基于机器学习算法的多因子选股策略

Machine Learning-Based Multi-Factor Strategy in A-Shares

陈纳川、王子卿、张涵石、吴信岩

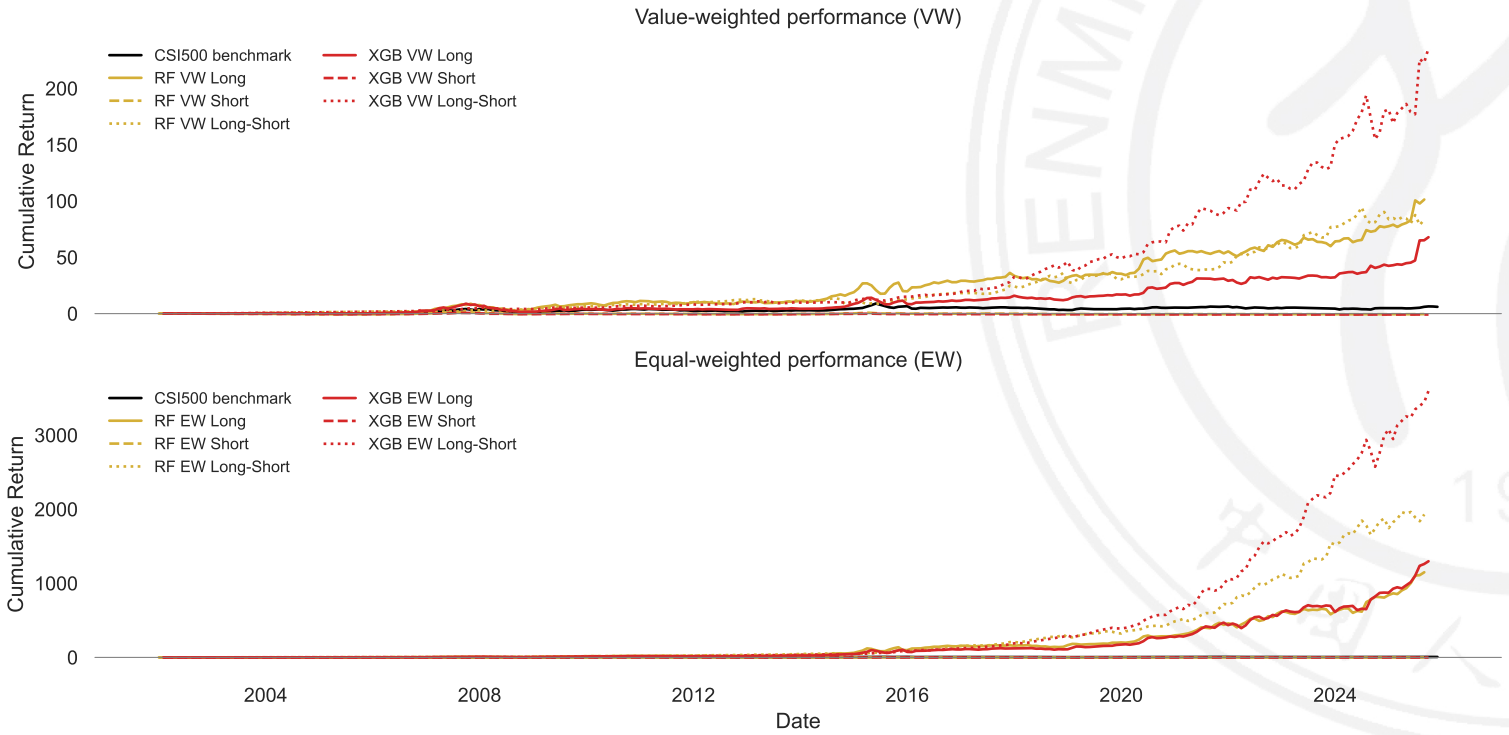
中国人民大学财政金融学院

2025 年 12 月 8 日

中国人民大學
RENMIN UNIVERSITY OF CHINA

项目概况

本项目构建了一个多因子库，并使用机器学习方法（随机森林和XGBoost）将因子作为特征进行选股。随后对以上策略使用多种指标和基准模型进行检验和收益归因，验证了该策略取得了显著的超额收益，并且稳健性较高。



目录

数据来源

- 数据来源：Tushare API
- 数据范围：2000 年 1 月至 2025 年 11 月
- 数据内容：
  - 个股行情数据（开盘价、收盘价、最高价、最低价、成交量等）
  - 指数行情数据（上证综指、深证成指、沪深 300 等）
  - 行业分类数据（申万行业分类）
  - 财务报表数据（资产负债表、利润表、现金流量表等）
  - 宏观经济数据（GDP、CPI、利率等）

