基于 DNN、GBT、RF 构建的标普 500 统计套利策略

赵润平
2021 级工程管理学院金融工程

日期：2024 年 8 月 20 日

摘　　要

本研究是对 Krauss, Xuan and Nicolas (2017) 的研究复现，旨在分析和实现深度神经网络（DNN）、梯度提升树（GBT）、随机森林（RF）等机器学习模型及其集成模型在统计套利策略中的应用和有效性 [19]。研究使用了标普 500 指数成分股的数据。除原文内容外，我们将数据拓展至 2024 年，且实现了 XGBOOST、LightGBM 模型作为补充，并使用网格搜索对各个模型的参数进行了调优。

原文研究背景强调了统计套利作为一种量化交易策略的重要性，总体上希望为弥合学术界（关注月度数据的高度透明模型）与专业金融界（关注盈利标准的高效“黑箱”模型）之间的差距做出贡献，采用了短期内的复杂模型来捕捉市场中的盈利机会。复现由赵浩瀚和我共同完成，赵浩瀚负责了数据处理、特征生成、模型构建和训练，以产生交易信号，我则完成了交易信号产生后行交易，不同规模的投资组合构建与回测分析。

在模型构建部分，研究使用了多种机器学习模型，并结合了集成学习方法。我负责的回测部分则根据模型信号执行交易，构建了不同规模的投资组合，并分析了考虑交易成本前后，策略的综合性能、收益-风险特征、系统性风险暴露来源。分时期的策略表现以及交易标的的行业拆解等方面。我的报告主要介绍以上内容，模型相关请查看另一份报告。

关键词: 机器学习；统计套利；标普 500

1　背景介绍

统计套利（Statistical Arbitrage，简称 StatArb）是一种广泛应用于对冲基金和专有交易部门的量化交易策略。“统计套利”策略具备以下特征：(i) 交易信号是系统化的或基于规则的，而非基于基本面；(ii) 交易组合为市场中性，即与市场的贝塔值为零；(iii) 产生超额收益的机制是统计性的 [1]。Lo [21] 指出，这种策略通常涉及大量证券（从几百到几千只不等）、极短的持有期（从几秒到几天），并依赖高度计算与信息技术（IT）基础设施。这些底层模型高度保密，不对公众和研究人员公开 [18]。

传统金融研究主要集中在识别具有高度“解释力”的资本市场异常，通常依赖线性模型或投资组合排序。这些研究大多基于月度数据，并未采用高级统计学习方法 [17]。相比之下，金融行业更倾向于使用短期内的复杂“黑箱”模型，这些模型依赖于非线性方法和大量计算来获取盈利机会，从而加剧了学术研究与实际交易实践之间的差距。一个明显的例子是，在金融领域权威期刊 *The Journal of Finance* 过去 30 年发表的约 2000 篇文章中，涉及神经网络的研究仅有 17 篇，且极少应用于实证分析。这反映出学术界在采用新兴技术进行短期市场研究方面的滞后性，也说明了传统方法在捕捉高频动态市场中的局限性。

原文研究旨在弥合这一差距，提出了一种针对 S&P 500 成分股的短期统计套利策略。研究采用了机器学习领域中的几种先进方法。首先，使用深度神经网络（Deep Neural Network），其多层结构能够进行特征抽象，已在语音识别、视觉物体识别等领域取得了突破 [20]。其次，采用梯度提升树（Gradient Boosting Trees），这是一种通过组合多个弱学习器来形成强学习器的技术 [14]。最后，依赖随机森林（Random Forests），它通过构建大量去相关的决策树实现模型增强 [14]。

研究策略的核心目标是通过模型预测股票是否能在未来的一个交易日内跑赢大盘，从而生成买入和卖空信号。数据使用从 1996 年到 2024 年的 S&P500 成分股数据。将所有成分股概率预测值降序排列，买入前 k 支股票并卖空后 k 支股票。

在模型构建部分，使用了深度神经网络、梯度提升树、随机森林、XGBoost 和 LightGBM 等模型，并结合了多模型集成方法。在策略执行与回测部分，依据不同模型信号，构建若干规模投资组合（$k = 10, 50 \ldots$），将整体表现进行对比。此外，回测还包括交易成本的影响，投资组合的“收益-风险”特征、系统性风险暴露来源以及分时期的策略表现等的分析。

交易信号产生后的投资组合构建与回测分析部分由我完成，先前的部分由赵浩瀚完成。本报告重点介绍我负责的部分。

2 数据与软件

这一部分介绍我负责的部分主要使用的数据来源、代码参考资料与第三方库等。

2.1 S&P500 成分股数据

从该 GitHub 仓库（更新至 2024 年 4 月 8 日）获取自 1996 年以来的 S&P 500 成分股列表。其包含 S&P 500 成分股的完整列表、行业分类和成分变动的历史记录。由 J. Aultman (fja0568@gmail.com) 版权所有，并定期维护。

结合该成分股信息，调用 Yahoo Finance 数据源 API，下载相关股价与日收益率数据。

2.2 回测数据与第三方库

策略表现的评价涉及若干指标，在此对需要特别说明的计算方法、工具和数据来源进行详细解释。

在收益-风险特征分析的表 3 和表 4 中，(i) 无特别说明指标使用 Python 中的 `empyrical`、`scipy.stats` 和 `statsmodels.api` 库计算，主要为金融数据中评估风险、收益和风险调整收益相关指标；(ii) VaR 和 CVaR 直接通过历史分位数计算得到；(iii) Pesaran-Timmermann 测试（一个包括 Hausman 统计量的假设检验）是基于文章 [25] 自行编写的函数，参考 GitHub 仓库开源代码；(iv) Newey-West 标准误和 t 统计量的计算同样自行编写，具体实现参考 CSDN 文章。

在系统性风险来源分析的表 5 中，(i) 使用了 Ken French 数据库 提供的关于资产定价因子模型（市场、FF3、FF5、动量和短期反转因子）的所有相关数据；(ii) VIX 数据来自 Cboe VIX® 指数历史数据网站。

3 方法论

这一部分详细阐释了模型产生信号后的交易算法，以及相关回测的原理与经济学含义。

3.1 统计套利交易策略

基于另一篇文档中产生预测结果，我接着进行排名与交易。对于每个时期 $t+1$，预测每只股票 s 在横截面上跑赢其中位数的概率 $\hat{\mathcal{P}}_{t+1,1}^{s,ML}$，其中 $ML \in \{DNN, GBT, RAF, XGBOOST, LightGBM, ENS\}$，且 $s \in \{1, \ldots, n\}$。根据预测分别对所有股票按降序排序，得到对应于 DNN、GBT、RAF、$XGBOOST$、$LightGBM$ 和 ENS 预测的排名。在排名的顶部，找到根据各自学习算法认为最被低估的股票，而在底部是根据这些算法预测出在 $t+1$ 期内最不可能跑赢横截面中位数、被高估的股票。因此，在每个排名中买入排名前 k 的股票，并卖空排名后 k 的股票，其中 $k \in \{1, \ldots, \lfloor n/2 \rfloor\}$。

3.2 回测概述

回测首先对模型综合性能进行对比，接着重点关注 $k = 10$ 的投资组合。为了简化分析，并考虑到不同组合的表现相似性，将分析限制在基础学习器和等权重集成模型（$ENS1$）上。同时，展示了在交易成本（每股每次买卖 0.05%[1]）前后的结果。该交易成本估算参考了 Avellaneda 和 Lee (2010) 的研究 [2]，结合我们的高流动性股票池和高周转率策略，这一估算是实用且合理的。

3.2.1 回测一：模型综合性能对比

回测一首先分析了由前 k 支股票组成的投资组合的表现，其中 $k \in \{10, 50, 100, 150, 200\}$。分析从每日回报率（交易成本前）、标准差以及每日方向准确性三个维度进行比较。图4展示了不同规模多空投资组合的每日绩效指标。

3.2.2 回测二：收益-风险特征分析

回测二旨在全面评估策略的收益与风险特征，具体包括每日回报特征、年化回报及风险指标等方面。

1. **每日回报特征**。
   结果展示于表3。评估了 k = 10 投资组合在交易成本前后的每日回报特征，大部分绩效指标选取参考 Bacon（2008）[3]。此外，使用了 Newey-West 标准误（NW）来修正滞后一阶的估计误差，并通过 Pesaran-Timmermann 检验（PT）分析策略的预测准确性 [25]。回报分布的尾部风险则采用 Mina 和 Xiao（2001）提出的 RiskMetrics 方法进行评估 [23]。
2. **年化回报与风险特征**。
   结果展示于表4。进一步计算了 k = 10 投资组合年化回报与风险指标，该分析包括夏普比率、最大回撤率以及其他风险调整后的回报指标，以全面评估策略的稳健性。

