基本面因子与股票收益：基于机器学习方法

2022 秋季学期《机器学习》课程论文

齐国皓[a], 夏京[b], 田鑫芳[c,*]

[a] 学号 *202018070207*，经济统计 *2001* 班，金融与统计学院，湖南大学
[b] 学号 *202018070211*，统计 *2001* 班，金融与统计学院，湖南大学
[c] 学号 *202118070112*，统计 *2101* 班，金融与统计学院，湖南大学

Abstract

本文基于 1985 年 1 月至 2022 年月的美股市场的 207 项异象 (因子)，采用线性回归、预测组合算法、Lasso 回归、岭回归、弹性网络回归、偏最小二乘回归、支持向量机、梯度提升树、极端梯度提升树、集成神经网络等 10 种机器学习算法，构建股票收益预测模型及投资组合。实证结果显示，机器学习算法能够有效地识别异象 (因子) 与收益之间的关系，能够获得优于市场组合 (以 SP500 为例) 的收益，非线性机器学习算法普遍优于线性机器学习算法，但 Ridge 和 LR 在扩展训练期后仍然有着较好的收益，推测异象-收益之间存在某种较强的 “线性关联”。本文进一步检验了异象 (因子) 在机器学习算法中的重要性，发现与收益 (Return) 相关的异象 (因子) 有较强的预测能力。

Keywords: 机器学习, 资产定价，异象因子，股票收益

1. 引言

预测问题是资产定价的核心。为了给股票定价，投资者必须预测公司未来的现金流量。寻求表现出色的交易策略的投资者寻找预测资产回报的信号。测试资产定价模型的研究人员寻找预测变量，这些变量可以预测资产之间的回报差异，或者捕捉不同时间回报的可预测变化。

学术界对因子的研究可以追溯到 20 世纪 30 年代。Graham and Dodd(1934) 在他们的著作 *Security Analysis*（证券分析）[1] 中提出了价值溢价的概念，60 年代和 70 年代，CAPM 和 APT 相继被提出，为研究因子提供了定量分析方法。Basu(1977)[2] 发现 “低价

*旁听课程，辅助我们进行研究与报告撰写
Email address: qgh124430@hnu.edu.cn (齐国皓)

股效应”，Banz(1981)[3] 发现“小市值效应”，这些现象与主流的有效市场假说相违背，因此被称为异象。Fama and French(1992)[4] 将这两个异象整合起来，与市场因子共同构成了经典的三因子模型，成为全球各国股票市场实证资产定价研究的基准模型。基于 Fama and French(1992)[4] 的研究，很多新的模型先后被提出：Carhart(1997)[5] 在 Fama-French 三因子模型的基础上加入了截面动量因子 (MOM)，提出了 Carhart 四因子模型。Novy-Marx(2013)[6] 指出盈利能力和未来预期收益密切相关，提出了他们的四因子模型。Hou et al. (2015)[7] 从实体投资经济学理论出发，提出了新的四因子模型 (q-因子模型)。Fama and French(2015)[8] 在 Fama-French 三因子模型的基础上添加了盈利和投资两个因子，提出了新的五因子模型。

在实证资产定价领域过去几十年的研究中，人们挖掘了数百个能够提供超额收益的异象 (因子) 以及能够较好预测股票收益的变量，也因此产生了“因子动物园”的说法 (Campbell and Liu , 2019[9])，许多学者使用组合排序 (Portfolio Sort) 和 Fama-MacBeth(FM) 回归等方法检验异象，例如，Hou, Xue and Zhang(2020) [10] 检验了 452 个异象，发现 65% 的异象无法通过 $|t| \geq 1.96$ 的单一测试。Chen and Zimmermann(2021) [11] 开源了 207 个因子，发现复现的 t 统计量对原始多空组合的 t 统计量的回归斜率为 0.88，R2 为 82%。现存的异象 (因子)、预测变量数量是众多的，但在过去很长的一段时间，资产定价研究一直仅仅关注低维的模型。例如，横截面股票收益预测的工作主要集中在具有少量公司特征的回归上，研究人员调查了大量公司特征的预测能力，但在任何个别研究中，研究人员考虑的预测因素的数量通常很少。同样，研究人员希望通过因子模型来总结股票收益横截面中的投资机会，并将重点放在具有极少数因子的模型上。例如，Hou, Xue and Zhang(2015) [7] 以及 Fama and French(2015) [8] 在其因子模型中除了价值加权市场投资组合超额收益外，仅包括三四个因子，这些因子是根据公司规模、盈利能力、投资或公司账面市值比等公司特征构建的投资组合。对模型施加极端的稀疏性，使得传统的统计方法表现良好，但缺遗漏了大量的信息 (单独的以及联合的影响)，从 Fama and French(1992)[4] 提出的三因子模型，到近期 Roy and Shijin(2018)[12] 提出的六因子模型，我们也可以发现学术界在慢慢适应“遗漏信息”的现实因素。

近年来，机器学习方法的兴起，提供了无需对预测问题施加极端的稀疏性限制的机会，

允许学者们考虑大量预测变量的联合效应, 推动了实证资产定价领域的发展。Jiang, Tang and Zhou (2019)[13] 使用了 FM 回归、主成分分析 (PCA)、偏最小二乘法 (PLS)、预测组合方法 (FC)，从中国市场上 75 个公司特征中提取信息，发现与交易摩擦、动量和盈利能力相关的公司特征是中国股市未来股票回报的最有效预测指标。Gu, Kelly and Xiu (2021)[14] 使用了自动编码器神经网络模型将来自资产特征等协变量与资产回报结合起来，取得了较小的资产定价模型的样本外定价误差。Zhang (2022)[15] 使用公司特征、系统性风险和宏观经济作为预测信号，发现具有记忆机制和 Transformer 的 RNN 在预测性方面具有最佳性能，而 CNN 的性能相对其他的 NN 模型较差。Gu, Kelly and Xiu (2020)[16] 使用机器学习预测向投资者展示了巨大的经济收益，并确定了性能最佳的方法 (树和神经网络)。李斌等 (2019)[17] 使用 12 种机器学习算法 (包括线性与非线性监督学习模型) 构建股票收益预测模型及投资组合，实证结果显示使用机器学习算法构建的投资策略能够获得比传统线性算法和所有单因子更好的投资绩效。

本文计划解决的第一个问题是：机器学习算法能否有效地识别出异象 (因子) 和超额收益间的线性和非线性关系，并依据预测构建的投资组合能够获得更好的绩效? 根据机器学习理论中的 “没有免费的午餐定理” (No Free Lunch Theorem) (Wolpert，1996)[18]，本文无法预先知道哪个算法在使用基本面异象 (因子) 预测股票收益的问题中会表现得更好，因此，基于 1985 年 1 月至 2020 年 11 月美股市场的 207 项异象 (因子)，本文采用线性回归、预测组合算法、Lasso 回归、岭回归、弹性网络回归、偏最小二乘回归、支持向量机、梯度提升树、极端梯度提升树、集成神经网络等 10 种机器学习算法，构建股票收益预测模型及投资组合，系统性地运用线性及非线性监督学习算法检验美国市场基本面异象与的股票收益的预测问题。

2011 年 John Cochrane 在美国金融协会主席演讲时 (Cochrane,2011)[19] 提出了三个至关重要的问题, 其一为: “哪些因子是重要的?” 因此，本文计划解决的第二个问题是如果机器学习能够获得更好的投资绩效，哪些因子起了重要作用? 由于时间和算力的限制，我们暂时仅采用 Linear Regression 方法进行检验，但可能由于 ols 自身效果不理想，导致我们检验的因子重要性差异不是很明显，需要进一步使用本文的全部算法对异象 (因子) 重要性进行检验。

本文的正文部分是这样安排的，根据我们的两个研究问题，分为三个部分：第二部分 (2) 为研究设计，阐述模型的训练过程、机器学习算法的简要说明，并阐述数据的预处理步骤，对数据进行简要的描述性统计；第三部分 (3) 为机器学习算法的有效性检验（即探究第一个问题），对算法的样本外预测情况进行说明，利用机器学习算法的预测结果进行投资，分析投资绩效，检验滑动窗口不同窗口期的设定，对模型训练以及投资绩效的影响；第四部分（4）为异象 (因子) 的重要性检验（即探究第二个问题），我们分析了较为重要的因子及其背后可能的成因。

2. 研究设计

2.1. 模型总体设计

我们使用美股市场的 207 个异象 (因子) 值，以收益率为目标进行监督学习算法的训练，标准的函数形式如下：

$$R_{i,t} = f(x_{i,t-1}; \theta) + \epsilon_{i,t} \tag{1}$$

其中，$f(\cdot)$ 定义为参数为 θ 的函数，在本文中为 10 种机器学习算法的函数形式，$R_{i,t}$ 为股票 i 第 t 期的收益，$x_{i,t-1} = (x_{i,t-1,1}, x_{i,t-1,2}, \cdots, x_{i,t-1,N})$ 为公司 i 在第 $t-1$ 期的异象因子向量，$\epsilon_{i,t}$ 为误差项，本文将在下一部分介绍所使用的 10 种机器学习算法。