数据预处理

1. 使用日频个股行情和复权因子，计算复权后的每日收益率
2. 频率对齐：将日频和季度数据转化为月频，计算相关指标
   - 按照财报发布日，填充季度数据，避免数据泄露
   - 修正了财报 YTD 数据重复计算的问题
   - 使用月度的开盘价和收盘价计算月度收益率
3. 股票池筛选：建立一个白名单，过滤一下风险过大的股票
   - ST 股、上市未满一年的次新股和停牌的股票
   - 市值分位数小于 30% 的股票（壳价值）
   - 涨停板等无法在调仓日无法买入的股票

因子库的构建和筛选

因子主要分为技术面和基本面两个大类，各分多个小类：

1. 技术面因子：
   - 技术因子
   - 动量因子
   - 波动率因子
   - 流动性因子
2. 基本面因子：
   - 质量因子
   - 价值因子
   - 成长因子
   - 规模因子

对因子的 IC、ICIR 以及分组回测指标进行筛选，最终选取了 99 个因子用于后续的模型训练。

其中包含 39 个基本面因子，60 个技术面因子

XGBoost 模型简介

XGBoost（极端梯度提升）算法是一种基于 Boosting 框架，串行训练 K 棵分类回归树 (CART) 的集成学习算法。通过在每一棵新树 f_k 都在拟合上一轮预测的残差，逐步逼近真实值。

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)$$

相比传统的树算法，XGBoost 在目标函数中引入了显式的正则化项：

$$\text{Obj} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

- l: 损失函数 (如 MSE)，衡量预测准确度。
- $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda||w||^2$: 正则项，惩罚树的复杂度（叶子节点数 T）和权重 (w)，有效防止过拟合。

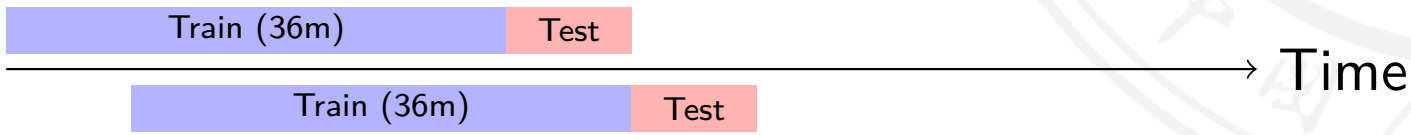
XGBoost 策略构建

模型选择动机

- 捕捉非线性：可以捕捉到因子和收益率之间的非线性关系
- 挖掘交互项：可以捕捉到不同因子之间的复杂非线性关系

训练与预测机制

- 滚动窗口
  - 训练集：过去 24 个月的数据。
  - 验证集：紧邻的下一个月的收益率 $R_{i,t+1}$
- 目标函数：最大化收益率
- 特征：经过标准化和去极值处理的因子库
- 关键超参数：1000 棵树，最大深度 6 层，学习率 0.05



Walk-Forward Validation

随机森林模型简介

随机森林是一种集成学习方法，通过构建多个决策树并结合其预测结果，提高模型的准确性和鲁棒性。与 XGBoost 的串行优化不同，RF 采用并行训练。
其特点是双重随机性

- 样本随机: 有放回地随机采样
- 特征随机: 在节点分裂时，只考虑特征子集

回归问题取所有树的平均值，分类问题取投票。其优势在于其极高的稳定性和抗过拟合能力。模型的方差由单棵树的方差 (σ^2) 和树之间的相关性 (ρ) 决定:

