技术因子的智能择股研究

利用机器学习技术建立智能策略分析框架

分析日期 2020 年 11 月 16 日

证券分析师： 丁竞渊
执业证书编号： S0630516080001
电话： 021-20333723
邮箱： djy@longone.com.cn

◎摘要：

技术分析方法具有悠久的发展历史，并在投资实践中被广泛采用。然而，许多基于技术分析方法的交易策略仍然依赖形态分析和主观判断，这在拥有数千只个股的 A 股市场显得效率低下。近年来，人工智能技术得到了快速发展，以 scikit-learn 为代表的机器学习工具大幅降低了智能技术的应用门槛。在本文所涉及的研究中，我们选择了常用的技术分析指标作为因子，通过训练机器学习模型实现对个股未来表现的预测目标。从模型训练结果来看，机器学习模型对于未来 5 到 10 个交易日的表现具有较好的预测效果，据此构建的投资组合有望获得较好的收益。但随着预测周期延长至 20 到 60 个交易日，预测准确度和相关投资组合收益水平显著下降。根据本文研究结果，我们认为在特定条件下，机器学习模型可以实现对个股未来表现的有效区分，可以用作选股策略的初步筛选器。同时我们还认为，要改进机器学习模型的预测准确度，关键在于进一步寻找区分效果明显的技术因子，以改善机器学习模型的训练数据质量。

相关研究报告

正文目录

1. 研究出发点与思路 5
2. 样本数据的基本情况 5
2.1. 通道突破指标 6
2.2. 均线系统 7
2.3. MACD 指标 8
2.4. 其他指标 8
3. 模型训练框架 12
3.1. 转换为分类问题 12
3.2. 模型训练方案 12
4. 模型训练结果评估 14
4.1. 交叉验证评分 14
4.2. 预测误差分析 16
4.3. 回归测试分析 23
5. 结论 29

图表目录

图 1 (各技术指标与未来涨跌幅、最大回撤之间的关系（一）) 10
图 2 (各技术指标与未来涨跌幅、最大回撤之间的关系（二）) 11
图 3 (样本点计算示意图) 13
图 4 (未来 5 交易日收益率预测交叉验证) 14
图 5 (未来 5 交易日最大回撤预测交叉验证) 14
图 6 (未来 10 交易日收益率预测交叉验证) 15
图 7 (未来 10 交易日最大回撤预测交叉验证) 15
图 8 (未来 20 交易日收益率预测交叉验证) 15
图 9 (未来 20 交易日最大回撤预测交叉验证) 15
图 10 (未来 60 交易日收益率预测交叉验证) 15
图 11 (未来 60 交易日最大回撤预测交叉验证) 15
图 12 (RF 未来 5 交易日回报预测误差) 17
图 13 (MLP 未来 5 交易日回报预测误差) 17
图 14 (RF 未来 10 交易日回报预测误差) 17
图 15 (MLP 未来 10 交易日回报预测误差) 17
图 16 (RF 未来 20 交易日回报预测误差) 17
图 17 (MLP 未来 20 交易日回报预测误差) 17
图 18 (RF 未来 60 交易日回报预测误差) 17
图 19 (MLP 未来 60 交易日回报预测误差) 17
图 20 (RF_0 未来 5 交易日回报分布) 17
图 21 (RF_0 未来 10 交易日回报分布) 17
图 22 (RF_0 未来 20 交易日回报分布) 17
图 23 (RF_0 未来 60 交易日回报分布) 17
图 24 (RF_1 未来 5 交易日回报分布) 18
图 25 (RF_1 未来 10 交易日回报分布) 18
图 26 (RF_1 未来 20 交易日回报分布) 18
图 27 (RF_1 未来 60 交易日回报分布) 18
图 28 (RF_2 未来 5 交易日回报分布) 18
图 29 (RF_2 未来 10 交易日回报分布) 18

图 30 (RF_2 未来 20 交易日回报分布)..... 18
图 31 (RF_2 未来 60 交易日回报分布)..... 18
图 32 (RF_3 未来 5 交易日回报分布)..... 18
图 33 (RF_3 未来 10 交易日回报分布)..... 18
图 34 (RF_3 未来 20 交易日回报分布)..... 18
图 35 (RF_3 未来 60 交易日回报分布)..... 18
图 36 (RF_4 未来 5 交易日回报分布)..... 19
图 37 (RF_4 未来 10 交易日回报分布)..... 19
图 38 (RF_4 未来 20 交易日回报分布)..... 19
图 39 (RF_4 未来 60 交易日回报分布)..... 19
图 40 (RF_5 未来 5 交易日回报分布)..... 19
图 41 (RF_5 未来 10 交易日回报分布)..... 19
图 42 (RF_5 未来 20 交易日回报分布)..... 19
图 43 (RF_5 未来 60 交易日回报分布)..... 19
图 44 (RF_6 未来 5 交易日回报分布)..... 19
图 45 (RF_6 未来 10 交易日回报分布)..... 19
图 46 (RF_6 未来 20 交易日回报分布)..... 19
图 47 (RF_6 未来 60 交易日回报分布)..... 19
图 48 (RF_7 未来 5 交易日回报分布)..... 20
图 49 (RF_7 未来 10 交易日回报分布)..... 20
图 50 (RF_7 未来 20 交易日回报分布)..... 20
图 51 (RF_7 未来 60 交易日回报分布)..... 20
图 52 (MLP_0 未来 5 交易日回报分布)..... 20
图 53 (MLP_0 未来 10 交易日回报分布)..... 20
图 54 (MLP_0 未来 20 交易日回报分布)..... 20
图 55 (MLP_0 未来 60 交易日回报分布)..... 20
图 56 (MLP_1 未来 5 交易日回报分布)..... 20
图 57 (MLP_1 未来 10 交易日回报分布)..... 20
图 58 (MLP_1 未来 20 交易日回报分布)..... 20
图 59 (MLP_1 未来 60 交易日回报分布)..... 20
图 60 (MLP_2 未来 5 交易日回报分布)..... 21
图 61 (MLP_2 未来 10 交易日回报分布)..... 21
图 62 (MLP_2 未来 20 交易日回报分布)..... 21
图 63 (MLP_2 未来 60 交易日回报分布)..... 21
图 64 (MLP_3 未来 5 交易日回报分布)..... 21
图 65 (MLP_3 未来 10 交易日回报分布)..... 21
图 66 (MLP_3 未来 20 交易日回报分布)..... 21
图 67 (MLP_3 未来 60 交易日回报分布)..... 21
图 68 (MLP_4 未来 5 交易日回报分布)..... 21
图 69 (MLP_4 未来 10 交易日回报分布)..... 21
图 70 (MLP_4 未来 20 交易日回报分布)..... 21
图 71 (MLP_4 未来 60 交易日回报分布)..... 21
图 72 (MLP_5 未来 5 交易日回报分布)..... 22
图 73 (MLP_5 未来 10 交易日回报分布)..... 22
图 74 (MLP_5 未来 20 交易日回报分布)..... 22
图 75 (MLP_5 未来 60 交易日回报分布)..... 22
图 76 (MLP_6 未来 5 交易日回报分布)..... 22
图 77 (MLP_6 未来 10 交易日回报分布)..... 22

图 78 (MLP_6 未来 20 交易日回报分布)..... 22
图 79 (MLP_6 未来 60 交易日回报分布)..... 22
图 80 (MLP_7 未来 5 交易日回报分布)..... 22
图 81 (MLP_7 未来 10 交易日回报分布)..... 22
图 82 (MLP_7 未来 20 交易日回报分布)..... 22
图 83 (MLP_7 未来 60 交易日回报分布)..... 22
图 84 (MLP 预测未来 5 个交易日的各分组组合走势)..... 23
图 85 (随机森林预测未来 5 个交易日的各分组组合走势)..... 23
图 86 (MLP 预测未来 10 个交易日的各分组组合走势)..... 24
图 87 (随机森林预测未来 10 个交易日的各分组组合走势)..... 24
图 88 (MLP 预测未来 20 个交易日的各分组组合走势)..... 25
图 89 (随机森林预测未来 20 个交易日的各分组组合走势)..... 25
图 90 (MLP 预测未来 60 个交易日的各分组组合走势)..... 26
图 91 (随机森林预测未来 60 个交易日的各分组组合走势)..... 26

表 1 涨跌幅和最大回撤分类..... 12
表 2 预测未来 5 个交易日的各分组表现..... 27
表 3 预测未来 10 个交易日的各分组表现..... 27
表 4 预测未来 20 个交易日的各分组表现..... 28
表 5 预测未来 60 个交易日的各分组表现..... 28

1.研究出发点与思路

基本面分析与技术面分析共同构成了证券投资经典分析方法的两大门派。一般而言，基本面分析更多地关注寻找更具价值的投资标的，而技术分析则侧重于在动态的，不完全均衡的市场中，抓住价格波动的机会。理论上，无论基本面分析还是技术分析都可以采用主观或量化的分析方法。但在实践中，技术分析所依赖的数据具有更高的及时性、准确性、完备性和量化程度，因此往往更易于被量化投资策略采用。

