基于多模型横截面 IC 的 ETF 收益预测研究：传统计量、机器学习与深度学习的比较分析

一、摘要

本研究选取了 2018-2024 年的 16 支美股主要 ETFs/ETNs 为样本，围绕横截面收益预测问题，系统评估了向量自回归模型（VAR）、多层感知机（MLP）、直方梯度提升机（HGBM）、卷积神经网络（CNN）、门控循环单元（GRU）、长短期记忆网络（LSTM）、Transformer 模型及其 Stacking 集成学习模型在横截面 IC 指标下的选股能力。通过构造收益率、振幅、尾盘强度、对数成交量四种特征变量，构造静态截面、时间序列两种滑动窗口，z-score 标准化完成特征工程构建。本研究以横截面 IC 为核心评价指标，结合 IC 均值、累计 IC、IC IR 与正向 IC 比例，探究各个模型的有效性和稳定性。实证结果显示，Transformer、LSTM 和 Stacking 模型在样本外具有显著的预测优势，累计 IC 分别超过 20、10 与 15，IR 均大于 0.1，显著优于传统计量线性模型与传统非线性机器学习方法。研究验证了深度学习与与集成模型在多资产选股中的建模优势，同时指出当前模型在预测稳定性（IR）方面仍有优化空间。

二、样本和数据来源

本研究选取了 2018 年 6 月 19 日[1]至 2024 年 12 月 31 日美股市场的 16 支 ETFs/ETNs 为研究样本，数据来源为开源金融数据库 Akshare，数据为日度频率，清洗空值后共有 1581 条有效数据。

研究所用 ETFs/ETNs、原始交易数据字段如下表所示。

类别	ETFs/ETNs	代表指数/行业	缺失值数量	样本数量
指数型 ETF	SPY	S&P 500	0	1645
	QQQ	NASDAQ-100	0	1645
	DIA	道琼斯工业指数	0	1645
	IWM	Russell 2000 小盘股指数	0	1645
行业 ETF	XLB	原材料	0	1645
	XLC	通信服务	0	1645
	XLE	能源	0	1645
	XLF	金融	0	1645
	XLI	工业	0	1645

[1] 16 支 ETF 中，XLC 最晚开始交易，首次交易日期为 2018 年 6 月 19 日，故选取这天为样本起始时间。

	XLK	信息技术	0	1645
	XLP	必需消费	0	1645
	XLRE	房地产	0	1645
	XLU	公共事业	0	1645
	XLV	医疗保健	0	1645
	XLY	可选消费	0	1645
ETN	VXX	VIX 期货，反映市场情绪	64	1581

表 1 研究所用 ETFs/ETNs

字段名	含义
DATE	交易日期
OPEN	开盘价
HIGH	最高价
LOW	最低价
CLOSE	收盘价
VOLUME	成交量

表 2 原始交易数据字段说明

三、特征工程构建

为了获取 ETF 交易数据中的市场信息，本研究构建了四个量化市场常用的特征变量，通过滑动窗口构造模型输入数据，并进行标准化，按时间顺序划分训练和测试集。

（一）特征变量设计

针对每一支 ETF 的日度交易数据，本研究构建了收益率、振幅、尾盘强度、对数成交量四个特征变量，并通过日期合并、构造特征矩阵，用以捕捉市场信息。本文使用的是基于收盘价（close to close）的收益率。具体构建过程如下表所示。

特征名称	计算方法	含义
收益率	$\mathrm{RETURN}_t = \frac{\mathrm{CLOSE}_t - \mathrm{CLOSE}_{t-1}}{\mathrm{CLOSE}_{t-1}}$	当日价格变动
振幅	$\mathrm{AMPLITUDE}_t = \frac{\mathrm{HIGH}_t - \mathrm{LOW}_t}{\mathrm{OPEN}_t}$	当日价格波动范围
尾盘强度	$\mathrm{TAIL\_STRENGTH}_t = \frac{\mathrm{CLOSE}_t - \mathrm{OPEN}_t}{\mathrm{OPEN}_t}$	当日收盘相对涨跌趋势
对数成交量	$\mathrm{LOG\_VOLUME}_t = \mathrm{LOG}(1 + \mathrm{VOLUME}_t)$	缩放成交量的尺度

表 3 特征变量设计过程

（二）滑动窗口

由于股票市场上的自变量具有时间序列结构，所以我们构建了定长为 7 天的滑动窗口输入样本，即用第 t-7 到 t-1 天的特征变量预测第 t 天的 ETF 收益率，使模型能够学习市场变量之间的时间依赖关系。

