机器学习驱动的基本面量化投资研究

李斌、邵新月和李玥阳

《中国工业经济》，2019年第8期

吕漫妮

2020年6月6日

目录

1. 引言
2. 研究设计
3. 实证结果与分析
4. 结论

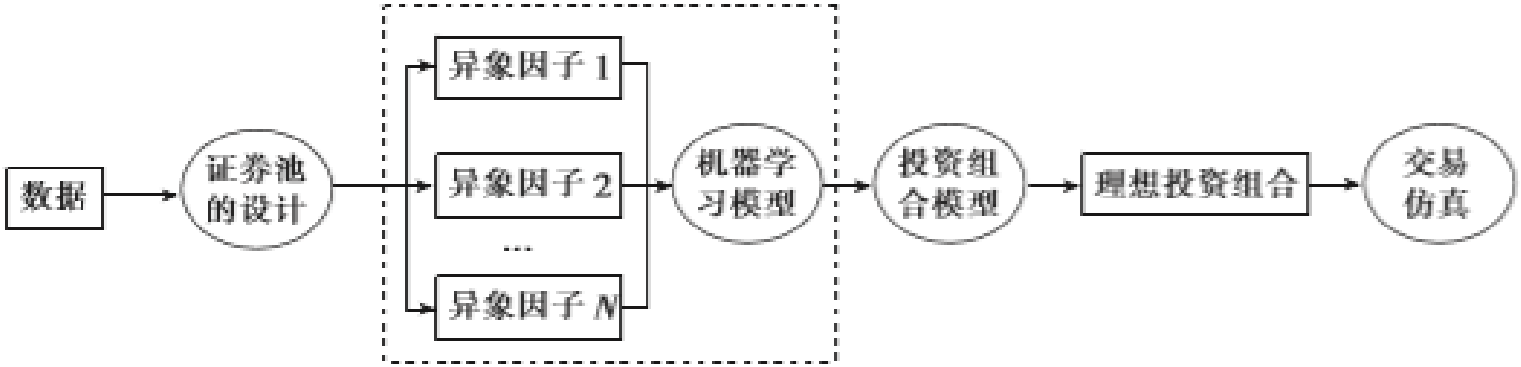
1. 引言

➢研究背景

- 1. 资产保值和增值是每个家庭和个人都会面临的问题。
- 2. 资产管理是金融服务实体经济的重要手段之一
- 3. 智能量化投资是我国金融业高质量发展的重要组成部分
- 4. 2017年，《新一代人工智能发展规划》，人工智能上升为国家战略
- 问题：如何发挥以机器学习为代表的人工智能技术优势，推动我国资产管理水平的提升？

1. 引言

➢研究背景



- 基本面量化投资融合了量化投资（算法驱动）与基本面投资（人为驱动），其核心是分析股票的基本面因素和风险溢价（或超额收益）之间的关系，或股票收益的准确预测。

1. 引言

➢研究动机

1. 金融研究提出数以百计的异象因子集合，但后续样本外检验发现大部分因子难以持续地提供超额收益，且因子间往往具有较强相关性；因子维度变大时，非线性因素使预测的复杂度急剧增加，亟待新方法介入
2. 传统的组合排序和FM回归并未综合考虑各因子及因子间的交互作用；现有研究方法并未提供高维因子与预测函数形式选择的建议。
3. 前美国金融学会会长Cochrane (2011)：在处理如此众多的因子时，将不得不使用“不同的研究工具”(“I suspect we will have to use different methods.”)

1. 引言

➢研究动机

- 机器学习和深度学习能够自动地寻找数据中的复杂结构和模式，从而提升预测能力：
  - 众多备选的预测函数形式；
  - 专门被设计用于逼近复杂的非线性关系；
  - 参数正则化和模型选择等技术有效降低过拟合风险。

研究问题

- 研究问题一：将机器学习算法与异象因子结合构建的股票收益预测模型能否通过有效识别数据间的线性关系和非线性关系，获得更好的预测效果？
- 研究问题二：若机器学习算法的运用能够提升预测绩效，究竟哪些因子能够更好地预测股票未来收益？ 基于机器学习算法筛选出的重要因子与传统单因子分析中显著的因子存在哪些差异？