技术分析方法已经有相当长的发展历史，但即使是一些出现很早的，基于技术分析的交易策略，至今仍然具有很强的生命力。这是因为技术分析更多地反映了市场力量博弈的一般规律，特别是市场情绪影响下投资者的一般行为模式，从而使得技术分析方法的有效性，不受或很少受到经济周期变化和历史发展的影响。一种基于技术分析方法的投资策略如果在一定的时间阶段失效了，其原因往往不在于策略本身，而在于策略为太多的投资者熟知，从而被太多地使用，或者出现了专门针对这类策略的对手。即使是这样，同一个策略仍然可能在一段时间以后再次变得有效。

事实上，已经有许多重要的量化投资策略源自传统的技术分析方法，并且往往进一步发展了它们。例如，量化投资策略要求策略是公式化的，甚至可程序化的，这就迫使技术分析中的交易条件必须是明确的，而不再依赖主观判断，从而提高了交易策略表现的一致性。再如，传统技术分析方法来自于经验法则，难以进行严格的验证，而量化投资策略可以通过回归测试方法对不同市场条件下的样本进行检验，从而可以获得对策略表现的定量化评估。

但是量化投资策略对交易条件的明确性要求，使之难以将主观技术分析方法中一些难以量化、公式化的部分纳入其中。此外，在构造量化策略的过程中，通常做法是先根据理论或经验设计交易规则和参数，然后通过回归测试加以验证，再根据测试结果调整规则或优化参数，如此反复迭代形成最终的可用策略。这种方式的缺陷在于，构建、测试、优化策略的效率较低，因子之间的一些隐含的关系难以被发掘，形成有效策略的难度较大。

近年来，人工智能技术实现了蓬勃发展和深入应用。尤其是以 Python 编程环境和 scikit-learn 为代表的机器学习库的普及，大幅降低了人工智能技术的应用门槛。如何将人工智能技术应用于投资实践，也成为当前金融领域中人们热衷尝试的课题。需要注意的是，当前人工智能技术应用较为成功的案例，集中在图像识别、人像识别、指纹识别、语音识别等物理特征识别领域。这类领域的共同特点是特征稳定，模型容易收敛，可以获得相当高的成功率。但是在资本市场，价格波动的运动特征要复杂得多，且极不稳定，这是由资本市场天然的博弈属性决定的，因此人工智能技术在投资领域的应用难度要大得多。

在本文中，我们针对一些常用的技术分析指标作为因子，采用 scikit-learn 中提供的多种机器学习进行训练，用于预测个股未来的收益、风险情况。之所以选择技术面因子而非基本面因子，是因为基本面因子的有效性容易受到市值、板块、行业、市场风格转换等因素的影响，需要进行更精细的模型训练和验证，这将是我们未来可能尝试的方向之一。而在本文中，我们首先采用表现更为稳定，普遍适用性更好，且更易获取数据的技术分析因子。本次研究的另一个目的是，建立起一套涵盖从样本采集，模型训练，模型验证，直至投资实践的完整流程的智能策略分析框架，以持续、高效地支持基于机器学习的策略研究。

2.样本数据的基本情况

在机器学习领域，样本数据是最关键的因素：数据决定了模型可能达到的最佳效果，而模型只是尽可能逼近这个效果。因此，我们选取在长期投资实践中，被市场中投资者广

泛使用，普遍认为有效的技术指标来构造样本数据。一般来说，技术分析策略可以分为趋势和反转两大交易类型。趋势交易策略的支持者认为，价格会沿着当前的方向持续运动，直至被取代，因此必须在趋势启动的时候建立头寸，以抓住获取超额利润的机会。而反转交易策略的逻辑则是，价格会沿着一个价值中枢（如均线）上下运动，因此应该在价格下跌时逢低买入，在价格上涨时逢高卖出。从策略表现上看，趋势交易策略具有抓住大行情的能力，盈利能力强，盈亏比高，但在震荡市中会反复承受亏损，胜率低。而反转交易策略的表现正好相反，适用于震荡市，但往往会错失大行情。趋势与反转交易策略的根本差异在于，对价格沿均值回复运动的频率特征有不同的预测。因此，一般而言，趋势交易策略更适用于“肥尾”特征明显的方向性交易；而套利交易、期权波动率交易，则可以视为对相关资产的价差序列，或波动率水平的反转交易。

在本文中，我们主要针对趋势交易策略所涉及的常用技术指标，考察机器学习技术能否基于这些技术指标进行有效的预测，在一定程度上替代形态分析和主观判断，从而提高选择投资标的时的效率和执行趋势交易策略的成功率。趋势交易策略的信号系统主要包括“通道”类，即以价格向上突破某一价格通道的上轨线作为做多信号，以价格向下突破某一价格通道的下轨线作为做空信号；以及均线交叉类，即以某一较短周期的平滑曲线向上穿越较长周期的平滑曲线作为做多信号，以较短周期的平滑曲线向下穿越较长周期的平滑曲线作为做空信号。因此，在本文中我们选择了比较有代表性的“海龟交易系统”、均线系统、MACD 系统中采用的主要技术指标，以及一些用于衡量趋势强度、交易活跃度、价格时间序列波动特征的其他指标，来构建机器学习样本的 x 向量。同时，我们依据历史数据统计了个股在未来 5，10，20，60 个交易日的表现，并将涨跌幅及最大回测作为样本的 y 向量。样本覆盖了 2017 年至 2019 年的 A 股全市场数据，共计近 300 万个数据点。图 1 和图 2 部分展示了样本中各技术指标和未来涨跌幅、最大回撤之间的对应关系。

通道突破指标

“海龟交易系统”是最典型的通道交易系统之一，其主要依据“唐奇安通道”(Donchian Channels）和“平均波幅”（Average True Range，ATR）作为其信号系统。唐奇安通道由上下两根轨线组成，其中上轨线（Donchian Upper）是过去 N 个交易周期[1]最高价的最大值；下轨线（Donchian Lower）是过去 N 个交易周期最低价的最小值；唐奇安通道还包括一根中线，为上下轨线的均值，但在“海龟交易系统”中并未使用。唐奇安通道的定义见公式 1。

$$Donchian\ Upper = \max_{N}(High)$$

$$Donchian\ Lower = \min_{N}(Low) \qquad \text{公式 1}$$

$$Donchian\ Middle = (Donchian\ Upper - Donchian\ Lower)/2$$

以多头交易为例，海龟交易系统通常采用一根较长周期（例如$N = 20$）的唐奇安通道上轨线，和一根较短周期（例如$N = 10$）唐奇安通道下轨线。当价格从下往上突破上轨线时，被视为入市建仓信号；当价格从上往下突破下轨线时，被视为平仓信号。如果是空头交易，则信号的使用正好相反，即向下突破长周期的下轨作为建仓信号，向上突破短周期的上轨作为平仓信号。海龟交易系统可以同时采用两套不同周期的唐奇安通道系统，在市场交易规则允许的情况下也可以同时进行多空交易。

[1] 根据交易频率不同，可以以交易日、周、小时、分钟作为一个交易周期。

除唐奇安通道外，ATR 指标也在海龟交易系统中具有重要作用。ATR 是“真实波幅”（True Range，TR）指标的 N 个周期的移动平均，而 TR 则是当前最高价、当前最低价和前一周期收盘价之差的绝对值的最大值，见公式 2。TR 衡量了一个交易周期中价格波动的幅度，而 ATR 则是对 N 个周期的 TR 值加以平滑。海龟交易系统将 ATR 用作设置止损位和头寸权重配比的重要参数。

$$TR = max\left((High - Low), \left|High - Close_{pre}\right|, \left|Low - Close_{pre}\right|\right)$$

$$ATR = \frac{1}{N}\sum_{i=1}^{N} TR_i$$

公式 2

图 1 第 1 行的 4 个子图中展示了当发生价格向上突破唐奇安通道上轨和向下突破唐奇安通道下轨时，样本中未来 5，10，20，60 个交易日的涨跌幅和最大回撤的分布情况。从图中可以看出，尽管发生了向上或向下突破，但涨跌幅样本仍然集中在 0 值附近，也就是突破信号对未来涨跌幅并没有实现有效的区分。而对于样本未来的最大回撤，向上与向下突破信号也没有带来分布上的明显差异。向上与向下突破信号对未来涨跌幅带来的唯一差异在于，随着日期延长至 20 或 60 个交易日，向上突破涨跌幅样本的右侧尾部明显长于向下突破。也就是说，发生向上突破后出现大幅上涨的样本，要远远多于发生向下突破的情况，尽管这些大幅上涨样本在整个样本中所占比例仍然是极低的。这实际上与趋势交易策略“截断亏损，让利润奔跑”的思想是一致的。趋势交易策略并不确保发生建仓信号的标的就能获得很好的表现，而是通过止损手段来控制亏损的幅度，或者通过头寸配比来控制单个标的亏损对整个账户造成的影响，以可控的风险来换取不放过大行情的机会。