另外，为了适配不同机器学习模型对输入样本的要求，我们构建了以下两种输入形式：

1.静态截面结构。把过去 7 天的 16 种 ETF 的 4 个特征按时间顺序拼接、展平为一维向量，构成（1581 * 448）的二维输入矩阵。适用于不考虑时间序列性的传统机器学习模型。

2.时间序列结构。保留 7 天的原始时间顺序与特征维度，组成二维张量输入，构成（1581 * 7 * 64）的三维输入矩阵。适用于考虑时间序列性的深度学习模型。

（三）标准化

为了统一量纲，本研究对输入数据采用了 z-score 标准化，使特征变量具有零均值和单位标准差。

（四）训练测试集划分

为了避免未来信息泄露，本研究统一按时间顺序划分训练和测试集，其中，训练集取时间序列的前 80%，到 2023 年 9 月 28 日为止；测试集取后 20%，自 2023 年 9 月 29 日开始。

四、模型选取与调参

（一）模型选取

本研究选取了 Var、MLP、HGBM、CNN、GRU、LSTM、Transformer、Stacking 集成学习共 8 种模型进行对比分析，涵盖了传统计量经济模型、非线性机器学习模型、深度学习模型和集成学习模型。

其中，Stacking 集成学习的基模型选取了样本外预测表现较好、且具有异质性的 MLP、LSTM、Transformer 三个模型，元模型选取了 Ridge 岭回归。另外，本研究尝试使用 LightGBM 进行预测，但在多个参数条件下，都因为有效增益不足导致模型空转迭代、训练停滞，导致树结构未能有效展开，故选取了同样基于梯度提升树的 HGBM 替代。

具体模型信息如下表所示。

模型类别	模型	英文简称	输入数据	多变量建模	捕捉因变量关系
计量经济	向量自回归模型	Var	仅基于收益率的时间序列	支持	能
非线性	多层感知机回归	MLP	静态截面	需要包装器	否
	直方梯度提升机	HGBM	静态截面	需要包装器	否
深度学习	卷积神经网络	CNN	时间序列	支持	否
	门控循环单元神经网络	GRU	时间序列	支持	能
	长短期记忆神经网络	LSTM	时间序列	支持	能
	Transformer	Transformer	时间序列	支持	能
集成学习	堆叠	Stacking	多模型输出结果	支持	能

表 4 模型信息

（二）模型参数

本研究对 MLP、HGBM 采用随机搜索调参，对 CNN、GRU、LSTM、Transformer 采用手动设定训练轮次、多次输出对比 IC 值调参。

另外，Transformer 还采用了因果注意力机制、Positional Encoding、AdamW 优化器等方法提升预测能力。

值得关注的是，由于使用 MultiOutputRegressor 包装的多变量随机搜索原生不支持以横截面 IC 值作为评估指标，所以 MLP、HGBM 的调参依旧以传统的决定系数R^2为评估指标，与默认参数对比，发现 MLP 表现不如默认参数，HGBM 反之，故应用相应的最佳参数组合进行预测。

调参过程如下表所示。

模型	参数名称	参数池	最优参数
VAR	lags	固定为 1	1
MLP	hidden_layer_sizes	[(64,), (128,), (64, 64), (128, 64)]	(128,)
	max_iter	[200, 300, 500]	500
HGBM	learning_rate	[0.01, 0.05, 0.1]	0.05

	max_iter	[100, 200, 300]	200
	max_leaf_nodes	[15, 31, 63]	31
	min_samples_leaf	[10, 20, 30]	20
	l2_regularization	[0.0, 0.1, 1.0]	0.1
	max_depth	[3, 5, None]	5
CNN	epochs	[20, 30, 50, 100]（手动调参）	50
GRU	pochs	[20, 30, 50, 100]（手动调参）	30
LSTM	epochs	[20, 30, 50, 100]（手动调参）	30
TRANSFORMER	lstm_epochs	[20, 30, 50, 100]（手动调参）	100
STACKING	alpha (Ridge)	固定为 1	1

表 5 模型参数

五、模型结果与实证分析

模型训练后得到的样本内外 IC、IR 统计量如下表所示；样本内外的累计 IC 图、每日 IC 值散点图如下表所示（随压缩包发送的表格有完整、清晰的 IC 图）。

按照量化行业的通用标准，IC 大于 0.02 表示该模型的选股能力有效、大于 0.05 则表示该模型的选股能力出色；IR 大于 0.3 该模型具有稳定性、可商用。

从样本外表现来看，Transformer 的平均 IC 大于 0.06、累计 IC 大于 20，LSTM 和 Stacking 集成学习模型的平均 IC 大于 0.02、累计 IC 分别大于 10、15，这三个模型 IR 也相对略大，都是具有泛化的选股能力的模型。