表1将指标分为四大类：收益分布、风险评估、风险-收益特征以及统计测试和模型验证。下面将说明指标如何量化策略的回报、波动性、风险管理能力及其预测准确性。

[1] 由于每次交易由一买和一卖构成，考虑交易成本后收益率减少 0.1%。

表 1: 策略表现度量 - 收益与风险特征

指标分类	指标名称	指标说明
收益分布	平均收益	策略在考虑交易成本前/后的平均收益率。
	描述性统计	最小值、四分位数、中位数、最大值。
	偏度和峰度	分析收益分布中的偏斜和肥尾现象。
风险评估	标准差	突出整体策略波动率。
	在险价值 (VaR) 和条件在险价值 (CVaR)	量化可能的最大损失。
	下行波动率	仅关注下行风险。
	最大回撤	最坏情况下的策略损失。
风险-收益特征	夏普 (Sharpe) 比率	每单位总风险的超额收益补偿。
	索提诺 (Sortino) 比率	相对于下行风险的收益补偿。
	卡尔玛 (Calmar) 比率	相对于最大回撤风险的收益补偿。
统计测试和模型验证	t 统计量 (Newey-West)	检验平均收益的显著性，考虑自相关性和异方差性调整。 H_0: “平均收益等于零。”
	PT(Pesaran-Timmermann) 测试统计量	验证模型的预测准确性。 H_0: “预测与实际响应是独立分布的。”

下面是对表1中指标的计算方式的具体说明，以及自主编写计算的指标代码展示。

1. **偏度：** 偏度衡量回报分布的非对称性。

$$S_S = \sum \left(\frac{r_i - \bar{r}}{\sigma_{Sp}} \right)^3 \times \frac{n}{(n-1) \times (n-2)} \tag{1}$$

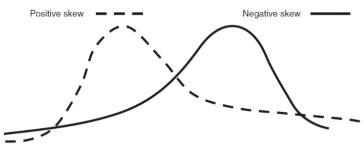

图 1: 分布偏度

当偏度为正时，表示收益分布倾向于频繁的小幅亏损和偶尔的大幅收益；而负偏度则意味着频繁的小幅收益和偶尔的大幅亏损。

2. **峰度：** 峰度衡量回报分布的“尾部厚度”。

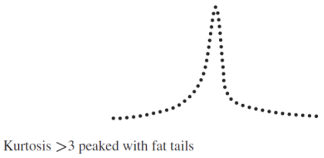

图 2: 分布峰度

$$K_S = \sum \left(\frac{r_i - \bar{r}}{\sigma_{Sp}} \right)^4 \times \frac{n \times (n+1)}{(n-1) \times (n-2) \times (n-3)} \tag{2}$$

高峰度（尖峰分布）表示存在更高的极端回报风险（正负两方面），表明有更高发生意外事件的可能性（“厚尾”）。低峰度（平峰分布）则表示分布更接近正态分布。

3. **在险价值（VaR, Value at Risk）：**估计在给定置信水平下的最大损失，在简便计算中 z_α 是置信水平 α 的 z 值（基于正态分布的回报）。

$$P(r_P < -VaR_\alpha) = 1 - \alpha \tag{3}$$

$$VaR_\alpha = -(\bar{r} - z_\alpha \times \sigma) \tag{4}$$

给出一个损失的阈值。

4. **条件在险价值（CVaR, Conditional Value at Risk）**[2]**：**衡量超出 VaR 的预期平均损失。

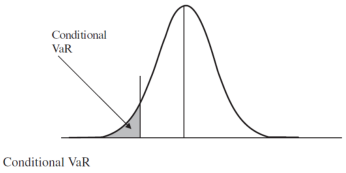

图 3: 条件风险价值（CVaR）图

$$CVaR_\alpha = -\mathbb{E}[r_P \mid r_P < -VaR_\alpha] \tag{5}$$

与 VaR 不同，VaR 仅提供不超过特定置信水平下的损失阈值，而 CVaR 估计所有超出该阈值的损失的平均值。

5. **Sharpe 比率：**衡量每单位风险的超额回报。

$$SR = \frac{r_P - r_F}{\sigma_P} \tag{6}$$

表示在承受更高波动性的情况下获得的超额回报。较高的值表明风险调整后的表现更好。

6. **Calmar 比率：**通过将回报与最大回撤相比较，衡量了投资的风险回报比。

$$CR = \frac{r_P - r_F}{D_{Max}} \tag{7}$$

高 Calmar 比率表明良好的风险回报特征，意味着较高的回报相对于大幅亏损的风险。

7. **Sortino 比率：**类似于夏普比率，但只考虑下行风险。

$$SortinoRatio = \frac{r_P - r_F}{\sigma_D} \tag{8}$$

专注于下行风险，即负收益的波动性，这对于关注下行风险的投资者更为相关。

8. **PT 测试：**PT（Pesaran-Timmermann）检验是一种非参数检验方法，主要用于验证预测模型对方向变化的准确性。该检验特别适用于复杂的非线性动态模型或定性数据的预测场景。

PT 检验的核心思想在于评估预测变量 x_t 是否能够正确预测实际结果 y_t 的符号方向。其核心步骤包括以下几个部分：

(a) **指示变量的定义：**将预测结果和实际结果转换为二元变量，以便进行符号匹配检验：

$$Y_t = 1 \text{ 若} y_t > 0, \text{ 否则} Y_t = 0; X_t = 1 \text{ 若} x_t > 0, \text{ 否则} X_t = 0; Z_t = 1 \text{ 若} y_t x_t > 0, \text{ 否则} Z_t = 0$$

其中，Y_t 是关于实际经济变量 y_t 的指示变量，而 X_t 则是预测变量 x_t 的指示变量。变量 Z_t 则表示预测与实际值的符号是否一致，若一致则取 1，否则为 0。通过衡量预测值与实际值之间符号匹配的比例，从而评估模型的预测能力。

(b) **概率估计：**通过估计 y_t 和 x_t 为正的概率来计算符号匹配的比例：

$$P_y = \Pr(y_t > 0), \quad P_x = \Pr(x_t > 0) \tag{9}$$

在实际操作中，$\hat{P} = \frac{1}{n}\sum_{t=1}^{n} Z_t = \bar{Z}$ 即为符号匹配的经验均值。

[2]CVaR 有时也被称为 Expected Shortfall，即预期损失。

(c) **Hausman 型统计量：**在零假设下，假定 y_t 和 x_t 独立分布，此时我们可以定义以下统计量来进行检验：

$$S_n = \frac{\hat{P} - \hat{P}_*}{\sqrt{\text{var}(\hat{P}) - \text{var}(\hat{P}_*)}} \sim N(0,1), \tag{10}$$

其中 $\hat{P}_*$ 为独立分布假设下符号匹配的期望值，它通过以下公式计算：

$$\hat{P}_* = \hat{P}_y \hat{P}_x + (1 - \hat{P}_y)(1 - \hat{P}_x) \tag{11}$$

其中 $\hat{P}_y = \bar{Y}$ 和 $\hat{P}_x = \bar{X}$ 分别为实际结果和预测结果为正的样本比例。

此外，统计量中的方差 $\text{var}(\hat{P})$ 和 $\text{var}(\hat{P}_*)$ 考虑了样本中不同符号组合的协方差：

$$\text{var}(\hat{P}) = n^{-1}\hat{P}_*(1 - \hat{P}_*) \tag{12}$$

$$\text{var}(\hat{P}_*) = n^{-1}((2\hat{P}_y - 1)^2 \hat{P}_x(1 - \hat{P}_x) + (2\hat{P}_x - 1)^2 \hat{P}_y(1 - \hat{P}_y)) + 4n^{-2}\hat{P}_y\hat{P}_x(1 - \hat{P}_y)(1 - \hat{P}_x) \tag{13}$$

显著的 S_n 值表明 x_t 对 y_t 具有预测能力，即变量之间存在依赖性，推翻了零假设。

特别地，PT 检验不依赖于具体的概率分布假设，适用于更广泛的场景，尤其是在仅有定性预测数据或模型为非线性时。PT 检验在超越传统符号检验的同时，更加关注符号一致性的比例，而非分布形态，从而提供了更加直接的预测准确性衡量。因此在本研究中，使用 PT 检验来检验模型对股票超过市场中位数的定性判断的正确性。

