

分析师：

徐寅

xuyinsh@xyzq.com.cn

S0190514070004

## 西学东渐--海外文献推荐系列之九十九

2020 年 10 月 22 日

## 报告关键点

目前预测金融市场未来收益通常采用线性多因子模型，本文选择信息透明、流动性高、交易成本低的美国商品期货市场，分别采用线性模型和决策树模型对未来收益进行预测。首先，本文发现决策树模型的预测能力远超线性模型，并由此推断数据结构可能存在非线性的特征，采用线性模型估计可能得到错误结果；其次，本文实证结果显示决策树模型对样本内收益变化的预测能力可以达到 56.21%，这与有效市场假说认为的无法基于历史信息预测未来收益相违背，并由此推断过往文献可能高估了市场的有效性；最后，本文发现采用决策树模型构建的投资组合在绝对收益和风险调整后收益上均优于线性模型，且该结果在不同时间跨度均成立，证明了结果的稳健性。

## 相关报告

《西学东渐--海外文献推荐系列之九十八》

《西学东渐--海外文献推荐系列之九十七》

《西学东渐--海外文献推荐系列之九十六》

## 投资要点

- 西学东渐，是指从明朝末年到近代，西方学术思想向中国传播的历史过程。西学东渐不仅推动了中国在科学技术和思想文化方面的发展，也有力地促进了社会与政治的大变革。在今天，西学东渐仍有其重要的现实意义。作为 A 股市场上以量化投资为研究方向的卖方金融工程团队，在平日的工作中，常常深感海外相关领域的研究水平之高、内容之新。而这也促使本文们通过大量的材料阅读，去粗取精，将认为最有价值的海外文献呈现在您的面前！
- 目前预测金融市场未来收益通常采用线性多因子模型，本文选择信息透明、流动性高、交易成本低的美国商品期货市场，分别采用线性模型和决策树模型对未来收益进行预测。首先，本文发现决策树模型的预测能力远超线性模型，并由此推断数据结构可能存在非线性的特征，采用线性模型估计可能得到错误结果；其次，本文实证结果显示决策树模型对样本内收益变化的预测能力可以达到 56.21%，这与有效市场假说认为的无法基于历史信息预测未来收益相违背，并由此推断过往文献可能高估了市场的有效性；最后，本文发现采用决策树模型构建的投资组合在绝对收益和风险调整后收益上均优于线性模型，且该结果在不同时间跨度均成立，证明了结果的稳健性。

团队成员：

风险提示：文献中的结果均由相应作者通过历史数据统计、建模和测算完成，在政策、市场环境发生变化时模型存在失效的风险。

请务必阅读正文之后的信息披露和重要声明



## 目录

1、引言 .....	3 -
2、实证方法 .....	5 -
2.1、Logit 回归 .....	5 -
2.2、LASSO 回归 .....	6 -
2.3、决策树模型 .....	6 -
3、实证研究 .....	8 -
3.1、数据 .....	8 -
3.2、预测因子 .....	9 -
3.3、参数 .....	11 -
3.4、主要发现 .....	11 -
4、探讨 .....	16 -
5、结论 .....	17 -
图表 1、决策树 .....	7 -
图表 2、4 周预测收益和 4 周实际收益的平方相关系数 .....	12 -
图表 3、不同模型的样本外表现 1：机器学习 vs. 线性模型 vs. 多头组合 .....	13 -
图表 4、不同模型的样本外表现 2：机器学习 vs. 线性模型 vs. 多头组合 .....	14 -
图表 5、预测因子重要性 .....	14 -
图表 6、样本外相关性：袋装树 vs. 线性模型 vs. 多头组合 .....	15 -
图表 7、样本外相关性：提升树 vs. 线性模型 vs. 多头组合 .....	15 -

## 报告正文

## 基于非线性模型预测商品期货收益

## 文献来源：

Struck, C. , & Cheng, E. . (2019). The cross-section of commodity returns: a non-parametric approach. The Journal of Financial Data Science.

## 推荐原因：

目前预测金融市场未来收益通常采用线性多因子模型，本文选择信息透明、流动性高、交易成本低的美国商品期货市场，分别采用线性模型和决策树模型对未来收益进行预测。首先，本文发现决策树模型的预测能力远超线性模型，并由此推断数据结构可能存在非线性的特征，采用线性模型估计可能得到错误结果；其次，本文实证结果显示决策树模型对样本内收益变化的预测能力可以达到 56.21%，这与有效市场假说认为的无法基于历史信息预测未来收益相违背，并由此推断过往文献可能高估了市场的有效性；最后，本文发现采用决策树模型构建的投资组合在绝对收益和风险调整后收益上均优于线性模型，且该结果在不同时间跨度均成立，证明了结果的稳健性。

## 我们的思考：

目前线性多因子模型已经广泛应用于金融市场资产定价，对因子挖掘以及单一因子分析的研究也已非常深入，但线性模型具有严格的参数假设，从而削弱了其收益预测能力。本文采用了决策树模型来规避这些假设，并在模型中纳入了因子之间的交互作用以及数据的非线性结构，从而使模型预测能力有较大提升。

## 1、引言

目前已有大量学术文献探讨过金融市场收益的可预测性，但关注的重点主要是单个收益因子和 Fama-French (1992) 提出的线性模型。在研究时，多数文章作出了三个严格的参数假设：(1) 用一个简单的函数描述预期收益和预测因子之间的关系，(2) 预测因子之间不相关，(3) 模型通常只包含有限的预测因子，这三点假设削弱了模型对收益的预测能力。

本文对比了标准线性模型和决策树模型的收益预测能力，其中决策树模型不需要进行特定的参数假设。具体而言，我们研究了 28 种商品期货的月度收益率与 59 个预测因子之间的关系。选择商品期货是因为这种资产更接近有效市场的假定：信息透明、高流动性、低交易成本。同时，与股票市场相比，商品期货市场不存在幸存者偏差。

本文有三个主要的研究结论：首先，采用决策树模型可以预测 3.74% 的月度收益，尽管这个数字很小，但与标准线性模型相比预测能力已经提高了 10 倍；其次，线性模型和决策树模型对样本内数据的预测能力分别为 0.58% 和 56.21%，对样本外数据的预测能力分别为 0.33% 和 3.74%；第三，不论从绝对收益还是风险

请务必阅读正文之后的信息披露和重要声明

- 3 -

调整后收益来考虑，采用决策树模型构建的投资组合表现都优于采用线性模型构建的投资组合。

有效市场假说认为，在有效市场中人们无法仅基于过去信息预测未来收益，除非采用的因子代表了某种风险。目前许多文献已经分离出很多单因子，可以分别解释一部分收益变动，例如 Fama 和 French (1993, 2016)、Longstaff (1995)、Asness, Moskowitz 和 Pedersen (2013)、Novy-Marx (2013)、Farhi 和 Gabaix (2016)。但大多数文章仅采用线性模型研究了一部分互相独立的因子。而我们的研究结果表明，在商品期货市场中，决策树模型的预测能力可以达到线性模型的 10 倍。

当我们将简单线性模型转换为决策树模型后，预测能力发生了巨大的变化，这一现象说明数据中具有特定的结构。因此，先前基于线性模型的研究可能会导致错误的定性解释。例如，多数文献对某一因子进行研究，发现其在统计学上显著，从而得出这个因子反映了一定风险的结论。然而，当加入其他因子后会发现，由于因子间的相互作用，该因子可能不再显著，也并不代表任何风险。

本文对线性模型的有效性提出了质疑，并认为一些异常的收益波动可能仅仅是统计上的结果。目前许多研究都把重点放在避免样本选择偏差和多重检验问题上，例如 Lo 和 MacKinlay (1990)、Black (1993)、Kothari, Shanken 和 Sloan (1995)、Jorion 和 Goetzmann (1999)、Schwert (2003)、Bailey 和 López de Prado (2014)、McLean 和 Pontiff (2016)、Harvey, Liu 和 Zhu (2016) 以及 Harvey 和 Liu (2019)。但我们认为不合理的模型选择也会导致最后的结论错误。

通常对因子进行检验是看该因子是否可以提升模型的解释能力，但本文提出了另一种检验新因子的方法：检验将因子加入现有因子集合后模型的解释力是否提升，这种方法将因子之间的相互作用和非线性也考虑在内。我们的观点与 Feng, Giglio 和 Xiu (2019) 类似，他们提出了一种线性的两阶段模型来筛选因子。因为我们发现因子之间的相互作用和非线性较强，我们认为线性模型对于预测未来收益并不适用。