由于 ATR 指标是直接使用价格进行计算的，对于价位差异很大的标的很难直接进行比较。因此在构造 ATR 指标样本时，我们进行了正规化处理，即用 N 个周期的 ATR 除以股价 N 个周期的移动平均线，得到“正规化 ATR”（Normalized ATR），见公式 3。在图 1 第 3 行第 2 列的子图中，展示了正规化 ATR 与未来涨跌幅之间的散点图。可以看出，未来大幅上涨或大幅下跌的样本均集中在正规化 ATR 较小值的附近，特别是在 0.025 至 0.05 左右（即波幅在均线的 2.5%至 5%）；而正规化 ATR 当前值越大的样本，未来大幅上涨和大幅下跌的可能均较小。可以说波动幅度呈现出了一种均值回复的现象。

$$Normalized\ ATR_N = \frac{ATR_N}{SMA_N}$$

公式 3

考虑到价格向上穿过唐奇安通道上轨，或向下穿过唐奇安通道下轨仅仅是一个“瞬时”的事件，而价格与通道上下轨线关系的其他情况并没有得到测量。为了表征这一信息，我们构造了最高价与唐奇安通道上轨的差值，和最低价与唐奇安通道下轨的差值这两个指标。同样，为了使不同价位的标的可以相互比较，我们也用移动平均线对其进行了正规化处理。在图 1 第 3 行第 3、4 列的子图中，展示了这两个指标与标的未来涨跌幅的关系。最高价与唐奇安通道上轨之差的样本集中在 0 值附近，且大幅上涨与大幅下跌的样本也集中于这一位置；最高价低于上轨的幅度越大，则未来价格波动幅度越低。最低价与下轨之差的样本也集中于 0 值附近，大幅上涨或下跌的样本也集中于此，但最低价低于下轨的情况很少；而最低价高于下轨的幅度越大，未来价格波动幅度也越低。

均线系统

均线是最常用的技术分析工具，典型的均线系统由若干不同周期的移动平均线组成，例如 5、10、20、60 日均线。均线往往会被纳入其他交易系统中，就均线本身的应用而言，

最常见的是均线多头排列，即短周期的均线依次高于长周期的均线，这通常被认为预示着单边上涨行情；相反，短周期的均线依次低于长周期均线，则被称为空头排列，预示着单边下跌行情。多头、空头排列通常在主观技术分析时使用，为了以量化的方式表征这一信息，我们计算了上述几根均线的差值作为指标，包括收盘价减 5 日均线，5 日均线减 10 日均线，10 日均线减 60 日均线等。同样，为了使不同价位的标的可以相互比较，对各均线之间的差值也采用除以收盘价进行标准化处理。除了多头、空头排列以外，通常也认为各均线之间的间隔越窄，越可能预示着单边行情。因此我们也计算了各均线之间的最大距离，并进行标准化处理，作为样本中的技术指标，见公式 4。

$$Max\ Distance = \frac{\max\limits_{i,j=5,10,20,60}\left(\left|SMA_i - SMA_j\right|\right)}{Close}$$ 公式 4

图 2 第 1 行各子图展示了各均线之间差值与未来涨跌幅之间的散点图，图 2 第 2 行第 1 列展示了各均线之间的最大距离与未来涨跌幅之间的散点图。可以看出，各均线价差在 0 值附近的样本，未来涨跌幅的绝对值最大，偏离 0 值越远的样本未来涨跌幅的绝对值越小。均线之间的最大距离也展现了同样的效果，即大幅上涨或下跌都出现在最大距离靠近 0 值的位置，均线间最大距离越大则涨跌幅度越小。

MACD 指标

“平滑异同移动平均线”（Moving Average Convergence Divergence，MACD）也是被广泛使用的技术分析工具。MACD 本质上也是一种均线系统，只是 MACD 基于指数移动平均线（EMA），相对于简单移动平均（SMA）而言，EMA 提高了最近周期价格的权重，从而提高了跟随价格的敏感性。一般而言，MACD 的计算分为几个步骤：首先计算 12 与 26 个周期的 EMA；然后计算 12 周期 EMA 减去 26 周期 EMA 的差值，被称为 MACD 线；在对 MACD 线计算 9 周期的 EMA，被称为 MACD Signal 线；最后计算 MACD 和 MACD Signal 的差值，被称为 MACD Histogram 线，见公式 5。从计算方法上看，MACD 考虑了不同周期 EMA 的差值，并对其进行平滑，测量的是 EMA 差值及其平滑线之间的关系。

$$MACD = EMA_{12} - EMA_{26}$$

$$MACD\ Signal = EMA_9(MACD)$$ 公式 5

$$MACD\ Histogram = MACD - MACD\ Signal$$

MACD 的应用方法非常多，在这里我们只选择最简单和易于量化的 MACD“金叉”（MACD Crossover），即 MACD 线自下而上穿过 MACD Signal 线；和 MACD“死叉”（MACD Crossunder），即 MACD 线自上而下穿过 MACD Signal 线，作为样本中的指标。MACD“金叉”通常被用作上涨行情的信号，MACD“死叉”则被用作下跌行情的信号。我们将 MACD“金叉”、“死叉”、MACD、MACD Signal 和 MACD Histogram 纳入样本指标。由于 MACD 依据 EMA 计算，同样受到股价水平的影响，因此我们也对 MACD、MACD Signal 和 MACD Histogram 进行了正规化处理。在图 1 第 2 行子图中展示了发生 MACD“金叉”、“死叉”情况下，未来涨跌幅和最大回撤的分布情况。与唐奇安通道突破类似，MACD“金叉”、“死叉”同样没有实现对未来涨跌幅和最大回撤的有效区分。图 2 的第 2 行第 3、4、5 列的子图展示了 MACD、MACD Signal 和 MACD Histogram 值与未来涨跌幅的关系。它们的共同特点是，当接近 0 值时与未来大幅上涨或下跌相对应，而偏离 0 值越远，则未来涨跌波动幅度越小。

其他指标

除了上述指标外，我们的样本中还包括了用于衡量交易活跃度的换手率（Turnover Ratio）指标；用于衡量趋势强度的“平均趋向指标”（Average Directional Indicator，ADX）和“商品通道指标”（Commodity Channel Index，CCI）；以及用于度量价格时间序列波动特征的 Hurst 指数。其中，ADX 的计算步骤见公式 6，可以看出 ADX 指标实际上是对价格波动幅度的一系列平滑和标准化处理[2]，以便一种标准化的方式来衡量当前趋势的强弱。CCI 指标的计算步骤见公式 7[3]，其意义在于提供了一个简单的数字表示，从统计上说明了最近价格偏离均线的程度。