下面是我用于计算 PT 检验统计量的 Python 代码：

```
## PT 检验统计量
def pt_test(Y, X):
    import numpy as np
    from scipy.stats import norm
    import collections

    Z = np.array([i for i in map(lambda x: 1 if np.sign(x[0]) == np.sign(x[1]) else 0,
        zip(Y, X))])

    n = len(Y)
    Py = sum(Y) / n
    Px = sum(X) / n
    PHat = sum(Z) / len(Z)
    PStar = Py * Px + (1 - Py) * (1 - Px)

    VarPHat = (PStar * (1 - PStar)) / n
    VarPStar = ((2 * Py - 1)**2 * Px * (1 - Px)) / n + ((2 * Px - 1)**2 * Py * (1 - Py))
        / n + (4 * Px * Py * (1 - Py) * (1 - Px)) / n**2
    s = (PHat - PStar) / np.sqrt(VarPHat - VarPStar)
    pValue = norm.sf(s)
    pt_return = collections.namedtuple('PT_test', 'PT_statistics␣p_value␣
        Directional_Accuracy')
    rt = pt_return(PT_statistics=s, p_value=pValue, Directional_Accuracy=PHat)
    return rt
pt_test_statistic = pt_test(nn["Target"], nn["pred"])[0]
```

9. **Newey-West t 检验：**通过修正回归模型中的自相关和异方差问题，提供更加稳健的系数估计，参考该统计量计算方式。回归模型中系数 β 的估计方差估计为：

$$Var[\hat{\beta}] = \left(\frac{1}{T}X'X\right)^{-1}\left(\frac{1}{T}X'\sigma^2\Omega X\right)\left(\frac{1}{T}X'X\right)^{-1} \tag{14}$$

在 Newey-West 修正中，方差矩阵 S 估计为：

$$S = \frac{1}{T}\left(\sum_{i=1}^{T} e_i^2 x_i x_i' + \sum_{t=1}^{L} w_l \sum_{t=l+1}^{T} e_t e_{t-l}(x_t x_{t-l}' + x_{t-l} x_t')\right) \tag{15}$$

其中，权重 $w_l = 1 - \frac{l}{L+1}$ 用于修正自相关影响，L 为滞后阶数。

最终修正后的 $\hat{\beta}$ 方差估计为：

$$Var[\hat{\beta}_{NW}] = T(X'X)^{-1} S (X'X)^{-1} \tag{16}$$

下面是我用于计算单变量 Newey-West t 检验统计量的 Python 代码：

```
## t -statistic (NW)
# 计算 Newey-West 标准误差
Y = nn_LSresult['long_short_return']
X = [1] * len(Y)
model = sm.OLS(Y, X).fit(cov_type='HAC', cov_kwds={'maxlags': 6})
std_error_nw = model.bse[0]  # 使用拟合模型中的标准误差
# t 统计量计算
t_statistic_nw = model.tvalues.values[0]
```

3.2.3 回测三：系统性风险敞口来源

回测三通过回归分析评估了集成策略对常见系统性风险来源的敞口。为简化分析，主要关注表现最优的 $k = 10$ 等权重集成策略（ENS1）。分析基于以下四种因子模型展开：Fama-French 三因子模型（FF3）、Fama-French 3+2 因子模型（FF3+2）、Fama-French 五因子模型（FF5）以及结合 VIX 指数的增强模型（FF VIX）。表5展示了 1999 年 7 月至 2023 年 12 月期间，策略在扣除交易成本后的系统性风险敞口。

因子代表不同的系统性风险来源。它们通过捕捉潜在的风险维度和行为偏差来解释横截面市场异常，这些因素驱动了偏离预期市场效率的情况。这里以三因子模型为例，对因子构造方式稍作说明。

在 Fama-French 三因子模型中，SMB（Small Minus Big）和 HML（High Minus Low）因子的计算依赖于对股票的规模和账面市值比（B/M）的分类，见表2。

表 2: 基于规模和账面市值比的股票分类

		账面市值比 (B/M)		
		高 (High)	中	低 (Low)
规模 (Size)	小盘股 (Small)	SH	SM	SL
	大盘股 (Big)	BH	BM	BL

$$SMB = \frac{1}{3}(SH + SM + SL) - \frac{1}{3}(BH + BM + BL), HML = \frac{1}{2}(SH + BH) - \frac{1}{2}(SL + BL) \tag{17}$$

SMB 和 HML 因子的计算见公式17，表示不同分类投资组合之间的平均收益之差，SMB 代表规模效应，HML 代表价值效应。

回测三共进行了四次回归分析。

1. **FF3：** 首先，我们使用 Fama-French 三因子模型（FF3），参考 Fama 和 French（1996）[9]。该模型捕捉了对整体市场、小盘减大盘股票（SMB）和高账面市值比减低账面市值比股票（HML）的敞口。回归公式为：

$$R_{i,t} - R_{f,t} = \alpha + \beta_{mkt}(R_{mkt,t} - R_{f,t}) + \beta_{SMB} SMB_t + \beta_{HML} HML_t + \epsilon_t \tag{18}$$

2. **FF3+2：**其次，我们在该模型基础上加入了动量因子（Momentum）和短期反转因子（Reversal），类似于 Gatev 等人（2006）[12]，称之为 Fama-French 3+2 因子模型（FF3+2）。回归公式为：

$$R_{i,t} - R_{f,t} = \alpha + \beta_{mkt}(R_{mkt,t} - R_{f,t}) + \beta_{SMB}SMB_t + \beta_{HML}HML_t + \beta_{Mom}Mom_t + \beta_{Rev}Rev_t + \epsilon_t \quad (19)$$

3. **FF5：**第三，我们使用最近发展起来的 Fama-French 五因子模型，参考 Fama 和 French（2015）[8]。该模型在三因子模型基础上增加了两个额外因子，即强劲减弱势盈利（RMW）和保守减激进投资行为（CMA）股票的组合。回归公式为：

$$R_{i,t} - R_{f,t} = \alpha + \beta_{mkt}(R_{mkt,t} - R_{f,t}) + \beta_{SMB}SMB_t + \beta_{HML}HML_t + \beta_{RMW}RMW_t + \beta_{CMA}CMA_t + \epsilon_t \quad (20)$$

4. **FF VIX：**最后，在 FF3+2 模型中加入了 VIX 指数 [10, 26]，即“投资者恐惧指标”，捕捉高波动期的市场情绪对策略表现的影响。具体来说，如果 VIX 大于 30（即 90% 分位数），将 VIX 作为哑变量标记为 1[3]。回归公式为：

$$R_{i,t} - R_{f,t} = \alpha + \beta_{mkt}(R_{mkt,t} - R_{f,t}) + \beta_{SMB}SMB_t + \beta_{HML}HML_t + \beta_{Mom}Mom_t + \beta_{Rev}Rev_t + \beta_{VIX}VIX_t + \epsilon_t \quad (21)$$

3.2.4 回测四：策略表现分时期对比

回测四评估了策略在不同时期内的表现，以更好地理解策略随时间变化的回报和风险特征。分析引入了五个子期间[4]。

前四个时期的具体划分方式与依据如下。

1. **1999 年 7 月 - 2001 年 3 月**：这一时期是在现代机器学习算法普及之前。
2. **2001 年 4 月 - 2008 年 8 月**：这一阶段机器学习技术的进步及其带来的市场效率提升。随机森林——最强的基础模型——由 Breiman（2001）发明，并在随后的几年中逐渐普及 [5]。同样，梯度提升树背后的主力方法出现——随机梯度提升（GBT），源自 Friedman（2002）的贡献 [11]。
3. **2008 年 9 月 - 2009 年 12 月**：这一时期对应全球金融危机。
4. **2010 年 1 月 - 2015 年 10 月**：这一时期强大机器学习算法的公开可用性增加以及技术投资的准入门槛降低。

基于原文，在子期间分析中同时关注集成策略的年度和月度异常值。

1. **1999 年**：对应于互联网泡沫破裂前的市场繁荣。
2. **2000 年**：正值即泡沫破裂并且科技股市值损失数十亿的时期。
3. **2008 年**：全球金融危机期间，尤其是在雷曼兄弟倒闭后的一个月（2008 年 10 月）。
4. **2011 年 10 月**：对应欧洲债务危机的高峰期，即希腊债务减记达成协议的时间点。