1. 引言

➢现有研究——机器学习在资产定价中的应用

- Feng, Giglio, and Xiu (2017)基于LASSO方法衡量因子对资产定价的贡献，发现盈利性和投资因子更具有统计上显著的解释力
- Light et al.（2017）采用“偏最小二乘法”（Partial Least Square，PLS）来检验公司 特征对期望收益的预测能力；Kozak et al.（2019）和 Kelly et al.（2019）分别运用 PCA 方法和 IPCA 方法（Instrumented PCA）提取因子中的共同因素
- Lewellen（2015）发现通过 FM 回归方法 综合 15 个因子能够很好地预测股票超额收益；DeMiguel et al.（2017）从投资者效用的角度预测截 面收益的公司特征
- 李斌等（2017）分别采用支持向量机、神经网络、Adaboost 算法预测股价涨跌方向，发现机器学习算法具有更高的准确率

1. 引言

➢现有研究——机器学习在资产定价中的应用

- 潘莉和徐建国（2011）检验了六个因子与股票收益的关系，构建了适用于 A 股市场的因子模型。
- 胡熠和顾明（2018）从安全性、便宜性和质量三个维度共选取 8 个异象因子构造综合性指标，并将其应用于中国 A 股市场，验证了巴菲特的价值投资策略在中国市场的适用性。
- Jiang et al.（2019）分别采用 FM 回归、PCA、PLS 和 FC 等线性方法整合 A 股市场中的 75 个异象因子，发现这些方法能够从因子中提取出有助于预测的信息。
- 李斌等（2017）分别采用支持向量机、神经网络、Adaboost等19种 算法预测股价涨跌方向，发现机器学习算法具有更高的准确率

1. 引言

➢研究内容

收集中国A股市场96个异象因子，采用12种代表性的线性或非线性机器学习算法构建异象因子-超额收益预测模型，并构建投资组合，以检验模型在A股市场的表现

➢研究结论

1. 系统性地对比各种机器学习算法的绩效，证明机器学习算法能够显著超越传统线性回归的绩效；且非线性算法的预测绩效明显超过线性算法。深度学习算法绩效提升最明显
2. 分析异象因子的重要性发现交易摩擦类因子在A股市场具有更强的预测能力，采用线性方法和非线性方法所得的重要因子区别于单因子检 验所得的重要因子

1. 引言

➢创新点

1. 因子集合的创新。本文构建的96个异象因子集合是目前中国A股市场相关研究中最大的因子集合，基于全面的数据集上的分析预测能够获取更有效的信息。
2. 预测模型的创新。将基本面量化投资与机器学习算法结合构建预测模型的设计在中国股票市场的研究中相对缺乏，本文使用12种线性/非线性机器学习算法构建多因子预测模型，是目前中国股票市场最全面的利用机器学习算法进行的异象因子研究。