$$+DM = High_{current} - High_{previous}$$

$$-DM = Low_{previous} - Low_{current}$$

$$+DI = \frac{SMA_N(+DM)}{ATR}$$

$$-DI = \frac{SMA_N(-DM)}{ATR}$$

公式 6

$$DX = \frac{|+DI - -DI|}{|+DI + -DI|}$$

$$ADX = SMA_N(DX)$$

$$TP = \frac{(High + Low + Close)}{3}$$

$$MD = \frac{\sum_{i=1}^{N}|TP_i - SMA_N(TP)|}{N}$$

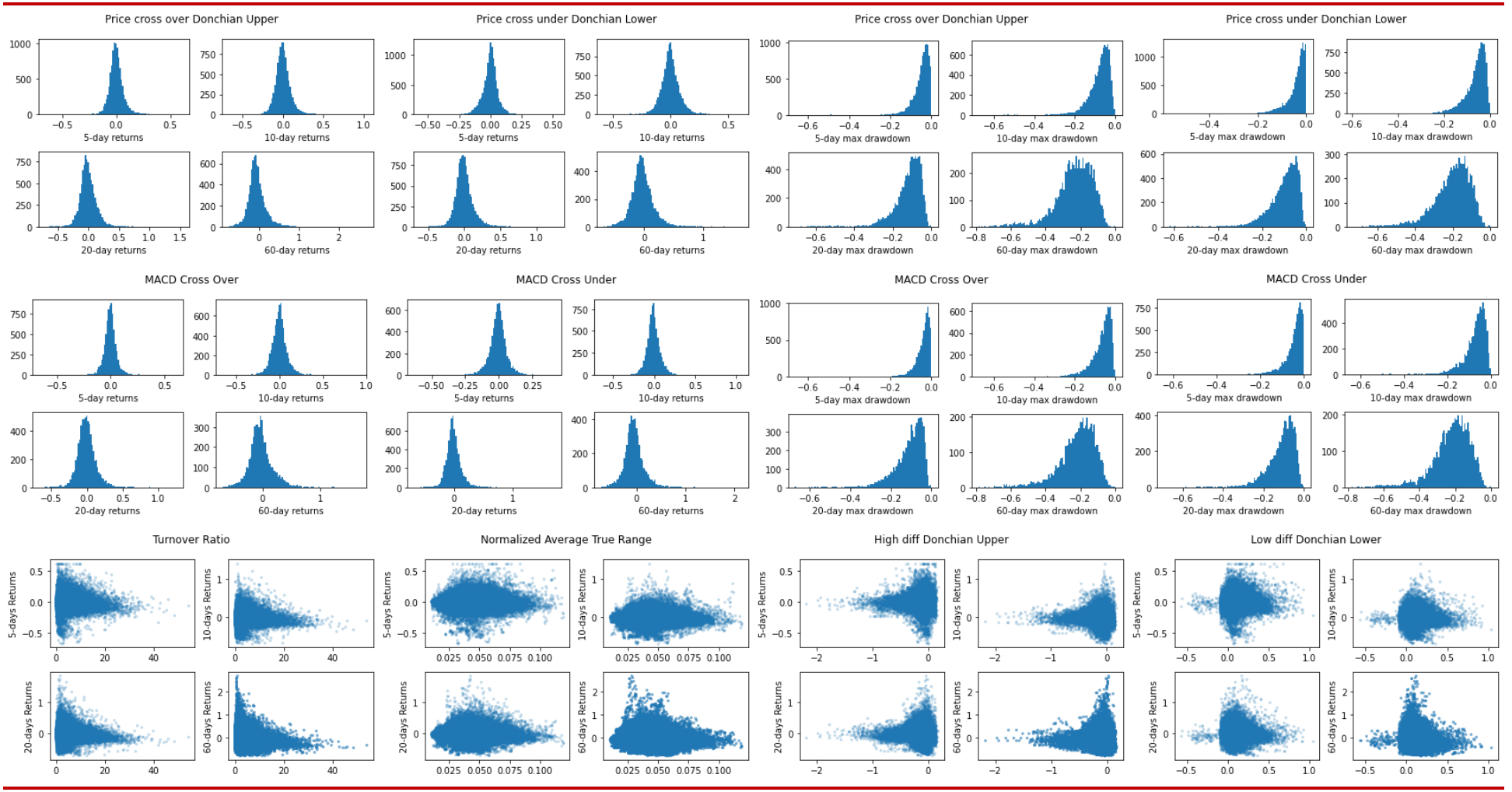
公式 7

$$CCI = \frac{TP - SMA_N(TP)}{0.015 \times MD}$$

图 1 第 3 行第 1 列，图 2 第 4 行各子图展示了换手率、50 和 100 周期的移动 Hurst 指数、14 周期的 ADX 和 CCI 等指标与未来涨跌幅的散点图。未来大幅上涨和大幅下跌的样本与换手率、ADX、CCI 接近 0 值相对应。这是因为作为活跃度和趋势强弱的衡量，这些指标在趋势接近结束时才会出现最大值，而趋势启动的时机则正好相反，位于 0 值附近。Hurst 指数衡量的是时间序列的持续性倾向，当 Hurst 指数值远大于 0.5 时意味着时间序列出现趋势，而接近 0.5 时则时间序列接近随机游走。图中，未来大幅涨跌行情与 Hurst 指数在 0.6 附近相对应，这是价格时间序列在震荡行情中最常见的取值。过高的 Hurst 指数往往意味着趋势行情的结束，而过低的 Hurst 指数则意味着价格仍然处于胶着，因而都与未来大行情无关。

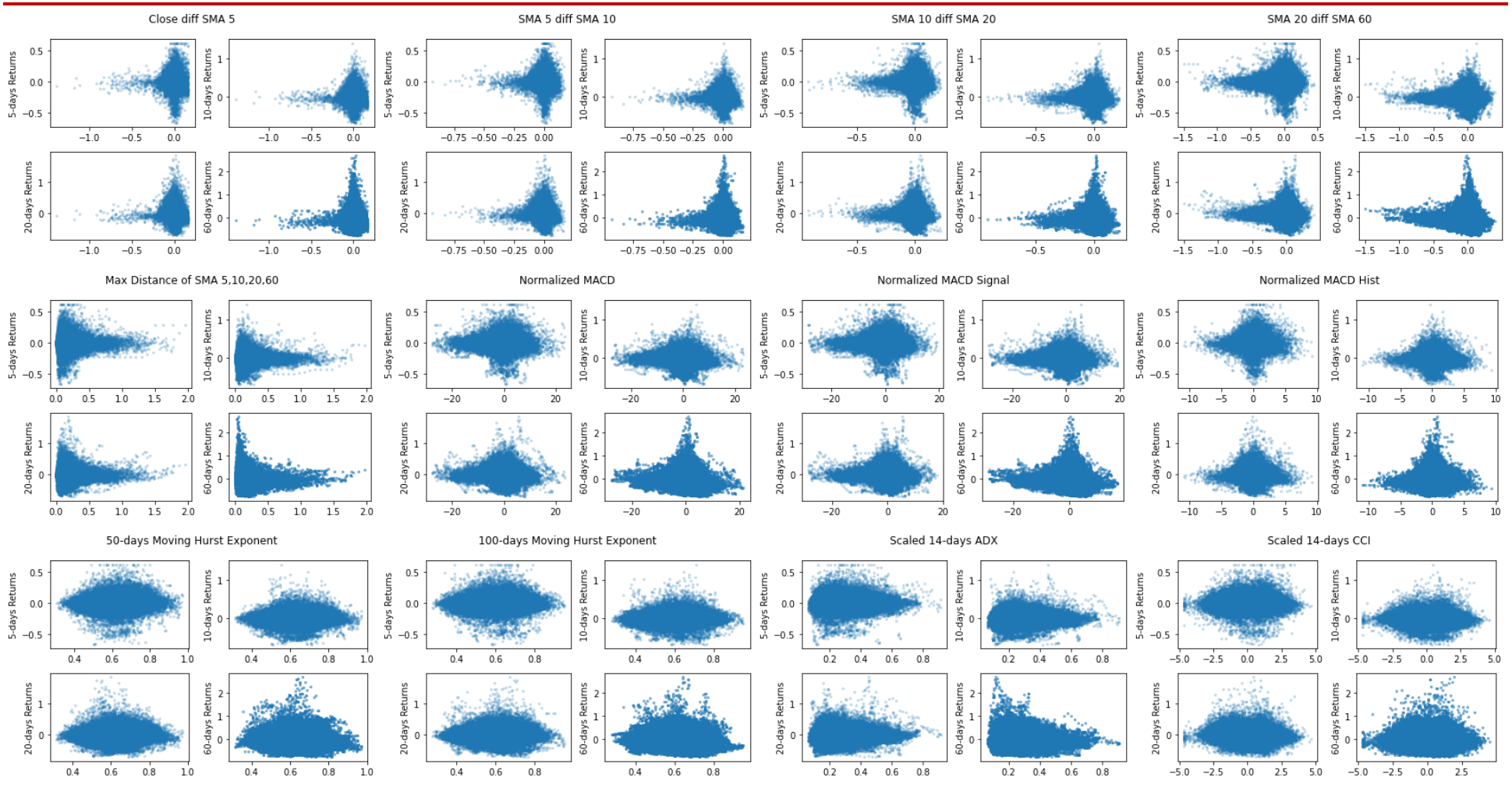
[2] 在标准的 ADX 计算中，上述公式还要乘以 100，让 ADX 值落入$(0,100)$的区间，以便在技术分析中符合人的阅读习惯。但本文所涉及的指标计算样本大多在$(-1,1)$的区间，为了让模型训练时各指标量纲比较接近，因此不作乘以 100 的处理。
[3] 公式 7 中 TP 为“典型价格”（Typical Price），MD 为“平均差”（Mean Deviation）。

图 1 (各技术指标与未来涨跌幅、最大回撤之间的关系（一）)



资料来源： 东海证券

图 2 (各技术指标与未来涨跌幅、最大回撤之间的关系（二）)



资料来源： 东海证券

3.模型训练框架

通过上述对数据的初步观察，我们可以看出任何单一的技术指标都不能完全地给出有关个股未来表现的信息，但每个技术指标又在某些侧面给出了一定程度上有效的信息。那么接下来的问题是，如果综合使用这些技术指标，是否可以形成对个股未来表现的有效判断，或者提高预测的准确率？这就是我们希望通过机器学习可以实现的目标。

转换为分类问题

机器学习模型可以被视为一个函数：通过输入一组数据 x，这通常被称为“特征”（feature），得到一个预测值 y 作为输出。所谓“训练”模型，就是通过学习大量的样本数据，得到这样的一个“函数”。机器学习系统的种类繁多，如果根据训练过程中是否有人监督（也就是是否拥有已知答案 y 的样本），受到监督的数量和监督类型，可以分为监督式学习、无监督式学习、半监督式学习和强化学习。由于价格时间序列数据完整、质量高、高度量化且易于获取，本文中所涉及的对个股未来表现的预测问题可以归为典型的监督式学习问题。

监督式学习通常又可以分为分类（Classification）任务和回归（Regression）任务，前者的输出是离散值，代表不同的类别；后者输出的是一个具体数值，预测得到的模型是一个连续函数。从我们的预测目标来看，刻画个股表现的数据都应该是连续的，例如涨跌幅、最大回撤、年化收益率、日均收益率、夏普比率等等。因此从直观上来看，这应该是一个典型的回归问题。然而，正如我们在前文中提到，由于资本市场本身的复杂性，价格时间序列的预测要比“物理”世界中的预测问题困难得多，成功率极低。为了尽可能获得有价值的训练结果，我们将未来 N 个交易日的涨跌幅和最大回撤按表 1 所示规则进行分类，从而将预测目标从一个回归问题转化为一个分类问题。根据表 1 的规则，未来涨跌幅按从低到高顺序被分为从 0 到 7 共 8 个级别，其中涨跌幅为负和涨跌幅为正各 4 个级别，级别 0 表示跌幅最大，级别 7 表示涨幅最大。未来最大回撤也按从低到高分为从 0 到 4 共 5 个级别，级别 0 表示回撤最大，级别 4 表示回撤最小。