总体来看，Transformer、LSTM 都是具备时间序列建模能力、能够捕捉多个因变量之间的相关关系的深度学习模型，Stakcing 则集合了 Transformer、LSTM 和 MLP 的特点。由此可见具备时序建模能力的深度学习模型在量化领域的突出竞争力。

个别来看，Transformer 表现出最优异的样本外泛化能力，累计 IC 曲线最陡、最大值超过 20，且持续上升、几乎没有大回撤，表明模型具备持续产生有效信号的能力。这不仅是因为模型自身具有多头注意力机制，可以学习不同维度的时序关系，还得益于本研究在

建模过程中集成了因果注意力机制、Positional Encoding 和 AdamW 优化器，提升了模型的跨资产、跨时间预测能力。

LSTM 的累计 IC 曲线在样本外初期呈现出一定程度的下降、震荡趋势，但中后期预测能力稳定上升，最大值达到 10，且回撤较小、波动有限、稳定性佳，这种后发式有效性说明模型具有市场结构适应能力和中期预测价值。这得益于其长期记忆能力，能够识别金融信号中的滞后效应。相似地，Stacking 的 IC 曲线也呈现出后发式有效性，但是初期震荡周期更短、降幅更小，且最大值接近 15，说明集成学习模型在应不同的市场状态时具有良好的鲁棒性。

相对而言，其他模型的表现则略显疲弱。GRU 虽然原理上与 LSTM 类似，但是 IC、IR 值均较低，且 IC 累计曲线图呈现大幅上下震荡趋势，表明在本研究中只捕捉到了浅层、短期的因子关系。CNN 作为前馈神经网络，则更适合做图像识别而不是选股预测，不具备时间记忆能力和捕捉因变量相关关系的能力，所以其 IC、IR、正分率都不如随机分配。MLP 则呈现明显的过拟合，样本内大多数点 IC 值为 1，样本外 IC、IR 值非常小，其效果接近随机分配。HGBM 虽然正分率略大于 50%，但是累计 IC 曲线上下震荡明显，说明模型偶尔命中有效因子，但是缺乏稳定提取能力，无法形成有效预测；可能是因为没有捕捉时间序列信息、因变量相关关系的能力，表明梯度提升树这一类树模型可能不适用于量化预测。传统计量模型 VAR 则完全失效，累计 IC 长期为负、持续下降，表明基于滞后资产收益的线性模型不适用于量化预测。

值得注意的是，当前模型的 IR 值都较低，仅有 Transformer、LSTM、Stacking 达到了 0.1 以上，具有一定的预测稳定性；但是总体上看，所有模型的 IR 都低于业内公认可用的水平（0.3），表明模型稳定性有待进一步提升。

MODEL	IN-SAMPLE				OUT-OF-SAMPLE			
	Mean IC	Std IC	IC IR	Positive IC Ratio	Mean IC	Std IC	IC IR	Positive IC Ratio
VAR	0.06	0.38	0.15	55.74%	-0.02	0.36	-0.06	47.78%
MLP	0.21	0.27	0.78	76.69%	0.01	0.26	0.03	52.70%
HGBM	0.31	0.27	1.16	86.32%	0.00	0.31	0.00	51.43%
CNN	0.05	0.27	0.20	57.44%	-0.02	0.28	-0.07	46.35%
GRU	0.07	0.29	0.25	59.03%	0.01	0.27	0.02	52.70%
LSTM	0.09	0.30	0.30	60.94%	0.03	0.29	0.12	52.06%
TRANSFORMER	0.07	0.38	0.19	58.79%	0.06	0.36	0.17	59.37%
STACKING	0.51	0.32	1.59	92.12%	0.04	0.39	0.10	55.56%

