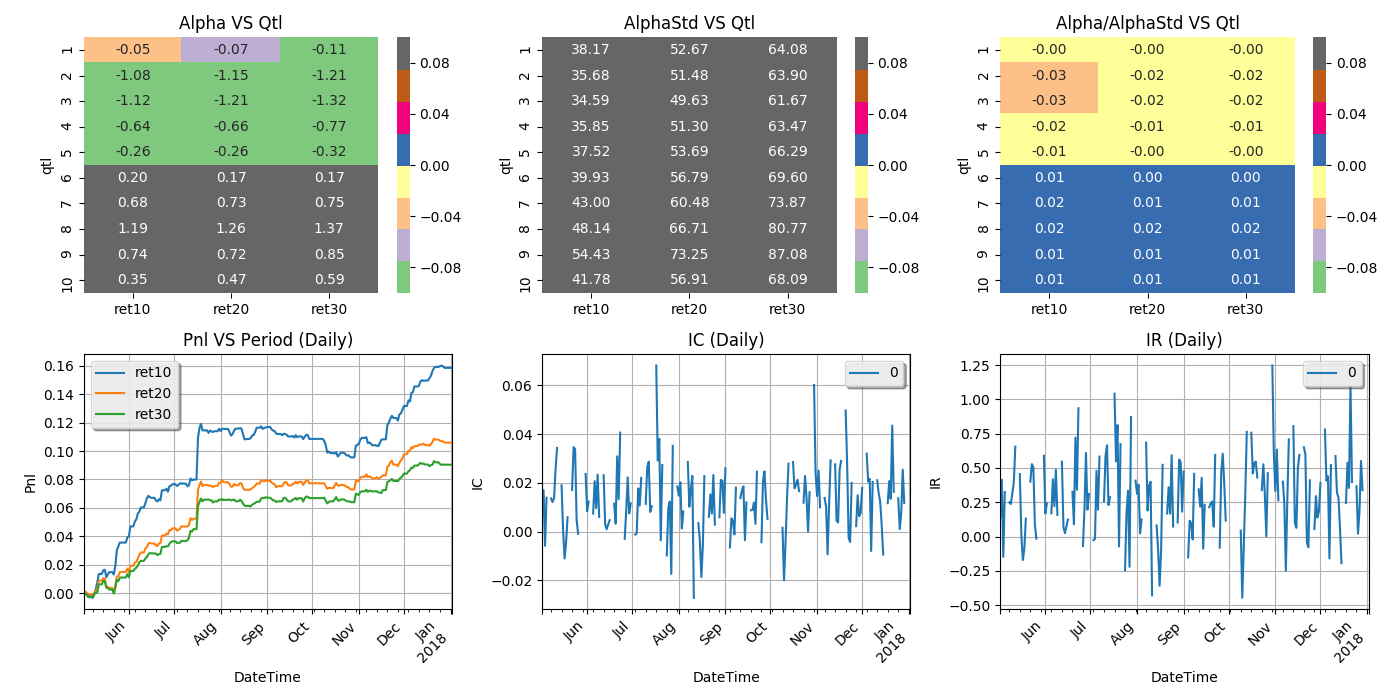


图3.全因子线性回归alpha因子检测结果

此时linear模型的sharpe指数仅有4.2 ，其他各个数据也都不理想。经过反思，我们认为可以选择强因子对模型进行拟合。考虑到模型的最终目标是回归的alpha指数的因子检测中，各个指标（IC/IR/Sharpe/MR/MDD/DDperiod）应强于每一个单独单因子，因此我们选择了sharpe指数最强的6个因子，并将其按照之前的所有顺序重新清理一遍。结果发现，24-1因子虽然sharpe指数高，但有nan值，因而同样需要抛弃。加入新选择的因子，此时所有因子分别为: 'TSX[100001\_2]', 'TSX[100035\_1]', 'TSX[100041\_4]', 'TSX[100041\_1]','TSX[100041\_2]','TSX[100041\_3]'。

