

基于 4 种机器学习算法的量化投资实证研究

杨佳琪

(湖南大学, 湖南省、长沙市, 410006)

摘要: 股票截面收益的研究不仅关乎实务界的投资决策, 而且影响着学术界的理论发展, 因而因子模型从诞生以来就备受瞩目, 围绕这一课题发现的定价因子也多达数百个, 如何有效地选择因子并设定预测函数成为研究的重难点。机器学习天然擅长高维的特征选择和复杂的预测问题, 越来越多资产定价领域的研究开始关注机器学习这一技术。本文以传统 OLS 回归为基准, 比较了 4 种主流机器学习算法在中国股票市场的绩效表现。实证结果表明, 所有的机器学习算法都能取得超过市场指数和传统 OLS 的投资绩效, 且非线性的梯度提升树模型的多空组合月平均收益率可达 2.64%, 夏普比率为 1.15。本文的研究结论对于将机器学习应用到因子投资领域具有重要参考价值。

关键词: 机器学习; 多因子模型; 投资组合

中图分类号: F8

文献标识码: A

1. 引言

对股票截面收益的研究一直是金融领域热议的话题, CAPM 模型以简单优美的数学式给出了资产预期收益率与市场因子之间的关系^[1]。Fama 和 French(1993)从实证角度出发, 构造了市值 (Small-Minus-Big, SMB) 因子和账面市值比 (High-Minus-Low, HML) 因子, 在 CAPM 模型的基础上提出了三因子模型, 拉开了多因子模型研究的序幕^[2]。此后, 越来越多的异象和定价因子被提出, Cochrane (2011) 将其形象地比作“因子动物园” (Factor Zoo) ^[3]。

一方面, 不断涌现的因子存在着数据挖掘、样本外失效等问题, 因子质量良莠不齐。Harvey et al.(2016) 列出了 300 多篇研究预期收益横截面的论文, 认为“金融经济学中大多数声称的研究发现可能是错误的”^[4]。Hou, Xue, Zhang(2020)对已发表文献中的 452 个因子进行检验, 结果表明大多数的异象无法复制, 并提出资本市场的效率比以前认识到的要高^[5]。传统的单因子检验, 如 Fama-Macbeth 回归、ICIR 检验等又忽略了因子之间的交互作用, 向前、向后回归也存在着严重的过拟合风险。另一方面, 不断出现的异象和定价因子也表明, 仅靠少数几个经典因子并不能够为预测提供长期、稳定的信息, 因子模型需要随时间动态地考虑大量因子。模型纳入的因子数增加后, 传统线性回归不仅面临着过拟合的风险, 考虑因子交互作用时其函数形式还存在着无数的可能, 从而难以人为指定。

随着多因子模型研究的深入, 传统定价方法在因子选择和模型形式上逐渐显现出各种局限和不足。而计算机技术的发展和多学科融合发展的背景下, 越来越多研究开始关注机器学习技术。黄乃静和于明哲 (2018) 认为机器学习中的监督学习技术具有灵活的函数形式, 不仅有较高的样本外预测能力而且可以避免过拟合^[6]。机器学习中的许多算法天然具备特征选择的能力, 并且能够以更灵活的形式逼近输入与输出之间的线性或非线性关系, 为多因子量化投资提供了更多的可能空间。

因此, 在多因子量化投资领域应用机器学习技术, 与其说是科技发展和多学科融合背景下的结果, 不如说是随着该领域研究和实践地不断深入, 机器学习正在成为一种必不可少的技术工具, 为传统方法面临的诸多挑战提供了有力的应对方法。

2. 实证设计

2.1. 样本选取与数据说明

本文选择 2017-2021 年的沪深 300 指数样本股作为初始研究样本。在阅读大量文献和券商研报后, 本文从优矿网 (<https://uqer.datayes.com>) 获取了共计 8 类, 244 个因子作为候选因子, 为模型提供了丰富的输入, 因子的计算及分类参见网站。股票月度收益率、无风险利率等数据来自 CSMAR 数据库。