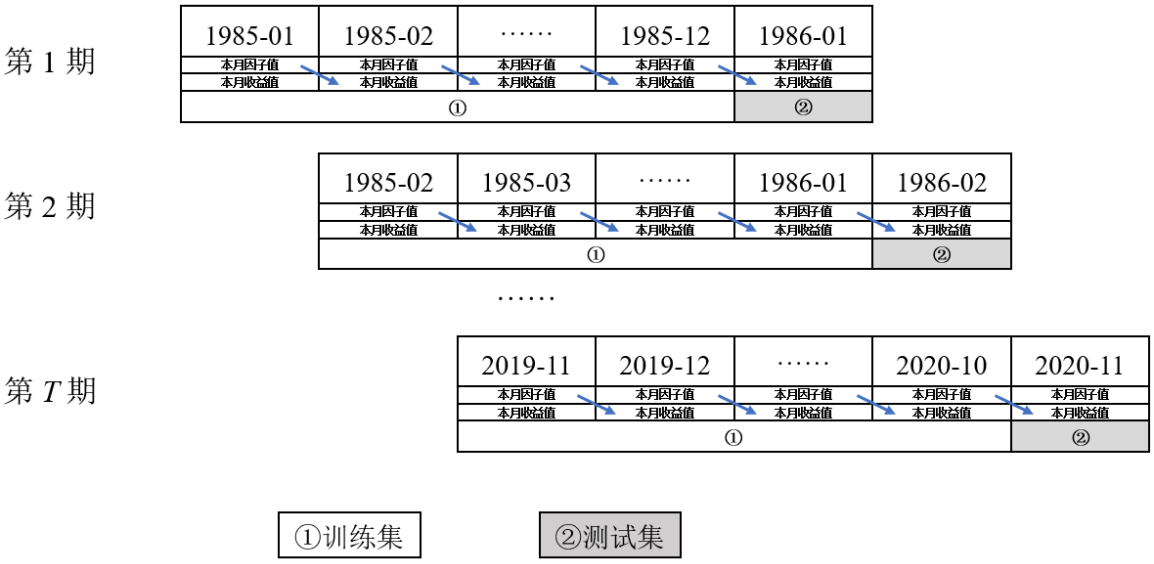


Figure 1: 滑动窗口法训练示意图 [窗口期 =12, 步长 =1]

对于每一个具体的函数形式 $f(\cdot)$，我们将使用历史数据对模型进行训练，采用滑动窗口法划分训练集和测试集，如 Figure 1 所示。模型的训练与测试步骤如下：(1) 当我们的滑动窗口设定窗口期为 12 个月，步长为 1 个月时，当我们要预测 1986 年 1 月的收益数据时，我们利用过去 12 个月（即 1985 年 1 月至 1985 年 12 月）的当月异象（因子）值匹配下月的收益率数据，得到训练集，训练我们的函数形式 $f(\cdot)$，得到模型参数。(2) 使用训练好的模型，将 1985 年 12 月的异象（因子）数据带入模型，得到模型对 1986 年 1 月的股票收益的预测。(3) 重复前两个步骤，共有 419 个训练集和测试集，训练后得到 1986 年 1 月至 2020 年 11 月每个月的每家公司当月股票收益的预测值。

在训练机器学习模型时，针对参数 θ 采用网格调参 (Grid Search) 的方法。首先设定一个初始参数池，然后在训练集上针对每个参数训练得到多空组合的收益，筛选得到最优参数。理论上，随着窗口的滑动，每个模型的最优参数也会随之改变。由于每一期网格搜索的计算成本较高，本文参考李斌等（2019）[17] 的做法，仅在第一个滑动窗口的训练集中进行调参。在此后的窗口滑动过程中，模型参数保持不变。所以，在本研究中，不同时期的模型最优参数是固定的，即为第一个滑动窗口训练所得的最优参数。

针对训练样本做一点说明，模型的训练集（以 window=12 月为例）是 12 个月的每家公司的 219 个异象与次月收益的组合，即假设每个月有 2000 家公司，则我们的训练集共有 2000×12 = 24000 个样本，共有 2000×12 = 24000 个收益率值，共有 2000×12×219 = 5256000 个异象（因子）值。而并不是用每一家公司的 12 个月数据训练一次模型，将所有公司的结果集合。这样做的好处是便于模型的训练，同时也便于我们探究机器学习模型是否能够识别异象与收益率之间的线性、非线性关系。

2.2. 机器学习算法说明

本部分将对本文所使用的 10 个线性与非线性监督学习算法的内容进行简要说明。

2.2.1. 线性回归 (Linear Regression, LR)

线性回归（Linear Regression）在因变量（Y）和一个或多个自变量（X）之间建立一种线性关系，$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n = \beta^T X$。对于给定的样本 X_i，因变量真实值

为 Y_i，预测值为 $\hat{Y}_i = \beta^T X_i$。将函数的损失函数定义为平方损失函数，

$$Loss(\beta) = \frac{1}{m}\sum_{i=1}^{m}(\widehat{Y_i} - Y_i)^2 \tag{2}$$

其中 m 为有效样本数量。通过最小化目标损失函数，即 $argmin \sum_{i=1}^{m}(\widehat{Y_i} - Y_i)^2$，即可求解方程最优拟合系数 β 值。本文将其简称为 OLS 算法，并作为基准模型与各类机器学习算法进行对比。

2.2.2. 预测组合模型 (Forecast Combination, FC)

预测组合模型 (Forecast Combination, FC) 的主要思想是通过对自变量不同的预测模型进行加权平均构成整体预测模型，其核心在于加权系数。目前常用的加权方法有算术平均法、最优权数法和方差倒数法等。在本文中，FC 模型由以单一因子作为自变量的 OLS 模型构成，具体构建和预测方式如下：

1. 在训练集上，分别训练单个因子为自变量的最小二乘模型 OLS_1，$OLS_2, \mathrm{OLS}_3, \ldots, \mathrm{OLS}_n$。
2. 在测试集上，运用所得的 n 个模型分别预测收益率，并取所有模型预测的均值作为最终预测。

尽管单变量 OLS 模型训练成本小，但其样本外预测不稳定；而将其组合后的 FC 模型能够提升样本外预测的稳定性。预测组合模型在金融研究中已有应用，如 Rapach et al.(2010)[20] 运用预测组合模型组合了基于各个因子的单变量回归模型，根据 FC 模型预测所构成的投资组合绩效优于基于 OLS 模型所构建的投资组合。

2.2.3. 岭回归 (ridge regression, Ridge)

岭回归 (ridge regression, Ridge) 同样是线性模型，其在标准线性回归损失函数的基础上加入 L_2 范数正则化项，即

$$Loss(\beta) = \frac{1}{m}\sum_{i=1}^{m}(\widehat{Y_i} - Y_i)^2 + \gamma \|\beta\|_2^2 \tag{3}$$

其中 $\gamma > 0$。通过在损失函数中增加 L_2 范数正则化项约束参数以降低模型复杂度，可以防止过拟合，并增强了模型的样本外预测能力。通过最小化损失函数即可求解方程最优拟合系数 β 值。

2.2.4. Lasso 回归 (Least absolute shrinkage and selection operator, Lasso)

Lasso 回归（Least absolute shrinkage and selection operator，Lasso）也是线性模型，其在损失函数中增加一个正则项 L_1 范数，即向量中各元素的绝对值之和作为正则化项，即：

$$Loss(\beta) = \frac{1}{m}\sum_{i=1}^{m}(\widehat{Y_i} - Y_i)^2 + \gamma||\beta||_1 \tag{4}$$

其中 $\gamma > 0$。相较于 L_2 范数而言，L_1 范数更易获得稀疏解，因此 Lasso 也常被用于高维数据的特征筛选。在回归过程中，越重要的特征对应的系数绝对值越大，而与输出变量相关性越低的特征，系数就越接近于 0。

在金融领域，Feng et al.(2020)[21] 运用 LASSO 进行的因子筛选方法，成功选出了具有更高统计显著性的因子类别；Messmer and Audrino(2017)[22] 则在美国市场采用了 LASSO 算法的变种 Adaptive LASSO 从 68 个公司特征因子中筛选出了 14 个公司特征，并保有了不亚于 68 个公司特征的解释能力。

2.2.5. 弹性网络回归 (ElasticNet Regression, Elastic)

弹性网络回归（ElasticNet Regression, Elastic）综合了 Lasso 和岭回归两种算法，同时使用 L1 和 L2 正则化，其损失函数可以表示为：

$$Loss(\beta) = \frac{1}{m}\sum_{i=1}^{m}(\widehat{Y_i} - Y_i)^2 + \alpha * l1_{ratio} * ||\beta||_1 + 0.5 * \alpha * (1 - l1_{ratio}) * ||\beta||_2^2 \tag{5}$$

不同于 Lasso 将部分系数清零的做法，弹性网络回归鼓励在高度相关变量时的群体效应。当多个特征和另一个特征相关的情形下弹性网络往往能够取得较好的预测效果，Lasso 倾向于随机选择其中一个特征，而弹性网络更倾向于选择两个特征。此外，上述回归正则化方法（岭回归、Lasso 回归和 Elastic 回归）往往在数据集中的变量具有高纬度以及变量间存在多重共线性时能够保持较好的预测效果。

2.2.6. 偏最小二乘回归 (Partial Least Squares Regression，PLS)

偏最小二乘回归（Partial Least Squares Regression，PLS）方法在普通多元回归的基础上结合了主成分分析 (Principal components analysis, PCA) 和典型相关分析 (Canonical Correlation Analysis，CCA) 的思想，以解决回归分析中自变量多重共线性的问题。

考虑存在 m 个自变量 $x_1, x_2, \ldots x_m$。偏最小二乘回归首先在自变量集合中提出第一主成分 t_1（t_1 是 $x_1, x_2, \ldots x_m$ 的线性组合，且尽可能多地提取原自变量集中的变异信息），然后建立因变量与 t_1 的回归方程，如果方程已达到满意的精度，则算法中止。否则继续第二主成分的提取，直到达到满意的精度。若最终提取了 r 个成分 $t_1, t_2, \ldots t_r$，偏最小二乘回归将通过建立因变量与 $t_1, t_2, \ldots t_r$ 的回归式，并还原为因变量与原自变量的回归方程。