我们的研究方法最近采用决策树方法研究股票市场收益的文章有相似之处。Gu, Kelly 和 Xiu (2020) 采用了多种回归模型，对 94 个独立因子和 74 个行业虚拟变量在美股市场进行了研究，最后发现由动量、流动性和波动性相关因子构成的因子集合对于预测资产收益更为有效，而且非线性模型可以更好预测未来收益。他们也证明了相比标准线性回归模型，决策树模型在对未来收益的预测能力上有较大提升。

尽管考虑了数据的非线性特征，我们的结果依然只能解释约 50% 的样本内收益变化。这可能是由于两个原因：首先，尽管我们的因子集合已经大于过往文献，但仍有进一步扩展的空间，例如可以进一步加入与宏观因子的交互项；其次，出于简化计算流程的考虑，我们构建的两个决策树模型都没有使用最复杂的结构。

我们在这两方面做的简化都表明模型的收益预测能力还有进一步提升的空间。但我们认为我们的模型已经接近预测能力上限：从标准线性模型转为决策树模型使样本内的收益预测能力从 0.58% 提高到 56.21%，但样本外的预测能力仅从 0.33% 提高到 3.74%，因此提高样本内预测能力对样本外预测能力的影响是有限的。我们认为即使将样本内预测能力由 56.21% 提高到 100%，也不会使样本外预测能力

有显著增加。

虽然我们的研究结果确实表明，以往的线性模型大大低估了收益的可预测性，但是使用了决策树模型后，依然有超过 95% 的样本外收益变化是无法被预测的。从这个角度来看，我们的研究结果与有效市场假说一致，即利用过去的信息无法预测未来的收益。

后文结构如下：第二章阐述了我们的实证方法，第三章描述了我们做出的假设以及实证结果，第四章对主要研究成果进行了解释，第五章为结论。

## 2、实证方法

我们可以将收益预测问题理解为在  $t$  时刻对  $t+1$  时刻商品期货  $i$  的收益  $E_t(\gamma_{i,t+1})$  进行预测， $g$  是关于一系列预测因子  $X_{i,t} = (x_{i,t,1}, x_{i,t,2}, \dots, x_{i,t,n})$  的函数，如下式：

$$E_t(\gamma_{i,t+1}) = g(x_{i,t,1}, x_{i,t,2}, \dots, x_{i,t,n}) = g(X_{i,t})$$

当我们考察某一预测因子是否具有长期预测能力时，我们可以把收益预测问题转化为对收益方向的预测，这个过程称为离散化。通过这一步可以降低极值的影响，参见 Fayyad 和 Irani (1993) 以及 Garcia 等人 (2013)。我们将资产  $i$  的未来收益  $r_{i,t+1}$  按照正负划分为 1 或 0。对预期收益变量  $\gamma_{i,t+1}$  定义如下：

$$\gamma_{i,t+1} = \begin{cases} 1, & \text{if } r_{i,t+1} > 0 \\ 0, & \text{otherwise} \end{cases}$$

接下来我们研究采用不同模型构建函数  $g$ （包括 Logit 回归、LASSO 回归以及决策树模型），并比较不同模型的预测能力。

### 2.1、Logit 回归

我们采用 Logit 模型作为对照。Logit 模型以对数线性函数的形式对因变量  $\gamma_{i,t+1}$  和预测因子  $X_{i,t}$  的关系进行建模。具体来说，Logit 模型中预期收益为正的的概率为：

$$\ln \left( \frac{\Pr(\gamma_{i,t+1} = 1 | X_{i,t}, \beta)}{1 - \Pr(\gamma_{i,t+1} = 1 | X_{i,t}, \beta)} \right) = \beta_1 x_{i,t,1} + \beta_2 x_{i,t,2} + \dots + \beta_N x_{i,t,N}$$

其中  $E_t(\gamma_{i,t+1}) = \Pr(\gamma_{i,t+1} = 1 | X_{i,t}, \beta)$ ， $\beta$  是  $(N \times 1)$  的参数向量，通过最大似然估计计算得到。由于  $\gamma_{i,t+1}$  是二元随机变量，所以  $\gamma_{i,t+1}$  的概率密度函数是：

$$f(\gamma_{i,t+1} | X_{i,t}, \beta) = p_{i,t}^{\gamma_{i,t+1}} (1 - p_{i,t})^{1 - \gamma_{i,t+1}}$$

通过概率密度函数，我们可以推导出 Logit 模型下预期收益为正的的概率：

$$p_{i,t} = \Pr(\gamma_{i,t+1} = 1 | X_{i,t}, \beta) = \frac{\exp(X'_{i,t} \beta)}{1 + \exp(X'_{i,t} \beta)}$$

其中  $\beta$  选择可以使上式概率最大的参数：

$$\max_{\beta} \mathcal{L} = \sum_{i=1}^I \sum_{t=1}^T \gamma_{i,t+1} \ln(p_{i,t}) + (1 - \gamma_{i,t+1}) \ln(1 - p_{i,t})$$

我们将在后文中的实证部分，采用样本外数据计算 Logit 模型的收益预测能



力。

## 2.2、LASSO 回归

Logit 模型的一个主要问题是，它需要我们首先挑选出有效的预测因子。为了简化运算，在实证部分我们使用整个预测因子集合来估计 Logit 模型，但是因子数量较多可能会导致模型存在多重共线性。为了系统性地筛选变量，我们对 Logit 模型进行了改进，采用 LASSO 方法进行回归。LASSO 是一种修正后的对数线性回归，将筛选变量作为回归目标之一。它在 Logit 的对数似然函数中加入了一个惩罚项，使贡献近乎为 0 的预测因子的  $\beta$  值为 0，从而实现变量筛选。LASSO 模型的对数似然函数形式为：

$$\max_{\beta} \mathcal{L} = \sum_{i=1}^I \sum_{t=1}^T y_{i,t+1} \ln(p_{i,t}) + (1 - y_{i,t+1}) \ln(1 - p_{i,t}) - \lambda \sum_{n=1}^N |\beta_n|$$

其中  $\lambda$  是一个比例参数，用于调节惩罚项对模型的影响，通常在 0 和 1 之间进行选择。

## 2.3、决策树模型

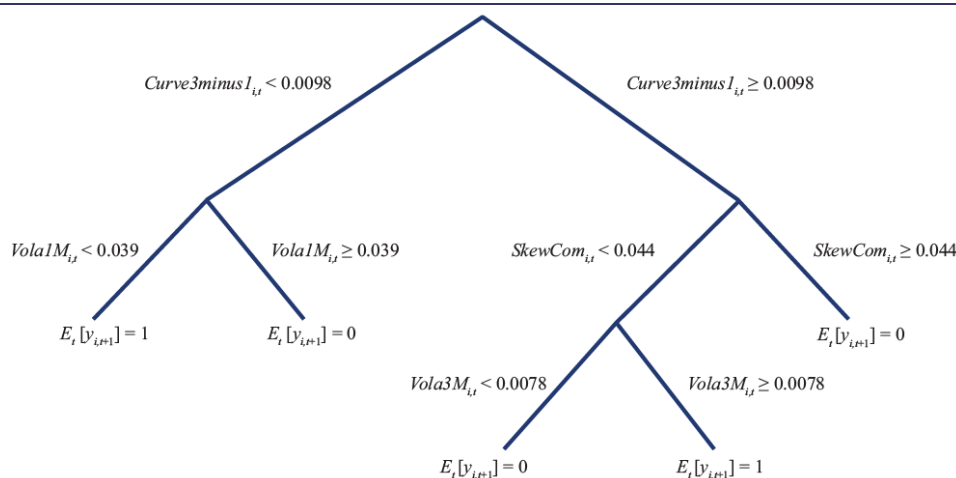
前两种方法的主要问题是它们没有考虑数据的非线性结构和预测因子之间的相互作用。决策树模型可以解决这个问题，目前也已应用在许多领域。本文使用了两种决策树模型：提升树（Boosted Tree）和袋装树（Bagged Tree）。我们首先描述了如何构建一个样本树，再对两种模型进一步解释。

