在 PTA 期货合约 1min 周期上的因子研究和组合策略

余康齐

2023 年 4 月 14 日

摘要

在这篇报告中，我简述了从特征工程开始，到模型训练，最后回测检验模型的 PTA 期货合约的研究全流程。特征工程我分为了两部分，先是依照个人经验挖掘出的 5 个基础因子，然后是利用遗传规划来挖掘简单的基础因子组合，并验证其有效性。模型训练部分我使用了线性模型，随机森林和神经网络三种模型，最后在日内回测中验证了模型的有效性，并体现了随机森林模型在此类问题的优越性。

关键词：遗传规划，因子组合，随机森林，神经网络，日内回测

目录

1 特征工程

1.1 基础因子

首先，我从 return，open interest，turnover，volume，bid1，ask1 等数据字段入手，构建了 5 个基础因子。因子 IC 值计算公式如下：

$$IC = correlation(\alpha, return)$$

值得注意的是，在因子表达式中的所有字段均指前一天的数据，因为我认为在 1min 周期上，因子的表达式应该是基于前一天的数据，而不是基于当天的数据。在构建因子和模型训练中我使用的数据是 2022-09-02 及之前的数据，后面五天的数据用于回测。

1.2 alpha1

第一个因子构建思路是利用了 return 的反转，我做出 1min 的 return 时序图，发现几乎每分钟 return 都会发生或大或小的反转，于是就有了如下因子：

$$alpha1 = -return$$

经计算在所使用的 10 天内，IC 值为 0.107，十分显著。以每天为频率计算一次 IC，10 天内 IC 值的变化如图所示，可以看出 IC 值在 10 天内波动不大，较为稳定。

1.3 alpha2

第二个因子构建思路是利用了 open interest 的变化。open interest 降低是因为有人平仓，而 open interest 增加是因为有人开仓，我认为其变化可以反映投资者们对走势的判断，于是就有了如下因子：