表 1　涨跌幅和最大回撤分类

涨跌幅（r）	涨跌幅级别	最大回撤（d）	最大回撤级别
$r < -20\%$	0	$d < -50\%$	0
$-20\% \le r < -10\%$	1	$-50\% \le d < -20\%$	1
$-10\% \le r < -5\%$	2	$-20\% \le d < -10\%$	2
$-5\% \le r < 0$	3	$-10\% \le d < -5\%$	3
$0 \le r < 5\%$	4	$d \ge -5\%$	4
$5\% \le r < 10\%$	5		
$10\% \le r < 20\%$	6		
$r \ge 20\%$	7		

资料来源：东海证券

模型训练方案

在将预测问题确定为分类问题后，就可以制定具体的模型训练方案。方案的第一步是构造满足模型训练要求的样本数据集。我们选取了 2017 年至 2019 年 A 股全市场数据，并剔除了停牌等不符合样本要求的情况。对于任意个股在任意交易日，只要能满足计算构

成样本特征值的条件，均可作为一个有效的样本数据点。对于每个样本数据点，我们可以按以下图 3 所示的方式构造样本：首先是用样本时间点之前的历史数据，计算各类技术指标作为特征值，也就是样本的 x 向量；然后用样本时间点之后的“未来”数据，计算涨跌幅、最大回撤等衡量个股未来表现的指标，并按表 1 所示规则将涨跌幅（Equity Return）转换为“涨跌幅级别”（Equity Return Level），将最大回撤（Max Drawdown）转换为“最大回撤级别”（Max Drawdown Level）作为样本的 y 向量。其中“未来”数据部分分别统计未来 5、10、20、60 个交易日的数据。

图 3 (样本点计算示意图)



资料来源： 东海证券

这样一来，我们将采用同样一组 x 向量分别去预测 y 向量中的涨跌幅级别和最大回撤级别这两个目标值，这实际上是两个不同的问题。同时我们预测的周期包括未来 5、10、20、60 个交易日的不同情况，因此我们实际的预测任务共有 8 个。

$$\begin{cases} F_{return\ level\ n}(\vec{x}) = y_{return\ level\ n} \\ F_{drawdown\ level\ n}(\vec{x}) = y_{drawdown\ level\ n} \end{cases}$$

公式 8

其中：$n = 5, 10, 20, 60$

初步完成样本数据的构造之后，机器学习模型训练之前往往还需要对样本数据进行一系列的预处理，以使数据更适于展开训练。例如，scikit-learn 中就提供了一系列数据预处理工具，如标准化（StandardScaler）、区间缩放（MinMaxScaler）、正态化（Normalization）等。除此之外还有一些特殊的预处理，比如将特征值转化为多项式特征（Polynomial Features），以便用高阶多项式对数据进行拟合。但从我们的实践上看，scikit-learn 中的一般性预处理工具并没有实现模型训练效果的改善。这可能是因为这类一般性的预处理方法并没有考虑问题的领域知识。因此在模型训练中我们主要采用一些自行设计的预处理方法，例如在上一章提到的，用股价或均线来实现技术指标的无量纲化等等。

我们在样本数据集中随机抽取了 10%的数据作为测试集，剩下的 90%的数据作为训练集数据。在分类测试集和训练集数据时，我们也没有直接使用 scikit-learn 中提供的 train_test_split 函数，而是自行进行了抽取。这是因为，除了按一般的机器学习模型评估方法外，我们还要在预测结果的基础上进行选股策略的回归测试，因而必须保留个股代码、时间点等信息，以免在测试集中混入训练集中的数据。

我们用于训练的分类模型包括：逻辑回归（Logistic Regression）、岭回归（Ridge regression）、K 邻近算法（K Neighbors）、高斯朴素贝叶斯（Gaussian NB）、决策树（Decision Tree）、随机森林（Random Forest）、极端树（Extra Tree）、多层感知机（Multilayer Perceptron）等等。

在训练过程中，我们首先采用 K 折交叉验证法将训练集数据进行拆分，进行交叉验证，以此来评估模型训练的结果。接下来再进一步利用整个训练集数据正式进行模型训练，并将完成训练的模型序列化。最后利用序列化的模型，利用测试集样本数据进行测试，评估模型的预测结果，并利用预测结果作为选股策略进行回归测试。

4.模型训练结果评估

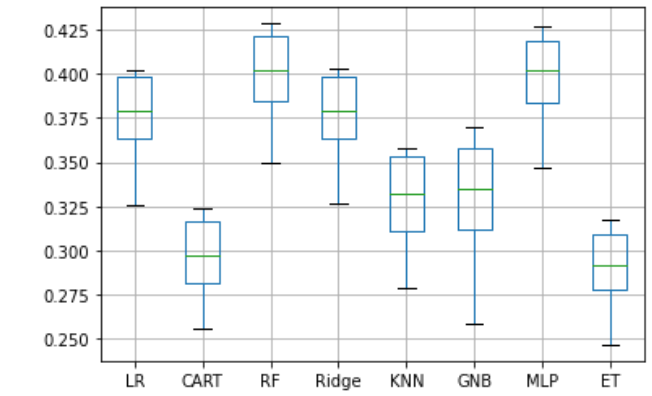
为了评估模型训练的结果，从多种分类模型中选择适当的模型进行应用，我们主要采用 3 个验证步骤：首先是在模型训练过程中，从训练集数据中拆分出部分数据作为评估数据，进行交叉验证，初步评价各模型的效果；其次是在利用整个训练集数据正式完成模型训练后，用测试集数据检验模型训练的结果；最后是针对选股这一特殊问题领域，将测试集数据按照涨跌幅级别的预测值分组，各组分别作为投资组合进行回归测试，以考察模型的实际应用潜力。

交叉验证评分

在 scikit-learn 中提供了多种分离训练数据集和评估数据集的算法，比较常用的包括 K 折交叉验证分离、弃一交叉验证分离、重复随机分离等等。在本文中，我们采用的是 K 折交叉验证分离，其原理是将训练集数据分成 K 组，将每个子集数据分别做一次验证集，剩余 K-1 组子集作为训练集，这样会得到 K 个模型；最后利用这 K 个模型最终的验证集的分类准确率的均值，作为此 K 折交叉验证下分类器的性能指标。

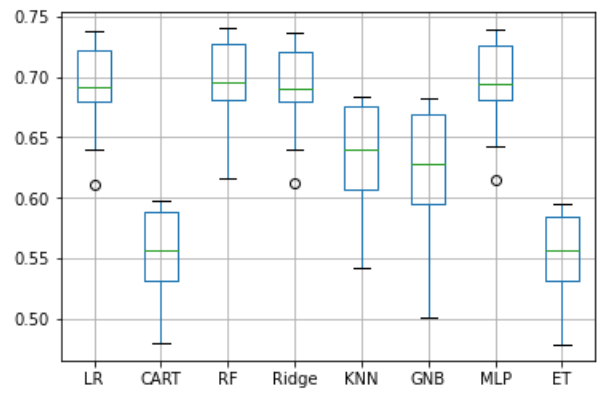
由于我们的训练方案中，有未来涨跌幅和最大回撤 2 个训练目标，涉及未来 5、10、20、60 个交易日的训练任务，因此我们考察的所有模型都经过了 8 次 K 折交叉验证。在所有模型的交叉验证中，K 都取值 10，因此在每个模型的每次交叉验证中都会得到 10 个准确度评分。以下图 4 至图 11 以箱线图的形式展示了对于每个训练任务，各模型交叉验证的得分情况。

图 4 （未来 5 交易日收益率预测交叉验证）



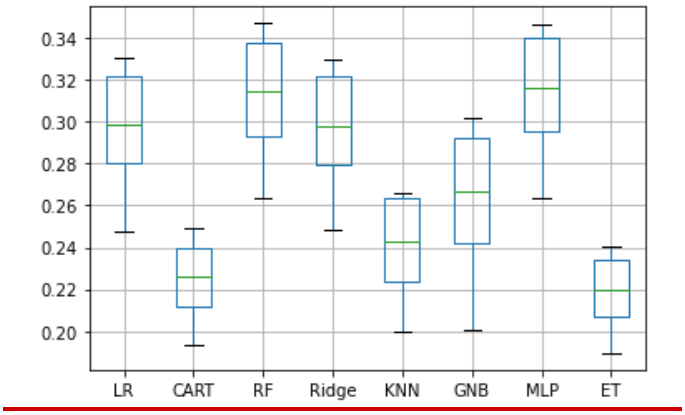
资料来源：东海证券

图 5 （未来 5 交易日最大回撤预测交叉验证）



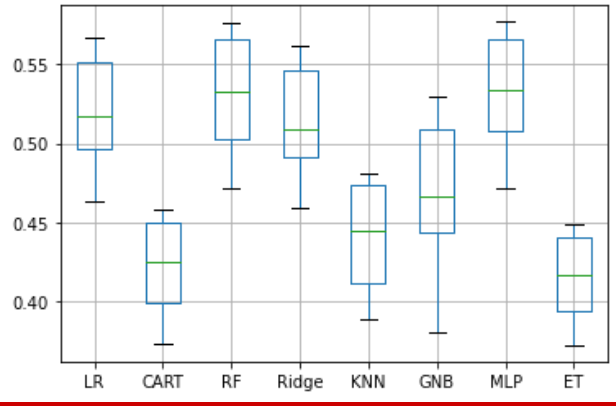
资料来源：东海证券

图 6 (未来 10 交易日收益率预测交叉验证)