图表 1 是一个决策树的示例。因为很难在全局范围内对决策树进行优化，所以我们采用了一个简化算法，叫做贪婪算法（Greedy Algorithm）。从顶部开始，每个节点  $j$  都与一种特征  $R_j$  相关联，并且可以进一步划分为两组（ $R'_j$  和  $R''_j$ ）。首先，我们在每个节点  $j$  上随机选择一个预测因子集合的子集，然后从这个子集中选择一个预测因子  $n$  及其划分区间的值  $s$ ，使得被错分的  $y_{i,t+1}$  最少。对于每一个观察值，可能落入  $R'_j$  组或  $R''_j$  组，而最终  $E_t(y_{i,t+1})$  的值是由组内较多的观察值决定。例如，如果在  $R'_j$  组中，观察值为 1 的数量比 0 多，那么  $E_t(y_{i,t+1}) = 1, \forall y_{i,t+1} \in R'_j$ 。最小化预测误差的公式如下：

$$\min_{n,s} \mathcal{L}_j = \frac{1}{N_{R'_j(n,s)}} \sum_{y_{i,t+1} \in R'_j(n,s)} 1(y_{i,t+1} \neq E_t(y_{i,t+1})) + \frac{1}{N_{R''_j(n,s)}} \sum_{y_{i,t+1} \in R''_j(n,s)} 1(y_{i,t+1} \neq E_t(y_{i,t+1}))$$

其中  $N_{R'_j(n,s)}$  是落入  $R'_j(n,s)$  组的  $y_{i,t+1}$  数量。对于  $y_{i,t+1}$ ，当  $x_{i,t,n} > s$  时分为  $R'$  组，当  $x_{i,t,n} < s$  时为  $R''$  组。模型会持续对观察值进行分组，直到组中的观察值小于一定数量或最后一组的观察值均为相同特征。

图表 1、决策树



资料来源：The Journal of Financial Data Science，兴业证券经济与金融研究院整理

图表 1 中的决策树由 4 个预测变量构成：3 月与 1 月期货价格之差（Curve3minus1）、过去 1 月收益的波动（Vola1M）、过去 3 个月收益的波动（Vola3M）和过去 260 个交易日的收益偏度（SkewCom）。按照 Curve3minus1 将初始集合分为两组。对左边节点采用 Vola1M 进一步分组，对右边节点采用 SkewCom 进一步分组，对其中一组更进一步采用 Vola3M 划分。

我们以一个示例展示如何计算最后每组的  $E_t(\gamma_{i,t+1})$ 。对于 Curve3minus1 < 0.0098 且 Vola1M < 0.039 的组，我们对落入该组的所有  $\gamma_{i,t+1}$  进行统计：如果多数  $\gamma_{i,t+1}$  的值为 1，则该组  $E_t(\gamma_{i,t+1}) = 1$ ；如果多数  $\gamma_{i,t+1}$  的值为 0，则  $E_t(\gamma_{i,t+1}) = 0$ 。该例对应的函数  $g(X_{i,t})$  如下：

$$\begin{aligned}
 E_t(\gamma_{i,t+1}) = & \beta_1 \times 1(\text{Curve3minus1} < 0.0098)1(\text{Vola1M} < 0.039) \\
 & + \beta_2 \times 1(\text{Curve3minus1} < 0.0098)1(\text{Vola1M} \geq 0.039) \\
 & + \beta_3 \times 1(\text{Curve3minus1} \geq 0.0098)1(\text{SkewCom} < 0.044) \\
 & 1(\text{Vola3M} < 0.0078) \\
 & + \beta_4 \times 1(\text{Curve3minus1} \geq 0.0098)1(\text{SkewCom} < 0.044) \\
 & 1(\text{Vola3M} \geq 0.0078) \\
 & + \beta_5 \times 1(\text{Curve3minus1} \geq 0.0098)1(\text{SkewCom} \geq 0.044)
 \end{aligned}$$

### 2.3.1、袋装树 (Bagged Trees)

传统决策树方法的一个主要缺点是稳定性低，即采用相同的数据集重复生成单个树，每次的结果可能不同。除了稳定性之外，决策树方法还会产生数据碎片。每次分割后节点上的观测值都会相应减少，平均来看每个路径只包含了一部分预测因子，这意味着最终预测是基于预测因子子集做出的。因此，我们可能在生成树时意外地忽略了有价值的预测因子。

目前已有文献提出了一些改进方法来解决上述问题，最常用的是集成方法，装袋算法 (Bagging) 就是一种集成方法。这种算法通过自举法 (Bootstrap) 从原始数据中创建出不同样本，采用每一个样本生成独立的决策树 (Breiman (1996)、Dietterich (2000)) 并构建一个森林 (Breiman (2001)、Criminisi, Shotton 和

请务必阅读正文之后的信息披露和重要声明

Konukoglu (2012) )。单个树的预测结果依然采用最后分组中的多数观测值，多个树的集合预测结果采用下式计算：

$$E_t(\gamma_{i,t+1}) = \begin{cases} 1, & \text{if } \frac{1}{B} \sum_{b=1}^B 1(E_{t,b}(\gamma_{i,t+1}) = 1) > 0.5 \\ 0, & \text{otherwise} \end{cases}$$

其中  $1(E_{t,b}(\gamma_{i,t+1}) = 1)$  是一个二元参数，如果决策树  $b$  的预测结果  $E_{t,b}(\gamma_{i,t+1}) = 1$ ，则  $1(E_{t,b}(\gamma_{i,t+1}) = 1)$  的值为 1，反之为 0。

构建森林可能会导致不同树之间选择了相同因子，我们通过随机选择预测因子的子集来解决这个问题。实践中我们选择的子集因子数量为预测因子总数的平方根。由于自举样本之间的差异，采用袋装树预测的结果可以减少过度拟合，从而更加稳健。

### 2.3.2、提升树 (Boosted Trees)

提升树是袋装树的变体，提升树中的自举样本并不是随机选出的。提升算法会依次构建决策树，当一个决策树生成后，会用整个样本对其进行检验。在下一个自举样本中，先前分类错误的观察值会再次出现，算法通过最小化以前所有树的错误预测来生成新的决策树，参见 Schapire 和 Singer (1999)。随着每棵树的生成，森林可以不断地提高预测精度。我们使用的提升算法叫做温和适应性提升。

## 3、实证研究

### 3.1、数据

我们选取了一个由所有流通的商品期货组成的固定样本集合，这些期货在 1987 年 1 月至 2019 年 10 月的样本区间内是可交易的。数据来自于彭博。通过选择固定的期货集合，我们避免了由幸存者偏差引起的问题。幸存者偏差是股票市场中普遍存在的现象。为了保持不同时间之间的可比性，我们只关注 1 月、2 月和 3 月期的合约，其中 2 月和 3 月期的合约只是为了构造因子，交易完全基于 1 月期合约。

商品期货包括：豆油 CBOT (BO)、玉米 CBOT (C)、糖 ICE (QW)、可可 ICE (CC)、WTI 原油 Nymex (CL)、布伦特原油 ICE (CO)、棉花#2 ICE (CT)、饲养牛 CME (FC)、黄金 Comex (GC)、铜 Comex (HG)、加热油 Nymex (HO)、咖啡 ‘C’ ICE (KC)、堪萨斯小麦 CBOT (KW)、铝 LME (LA)、活牛 CME (LC)、生猪 CME (LH)、铅 LME (LL)、镍 LME (LN)、锌 LME (LX)、天然气 Nymex (NG)、铂金 Nymex (PL)、柴油 Nymex (QS)、大豆 CBOT (S)、糖#11 ICE (SB)、银 Comex (SI)、豆粕 CBOT (SM)、小麦 CBOT (W) 以及 RBOB 汽油 Nymex (XB)。

我们采用了标准的期货滚动流程来计算每种商品期货日收盘价的连续收益序列。如果合约的最后交易日期在合约月之前，我们将在最后交易日期前 10 天对期货合约进行滚动。对于一些商品期货，第一个通知日是在最后一个交易日之前，



在第一个通知日之后该合约的交易量会显著下降。因此，我们选择在最后一个交易日前 10 天对期货合约进行换手，以确保合约流动性。例如，对于一份 2019 年 9 月到期的布伦特原油合约，9 月合约的最后交易日期是 2019 年 7 月 31 日。因为最后交易日期在合约月之前，所以我们会在合约最后一个交易日（2019 年 7 月 31 日）的前 10 天进行换手。如果合约的最后交易日期在合约月内或之后的月份，我们将在合约月第一天的前 10 天进行滚动换手。例如，对于一份 2019 年 12 月到期的玉米合约，12 月合约的最后交易日期是 2019 年 12 月 13 日。因为最后的交易日是在对应的合约月份，所以我们将在合约月的第一天（即 2019 年 12 月 1 日）前 10 天对合约进行滚动。如果合约月的第一天是非交易日，我们将在下一个交易日的前 10 天进行换手。