$$alpha2 = -delta(openinterest, 1)$$

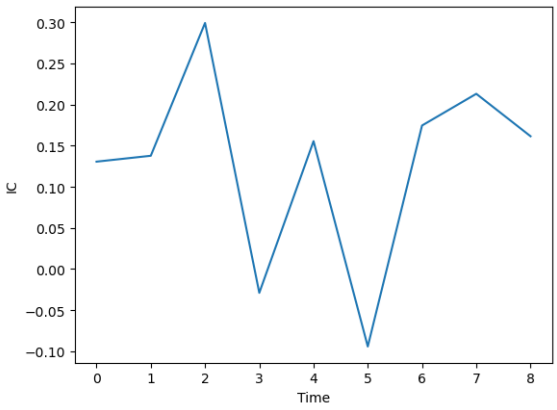

图 1: alpha1 IC 值随时间变化图

经计算在所使用的 10 天内，IC 值为 0.112，十分显著。以每天为频率计算

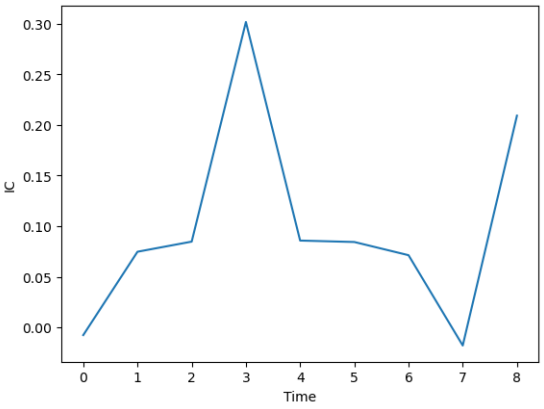

图 2: alpha2 IC 值随时间变化图

一次 IC，10 天内 IC 值的变化如图所示，可以看出 IC 值在 10 天内波动不大，且大多在 0 以上。

1.4 alpha3

第三个因子构建思路借鉴了股票研究中的 VWAP 思想，在此处，我假设 1min 内的 VWAP 是此资产在上一分钟的正确定价，我用 VWAP 和 last

price 比值作为可盈利的空间，于是就有了如下因子：

$$VWAP = \frac{turnover}{volume}$$

$$alpha3 = lastprice/VWAP$$

经计算在所使用的 10 天内，IC 值为 0.089，较为显著。以每天为频率计算

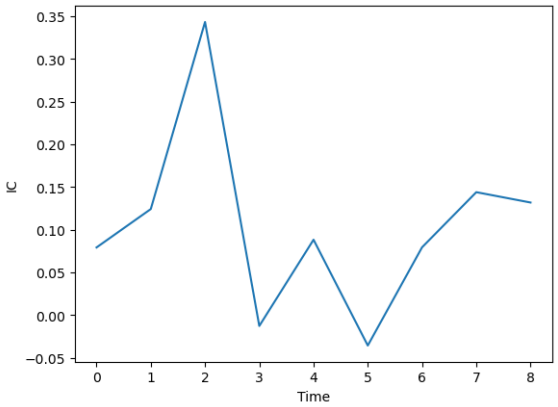

图 3: alpha3 IC 值随时间变化图

一次 IC，10 天内 IC 值的变化如图所示，可以看出 IC 值在 10 天内波动不大，且都在-0.05 以上。

1.5 alpha4

第四个因子构建思路是利用了在一分钟内成交价最高价，最低价和最终价的关系。因子表达式如下：

$$alpha4 = \frac{high - lastprice}{lastprice - low}$$

经计算在所使用的 10 天内，IC 值为 0.067，较为显著。以每天为频率计算一次 IC，10 天内 IC 值的变化如图所示，可以看出 IC 值在 10 天内波动不大，较为稳定。

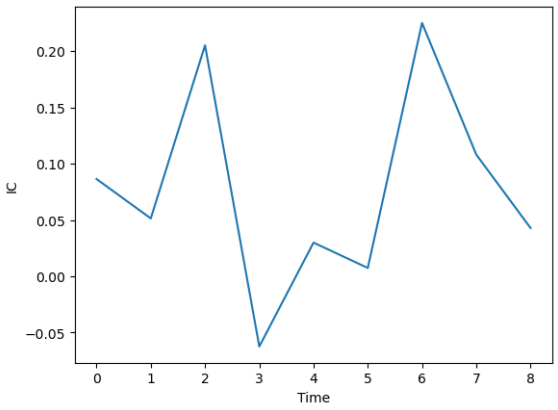

图 4: alpha4 IC 值随时间变化图

1.6 alpha5

第五个因子构建思路是假设 Bid1 和 Ask1 是理性的，那么如果 Ask1 和 Askvol1 乘积大于 Bid1 和 Bidvol1 乘积，那么说明市场方向为向上，反之向下，于是就有了如下因子：

$$alpha5 = \frac{Ask1 * Askvol1}{Bid1 * Bidvol1}$$

经计算在所使用的 10 天内，IC 值为 0.099，较为显著。以每天为频率计算

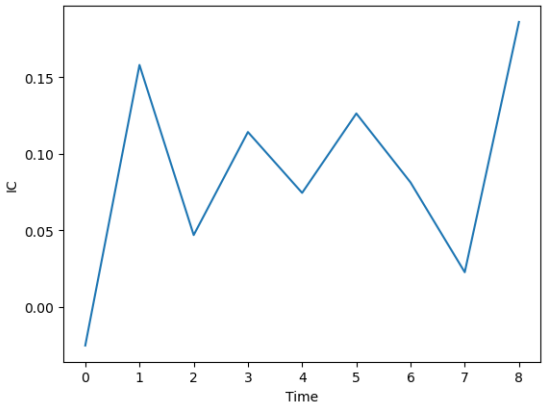

图 5: alpha5 IC 值随时间变化图

一次 IC，10 天内 IC 值的变化如图所示，可以看出 IC 值在 10 天十分稳定。

1.7 因子相关性

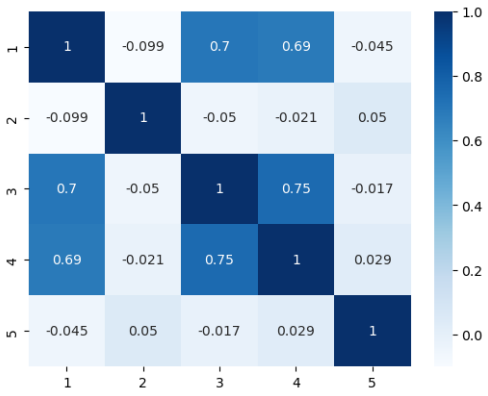

图 6: 因子相关性图

因子相关性图如图六所示，alpha2 和 alpha5 与其他三个因子相关性低，是在 5 个因子里相对好的因子。alpha1，alpha3 和 alpha4 相关性高，可能是内在思路和数据源相似。