Light et al.(2017) 选用 PLS 检验了公司特征对股票截面收益的预测能力。同样，Gu et al.(2020)[16] 也发现 PLS 算法在美国市场股票收益预测要好于传统的 OLS 算法。

2.2.7. 支持向量机 (support vector machine)

支持向量机 (support vector machine) 是一种常用的分类算法，但当标签为连续值时，也可用于拟合回归问题。模型通过寻求结构化风险最小来提高学习机泛化能力，实现经验风险和置信范围的最小化，从而达到在统计样本量较少的情况下，亦能获得良好统计规律的目的。通俗来讲，其基本模型定义为特征空间上的间隔最大的线性分类器，即支持向量机的学习策略便是间隔最大化，最终可转化为一个凸二次规划问题的求解。SVM 算法拥有低泛化误差，可以解决高维问题等优点，但同时模型的预测结果对参数和核函数的选取非常敏感，模型的主要参数包括：

1. 核函数类型（kernel），备选核函数类型包含线性核（linear），高斯核（rbf），多项式核（poly）等；
2. 惩罚因子（C）；
3. 核函数对应的核系数（γ）。

2.2.8. 梯度提升树 (Gradient Boosting Decision Tree，GBDT)

梯度提升树 (Gradient Boosting Decision Tree，GBDT) 是一种迭代的决策树算法，由多棵决策树组成，综合所有树的预测作为最终预测。该算法的核心在于每棵树学习之前所

有树的残差；而为了消除残差，模型在残差减少的梯度 (Gradient) 方向上建立一个新的模型。因此在 GBDT 中，每个新树的建立是为了使得之前模型的残差沿梯度方向减少。此外，决策树进行分支时以最小化平方误差为标准，对每一个特征，每一个阈值进行穷举以寻求最优的分割点。对于训练集 $Train = \{(x_1, y_1), (x_1, y_1), \ldots, (x_N, y_N)\}$，构建梯度提升树 $f_M(x)$ 的算法流程如下：

1. 初始化 $f_o(x) = 0$；
2. 对于 $m = 1, 2, \ldots, M$，计算其预测残差：

$$r_{mi} = y_i - f_{m-1}(x_i) \qquad i = 1, 2, \ldots, N \tag{6}$$

3. 对于 N 个残差学习得到一个回归树 $T(x; \emptyset_m)$；
4. 更新 $f_m(x) = f_{m-1}(x) + Train(x; \emptyset_m)$；
5. 最终得到梯度提升树：$f_M(x) = \sum_{m=1}^{M} T(x; \emptyset_m)$。

GBDT 在较少的调参时间情况下能够获得相对较高的预测准确率。同时由于使用的损失函数相对稳健，GBDT 算法对异常值的鲁棒性非常强，算法的主要参数包含：每个弱学习器的权重缩减系数，即学习率（δ）；弱学习器的最大迭代次数（N）；决策树最大深度 (maxdep)。现有研究中，Krauss et al.(2017)[23] 运用梯度提升树算法进行标准普尔 500 指数成分股的运动方向预测，并根据预测结果构建投资组合，绩效明显好于市场投资组合。

2.2.9. 极端梯度提升树 (Extreme Gradient Boosting, Xgboost)

Boosting 算法以集成弱分类器的方式提高预测的稳定性和准确性，是机器学习领域中被广泛使用的算法 (Wu et al., 2008[24])。代表性算法是由 Chen and Guestrin(2016)[25] 提出的极端梯度提升树 (Extreme Gradient Boosting, Xgboost)。Xgboost 通过 Boosting 算法来聚合作为基学习器的 CART 树算法。因此 Xgboost 具有 Boosting 算法的优点，但训练成本低且结果更为精确。具体算法流程如下：

1. 基于训练集构建第一棵 CART 回归树，并计算出模型的残差；
2. 通过第一步计算出的残差训练下一棵 CART 回归树，再次进行残差计算；

3. 重复 (2.) 直到最大迭代次数。

单就上述步骤而言，Xgboost 同 GBDT 较为相似，都是以前一次预测的残差作为下一步训练目标，且模型参数类型设置与 GBDT 算法基本一致，但两者存在以下不同： 在拟合目标的设置上，Xgboost 在 GBDT 的基础上加入了正则化项，使模型具有更好的泛化能力； 在计算残差过程中，XGBoost 在 GBDT 的基础上加入了二阶导数，提升了残差估计的准确性； 在 CART 回归树的叶节点划分时，GBDT 算法采用的是最小化均方差，而 XGBoost 算法则是最大化上述方程中的正则化项。这些细节处理的不同使得 XGBoost 独立于 GBDT 算法，在常见的机器学习任务中取得了不俗的成果。

2.2.10. 神经网络集成模型 (Ensemble Artificial Neural Network，EN-ANN)

神经网络集成模型（Ensemble Artificial Neural Network，EN-ANN）是一种基于人工神经网络（Artificial Neural Network，ANN）的集成学习算法，主要针对单一 ANN 模型由于初始化问题而产生的预测不稳定现象。其核心是构建多个不同初始化状态下的简单神经网络，使得模型集成更多的可能性以提升预测结果的稳定性。在训练过程中神经网络集成模型对每一个 ANN 都进行单独的训练和优化实现单个模型的近似最优化。本文中神经网络集成模型的算法流程如下：

1. 初始化每一个神经网络 $NN_1, NN_2, NN_3, \ldots, NN_K$；
2. 在训练集中，对每一个神经网络模型以最小化均方误差作为优化目标进行参数拟合；
3. 选取训练集上均方误差前 50% 的神经网络作为预测池构建神经网络集成模型，并输出预测池的平均预测值。

神经网络集成模型在有效提升模型稳定性的同时能对由于初始化数值选取存在偏差而陷入局部最优解的训练器进行甄别，筛选出相对稳定的训练器进行预测，在很大程度上能够提升了预测的鲁棒性。

2.3. 数据说明与预处理

本文选取 1985 年 1 月至 2020 年 11 月美国纽约证券交易所（NYSE）、美国证券交易所（AMSE）、全国证券交易商自动报价系统协会（NASDAQ）的上市公司为研究样本，数

据为月度频率。美国公司的收益率数据（月度）获取自 CRSP 数据库，我们选取 “Holding Period Return” 的回报率。异象（因子）数据使用 Chen and Zimmermann(2021) [11] 开源的美股市场 207 个异象（因子）数据[1]。

初始的股票池为三个市场的所有股票，我们剔除了金融、保险和房地产行业（SIC 代码为 6000-6799 的公司），剔除了交易状态为 “halted” 或 “suspended” 或 “unknown” 的公司。同时，数据存在着一定的缺失值，我们采取如下步骤进行处理：(1) 对于异象 (因子) 数据, 若存在 Inf 值，将其替换成空值 (nan)，然后使用每个异象 (因子) 的横截面均值替换空值 (nan), 若此步填充后仍然存在空值 (nan), 即将其替换为 0。(2) 若 t 月的收益数据存在缺失值，则直接删去 t 月的所有异象 (因子) 数据与收益数据。

在剔除异常交易状态的股票、金融、保险和房地产行业的股票、处理好缺失值后，1985 年 1 月至 2020 年 11 月的有效样本为 835961 条，对其进行描述性统计如下：

[样本数据的描述性统计表, 见附录][A.6]

因子的描述性统计显示，不同因子的取值在数量级及分布上存在显著差异，可能导致预测偏差，比如： 量级较大的特征在预测时占据主导地位； 数量级的差异会引起部分机器学习算法迭代收敛速度减慢。由此，本文将训练集数据标准化，假设这些样本来自某一均值为 0，方差为 1 的随机变量。标准化方式为：

$$X_{scale} = \frac{X - \overline{X}}{\sigma_X} \tag{7}$$

其中，$\overline{X}$ 和 σ_X 分别是变量 X 的均值和标准差。

3. 机器学习算法的有效性检验

3.1. 算法的样本外预测情况 (模型评估)

为了评估在 10 种机器学习算法下，超额异象 (因子) 预测股票未来收益的预测性能，我们依照 Gu et al. (2020)[16] 的做法，用下式计算样本外 R^2:

$$\mathcal{R}^2_{oos} = 1 - \frac{\sum_{(i,t)}(r_{i,t+1} - \hat{r}_{i,t+1})^2}{\sum_{(i,t)} r^2_{i,t+1}} \tag{8}$$

[1]Chen and Zimmermann 的开源数据于此网站下载：https://www.openassetpricing.com/data/

这样构造 $\mathcal{R}^2_{oos}$ 的好处在于，对于个股的分析，用历史平均值预测未来的超额股票回报时，其更偏向正向。将窗口期设定为 12 个月的情况下训练模型得到的 $\mathcal{R}^2_{oos}$ 如下表所示：

Table 1: 机器学习算法的 $\mathcal{R}^2_{oos}$[window=12]

OLS	FC	Ridge	Lasso	EN	PLS	SVR	EN-ANN	XGBoost	GBDT
-5.6E+22	-3.3E+6	-340.9%	-6.0%	-7.0%	-3.6E+05	-9.21%	-13.3%	-8.27%	-10.08%

由表所示，我们可以发现线性机器学习算法的 $\mathcal{R}^2_{oos}$ 十分大，而非线性机器学习算法 $\mathcal{R}^2_{oos}$ 较小，说明非线性机器学习算法更能刻画异象 (因子) 与收益之间的关系。