5.Rolling method

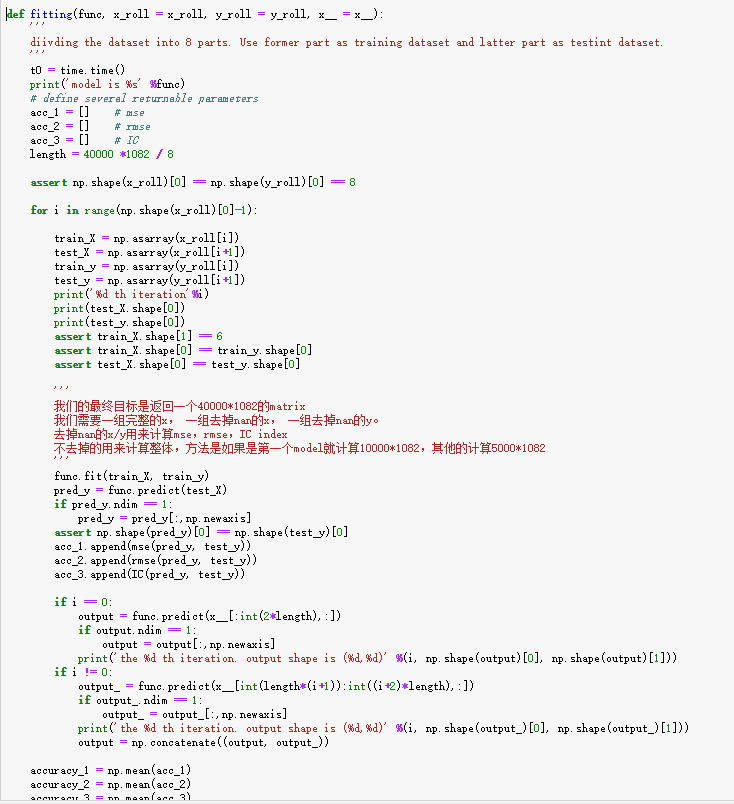
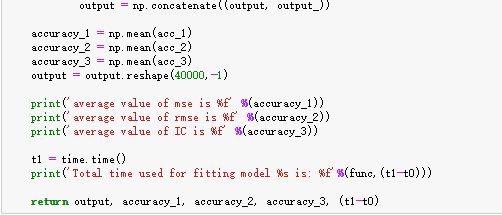
在上周的实验中我们用35000-T个时间点进行模型预测，并用5000个时间点的数据做模型检验，并计算mse/rmse/IC指数。但是我们在因子测算前，会把4000\*1829支股票完整的输入到模型中。此时前35000个时间点的预测值会非常非常接近真实值，从而使得模型预测变得毫无意义。为此，我们使用滚动的方法来进行动态的模型预测，即用1-5000时间点数据预测5001-10000，再用5001-10000预测10001-15000……这样我们会得到一个training:testing = 5000:35000的数据结果，极大的提升了模型预测的可靠性。其具体代码如下:  

图4. Rolling method to generating model-fitting result.

通过循环的方法，对于给定的函数(func)，不断产生train\_test dataset 并通过已构建好评判函数生成评判结果，同时返回40000\*1028的拟合值，mse/rmse/IC评估标准的大小和总运行时间。通过这样的方法，各个模型的性能被极大的提升了。

6. Greedy Research for parameter selection

这是一种通过循环同一列表不同参数，放入模型中进行拟合，从而确认模型好坏的方法，在实际运行过程中使用的时间非常长，效率较低而未对所有模型使用。

1. **实验步骤**
2. 导入数据集-确定导入因子
3. 规范化
4. Pearson检验
5. Rolling方法分割数据，训练并测试模型
6. 生成因子检验数据
7. 导出数据
8. **结果**

直接讨论所有数据处理后的结果。此时我们对linear/lasso/Decision Tree/Random Forest/GBDT模型输出数据并得到其因子检测结果。需要比较的结果为：最强6因子因子检测结果，非rolling模型结果，rolling下模型结果。

