九坤 Kaggle 量化大赛有哪些启示？

华泰研究
2023 年 1 月 30 日 | 中国内地　　深度研究

研究员　林晓明
SAC No. S0570516010001　linxiaoming@htsc.com
SFC No. BPY421　+(86) 755 8208 0134

研究员　李子钰
SAC No. S0570519110003　liziyu@htsc.com
SFC No. BRV743　+(86) 755 2398 7436

研究员　何康，PhD
SAC No. S0570520080004　hekang@htsc.com
SFC No. BRB318　+(86) 21 2897 2039

人工智能系列之 64：从九坤 Kaggle 量化大赛高分方案中寻找借鉴

本文梳理 2022 年九坤 Kaggle 量化大赛高分队伍解决方案，提炼出特征工程、损失函数、交叉验证、模型集成四个主要方向，并应用于华泰人工智能中证 500 指数增强策略改进。结果表明：(1)特征工程引入均值因子对神经网络有效；(2)CCC 损失优于 MSE 损失和 IC 损失；(3)时序交叉验证作用不明显；(4) 集成神经网络和决策树类模型提升较稳定。对比整合多项改进的模型与基线模型，回测期 2011 年至 2022 年内，年化超额收益从 14.2%提升至 17.0%，信息比率从 2.3/2.4 提升至 2.7。

多家头部量化机构在 Kaggle 发布竞赛，九坤竞赛贴近实际量化选股场景

随着数据科学在线社区日益成熟，越来越多的爱好者投身于网络编程竞赛之中。Kaggle 是全球知名的数据科学在线平台之一，Two Sigma、Optiver 等头部量化机构曾在 Kaggle 发布挑战竞赛。国内量化私募九坤投资于 2022 年 1 月启动 Kaggle 竞赛，吸引两千多只队伍参赛。比赛具体任务为基于给定的 A 股匿名特征，预测股票未来短期收益，最终评价指标为预测收益和真实收益的 IC 值，属于典型的监督学习问题，和实际量化选股场景较贴近。

四个改进方向：引入均值因子，引入 CCC 损失，时序交叉验证，模型集成

我们梳理九坤 Kaggle 量化大赛高分队伍解决方案，提炼出四个改进方向。(1)特征工程引入截面上全部股票因子的均值，均值因子可能反映原始因子整体分布的时变特性，是市场环境的一种简单表达。(2)损失函数引入一致性相关系数 CCC，可视作 IC 和 MSE 的融合，兼顾相关性和距离。(3)采用时序交叉验证选取最优超参数。(4)集成不同类型机器学习模型。以神经网络和 XGBoost 构建中证 500 指数增强策略作为基线，测试上述技巧的改进效果。

均值因子对神经网络有效，加权 CCC 损失回测表现好，模型集成提升稳定

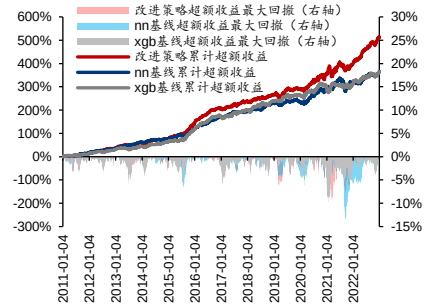
四项改进技巧效果各异。特征工程引入的均值因子对神经网络有提升，但削弱了 XGBoost 模型。损失函数中，MSE 表现不突出；IC 损失单因子测试表现好，但指增组合回测表现差；CCC 损失在单因子测试表现一般，但指增组合回测表现较好；加权均优于等权。交叉验证调参改进不显著，考虑到时间开销大，性价比不高，算力有限前提下，使用经验超参数即可。模型集成提升较稳定，神经网络类和决策树类模型有互补效果。

讨论：(1)如何使用弱因子；(2)因子合成和组合优化的目标错配问题

研究发现均值因子对神经网络有效但对 XGBoost 无效。均值因子属于弱因子，有用但比重不宜过大。XGBoost 引入弱因子后，特征采样使原始因子可能被排除在外，从而削弱模型。神经网络可通过预处理缩小取值，有限度地使用弱因子。研究还发现 IC 损失单因子测试优于 MSE 损失，但指增组合表现差，本质是因子合成和组合优化的目标错配。IC 属于全局统计量，不会侧重于个别头部样本，但这些样本可能对组合优化影响较大。MSE 的特点之一是给予极端误差较大惩罚，恰好弥补 IC 弱点。CCC 融合 IC 和 MSE，兼顾共性和个性，是一类理想的损失函数。

风险提示：人工智能挖掘市场规律是对历史的总结，市场规律在未来可能失效。人工智能技术存在过拟合风险。深度学习模型受随机数影响较大，本文未进行随机数敏感性测试。本文测试的选股模型调仓频率较高，假定以 vwap 价格成交，忽略其他交易层面因素影响。

基于九坤大赛改进策略超额收益表现



注：回测期 2011-01-04 至 2022-12-30，基准为中证 500
资料来源：朝阳永续，Wind，华泰研究

正文目录

研究导读

得益于数据科学在线社区日益成熟，机器学习和大数据的学习门槛逐渐降低，全球的爱好者都可以通过在线平台参与编程训练和竞赛项目，和顶尖团队进行较量和探讨。Kaggle 正是影响力较大的平台之一，囊括了超过 500 项竞赛、5 万个数据库和 40 万组代码。美国白宫、斯坦福大学、北京大学、微软、谷歌等机构和企业都曾在 Kaggle 发布竞赛，征集解决方案。

量化投资和机器学习、大数据关系紧密，多家量化投资机构也在 Kaggle 平台发起挑战竞赛，发布方不乏 Winton、Two Sigma 等知名对冲基金，也包含 Jane Street、Optiver 等头部做市商。项目内容大多是基于资产历史行情、新闻数据或匿名特征，预测未来收益率或波动率。下表整理了 Kaggle 平台量化投资相关竞赛。2022 年 1 月，国内量化私募九坤投资也上线 Kaggle 竞赛，受到市场关注，2893 支队伍参赛，最终前 10 名队伍获得 10 万美元奖金。

图表1：Kaggle 平台量化投资相关竞赛

发布时间	发布机构	竞赛描述	网址
2015 年 10 月	Winton	利用股票 T-2 至 T 日中行情等数据，预测 T 日中至 T+2 日收益率	https://www.kaggle.com/competitions/the-winton-stock-market-challenge
2016 年 12 月	Two Sigma	利用资产匿名特征，预测价格	https://www.kaggle.com/competitions/two-sigma-financial-modeling/
2018 年 9 月	Two Sigma	利用新闻数据，预测股票价格	https://www.kaggle.com/competitions/two-sigma-financial-news
2020 年 11 月	Jane Street	利用股票匿名特征，制定交易策略	https://www.kaggle.com/competitions/jane-street-market-prediction
2021 年 6 月	Optiver	利用股票订单簿数据，预测未来 10 分钟波动率	https://www.kaggle.com/competitions/optiver-realized-volatility-prediction/
2021 年 11 月	G-research	利用数字货币行情数据。预测未来 15 分钟的残差收益率	https://www.kaggle.com/competitions/g-research-crypto-forecasting/
2022 年 1 月	九坤投资	利用股票匿名特征，预测收益率，最大化 IC 值	https://www.kaggle.com/competitions/ubiquant-market-prediction/
2022 年 4 月	日本交易所集团	利用股票行情、财报等数据，预测未来收益率排序，最大化多空组合夏普比率	https://www.kaggle.com/competitions/jpx-tokyo-stock-exchange-prediction

资料来源：Kaggle，华泰研究

本文的主题是“抄作业”，九坤 Kaggle 量化大赛高手云集，高分队伍是否有经验值得借鉴？我们梳理了部分高分队伍公布的解决方案，提炼出有共性的四个方向——**特征工程、损失函数、交叉验证和模型集成**，并应用于中证 500 指数增强策略的改进。结果显示，改进策略相比基线策略有稳定提升，回测期 2011 年至 2022 年内，年化超额收益从 14.2%提升至 17.0%，信息比率从 2.3/2.4 提升至 2.7。测试的改进技巧中，神经网络引入均值因子、CCC 损失、模型集成提升作用较显著。

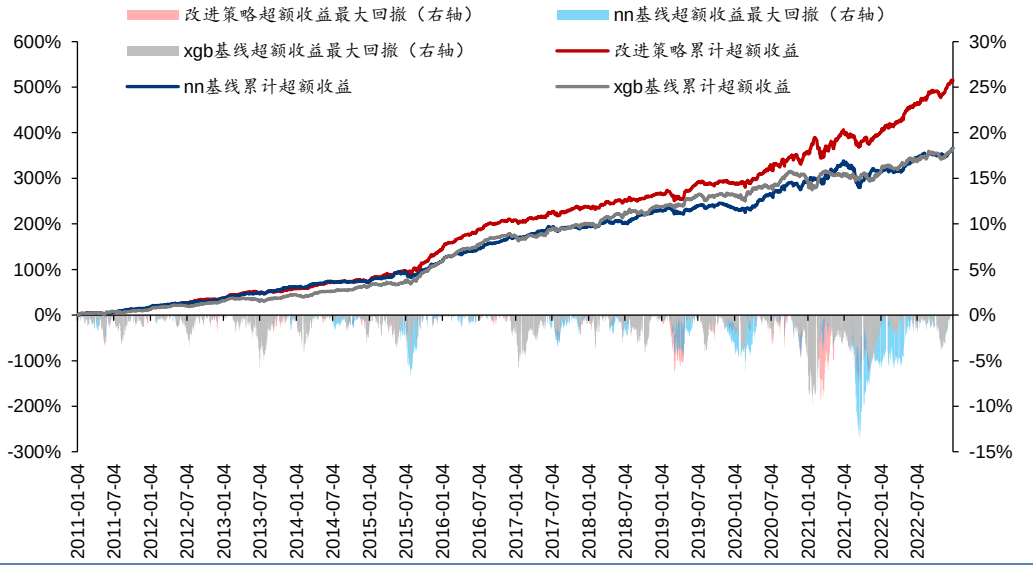
图表2：部分测试模型回测绩效

	年化收益率	年化波动率	夏普比率	最大回撤	Calmar 比率	年化超额收益率	年化跟踪误差	信息比率	超额收益最大回撤	超额收益 Calmar 比率	相对基准月胜率	年化双边换手率
							基线策略					
nn	15.94%	25.69%	0.62	50.25%	0.32	14.24%	5.99%	2.38	13.36%	1.07	77.08%	16.18
xgb	15.82%	26.07%	0.61	46.94%	0.34	14.22%	6.28%	2.26	9.70%	1.47	68.75%	16.26
							改进策略					
nn_fe+nn_wccc+xgb	18.56%	26.33%	0.70	48.96%	0.38	17.00%	6.24%	2.73	9.32%	1.82	76.39%	16.31
nn_fe+nn_wccc+xgb_cv	18.57%	26.38%	0.70	49.46%	0.38	17.03%	6.36%	2.68	9.54%	1.79	73.61%	16.31