3.2. 美国市场的投资绩效

我们依据模型预测出的收益率，在每一个月的截面上对股票排序并分 10 组，分别构建多头组合 (等权持有前 10% 的股票)、空头组合 (等权持有后 10% 的股票)、多空对冲组合 (多头-空头) 的年化收益率，并计算每个组合的夏普比率，如下表所示:

Table 2: 投资组合绩效 (window：12months)

	多头组合		多空组合		空头组合	
	Mean(%)	夏普比率	Mean(%)	夏普比率	Mean(%)	夏普比率
OLS	20.02%	1.0337	18.94%	1.1962	1.08%	0.0351
	(5.57)		(6.76)		(0.29)	
FC	19.89%	0.9271	16.56%	0.6928	3.34%	0.1375
	(5.15)		(4.11)		(0.84)	
Ridge	20.31%	1.0504	19.52%	1.2255	0.79%	0.0209
	(5.66)		(6.91)		(0.21)	
Lasso	22.14%	1.1265	21.70%	1.2960	0.44%	0.0035
	(6.04)		(7.26)		(0.12)	
Elastic	21.79%	1.1157	21.63%	1.3006	0.17%	-0.0098

	(5.97)		(7.28)		(0.04)	
PLS	21.23%	1.1000	19.29%	1.1387	1.94%	0.0759
	(5.92)		(6.74)		(0.51)	
SVR	21.04%	1.2594	20.91%	1.3437	0.13%	-0.0107
	(6.72)		(7.59)		(0.03)	
EN-ANN	21.13%	1.0643	19.01%	1.3330	2.12%	0.0848
	(5.71)		(7.57)		(0.01)	
XGBoost	22.46%	1.1000	22.81%	1.4308	-0.36%	-0.0320
	(5.99)		(8.48)		(-0.09)	
GBDT	21.47%	1.1861	19.58%	1.0349	1.89%	0.0628
	(5.98)		(8.11)		(0.01)	

观察分析 Table 1 可以发现:(1) 同为线性模型，除了 FC 模型外，线性机器学习算法 (Ridge,Lasso,ElasticNet,PLS) 均能获得较基准 OLS 回归更高的多空组合收益，且带有惩罚项的线性机器学习算法 (Ridge,Lasso,Elastic) 有高于基准 OLS 回归的夏普比率。(2) 非线性机器学习算法 (SVR,EN-ANN,XGBoost,GBDT) 均能获得较基准 OLS 回归更高的多空组合收益，分别提升了 10.4%,0.3%,20.4%,3.4%, 除 GBDT 外，夏普比率也分别提升了 12.33%,11.44%,19.6%, 显示了异象 (因子) 间非线性模式的存在，其中 XGBoost 模型表现地尤为优秀。(3) 括号中展示的是投资收益的 Newey and West(1987)[26]t 值，可以发现，多空组合的收益均为 1% 显著。观察多头组合的收益，其远高于空头组合，表明多空组合的收益主要来源于多头头寸，表现最好的基于 XGBoost 预测而构建的投资组合的收益较基准 OLS 提升了 12.18%。

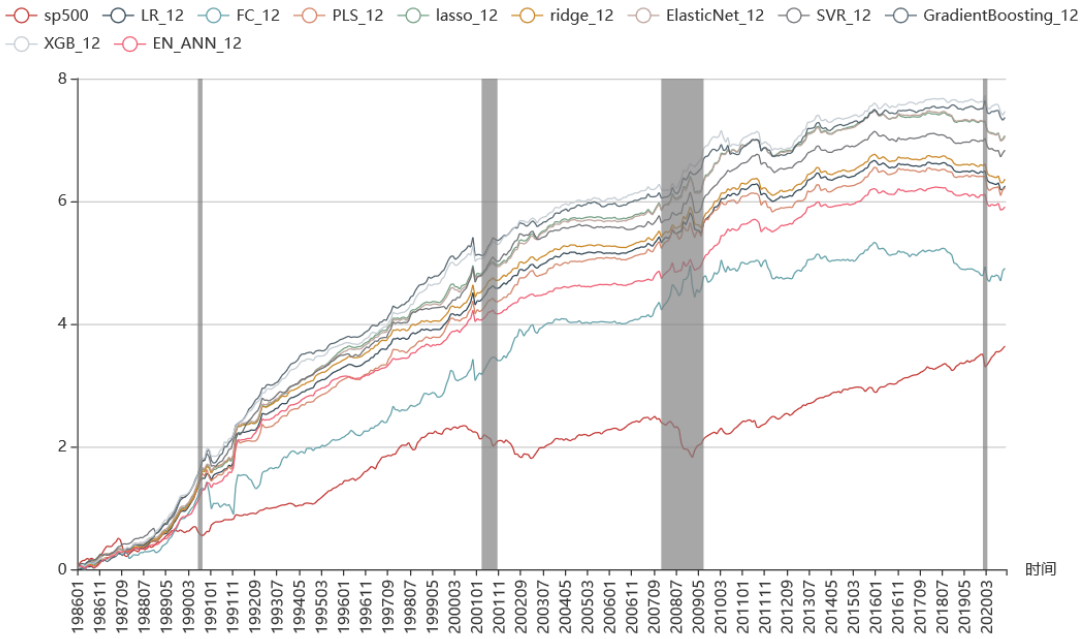


Figure 2: 累计收益率曲线 (window=12)

绘制 10 种机器学习算法所构建的多空投资组合的累计对数收益率，与 SP500 的累计收益率曲线做对比可以发现，整体来看由机器学习算法构建的投资组合的累计收益率曲线更为平缓，波动较小，且每一种模型的累计收益率均大于 SP500 的累计对数收益率，非线性机器学习算法的收益明显优于机器学习算法。图中阴影部分为 National Bureau of Economic Research 确定的衰退时期，在这几个阶段，我们可以发现 SP500 的收益率呈现下跌，而我们机器学习算法所构造的投资组合的累计收益率是波动上升的，可以说明机器学习算法的有效性。

3.3. 检验划分不同窗口期对模型训练与投资绩效的影响

为了验证模型的稳定性，我们也检验了窗口期 (即训练集长度) 为 3 个月，24 个月时的模型表现，首先展示多头组合、空头组合、多空对冲组合 (多头-空头) 的年化收益率与每个组合的夏普比率，如 Table 2 所示:

Table 3: 投资组合绩效 (window：3months)

	多头组合		多空组合		空头组合	
	Mean(%)	夏普比率	Mean(%)	夏普比率	Mean(%)	夏普比率
OLS	18.87%	1.1151	13.74%	0.9321	5.12%	0.2286
	(6.18)		(6.08)		(1.38)	
FC	18.41%	1.0176	14.12%	0.5723	4.30%	0.1653
	(5.65)		(3.85)		(1.07)	
Ridge	19.32%	1.1436	14.57%	0.9789	4.75%	0.2110
	(6.32)		(6.28)		(1.29)	
Lasso	20.55%	1.0316	17.33%	1.1870	3.21%	0.1339
	(6.49)		(6.60)		(0.86)	
Elastic	20.27%	1.1631	17.05%	0.9245	3.23%	0.1332
	(6.35)		(5.95)		(0.86)	
PLS	19.21%	1.0919	15.34%	0.8146	3.88%	0.1577
	(6.02)		(5.43)		(0.99)	
SVR	18.92%	1.2119	13.73%	0.8444	5.20%	0.2145
	(6.65)		(5.26)		(1.28)	
EN-ANN	18.11%	1.0386	11.96%	0.8864	6.15%	0.2698
	(5.65)		(5.72)		(1.61)	
XGBoost	21.13%	1.1691	19.18%	0.9962	1.96%	0.0660
	(6.40)		(6.33)		(0.46)	
GBDT	21.47%	1.1861	19.58%	1.0349	1.89%	0.0628
	(6.49)		(6.56)		(0.44)	

观察多空组合的收益可以发现，训练期较短时，(1) 线性机器学习算法 (Ridge,Lasso, ElasticNet,PLS) 均能获得较基准 OLS 回归更高的多空组合收益，其中 Lasso 和 Elastic 尤

为优秀，较基准 OLS 回归提升了 26.1%,24.1%。(2) 非线性机器学习算法表现得差异较大，其中 SVR 与 EN-ANN 投资绩效较差，可能的原因是由于训练期过短，样本数据过少，这两个模型效果不理想，但 XGBoost 与 GBDT 表现较好，多空收益较基准 OLS 回归提升了 40%,42.5%。(3) 综合所有模型来看，夏普比率的差异不大，window=3 与 window=12 相比，模型训练效果较差。