由于不同类型的因子在取值范围上有较大差异, 对于线性模型而言, 量级大的因子可能在预测中占据更重要的位置; 对于非线性模型而言, 量级差异可能影响模型的收敛速度。为了提高模型的训练速度和预测的准确性, 本文对因子数据进行去极值、填充缺失值、标准化等预处理步骤。具体而言, 本文采用绝对值差中位数法 (Median Absolute Deviation, MAD) 去极值, 以行业均值填充缺失值, 最后采用均值方差标准化对去极值和填充缺失值后的因子数据进行标准化处理。

2.2. 模型选择

在机器学习算法的选择方面, 首先, 机器学习可分为传统机器学习和深度学习两种, 本文没有办法穷尽所有两大类的所有算法, 而神经网络这类深度学习算法通常用于图像、语音识别等领域, 相比之下, 中国股票市场的历史数据有限, 对于深度学习而言仍属于小样本, 较难满足复杂模型的训练需求, 从时间成本和样本量出发, 本文仅采用了传统机器学习算法; 其次, 为了应对因子选择和函数形式两个难点, 选择了具有嵌入式特征选择特性的算法, 并且考虑了线性和非线性两大类别的算法丰富函数形式。具体而言, 本文选择了 5 种机器学习算法, 包含两种线性机器学习算法——套索回归 (Least absolute shrinkage and selection operator, Lasso) 和弹性网络回归 (Elastic Net), 两种非线性机器学习算法——随机森林 (Random Forest, RF) 和梯度提升树 (Gradient Boosting Decision Tree, GBDT)。这几种算法在大量研究中被证实具有稳健的预测能力。以套索和弹性网络算法为例, 谢合亮和胡迪 (2017) 以沪深 300 指数成分股为研究对象, 将 OLS 回归、套索算法、弹性网络算法应用与量化投资, 研究发现, 基于套索和弹性网络方法构造的投资组合的绩效表现优于普通最小二乘法, 且弹性网络算法相比套索方法更能筛选出有效因子^[7]。以随机森林和梯度提升树算法为例, Booth (2013) 构建了基于 RF 性能加权的自动交易系统, 该系统在盈利能力和预测精度方面都优于其他集成技术^[8]。Ładyżyński et al. (2013) 运用 RF 算法对标普 500 的成分股预测股价取得了不错的效果^[9]。本文所选择的模型在训练完成后可以通过系数或是特征重要性来评价因子对模型的贡献, 一定程度上避免了机器学习算法的“黑箱”问题, 增强了模型结果的可解释性。

2.3. 模型的训练与预测

在模型的训练与预测方面, 考虑到金融序列具有时序性, 常见的交叉验证法存在数据泄露的风险, 而滑动窗口法不仅可以保留时序性, 而且运用决策时点前的数据训练模型也更符合真实投资的场景。本文在实证中将滑动窗口的训练集长度设置为 24 个月, 并在稳健性检验中考虑了 12 个月、36 个月的情形。具体而言, 本文数据集的构建方式为 t 期个股因子配对 $t+1$ 期股票月度收益率。在滑动窗口的第一期用 2007 年和 2008 年两年的数据集进行模型的训练, 将 2009 年 1 月的截面数据作为测试集; 滑动窗口的第二期用 2007 年 2 月至 2009 年 1 月的数据集进行模型的训练, 将 2009 年 2 月的截面数据作为测试集, 以此类推, 直至最后一期共滚动训练和预测 155 次。考虑时间成本和可行性, 本文仅在滑动窗口的第一期采用网格调参的方式选择模型参数, 此后各期保持不变, 参数选择结果如下表所示。

表 2.1 模型参数设置

模型	参数名称	参数值 Lenth = 12	参数值 Lenth = 24	参数值 Lenth = 36
Lasso	alpha	0.0001	0.0003	0.00002
Elastic	alpha	0.0001	0.0005	0.00005
	l_1 _ratio	0.6	0.5	0.5
RF	max_depth	7	5	5
	n_estimators	400	200	250
	max_features	sqrt	log2	log2
	min_samples_split	5	5	5
GBDT	learning_rate	0.05	0.05	0.1
	max_depth	3	3	3
	n_estimators	100	100	100
	max_features	log2	log2	log2