注：回测期 2011-01-04 至 2022-12-30，基准为中证 500 指数

资料来源：朝阳永续，Wind，华泰研究

图表3：基于九坤 Kaggle 量化大赛的改进策略超额收益表现



注：回测期 2011-01-04 至 2022-12-30，基准为中证 500 指数；展示的改进策略为 nn_fe+nn_wccc+xgb

资料来源：朝阳永续，Wind，华泰研究

九坤 Kaggle 量化大赛高分方案解析

九坤 Kaggle 量化大赛的具体任务为：基于给定的 A 股匿名特征，采用机器学习或深度学习算法，预测股票未来短期收益，评价指标为预测收益和真实收益的 IC 值均值，属于典型的监督学习问题。

大赛提供的训练数据超过 18GB，每条样本为一只股票在一个交易日的数据，包含如下字段：

1. time_id：时间 id，为有序数据。
2. investment_id：股票 id。
3. [f_0:f_299]：300 个匿名特征。
4. target：预测目标，股票未来一段时间的收益率，但未公布具体区间。

图表4：九坤 Kaggle 量化大赛排名靠前方案

排名	模型架构	训练数据	特征工程	损失函数	交叉验证	其他
1	5 LGBM + 5 TabNet 的平均	train 和 supplemental_train 合并后的 2400000 行	300 个官方提供的特征 100 个按 time_id 求平均的特征	RMSE MSE IC	KFold PurgedGroupTimeSeries TimeSeriesSplit	
2	5 LGBM 的平均	train 和 supplemental_train 所有数据	300 个官方提供的特征 100 个按 time_id 求平均的特征 300 个特征的均值、标准差及分位数	RMSE IC	Purged KFold	AE MLP、特征中性化、PCA 并没有起到作用
3	5 个不同随机数种子 Transformer（6 层，最大长度 3500）的平均	train 和 supplemental_train 所有数据	使某些特征随机变为 0 对序列随机掩码	IC	Last K (K=100,200,300) Validation	特征裁剪、按 time_id 分组取平均的特征、按相关性找出来的特征、样本筛选和加权、LGBM、MLP、1DCNN 效果不如 Transformer
5	1 个 NN（4 层）	time_id > 599 的 train 和 supplemental_train 所有数据	Quantile Transformer	MSE	Purged KFold	Target 取对数并剔除离群值
7	1 个 LGBM	train 的所有数据	从大的特征池中产生 900 多特征		TimeSeriesSplit	
8	10 个 LGBM + 30 个 NN + 1 个自制模型的加权平均		300 个官方提供的特征 每个 time_id 平均的特征 其他特征			
17	1 个 NN（2 层）		276 个官方提供的特征 1 个构造的 A 股市场“停牌”特征	L1 L2		

注：LGBM 为 LightGBM 的缩写，NN 为神经网络的缩写

资料来源：Kaggle，华泰研究

比赛于 2022 年 1 月开始，7 月公布最终成绩，部分排名靠前队伍公开了解决方案，如上表。我们从众多方案中提炼具有共性的技巧，从特征工程、损失函数、交叉验证和模型集成四个方向展开介绍。

特征工程

特征工程是模型搭建前的数据预处理和新特征构造工作，特征工程的质量一定程度上决定了预测结果的好坏。数据预处理主要包括缺失值填充、异常值剔除和标准化；新特征的构造则依赖投资者的经验和对市场的理解。原始特征可能无法很好反映样本和潜在问题的关系，通过引入对原始特征处理和组合后的新特征，或可提升模型训练效果。

九坤量化大赛中，原始数据进行了预处理和匿名处理，无法基于因子含义构造新特征，这给特征构造增加了困难。我们发现多数高分队伍都进行了新特征的构造，仅有少数方案只使用原始的 300 维特征。第 1、2、8 名队伍都提到了构造**“按照时间 ID 取平均的均值因子”**，他们指出均值因子的引入对于模型效果有显著提升。

具体而言，假设 f_0 为某个因子，在每个交易日对全部股票的 f_0 求均值，即得到该交易日股票的均值因子 f_mean_0。该交易日全部股票的 f_mean_0 取值相同，在交易日间有差异，反映 f_0 因子整体分布的时变特性。这一操作在传统机器学习中似乎不常见，构造一个全部股票取值相同的因子也略显反常，但在九坤量化大赛中有效。

关于均值因子的有效性和背后的含义，我们猜想：与其他领域的预测问题相比，股票收益率预测有其特殊性——未来表现不仅和股票本身特征相关，还与市场整体环境（如宏观状态、市场风格等）相关，**规律存在时变特性**，因此有必要引入特征刻画市场环境变化。**均值因子反映原始因子整体分布的时变特性**，是市场环境的一种简单表达，可能具备一定信息量。

损失函数

损失函数决定了模型的优化方向，损失函数的选择取决于评价指标、下游任务等因素。九坤量化大赛的最终评价指标为预测收益和真实收益的 Pearson 相关系数，即 IC 值，衡量预测值和真实值的线性相关程度，部分高分队伍直接采用 IC 值的相反数作为损失函数：

$$IC(y,\hat{y}) = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\hat{\bar{y}})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i-\hat{\bar{y}})^2}}$$

均方误差 MSE 是回归任务中常用的损失函数之一，衡量预测值和真实值间的距离，对偏离真实值的预测给予较大惩罚。部分高分队伍也采用 MSE 或 RMSE（MSE 的平方根）作为损失函数：

$$MSE(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

总结 IC 和 MSE 作为损失函数的优缺点：

1. **IC 衡量预测值和真实值的相关性**。优点是和比赛最终的评价指标直接挂钩，也是量化机构都会考察的指标，不受量纲影响从而在模型间可比。缺点是非凸不保证收敛，可能导致训练不稳定。
2. **MSE 衡量预测值和真实值的距离**。优点是易于计算和求导，具有凸性从而保证收敛，在数据噪声较小的情况下可作为 IC 的替代。缺点是受数据量纲影响。

九坤量化大赛的讨论区里，有选手提出使用**一致性相关系数 CCC（concordance correlation coefficient）作为 IC 和 MSE 的融合，同时考虑相关性和距离**。CCC 由 Lawrence I-Kuei Lin 在 1989 年于 Biometrics 发表的论文 *A concordance correlation coefficient to evaluate reproducibility* 中提出：

$$CCC = \frac{2\rho_{xy}\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2+(\mu_x-\mu_y)^2} \quad (1)$$

Pandit 和 Schuller 在 2019 年于 arXiv 平台发布的论文 *The many-to-many mapping between the concordance correlation coefficient and the mean square error* 推导了其等价形式：

$$CCC = \frac{2\sigma_{xy}}{MSE+2\sigma_{xy}} \quad (2)$$

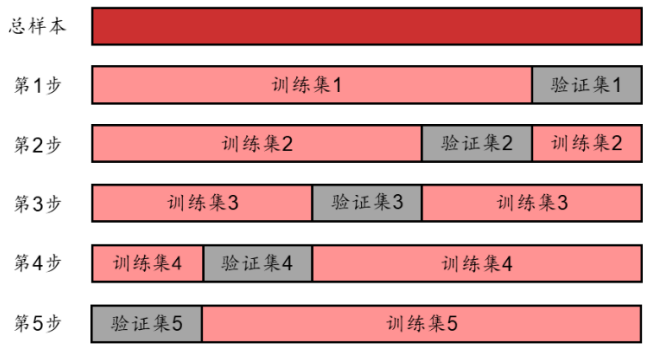
观察 CCC 的定义，(1)式分子中的 ρ_{xy} 代表 x 和 y 的 Pearson 相关系数，即 IC；(2)式分母包含 MSE。直观来看，分子考虑两组数据的相关性，分母对两组数据均值的偏离度进行了惩罚。实际使用中，可以取 CCC 的相反数作为损失函数。尽管高分队伍未使用 CCC 损失，我们仍可以从讨论区中获得启发。

交叉验证

交叉验证主要用于选择模型超参数。最简单的方式是单次验证，即选择固定比例的训练集和验证集。常用的方式是 K 折交叉验证，将原始数据分成 K 份，每次使用 K-1 份训练模型，使用剩余 1 份评价模型，对 K 次评价取平均作为该组超参数的整体评价。但 K 折的缺点是可能使用未来信息，第 1、7 名队伍均提到该问题，并提出使用时序交叉验证。

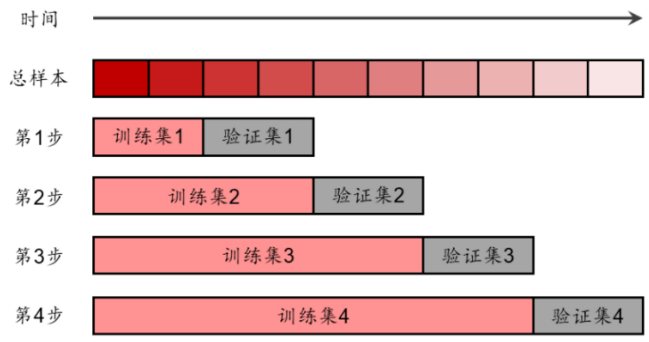
我们在《人工智能 14：对抗过拟合：从时序交叉验证谈起》(2018-11-28) 中介绍过该方法。时序交叉验证将原始数据按时间顺序划分为 K 份，第 i 次验证时，使用 1 至 i 份训练模型，第 i+1 份评价模型，避免未来信息，使用数据量约是 K 折交叉验证的一半。总的来看，**时序交叉验证的优点是无未来信息，且使用数据量少时间开销低，缺点是可能存在欠拟合风险。**

图表5：K 折交叉验证示意图



资料来源：华泰研究

图表6：时序交叉验证示意图



资料来源：华泰研究

模型集成

模型集成可以看成机器学习中“免费的午餐”。完美训练单个模型难度很大，模型集成通过融合多个子模型，实现取长补短，为比赛中多数高分队伍采用。第 1、8、17 名集成了多种不同类型的子模型，如决策树类模型和神经网络模型；第 2、3 名集成了多个同类型的子模型。尽管投票法、Stacking 等模型集成方法层出不穷，比赛中仍主要采用最简单的等权法。