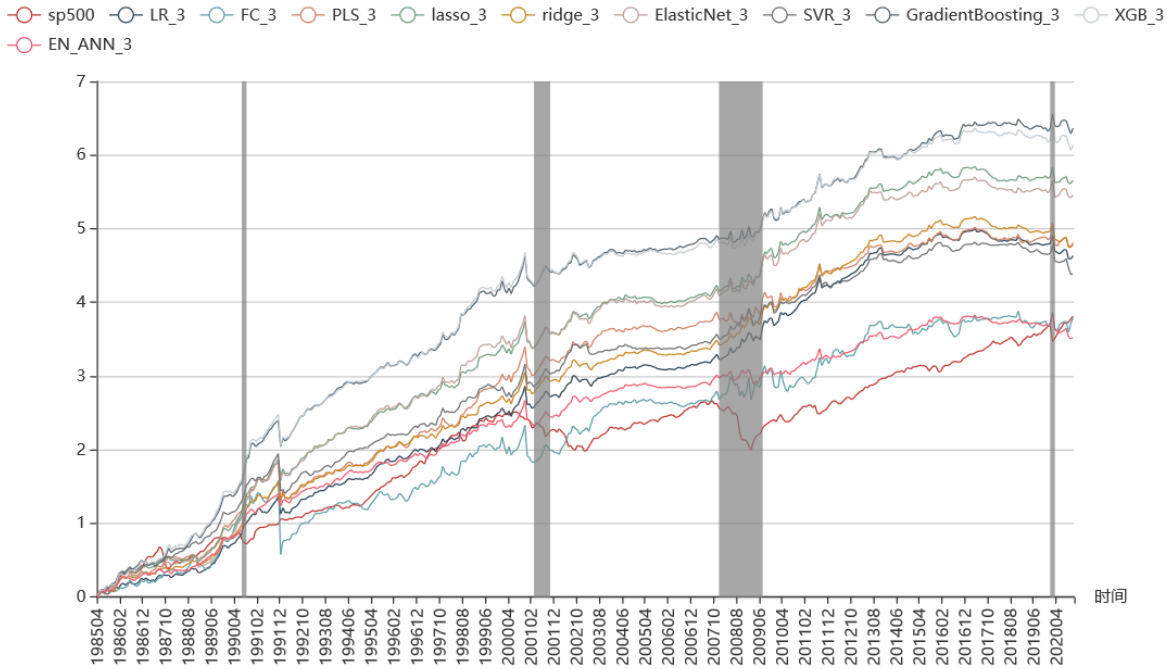


Figure 3: 累计收益率曲线 (window=3)

从累计收益率曲线可以发现，非线性机器学习算法 (SVR,XGBoost,GBDT) 与线性机器学习算法 (lasso,ridge,ElasticNet) 的累计收益率均能够跑赢 SP500 指数的累计收益率，特别是 Lasso 和 ElasticNet 算法，是表现最好的两个线性机器学习算法，特别的，非线性机器学习算法 EN-ANN 表现较差，与前文所述的其多空组合收益较差一致，我们判断是由于训练期过短，样本数据不足以使复杂的神经网络训练出较好的结果。

Table 4: 美国市场的投资绩效 (window：24months)

	多头组合		多空组合		空头组合	
	Mean(%)	夏普比率	Mean(%)	夏普比率	Mean(%)	夏普比率
OLS	21.24%	1.1405	21.04%	1.5533	0.20%	-0.0078
	(6.13)		(8.44)		(0.05)	
FC	19.35%	1.0352	15.42%	0.7284	3.93%	0.1617
	(5.68)		(4.28)		(0.94)	
Ridge	21.48%	1.1564	21.22%	1.5422	0.26%	-0.0046
	(6.21)		(8.32)		(0.07)	
Lasso	13.61%	0.8210	4.96%	0.3105	8.65%	0.4668
	(4.49)		(1.81)		(2.51)	
Elastic	15.92%	0.8574	10.50%	0.5908	5.42%	0.2738
	(4.63)		(3.29)		(1.53)	
PLS	19.50%	1.0456	15.76%	0.7409	3.73%	0.1521
	(5.74)		(4.34)		(0.89)	
SVR	20.41%	1.3265	19.29%	0.9719	1.13%	0.0321
	(7.12)		(5.54)		(0.25)	
EN-ANN	21.79%	1.0782	20.56%	1.5217	1.23%	0.0434
	(5.69)		(8.25)		(0.32)	
XGBoost	20.76%	1.1046	20.05%	1.2679	0.71%	0.0164
	(5.94)		(7.10)		(0.17)	
GBDT	21.10%	1.1244	19.25%	1.2409	1.85%	0.0705
	(6.05)		(7.03)		(0.46)	

将窗口期扩展至 24 个月，与先前规律所不一致的是，OLS 基准回归的多空组合年化收益是除 Ridge 外最大的，同时，非线性机器学习算法的效果要显著优于线性机器学习算

法，可以说明异象 (因子) 之间的确存在着非线性关系，同时，线性模型所带来的投资绩效不比非线性模型的要差。

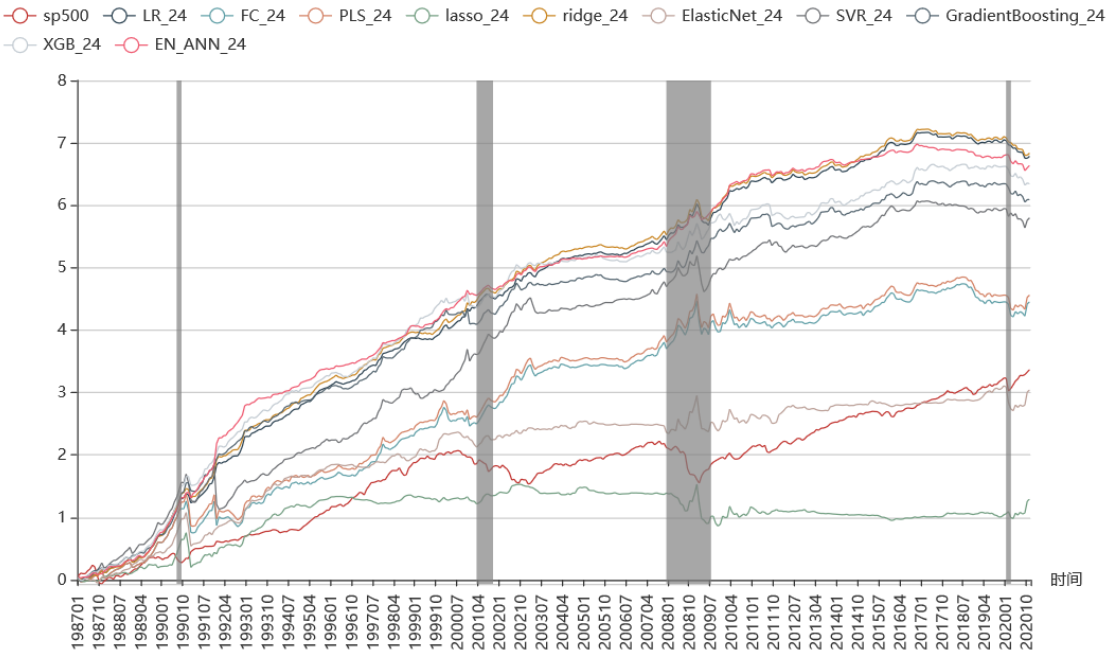
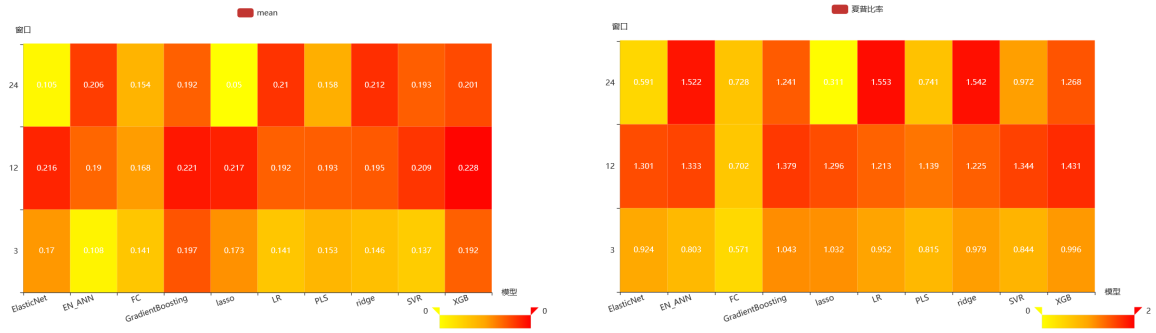


Figure 4: 累计收益率曲线 (window=24)

观察累计收益率曲线可以发现，Laso,ElasticNet 均有低于 SP500 收益率曲线的阶段，不同的模型收益率曲线的高低排序，与 window=3，window=12 相比有很大差异，Ridge 和 LR 和复杂的 EN-ANN 模型同处于最高层，但线性模型 (Ridge,LR) 的累计收益率曲线更为平缓、稳定，而神经网络模型起伏波动较大，可能是由于过拟合现象的存在。



(a) 年化收益率比较 (b) 夏普比率比较

Figure 5: 不同窗口期的投资绩效比较

我们利用 window=3、12、24 三种情况下的年化收益率与夏普比率数据绘制热图，以可视化他们的差异情况，观察热图，纵向来看，当我们扩展训练期从 1 个季度至 1 年时，夏普比率与年化收益率均有显著提高，但继续扩展至 2 年时，不同的模型有不同的表现，复杂的非线性机器学习算法有明显的提升或维持了现状，但线性机器学习算法除了 Ridge、LR、FC 外，效果均变差。横向来看，非线性机器学习算法总体优于线性机器学习算法，但 Ridge 和 LR 仍然是一个特例，可能是我们的异象 (因子) 与收益间的线性关系已经可以带来较高的收益，但这仍然需要我们继续延长窗口期，或者进行变换样本等方式来检验这个结论。

Table 5: 不同窗口期下机器学习算法的 $\mathcal{R}^2_{oos}$

	3months	12months	24months
OLS	-1.33E+23	-5.60E+22	-7.64E+22
FC	-1.25E+06	-3.26E+06	-2.05E+06
Ridge	-99.54%	-340.89%	-167.64%
Lasso	-17.23%	-6.01%	-0.09%
Elastic	-10.51%	-7.00%	-0.13%
PLS	-1.36E+05	-3.56E+05	-2.46E+05
SVR	-28.45%	-9.21%	-1.32%
EN-ANN	-50.62%	-13.30%	-7.56%
XGBoost	-12.42%	-8.27%	-2.61%
GBDT	-14.66%	-10.08%	-3.57%