2.4. 投资组合的构建与绩效评价

为了检验各机器学习算法的预测能力，本文将分别基于各模型的预测结果构建投资组合并进行绩效检验。

投资组合的构建方面，以 24 个月的滑动窗口长度为例，2007 年 1 月至 2008 年 12 月作为初始滚动区间，选取 2009 年 1 月至 2021 年 11 月的因子数据进行滚动预测。在每个回测区间的最后一个交易日，以过去 24 个月的数据作为训练集对模型进行训练，以当期因子数据作为训练所得模型的输入数据，预测得到下期股票的收益率，基于预测结果构建投资组合，即等权重地买入预期收益率排名前 10 位的股票，卖出排名后 10 位的股票。通常，与股票量、价有关的因子的调仓频率为月度，而基于财务指标构建的因子调仓频率为年度，考虑到本文两类因子均有，且为了保证因子数据的时效性，减少滞后的影响，在整个回测区间内均按月调仓，即每月最后一个交易日买入股票，持有一个月，至下月最后一个交易日换仓，忽略交易手续费。

在模型评价方面，本文通过构建投资组合进行投资绩效的检验。每月末根据预测结果，做多预期收益最高的前 10 只股票，并做空预期收益最低的后 10 只股票，并计算在整个回测区间内的月平均收益率、经 Fama-French 五因子模型调整后的阿尔法、夏普比率和最大回撤等指标进行绩效评价。

3. 实证结果

3.1. 模型投资绩效结果对比

表 3.1 列示了基于 4 种机器学习算法和基准 OLS 所构建的投资组合的绩效表现。

首先，考察多因子模型的投资绩效能否超越指数表现。从收益表现来看，在回测区间内，沪深 300 指数的月平均收益率仅为 0.83%，而传统 OLS 回归的月均收益率为 1.90%，经 Fama-French 调整后的阿尔法仍有 1.69%，收益表现十分显著；从经风险调整后的收益来看，同期沪深 300 指数的夏普比率仅为 0.32，而传统 OLS 回归的夏普比率为 0.91，每承受一单位的风险，基于传统 OLS 构建的投资组合能够获得更高的收益回报；从风险的角度来看，由于回测区间跨越了两次大的金融危机，沪深 300 指数的最大回撤高达 42.72%，如果在最高点买入指数基金，投资者将承受极高的损失，而基于传统 OLS 回归构建的投资组合的最大回撤为 32.90%，抗风险能力远好于沪深 300 指数。不难看出，即便是最简单线性多因子

模型也能够取得超过市场的投资表现, 说明了多因子模型的有效性。

其次, 以传统 OLS 为基准, 基于机器学习算法构建的投资组合的收益表现和抗风险能力均优于传统 OLS 回归。值得注意的是, 套索算法、弹性网络算法均属于线性模型, 在加入惩罚项后, 月平均收益相比 OLS 回归分别提升了 14.89%、22.34%, $FF5-\alpha$ 分别提升了 2.42%、18.18%, 最大回撤分别下降了 15.22%、20.59%, 表明惩罚项能够一定程度上克服噪音、相关性和无效因子对线性模型拟合效果的影响, 从而提升预测的准确性。

最后, 在机器学习模型之中, 非线性模型的收益表现整体优于线性模型, 这意味着因子与因子之间、因子与收益率之间可能存在着线性模型无法捕捉的非线性的关系。其中, 梯度提升树模型构建的投资组合的月均收益率达到 2.64%, 相比 OLS 回归和弹性网络算法分别提升了 38.95%和 8.64%, 夏普比率达 1.15, 分别提升了 26.37%和 13.86%, 最大回撤方面的提升较小。

表3.1 各模型投资绩效结果

	月均收益率 (%)	NW-t	$FF5-\alpha$ (%)	NW-t	夏普比率	最大回撤 (%)
OLS	1.90	3.86***	1.69	2.91***	0.91	25.92
Lasso	2.28	4.12***	1.80	2.87***	0.96	32.90
Elastic Net	2.43	4.27***	2.09	3.27***	1.01	27.74
RF	2.51	5.01***	2.23	3.73***	1.09	23.92
GBDT	2.64	4.99***	2.41	3.67***	1.15	25.74