高分队伍使用决策树类模型和神经网络模型作为子模型，两者有各自优势，集成能起到互补效果。决策树类模型对于数据的要求相对较低，对异常值、缺失值和特征间数量级不敏感，是在实操中较常用的一类模型。神经网络一般要求数据数量级一致、不能有缺失值，但可以通过批量训练将多个截面的信息一并地输入到模型中，自动构造出有效的新特征。

除上述四项外，高分队伍在模型架构上亦有可取之处，如第 1 名采用 TabNet，第 3 名采用 Transformer，但模型本身不是本研究关注的重点，故不作进一步测试。有少数队伍采用了独特的训练技巧，如第 3 名使某些特征随机变为 0，第 5 名对预测目标取对数，对特征做分位数转换（未指明转换成何种分布），上述个性化处理也不在后文讨论之列。

方法

本研究在现有周频中证 500 指增模型基础上，引入九坤量化大赛中的技巧，测试改进效果。全部测试模型如下表。

图表7：全部测试模型

	机器学习模型	特征	损失函数	交叉验证
			基线	
nn	全连接神经网络	42 个因子	加权 mse	单次验证，仅调迭代次数
xgb	XGBoost	42 个因子	加权 mse	单次验证，仅调迭代次数
			特征工程	
nn_fe	全连接神经网络	42 个因子+42 个均值因子	加权 mse	单次验证，仅调迭代次数
xgb_fe	XGBoost	42 个因子+42 个均值因子	加权 mse	单次验证，仅调迭代次数
			损失函数	
nn_mse	全连接神经网络	42 个因子	mse	单次验证，仅调迭代次数
nn_wmse（基线）	全连接神经网络	42 个因子	加权 mse	单次验证，仅调迭代次数
nn_ic	全连接神经网络	42 个因子	ic	单次验证，仅调迭代次数
nn_wic	全连接神经网络	42 个因子	加权 ic	单次验证，仅调迭代次数
nn_ccc	全连接神经网络	42 个因子	ccc	单次验证，仅调迭代次数
nn_wccc	全连接神经网络	42 个因子	加权 ccc	单次验证，仅调迭代次数
			交叉验证	
xgb_cv	XGBoost	42 个因子	加权 mse	5 折时序交叉验证，调迭代次数及 XGBoost 超参数
			模型集成	
nn+xgb	50%*nn+50%*xgb			
nn_wccc+xgb	50%*nn_wccc+50%*xgb			
nn_fe+nn_wccc+xgb	25%*nn_fe+25%*nn_wccc+50%*xgb			
nn_fe+nn_wccc+xgb_cv	25%*nn_fe+25%*nn_wccc+50%*xgb_cv			

资料来源：华泰研究

基线模型为全连接神经网络（nn）和 XGBoost（xgb），特征为 42 个常规的基本面和量价因子，标签为未来 10 个交易日收益率在截面上的排序，损失函数为加权 mse（wmse），以截面上个股收益率排序进行衰减加权。交叉验证方法为单次验证，以 252*6 个交易日为训练集，252*2 个交易日为验证集，252*0.5 个交易日为测试集，相当于约半年滚动训练一次。交叉验证配合早停，仅用于确定模型的迭代次数，其余超参数均为固定值。

下面介绍四个方向的改进技巧：

1. 特征工程：除原始 42 个因子外，增加 42 个均值因子。针对每个原始因子，首先进行去极值；其次在截面上将因子**转换为标准差等于 1 的分布**，避免因子间量纲差异的影响；随后对截面上全部股票求均值；最后**整体乘以 0.01**，突出原始因子作用，弱化均值因子影响。整体乘以 0.01 对模型的影响将在后文讨论。
2. 损失函数：测试 MSE、IC、CCC 三类损失函数，每类损失又分为等权和加权两种情况。其中加权 CCC 定义为：

$$\mu_x = \frac{1}{N}\sum\nolimits_N w_i x_i, \quad \mu_y = \frac{1}{N}\sum\nolimits_N w_i y_i$$

$$\sigma_x^2 = \frac{1}{N}\sum\nolimits_N w_i (x_i - \mu_x)^2, \quad \sigma_y^2 = \frac{1}{N}\sum\nolimits_N w_i (y_i - \mu_y)^2$$

$$Weighted\_CCC = \frac{\frac{1}{N}\sum_N w_i x_i y_i - \mu_x \mu_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

   XGBoost 不便于自定义此类损失函数，故测试仅针对全连接神经网络。
3. 交叉验证：采用 5 折时序交叉验证结合网格搜索，确定 XGBoost 学习率和最大树深。神经网络训练时间开销大，故测试仅针对 XGBoost。同样受限于时间开销，本文未测试 K 折交叉验证，网格搜索颗粒度也较粗。超参数搜索方式的优化有待进一步研究。

4. 模型集成：直接对全连接神经网络和 XGBoost 预测值取均值，两类模型等权。若神经网络有细分子模型，则内部再进行等权平均。

选股因子、模型构建方法及网络结构如下列图表。具体细节可参考华泰金工研报《人工智能 55：图神经网络选股的进阶之路》（2022-04-11）。

图表8：选股模型使用的 42 个因子

类别	名称	计算方式
估值	bp_lf	1/市净率
	ep_ttm	1/市盈率(TTM)
	ocfp_ttm	1/净经营性现金流(TTM)
	dyr12	近 252 日股息率
预期	con_eps_g	一致预期 EPS(FY1)近 63 日增长率
	con_roe_g	一致预期 ROE(FY1)近 63 日增长率
	con_np_g	一致预期归母净利润(FY1)近 63 日增长率
反转	ret_5d	近 5 日区间收益率
	ret_1m	近 21 日区间收益率
	exp_wgt_return_3m	近 63 日收益率以换手率指数衰减加权
波动率	std_1m	收益率近 21 日标准差
	vstd_1m	成交量近 21 日标准差
	ivr_ff3factor_1m	残差收益率（收益率对万得全 A、市值、BP 因子收益率回归）近 21 日标准差
换手率	turn_1m	换手率近 21 日均值
	std_turn_1m	换手率近 21 日标准差
	bias_turn_1m	换手率近 21 日均值/近 504 日均值
日间技术	std_ret_10d	收益率近 10 日标准差
	std_vol_10d	成交量近 10 日标准差
	std_turn_10d	换手率近 10 日标准差
	corr_ret_close	收益率和收盘价近 10 日相关系数
	corr_ret_open	收益率和开盘价近 10 日相关系数
	corr_ret_high	收益率和最高价近 10 日相关系数
	corr_ret_low	收益率和最低价近 10 日相关系数
	corr_ret_vwap	收益率和均价近 10 日相关系数
	corr_ret_vol	收益率和成交量近 10 日相关系数
	corr_ret_turn	收益率和换手率近 10 日相关系数
	corr_vol_close	成交量和收盘价近 10 日相关系数
	corr_vol_open	成交量和开盘价近 10 日相关系数
	corr_vol_high	成交量和最高价近 10 日相关系数
	corr_vol_low	成交量和最低价近 10 日相关系数
	corr_vol_vwap	成交量和均价近 10 日相关系数
日内技术	low2high	low/high
	vwap2close	vwap/close
	kmid	(close-open)/open
	klen	(high-low)/open
	kmid2	(close-open)/(high-low)
	kup	(high-greater(open,close))/open
	kup2	(high-greater(open,close))/(high-low)
	klow	(less(open,close)-low)/open
	klow2	(less(open,close)-low)/(high-low)
	ksft	(2*close-high-low)/open
	ksft2	(2*close-high-low)/(high-low)

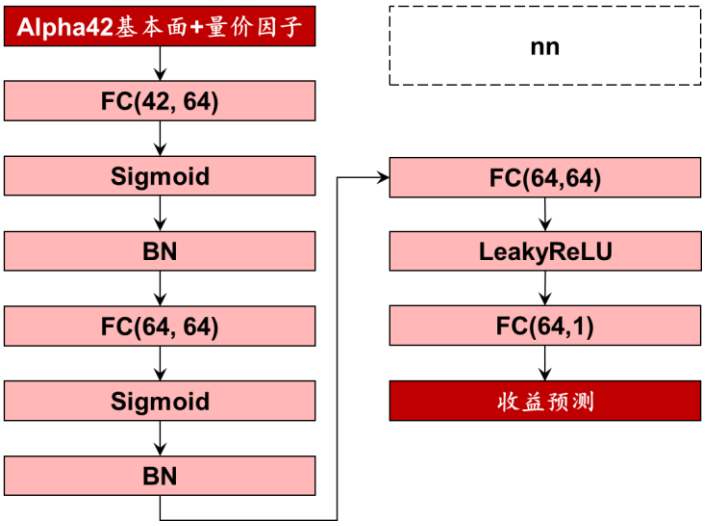
资料来源：朝阳永续，Wind，华泰研究

图表9：选股模型构建方法

步骤	参数	参数值
构建股票池	股票池	全 A 股；剔除上市未满 63 个交易日个股，剔除 ST、*ST、退市整理期个股；每个季末截面期，在未停牌个股中，筛选过去 1 年日均成交额和日均总市值均排名前 60%个股
构建数据集	特征	T 日 42 个基本面和量价因子
	标签	T+11 日相对于 T+1 日 vwap 收益率
因子预处理	特征	5 倍 MAD 缩尾；zscore 标准化；缺失值填为 0；不做中性化
	标签	剔除缺失值；截面排序数标准化
训练流程	测试集完整区间	20110104~20221230
	训练、验证、测试集划分	训练集 252*6 个交易日，验证集 252*2 个交易日，测试集 126 个交易日； 如第 1 期训练集 20020910~20081205，验证集 20081208~20101231，测试集 20110104~20110711； 第 2 期训练集 20030325~20090616，验证集 20090617~20110711，测试集 20110712~20120113
	特殊处理	剔除训练集、验证集最后 10 个交易日样本，防止信息泄露
设置模型（全连接神经网络）	网络结构	全连接神经网络
	隐单元数	64
	损失函数	基准为加权 mse 测试 mse、ic 或 ccc，采用等权或根据收益率衰减加权
	batch size	每个交易日的全体股票视作一个 batch
	学习率	0.0001
	优化器	adam
	早停次数	20
设置模型（XGBoost）	模型	XGBoost
	损失函数	加权 mse
	学习率	0.05 交叉验证测试 0.05、0.075、0.1
	最大迭代次数	1000
	早停次数	20
	最大树深	3 交叉验证测试 3、5、7
	行列采样比例	0.8
构建组合	基准	中证 500 指数
	优化目标	最大化预期收益
	组合仓位	1
	个股权重下限	0
	个股偏离权重约束	[-1%, 1%]
	行业偏离权重约束	[-1%, 1%]
	风格偏离标准差约束	[-1%, 1%]
	风格因子	对数流通市值（预处理：5 倍 MAD 缩尾，zscore 标准化）
	调仓周期	每 5 个交易日
	单次调仓单边换手率上限	15%
	成分股权重约束	无
回测	单边费率	0.002
	交易价格	vwap
	特殊处理	停牌不买入/卖出；一字板涨停不买入；一字板跌停不卖出；其余可交易股票重新分配权重