1. 总比较图

下图列出依照sharpe指数倒叙排列的模型选择结果，未标注\_R的模型是非rolling模型的结果，下标为R的则是rolling模型的拟合alpha结果。

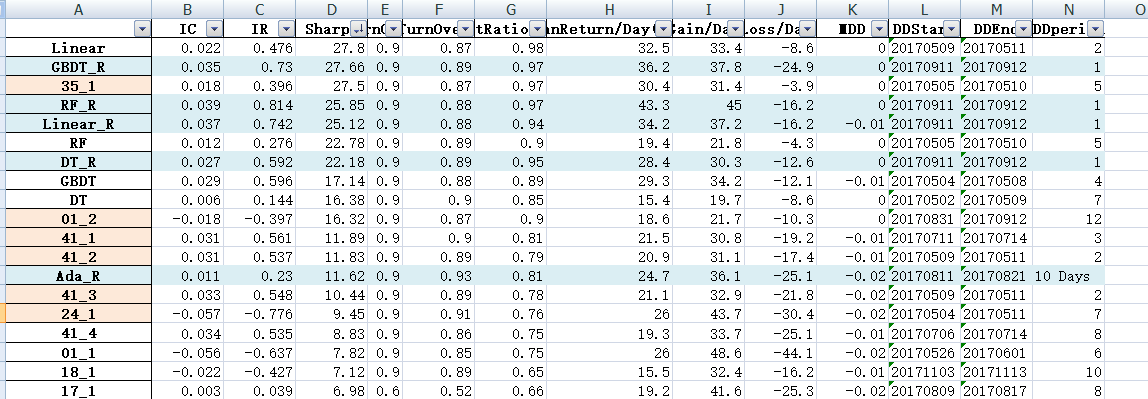


表3.全因子全模型比较结果图

1. 最强因子检测结果