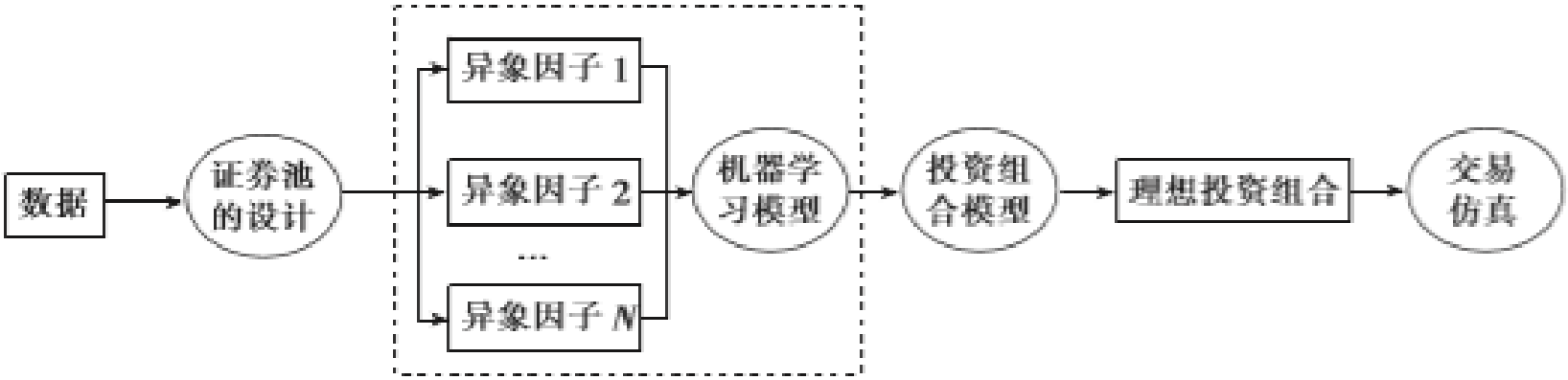
1. 引言

- **研究意义**

1. 丰富了经济学和管理学研究的工具箱。
2. 丰富了量化投资的理论和实践研究。
3. 丰富了中国市场中股票截面收益影响因素的研究。

2. 研究设计

➢模型的总体设计



$$\mathrm{R}_{t,i} = \mathrm{f}\left(\boldsymbol{x}_{t-1,i}; \theta\right) + \epsilon_{t,i}$$

- f(·)定义一个参数为θ的函数，在本文中为丰富的机器学习和深度学习方法中的函数形式
- $\mathrm{R}_{t,i}$为股票i第t期的超额收益
- $\boldsymbol{x}_{t-1,i} = \left(x_{t-1,i,1}, x_{t-1,i,2}, \ldots, x_{t-1,i,N}\right)$为公司$i$第$t-1$期的N个异象因子
- $\epsilon_{t,i}$为误差项

2. 研究设计 – 数据

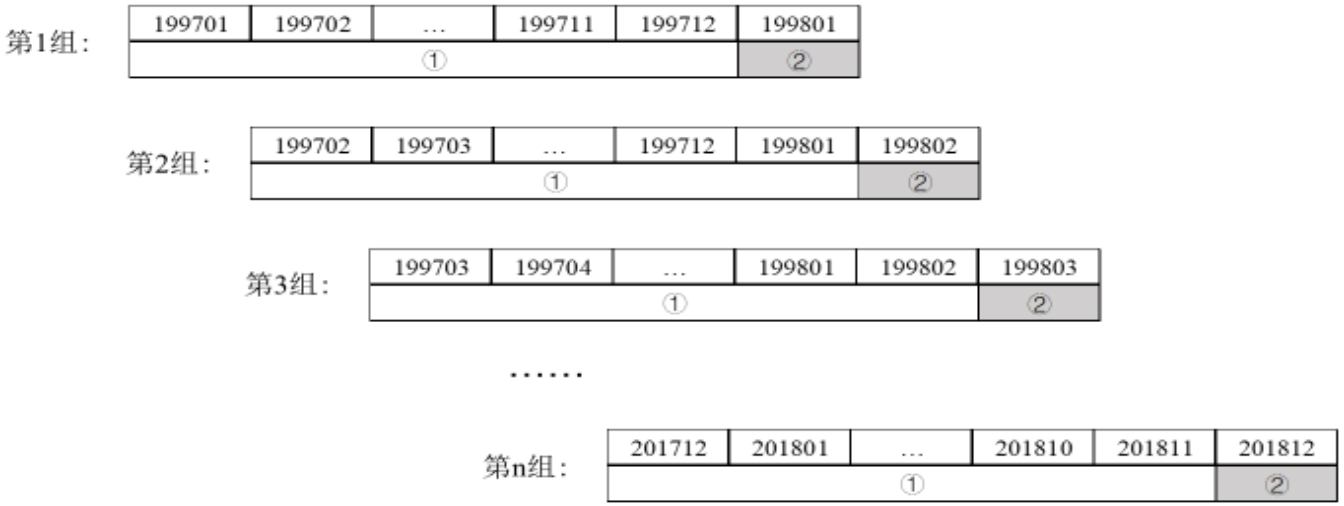
- **时间区间：** 1997年1月-2018年10月A股市场月频数据
- **因子集合：** 借鉴Green, Hand and Zhang (2017)，选取了96个公司特征代理异象因子，分为交易摩擦因子、动量因子、价值因子、成长因子、盈利因子、财务流动因子共六大类。
- **输出变量：** 考虑现金股利再投资的股票月度收益率
- **数据处理**
  - 因子：对于季度财务数据均进行月度填充，数据来源于CSMAR。
  - 不同因子的取值在数量级及分布上存在显著差异，将训练集数据标准化