资料来源：华泰研究

图表10： 网络结构



资料来源：华泰研究

结果

全部测试模型因子评价指标及回测绩效如下列图表。核心结论如下：

1. 特征工程引入的均值因子对神经网络有提升，但削弱了 XGBoost。
2. 损失函数中，MSE 表现不突出；IC 损失单因子测试表现好，但指增组合回测表现差；CCC 损失在单因子测试表现一般，但指增组合回测表现较好；加权均优于等权。
3. 交叉验证调参改进不显著，考虑到时间开销大，性价比并不高，算力有限前提下，使用经验超参数即可。
4. 模型集成提升较稳定，神经网络类和决策树类模型有互补效果。

图表11： 全部测试模型合成因子评价指标（回测期 2011-01-04 至 2022-12-30）

	IC 均值	RankIC 均值	加权 IC 均值	加权 RankIC 均值	ICIR	RankICIR	加权 ICIR	加权 RankICIR	Top 组精确率	Bottom 组精确率	Top 组年化收益率	Bottom 组年化收益率	多空对冲年化收益率	基准收益率
							基线							
nn	8.5%	10.6%	7.1%	9.6%	0.83	0.99	0.70	0.90	56.3%	59.8%	24.7%	-28.1%	26.4%	6.3%
xgb	8.7%	11.3%	6.5%	9.7%	0.75	0.95	0.57	0.82	56.5%	60.2%	23.3%	-30.1%	26.7%	6.3%
							特征工程							
nn_fe	8.5%	10.7%	7.1%	9.7%	0.83	0.99	0.70	0.91	56.3%	59.9%	25.5%	-28.3%	26.9%	6.3%
xgb_fe	6.8%	8.8%	5.1%	7.6%	0.58	0.72	0.44	0.62	54.8%	57.8%	20.6%	-21.9%	21.3%	6.3%
							损失函数							
nn_mse	8.6%	11.1%	6.7%	9.7%	0.81	1.02	0.64	0.89	56.6%	60.0%	23.7%	-29.4%	26.6%	6.3%
nn_wmse（基线）	8.5%	10.6%	7.1%	9.6%	0.83	0.99	0.70	0.90	56.3%	59.8%	24.7%	-28.1%	26.4%	6.3%
nn_ic	8.6%	11.2%	6.6%	9.7%	0.76	0.95	0.59	0.84	56.6%	60.1%	23.3%	-29.8%	26.6%	6.3%
nn_wic	8.7%	10.9%	7.2%	9.8%	0.81	0.97	0.68	0.88	56.3%	60.0%	25.0%	-29.3%	27.2%	6.3%
nn_ccc	8.5%	11.1%	6.5%	9.7%	0.77	0.97	0.59	0.85	56.4%	60.2%	21.7%	-30.1%	25.9%	6.3%
nn_wccc	8.7%	10.9%	7.2%	9.8%	0.82	0.97	0.68	0.89	56.2%	60.1%	24.2%	-29.6%	26.9%	6.3%
							交叉验证							
xgb_cv	8.7%	11.2%	6.5%	9.6%	0.75	0.94	0.57	0.81	56.4%	60.3%	23.0%	-29.8%	26.4%	6.3%
							模型集成							
nn+xgb	9.1%	11.5%	7.2%	10.2%	0.83	1.02	0.66	0.91	56.8%	60.5%	25.2%	-31.4%	28.3%	6.3%
nn_wccc+xgb	8.9%	11.1%	7.3%	10.0%	0.83	0.99	0.68	0.90	56.3%	60.3%	24.3%	-30.4%	27.3%	6.3%
nn_fe+nn_wccc+xgb	9.0%	11.3%	7.3%	10.1%	0.83	1.00	0.68	0.91	56.5%	60.5%	24.9%	-31.0%	27.9%	6.3%
nn_fe+nn_wccc+xgb_cv	9.0%	11.3%	7.3%	10.1%	0.82	0.99	0.68	0.90	56.5%	60.5%	24.9%	-31.0%	27.9%	6.3%

资料来源：朝阳永续，Wind，华泰研究

图表12： 全部测试模型回测绩效（回测期 2011-01-04 至 2022-12-30，基准为中证 500 指数）

	年化收益率	年化波动率	夏普比率	最大回撤	Calmar 比率	年化超额收益率	年化跟踪误差	信息比率	超额收益最大回撤	超额收益 Calmar 比率	相对基准月胜率	年化双边换手率
							基线					
nn	15.94%	25.69%	0.62	50.25%	0.32	14.24%	5.99%	2.38	13.36%	1.07	77.08%	16.18
xgb	15.82%	26.07%	0.61	46.94%	0.34	14.22%	6.28%	2.26	9.70%	1.47	68.75%	16.26
							特征工程					
nn_fe	17.41%	26.25%	0.66	48.20%	0.36	15.83%	6.49%	2.44	15.25%	1.04	77.78%	16.18
xgb_fe	12.79%	26.68%	0.48	49.54%	0.26	11.37%	6.80%	1.67	12.37%	0.92	68.06%	16.26
							损失函数					
nn_mse	14.43%	25.91%	0.56	47.44%	0.30	12.83%	5.87%	2.19	14.65%	0.88	75.69%	16.44
nn_wmse（基线）	15.94%	25.69%	0.62	50.25%	0.32	14.24%	5.99%	2.38	13.36%	1.07	77.08%	16.18
nn_ic	10.03%	23.94%	0.42	52.90%	0.19	7.92%	6.52%	1.21	19.04%	0.42	61.81%	14.32
nn_wic	10.32%	23.98%	0.43	52.71%	0.20	8.26%	5.89%	1.40	11.85%	0.70	59.72%	14.21
nn_ccc	15.08%	25.78%	0.59	48.50%	0.31	13.44%	5.85%	2.30	7.64%	1.76	72.92%	16.35
nn_wccc	17.06%	25.97%	0.66	47.93%	0.36	15.42%	6.10%	2.53	9.82%	1.57	75.00%	16.32
							交叉验证					
xgb_cv	15.82%	26.32%	0.60	47.90%	0.33	14.27%	6.50%	2.20	6.40%	2.23	71.53%	16.27
							模型集成					
nn+xgb	16.78%	25.95%	0.65	48.29%	0.35	15.15%	5.99%	2.53	9.36%	1.62	79.17%	16.20
nn_wccc+xgb	17.57%	26.06%	0.67	47.97%	0.37	15.95%	6.10%	2.62	9.20%	1.73	74.31%	16.30
nn_fe+nn_wccc+xgb	18.56%	26.33%	0.70	48.96%	0.38	17.00%	6.24%	2.73	9.32%	1.82	76.39%	16.31
nn_fe+nn_wccc+xgb_cv	18.57%	26.38%	0.70	49.46%	0.38	17.03%	6.36%	2.68	9.54%	1.79	73.61%	16.31

资料来源：朝阳永续，Wind，华泰研究

特征工程

对比引入均值因子前后的表现。神经网络无论在 Top 组收益，还是在指增组合年化超额收益、信息比率方面，均有显著提升。但 XGBoost 在上述指标均有较大削弱。原因可能是 XGBoost 对均值因子的“过度”使用，具体将在后文探讨。

图表13： 特征工程测试模型合成因子评价指标（回测期 2011-01-04 至 2022-12-30）

	IC 均值	RankIC 均值	加权 IC 均值	加权 RankIC 均值	ICIR	RankICIR	加权 ICIR	加权 RankICIR	Top 组精确率	Bottom 组精确率	Top 组年化收益率	Bottom 组年化收益率	多空对冲年化收益率	基准收益率
							基线							
nn	8.5%	10.6%	7.1%	9.6%	0.83	0.99	0.70	0.90	56.3%	59.8%	24.7%	-28.1%	26.4%	6.3%
xgb	8.7%	11.3%	6.5%	9.7%	0.75	0.95	0.57	0.82	56.5%	60.2%	23.3%	-30.1%	26.7%	6.3%
							特征工程							
nn_fe	8.5%	10.7%	7.1%	9.7%	0.83	0.99	0.70	0.91	56.3%	59.9%	25.5%	-28.3%	26.9%	6.3%
xgb_fe	6.8%	8.8%	5.1%	7.6%	0.58	0.72	0.44	0.62	54.8%	57.8%	20.6%	-21.9%	21.3%	6.3%

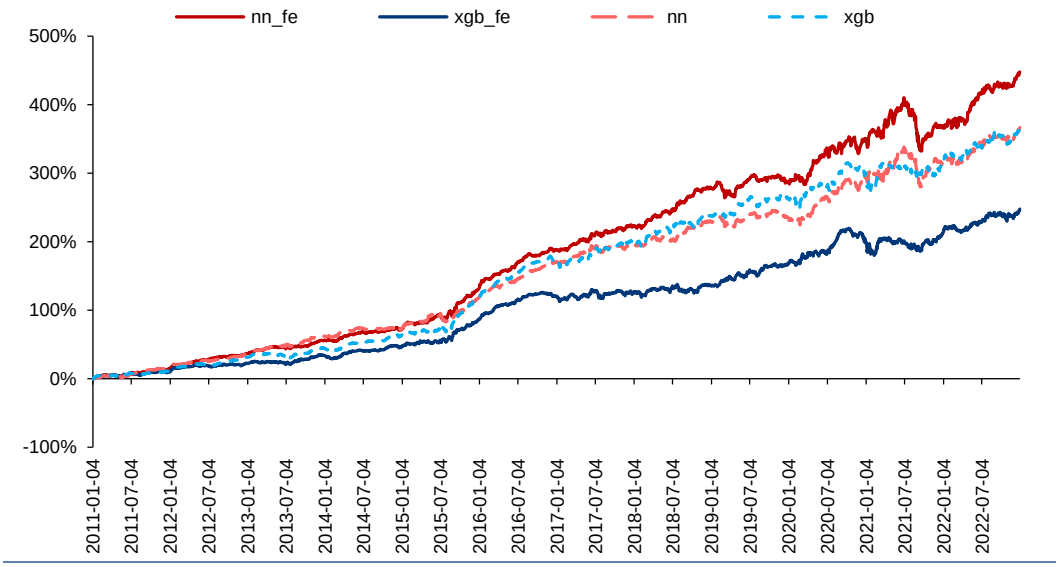
资料来源：朝阳永续，Wind，华泰研究

图表14： 特征工程测试模型回测绩效（回测期 2011-01-04 至 2022-12-30，基准为中证 500 指数）

	年化收益率	年化波动率	夏普比率	最大回撤	Calmar 比率	年化超额收益率	年化跟踪误差	信息比率	超额收益最大回撤	超额收益 Calmar 比率	相对基准月胜率	年化双边换手率
						基线						
nn	15.94%	25.69%	0.62	50.25%	0.32	14.24%	5.99%	2.38	13.36%	1.07	77.08%	16.18
xgb	15.82%	26.07%	0.61	46.94%	0.34	14.22%	6.28%	2.26	9.70%	1.47	68.75%	16.26
						特征工程						
nn_fe	17.41%	26.25%	0.66	48.20%	0.36	15.83%	6.49%	2.44	15.25%	1.04	77.78%	16.18
xgb_fe	12.79%	26.68%	0.48	49.54%	0.26	11.37%	6.80%	1.67	12.37%	0.92	68.06%	16.26