通过对比计算得出的不同窗口期下机器学习算法的 $\mathcal{R}^2_{oos}$，我们可以发现，虽然三个 window 的 $\mathcal{R}^2_{oos}$ 均为负数，但总体来看，扩展训练期会提高模型的拟合水平，$\mathcal{R}^2_{oos}$ 有明显地向 0 接近的形势。纵向来看，线性机器学习算法 (LR,FC,PLS,Ridge) 的 $\mathcal{R}^2_{oos}$ 非常大，说明其预测的效果并不是十分理想，而非线性机器学习算法 (SVR,EN-ANN,XGBoost,GBDT) 的 $\mathcal{R}^2_{oos}$ 较小，说明预测预测效果较好。综合 $\mathcal{R}^2_{oos}$ 与前文所述的投资绩效，可以发现 Ridge 的综合多空收益较大，但其 $\mathcal{R}^2_{oos}$ 较小，说明 Ridge 方法作为线性机器学习算法，预测性能

并没有达到最优，但其选股能力较为优秀。

4. 异象 (因子) 的重要性检验

本部分探究我们的第二个研究问题：“哪个因子更重要?”，由于时间与算力的限制，仅采用 Linear Regresion(LR) 一种检验方法，计算除去某一单个因子后的年化收益的损失来衡量该因子的重要性，将收益损失最大的因子重要性记为 100%，剩余因子的重要性数值则按照收益损失最大值折算得到。本文筛选出重要性数值位于前 20 的因子，如条形图所示:

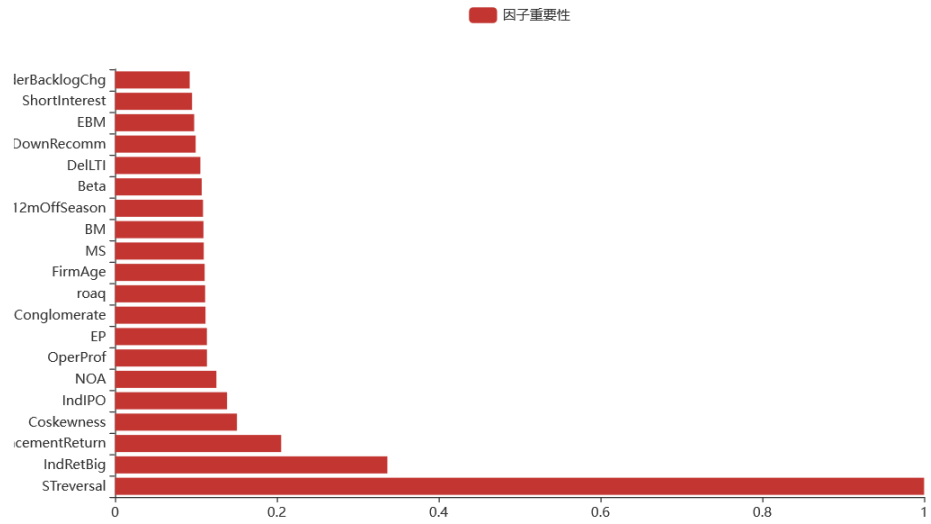


Figure 6: Top 20 重要的异象 (因子)

这 20 个重要的因子按顺序分别为:STreversal, IndRetBig, Announcement Return, Coskewness, IndIPO, NOA, OperProf, EP, Conglomerate, roaq, FirmAge, MS, BM, Mom12mOffSeason, Beta, DelLTl, DownRecomm, EBM, ShortInterest, OrderBacklogChg，我们参考Appendix B对这 20 个异象 (因子) 的详细定义，可以发现，STreversal, IndRetBig, AnnouncementReturn,Coskewness 均与收益率直接或间接相关，可以发现与预测收益率相关的重要因子可能直接地与收益相关，但这个结论仍需要检验，进一步可以使用本文的其余 9 种机器学习算法，以相同的思想进行检验，查看哪些因子被纳入 Top20 的次数较多，可以较有把握地说明那些因子更为重要。

References

[1] B. Graham, D. L. F. Dodd, S. Cottle, et al., Security analysis, Vol. 452, McGraw-Hill New York, 1934.

[2] S. Basu, Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis, The journal of Finance 32 (3) (1977) 663–682.

[3] R. W. Banz, The relationship between return and market value of common stocks, Journal of financial economics 9 (1) (1981) 3–18.

[4] E. F. Fama, K. R. French, The cross-section of expected stock returns, the Journal of Finance 47 (2) (1992) 427–465.

[5] M. M. Carhart, On persistence in mutual fund performance, The Journal of finance 52 (1) (1997) 57–82.

[6] R. Novy-Marx, The other side of value: The gross profitability premium, Journal of financial economics 108 (1) (2013) 1–28.

[7] K. Hou, C. Xue, L. Zhang, Digesting anomalies: An investment approach, The Review of Financial Studies 28 (3) (2015) 650–705.

[8] E. F. Fama, K. R. French, A five-factor asset pricing model, Journal of financial economics 116 (1) (2015) 1–22.

[9] C. R. Harvey, Y. Liu, A census of the factor zoo, Available at SSRN 3341728 (2019).

[10] K. Hou, C. Xue, L. Zhang, Replicating anomalies, The Review of Financial Studies 33 (5) (2020) 2019–2133.

[11] A. Y. Chen, T. Zimmermann, Open source cross-sectional asset pricing, Critical Finance Review, Forthcoming (2021).

[12] R. Roy, S. Shijin, A six-factor asset pricing model, Borsa Istanbul Review 18 (3) (2018) 205–217.

[13] F. Jiang, G. Tang, G. Zhou, Firm characteristics and chinese stocks, Journal of Management Science and Engineering 3 (4) (2018) 259–283.

[14] S. Gu, B. Kelly, D. Xiu, Autoencoder asset pricing models, Journal of Econometrics 222 (1) (2021) 429–450. doi:10.1016/j.jeconom.2020.07.009.

[15] C. Zhang, Asset Pricing and Deep Learning (Sep. 2022). arXiv:2209.12014.

[16] S. Gu, B. Kelly, D. Xiu, Empirical Asset Pricing via Machine Learning, The Review of Financial Studies 33 (5) (2020) 2223–2273. doi:10.1093/rfs/hhaa009.

[17] 李斌, 邵新月, 李王月阳, 机器学习驱动的基本面量化投资研究, 中国工业经济 (8) (2019) 61–79. doi:10.19581/j.cnki.ciejournal.2019.08.004.

[18] D. H. Wolpert, The lack of a priori distinctions between learning algorithms, Neural computation 8 (7) (1996) 1341–1390.

[19] J. H. Cochrane, Presidential address: Discount rates, The Journal of finance 66 (4) (2011) 1047–1108.

[20] D. E. Rapach, J. K. Strauss, G. Zhou, Out-of-sample equity premium prediction: Combination forecasts and links to the real economy, The Review of Financial Studies 23 (2) (2010) 821–862.

[21] G. Feng, S. Giglio, D. Xiu, Taming the factor zoo: A test of new factors, The Journal of Finance 75 (3) (2020) 1327–1370.

[22] M. Messmer, F. Audrino, The (adaptive) lasso in the zoo-firm characteristic selection in the cross-section of expected returns, Working paper (2017).

[23] C. Krauss, X. A. Do, N. Huck, Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the s&p 500, European Journal of Operational Research 259 (2) (2017) 689–702.

[24] Q. Wu, C. J. Burges, K. M. Svore, J. Gao, Ranking, boosting, and model adaptation, Tech. rep., Technical report, Microsoft Research (2008).

[25] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

[26] W. K. Newey, K. D. West, Hypothesis testing with efficient method of moments estimation, International Economic Review (1987) 777–787.

Appendix A. 样本数据的描述性统计

	count	mean	std	min	50%	max
AbnormalAccruals	1935	-0.006	0.058	-0.826	-0.006	0.778
Accruals	1935	0.005	0.085	-1.197	0.005	1.038
AccrualsBM	1935	0.722	0.051	0.005	0.722	1.000
Activism1	1935	15.289	0.242	12.227	15.290	18.200
Activism2	1935	13.256	1.101	7.211	13.255	42.112
AdExp	1935	0.036	0.039	0.000	0.038	1.290
AgeIPO	1935	27.838	6.083	1.252	27.693	117.972
AM	1935	12.391	26.410	0.050	12.580	788.313
AnalystRevision	1935	1.001	0.650	-12.984	1.001	16.838
AnalystValue	1935	0.885	0.277	-2.424	0.891	5.907
AnnouncementReturn	1935	0.002	0.048	-0.452	0.002	0.625
AOP	1935	-6.533	52.547	-2162.342	-6.563	5.625
AssetGrowth	1935	-0.174	0.786	-28.508	-0.174	0.854
Beta	1935	0.650	0.500	-1.542	0.578	5.093
BetaFP	1935	0.656	0.431	0.001	0.613	3.645
BetaLiquidityPS	1935	-0.026	0.245	-2.014	-0.024	2.122
BetaTailRisk	1935	0.598	0.223	-0.729	0.596	2.614
betaVIX	1935	0.000	0.010	-0.115	0.000	0.114
BidAskSpread	1935	0.014	0.021	0.000	0.008	0.372
BM	1935	-0.353	0.486	-5.096	-0.350	2.908
BMdec	1935	3.601	27.667	-28.422	2.910	1017.007
BookLeverage	1935	-10.146	51.146	-1762.908	-10.175	436.998
BPEBM	1935	1.769	51.827	-359.167	1.682	2007.383