$$\mathrm{Var}(\mathrm{Ensemble}) = \rho\sigma^2 + \frac{1-\rho}{K}\sigma^2$$

随机森林策略构建

模型选择动机

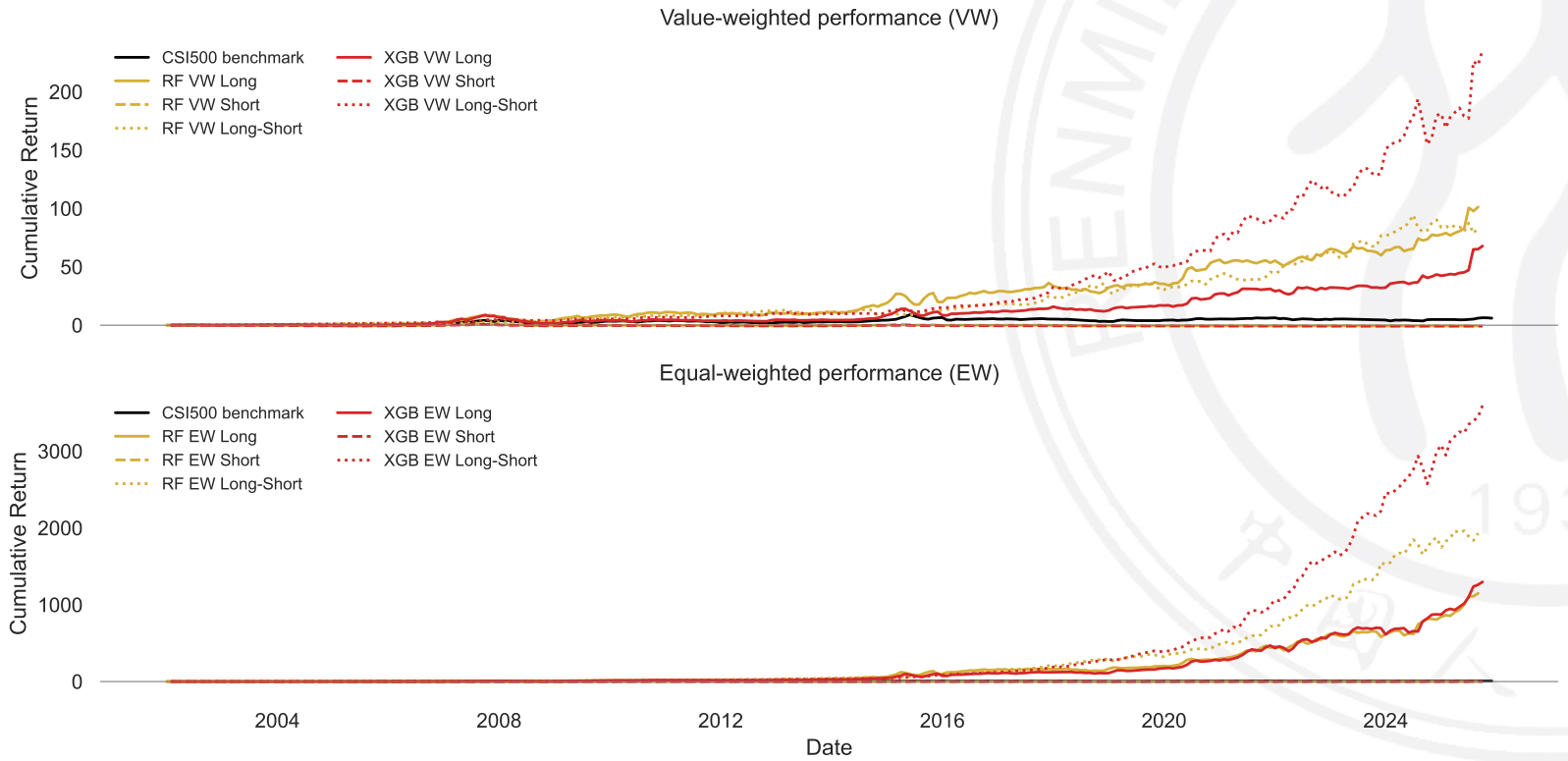
- 抗过拟合：A 股市场充斥着大量随机噪音，RF 更不容易过拟合
- 非线性与交互：能够捕捉因子间的非线性关系和交互作用
- 特征重要性：利用 OOB 数据进行特征重要性排序，可以进行因子筛选

训练与预测机制

- 滚动窗口
  - 训练集：过去 24 个月的数据。
  - 验证集：紧邻的下一个月的收益率 $R_{i,t+1}$
- 目标函数：最大化收益率
- 特征：经过标准化和去极值处理的因子库
- 关键超参数：100 棵树，最大深度 6 层，叶节点最小样本：20

策略表现：累计收益曲线

- 长期显著跑赢基准：策略在回测区间内持续产生超额收益，市值加权累计收益约 200 倍。
- 风格效应分析：等权累计收益超 3000 倍，反映 A 股小市值因子红利与复利效应叠加。