注: *, **, ***分别表示在 10%、5%、1%的显著性水平下显著。

进一步地, 表 3.2 以 2015 年为分界点, 考察了模型在不同时间区间内的投资绩效差异, 从而验证模型投资表现的稳定性。选择以 2015 年作为分界点的原因有两点, 其一是 2015 年中国股市经历了一次大涨大跌, 沪深 300 指数从 4840 点跌至 2877 点, 此后的市场风格可能发生转变; 其二是以 2015 年为界可以将整个回测区间划分为两段时长大致相当的区间, 且两段区间内都涵盖了一次完整的牛熊区间。

表3.2 各模型分时间区间的投资绩效结果

	月均收益率 (%)	NW-t	$FF5-\alpha$ (%)	NW-t	夏普比率	最大回撤 (%)
Panel A: 2009.02-2015.12						
OLS	1.65	2.21**	1.53	1.61	0.67	25.92
Lasso	1.94	2.40**	1.47	1.42	0.71	28.67
Elastic Net	2.04	2.47**	1.74	1.63	0.73	24.41
RF	2.26	3.09***	1.51	1.69*	1.00	23.93
GBDT	2.13	2.84***	1.76	1.89*	1.03	26.77
Panel B: 2016.01-2021.12						
OLS	2.18	3.59***	1.94	2.66***	1.31	15.42
Lasso	2.67	3.71***	2.18	2.99***	1.36	21.27
Elastic Net	2.89	3.85***	2.33	3.17***	1.45	19.58
RF	2.89	4.01***	2.48	2.83***	1.17	20.20
GBDT	3.09	4.20***	2.89	2.87***	1.25	25.74

注: *, **, ***分别表示在 10%、5%、1%的显著性水平下显著。

观察表 3.2 不难发现, 其一, 在两段区间内沪深 300 指数的月平均收益率分别为 0.74%和 0.86%, 最大回撤分别为 70.75%和 40.56%, 而所有模型多空组合的月均收益率均

超过了 1.6% 且是显著的, 最大回撤均控制在 30% 以内, 表明模型在收益表现和抗风险能力方面都能够超越沪深 300 指数, 且各模型间投资绩效的比较结果与整个区间内回测的结论大致相当, 即所有机器学习算法的绩效表现均优于传统线性回归, 而非线性机器学习算法进一步优于线性机器学习算法。其二, 在 2015 年及以前, 非线性模型多空投资组合经 Fama 和 French 五因子调整后的收益在 10% 的显著性水平下显著, 而所有线性模型的多空投资组合经 Fama 和 French 五因子调整后的收益均不显著, 这一现象的原因可能是非线性模型能够更好地刻画因子与股票截面收益之间非线性的关系, 因而提供了五因子模型无法解释的超额收益; 其三, 在 2016 年及以后, 模型产生的收益大部分无法被五因子模型解释, 这一发现也说明了因子的有效性会随时间变化, 少数几个经典因子很难长期为收益预测提供信息, 影响收益的因素可能是丰富且动态变化的。

4. 稳健性检验

为检验以上实证结果是否可靠, 本文还通过改变训练集长度进行了稳健性检验。表 4.1 展示了当训练集长度为 12 个月和 36 个月时各模型的投资绩效。结果表明, 本文的主要结论并未发生显著变化, 即多因子模型能够取得超越市场指数的投资收益, 而机器学习算法, 特别是非线性机器学习算法则进一步优于传统线性回归模型。

表4.1 不同训练集长度下各模型的回测结果

	月均收益率 (%)	NW-t	FF5- α (%)	NW-t	夏普比率	最大回撤 (%)
Panel A: 12 个月滑动窗口						
OLS	1.19	2.46***	0.90	1.81*	0.61	30.76
Lasso	1.62	3.04***	1.30	2.16**	0.69	35.04
Elastic Net	1.65	3.39***	1.56	2.74**	0.75	31.79
RF	1.87	3.11***	1.75	2.52**	0.68	32.20
GBDT	2.03	3.56***	1.97	3.00***	0.82	31.25
Panel B: 36 个月滑动窗口						
OLS	0.98	1.77**	0.65	1.07	0.41	35.89
Lasso	2.21	4.07***	2.12	3.18**	0.97	28.32
Elastic Net	2.32	4.42***	2.13	3.51***	1.04	23.41
RF	2.44	4.48***	2.23	3.72***	1.10	27.78
GBDT	2.49	4.77***	2.32	3.66***	1.14	26.60