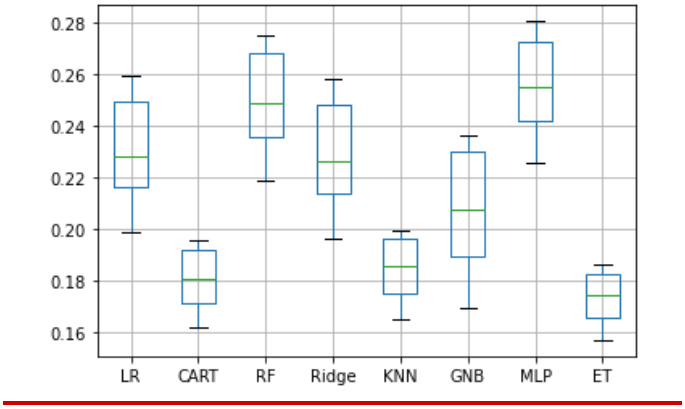
资料来源：东海证券

图 7 (未来 10 交易日最大回撤预测交叉验证)



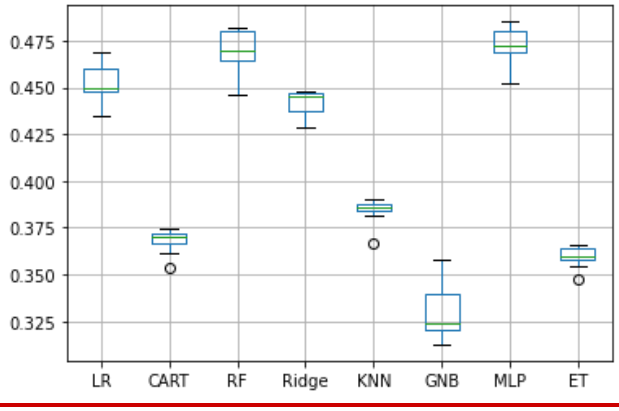
资料来源：东海证券

图 8 (未来 20 交易日收益率预测交叉验证)



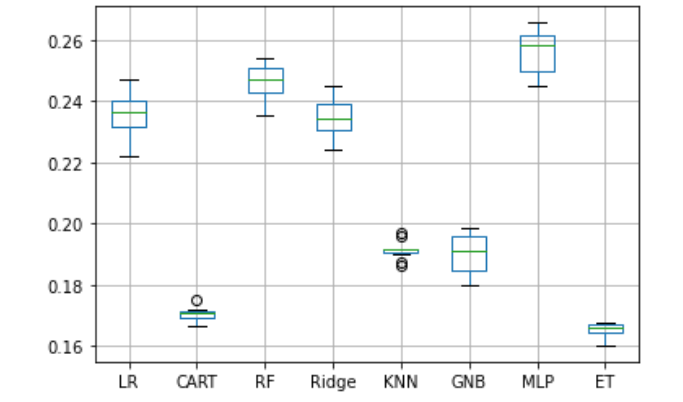
资料来源：东海证券

图 9 (未来 20 交易日最大回撤预测交叉验证)



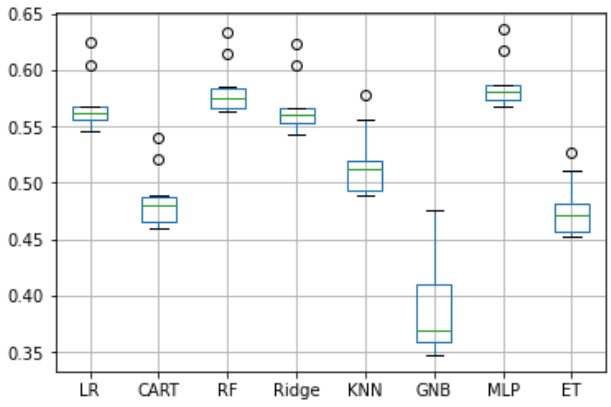
资料来源：东海证券

图 10 (未来 60 交易日收益率预测交叉验证)



资料来源：东海证券

图 11 (未来 60 交易日最大回撤预测交叉验证)



资料来源：东海证券

从图 4 至图 11 可以看出交叉验证结果呈现出几个特征：对涨跌幅的预测准确率远低于对最大回撤的预测准确率，涨跌幅预测准确率最高的也只能达到略微超过 40%，而最大回撤的预测准确率可以达到接近 70%；预测准确率随预测周期延长而下降；随机森林（RF）和多层感知器（MLP）两个模型的预测准确率在所有预测任务中都效果最好。

预测误差分析

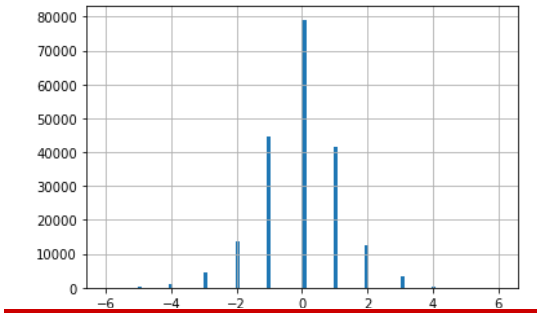
尽管单纯从交叉验证结果来看，尤其是涨跌幅预测任务的交叉验证结果并不理想，但用作选股策略的预测任务有其特殊性：在一般的分类预测任务中，只要预测的类别不同于实际的类别，就算一次预测错误。但不要忘记我们的分类问题实际是从连续的涨跌幅或最大回撤数据转换而来的。也就是说，即使没能准确地预测出一个样本数据点对应的未来涨跌幅或最大回撤级别，但实际偏差不大，那么这样的预测仍然是具有价值的。特别地，对于涨跌幅来说，如果能正确地预测其方向，预测幅度的偏差对于整个投资组合来说也不会造成太大的影响。所以，我们除了要从机器学习的角度去看待交叉验证的结果，还要从问题领域的角度进一步去考察这一问题。

为此，我们接下来用整个训练集数据完成了对涨跌幅和最大回撤在各个预测周期的训练任务，并用最初分离出的测试集数据（测试集数据既没参与模型训练也没参与交叉验证）完成对各模型的检验。通过测试集数据检验得到的各模型评分与交叉验证得到的平均分基本一致。然后，针对选股这一问题领域，我们采用了两种特殊的预测误差分析方法：

一是比较涨跌幅和最大回撤级别的预测值与实际级别之间的差异，具体做法就是用预测级别减去实际级别之差作为预测误差。如果差值为 0，说明实现了准确预测；如果差值为正，说明发生了高估；如果差值为负，说明发生的低估。图 12 至图 19 展示了评分最高的随机森林（RF）模型和多层感知器（MLP）模型，对未来 5、10、20、60 个交易日的涨跌幅的预测误差。从图中可以看出，尽管预测准确率不高，且随预测周期的延长而下降，但是所有预测任务的预测误差分布，均呈现出以 0 值为中心向两侧快速递减的现象。也就是说，即使从分类任务的角度看预测准确率不高，但预测误差在 1 个级别以内的仍然占大多数。还有一个现象是，在预测准确率较低的 20 和 60 个交易日，0 值左侧的分布明显多于右侧。这意味着涨跌幅被低估的情况明显多于被高估的情况。

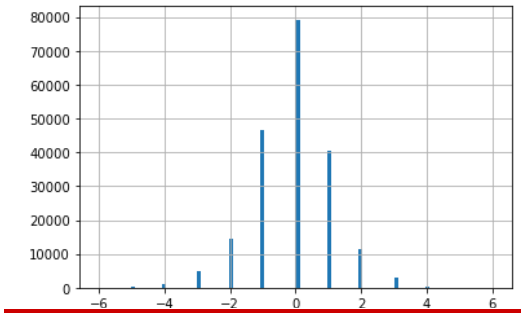
再一个是逐个比较各个预测级别中实际的涨跌幅的分布情况。图 20 至图 51 展示了随机森林（RF）在各个预测级别的实际涨跌幅分布，图 52 至图 83 展示了多层感知器（MLP）在各个预测级别的实际涨跌幅分布。具体而言，每行的子图代表了 1 个预测级别的分布情况，每列代表了 1 个预测周期的分布情况。就各个预测级别的涨跌幅分布变化来看，从 0 级到 7 级，分布的中心值逐步从 0 值的左侧向右侧移动。这说明无论是随机森林还是多层感知器，均实现了对未来涨跌幅比较有效的区分。另须注意的是，对于 5 个交易日和 10 个交易日的预测任务，其级别为 0（跌幅最大级别）和级别为 7（涨幅最大级别）的预测值总体均比较少。这与大多数个股的股价在较短的时间周期内变化幅度有限是一致的。

图 12 (RF 未来 5 交易日回报预测误差)



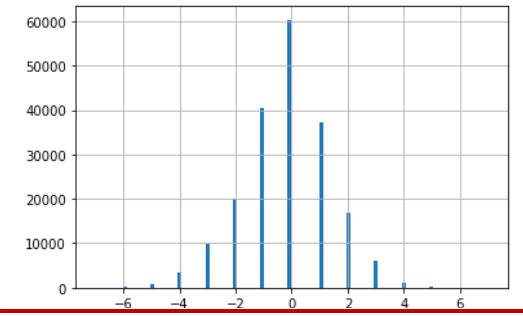
资料来源：东海证券

图 13 (MLP 未来 5 交易日回报预测误差)