图6展示了 $k = 10$ 投资组合在 1999 年 7 月至 2023 年 12 月期间扣除交易成本后的分时期表现，比较了 DNN、GBT、RAF、ENS1 与整体市场（MKT）及 VIX 指数的表现。表6汇总了不同子时期内各策略的年化风险-回报特征。

[3]VIX 的这个阈值（> 30）表示高波动期，约占所有交易日的 10%。在此回归中，由于哑变量不可投资，截距不再被解释为超额回报。

[4]我们拓展了原文的分析时间段，前四个时期划分和原文保持一致，新增加的第五个时期为 2015 年 11 月至 2023 年 12 月。

3.2.5 回测五：交易标的行业拆解

回测五进一步分析了策略在不同行业的表现，主要通过行业拆分来探讨策略的具体投资行为。表7展示了结果，对 1999 年 7 月至 2023 年 12 月期间的 S&P 500 成分股进行了分类，获取了全球行业分类标准（GICS）代码，并比较了等权重集成策略（ENS1）的多头和空头投资组合在不同行业中的分布，对比显示了各行业在总数据集中的相对权重与集成策略的投资行为之间的偏差。

4 结果与分析

我根据模型信号执行交易算法后，进行回测。这一部分对策略表现展开分析。

4.1 结果一：模型综合性能对比

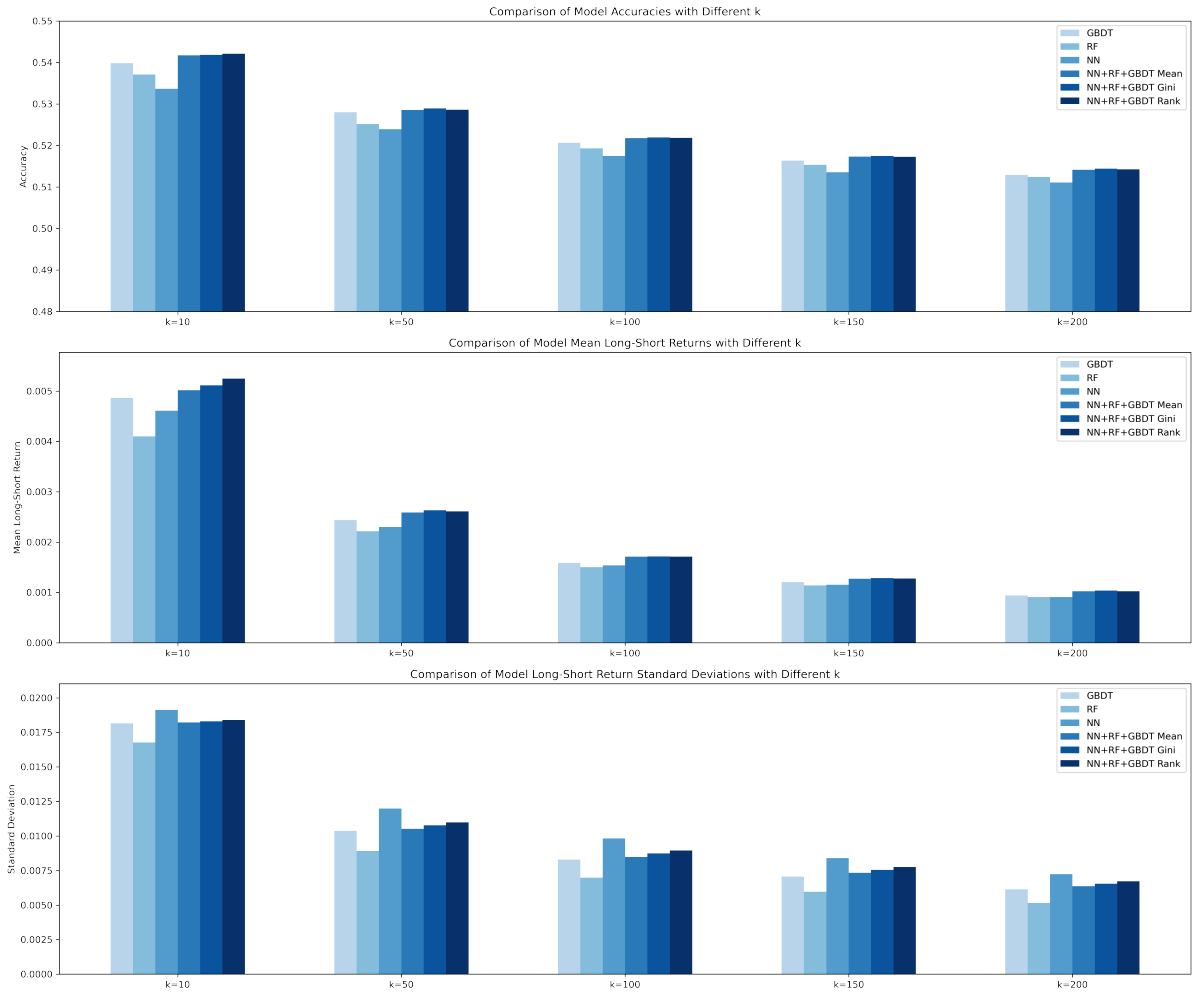

图 4: 不同规模多空投资组合的每日绩效指标：1999 年 7 月至 2023 年 12 月期间的平均回报率、标准差和方向准确性。

结果展示于图4。在对前 k 支股票（$k \in \{10, 50, 100, 150, 200\}$）的投资组合进行分析时，比较了交易成本前的日回报率、标准差和每日方向准确性。总体结果表明，随着 k 值的增加，组合的方向准确性和回报率均呈下降趋势。这一趋势表明，虽然较大的组合能够降低个体股票的不确定性，但其整体收益和准确性会因此受到影响。具体而言，方向准确性的变化几乎与回报率的模式保持一致，显示出这两者之

间的紧密关联。同时，k 值的增加导致组合标准差下降，这与经典投资组合理论的预测相符，即通过资产分散化可以有效降低风险。

集成模型在所有 k 值下的方向准确性均优于单一基础模型。根据 Hansen 和 Salamon (1990) 以及 Dietterich (2000) 的研究，集成模型的高准确性依赖于两个条件：基础学习器的多样性和足够的准确性 [6, 13]，推测本研究原因是多样性，模型误差间相关性低[5]。

在基础学习器的比较中，梯度提升树（GBT）表现优于深度神经网络（DNN）和随机森林（RAF），与原文的排序（RAF > GBT > DNN）略有不同。此外，三种集成模型（ENS1(Equal-Weighted), ENS2(Performance-Based), and ENS3(Rank-Based)）在回报率上的表现非常接近，日回报率均约为 0.50%，三者之间的差异小于每日 0.01%。

基于上述策略表现特点，后续分析集中在 k = 10 的投资组合上，且由于模型表现相似[6]，主要限于基础学习器和 ENS1（等权重集成模型）的表现。

4.2 结果二：收益-风险特征分析

4.2.1 结果二：每日回报特征

表 3: k = 10 组合的日收益特征，比较 DNN、GBT、RAF、ENS1 与整体市场（MKT），分别在扣除交易成本前后的表现（时间范围：1999 年 7 月到 2023 年 12 月）。NW 表示 Newey-West 标准误差（带有一阶滞后修正），PT 表示 Pesaran-Timmermann 测试。