### 3.2、预测因子

商品期货的一个特点是价格几乎完全反映了经济基本面的信息，包括现货价格、仓储成本以及未来供需情况。我们方法的一个关键点是选择预测因子，这些预测因子是根据可观测的历史市场价格构建的。我们主要参考了过往文献中提出的一些因子，比如波动率、动量、价值、规模、流动性和隐含尾部风险等，参见 Sharpe (1964)、Fama 和 French (1992)、Longstaff (1995)、Asness, Moskowitz 和 Pedersen (2013)、Farhi 和 Gabaix (2016) 等。Fuertes, Miffre 和 Fernandez-Perez (2015) 研究了商品期货市场的风险因子，他们在过去研究发现的这些因子中还加入了历史收益的偏度和峰度。

#### 3.2.1、动量

我们的第一组预测因子  $x_{i,t,1-4}$  是商品期货历史收益的动量（参见 Moskowitz, Ooi 和 Pedersen (2012)）。公式如下：

$$x_{i,t,1-4} = \frac{P_{i,t}}{P_{i,t-q}} \quad \forall q \in \{10, 20, 60, 180\}$$

其中  $q$  为回顾的时间窗口长度， $P_{i,t}$  是  $t$  时刻商品期货  $i$  的收益。我们的第二组预测因子也是对动量的考察，采用移动平均方法构造，衡量短期和长期移动平均线的收敛程度（参见 Brock, Lakonishok 和 Lebaron (1992)），公式如下：

$$x_{i,t,5-8} = \frac{(\frac{1}{Q} \sum_{q=0:Q} P_{i,t-q} - \frac{1}{Z} \sum_{z=0:Z} P_{i,t-z}) / (\frac{1}{Z} \sum_{z=0:Z} P_{i,t-z})}{(\frac{1}{Q} \sum_{q=10:(Q+10)} P_{i,t-q} - \frac{1}{Z} \sum_{z=10:(Z+10)} P_{i,t-z}) / (\frac{1}{Z} \sum_{z=10:(Z+10)} P_{i,t-z})} - 1$$

其中， $Z$  为长期移动平均线回顾的时间窗口，取 260 个或 360 个交易日， $Q$  为短期移动平均线回顾的时间窗口，取 30 个或 60 个交易日。

#### 3.2.2、波动率

我们的第三组预测因子衡量的是期货日度收益的波动率，与 Sharpe (1964) 提出的风险度量指标相似。公式如下：

$$x_{i,t,9-12} = \sqrt{\frac{1}{Q-1} \sum_{q=0:Q} \left( \frac{P_{i,t-q} - P_{i,t-1-q}}{P_{i,t-1-q}} - \mu_Q \right)^2} \quad \forall Q \in \{10, 20, 60, 180\}$$

请务必阅读正文之后的信息披露和重要声明

其中  $\mu_Q$  表示日度收益均值， $Q$  代表回顾的时间窗口，以交易日表示。

### 3.2.3、价值

我们采用了三个因子来反映价值。首先，我们采用了两个变量来衡量期货曲线的陡峭程度。期货价格曲线通常是以升水（向上倾斜）来解释储存成本，更陡峭的曲线可能意味着商品供应暂时过剩，商品价格也会相应的暂时走低。反之亦然，处于贴水（向下倾斜）的期货曲线可能表明商品的暂时短缺，商品价格也相应的出现上涨。前两个价值因子的公式如下：

$$x_{i,t,13-14} = \frac{F_{q,t} - F_{1,t}}{F_{1,t}} \quad \forall q \in \{2,3\}$$

其中  $q$  表示合约月份（例如  $F_{3,t}$  表示  $t$  时刻 3 月期的期货合约价格）。Schwerhoff 和 Stuermer (2019) 的研究表明，大宗商品的长期价格相当稳定。因此，我们还采用了一种衡量方法，来评估当前价格与长期历史平均价格的关系。公式如下：

$$x_{i,t,15} = \frac{F_{1,t}}{\frac{1}{4 \times 260} \sum_{q=0:4 \times 260} F_{1,t-q}}$$

### 3.2.4、其他

Farhi 和 Gabaix (2016) 在货币市场中研究了一个风险因子，关注的是 95% 看跌和 105% 看涨的虚值期权 30 天隐含波动率的差异。期权隐含波动率的差异反映了期权价格的差异，进而体现 1 月期合约的收益风险是左偏还是右偏。公式如下：

$$x_{i,t,16} = iVol_{95,i,t}^{30days} - iVol_{105,i,t}^{30days}$$

按照 Fuertes, Miffre 和 Fernandez-Perez (2015) 的方法，我们还使用了过去一年日收益的偏度和峰度，公式如下：

$$x_{i,t,17} = \frac{\frac{1}{Q} \sum_{q=0:Q} \left( \frac{P_{i,t-q} - P_{i,t-1-q}}{P_{i,t-1-q}} - \mu_Q \right)^3}{\left( \sqrt{\frac{1}{Q} \sum_{q=0:Q} \left( \frac{P_{i,t-q} - P_{i,t-1-q}}{P_{i,t-1-q}} - \mu_Q \right)^2} \right)^3}$$

$$x_{i,t,18} = \frac{\frac{1}{Q} \sum_{q=0:Q} \left( \frac{P_{i,t-q} - P_{i,t-1-q}}{P_{i,t-1-q}} - \mu_Q \right)^4}{\left( \sqrt{\frac{1}{Q} \sum_{q=0:Q} \left( \frac{P_{i,t-q} - P_{i,t-1-q}}{P_{i,t-1-q}} - \mu_Q \right)^2} \right)^4}$$

其中  $Q$  为 260 个工作日。

最后一个指标是每种商品相对于其他商品的市场规模，采用商品期货过去一段时间的平均交易额与八种核心商品期货平均交易额的比值，公式如下：

$$x_{i,t,19} = \frac{1}{Z} \sum_{z=0:Z} relVol_{i,t-z}$$

其中：

$$relVol_{i,t} = \frac{\frac{1}{Q} \sum_{q=0:Q} Vol_{i,t-q}}{\frac{1}{Q} \sum_{q=0:Q} \sum_j Vol_{j,t-q}} \quad j \in \{BO, C, CL, HO, LC, S, SM, W\}$$

请务必阅读正文之后的信息披露和重要声明

$Vol_{i,t}$  为商品期货  $i$  在  $t$  日的交易额，以美元计价。我们将  $Q$  和  $Z$  设置为 40 个工作日。

### 3.2.5、虚拟变量

最后，我们还加入了一些虚拟变量。首先，考虑到一些商品期货的季节性（比如天然气），我们加入了 12 个虚拟变量代表不同月份。其次，为了代表商品期货特有的特征，我们加入了 28 个虚拟变量，每个虚拟变量对应一种商品期货。

我们总共使用了 59 个预测因子，并按照 Struck 和 Cheng (2019) 的方法对数据进行标准化。该方法首先观察每个预测因子的实际分布，再按照逆分布函数将因子标准化为服从  $(0,1)$  的正态分布。数据标准化对量化结果的影响非常小，不会改变我们的任何结论。

### 3.3、参数

由于标准袋装树估计对采用的随机数序列高度敏感，我们从不同点开始采用相同的随机数序列重复估计了近 100 次，以确保我们的结果是稳健的。最后我们对所有估计结果进行平均。

为了避免多重检验问题，我们使用的样本内和样本外数据没有重叠，样本内区间为 1987 年 1 月至 2013 年 12 月，样本外区间为 2014 年 1 月至 2019 年 10 月。我们选择的样本内区间更长一些，主要有 2 个原因：第一，如 Rietz (1988)、Barro (2006)、Farhi 和 Gabaix (2016)、Struck 和 Cheng (2019) 所述，尾部风险是影响许多因子策略表现的关键，因此样本内数据应该包括尽可能多的尾部事件；第二，我们希望表明，我们的研究结果在本文发表的前几年依然是可靠的。

由于长期大宗商品价格在整个抽样期间相对稳定（虽然全球对大宗商品的需求不断增加，但科技发展也使供应量有所提升，因此价格相对稳定，参见 Acemoglu (2002)），我们根据模型对大宗商品的收益预测进行了排序，来构建市场中性的投资组合。我们按照排序选择前五的商品期货等权构建多头组合，并选择后五的商品期货等权构建空头组合。为了对不同的风险因子策略进行检验，我们还使用 28 个商品期货等权构建了一个多头组合。