1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
t-1年三季度报报表填充				t-1年年度报表填充				t年半年度报表填充		t年三季度报表填充	

  - 剔除掉ST、金融股及上市一年内数据后，共381062条有效样本
  - 缺失值处理：①若股票在第t月收益数据存在缺失，则剔除该股票在月份t上的所有数据；②若某只股票的因子值缺失，则以0填充。

2. 研究设计 – 研究方法

- **机器学习算法，采用监督学习算法**
  0. 基准算法：线性回归模型
  1. 线性算法：FC、Ridge、Lasso、ElasticNet、PLS
  2. 非线性算法-传统机器学习：SVM、GBDT、XGBoost、EN-ANN
  3. 非线性算法-深度学习算法：DFN、RNN、LSTM
- **训练、测试集划分：**滑动窗口法，网格调参



2. 研究设计 – 评估

➢**模型绩效衡量指标**

依据模型预测构建多空组合（做多前 10% 、做空后10%，等权）、多头组合、空头组合的月度收益率和风险调节收益（包括 Fama-French 三因子调节的阿尔法、Fama-French 五因子调节的阿尔法和夏普比率等）

➢**异象因子重要性衡量**

单因子	统计学意义上显著的因素（多空组合收益t-statistic>1.96）
机器学习方法	从全变量中剔除单一因子之后的收益损失

3. 实证结果1：机器学习模型在A股市场的实证绩效

➢ 机器学习算法全变量预测结果对比(12月滚动窗口)

Panel A:组合投资绩效

	多头组合			空头组合			多空组合		
	Mean (%)	FF5-α (%)	夏普比率	mean (%)	FF5-α (%)	夏普比率	mean (%)	FF5-α (%)	夏普比率
OLS	2.35	0.82	0.7108	0.34	-0.97	0.0460	2.01	1.58	1.5088
FC	2.62	1.00	0.7853	0.34	-0.76	0.0463	2.28	1.55	1.2931
Ridge	2.41	0.88	0.7334				2.08	1.65	1.5469
Lasso	2.44	0.87	0.7409				2.08	1.64	1.5300
Elastic	2.46	0.90	0.7489	0.34	-1.00	0.0473	2.12	1.70	1.5694
PLS	2.48	1.00	0.7549	0.19	-1.13	-0.0084	2.30	1.92	1.5709
SVM	2.52	1.00					2.25	1.86	1.7378
EN-ANN	2.59	1.01					2.34	2.34	1.8082
XGBoost	2.72	1.04					2.73	2.15	2.0066
GBDT	2.67	1.03					2.68	2.13	1.9264
DFN	2.86	1.25	0.8595				2.78	2.14	2.0150
RNN	2.52	1.10	0.7871				2.10	1.79	1.9794
LSTM	2.86	1.18	0.8584				2.57	2.01	1.9670
SIZE	2.45	0.51	0.7068	0.72	0.02	0.2034	1.73	0.27	0.6823
MKT	0.61	-0.01	0.1795						