核心绩效指标

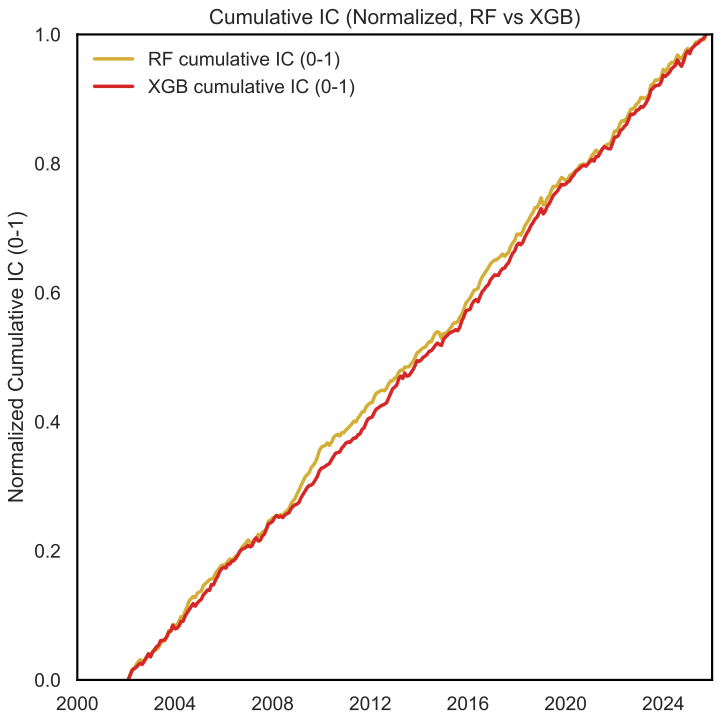
1. 核心绩效指标对比 (2000 - 2025 Long-Only)

Indicator	Random Forest (RF)		XGBoost (XGB)		Benchmark
	VW	EW	VW	EW	*CSI 500*
Ann. Return	23.16%	33.49%	22.17%	35.27%	~8.5%
Sharpe Ratio	0.8618	1.1374	0.7488	1.1317	0.35
Max Drawdown	-64.49%	-62.61%	-74.68%	-65.04%	-72.0%
Win Rate	60.70%	62.81%	60.70%	63.16%	52.1%

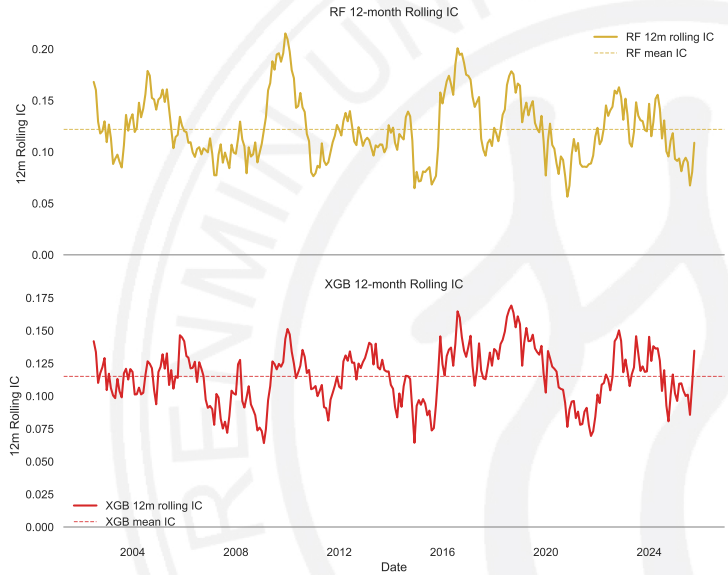
2. 模型特性分析

- 权重影响: *EW* 策略普遍优于 *VW*，表明模型在中小市值股票上的选股能力更为显著
- 模型差异: *XGBoost* 在 *EW* 中收益率更高，体现了其对非线性 *Alpha* 的挖掘潜力；而 *Random Forest* 回撤控制更优，表现出更强的抗噪性与稳健性。

主要指标汇总: 有效性分析



- 趋势: 曲线呈近乎线性的上升趋势。
- 回撤: 历史上未出现显著的长期回撤，适应多种市场风格。



- 稳定性: 滚动均值长期位于 0 轴上方，极少击穿安全线。
- 高信噪比: **ICIR > 1.0** 表现优秀。

CH3&CH4 模型分析: Random Forest (RF)

- **CH-3**:

$$R_p = \alpha + \beta_{MKT} R_{MKT} + \beta_{SMB} R_{SMB} + \beta_{VMG} R_{VMG} + \epsilon$$

- **CH-4**:

$$R_p = \alpha' + \beta_{MKT} R_{MKT} + \beta_{SMB} R_{SMB} + \beta_{VMG} R_{VMG} + \beta_{PMO} R_{PMO} + \epsilon$$