1.8 遗传规划

遗传规划是一种基于符号树的机器学习算法，它可以自动地从数据中拟合出一个表达式，这个表达式可以用来预测未来的数据。
为避免过拟合，我选取的树最大深度为 2。在使用基础算子训练前五个 alpha 后，我得到了以下新的特征：

$$alpha6 = alpha2^2$$

测试集 IC 值为 0.133，训练集 IC 达到 0.249。

$$alpha7 = alpha1 \times alpha2 \times abs(alpha4)$$

测试集 IC 值为 0.101，训练集 IC 达到 0.106。

$$alpha8 = min(min(alpha3, alpha4), 0.174 \times alpha1)$$

测试集 IC 值为 0.073，训练集 IC 达到 0.170。

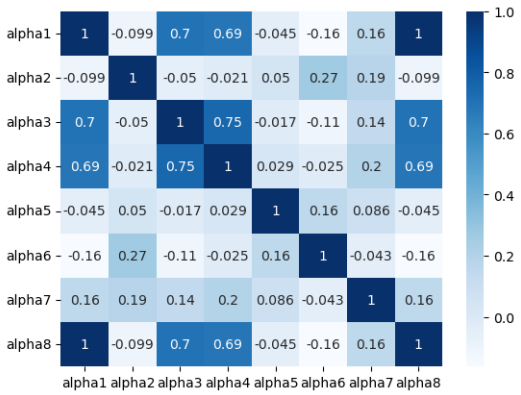

图 7: 因子相关性图

可以看到，新生成的特征 alpha6 和 alpha7 与其他特征相关性较低，是在 5 个因子里相对好的因子。alpha8 与 alpha1 完全一致，剔除。

2 模型训练

2.1 线性模型

Dep. Variable:	return	**R-squared:**	0.059
Model:	OLS	**Adj. R-squared:**	0.054
Method:	Least Squares	**F-statistic:**	14.49
Date:	Thu, 13 Apr 2023	**Prob (F-statistic):**	2.13e-18
Time:	21:07:18	**Log-Likelihood:**	-2277.6
No. Observations:	1640	**AIC:**	4571.
Df Residuals:	1632	**BIC:**	4614.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	4.982e-15	0.024	2.07e-13	1.000	-0.047	0.047
alpha1	0.1395	0.037	3.813	0.000	0.068	0.211
alpha2	0.0632	0.026	2.472	0.014	0.013	0.113
alpha3	0.0241	0.040	0.608	0.543	-0.054	0.102
alpha4	-0.0600	0.040	-1.520	0.129	-0.138	0.017
alpha5	0.1028	0.024	4.198	0.000	0.055	0.151
alpha6	0.1459	0.026	5.665	0.000	0.095	0.196
alpha7	0.0663	0.025	2.626	0.009	0.017	0.116

Omnibus:	101.690	**Durbin-Watson:**	2.003
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	423.705
Skew:	-0.076	**Prob(JB):**	9.85e-93
Kurtosis:	5.485	**Cond. No.**	3.23

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

将全部数据放入回归得到的拟合 R squared 为 0.059，说明模型拟合

程度不高。但 alpha1，alpha2，alpha5，alpha6 和 alpha7 极为显著。我还使用了 10 组交叉验证来评价模型，得到的平均 R squared 为 0.021。由于 alpha1，alpha3 和 alpha4 相关性高，可能会出现多重共线性的问题，所以我单独使用 alpha1，alpha3 和 alpha4 来训练模型，并交叉验证得到的最优的平均 R squared 为 0.023，相比全部使用稍有提升。