我们列出sharpe指数最高的三个因子的因子检测结果，并以这些结果为基础分析其他模型的表现情况：

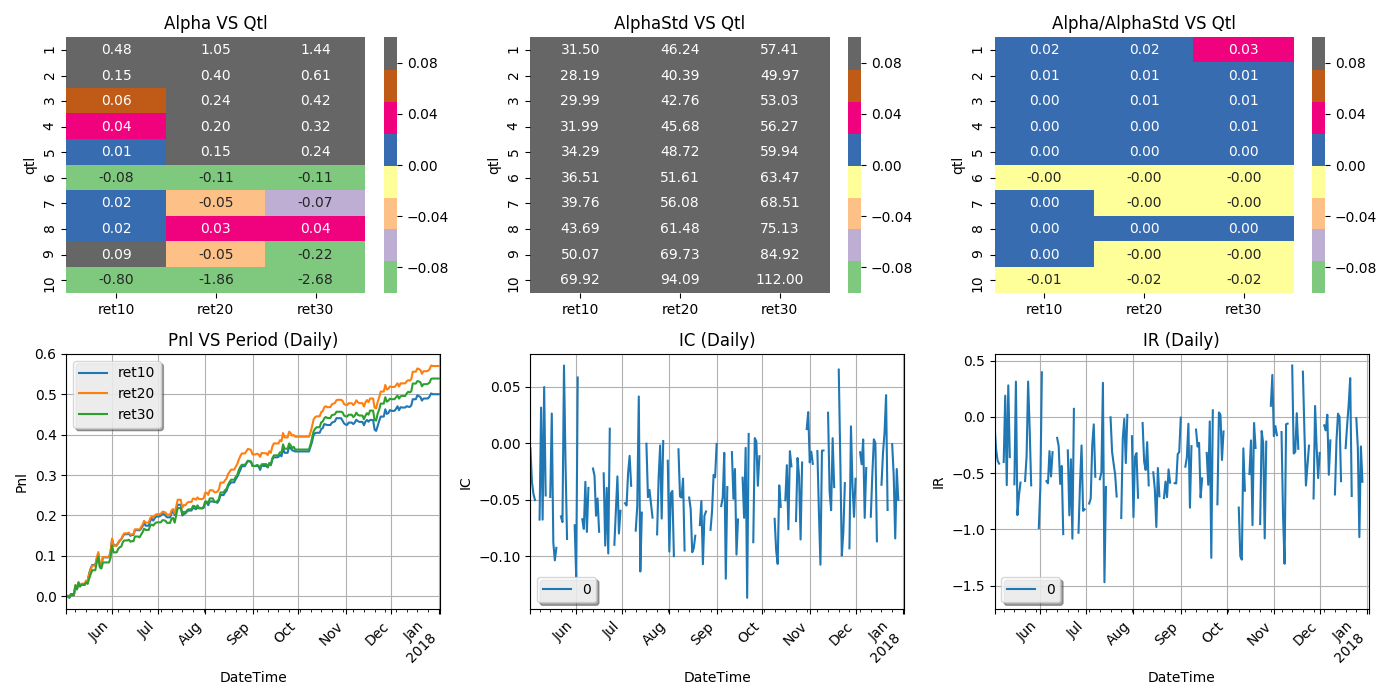


图5：Factor 43-1 Analysis result

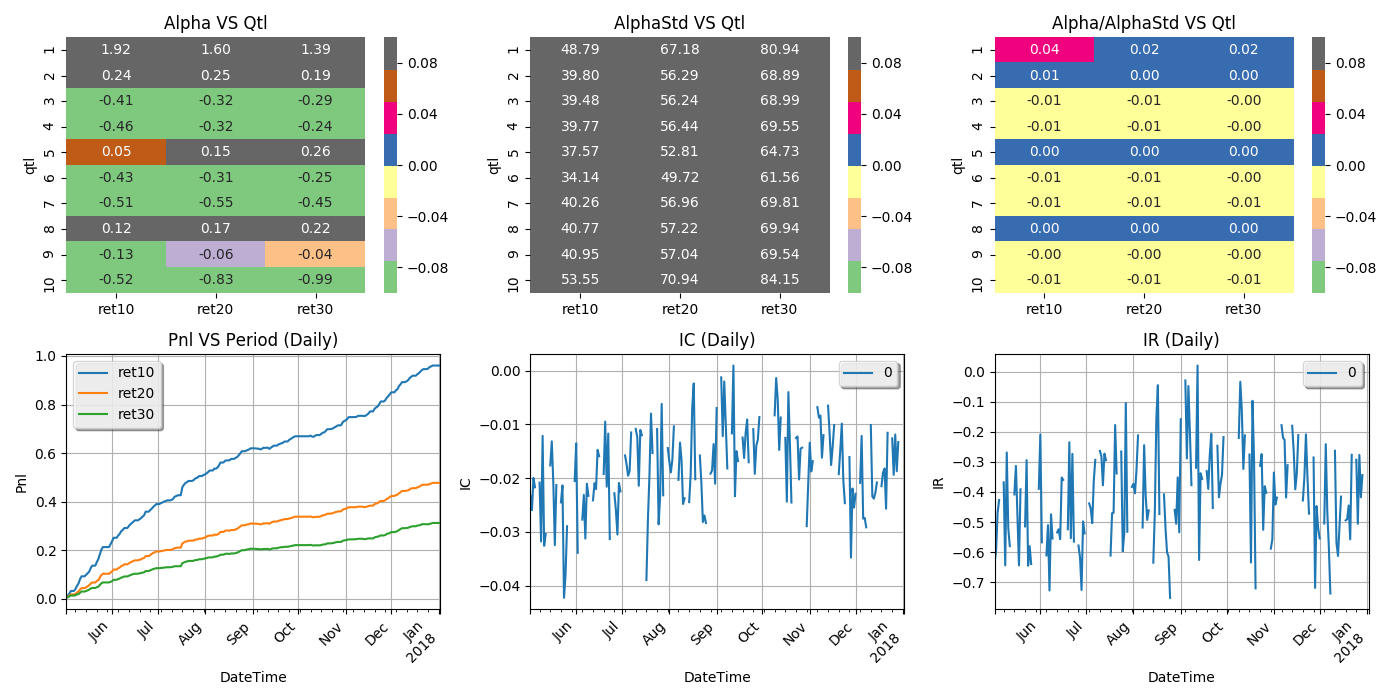