机器学习算法能够更好地识别因子间的线性关系

非线性传统机器学习算法总体而言能够获得比线性算法更好的绩效

深度学习算法能够更好地识别非线性关系

- 不同策略间收益差异性检验（12 个月滚动窗口）

	OLS 与其他算法				DFN 与其他算法		
	多头组合	空头组合	多空组合		多头组合	空头组合	多空组合
FC	1.8044	1.8044	0.0201	FC	1.5921	−1.6284	1.8126
Ridge	1.4412	1.4412	−0.8061	Ridge	4.0901	−1.8952	3.4790
Lasso	1.4247	0.4539	0.7950	Lasso	3.8110	−2.1672	3.5617
ElasticNet	2.1662	0.1431	1.5268	ElasticNet	3.5755	−2.0009	3.2768
PLS	1.1332	−1.3566	1.4788	PLS	3.1101	−0.8950	2.3957
SVM	2.5459	−1.1565	2.4288	SVM	2.9808	−1.5177	2.6016
EN-ANN	1.6926	−0.7538	1.4916	EN-ANN	1.8888	−1.2340	1.8226
XGBoost	2.6377	−2.7590	3.2382	XGBoost	1.0217	0.6688	0.2266
GBDT	2.3413	−2.6646	3.0821	GBDT	1.3841	0.6528	0.4485
DFN	4.2601	−1.9629	3.6477				
RNN	1.1670	0.6897	0.3742				
LSTM	3.5156	−0.3451	2.5548				

- 机器学习算法都较OLS回归投资绩效存在明显提升→非线性模型的识别的重要性
- DFN能够显著超越Ridge、Lasso、ElasticNet、PLS和SVM算法获得更高的投资绩效→深度学习对非线性模型的识别有效性

➢ 剔除市值因子后的预测绩效（12个月滑动窗口）

检验模型是否受到市值因子的驱动而并非多个因子的聚合效果

全因子

	多空组合			剔除市值因子 多空组合		
	mean (%)	FF5-α (%)	夏普比率	mean	FF5α	夏普比率
OLS	2.01	1.58	1.5088	2.03%	1.60%	1.5172
FC	2.28	1.55	1.2931	2.24%	1.54%	1.2848
Ridge	2.08	1.65	1.5469	2.07%	1.63%	1.5467
Lasso	2.08	1.64	1.5300	2.08%	1.65%	1.5114
Elastic	2.12	1.70	1.5694	2.11%	1.68%	1.5435
PLS	2.30	1.92	1.5709	2.28%	1.88%	1.5725
SVM	2.25	1.86	1.7378	2.21%	1.83%	1.6674
EN-ANN	2.34	2.34	1.8082	2.24%	1.75%	1.8326
XGBoost	2.73	2.15	2.0066	2.60%	2.12%	1.9610
GBDT	2.68	2.13	1.9264	2.61%	2.09%	1.9219
DFN	2.78	2.14	2.0150	2.82%	2.40%	2.0971
RNN	2.10	1.79	1.9794	2.49%	1.94%	1.7429
LSTM	2.57	2.01	1.9670	1.75%	1.41%	1.5037
SIZE	1.73	0.27	0.6823	1.73%	0.27%	0.6823

结论：
机器学习驱动的基本面量化投资模型的绩效并非由 size 因子主导，而是多因子共同作用的结果。

➢ 集成各类算法的绩效

为了更直观的说明机器学习算法相对于传统线性模型的绩效提升，本文简单加权11种机器学习算法（其中未包括线性组合的FC）构建集成预测模型：$R_{t,i}^{ensemble} = \frac{1}{11}\sum_{j=1}^{11} R_{t,i}^{j}$，结果显示集成算法的基本面量化投资策略在多空组合和单独做多/空组合上均能够获得显著优于 OLS 算法的收益和风险调节收益。

	3 个月滑动窗口			12 个月滑动窗口		
	多空组合	多头组合	空头组合	多空组合	多头组合	空头组合
Mean(%)	2.56***	2.60***	0.04	2.98***	2.89***	–0.09
	(7.0439)	(3.5751)	(0.0683)	(9.2611)	(3.8840)	(–0.1451)
FF3–α(%)	2.38***	1.34***	–1.20***	2.67***	1.41***	–1.46***
	(6.7331)	(4.7965)	(–6.8497)	(8.8577)	(5.8175)	(–8.1355)
FF5–α(%)	2.18***	1.20***	–1.21***	2.55***	1.28***	–1.48***
	(6.2674)	(4.7571)	(–6.2301)	(9.2451)	(6.1019)	(–7.7384)
夏普比率	1.4698	0.7715	–0.0651	2.1736	0.8720	–0.1085
	24 个月滑动窗口			36 个月滑动窗口		
	多空组合	多头组合	空头组合	多空组合	多头组合	空头组合
Mean(%)	2.69***	2.78***	0.09	1.96***	2.41***	–0.45
	(8.0896)	(3.5986)	(0.1321)	(5.8373)	(3.0867)	(0.6592)
FF3–α(%)	2.37***	1.33***	–1.24***	1.70***	1.06***	–0.84***
	(7.6746)	(5.1849)	(–7.3478)	(6.2174)	(4.4701)	(–4.6277)
FF5–α(%)	2.17***	1.13***	–1.24***	1.48***	0.85***	–0.83***
	(8.3591)	(5.4178)	(–6.7986)	(5.1328)	(3.3757)	(–4.0735)
夏普比率	1.9439	0.8296	–0.0398	1.4439	0.7216	–0.0912