	Before transaction costs				After transaction costs				
	DNN	GBT	RAF	ENS1	DNN	GBT	RAF	ENS1	MKT
Mean Return (long)	0.0021	0.0022	0.0020	0.0023	0.0011	0.0012	0.0010	0.0013	-
Mean Return (short)	-0.0025	-0.0024	-0.0021	-0.0027	-0.0015	-0.0017	-0.0013	-0.0019	-
Mean Return	0.0046	0.0041	0.0050	0.0050	0.0026	0.0029	0.0023	0.0030	0.0004
Standard Error (NW)	0.0020	0.0019	0.0023	0.0024	0.0017	0.0017	0.0016	0.0017	0.0002
t-statistic (NW)	19.6369	20.3608	18.8207	21.2466	11.1158	11.5908	9.6383	12.7802	2.7103
PT Test Statistic	36.9798	40.4204	40.4136	44.5175	-	-	-	-	-
Minimum	-0.0925	-0.1012	-0.0945	-0.0945	-0.1002	-0.0844	-0.0558	-0.0844	-0.1199
Quartile 1	-0.0067	-0.0059	-0.0065	-0.0057	-0.0087	-0.0082	-0.0085	-0.0082	-0.0049
Median	0.0044	0.0045	0.0046	0.0046	0.0029	0.0035	0.0017	0.0031	0.0025
Quartile 3	0.0156	0.0105	0.0126	0.0136	0.0136	0.0136	0.0158	0.0157	0.0156
Maximum	0.1236	0.1257	0.1062	0.1052	0.0836	0.0832	0.0848	0.0832	0.0432
Standard Deviation	0.0191	0.0181	0.0202	0.0201	0.0131	0.0125	0.0123	0.0125	0.0125
Skewness	0.9406	0.6070	0.5753	0.5775	0.0313	0.0237	0.0147	0.0153	-0.0123
Kurtosis	1.9590	1.9107	1.1845	1.1491	1.9590	1.5912	1.7037	1.0845	8.8761
Historical 1-percent VaR	-0.1547	-0.1424	-0.1361	-0.1389	-0.0883	-0.0942	-0.0536	-0.0942	-0.0494
Historical 1-percent CVaR	-0.0653	-0.0518	-0.0547	-0.0547	-0.0583	-0.0488	-0.0249	-0.0488	-0.0444
Historical 5-percent VaR	-0.0258	-0.0320	-0.0351	-0.0351	-0.0340	-0.0249	-0.0247	-0.0249	-0.0494
Historical 5-percent CVaR	-0.0370	-0.0351	-0.0351	-0.0351	-0.0377	-0.0365	-0.0365	-0.0365	-0.0297
Maximum Drawdown	-0.3102	-0.2884	-0.2623	-0.2524	-0.4011	-0.3462	-0.1467	-0.3462	-0.1154
Calmar Ratio	6.7861	7.8257	6.5080	7.3353	4.9011	4.1478	4.8712	4.1478	0.4762
Share with return > 0	0.6106	0.6142	0.6011	0.6270	0.5635	0.5640	0.5442	0.5730	0.5400

[5]研究结合了三种截然不同的模型类型——深度神经网络、提升的浅层树和高深度去相关的树。因此，即使在相同数据上训练，基础学习器也具有一定的多样性。

[6]在此处没有列出 XGBOOST、LightGBM 模型表现，回测结果表明树模型之间差距很小，故最终分析也只保留 GBT 相关分析。

表 4: k = 10 组合的年化收益和风险指标，分别在扣除交易成本前后比较 DNN、GBT、RAF、ENS1 与整体市场（MKT），时间范围为 1999 年 7 月到 2015 年 10 月。

	Before transaction costs				After transaction costs				
	DNN	GBT	RAF	ENS1	DNN	GBT	RAF	ENS1	MKT
Mean Return	2.0440	2.2582	1.7070	2.3857	0.8419	0.9718	0.6377	1.0520	0.0807
Mean excess return	2.1966	2.4125	1.8171	2.5480	0.9346	1.0655	0.7045	1.1040	0.0174
Standard deviation	0.3035	0.2879	0.2661	0.2895	0.3035	0.2879	0.2661	0.2895	0.1977
Downside deviation	0.1764	0.1633	0.1518	0.1629	0.1916	0.1783	0.1674	0.1778	0.1405
Sharpe ratio	3.8267	4.2550	3.8827	4.3673	2.1662	2.5046	1.9884	2.6305	0.4915
Sortino ratio	6.5837	7.5011	6.8035	7.7624	3.4315	4.0445	3.1601	4.2844	0.6917

表3报告了 1999 年 7 月至 2023 年 12 月期间 k = 10 投资组合的每日回报特征。对于等权重集成模型（ENS1），交易成本前的平均日回报为 0.50%，略高于 GBT 的 0.49%、DNN 的 0.46% 和 RAF 的 0.41%。[7]在交易成本后，ENS1 的日回报降至 0.03%。其中，空头仓位的回报贡献（51%-54%）高于多头仓位（41%-47%）。[8]模型的预测准确性通过 Pesaran-Timmermann (PT) 测试得到验证，统计量超过 13.0，表明预测在统计上显著。[9]

在风险特征方面，策略的历史 1% VaR 介于-4.0% 到-4.2% 之间，显著高于整体市场的-3.4%，显示出较高的尾部风险。[10]相比其他策略，如经典配对交易，该策略尾部风险相当大。Gatev 等人 (2006) 发现 k = 5 组合的日 VaR 为 1.24%，k = 20 组合为 0.65%[12]，明显低于本研究。最大回撤方面，GBT 的回撤为 84%，RAF 为 73%。尽管回撤显著，ENS1 的 Calmar 比率达到 171%，远高于市场的 14%，这表明该策略在回撤后的恢复能力强于市场[11]，ENS1 策略在较大回撤后的恢复时间大约需要一年半，市场则需六年。

尽管策略存在较高的尾部风险和较大回撤，这些策略仍提供了可观的平均回报和快速恢复时间，显示出其在高风险调整回报方面的潜在吸引力。

4.2.2 结果二：年化回报与风险特征

表4展示了年化风险-回报指标。在交易成本之后，集成模型（ENS1）的年化回报为 105%，显著高于 GBT 的 97.2%、DNN 的 84.2% 和 RAF 的 63.8%, 优于原文表现[12]。

夏普比率衡量了单位风险下的超额回报，集成模型的夏普比率为 4.28，显著优于整体市场，表明每单位风险的回报率比市场高出约 5.36 倍（2.63/0.49）。这使得集成模型在与其他统计套利策略的比较中表现出色：经典配对交易在 1962 年至 2002 年间前 20 对股票中实现了 0.59 的夏普比率 [12]，广义配对交易在 1997 年至 2007 年间的夏普比率为 1.44[1]，而深度条件投资组合排序在 1968 年至 2012 年间的夏普比率为 2.96——尽管是在交易成本前且基于更大且流动性较低的股票池 [24]。Huck（2009）使用 Elman 神经网络和 ELECTRE III 在 1992 年至 2006 年间实现了约 1.5 的夏普比率——同样是在交易成本前 [16]。Sortino 比率通过下行偏差来衡量回报，集成模型的下行偏差为 0.16，大约是整体市场的 6.20 倍（4.28/0.69）。

总体来说，在风险-回报表现中，集成模型表现最佳，其次依次是 GBT、RAF 和 DNN。

[7]原文中等权重集成的平均日回报为 0.45%，优于 DNN 的 0.33%、GBT 的 0.37% 和 RAF 的 0.43%。

[8]由于对数据进行了 5% 的 Winsor 缩尾处理并采用对数回报率，数据中的偏度和峰度基本消失。

[9]对于 H0: “预测与实际结果独立分布”，在 5% 显著性水平下的临界值为 1.96。

[10]原文中历史 1% VaR 在-5.9% 到-6.9% 之间。

[11]原文中的 ENS1 的 Calmar 比率为 99%。

[12]原文中集成模型的年化回报为 73%，其次是 RAF（67%）、GBT（46%）和 DNN（27%）。

表 5: 等权重集成策略 ENS1（k = 10）的系统性风险暴露（扣除交易成本后），时间范围为 1999 年 7 月到 2023 年 12 月。标准误差在括号中显示。

	FF3	FF3+2	FF5	FF VIX
(Intercept)	0.0030***	0.0027***	0.0029***	0.0026***
	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Market	0.2130***	0.1815***	0.2446***	0.1861***
	(0.0185)	(0.0194)	(0.0207)	(0.0195)
SMB	-0.0410	-0.0646*		-0.0631*
	(0.0361)	(0.0344)		(0.0344)
HML	-0.0662**	0.1361***		0.1384***
	(0.0294)	(0.0296)		(0.0296)
Momentum		0.4151***		0.4168***
		(0.0224)		(0.0224)
Reversal		0.0043***		0.0043***
		(0.0002)		(0.0002)
SMB5			0.6038*	
			(0.3085)	
HML5			-0.1251***	
			(0.0368)	
RMW5			0.1153**	
			(0.0511)	
CMA5			0.1529**	
			(0.0643)	
VIX				0.0017**
				(0.0007)
R^2	0.0214	0.1145	0.0245	0.1152
Adj. R^2	0.0209	0.1138	0.0235	0.1144
Num. obs	6226	6226	6226	6226
RMSE	0.0180	0.0172	0.0180	0.0172

*** p<0.001, ** p<0.05, * p<0.1

4.3 结果三：系统性风险敞口来源

表5展示了系统性风险敞口分析的结果。集成策略在所有模型中的日 Alpha 收益范围为 0.26% 到 0.30%，并且在统计上显著。[13]市场行为的分析结果显示，五个因子模型的载荷正负不一致，且不同模型之间也存在差异，这表明市场行为并不十分确定。