图6：Factor 02-1 Analysis result

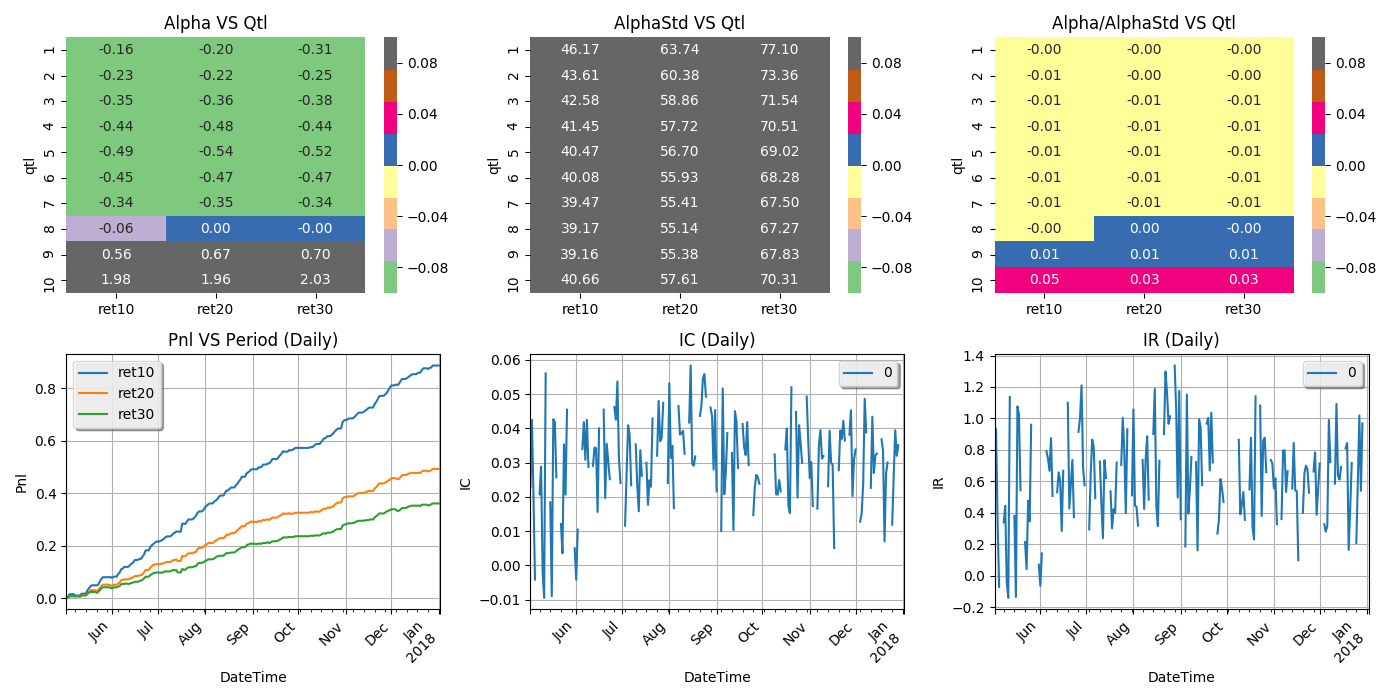


图7：Factor 41-1 Analysis result

1. 模型拟合结果：

选取linear，GBDT\_Rolling，

1. **讨论**
2. **附录**