资料来源：朝阳永续，Wind，华泰研究

图表15： 特征工程测试模型超额收益表现（回测期 2011-01-04 至 2022-12-30，基准为中证 500 指数）



资料来源：朝阳永续，Wind，华泰研究

损失函数

对比神经网络模型 MSE、IC、CCC 三类损失函数，以及等权和多头加权两种方式的表现。

单因子测试结果可概括为“种瓜得瓜，种豆得豆”。IC 损失下的 IC 均值和 Rank IC 均值较高，加权 IC 损失下的加权 IC 均值和加权 Rank IC 均值较高。多头加权损失的 Top 组收益均高于对应的等权损失。

但单因子测试和指增组合测试存在错位。加权 IC 损失的单因子多头收益和对冲收益均高于其余损失，但指增组合表现却低于除等权 IC 损失外的其余损失。CCC 损失的单因子表现不算突出，但从指增组合表现看，无论是等权和加权，均优于对应的 MSE 和 IC 损失。加权 CCC 损失的年化超额收益和信息比率较出色。

图表16：损失函数测试模型合成因子评价指标（回测期 2011-01-04 至 2022-12-30）

	IC 均值	RankIC 均值	加权 IC 均值	加权 RankIC 均值	ICIR	RankICIR	加权 ICIR	加权 RankICIR	Top 组精确率	Bottom 组精确率	Top 组年化收益率	Bottom 组年化收益率	多空对冲年化收益率	基准收益率
nn_mse	8.6%	11.1%	6.7%	9.7%	0.81	1.02	0.64	0.89	56.6%	60.0%	23.7%	-29.4%	26.6%	6.3%
nn_wmse（基线）	8.5%	10.6%	7.1%	9.6%	0.83	0.99	0.70	0.90	56.3%	59.8%	24.7%	-28.1%	26.4%	6.3%
nn_ic	8.6%	11.2%	6.6%	9.7%	0.76	0.95	0.59	0.84	56.6%	60.1%	23.3%	-29.8%	26.6%	6.3%
nn_wic	8.7%	10.9%	7.2%	9.8%	0.81	0.97	0.68	0.88	56.3%	60.0%	25.0%	-29.3%	27.2%	6.3%
nn_ccc	8.5%	11.1%	6.5%	9.7%	0.77	0.97	0.59	0.85	56.4%	60.2%	21.7%	-30.1%	25.9%	6.3%
nn_wccc	8.7%	10.9%	7.2%	9.8%	0.82	0.97	0.68	0.89	56.2%	60.1%	24.2%	-29.6%	26.9%	6.3%

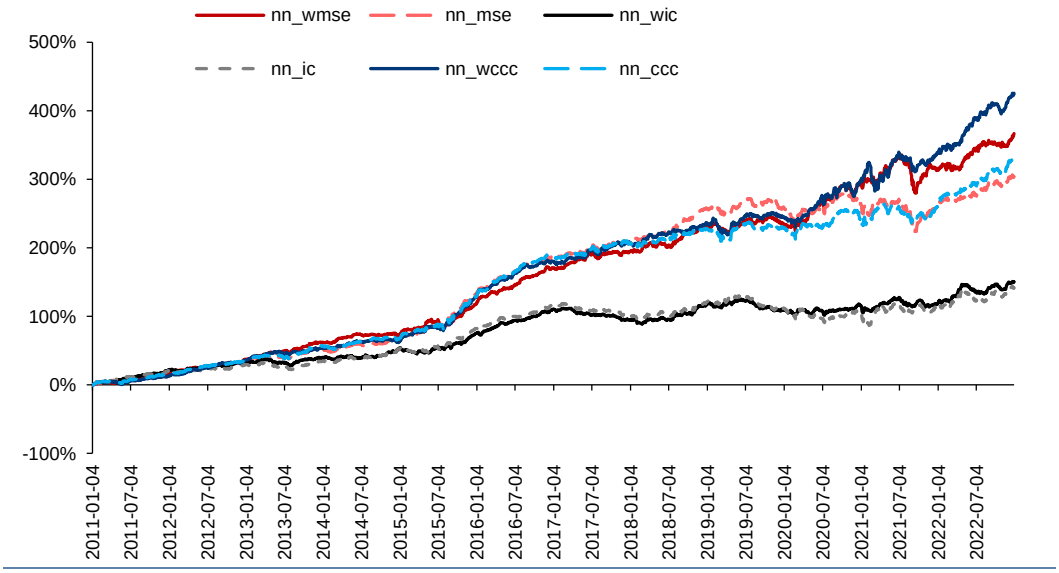
资料来源：朝阳永续，Wind，华泰研究

图表17：损失函数测试模型回测绩效（回测期 2011-01-04 至 2022-12-30，基准为中证 500 指数）

	年化收益率	年化波动率	夏普比率	最大回撤	Calmar 比率	年化超额收益率	年化跟踪误差	信息比率	超额收益最大回撤	超额收益 Calmar 比率	相对基准月胜率	年化双边换手率
nn_mse	14.43%	25.91%	0.56	47.44%	0.30	12.83%	5.87%	2.19	14.65%	0.88	75.69%	16.44
nn_wmse（基线）	15.94%	25.69%	0.62	50.25%	0.32	14.24%	5.99%	2.38	13.36%	1.07	77.08%	16.18
nn_ic	10.03%	23.94%	0.42	52.90%	0.19	7.92%	6.52%	1.21	19.04%	0.42	61.81%	14.32
nn_wic	10.32%	23.98%	0.43	52.71%	0.20	8.26%	5.89%	1.40	11.85%	0.70	59.72%	14.21
nn_ccc	15.08%	25.78%	0.59	48.50%	0.31	13.44%	5.85%	2.30	7.64%	1.76	72.92%	16.35
nn_wccc	17.06%	25.97%	0.66	47.93%	0.36	15.42%	6.10%	2.53	9.82%	1.57	75.00%	16.32

资料来源：朝阳永续，Wind，华泰研究

图表18：损失函数测试模型超额收益表现（回测期 2011-01-04 至 2022-12-30，基准为中证 500 指数）



资料来源：朝阳永续，Wind，华泰研究

交叉验证

对比 XGBoost 模型时序交叉验证调参的表现，调参后的模型仅在指增组合超额收益回撤比指标上有显著提升，其余重要指标反而略有削弱。但交叉验证调参的时间开销（近 19 小时）远高于不调参（近 5 分钟），在算力有限情况下性价比不高。

需要说明的是，本文采用网格搜索的调参方法效率较低，从而导致调参颗粒度较粗糙。基于贝叶斯优化的调参方法可以提升搜索效率，有待进一步测试。

图表19：交叉验证测试模型合成因子评价指标（回测期 2011-01-04 至 2022-12-30）

	IC 均值	RankIC 均值	加权 IC 均值	加权 RankIC 均值	ICIR	RankICIR	加权 ICIR	加权 RankICIR	Top 组精确率	Bottom 组精确率	Top 组年化收益率	Bottom 组年化收益率	多空对冲年化收益率	基准收益率
xgb	8.7%	11.3%	6.5%	9.7%	0.75	0.95	0.57	0.82	56.5%	60.2%	23.3%	-30.1%	26.7%	6.3%
xgb_cv	8.7%	11.2%	6.5%	9.6%	0.75	0.94	0.57	0.81	56.4%	60.3%	23.0%	-29.8%	26.4%	6.3%

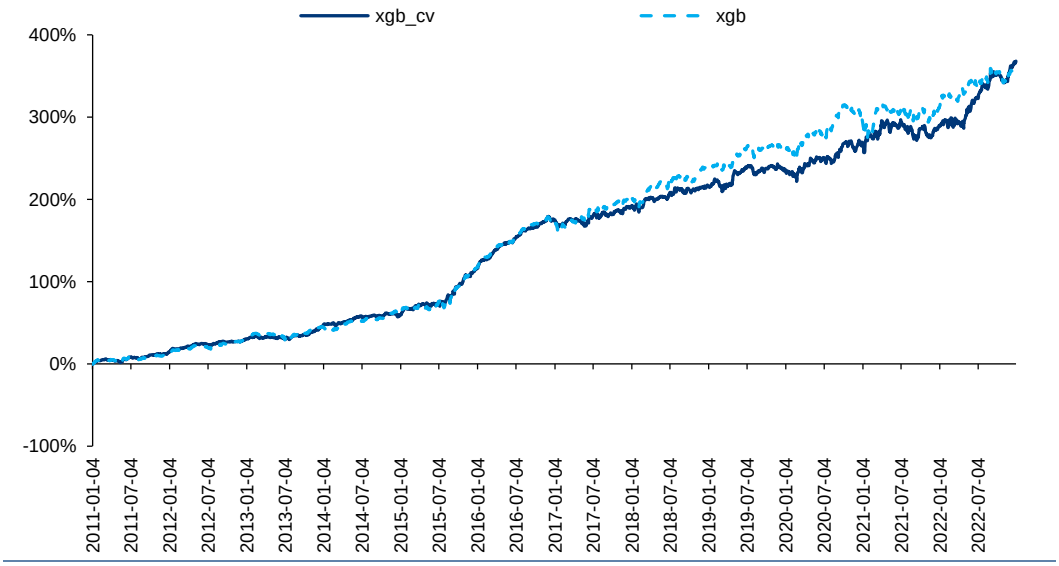
资料来源：朝阳永续，Wind，华泰研究

图表20：交叉验证测试模型回测绩效（回测期 2011-01-04 至 2022-12-30，基准为中证 500 指数）

	年化收益率	年化波动率	夏普比率	最大回撤	Calmar 比率	年化超额收益率	年化跟踪误差	信息比率	超额收益最大回撤	超额收益 Calmar 比率	相对基准月胜率	年化双边换手率
xgb	15.82%	26.07%	0.61	46.94%	0.34	14.22%	6.28%	2.26	9.70%	1.47	68.75%	16.26
xgb_cv	15.82%	26.32%	0.60	47.90%	0.33	14.27%	6.50%	2.20	6.40%	2.23	71.53%	16.27