FF5 中的 α 收益最高，为 0.24%。该模型显示出在 HML 因子上负向且显著的载荷，而在 RMW 和 CMA 因子上则为正向且显著的载荷，这表明策略倾向于投资利润较强和温和投资行为的股票。FF3+2 解释能力大大提高，调整后的 R^2 为 0.11。集成策略展示了对动量和短期反转模式的敏感性，这表明机器学习算法能够提取这些市场行为，与原文的发现一致。此外根据 FF VIX 模型，VIX 因子上显示出统计显著的载荷，这表明集成策略在高市场动荡时期表现更好，与原文一致。

总体而言，集成策略在所有因子模型中均能产生 0.26% 到 0.30% 的统计和经济上显著的日 Alpha 收益。尽管部分回报受到系统性风险因子的影响，但这种影响并不一致，表明策略部分反映了基于回报的

[13]原文中的日 Alpha 收益范围为 0.14% 到 0.24%。

市场异常。

4.4 结果四：策略表现分时期对比

在子期间分析中，策略表现显示出随时间变化的显著差异。图6展示了分时期分析结果，图6和图7为市场和策略表现与原文的对比图，相同的趋势表现证明了本研究数据来源可靠、复现结果正确。

第一个子期间（1999 年 7 月至 2008 年 4 月）标志着机器学习算法广泛应用于金融领域之前的强劲超额表现。此期间的回报在统计和经济上都非常显著，年化回报在扣除交易成本后超过 900%，集成策略的夏普比率达到 6.5，策略此时在使用过去尚未公开的强大技术来发现并有效利用金融时间序列中的结构。

第三个子期间（2008 年 9 月至 2009 年 12 月）与全球金融危机相吻合。在这一市场动荡剧烈的时期，策略表现出色，特别是在波动性显著增加的情况下，策略展现了其在高波动市场中的稳健性。集成策略在此期间的年化回报超过 160%，夏普比率为 2.4，高于市场但低于第一个子期间的表现。这一现象与 Do 和 Faff（2010）、Bogomolov（2013）和 Huck（2015）[4, 7, 15] 的研究结果一致，表明在金融危机期间，提供流动性的配对交易策略表现良好。

第四个子期间（2010 年 1 月至 2015 年 10 月）见证了机器学习技术的普及，尤其是随机梯度提升（GBT）的广泛应用，导致市场中未见结构的进一步检测。在此阶段，神经网络和随机森林（考虑交易成本后年化超额收益分别为 62% 和 59%）模型的表现开始相对恶化，并逐渐被 GBT 超越。Breiman（2001）和 Friedman（2002）的研究展示了这些高性能方法的发明和逐步普及，特别是在廉价计算资源可获得性的推动下，进一步加速了市场效率的提升 [5, 11]。

在最后一个子期间（2015 年 11 月至 2023 年 12 月），策略表现持续恶化，年化平均回报在扣除交易成本后介于-15% 到 7% 之间，值得注意的是在其余模型表现大幅回调时（均为负数），神经网络模型表现持续强劲（唯一为正）。技术进步和公共可访问性增加导致市场中利润的下降，尤其是对于公开可用的机器学习算法。然而，专有算法在对冲基金中的潜在应用可能仍具有一定的盈利潜力。

总结来看，本研究模型表现总体上滞后于原始论文，特别是在 2010 年至 2015 年期间表现不一致。尽管在 2015 年后表现出弱点，但在趋势上与原文一致。策略表现显示出初期的显著超额回报、在全球金融危机期间回报激增，随后随着机器学习技术的进步和市场效率的提升进入适度阶段，最终由于技术进步和算法的广泛可访问性导致盈利下降。

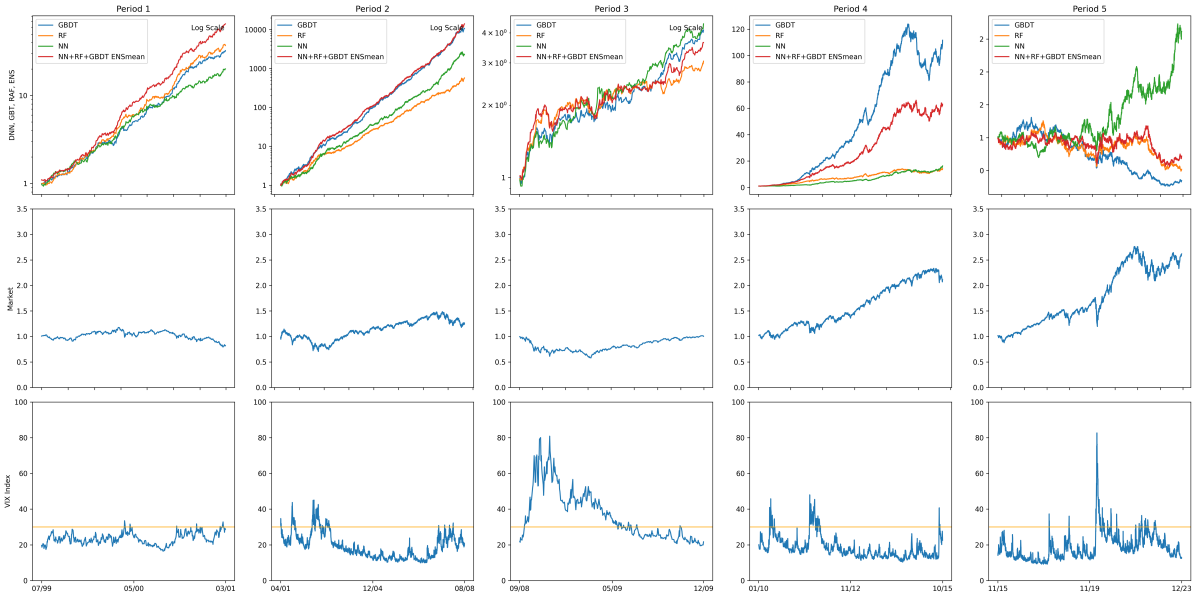

图 5: k = 10 投资组合在 1999 年 7 月至 2023 年 12 月期间扣除交易成本后的分时期表现，比较了 DNN、GBT、RAF、ENS1 与整体市场（MKT）及 VIX 指数的表现。

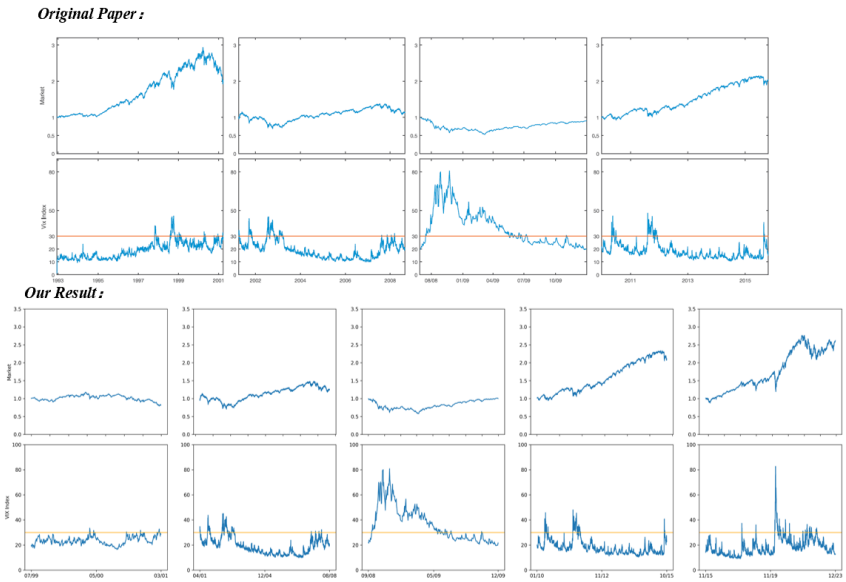

图 6: 原始论文与我们的结果比较（1999 年 7 月至 2023 年 12 月期间整体市场（MKT）及 VIX 指数）。

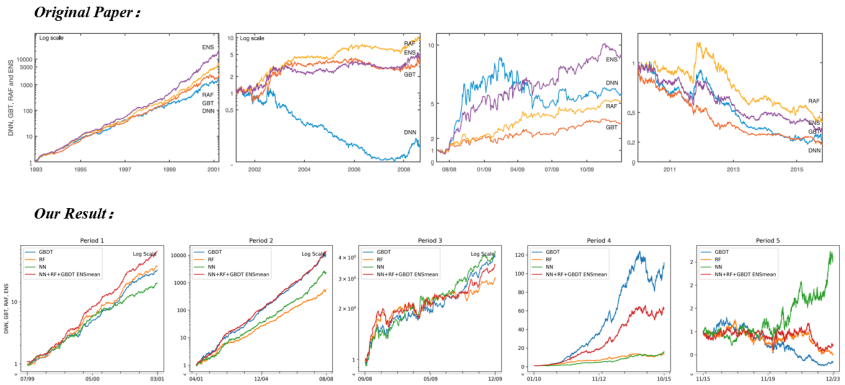

图 7: 原始论文与我们的结果比较（1999 年 7 月至 2023 年 12 月期间扣除交易成本后 k = 10 投资组合的分时期表现，比较了 DNN、GBT、RAF、ENS1 的表现）。