➢ 考虑交易成本的绩效

考虑交易成本为单边 0.50%、0.75%和 1.00%的三种情形

	transcost=0.50%			*transcost*=0.75%			*transcost*=1.00%		
	Mean (%)	FF5−α (%)	夏普比率	Mean (%)	FF5−α (%)	夏普比率	Mean (%)	FF5−α (%)	夏普比率
OLS	0.80	0.58	0.6710	0.30	0.08	0.2521	−0.20	−0.42	−0.1668
FC	1.07	0.55	0.6675	0.57	0.05	0.3547	0.07	−0.45	0.0419
Ridge	0.87	0.65	0.7209	0.37	0.15	0.3078	−0.13	−0.35	−0.1052
Lasso	0.87	0.64	0.7109	0.37	0.14	0.3014	−0.13	−0.36	−0.1081
Elastic	0.91	0.70	0.7474	0.41	0.20	0.3364	−0.09	−0.30	−0.0746
PLS	1.09	0.92	0.8176	0.59	0.42	0.4409	0.09	−0.08	0.0643
SVM	1.04	0.86	0.8846	0.54	0.36	0.4580	0.04	−0.14	0.0314
EN−ANN	1.13	0.87	0.9586	0.63	0.37	0.5338	0.13	−0.13	0.1090
XGBoost	1.52	1.15	1.2099	1.02	0.65	0.8115	0.52	0.15	0.4131
GBDT	1.47	1.13	1.1465	0.97	0.63	0.7565	0.47	0.13	0.3665
DFN	1.57	1.14	1.2311	1.07	0.64	0.8392	0.57	0.14	0.4473
RNN	0.89	0.79	0.9306	0.39	0.29	0.4062	−0.11	−0.21	−0.1182
LSTM	1.36	1.01	1.1327	0.86	0.51	0.7156	0.36	0.01	0.2984

结论：本文所提出的基本面量化投资策略在承担合理交易成本时仍能获得显著的收益与风险调节收益。

➢ 变动滑动窗口的绩效

除采用 12 个月滑动窗口外，本文也分别采用了 3 个月、24 个月和 36 个月的滑动窗口。

结论：

1. 除 36 个月滑动窗口外，DFN 算法均优于其他机器学习算法和 OLS 的预测效果（包括多空 组合和做多组合），验证了 12 个月滑动窗口所得结果的一般性。
2. 采用较长的训练区间（如 36 个月）时的绩效相对较弱，而 3 个月、12 个月和 24 个月滑动窗口时 的绩效不存在明显差异。

3. 实证结果2：机器学习的视角下因子的重要性。

- 本文拟采用 14 种检验方法， 包括单因子检验、OLS、FC、PLS、Lasso、Ridge、ElasticNet、 SVM、EN-ANN、XGBoost、GBDT、DFN、RNN 和 LSTM 等。
- 其中单因子检验是传统资产定价研究中所采用的检验方法（Bali et al.，2016）。 对于其他方法，本文计算去除某一因子后的收益损失来衡量该因子的重要性。
- 本文将筛选出重要性位于前 20 的因子作为重要因子。