资料来源：朝阳永续，Wind，华泰研究

图表21：交叉验证测试模型超额收益表现（回测期 2011-01-04 至 2022-12-30，基准为中证 500 指数）



资料来源：朝阳永续，Wind，华泰研究

模型集成

对比模型集成后的表现，各集成模型在单因子加权 RankIC 均值、多空收益、指增组合年化超额收益、信息比率上均有显著提升。并且子模型为改进模型（后 3 组）的表现优于子模型为原始模型（nn+xgb）。对比前述特征工程、损失函数、交叉验证的技巧，模型集成带来的提升幅度更大且效应更稳定。

图表22：模型集成测试模型合成因子评价指标（回测期 2011-01-04 至 2022-12-30）

	IC 均值	RankIC 均值	加权 IC 均值	加权 RankIC 均值	ICIR	RankICIR	加权 ICIR	加权 RankICIR	Top 组精确率	Bottom 组精确率	Top 组年化收益率	Bottom 组年化收益率	多空对冲年化收益率	基准收益率
基线														
nn	8.5%	10.6%	7.1%	9.6%	0.83	0.99	0.70	0.90	56.3%	59.8%	24.7%	-28.1%	26.4%	6.3%
xgb	8.7%	11.3%	6.5%	9.7%	0.75	0.95	0.57	0.82	56.5%	60.2%	23.3%	-30.1%	26.7%	6.3%
模型集成														
nn+xgb	9.1%	11.5%	7.2%	10.2%	0.83	1.02	0.66	0.91	56.8%	60.5%	25.2%	-31.4%	28.3%	6.3%
nn_wccc+xgb	8.9%	11.1%	7.3%	10.0%	0.83	0.99	0.68	0.90	56.3%	60.3%	24.3%	-30.4%	27.3%	6.3%
nn_fe+nn_wccc+xgb	9.0%	11.3%	7.3%	10.1%	0.83	1.00	0.68	0.91	56.5%	60.5%	24.9%	-31.0%	27.9%	6.3%
nn_fe+nn_wccc+xgb_cv	9.0%	11.3%	7.3%	10.1%	0.82	0.99	0.68	0.90	56.5%	60.5%	24.9%	-31.0%	27.9%	6.3%

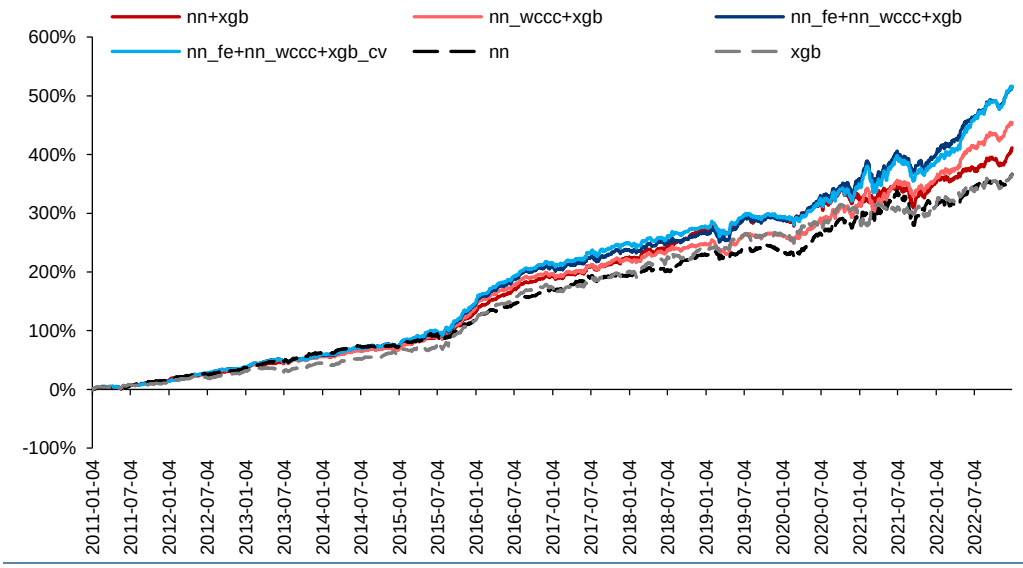
资料来源：朝阳永续，Wind，华泰研究

图表23：模型集成测试模型回测绩效（回测期 2011-01-04 至 2022-12-30，基准为中证 500 指数）

	年化收益率	年化波动率	夏普比率	最大回撤	Calmar 比率	年化超额收益率	年化跟踪误差	信息比率	超额收益最大回撤	超额收益 Calmar 比率	相对基准月胜率	年化双边换手率
基线												
nn	15.94%	25.69%	0.62	50.25%	0.32	14.24%	5.99%	2.38	13.36%	1.07	77.08%	16.18
xgb	15.82%	26.07%	0.61	46.94%	0.34	14.22%	6.28%	2.26	9.70%	1.47	68.75%	16.26
模型集成												
nn+xgb	16.78%	25.95%	0.65	48.29%	0.35	15.15%	5.99%	2.53	9.36%	1.62	79.17%	16.20
nn_wccc+xgb	17.57%	26.06%	0.67	47.97%	0.37	15.95%	6.10%	2.62	9.20%	1.73	74.31%	16.30
nn_fe+nn_wccc+xgb	18.56%	26.33%	0.70	48.96%	0.38	17.00%	6.24%	2.73	9.32%	1.82	76.39%	16.31
nn_fe+nn_wccc+xgb_cv	18.57%	26.38%	0.70	49.46%	0.38	17.03%	6.36%	2.68	9.54%	1.79	73.61%	16.31

资料来源：朝阳永续，Wind，华泰研究

图表24：模型集成测试模型超额收益表现（回测期 2011-01-04 至 2022-12-30，基准为中证 500 指数）



资料来源：朝阳永续，Wind，华泰研究

讨论

均值因子在神经网络和 XGBoost 间的差异，兼谈如何使用弱因子

特征工程引入均值因子提升神经网络表现，但削弱 XGBoost 表现。要弄清此中原因，相当于用线性的人脑理解非线性模型的工作机理，难度颇大，我们尝试从下列角度分析。

首先，我们提出一个假设：**均值因子属于弱因子，有用，但比重不宜过大**。前文提到，均值因子是对市场环境的刻画，有一定信息量；高分队伍的实践也表明该因子有效。但我们同时观察到，比赛中第 1、2 名未使用全部 300 个匿名特征构建均值因子，而是筛选 IC 前 100 个特征构建，均值因子从数量上比重不高，模型对均值因子的使用是有限度的。并且从理论上分析，选股模型应侧重于个股信息的挖掘，市场环境信息只起到辅助作用。

其次，考察 XGBoost 模型的特征重要性，计算各期特征重要性均值，部分结果如下表。在引入均值因子的 xgb_fe 模型中，重要性最高的特征为 bp_lf_mean，即 bp_lf 的均值因子。重要性排名前 10 的特征中，均值因子占据 4 位。**XGBoost 模型全部均值因子重要性之和占比 44%**，接近一半水平，比重较高。

图表25：xgb_fe 模型各期特征重要性均值前 10 的特征

排序	因子	xgb	xgb_fe
1	bp_lf_mean	-	4.38%
2	corr_vol_high	3.75%	4.12%
3	bp_lf	5.11%	3.50%
4	ep_ttm	6.28%	3.45%
5	turn_1m	4.14%	3.18%
6	std_1m_mean	-	3.11%
7	exp_wgt_return_3m	5.15%	2.95%
8	exp_wgt_return_3m_mean	-	2.83%
9	vwap2close	2.90%	2.81%
10	con_roe_g_mean	-	2.74%

资料来源：朝阳永续，Wind，华泰研究

再次，测试均值因子缩放系数的影响。预处理环节，我们对均值因子整体乘以 0.01，突出原始因子作用，弱化均值因子影响。需要说明的是，从两类模型的原理看，特征的相对量级对神经网络有效，对 XGBoost 影响不大。我们进一步测试均值因子缩放系数为 1e-4 和 1 时的表现以验证上述猜想。

结果如下表所示，当缩放系数为 1，即不对均值因子做缩放处理时，神经网络和 XGBoost 均表现较差。随着缩放系数的降低，神经网络有显著提升，在系数为 0.01 时已能战胜原始模型；XGBoost 变化不大，都不能战胜原始模型。由此可见，**神经网络主动降低均值因子值有一定效果**，但对 XGBoost 不起作用。

图表26：均值因子缩放系数测试模型合成因子评价指标（回测期 2011-01-04 至 2022-12-30）

	IC 均值	RankIC 均值	加权 IC 均值	加权 RankIC 均值	加权 ICIR	RankICIR	加权 ICIR	加权 RankICIR	Top 组精确率	Bottom 组精确率	Top 组年化收益率	Bottom 组年化收益率	多空对冲年化收益率	基准收益率
nn	8.5%	10.6%	7.1%	9.6%	0.83	0.99	0.70	0.90	56.3%	59.8%	24.7%	-28.1%	26.4%	6.3%
nn_fe_scale=1e-4	8.6%	10.7%	7.1%	9.7%	0.83	0.99	0.70	0.91	56.3%	59.9%	25.6%	-28.4%	27.0%	6.3%
nn_fe_scale=0.01	8.5%	10.7%	7.1%	9.7%	0.83	0.99	0.70	0.91	56.3%	59.9%	25.5%	-28.3%	26.9%	6.3%
nn_fe_scale=1	6.2%	7.8%	5.1%	7.1%	0.65	0.78	0.54	0.71	53.9%	57.9%	17.8%	-20.3%	19.0%	6.3%
xgb	8.7%	11.3%	6.5%	9.7%	0.75	0.95	0.57	0.82	56.5%	60.2%	23.3%	-30.1%	26.7%	6.3%
xgb_fe_scale=1e-4	7.7%	10.3%	5.6%	8.7%	0.62	0.80	0.45	0.68	55.7%	59.6%	20.5%	-27.2%	23.9%	6.3%
xgb_fe_scale=0.01	6.8%	8.8%	5.1%	7.6%	0.58	0.72	0.44	0.62	54.8%	57.8%	20.6%	-21.9%	21.3%	6.3%
xgb_fe_scale=1	7.6%	10.2%	5.6%	8.6%	0.63	0.80	0.46	0.68	55.5%	59.5%	20.5%	-26.8%	23.6%	6.3%