表 6: DNN、GBT、RAF、ENS1 在不同子时期的年化风险-收益特征。

	Before transaction costs				After transaction costs				
	DNN	GBT	RAF	ENS1	DNN	GBT	RAF	ENS1	MKT
Period 07/99-03/01									
Mean return	7.5049	10.2080	11.1282	14.4298	4.1566	5.7993	6.3588	9.0945	-0.0808
Mean excess return	7.9396	10.7812	11.6946	15.1159	4.4214	6.1486	6.7040	9.9848	0.0550
Standard deviation	0.3621	0.3489	0.3331	0.3596	0.3621	0.3489	0.3331	0.3652	0.2278
Sharpe ratio	6.1159	7.1328	7.6932	7.8280	4.7241	5.6882	6.1802	6.5402	-0.3625
Maximum drawdown	-0.0821	-0.1179	-0.0844	-0.0815	-0.0932	-0.0960	-0.1338	-0.0900	-0.3080
Calmar ratio	91.4134	86.6181	131.9135	177.0372	44.5915	43.3339	70.6483	104.3364	-0.2625
Period 04/01-08/08									
Mean return	3.4905	4.5655	2.7763	4.7873	1.7193	2.3718	1.2861	2.5025	0.0435
Mean excess return	3.6849	4.7739	2.9015	5.0080	1.8375	2.4985	1.3622	2.6276	0.0271
Standard deviation	0.2932	0.2709	0.2474	0.2676	0.2932	0.2709	0.2474	0.2676	0.1697
Sharpe ratio	5.2831	6.4935	5.5083	6.7160	3.5643	4.6328	3.4710	4.8287	0.3357
Maximum drawdown	-0.1971	-0.1684	-0.1567	-0.1296	-0.2181	-0.1835	-0.1831	-0.1631	-0.4637
Calmar ratio	17.7068	27.1184	17.7175	36.9418	7.8827	12.9240	7.0059	16.7558	0.1124
Period 09/08-12/09									
Mean return	3.5968	3.3903	2.5852	3.1220	1.7836	1.6583	1.1700	1.7836	0.0278
Mean excess return	3.9539	3.7419	2.7966	3.4620	2.0008	1.8721	1.2985	1.6268	0.0027
Standard deviation	0.3903	0.3493	0.3309	0.4011	0.3903	0.3493	0.3309	0.3930	0.1930
Sharpe ratio	4.1125	3.9494	3.9239	3.7393	2.8212	2.6744	2.4454	2.4245	0.9127
Maximum drawdown	-0.1629	-0.1215	-0.1504	-0.2087	-0.1798	-0.1460	-0.1905	-0.2310	-0.4637
Calmar ratio	22.0747	27.9146	17.1954	14.9558	9.9184	11.3574	14.1426	12.3714	0.3360
Period 01/10-10/15									
Mean return	1.6020	2.6293	1.5450	2.2725	0.5740	1.1969	0.5395	0.9893	0.1241
Mean excess return	1.6917	2.7725	1.6425	2.4021	0.6285	1.2839	0.5987	0.9242	0.0004
Standard deviation	0.2532	0.2568	0.2474	0.2562	0.2532	0.2568	0.2474	0.2568	0.1855
Sharpe ratio	3.9100	5.1594	4.1235	3.9248	1.7565	1.9190	1.6684	3.0834	0.4751
Maximum drawdown	-0.1392	-0.1915	-0.1715	-0.1440	-0.1923	-0.3582	-0.2653	-0.2476	-0.2035
Calmar ratio	11.5074	13.7297	9.0073	15.7812	2.9854	3.3411	3.0337	3.9953	0.6107
Period 11/15-12/23									
Mean return	0.7749	0.3966	0.4760	0.5347	0.0729	-0.1562	-0.1081	-0.0772	0.1212
Mean excess return	0.8859	0.4692	0.5410	0.6086	0.1423	-0.0687	0.0876	0.0140	0.0199
Standard deviation	0.3109	0.2778	0.2578	0.2812	0.3109	0.2778	0.2578	0.2812	0.1990
Sharpe ratio	2.0024	1.3421	1.6435	1.6648	0.3815	-0.4719	-0.3150	-0.1449	0.6935
Maximum drawdown	-0.3012	-0.2884	-0.2623	-0.3247	-0.4011	-0.8461	-0.6890	-0.5903	-0.3358
Calmar ratio	2.5728	1.3752	1.8147	1.6470	0.1816	-0.1570	-0.1274	-0.1274	0.3538

受到全球金融危机期间强劲回报的启发，我们分析了集成策略的年度和月度异常值。结果表明在互联网泡沫、全球金融危机、欧债危机期间等市场动荡和危机时期，模型表现尤为强劲。原文给出的解释是，在市场动荡严重的时期（不对称/极端），即在投资者注意力从个股转向重大市场事件时，错误定价产生，机器学习模型有效捕捉到了证券之间的相对错误定价 [17, 22]。

表 7: 等权重集成策略（k = 10）在 1999 年 7 月到 2023 年 12 月期间的 S&P 500 成分股按行业划分情况与 ENS1 多空组合持仓按行业划分情况的百分比对比。

GICS Sector	Share in S&P 500	Share long	Share short
Communication Services	3.6586	4.4573	4.6122
Consumer Discretionary	10.1774	14.2016	13.7509
Consumer Staples	7.6237	5.4502	3.4683
Energy	4.6801	4.9679	4.5938
Financials	14.4489	9.5812	7.6217
Health Care	12.9631	15.9533	17.2131
Industrials	15.0525	12.5457	11.9158
Information Technology	12.7867	20.5064	25.9418
Materials	5.4973	4.6878	5.1111
Real Estate	6.5466	4.4396	4.6183
Utilities	6.5651	3.2091	2.7465

4.5 结果五：交易标的行业拆解

表7展示了交易标的行业拆解的结果。显然，金融板块的权重在投资组合中被低估了——尽管金融板块在数据集中占比为 14.45%，但在多头仓位中的占比为 9.58%，在空头仓位中的占比仅为 7.62%。同样，工业和公用事业板块在投资组合中的权重也较低，分别在 S&P 500 中占比为 15.05% 和 6.57%，但在多头仓位中的占比为 12.55% 和 3.21%，在空头仓位中的占比为 11.92% 和 2.75%。这种权重差距被信息技术板块的高占比所填补，该板块在 S&P 500 中的占比为 12.79%，但在多头仓位和空头仓位中的占比分别为 20.51% 和 25.95%。这一现象表明，投资组合在选择高贝塔值股票（如科技股）时，倾向于牺牲低贝塔值股票（如工业、公用事业及金融股）。总体而言，各行业中的多头和空头投资比例相对均衡，但在科技股方面表现出明显的偏好。结论与原文一致。[14]

5 总结

本研究是对 Krauss, Xuan and Nicolas (2017) 的研究复现，旨在分析和实现深度神经网络（DNN）、梯度提升树（GBT）、随机森林（RF）等机器学习模型及其集成模型在统计套利策略中的应用和有效性 [19]。研究使用了标普 500 指数成分股的数据，并相较将数据集扩展至 2024 年，除了基础模型外，还补充了 XGBoost 和 LightGBM 模型，并通过网格搜索对模型参数进行了优化。

赵浩瀚负责了数据处理、特征生成、模型构建和训练以产生交易信号，我则完成了交易信号产生后的投资组合构建与回测分析。原文研究背景强调了统计套利作为一种量化交易策略的重要性，总体上希望为弥合学术界（关注月度数据的高度透明模型）与专业金融界（关注盈利标准的高效“黑箱”模型）之间的差距做出贡献，采用了短期内的复杂模型来捕捉市场中的盈利机会。