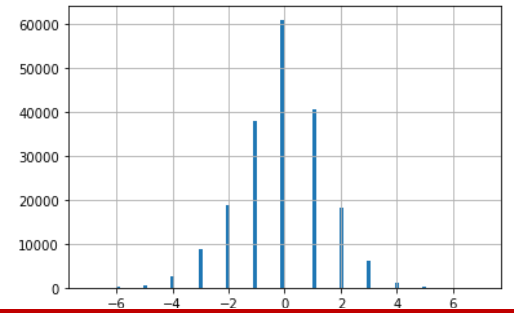
资料来源：东海证券

图 14 (RF 未来 10 交易日回报预测误差)



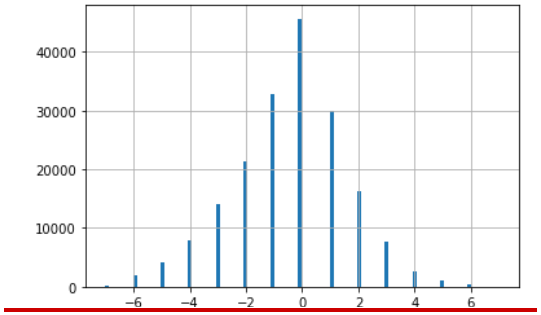
资料来源：东海证券

图 15 (MLP 未来 10 交易日回报预测误差)



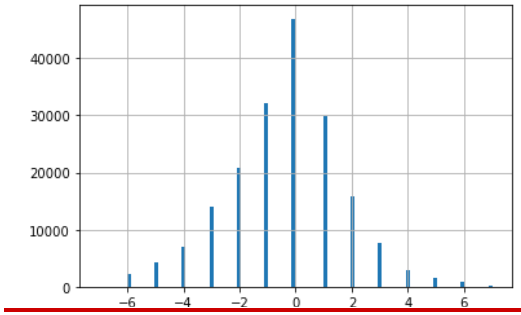
资料来源：东海证券

图 16 (RF 未来 20 交易日回报预测误差)



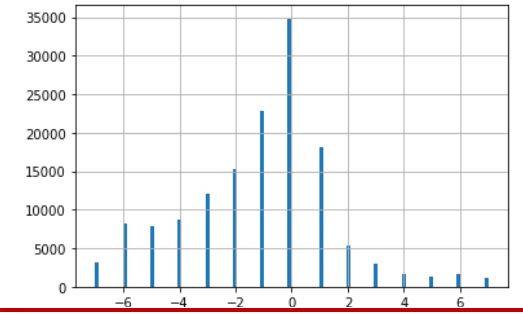
资料来源：东海证券

图 17 (MLP 未来 20 交易日回报预测误差)



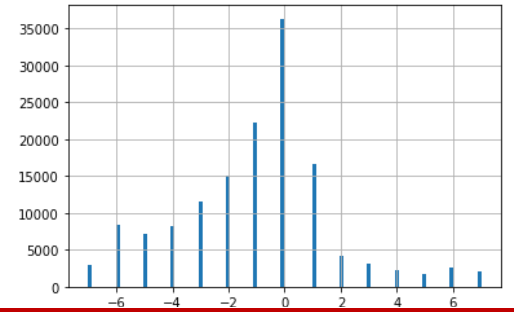
资料来源：东海证券

图 18 (RF 未来 60 交易日回报预测误差)



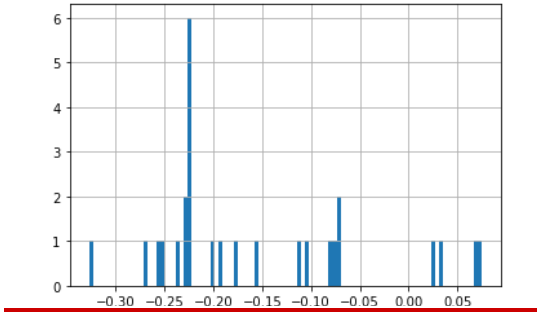
资料来源：东海证券

图 19 (MLP 未来 60 交易日回报预测误差)



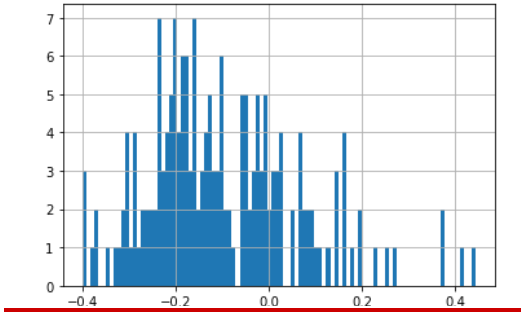
资料来源：东海证券

图 20 (RF_0 未来 5 交易日回报分布)



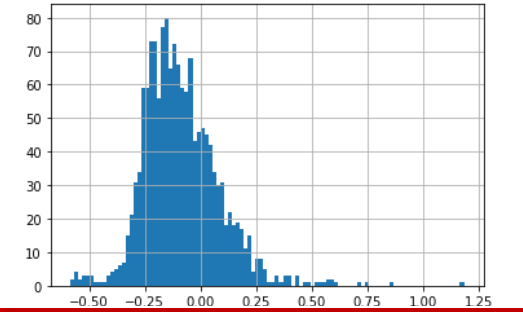
资料来源：东海证券

图 21 (RF_0 未来 10 交易日回报分布)



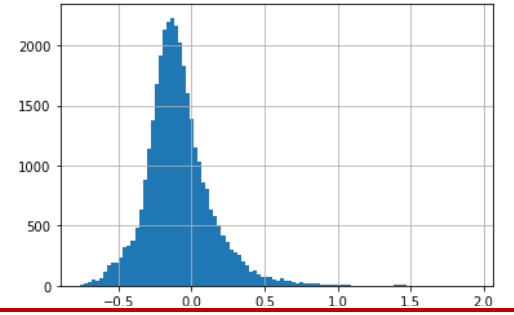
资料来源：东海证券

图 22 (RF_0 未来 20 交易日回报分布)



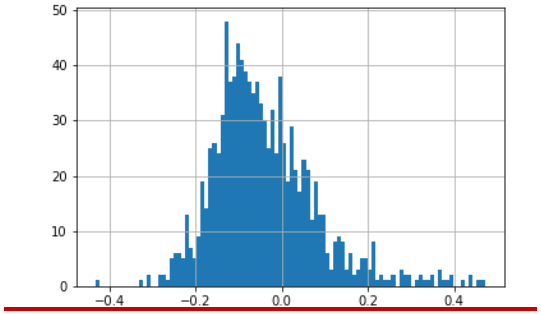
资料来源：东海证券

图 23 (RF_0 未来 60 交易日回报分布)



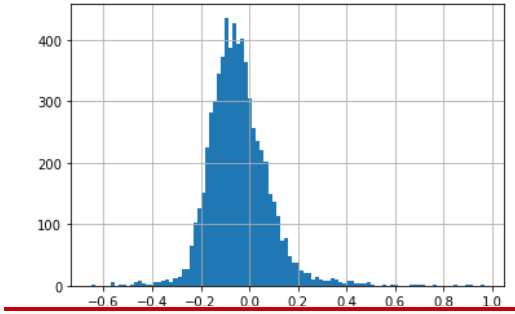
资料来源：东海证券

图 24 (RF_1 未来 5 交易日回报分布)



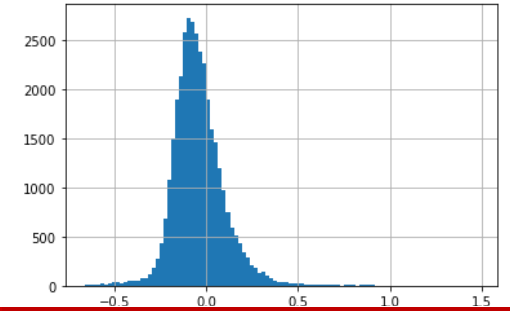
资料来源：东海证券

图 25 (RF_1 未来 10 交易日回报分布)



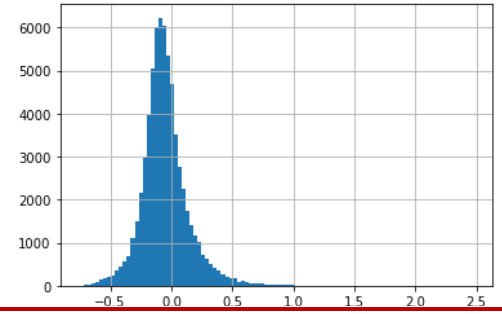
资料来源：东海证券

图 26 (RF_1 未来 20 交易日回报分布)



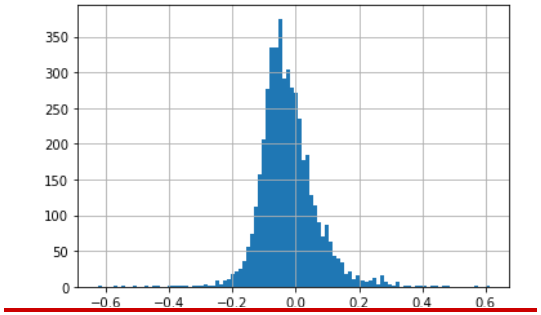
资料来源：东海证券

图 27 (RF_1 未来 60 交易日回报分布)