资料来源：朝阳永续，Wind，华泰研究

图表27：均值因子缩放系数测试模型回测绩效（回测期 2011-01-04 至 2022-12-30，基准为中证 500 指数）

	年化收益率	年化波动率	夏普比率	最大回撤	Calmar比率	年化超额收益率	年化跟踪误差	信息比率	超额收益最大回撤	超额收益Calmar比率	相对基准月胜率	年化双边换手率
nn	15.94%	25.69%	0.62	50.25%	0.32	14.24%	5.99%	2.38	13.36%	1.07	77.08%	16.18
nn_fe_scale=1e-4	17.33%	26.36%	0.66	48.40%	0.36	15.78%	6.47%	2.44	14.66%	1.08	76.39%	16.21
nn_fe_scale=0.01	17.41%	26.25%	0.66	48.20%	0.36	15.83%	6.49%	2.44	15.25%	1.04	77.78%	16.18
nn_fe_scale=1	8.20%	25.35%	0.32	60.30%	0.14	6.53%	5.87%	1.11	18.27%	0.36	66.67%	16.41
xgb	15.82%	26.07%	0.61	46.94%	0.34	14.22%	6.28%	2.26	9.70%	1.47	68.75%	16.26
xgb_fe_scale=1e-4	12.02%	26.11%	0.46	48.35%	0.25	10.45%	6.76%	1.55	14.15%	0.74	61.81%	16.19
xgb_fe_scale=0.01	12.79%	26.68%	0.48	49.54%	0.26	11.37%	6.80%	1.67	12.37%	0.92	68.06%	16.26
xgb_fe_scale=1	12.40%	25.72%	0.48	46.53%	0.27	10.72%	6.62%	1.62	10.42%	1.03	64.58%	16.18

资料来源：朝阳永续，Wind，华泰研究

最后，从理论角度分析两类模型训练过程。XGBoost 对特征进行随机采样，在采样的候选特征中寻找最优划分方式，并非从全部特征中搜索。**XGBoost 引入均值因子这类弱因子后，原始特征被采样到的概率下降，可能被排除在候选特征外，导致模型预测效果下降。神经网络不涉及特征采样操作，因此可以通过缩小取值的方式，在合理限度内使用弱因子。**

MSE 和 IC 损失函数的差异，兼谈因子合成和组合优化的目标错配问题

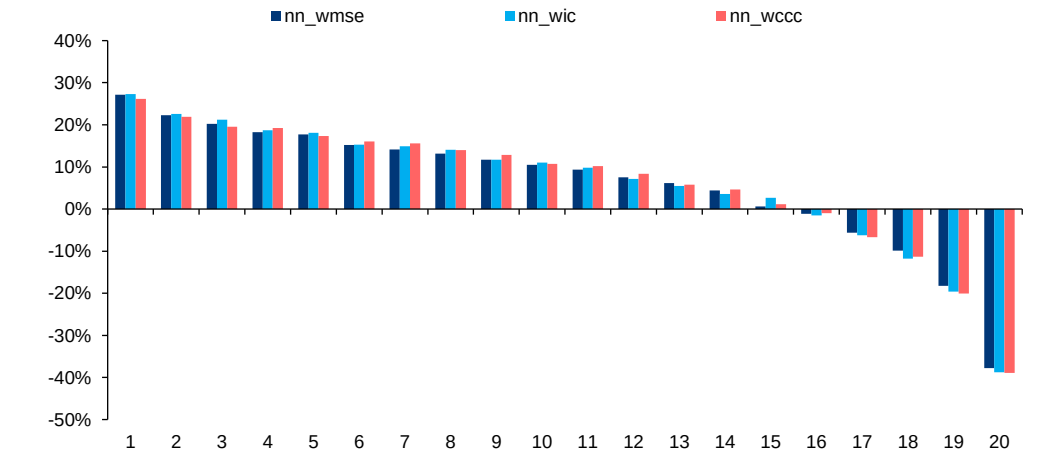
本文另一个“反直观”的结论是 IC 和加权 IC 作为损失函数，单因子测试表现较好，但指增组合表现较差，弱于 MSE 和 CCC 损失。这也是因子投资长期存在的痛点，难度同样颇大，我们尝试做如下分析。

一个不会出错的回答是：单因子测试和策略回测有差异，其背后是因子合成和组合优化两步的目标错配。该问题早已为研究者关注，常用的样本多头加权，正是针对指数增强多头组合，在因子合成这步进行的修正（尽管未必是最佳解决方法），使因子合成的目标尽可能向组合优化的现实场景靠拢。

然而这并不能解答本文遇到的问题，即使是加权 IC 损失，在分 10 层多头收益高于加权 MSE 和加权 CCC 的情况下（25.0%，高于 24.7%和 24.2%），指增组合超额收益仍然大幅落后（8.3%，低于 14.2%和 15.4%），这又如何解释？

容易想到的解释是分 10 层还不够细。随着 A 股市场的扩容，指增选股数量在股票池的占比进一步减小，选股更加向头部集中。目前行业里常见的做法是分 20 层测试，结果如下图，加权 IC 多头端收益仍然优于加权 MSE 和加权 CCC。看来问题不在于评价指标是分 10 层还是 20 层。

图表28：部分测试模型合成因子分 20 层回测年化收益率

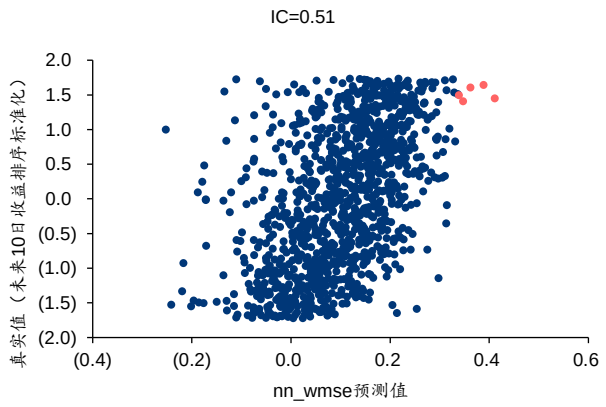


资料来源：朝阳永续，Wind，华泰研究

可否考察加权 IC 超额收益低于加权 MSE 的交易日，从个案出发寻找线索？我们统计加权 MSE 相比加权 IC 近 10 日超额收益，差距最大的交易日为 2015 年 1 月 19 日。由于预测区间为未来 10 个交易日，前推 10 日为 2015 年 1 月 5 日。我们对比该截面日下，各模型的预测值和真实值，如下图所示。

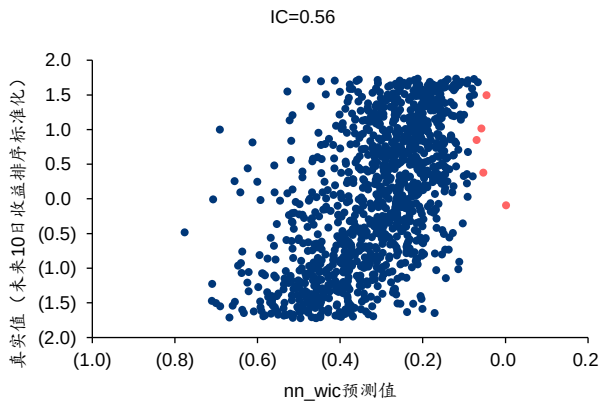
左图加权 MSE 模型的 IC 值为 0.51，低于右图加权 IC 模型的 IC 值 0.56。但观察预测值（横轴）排名靠前的样本（红点为前 5 名），这些点在左图对应的纵坐标也较大，表明真实收益高，但在右图的纵坐标不算高。这些预测值靠前的股票有大概率会被选中，且分配较高权重。该截面日股票池有 1127 只有效样本，即使分 20 层测试，由于分层组合收益通常为等权重计算，这些点也会被淹没在 Top 组的 50 多只股票中，但恰恰是这些样本很大程度上左右了策略的收益。

图表29： nn_wmse 模型预测值和真实值对比（2015-01-05 截面日）



资料来源：朝阳永续，Wind，华泰研究

图表30： nn_wic 模型预测值和真实值对比（2015-01-05 截面日）



资料来源：朝阳永续，Wind，华泰研究

至此我们得到下列关键结论：

1. 评价指标的角度：传统的单因子测试评价指标可能不适用于目前多头选股场景。IC 值、t 统计量等反映预测值的全局性，但对多头侧重不够，即使是加权 IC、分 20 层回测多头收益，也会和最终选股组合表现脱节。
2. 损失函数的角度：以 IC 为损失函数，可以让评价指标变得很好看。但 IC 作为全局性的统计量，不会侧重于个别头部样本，但这些少数样本可能对组合优化影响很大。而 MSE 的特点之一是给予极端误差较大惩罚，恰好可以弥补 IC 的弱点。CCC 结合 IC 和 MSE 的特点，同时学习数据中的共性和个性，在本文中表现好也就不难理解。
3. 因子合成和组合优化的目标错配：理想的解决方案是将两步放在相同的网络中进行优化，实现真正的端到端训练。但当前技术尚未成熟，我们也暂没有特别好的思路。相对现实的方案是设计因子合成的损失函数和评价指标，尽可能向组合优化场景靠拢。例如根据多头选股数量，确定损失函数和评价指标的多头倾斜程度；融合多种损失函数，或采用多目标训练等。

总结

本文梳理 2022 年九坤 Kaggle 量化大赛高分队伍解决方案，提炼出特征工程、损失函数、交叉验证、模型集成四个主要方向，并应用于华泰人工智能中证 500 指数增强策略改进。结果表明：(1)特征工程引入均值因子对神经网络有效；(2)CCC 损失优于 MSE 损失和 IC 损失；(3)时序交叉验证作用不明显；(4) 集成神经网络和决策树类模型提升较稳定。对比整合多项改进的模型与基线模型，回测期 2011 年至 2022 年内，年化超额收益从 14.2%提升至 17.0%，信息比率从 2.3/2.4 提升至 2.7。

随着数据科学在线社区日益成熟，越来越多的爱好者投身于网络编程竞赛之中。Kaggle 是全球知名的数据科学在线平台之一，Two Sigma、Optiver 等头部量化机构曾在 Kaggle 发布挑战竞赛。国内量化私募九坤投资于 2022 年 1 月启动 Kaggle 竞赛，吸引两千多只队伍参赛。比赛具体任务为基于给定的 A 股匿名特征，预测股票未来短期收益，最终评价指标为预测收益和真实收益的 IC 值，属于典型的监督学习问题，和实际量化选股场景较贴近。