此外，我还使用了岭回归，发现表现稍优于多元线性回归。最后我选选取了线性模型中最好的模型：对 alpha1，alpha2，alpha5，alpha6 和 alpha7 用岭回归，在测试集中 R squared 为 0.064。其在测试集上的拟合价格趋势图和原价格趋势（起始价设为 1）的拟合效果如下：

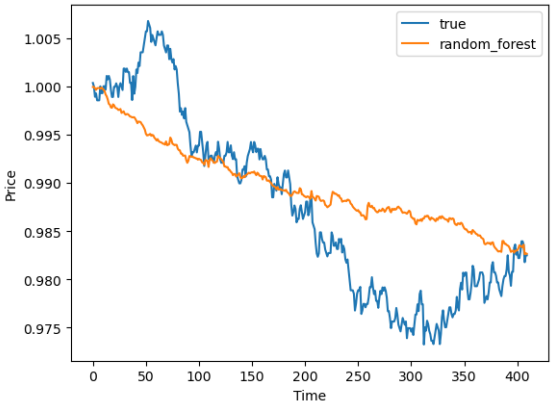

图 8: 岭回归拟合效果

2.2 非线性模型

2.3 随机森林

在这节我使用了随机森林模型，经调参后，最优模型的训练集 R squared 为 0.05，测试集 R squared 为 0.04。随机森林所依照的树结构在文件夹 tree 中。

下面是在测试集上的拟合价格趋势图和原价格趋势（起始价设为 1）：

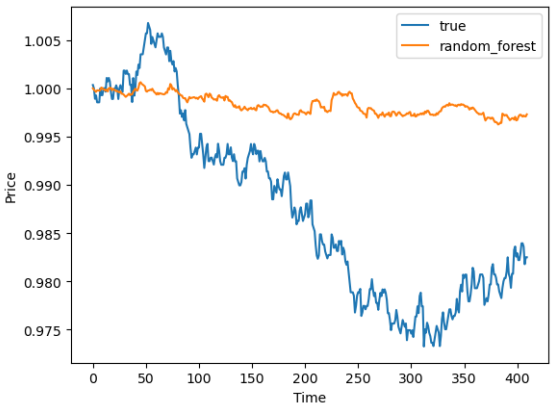

图 9: 随机森林拟合效果

2.4 神经网络

在这节我使用两层隐藏层的神经网络模型，经调参后，最优模型的训练集 R squared 为 0.01，测试集 R squared 为 0.04。

下面是在测试集上的拟合价格趋势图和原价格趋势（起始价设为 1）：

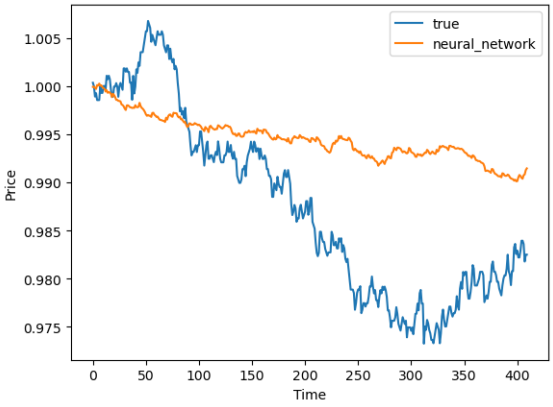

图 10: 神经网络拟合效果

2.5 模型总结

加上岭回归一共使用了三种模型，三种模型在预测集的结果都比真实数据平缓，其中岭回归在测试集表现的最好。

3 日内交易回测

3.1 回测数据

回测数据使用的是最后五天的交易数据，为的是不使用前 10 天用于训练模型的数据，模拟真实的交易情形。

3.2 回测架构

回测类主要架构是从 dataset 中逐个获取 tick 级数据，每分钟结束时获取当前分钟的 alpha 值，然后根据 alpha 值来预测下一分钟的收益，最后根据其来执行买卖策略。成交逻辑是下单后在下一 tick 撮合成交，取消订单立刻执行。如下单的 volume 违反了 position limit，则会拒绝下单。每天最后一到两分钟会强制平仓。