在模型构建部分，研究使用了多种机器学习模型，并结合了集成学习方法。回测部分则根据模型信号构建了不同规模的投资组合，并分析了考虑交易成本前后，策略的综合性能、收益-风险特征、系统性风险暴露来源。分时期的策略表现以及交易标的的行业拆解等方面。

[14] 原文结果中，金融板块在 S&P 500 中的占比为 14.45%，但在多头和空头仓位中的占比分别为 9.58% 和 7.62%。工业板块占比 15.05%，在多头和空头仓位中的占比分别为 12.55% 和 11.92%。公用事业板块占比 6.57%，对应的多头和空头仓位占比分别为 3.21% 和 2.75%。信息技术板块占比 12.79%，多头和空头仓位占比分别为 20.51% 和 25.95%。

1. **模型综合性能对比**

   分析显示，随着 k 值（投资组合中的股票数量）的增加，组合的方向准确性和回报率呈下降趋势，虽然较大组合可以分散个股波动，但整体收益和准确性随之降低。具体而言，集成模型在各 k 值下的方向准确性显著优于单一模型。集成模型的高准确性源于基础学习器的多样性和足够的准确性，这在模型的误差相关性较低的情况下尤为明显。基础学习器中，梯度提升树（GBT）表现优于深度神经网络（DNN）和随机森林（RAF），且三种集成模型（ENS1, ENS2, ENS3）在回报率上的表现接近，日回报率均约为 0.50%。基于结果，后续分析主要聚焦在 $k = 10$ 的投资组合以及基础学习器与等权重集成模型（ENS1）的表现。

2. **收益-风险特征分析**

   在收益-风险特征分析中，ENS1 在扣除交易成本前的平均日回报为 0.50%，略高于 GBT、DNN 和 RAF，且在扣除交易成本后的回报仍保持在 0.03% 左右。空头仓位的回报贡献更为显著，占比约为 51%-54%。Pesaran-Timmermann (PT) 测试结果验证了模型的预测准确性。尽管策略存在较高的尾部风险（1% VaR 介于-4.0% 至-4.2%），ENS1 在最大回撤后的恢复能力较强，Calmar 比率高达 171%。年化风险-回报指标表明，扣除交易成本后，ENS1 的年化收益为 105%，显著高于其他基础学习器。集成模型的夏普比率（4.28）和 Sortino 比率（7.76）显示出每单位风险下的超额回报大幅领先市场，表现优于其他经典统计套利策略。

3. **系统性风险敞口分析**

   系统性风险分析揭示了集成策略在多因子模型下均能产生 0.26%-0.30% 的显著日 Alpha 收益，且在统计上显著。五因子模型（FF5）显示，集成策略在投资利润强劲和温和投资行为的股票上具有较高的载荷，而在高市场波动（如 VIX 因子）期间表现更好。整体分析表明，集成策略能够有效捕捉市场异常，并在不确定的市场环境中取得超额收益。

4. **策略表现的分时期对比**

   在子期间分析中，策略表现随时间显著变化。1999 年至 2008 年期间，策略表现尤为强劲，年化回报高达 900% 以上，夏普比率达到 6.5，得益于机器学习算法在金融领域的早期应用。2008 年全球金融危机期间，策略在高波动市场中表现稳健，年化回报超过 160%。2010 年至 2015 年，随着 GBT 等高性能方法的普及，市场效率提升，策略表现开始逐步减弱。2015 年后，策略盈利大幅下降，年化回报在扣除交易成本后仅为-15% 至 7%，技术进步和算法的广泛可访问性被认为是导致盈利下降的主要原因。

5. **交易标的行业拆解**行业拆解分析显示，金融、工业和公用事业板块在投资组合中的配置比例相对较低，科技板块则占据较高比重。具体而言，科技板块在 S&P 500 中的占比为 12.79%，但在多头和空头仓位中的占比分别高达 20.51% 和 25.95%。这反映了投资组合倾向于选择高贝塔值股票（如科技股），在一定程度上牺牲了低贝塔值股票（如金融和公用事业股）。这一偏好与原始研究结论一致。

总结来说，本文通过复现和扩展原有研究，验证了机器学习模型在统计套利策略中的有效性，并展示了策略在不同市场环境下的表现。模型在早期阶段展示了显著的超额收益和稳健的风险控制能力，尽管随技术进步和市场效率的提升，表现逐渐减弱。尽管策略在 2015 年后表现出弱化趋势，但其在市场动荡时期的超额回报表现尤为突出。

参考文献

[1] Marco Avellaneda and Jeong-Hyun Lee. “Statistical Arbitrage in the U.S. Equities Market”. In: *Quantitative Finance* 10 (2010), pp. 761–782.

[2] Marco Avellaneda and Jeong-Hyun Lee. “Statistical arbitrage in the US equities market”. In: *Quantitative Finance* 10.7 (2010), pp. 761–782.

[3] Carl R. Bacon. *Practical portfolio performance: Measurement and attribution*. John Wiley & Sons, 2008.

[4] Tim Bogomolov. “Pairs trading in the land down under”. In: *International Review of Financial Analysis* 27 (2013), pp. 21–29.

[5] Leo Breiman. “Random forests”. In: *Machine learning* 45.1 (2001), pp. 5–32.

[6] Thomas G Dietterich. “Ensemble methods in machine learning”. In: *International workshop on multiple classifier systems* (2000), pp. 1–15.

[7] Ba Chu Do and Robert W Faff. “Does simple pairs trading still work?” In: *Financial Analysts Journal* 66.4 (2010), pp. 83–95.

[8] Eugene F Fama and Kenneth R French. “A five-factor asset pricing model”. In: *Journal of Financial Economics* 116.1 (2015), pp. 1–22.

[9] Eugene F Fama and Kenneth R French. “Multifactor explanations of asset pricing anomalies”. In: *Journal of Finance* 51.1 (1996), pp. 55–84.

[10] Marcelo Fernandes, Marcelo C Medeiros, and Marcelo Scharth. “The VIX risk premium: A new measure for international stock markets”. In: *Journal of Banking & Finance* 40 (2014), pp. 1–13.

[11] Jerome H Friedman. “Stochastic gradient boosting”. In: *Computational statistics & data analysis* 38.4 (2002), pp. 367–378.

[12] Evan Gatev, William N Goetzmann, and K Geert Rouwenhorst. “Pairs trading: Performance of a relative-value arbitrage rule”. In: *The Review of Financial Studies* 19.3 (2006), pp. 797–827.

[13] Lars Kai Hansen and Peter Salamon. “Neural network ensembles”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12.10 (1990), pp. 993–1001.

[14] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2009.

[15] Nicolas Huck. “Pairs selection and outranking: An application to the S&P 100 index”. In: *European Journal of Operational Research* 243.2 (2015), pp. 703–712.

[16] Nicolas Huck. “Pairs trading and outranking: A multi-criteria decision approach”. In: *European Journal of Operational Research* 196.2 (2009), pp. 653–661.

[17] Heiko Jacobs. “What Do We Learn From Financial Anomalies?” In: *European Financial Management* 21 (2015), pp. 22–67.

[18] Amir E. Khandani and Andrew W. Lo. “What Happened To The Quants In August 2007?” In: *Journal of Investment Management* 5 (2011), pp. 29–78.

[19] Christopher Krauss, Xuan Anh Do, and Nicolas Huck. “Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500”. In: *Eur. J. Oper. Res.* 259 (2017), pp. 689–702.

[20] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep Learning”. In: *Nature* 521 (2015), pp. 436–444.

[21] Andrew W Lo. *Hedge Funds: An Analytic Perspective*. Princeton University Press, 2010.

[22] Francois Longin and Bruno Solnik. “Extreme correlation of international equity markets”. In: *The Journal of Finance* 56.2 (2001), pp. 649–676.

[23] Jorge Mina and Jerry Y Xiao. *Return to RiskMetrics: The evolution of a standard*. Tech. rep. RiskMetrics Group, 2001.

[24] Benjamin Moritz and Hato Zimmermann. “Deep conditional portfolio sorting: A new approach for risk-based investing”. In: *Proceedings of the 7th International Conference on Computational Finance*. 2014.

[25] M Hashem Pesaran and Allan Timmermann. “A simple nonparametric test of predictive performance”. In: *Journal of Business & Economic Statistics* 10.4 (1992), pp. 461–465.

[26] Robert E Whaley. “The investor fear gauge”. In: *The Journal of Portfolio Management* 26.3 (2000), pp. 12–17.