➢ 机器学习算法中累计被选中次数超过 5 次的重要因子

序号	因子	因子名称	因子类别	N
1	*aeavol*	收益公告异常交易量	交易摩擦因子	10
2	*egr*	股东权益变化	成长因子	10
3	*turnsd*	换手率的波动率	交易摩擦因子	10
4	*LM*	标准化的换手率	交易摩擦因子	9
5	*retvol*	总波动率	交易摩擦因子	9
6	*skewness*	总偏态	交易摩擦因子	9
7	*vold*	交易额	交易摩擦因子	8
8	*CFdebt*	现金流负债比	财务流动性因子	7
9	*idvol*	异质波动率	交易摩擦因子	7
10	*illq*	非流动性风险	交易摩擦因子	7
11	*tang*	偿债能力/总资产	财务流动性因子	7
12	*chfeps*	预期每股收益的变化	盈利因子	6
13	*lagretn*	短期反转	动量因子	6
14	*retmax*	最大日收益率	交易摩擦因子	5
15	*sharechg*	股本增长率	交易摩擦因子	5
16	*stdvold*	交易额的波动率	交易摩擦因子	5

结论：

1、交易摩擦类因子在 A 股市场具有极强的预测能力。

2、财务因子在中国市场的预测能力相对较弱，筛选出的重要因子多为采用股票交易数据计算所得的异象因子。

结论：交易摩擦类因子被选为重要因 子的比例为 52%，强的预测能力并非是交易摩擦因子本身占总体比例较大导致。

因子类别	因子总数	重要因子数	占比(%)
交易摩擦因子	21	11	52
财务流动性因子	10	2	20
动量因子	6	1	17
盈利因子	14	1	7
成长因子	35	1	3
价值因子	10	0	0

交易摩擦因子具有较强预测能力的可能原因：

1、交易摩擦因子主要是依据交易数据构 造的月频数据，及时性更高并包含了更多的市场信息。

2、交易摩擦因子稳定性相对更高。

➢ 以 16 项重要因子作为输入构建投资组合的绩效

	多头组合			空头组合			多空组合		
	Mean (%)	FF5–α (%)	夏普比率	Mean (%)	FF5–α (%)	夏普比率	Mean (%)	FF5–α (%)	夏普比率
OLS	2.65	1.08	0.8192	0.10	–1.19	–0.0405	2.55	2.06	1.6652
FC	2.84	1.18	0.8612	–0.15	–1.31	–0.1283	2.98	2.28	1.7573
Ridge	2.65	1.08	0.8192	0.10	–1.19	–0.0405	2.55	2.06	1.6652
Lasso	2.68	1.10	0.8292	0.06	–1.20	–0.0527	2.62	2.09	1.6744
Elastic	2.67	1.09	0.8240	0.07	–1.20	–0.0483	2.59	2.07	1.6652
PLS	2.73	1.16	0.8462	0.04	–1.16	–0.0619	2.69	2.10	1.7739
SVM	2.51	1.00	0.7811	0.12	–1.18	–0.0335	2.39	1.97	1.5760
EN–ANN	2.46	0.85	0.7570	0.30	–0.98	0.0322	2.16	1.61	1.7334
XGBoost	2.85	1.21	0.8789	–0.01	–1.29	–0.0783	2.86	2.30	2.0182
GBDT	2.81	1.15	0.8610	0.07	–1.22	–0.0519	2.74	2.16	1.8468
DFN	3.24	1.44	0.9533	–0.17	–1.40	–0.1394	3.41	2.63	2.0182
RNN	2.42	0.88	0.7547	0.46	–0.90	0.0882	1.95	1.57	1.5986
LSTM	2.91	1.22	0.8651	0.06	–1.21	–0.0538	2.85	2.22	2.0861

结论：所有算法较全变量均存在不同幅度的提升。

4. 结论

1. 线性机器学习算法表现优于单因子和线性回归模型；非线性机器学习算法的绩效表现总体优于线性算法；深度学习算法(DFN和LSTM)获得了最好的投资绩效。
2. 机器学习算法能够识别异象因子间的非线性关系而获得更好的投资收益，即使考虑交易成本和做空限制也能获得显著的超额收益。
3. 因子重要性显示，机器学习能够发现与传统单因子检验不一样的重要因子，交易摩擦类因子对股票收益具有较强的预测能力。
4. 基于重要因子集合的投资策略绩效能够超越基于全变量集合的。