3.3 回测策略

在前文中我们提到，三个模型预测的收益率均比真实收益率平缓，所以我们不能直接使用预测收益率建立模型。我选择以前 10 天预测收益率的分布为参照，再判断预测出的收益率是否值得进行交易。由于此次回测极为理想化，没有考虑信息传输时间，冲击成本等因素，并且可以不限回合数，所以我选择了数据的均值作为参考，当预测收益率大于此均值做多，反之做空。但在现实交易市场中，我会考虑到这些因素，所以我会在均值上下各加一个标准差，即在均值上下各一个标准差（预测收益率的分布大致符合正态分布）的预测收益率我会选择不做出交易决定，来减少交易次数。此

外，当预测值十分显著（大于均值两个标准差或小于均值两个标准差），且 position 满足条件，我会下 volume 为 2 的订单，来获取更多更稳定的利润。

3.4 回测结果

下三图为三个模型在 5 天的回测结果：

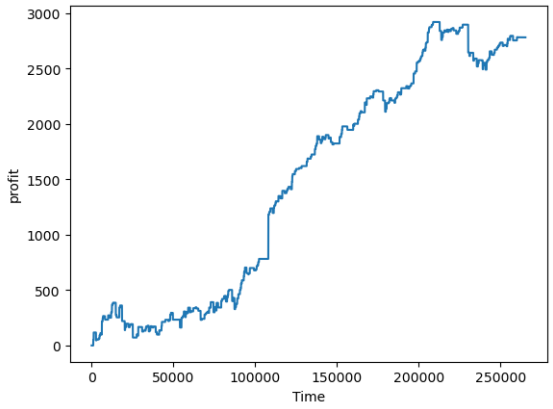

图 11: 岭回归回测结果

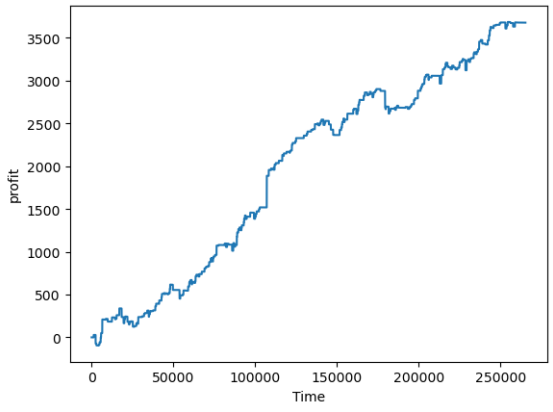

图 12: 随机森林回测结果

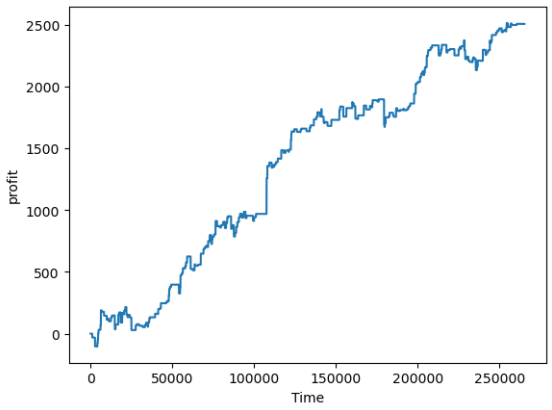

图 13: 神经网络回测结果

下表为三个模型在 5 天的回测指标统计：值得注意的是，我设置本金为交易开始时的第一个 tick 的 last price 的五倍（因为一手五吨），本金在计算以下指标时有使用。sharp ratio 过低是因为我采用的是分钟级的 return 计算，数据量过大。

表 1: backtest result

	annually return	sharp ratio	profit withdraw ratio	win rate
ridge regression	942.003370	0.013273	0.667383	0.728571
random forest regression	7568.973185	0.018182	1.317724	0.749020
neural network	493.579812	0.013301	0.870338	0.695833

从指标和回测收益图来看，随机森林模型的表现最好，主要体现在收益回测比高，收益率高，胜率高。岭回归与随机森林相比差在回测过大，神经网络与随机森林相比收益率太低。