BrandInvest	1935	-1904.519	375.182	-15789.007	-1905.982	-10.475
Cash	1935	0.099	0.075	0.000	0.100	0.904
CashProd	1935	91.576	1319.564	-4299.749	87.925	56473.076
CBOperProf	1935	-0.001	0.000	-0.001	-0.001	-0.001
CF	1935	-0.052	0.991	-32.534	-0.051	4.279
cfp	1935	0.090	0.890	-18.520	0.085	21.140
ChangeInRecommendation	1935	-0.017	0.239	-2.112	-0.017	2.114
ChAssetTurnover	1935	0.236	32.251	-725.898	0.264	769.918
ChEQ	1935	-1.258	2.447	-99.017	-1.252	-0.079
ChForecastAccrual	1935	0.492	0.174	0.000	0.490	1.000
ChInv	1935	-0.003	0.038	-0.689	-0.003	0.637
ChInvIA	1935	-310349 688	213961 0217399	-408255 03316422	-309701 727	610562 34554270
ChNAnalyst	1935	-0.099	0.030	-0.618	-0.099	-0.012
ChNNCOA	1935	-0.002	0.101	-1.472	-0.003	1.366
ChNWC	1935	-0.003	0.081	-1.164	-0.003	1.025
ChTax	1935	0.000	0.015	-0.390	0.000	0.248
CitationsRD	1935	0.056	0.001	0.043	0.056	0.084
CompEquIss	1935	0.609	1.066	-2.434	0.612	32.629
CompositeDebtIssuance	1935	-0.534	0.627	-7.088	-0.542	5.654
ConsRecomm	1935	-0.308	0.052	-0.830	-0.308	-0.075
ConvDebt	1935	-0.067	0.178	-1.000	-0.033	0.000
CoskewACX	1935	0.172	0.200	-0.563	0.164	0.909
Coskewness	1935	0.273	0.287	-0.916	0.281	1.306
CredRatDG	1935	-0.019	0.072	-0.864	-0.017	-0.002
CustomerMomentum	1935	0.011	0.012	-0.166	0.011	0.205

DebtIssuance	1935	-0.405	0.296	-1.000	-0.412	0.000
DelBreadth	1935	0.122	0.356	-3.934	0.120	6.559
DelCOA	1935	0.002	0.062	-0.900	0.002	0.899
DelCOL	1935	0.004	0.069	-0.887	0.004	1.001
DelDRC	1935	0.007	0.002	-0.014	0.007	0.055
DelEqu	1935	-0.016	0.078	-1.129	-0.016	1.084
DelFINL	1935	-0.026	0.095	-1.231	-0.026	1.197
DelLTI	1935	-0.030	0.099	-1.221	-0.029	1.036
DelNetFin	1935	0.006	0.105	-1.162	0.005	1.168
DivInit	1935	0.031	0.121	0.000	0.013	1.000
DivOmit	1935	-0.006	0.053	-0.956	-0.003	0.000
DivSeason	1935	0.443	0.294	0.000	0.441	1.000
DivYieldST	1935	0.589	0.549	0.000	0.587	3.000
dNoa	1935	-0.085	0.575	-20.695	-0.084	1.313
DolVol	1935	-1.933	2.407	-9.919	-1.880	6.551
DownRecomm	1935	-0.359	0.107	-0.843	-0.359	-0.087
EarningsConsistency	1935	0.131	0.528	-8.449	0.130	12.675
EarningsForecastDisparity	1935	38.031	160.597	-1122.297	37.359	5208.593
EarningsStreak	1935	-0.035	0.151	-5.531	-0.036	0.261
EarningsSurprise	1935	-310082	259935	-693935	-428406	693519
		746865	15273811	483357427	491915	000448200
EarnSupBig	1935	-5927.539	7160.646	-17230.396	-5213.823	0.527
EBM	1935	2.723	51.501	-352.017	2.625	2001.085
EntMult	1935	-15.354	40.621	-1543.817	-15.350	53.778
EP	1935	0.086	0.056	0.001	0.086	1.603
EquityDuration	1935	-443.299	7993.193	-342059.352	-269.138	4387.171

ExchSwitch	1935	-0.010	0.067	-0.945	-0.004	0.000
ExclExp	1935	-0.011	0.175	-2.182	-0.011	1.486
FEPS	1935	9.848	100.266	-9.522	10.223	4282.303
fgr5yrLag	1935	-13.687	2.008	-60.449	-13.688	1.334
FirmAge	1935	-195.781	55.000	-811.722	-195.600	-2.903
FirmAgeMom	1935	0.073	0.060	-0.448	0.073	1.091
ForecastDispersion	1935	-0.148	0.321	-11.051	-0.147	0.000
FR	1935	-0.008	0.053	-1.243	-0.007	0.899
Frontier	1935	-0.046	0.141	-2.646	-0.046	2.243
Governance	1935	-9.181	0.426	-11.379	-9.182	-7.275
GP	1935	0.191	0.068	-0.936	0.191	1.597
GrAdExp	1935	-0.097	0.128	-2.074	-0.099	1.586
grcapx	1935	-1.334	22.663	-644.919	-1.355	358.991
grcapx3y	1935	-1.746	11.666	-379.768	-1.729	166.265
GrLTNOA	1935	0.011	0.092	-1.308	0.011	1.507
GrSaleToGrInv	1935	-6.114	54.537	-2223.692	-6.006	17.376
GrSaleToGrOverhead	1935	0.043	0.946	-6.161	0.042	35.071
Herf	1935	-0.219	0.168	-1.493	-0.217	-0.021
HerfAsset	1935	-0.221	0.131	-1.000	-0.223	-0.032
HerfBE	1935	-0.232	0.334	-9.247	-0.235	-0.028
High52	1935	0.849	0.164	0.063	0.884	1.616
hire	1935	-0.043	0.177	-1.993	-0.042	1.881
IdioRisk	1935	-0.020	0.021	-0.348	-0.015	0.000
IdioVol3F	1935	-0.018	0.018	-0.301	-0.013	0.000
IdioVolAHT	1935	-0.021	0.017	-0.258	-0.017	-0.001
Illiquidity	1935	0.000	0.000	0.000	0.000	0.002

IndIPO	1935	-0.081	0.192	-1.000	-0.038	0.000
IndMom	1935	0.075	0.063	-0.035	0.080	0.209
IndRetBig	1935	0.016	0.022	-0.020	0.015	0.055
IntanBM	1935	0.111	0.298	-3.291	0.108	2.598
IntanCFP	1935	-0.022	0.203	-5.207	-0.024	1.681
IntanEP	1935	0.010	0.227	-5.779	0.008	1.840
IntanSP	1935	-0.221	0.583	-6.855	-0.232	4.121
IntMom	1935	0.065	0.207	-0.784	0.059	3.290
Investment	1935	-1.064	0.654	-11.122	-1.068	8.616
InvestPPEInv	1935	-0.045	0.165	-5.780	-0.044	0.813
InvGrowth	1935	-0.772	0.599	-23.564	-0.775	0.870
IO_ShortInterest	1935	74.370	1.370	46.636	74.369	102.693
iomom_cust	1935	0.739	0.085	0.043	0.738	0.992
iomom_supp	1935	0.769	0.077	0.048	0.768	0.992
Leverage	1935	10.059	20.102	0.003	10.167	610.963
LRreversal	1935	-0.340	0.594	-14.025	-0.342	0.929
MaxRet	1935	-0.048	0.061	-1.211	-0.034	-0.001
MeanRankRevGrowth	1935	2598.124	477.070	385.449	2600.611	4763.221
Mom12m	1935	0.127	0.339	-0.883	0.117	6.564
Mom12mOffSeason	1935	0.011	0.028	-0.183	0.011	0.298
Mom6m	1935	0.056	0.198	-0.795	0.050	2.989
Mom6mJunk	1935	0.058	0.132	-0.658	0.057	2.175
MomOffSeason	1935	-0.012	0.012	-0.139	-0.012	0.088
MomOffSeason06YrPlus	1935	-0.013	0.012	-0.139	-0.013	0.202
MomOffSeason11YrPlus	1935	-0.013	0.009	-0.123	-0.013	0.094
MomOffSeason16YrPlus	1935	-0.014	0.005	-0.075	-0.014	0.029

MomRev	1935	0.570	0.062	0.001	0.570	1.000
MomSeason	1935	0.012	0.045	-0.310	0.012	0.583
MomSeason06YrPlus	1935	0.013	0.035	-0.277	0.013	0.468
MomSeason11YrPlus	1935	0.012	0.029	-0.251	0.013	0.419
MomSeason16YrPlus	1935	0.012	0.024	-0.225	0.012	0.360
MomSeasonShort	1935	0.011	0.079	-0.496	0.010	1.084
MomVol	1935	5.855	0.772	1.025	5.852	9.995
MRreversal	1935	-0.067	0.200	-3.228	-0.063	0.766
MS	1935	3.775	0.224	1.007	3.775	5.986
NetDebtFinance	1935	-0.015	0.064	-0.780	-0.015	0.663
NetDebtPrice	1935	-1.920	1.225	-41.555	-1.918	9.780
NetEquityFinance	1935	-0.010	0.044	-0.745	-0.008	0.367
NetPayoutYield	1935	-0.001	0.025	-0.617	-0.001	0.486
NOA	1935	-0.378	0.627	-19.044	-0.378	2.165
NumEarnIncrease	1935	1.285	1.238	0.000	1.287	7.993
OperProf	1935	0.052	0.490	-16.742	0.052	5.564
OperProfRD	1935	-0.002	0.000	-0.002	-0.002	-0.002
OPLeverage	1935	0.226	0.397	-0.016	0.219	9.868
OptionVolume1	1935	-37589.433	209138.931	-8655377.522	-36277.785	-11169.285
OptionVolume2	1935	-1.447	1.271	-46.694	-1.438	-0.452
OrderBacklog	1935	-0.344	0.052	-2.112	-0.344	-0.008
OrderBacklogChg	1935	-0.005	0.021	-0.478	-0.005	0.492
OrgCap	1935	-0.032	0.128	-0.836	-0.032	3.598
OScore	1935	-0.174	0.031	-0.904	-0.174	0.000
PatentsRD	1935	0.119	0.002	0.090	0.119	0.170
PayoutYield	1935	0.155	0.068	0.001	0.154	2.574