表 6 IC & IR 统计量

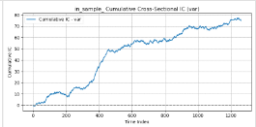
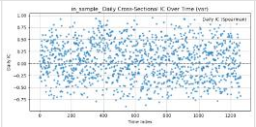
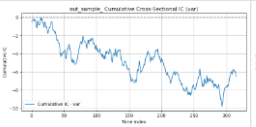
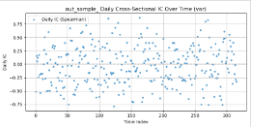
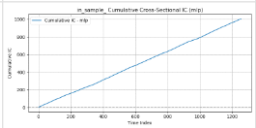
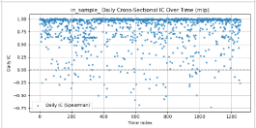
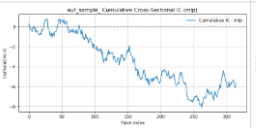
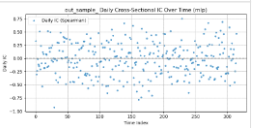
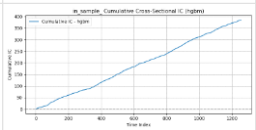
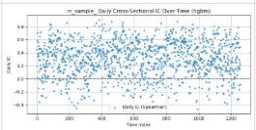
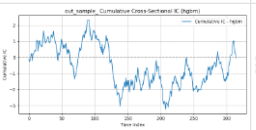
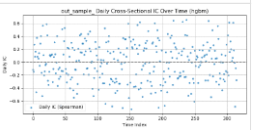
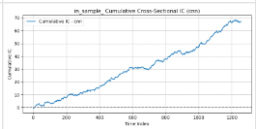
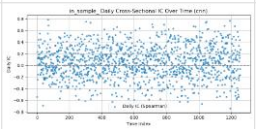
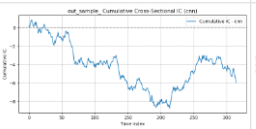
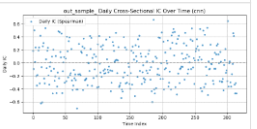
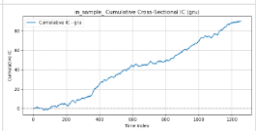
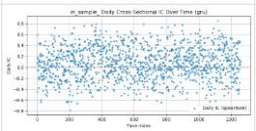
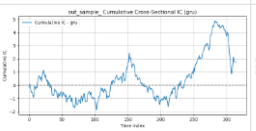
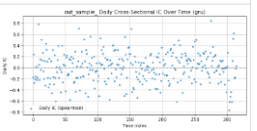
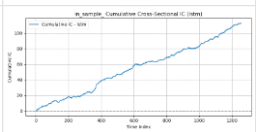
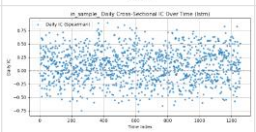
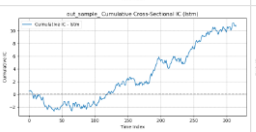
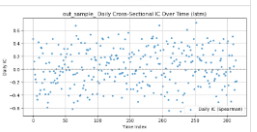
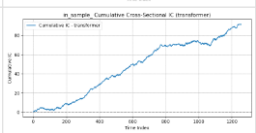
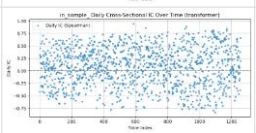
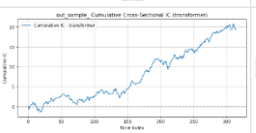
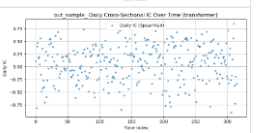
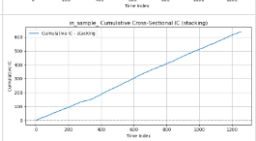
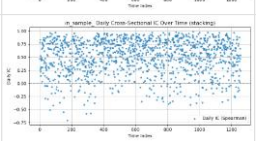
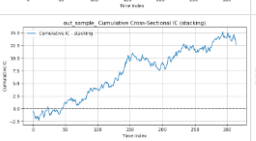
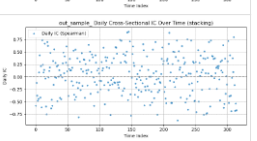
Model	In-Sample		Out-of-Sample	
	Cumulative IC	Daily IC	Cumulative IC	Daily IC
Var				
MLP				
HGBM				
CNN				
GRU				
LSTM				
Transformer				
Stacking				

表 7 累计 IC 图、IC 散点图

六、结论

1.深度学习模型具备显著优势：LSTM 与 Transformer 在捕捉横截面资产收益结构与时间依赖性方面表现最优，样本外累计 IC 和 IR 显著领先。

2.Stacking 集成学习模型提升了鲁棒性：将 LSTM、Transformer 与 MLP 集成后，Stacking 模型在不同市场状态下具备更稳定的表现，累计 IC 和正向 IC 比例较高，展示出多模型融合带来的泛化增强效果。

3.传统模型和浅层机器学习模型表现不佳：VAR、CNN、HGBM 等模型由于缺乏时间记忆或横截面建模能力，导致样本外预测失效，IC 长期为负或震荡回落。

4.IC IR 值普遍偏低：尽管部分模型具备有效的选股能力，但所有模型的样本外 IR 均未超过量化行业标准（0.3），说明预测稳定性仍需进一步提升。