我们采用 4 周的标准预测窗口，同时我们也选择每 4 周对组合进行再平衡。最后，考虑到我们以比较低的频率交易期货合约，我们估计每年的交易成本低于 8 个基点。

### 3.4、主要发现

我们的实证分析分为四个步骤：首先，我们计算了四种模型（Logit、LASSO、袋装树、提升树）对样本内和样本外数据的收益预测能力，结果见图表 2；其次，我们比较了四种模型的样本外表现及构建的多头组合表现，结果见图表 3 和图表 4；接着，我们关注每个模型赋予单个预测因子的权重，排名见图表 5；最后，我们重点研究了基于决策树构建的投资组合与其他资产之间的相关性，结果见图表 6 和图表 7。

图表 2 展示了四种模型预测收益和实际收益之间的相关性（均以 4 周为预测长度）。第 i 行和第 ii 行是包含了所有 28 个商品期货的投资组合收益之间的相关性。第 iii 行和第 iv 行是仅使用每段时间前五个和后五个商品期货（采用预测收益排序）构建的多空组合收益之间的相关性。第 v 行和第 vi 行是仅使用每段时间前三个和后三个商品期货（采用预测收益排序）构建的多空组合收益之间的相关性。

图表 2 中有五个关键点：首先，对于样本外数据，Top 3/Bottom 3 的平方相关系数高达 3.74%；其次，对于样本外数据，袋装树模型构建的组合相关性是线性模型（Logit、LASSO）的 10 倍以上；第三，所有模型对样本外数据的预测能力都随着投资组合选取资产数量的减少而增强，例如袋装树模型对 Top 3/Bottom 3 的预测能力为 3.74%，高于对 Top 5/Bottom 5 的预测能力（2.02%）和全样本的预测能力（0.99%）；第四，我们采用的决策树模型获得了远超线性模型的样本内相关性；最后，决策树模型具有较高的样本内相关性，但样本外相关性并没有这么高。

**图表 2、4 周预测收益和 4 周实际收益的平方相关系数**

	Logit	LASSO	Bagged Trees	Boosted Trees
i. In-Sample, All Commodities	0.42%	0.42%	50.26%	51.87%
ii. Out of Sample, All Commodities	0.00%	0.00%	0.99%	0.33%
iii. In Sample, Top 5/Bottom 5	0.61%	0.66%	55.41%	54.59%
iv. Out of Sample, Top 5/Bottom 5	0.12%	0.10%	2.02%	1.40%
v. In Sample, Top 3/Bottom 3	0.58%	0.57%	56.21%	56.23%
vi. Out of Sample, Top 3/Bottom 3	0.33%	0.25%	3.74%	1.44%

资料来源：The Journal of Financial Data Science，兴业证券经济与金融研究院整理

图表 3 展示了不同模型的样本外表现，并将样本外区间分为五个阶段：整个样本外时期、过去 3 年、过去 2 年、过去 1 年及年初至今。表中分别展示了不同区间每种模型的年化平均日收益和波动率，以及最大回撤，还对每种模型计算了两个风险调整后收益指标：夏普比率（根据日收益和波动率计算）和收益回撤比。图表 4 是样本外表现走势。

图表 3 和 4 中有两个值得注意的结果。首先，从风险调整收益（夏普比率和收益回撤比）来看，两种决策树模型（袋装树和提升树）表现都显著优于线性模型（Logit 和 LASSO）和多头组合，且这个结果在所有样本时期都成立（整个样本外区间、过去 3 年、过去 2 年、过去 1 年和年初至今）。线性模型的夏普比率介于-0.461 和 0.526 之间，多头组合的夏普比率介于-0.609 和 0.487 之间，而决策树模型的夏普比率介于 0.736 和 1.562 之间。此外，线性模型的收益回撤比在-0.349 至 0.274 之间，多头组合的收益回撤比在-0.182 和 0.140 之间，而决策树模型的收益回撤比在 0.75 至 2.007 之间。Moritz 和 Zimmermann (2016)、Gu, Kelly 和 Xiu (2020) 都发现美股市场中决策树模型的风险调整收益高于线性模型。虽然研究的是不同市场和不同时间跨度，很难直接进行比较，但他们的定量发现与我们的发现大致相同。

其次，仔细研究图表 3 和图表 4 可以发现，决策树模型的风险调整收益较高主要是由于组合收益较高，而不是由于更低的波动率或回撤。在相似的回撤或波

请务必阅读正文之后的信息披露和重要声明

动率水平下，两种决策树模型的年化收益比线性模型和多头组合高至少 8%。但决策树模型的波动率与线性模型相比没有显著优势，以年初至今来看，袋装树模型的波动率为 11.4%，而 Logit 模型只有 9.3%；以过去一年来看，提升树模型的波动率为 9.1%，低于 LASSO 模型的 9.9%。类似地，决策树模型的最大回撤可能高于线性模型（整个样本外区间提升树的最大回撤为 15%，而 Logit 模型只有 13.7%），也可能低于线性模型（过去 3 年袋装树模型的最大回撤为 6.8%，低于 LASSO 的最大回撤 13.9%）。

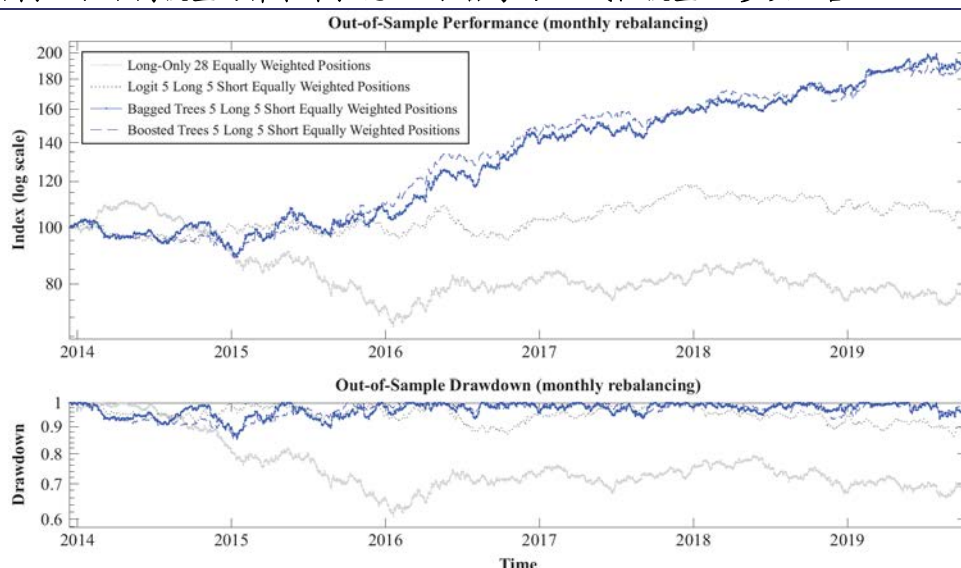
图表 3、不同模型的样本外表现 1：机器学习 vs.线性模型 vs.多头组合

	Sample	Last 3 Years	Last 2 Years	Last 1 Year	Year-to- Date
<b>Boosted Trees:</b> Sharpe Ratio	1.296	1.343	1.224	0.736	1.562
Return-to- MaxDD	0.75	1.644	1.443	0.936	2.007
Volatility	0.087	0.088	0.085	0.091	0.092
Return	0.112	0.118	0.104	0.067	0.144
Max Drawdown	-0.15	-0.072	-0.072	-0.072	-0.072
<b>Bagged Trees:</b> Sharpe Ratio	1.101	1.34	1.193	0.79	1.118
Return-to- MaxDD	0.836	1.916	1.708	1.33	1.86
Volatility	0.106	0.098	0.098	0.115	0.114
Return	0.116	0.131	0.117	0.091	0.127
Max Drawdown	-0.139	-0.068	-0.068	-0.068	-0.068
<b>LASSO:</b> Sharpe Ratio	0.174	0.443	-0.221	-0.489	-0.407
Return-to- MaxDD	0.115	0.274	-0.136	-0.349	-0.275
Volatility	0.092	0.086	0.085	0.099	0.094
Return	0.016	0.038	-0.019	-0.048	-0.038
Max Drawdown	-0.139	-0.139	-0.139	-0.139	-0.139
<b>Logit:</b> Sharpe Ratio	0.187	0.526	-0.168	-0.439	-0.461
Return-to- MaxDD	0.125	0.328	-0.104	-0.309	-0.312
Volatility	0.092	0.086	0.085	0.097	0.093
Return	0.017	0.045	-0.014	-0.042	-0.043
Max Drawdown	-0.137	-0.137	-0.137	-0.137	-0.137
<b>Long-Only:</b> Sharpe Ratio	-0.353	-0.032	-0.218	-0.609	0.487
Return-to- MaxDD	-0.097	-0.009	-0.063	-0.182	0.14
Volatility	0.107	0.096	0.098	0.102	0.097
Return	-0.038	-0.003	-0.021	-0.062	0.047
Max Drawdown	-0.391	-0.339	-0.339	-0.339	-0.339