PctAcc	1935	2.476	36.229	-620.124	2.425	1042.457
PctTotAcc	1935	-3.098	39.889	-1180.752	-3.033	605.859
PredictedFE	1935	-0.047	0.008	-0.106	-0.047	-0.001
PriceDelayRsq	1935	0.414	0.310	0.002	0.332	1.000
PriceDelaySlope	1935	1.846	131.412	-1978.953	0.658	4274.398
PriceDelayTstat	1935	1.840	1.291	-2.179	1.850	5.535
ProbInformedTrading	1935	0.286	0.010	0.194	0.286	0.439
PS	1935	5.375	0.343	1.419	5.375	8.197
RD	1935	0.031	0.010	0.000	0.031	0.350
RDAbility	1935	-1.091	0.051	-2.150	-1.092	0.045
RDcap	1935	0.008	0.016	0.000	0.008	0.596
RDIPO	1935	-0.007	0.055	-0.929	-0.003	0.000
RDS	1935	-72.593	1433.258	-25644.704	-69.370	31377.821
realestate	1935	0.000	0.045	-0.408	0.000	0.588
Recomm_ShortInterest	1935	0.797	0.031	0.218	0.796	0.950
ResidualMomentum	1935	-0.026	0.264	-1.513	-0.025	1.213
retConglomerate	1935	0.012	0.013	-0.085	0.012	0.137
ReturnSkew	1935	-0.097	0.818	-4.088	-0.068	3.708
ReturnSkew3F	1935	-0.091	0.749	-3.822	-0.069	3.510
REV6	1935	-0.030	0.032	-1.018	-0.030	0.201
RevenueSurprise	1935	0.268	3.681	-47.085	0.264	115.685
RIO_Disp	1935	3.468	0.218	1.063	3.468	5.000
RIO_MB	1935	2.596	0.161	1.000	2.596	4.947
RIO_Turnover	1935	3.214	0.211	1.049	3.214	4.998
RIO_Volatility	1935	3.327	0.188	1.016	3.327	4.998
roaq	1935	0.007	0.087	-0.515	0.007	3.123

RoE	1935	0.345	6.666	-36.744	0.249	272.519
sfe	1935	0.075	0.046	-1.203	0.075	0.360
ShareIss1Y	1935	-0.401	3.292	-128.110	-0.346	0.985
ShareIss5Y	1935	-4.044	24.426	-967.502	-4.028	0.993
ShareRepurchase	1935	0.362	0.334	0.000	0.318	1.000
ShareVol	1935	-0.188	0.140	-1.000	-0.193	0.000
ShortInterest	1935	-21008.465	17733.383	-300431.875	-20091.190	-1.625
sinAlgo	1935	1.000	0.000	1.000	1.000	1.000
skew1	1935	-0.073	0.016	-0.371	-0.073	0.057
SmileSlope	1935	0.001	0.035	-0.555	0.001	0.557
SP	1935	1.567	3.434	-1.333	1.609	108.522
Spinoff	1935	0.025	0.109	0.000	0.012	1.000
std_turn	1935	-0.101	0.694	-27.174	-0.089	-0.001
SurpriseRD	1935	0.047	0.060	0.000	0.047	1.000
tang	1935	0.598	0.020	0.189	0.598	0.903
Tax	1935	0.873	4.288	-56.006	0.882	142.991
TotalAccruals	1935	-0.022	0.153	-4.051	-0.022	1.394
TrendFactor	1935	0.194	0.044	0.116	0.189	0.278
UpRecomm	1935	0.347	0.106	0.084	0.347	0.840
VarCF	1935	-1.772	16.099	-629.953	-1.891	0.000
VolMkt	1935	-0.136	1.029	-35.527	-0.054	0.000
VolSD	1935	-3.826	29.020	-977.293	-0.502	-0.002
VolumeTrend	1935	-0.006	0.017	-0.064	-0.005	0.055
XFIN	1935	0.003	0.238	-1.244	0.004	7.896
zerotrade	1935	1.728	3.052	0.000	0.318	17.370
zerotradeAlt1	1935	1.817	3.416	0.000	0.156	19.108

zerotradeAlt12	1935	1.662	2.852	0.000	0.457	16.571
STreversal	1935	-0.967	9.955	-129.659	-0.551	59.768
Price	1935	-2.707	0.945	-10.788	-2.725	1.960
Size	1935	-11.995	1.887	-18.389	-11.900	-5.315

Appendix B. 重要性排名前 20 的异象 (因子) 的定义

- STreversal：Stock return (ret) over the previous month.
- IndRetBig：Average monthly return (ret) of the 30% largest companies by market value of equity in the same Fama-French 48 industry. Exclude the largest 30% of companies for IndRetBig (not to compute the anomaly!)
- AnnouncementReturn：Get announcement date for quarterly earnings from IBES (fpi = 6). AnnouncementReturn is the sum of (ret - mktrf + rf) from two days before an earnings announcement to 1 days after the announcement.
- Coskewness：Signal is the sample counterpart of $E[\tilde{r}_{it}\tilde{r}_{mt}^2]/(SD[\tilde{r}_{it}]SD[\tilde{r}_{mt}]^2$ where $\tilde{r}_{it}$ is the de-meaned stock return and $\tilde{r}_{mt}$ is the de-meaned market excess return. Signal is computed using the past year of daily data, and using the NYSE CRSP VW index for the market (dsia), with returns continuously compounded. See code for details.
- IndIPO：1 if IPO in the past 6-36 months. 0 otherwise. IPO dates are taken from Jay Ritter's IPO data available at: http://bear.warrington.ufl.edu/ritter/ipodata.htm. Missing IPO dates imply IndIPO = 0
- NOA：Difference between operating assets and operating liabilities, scaled by lagged total assets. Operating assets are total assets (at) minus cash- and short-term investments (che), operating liabilities are total assets minus long-term debt (dltt), minority interest (mib), deferred charges (dc) and book equity (ceq).
- OperProf：Revenue (revt) minus cost (cogs) - administrative expenses (xsga) - interest expenses (xint), scaled by book value of equity (ceq). Exclude smallest size tercile.
- EP:ib / lag(market value of equity, 6 months). NYSE stocks only. Exclude if EP $<$ 0. Lag simulates the Dec 31 market equity used in original paper

- Conglomerate:Identify conglomerate firms as those with multiple OPSEG or BUSSEG entries in the Compustat segment data (and require that at least 80% of firm's total assets are covered by segment data). Compute monthly stock return at the 2-digit SIC level for stand-alone (non-conglomerate) firms only, and match those returns to conglomerates' segments. Compute weighted conglomerate return as the industry return of stand-alone companies, weighted with a conglomerate's total sales in each industry.

- roaq：This is like a more timely version of the other profitability measures. Interestingly, they don't cite Fama French 2006, nor Novy Marx 2013. MP have a very slightly different formulation.

- FirmAge：OP uses special NYSE archive data that we lack.

- MS：MS is only evaluated for low BM firms and comes from combining three signals related to profitability and cash flow, two signals related to income volatility, and three signals related to investment.

- BM：Log of annual book equity (ceq) over market equity.

- Mom12mOffSeason：This acronym has a different form than the other off season Heston and Sadka ones because its behavior is distinct. The other off season signals behave like long-term reversal.

- Beta：Coefficient of a 60-month rolling window regression of monthly stock returns minus the riskfree rate on market return minus the risk free rate (ewretd - rf). Exclude if estimate based on less than 20 months of returns.

- DelLTl：Difference in investment and advances (ivao) between years t-1 and t, scaled by average total assets (at) in years t-1 and t.

- DownRecomm：Keep fpi = 1. Binary variable equal to 1 if mean analyst earnings forecast for the next quarter (meanest) has improved over the previous month, and 0 otherwise.

- EBM：(ceq + che - dltt - dlc - dc - dvpa+ tstkp) / (mve_c + che - dltt - dlc - dc - dvpa+ tstkp). Exclude if price less than 5.

- ShortInterest：Short-interest from Compustat (shortint) scaled by shares outstanding (shrout). Short-interest data are available bi-weekly with a four day lag. We use the mid-month observation to make sure data would be available in real time. OP uses Asquith and Meulbroek's database, which covers all of NYSE and AMEX. We're unsure of the quality of our Compustat data, especially since it is missing many values pre-2003. However, the missing pre-2003 is mostly NASDAQ. According to Rapach, Ringgenberg, and Zhou 2016, Compustat added the short interest to their dataset in 2014.

- OrderBacklogChg：Define normalized order backlog as order backlog (ob) divided by average total assets (at) in years t-1 and t. Exclude if order backlog is 0. Signal is normalized order backlog minus normalized order backlog one year ago.