我们梳理九坤 Kaggle 量化大赛高分队伍解决方案，提炼出四个改进方向。(1)特征工程引入截面上全部股票因子的均值，均值因子可能反映原始因子整体分布的时变特性，是市场环境的一种简单表达。(2)损失函数引入一致性相关系数 CCC，可视作 IC 和 MSE 的融合，兼顾相关性和距离。(3)采用时序交叉验证选取最优超参数。(4)集成不同类型机器学习模型。以神经网络和 XGBoost 构建中证 500 指数增强策略作为基线，测试上述技巧的改进效果。

四项改进技巧效果各异。特征工程引入的均值因子对神经网络有提升，但削弱了 XGBoost。损失函数中，MSE 表现不突出；IC 损失单因子测试表现好，但指增组合回测表现差；CCC 损失在单因子测试表现一般，但指增组合回测表现较好；加权均优于等权。交叉验证调参改进不显著，考虑到时间开销大，性价比不高，算力有限前提下，使用经验超参数即可。模型集成提升较稳定，神经网络类和决策树类模型有互补效果。

研究发现均值因子对神经网络有效但对 XGBoost 无效。均值因子属于弱因子，有用但比重不宜过大。XGBoost 引入弱因子后，特征采样使原始因子可能被排除在外，从而削弱模型。神经网络可通过预处理缩小取值，有限度地使用弱因子。研究还发现 IC 损失单因子测试优于 MSE 损失，但指增组合表现差，本质是因子合成和组合优化的目标错配。IC 属于全局统计量，不会侧重于个别头部样本，但这些样本可能对组合优化影响较大。MSE 的特点之一是给予极端误差较大惩罚，恰好弥补 IC 弱点。CCC 融合 IC 和 MSE，兼顾共性和个性，是一类理想的损失函数。

参考文献

Lin, I. K. . (1989). A concordance correlation-coefficient to evaluate reproducibility. Biometrics, 45(1), 255-268.

Pandit, V. , & Schuller, B. . (2019). The many-to-many mapping between the concordance correlation coefficient and the mean square error.

风险提示

人工智能挖掘市场规律是对历史的总结，市场规律在未来可能失效。人工智能技术存在过拟合风险。深度学习模型受随机数影响较大，本文未进行随机数敏感性测试。本文测试的选股模型调仓频率较高，假定以 vwap 价格成交，忽略其他交易层面因素影响。

免责声明

分析师声明

本人，林晓明、李子钰、何康，兹证明本报告所表达的观点准确地反映了分析师对标的证券或发行人的个人意见；彼以往、现在或未来并无就其研究报告所提供的具体建议或所表达的意见直接或间接收取任何报酬。

一般声明及披露

本报告由华泰证券股份有限公司（已具备中国证监会批准的证券投资咨询业务资格，以下简称“本公司”）制作。本报告所载资料是仅供接收人的严格保密资料。本报告仅供本公司及其客户和其关联机构使用。本公司不因接收人收到本报告而视其为客户。

本报告基于本公司认为可靠的、已公开的信息编制，但本公司及其关联机构(以下统称为“华泰”)对该等信息的准确性及完整性不作任何保证。

本报告所载的意见、评估及预测仅反映报告发布当日的观点和判断。在不同时期，华泰可能会发出与本报告所载意见、评估及预测不一致的研究报告。同时，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动。以往表现并不能指引未来，未来回报并不能得到保证，并存在损失本金的可能。华泰不保证本报告所含信息保持在最新状态。华泰对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

本公司不是 FINRA 的注册会员，其研究分析师亦没有注册为 FINRA 的研究分析师/不具有 FINRA 分析师的注册资格。

华泰力求报告内容客观、公正，但本报告所载的观点、结论和建议仅供参考，不构成购买或出售所述证券的要约或招揽。该等观点、建议并未考虑到个别投资者的具体投资目的、财务状况以及特定需求，在任何时候均不构成对客户私人投资建议。投资者应当充分考虑自身特定状况，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。对依据或者使用本报告所造成的一切后果，华泰及作者均不承担任何法律责任。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

除非另行说明，本报告中所引用的关于业绩的数据代表过往表现，过往的业绩表现不应作为日后回报的预示。华泰不承诺也不保证任何预示的回报会得以实现，分析中所做的预测可能是基于相应的假设，任何假设的变化可能会显著影响所预测的回报。

华泰及作者在自身所知情的范围内，与本报告所指的证券或投资标的不存在法律禁止的利害关系。在法律许可的情况下，华泰可能会持有报告中提到的公司所发行的证券头寸并进行交易，为该公司提供投资银行、财务顾问或者金融产品等相关服务或向该公司招揽业务。

华泰的销售人员、交易人员或其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。华泰没有将此意见及建议向报告所有接收者进行更新的义务。华泰的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。投资者应当考虑到华泰及/或其相关人员可能存在影响本报告观点客观性的潜在利益冲突。投资者请勿将本报告视为投资或其他决定的唯一信赖依据。有关该方面的具体披露请参照本报告尾部。

本报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布的机构或人员，也并非意图发送、发布给因可得到、使用本报告的行为而使华泰违反或受制于当地法律或监管规则的机构或人员。

中国香港

本报告由华泰证券股份有限公司制作,在香港由华泰金融控股（香港）有限公司向符合《证券及期货条例》及其附属法律规定的机构投资者和专业投资者的客户进行分发。华泰金融控股（香港）有限公司受香港证券及期货事务监察委员会监管，是华泰国际金融控股有限公司的全资子公司，后者为华泰证券股份有限公司的全资子公司。在香港获得本报告的人员若有任何有关本报告的问题,请与华泰金融控股（香港）有限公司联系。

香港-重要监管披露

- 华泰金融控股（香港）有限公司的雇员或其关联人士没有担任本报告中提及的公司或发行人的高级人员。
- 有关重要的披露信息，请参华泰金融控股（香港）有限公司的网页 https://www.htsc.com.hk/stock_disclosure

其他信息请参见下方 **“美国-重要监管披露”**。

美国

在美国本报告由华泰证券（美国）有限公司向符合美国监管规定的机构投资者进行发表与分发。华泰证券（美国）有限公司是美国注册经纪商和美国金融业监管局（FINRA）的注册会员。对于其在美国分发的研究报告，华泰证券（美国）有限公司根据《1934 年证券交易法》（修订版）第 15a-6 条规定以及美国证券交易委员会人员解释，对本研究报告内容负责。华泰证券（美国）有限公司联营公司的分析师不具有美国金融监管（FINRA）分析师的注册资格，可能不属于华泰证券（美国）有限公司的关联人员，因此可能不受 FINRA 关于分析师与标的公司沟通、公开露面和所持交易证券的限制。华泰证券（美国）有限公司是华泰国际金融控股有限公司的全资子公司，后者为华泰证券股份有限公司的全资子公司。任何直接从华泰证券（美国）有限公司收到此报告并希望就本报告所述任何证券进行交易的人士，应通过华泰证券（美国）有限公司进行交易。

美国-重要监管披露

- 分析师林晓明、李子钰、何康本人及相关人士并不担任本报告所提及的标的证券或发行人的高级人员、董事或顾问。分析师及相关人士与本报告所提及的标的证券或发行人并无任何相关财务利益。本披露中所提及的“相关人士”包括 FINRA 定义下分析师的家庭成员。分析师根据华泰证券的整体收入和盈利能力获得薪酬，包括源自公司投资银行业务的收入。
- 华泰证券股份有限公司、其子公司和/或其联营公司，及/或不时会以自身或代理形式向客户出售及购买华泰证券研究所覆盖公司的证券/衍生工具，包括股票及债券（包括衍生品）华泰证券研究所覆盖公司的证券/衍生工具，包括股票及债券（包括衍生品）。
- 华泰证券股份有限公司、其子公司和/或其联营公司，及/或其高级管理层、董事和雇员可能会持有本报告中所提到的任何证券（或任何相关投资）头寸，并可能不时进行增持或减持该证券（或投资）。因此，投资者应该意识到可能存在利益冲突。

评级说明

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力（含此期间的股息回报）相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数），具体如下：

行业评级

增持：预计行业股票指数超越基准
中性：预计行业股票指数基本与基准持平
减持：预计行业股票指数明显弱于基准

公司评级

买入：预计股价超越基准 15%以上
增持：预计股价超越基准 5%~15%
持有：预计股价相对基准波动在-15%~5%之间
卖出：预计股价弱于基准 15%以上
暂停评级：已暂停评级、目标价及预测，以遵守适用法规及/或公司政策
无评级：股票不在常规研究覆盖范围内。投资者不应期待华泰提供该等证券及/或公司相关的持续或补充信息

法律实体披露

中国：华泰证券股份有限公司具有中国证监会核准的“证券投资咨询”业务资格，经营许可证编号为：91320000704041011J

香港：华泰金融控股（香港）有限公司具有香港证监会核准的“就证券提供意见”业务资格，经营许可证编号为：AOK809

美国：华泰证券（美国）有限公司为美国金融业监管局（FINRA）成员，具有在美国开展经纪交易商业务的资格，经营业务许可编号为：CRD#:298809/SEC#:8-70231

华泰证券股份有限公司

南京
南京市建邺区江东中路 228 号华泰证券广场 1 号楼/邮政编码：210019

电话：86 25 83389999/传真：86 25 83387521
电子邮件：ht-rd@htsc.com

北京
北京市西城区太平桥大街丰盛胡同 28 号太平洋保险大厦 A 座 18 层/邮政编码：100032
电话：86 10 63211166/传真：86 10 63211275
电子邮件：ht-rd@htsc.com

深圳
深圳市福田区益田路 5999 号基金大厦 10 楼/邮政编码：518017
电话：86 755 82493932/传真：86 755 82492062
电子邮件：ht-rd@htsc.com

上海
上海市浦东新区东方路 18 号保利广场 E 栋 23 楼/邮政编码：200120
电话：86 21 28972098/传真：86 21 28972068
电子邮件：ht-rd@htsc.com

华泰金融控股（香港）有限公司
香港中环皇后大道中 99 号中环中心 58 楼 5808-12 室
电话：+852-3658-6000/传真：+852-2169-0770
电子邮件：research@htsc.com
http://www.htsc.com.hk

华泰证券（美国）有限公司
美国纽约公园大道 280 号 21 楼东（纽约 10017）
电话：+212-763-8160/传真：+917-725-9702
电子邮件: Huatai@htsc-us.com
http://www.htsc-us.com