注: *, **, *** 分别表示在 10%、5%、1% 的显著性水平下显著

5. 结论

本文选择了沪深 300 指数成分股 2007 年 1 月至 2021 年 12 月的月数据作为分析对象, 选取八大类共 244 个因子, 以 OLS 回归为基准, 运用套索、弹性网络、随机森林和梯度提升树四种机器学习算法进行股票截面收益的滚动预测, 并基于预测结果构建投资组合, 进而评价投资绩效。实证结果表明, 机器学习算法, 特别是非线性的机器学习算法能够取得超越市场指数和传统 OLS 的投资绩效。这也表明, 机器学习算法能够更高效地进行因子的选择, 有效的降低过拟合的风险, 进一步逼近因子与截面收益之间非线性关系, 从而获得更优的投资绩效。本文的研究结论对投资实践有一定的指导价值, 对机器学习在资产定价领域的研究应用也有一定的补充。

参考文献

- [1] Sharpe W F. Capital asset prices: A theory of market equilibrium under conditions of risk[J]. The journal of finance, 1964, 19(3): 425-442.
- [2] Fama E F, French K R. Common risk factors in the returns on stocks and bonds[J]. Journal of, 1993.
- [3] Cochrane J H. Presidential address: Discount rates[J]. The Journal of finance, 2011, 66(4): 1047-1108.
- [4] Harvey C R, Liu Y, Zhu H. ... and the cross-section of expected returns[J]. The Review of Financial Studies, 2016, 29(1): 5-68.
- [5] Hou K, Xue C, Zhang L. Replicating anomalies[J]. The Review of Financial Studies, 2020, 33(5): 2019-2133.
- [6] 黄乃静,于明哲.机器学习对经济学研究的影响研究进展[J].经济学动态,2018(07):115-129.
- [7] 谢合亮,胡迪.多因子量化模型在投资组合中的应用——基于套索与 Elastic Net 的比较研究[J].统计与信息论坛,2017,32(10):36-42.
- [8] Booth A, Gerding E, McGroarty F. Automated trading with performance weighted random forests and seasonality[J]. Expert Systems with Applications, 2014, 41(8): 3651-3661.
- [9] Ładyżyński P, Żbikowski K, Grzegorzewski P. Stock trading with random forests, trend detection tests and force index volume indicators[C]//International Conference on Artificial Intelligence and Soft Computing. Springer, Berlin, Heidelberg, 2013: 441-452.

An empirical study on quantitative investment based on four machine learning algorithms

Jiaqi Yang

(School of Finance and Statistics, Hunan University, Changsha / Hunan, 410006)

Abstract: The prediction of stock sectional return is not only related to the investment decision of the industry, but also affects the theoretical development of the academic world. Therefore, the factor model has been attracting much attention since its birth. There are hundreds of abnormal phenomena and pricing factors found around this topic. Machine learning is naturally good at high-dimensional feature selection and complex prediction problems. More and more researches in the field of asset pricing begin to focus on machine learning technology. Based on traditional OLS regression, this paper systematically compares the performance of four mainstream machine learning algorithms in Chinese stock market prediction. The empirical results show that all the machine learning algorithms can achieve better investment performance than traditional OLS, and the average monthly return rate of the nonlinear gradient ascending tree model of long-short portfolio can reach 2.64%, and the Sharpe ratio is 1.15. The conclusion of this paper has important reference value for applying machine learning to factor investment.

Keywords: Machine learning; Multi-factor model; portfolio

作者简介:

杨佳琪(1996—), 女, 四川西昌人, 湖南大学金融与统计学院硕士研究生。研究方向:金融工程与金融科技。