资料来源：The Journal of Financial Data Science，兴业证券经济与金融研究院整理



图表 4、不同模型的样本外表现 2：机器学习 vs.线性模型 vs.多头组合



资料来源：The Journal of Financial Data Science，兴业证券经济与金融研究院整理

当我们考察预测因子的重要性时，我们发现不同方法之间几乎没有相似性（见图表 5）。价值（ $x_{i,t,16}$ ）、1 个月动量（ $x_{i,t,2}$ ）、3 个月动量（ $x_{i,t,3}$ ）、收益率偏度（ $x_{i,t,17}$ ）和峰度（ $x_{i,t,18}$ ）是线性模型中重要程度排名前五的预测因子。在提升树模型中，前述 5 个因子有 4 个的重要程度排在前 10（1 月和 3 月动量、收益偏度和峰度）。在袋装树模型中，前述 5 个因子只有 2 个的重要程度排在前 5（价值和收益峰度）。整体来看，不同模型在预测因子重要性排序上差异较大。这与 Gu, Kelly 和 Xiu（2020）的发现有所差异，他们发现采用不同方法对美股市场研究得到的因子重要性排名基本一致。

图表 5、预测因子重要性

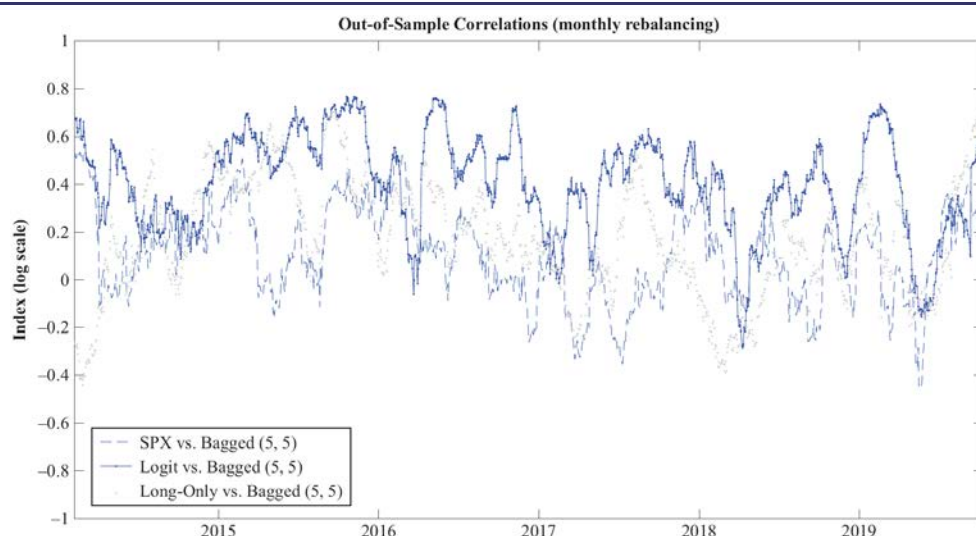
	Logit/LASSO	Boosted Trees	Bagged Trees
1	Ret3M ( $x_{i,t,3}$ )	Curve3minus1	Comvalue
2	Comvalue ( $x_{i,t,15}$ )	Curve2minus1	KurtCom
3	KurtCom ( $x_{i,t,18}$ )	Ret2W	Vola3M
4	Ret1M ( $x_{i,t,2}$ )	Vola2W	Vola9M
5	SkewCom ( $x_{i,t,17}$ )	KurtCom	SIMovAvg60D260D
6	Vola3M ( $x_{i,t,11}$ )	Ret1M	VolumeCom
7	Vola1M ( $x_{i,t,10}$ )	SkewCom	Vola1M
8	Skewivol ( $x_{i,t,16}$ )	Vola1M	SIMovAvg60D360D
9	VolumeCom ( $x_{i,t,19}$ )	Vola3M	Vola2W
10	Vola9M ( $x_{i,t,12}$ )	Ret3M	Skewivol
11	Vola2W ( $x_{i,t,9}$ )	VolumeCom	SIMovAvg30D260D
12	Curve3minus1 ( $x_{i,t,14}$ )	Vola9M	SkewCom
13	SIMovAvg30D260D ( $x_{i,t,5}$ )	Ret9M	SIMovAvg30D360D
14	Ret2W ( $x_{i,t,1}$ )	SIMovAvg60D260D	Ret2W
15	SIMovAvg60D360D ( $x_{i,t,8}$ )	SIMovAvg30D260D	Curve3minus1
16	Ret9M ( $x_{i,t,4}$ )	Comvalue	Ret1M
17	SIMovAvg60D260D ( $x_{i,t,6}$ )	SIMovAvg60D360D	Ret9M
18	Curve2minus1 ( $x_{i,t,13}$ )	SIMovAvg30D360D	Ret3M
19	SIMovAvg30D360D ( $x_{i,t,7}$ )	Skewivol	Curve2minus1

资料来源：The Journal of Financial Data Science，兴业证券经济与金融研究院整理

因子重要性排序的差异进一步导致了不同模型结果的相关性较低。图表 6 和图表 7 分别为袋装树和提升树模型与其他模型的 40 天滚动相关性。图表 6 中袋装树和 Logit 模型的相关系数在-0.2 至 0.8 之间，长期均值接近 0.35。图表 7 中提升树和 Logit 模型的相关系数在-0.6 至 0.8 之间，长期均值接近 0.2。相关性和预测因子重要性的差异也表明数据的实际结构并不是简单的线性结构。

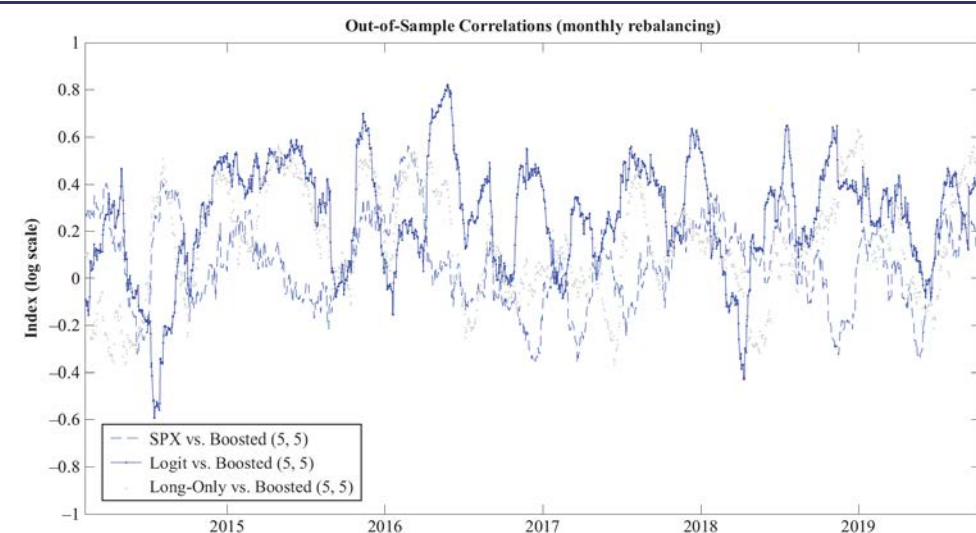
此外，这两个图表还表明，决策树模型与多头组合和股票市场（标普 500）的相关性也较低。因此，采用决策树模型构建的组合可以为股票多头组合提供额外收益，并且这种收益既不能被线性风险因子完全解释，也不能被商品期货多头策略完全解释。

**图表 6、样本外相关性：袋装树 vs.线性模型 vs.多头组合**



资料来源：The Journal of Financial Data Science，兴业证券经济与金融研究院整理

**图表 7、样本外相关性：提升树 vs.线性模型 vs.多头组合**



资料来源：The Journal of Financial Data Science，兴业证券经济与金融研究院整理

## 4、探讨

基于我们的定量结果，我们提出了两个观点：首先，市场的有效性可能低于历史文献认为的水平；其次，对因子的孤立研究可能会导致对数据的错误解释。

我们的第一个观点是基于对商品期货市场的理解，这个市场信息透明、流动性高、交易成本低，非常接近有效市场假说的定义。同时因为我们研究的是固定的商品期货集合，所以避免了影响股票市场研究的幸存者偏差。此外，与股票市场研究不同的是，我们只使用了基于历史市场价格的因子。