Strategy	**CH-3 Model**			**CH-4 Model**		
	Ann. Alpha	**t-stat**	R^2	**Ann. Alpha**	**t-stat**	R^2
EW (等权)	32.78%	**11.26**	0.035	33.31%	**11.23**	0.042
VW (市值加权)	19.43%	**4.29**	0.003	20.52%	**4.45**	0.009

结果解读:

- RF 策略在 CH-3 和 CH-4 模型下均获得显著的 Alpha。
- VW 策略的年化 Alpha 依然达到 **20%** 左右，证明收益并非主要来自风险因子暴露。

CH3&CH4 模型分析: XGBoost (XGB)

Strategy	CH-3 Model			CH-4 Model		
	Ann. Alpha	t-stat	R^2	Ann. Alpha	t-stat	R^2
EW (等权)	35.19%	**13.44**	0.043	36.50%	**13.81**	0.067
VW (市值加权)	23.32%	**5.36**	0.022	25.07%	**5.67**	0.034

- **Alpha** 显著性: 两个模型在引入动量因子 (CH-4) 后，Alpha 依然显著且甚至略有提升，说明策略不仅仅是捕捉了动量效应。
- **XGB** 优势: XGBoost 在 VW 组合下的 Alpha t-stat 达到 **5.67** ，表明其在大资金容量下的选股确定性更高，更有效地剥离了常见的风险因子。
- 低 R^2: 极低的 R^2 (不到 0.1) 说明策略收益与常见风格因子相关性极低，具备独特的特质收益源。

Fama-MacBeth 横截面回归分析

- 回归方程 (每月 t 进行横截面回归):

$$R_{i,t+1} = \lambda_0 + \lambda_{ML} \cdot \hat{y}_{i,t} + \lambda_{Size} \cdot \text{Size}_{i,t} + \lambda_{Bm} \cdot \text{Bm}_{i,t} + \cdots + \epsilon_{i,t}$$

- 控制变量: 市值, 账面市值比, 12 个月动量, 非流动性。

Factor	XGBoost (XGB)		Random Forest (RF)	
	Coeff (λ)	t-stat	Coeff (λ)	t-stat
Intercept	-0.0544	-8.46	-0.1145	-13.49
Model Score (pred_ret)	**0.1363**	**18.22**	**0.2559**	**21.24**

结论:

- 两个模型的预测分数 t 值均远超临界值，分别为 **18.22** 和 **21.24**，表明其具有极强的独立预测能力。
- 即使在控制了显著的小市值效应 ($\lambda_{Size} < 0$) 后，机器学习因子的显著性依然未被削弱。

GRS 联合显著性检验

- 目的：检验 5 组投资组合 (Q1 ~ Q5) 的定价误差 (α) 是否联合为零。
- 零假设 (H_0)：$\alpha_{Q1} = \alpha_{Q2} = \alpha_{Q3} = \alpha_{Q4} = \alpha_{Q5} = 0$
- 基准模型：Fama-French 三因子模型 (CH-3)。

Model	GRS Statistic	P-Value	Result
Random Forest (RF)	5.64	0.0001	Reject H_0
XGBoost (XGB)	7.22	< 0.0001	Reject H_0

结论

- 统计学意义：强烈拒绝零假设。这证明 Q1-Q5 组合中存在无法被市场、市值和价值因子解释的显著 Alpha。
- 经济学意义：机器学习模型确实挖掘到了独立于传统因子之外的有效选股逻辑。

CNE-6 模型与行业收益归因

- 回归方程 (每月横截面回归):

$$R_{i,t+1} = \gamma_{ML} \cdot \hat{y}_{i,t} + \sum_{k=1}^{10} \gamma_{Style,k} \cdot F_{Style,k} + \sum_{j=1}^{110} \gamma_{Ind,j} \cdot D_{Ind,j} + \epsilon_{i,t}$$

- 控制变量: 10 个 Barra 因子 + 110 个申万行业变量

Factor	Return (γ)	t-stat
ML Score	22.80%	20.65
Size	-1.14%	-7.59
Momentum	0.17%	2.30
Liquidity	-0.46%	-4.31
Earnings Yield	0.25%	3.90
Beta	0.17%	2.07

结论:
在剥离了所有风格和行业干扰后，ML 因子的纯收益 t-stat 高达 **20.65**，证明策略拥有极其纯粹且显著的 **Alpha**，而非简单的风格暴露。

请大家批评指正！

本项目代码已开源于 https://github.com/nachuanchen/AMQI