到目前为止，历史文献已经研究了许多可以解释部分收益变化的因子，然而大多数文献都只关注了少量线性独立的因子。有效市场假说认为，人们无法基于过去信息预测未来收益，除非过去信息反映了某种风险。但我们的研究表明在模型中加入因子的交互作用和非线性后，模型的预测能力比线性模型提高了 10 倍，并且该结果通过了大量的稳健性检验。

考虑到我们使用的方法虽然在统计上比较直观，但难以解释，我们无法确定我们的模型是否反映了某种市场风险。但结果显示模型对收益率的预测能力远远大于先前的假设，这与有效市场假说不一致。

我们的第二个观点也基于我们的发现，即收益的可预测性比过往文献发现的高得多。正如前文所述，绝大多数文献在研究收益可预测性时都做了过于简化的统计假设，其中许多文献都只是孤立地研究一个预测因子。简化的统计模型使这些文章的结果易于解释，但根据我们的研究，因子之间的相互作用和非线性对于预测收益也非常重要。

没有考虑相互作用和非线性的研究可能会对数据做出错误的解释。例如，一些文献在单独研究某一因子时发现其在统计学上显著，因此得出该因子可能反映某种风险或行为偏差的结论。但当和其他因子共同研究时，该因子可能不再显著，表明其并不代表某种风险或行为偏差。

当我们从简单线性模型转换为决策树模型时，模型预测能力有显著提升，这表明数据结构具有非线性。因此，我们的模型不仅能够更好地预测未来收益，也对过往针对几个甚至单一因子进行研究的文献提出了质疑。

我们研究结果的局限性是什么？尽管考虑了数据中的复杂结构，图表 2 的结果表明我们依然只能解释约 50% 的样本内收益变化。这可能是由于两个原因：首先，尽管我们的因子集合比过往文献更大，但仍有进一步扩展的空间——例如，纳入与汇率变动等宏观因子之间的交互项；其次，我们构建的两个决策树模型没有使用非常复杂的结构。如前文所述，我们使用了贪婪算法来提高运算效率，而不是寻找全局最优树。

对模型的这两点简化说明收益预测能力还有进一步提升的潜力。但我们依然认为我们的结果接近预测能力上限，理由如下：从线性模型转换为决策树模型使样本内预测能力由 0.58% 提高至 56.21%，但样本外预测能力仅从 0.33% 提高至 3.74%。因此，样本内预测能力的提高对样本外预测能力的影响是有限的。所以我们认为即使将样本内预测能力由 56.21% 提高至 100%，样本外预测能力可能也不会有显著提升。

虽然我们的研究结果表明，以往的线性模型大大低估了收益率的可预测性，但是使用了决策树模型后，依然有超过 95% 的样本外收益率变化是无法预测的。从这个角度来看，我们的研究结果与有效市场假说一致，即利用过去信息无法预测未来收益。

## 5、结论

金融市场的收益率是否可被预测？常规的线性模型对数据添加了三个特别假设，削弱了其收益预测能力。本文对比了标准线性模型和决策树模型在商品期货市场中的收益预测能力。本文选择商品期货市场，是因为其信息透明、流动性高、交易成本低，接近有效市场假说对理想环境的定义。

本文有 3 个主要发现：首先，采用决策树模型可以预测 3.74% 的月度收益，是线性模型的 10 倍；其次，从线性模型转换为决策树模型后，样本内（样本外）预测能力由 0.58%（0.33%）提升至 56.21%（3.74%）；第三，采用决策树模型构建的投资组合在绝对收益和风险调整后收益上均优于线性模型。

基于我们的研究结果，我们提出了两个观点：首先，决策树模型的收益预测能力远高于线性模型，对此一种可能的解释是，过往文献高估了市场的有效性；其次，当我们从简单线性模型转换为决策树模型时，模型预测能力有显著提升，这表明数据结构具有非线性。因此，当我们对新因子进行检验时，不能对其单独检验来判断是否具有解释能力，也不能将新因子加入到一个简单地线性模型中判断是否增加了解释能力，而应该将新因子加入到考虑因子相互作用和非线性的因子集合中来检验其解释能力。



## 参考文献

- [1] Acemoglu, D. 2002. "Directed Technical Change." *The Review of Economic Studies* 69 (4): 781–809.
- [2] Asness, C. S., T. J. Moskowitz, and L. H. Pedersen. 2013. "Value and Momentum Everywhere." *The Journal of Finance* 68 (3): 929–985.
- [3] Athey, S. *The Impact of Machine Learning on Economics*. Chicago: University of Chicago Press, 2019.
- [4] Athey, S., and G. W. Imbens. 2019. "Machine Learning Methods That Economists Should Know About." *Annual Review of Economics* 11 (1): 685–725.
- [5] Bailey, D. H., and M. López De Prado. 2014. "The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting, and Non-Normality." *The Journal of Financial Data Science* 40 (5): 94–107.
- [6] Barro, R. J. 2006. "Rare Disasters and Asset Markets in the Twentieth Century\*." *The Quarterly Journal of Economics* 121 (3): 823–866.
- [7] Bishop, C. M. *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press, Inc., 1995.
- [8] Black, F. 1993. "Beta and Return." *The Journal of Financial Data Science* 20 (1): 8.
- [9] Breiman, L. 1996. "Bagging Predictors." *Machine Learning* 24 (2): 123–140.
- [10] Breiman, L. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.
- [11] Brock, W., J. Lakonishok, and B. Lebaron. 1992. "Simple Technical Trading Rules and the Stochastic Properties of Stock Returns." *The Journal of Finance* 47 (5): 1731–1764.
- [12] Criminisi, A., J. Shotton, and E. Konukoglu. 2012. "Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning." *Foundations and Trends in Computer Graphics and Vision* 7 (2–3): 81–227.
- [13] De Bondt, W. F. M., and R. Thaler. 1990. "Do Security Analysts Overreact?" *American Economic Review* 80 (2): 52–57.
- [14] De Long, J. B., A. Shleifer, L. H. Summers, and R. J. Waldmann. 1990. "Noise Trader Risk in Financial Markets." *Journal of Political Economy* 98 (4): 703–738.
- [15] DellaVigna, S., and J. M. Pollet. 2009. "Investor Inattention and Friday Earnings Announcements." *The Journal of Finance* 64 (2): 709–749.
- [16] Dietterich, T. G. 2000. "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization." *Machine Learning* 40 (2): 139–157.
- [17] Elton, E. J., M. J. Gruber, and C. R. Blake. 1996. "Survivorship Bias and Mutual Fund Performance." *The Review of Financial Studies* 9 (4): 1097–1120.
- [18] Fama, E. F. 1970. "Efficient Capital Markets: A Review of Theory and Empirical Work." *The Journal of Finance* 25 (2): 383–417.
- [19] Fama, E. F., and K. R. French. 1992. "The Cross-Section of Expected Stock Returns." *The Journal of Finance* 47 (2): 427–465.
- [20] Fama, E. F., and K. R. French. 1993. "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics* 33 (1): 3–56.
- [21] Fama, E. F., and K. R. French. 2016. "Dissecting Anomalies with a Five-Factor Model." *The Review of Financial Studies* 29 (1): 69–103.
- [22] Farhi, E., and X. Gabaix. 2016. "Rare Disasters and Exchange Rates." *Quarterly Journal of Economics* 131 (1): 1–52.
- [23] Fayyad, U. M., and K. B. Irani. "Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning." *Proceedings of the 13th International Joint Conference on Artificial Intelligence*. Chambéry, France, August 28–September 3, 1993, pp. 1022–1029.
- [24] Feng, G., S. Giglio, and D. Xiu. "Taming the Factor Zoo: A Test of New Factors." Working paper no. 25481, National Bureau of Economic Research Working Paper Series, 2019.
- [25] Fuertes, A. M., J. Miffre, and A. Fernandez-Perez. 2015. "Commodity Strategies Based on Momentum, Term Structure, and Idiosyncratic Volatility." *Journal of Futures Markets* 35 (3): 274–297.
- [26] Garcia, S., J. Luengo, J. A. Saez, V. Lopez, and F. Herrera. 2013. "A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning." *IEEE Transactions on Knowledge and Data Engineering* 25 (4): 734–750.
- [27] Grinblatt, M., and B. Han. 2005. "Prospect Theory, Mental Accounting, and Momentum." *Journal of Financial Economics* 78 (2): 311–339.
- [28] Gu, S., B. T. Kelly, and D. Xiu. 2020. "Empirical Asset Pricing via Machine Learning." *The Review of Financial Studies* 35 (5): 2223–2273.
- [29] Harvey, C. R., and Y. Liu. 2019. "False (and Missed) Discoveries in Financial Economics." *The Journal of Finance* (forthcoming).
- [30] Harvey, C. R., Y. Liu, and H. Zhu. 2016. "... and the Cross-Section of Expected Returns." *The Review of Financial Studies* 29 (1): 5–68.

请务必阅读正文之后的信息披露和重要声明



- [31] Hirshleifer, D. 2015. "Behavioral Finance." Annual Review of Financial Economics 7 (1): 133–159.
- [32] Jorion, P., and W. N. Goetzmann. 1999. "Global Stock Markets in the Twentieth Century." The Journal of Finance 54 (3): 953–980.
- [33] Kothari, S. P., J. Shanken, and R. G. Sloan. 1995. "Another Look at the Cross-Section of Expected Stock Returns." The Journal of Finance 50 (1): 185–224.
- [34] Lo, A., and A. C. MacKinlay. 1990. "Data-Snooping Biases in Tests of Financial Asset Pricing Models." The Review of Financial Studies 3 (3): 431–467.
- [35] Longstaff, F. A. 1995. "How Much Can Marketability Affect Security Values?" The Journal of Finance 50 (5): 1767–1774.
- [36] López de Prado, M. L. Advances in Financial Machine Learning. Hoboken: Wiley, 2018.
- [37] MacKinlay, A. C. 1995. "Multifactor Models Do Not Explain Deviations from the CAPM." Journal of Financial Economics 38 (1): 3–28.
- [38] Mclean, R. D., and J. Pontiff. 2016. "Does Academic Research Destroy Stock Return Predictability?" The Journal of Finance 71 (1): 5–32.
- [39] Moritz, B., and T. Zimmermann. "Tree-Based Conditional Portfolio Sorts: The Relation between Past and Future Stock Returns." Working paper, 03, SSRN, 2016.
- [40] Moskowitz, T. J., Y. H. Ooi, and L. H. Pedersen. 2012. "Time Series Momentum." Journal of Financial Economics 104 (2): 228–250.
- [41] Novy-Marx, R. 2013. "The Other Side of Value: The Gross Profitability Premium." Journal of Financial Economics 108 (1): 1–28.
- [42] Rietz, T. A. 1988. "The Equity Risk Premium a Solution." Journal of Monetary Economics 22 (1): 117–131.
- [43] Schapire, R. E., and Y. Singer. 1999. "Improved Boosting Algorithms Using Confidence-Rated Predictions." Machine Learning 37 (3): 297–336.
- [44] Schwerhoff, G., and M. Stuermer. "Non-Renewable Resources, Extraction Technology, and Endogenous Growth." Working paper 2015(09), Federal Reserve Bank of Dallas, 2019.
- [45] Schwert, G. W. 2003. Chapter 15 Anomalies and Market Efficiency, Vol. 1, pp. 939–974. Elsevier, 2003.
- [46] Sharpe, W. F. 1964. "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." The Journal of Finance 19 (3): 425–442.
- [47] Shleifer, A., and R. W. Vishny. 1997. "The Limits of Arbitrage." The Journal of Finance 52 (1): 35–55.
- [48] Struck, C., and E. Cheng. 2019. "Time-Series Momentum: A Monte Carlo Approach." The Journal of Financial Data Science 1 (4): 103–123.

**风险提示：**文献中的结果均由相应作者通过历史数据统计、建模和测算完成，在政策、市场环境发生变化时模型存在失效的风险。

## 分析师声明

本人具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度，独立、客观地出具本报告。本报告清晰准确地反映了本人的研究观点。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

## 投资评级说明

投资建议的评级标准	类别	评级	说明
报告中投资建议所涉及的评级分为股票评级和行业评级(另有说明的除外)。评级标准为报告发布日后的12个月内公司股价(或行业指数)相对同期相关证券市场代表性指数的涨跌幅。其中：A股市场以上证综指或深圳成指为基准，香港市场以恒生指数为基准；美国市场以标普500或纳斯达克综合指数为基准。	股票评级	买入	相对同期相关证券市场代表性指数涨幅大于15%
		审慎增持	相对同期相关证券市场代表性指数涨幅在5%~15%之间
		中性	相对同期相关证券市场代表性指数涨幅在-5%~5%之间
		减持	相对同期相关证券市场代表性指数涨幅小于-5%
		无评级	由于本文无法获取必要的资料，或者公司面临无法预见结果的重大不确定性事件，或者其他原因，致使本文无法给出明确的投资评级
	行业评级	推荐	相对表现优于同期相关证券市场代表性指数
		中性	相对表现与同期相关证券市场代表性指数持平
		回避	相对表现弱于同期相关证券市场代表性指数

## 信息披露

本公司在知晓的范围内履行信息披露义务。客户可登录 [www.xyzq.com.cn](http://www.xyzq.com.cn) 内幕交易防控栏内查询静默期安排和关联公司持股情况。

## 使用本研究报告的风险提示及法律声明

兴业证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。本报告仅供兴业证券股份有限公司（以下简称“本公司”）的客户使用，本公司不会因接收人收到本报告而视其为客户。本报告中的信息、意见等均仅供客户参考，不构成所述证券买卖的出价或征价邀请或要约。该等信息、意见并未考虑到获取本报告人员的具体投资目的、财务状况以及特定需求，在任何时候均不构成对任何人的个人推荐。客户应当对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特定需求，必要时就法律、商业、财务、税收等方面咨询专家的意见。对依据或者使用本报告所造成的一切后果，本公司及/或其关联人员均不承担任何法律责任。

本报告所载资料的来源被认为是可靠的，但本公司不保证其准确性或完整性，也不保证所包含的信息和建议不会发生任何变更。本公司并不对使用本报告所包含的材料产生的任何直接或间接损失或与此相关的其他任何损失承担任何责任。

本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可升可跌，过往表现不应作为日后的表现依据；在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告；本公司不保证本报告所含信息保持在最新状态。同时，本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

除非另行说明，本报告中所引用的关于业绩的数据代表过往表现。过往的业绩表现亦不应作为日后回报的预示。本文并不承诺也不保证，任何所预示的回报会得以实现。分析中所做的回报预测可能是基于相应的假设。任何假设的变化可能会显著地影响所预测的回报。

本公司的销售人员、交易人员以及其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。本公司没有将此意见及建议向报告所有接收者进行更新的义务。本公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。

本报告并非针对或意图发送予或为任何就发送、发布、可得到或使用此报告而使兴业证券股份有限公司及其关联子公司等违反当地的法律或法规或可致使兴业证券股份有限公司受制于相关法律或法规的任何地区、国家或其他管辖区域的公民或居民，包括但不限于美国及美国公民（1934年美国《证券交易所》第15a-6条例定义为本「主要美国机构投资者」除外）。

本报告的版权归本公司所有。本公司对本报告保留一切权利。除非另有书面显示，否则本报告中的所有材料的版权均属本公司。未经本公司事先书面授权，本报告的任何部分均不得以任何方式制作任何形式的拷贝、复印件或复制品，或再次分发给任何其他人，或以任何侵犯本公司版权的其他方式使用。未经授权的转载，本公司不承担任何转载责任。

## 特别声明

在法律许可的情况下，兴业证券股份有限公司可能会利差本报告中提及公司所发行的证券头寸并进行交易，也可能为这些公司提供或争取提供投资银行业务服务。因此，投资者应当考虑到兴业证券股份有限公司及/或其相关人员可能存在影响本报告观点客观性的潜在利益冲突。投资者请勿将本报告视为投资或其他决定的唯一信赖依据。

## 兴业证券研究

上海	北京	深圳
地址：上海浦东新区长柳路36号兴业证券大厦15层	地址：北京西城区锦什坊街35号北楼601-605	地址：深圳市福田区皇岗路5001号深业上城T2座52楼
邮编：200135	邮编：100033	邮编：518035
邮箱：research@xyzq.com.cn	邮箱：research@xyzq.com.cn	邮箱：research@xyzq.com.cn