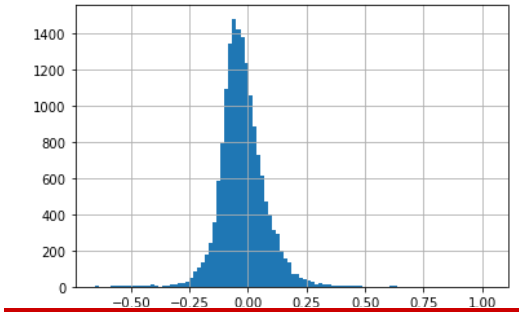
资料来源：东海证券

图 28 (RF_2 未来 5 交易日回报分布)



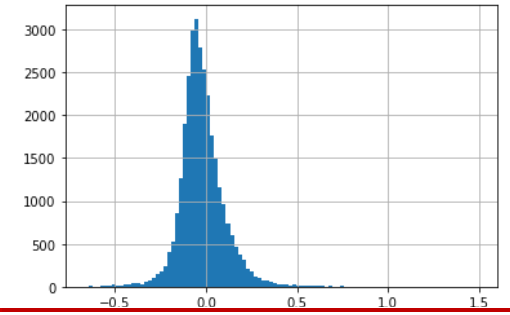
资料来源：东海证券

图 29 (RF_2 未来 10 交易日回报分布)



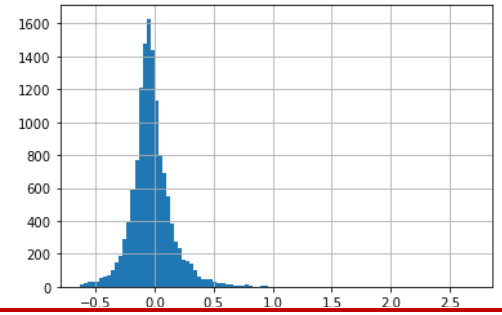
资料来源：东海证券

图 30 (RF_2 未来 20 交易日回报分布)



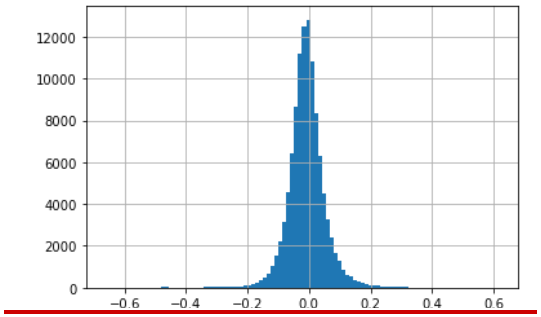
资料来源：东海证券

图 31 (RF_2 未来 60 交易日回报分布)



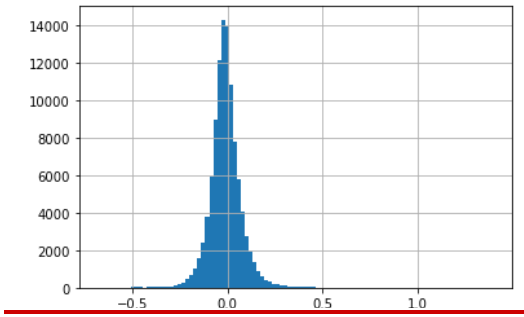
资料来源：东海证券

图 32 (RF_3 未来 5 交易日回报分布)



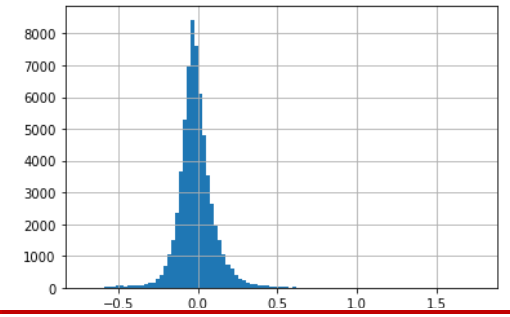
资料来源：东海证券

图 33 (RF_3 未来 10 交易日回报分布)



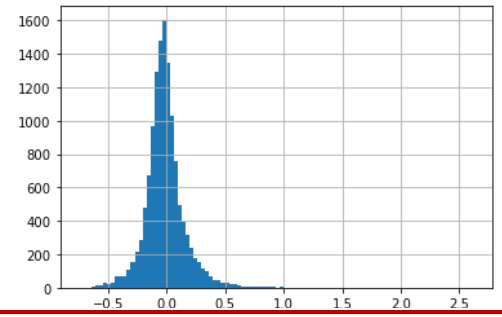
资料来源：东海证券

图 34 (RF_3 未来 20 交易日回报分布)



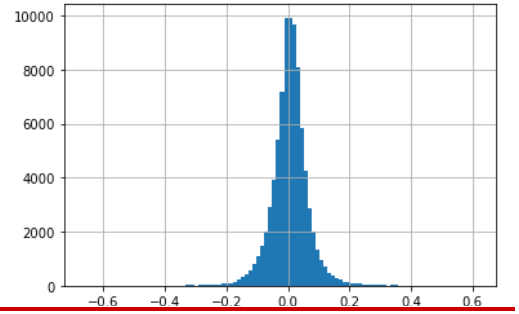
资料来源：东海证券

图 35 (RF_3 未来 60 交易日回报分布)



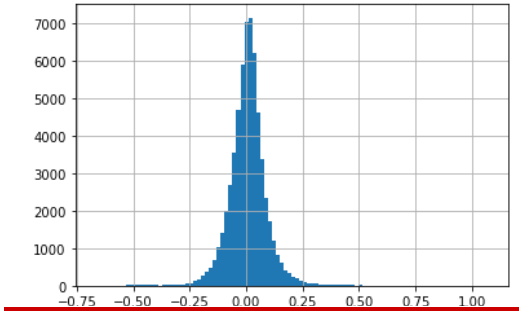
资料来源：东海证券

图 36 (RF_4 未来 5 交易日回报分布)



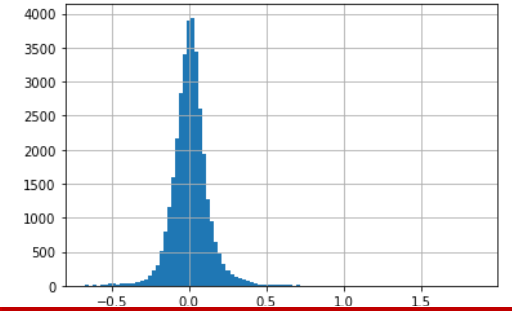
资料来源：东海证券

图 37 (RF_4 未来 10 交易日回报分布)



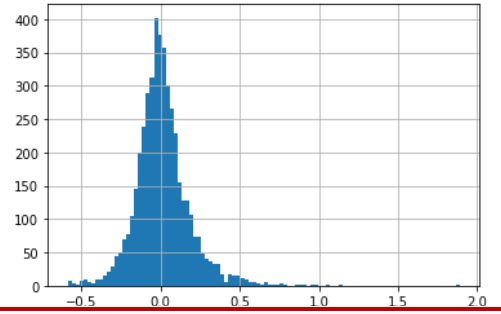
资料来源：东海证券

图 38 (RF_4 未来 20 交易日回报分布)



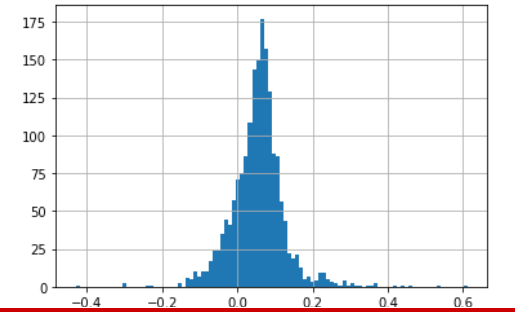
资料来源：东海证券

图 39 (RF_4 未来 60 交易日回报分布)



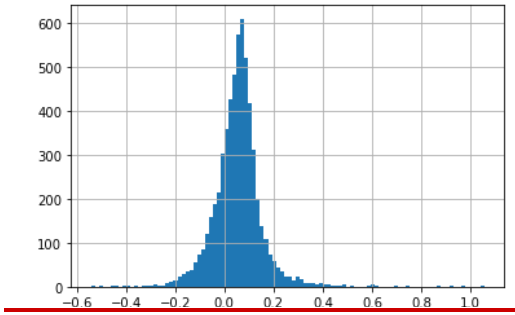
资料来源：东海证券

图 40 (RF_5 未来 5 交易日回报分布)



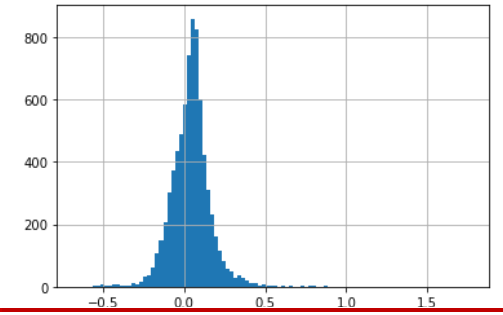
资料来源：东海证券

图 41 (RF_5 未来 10 交易日回报分布)



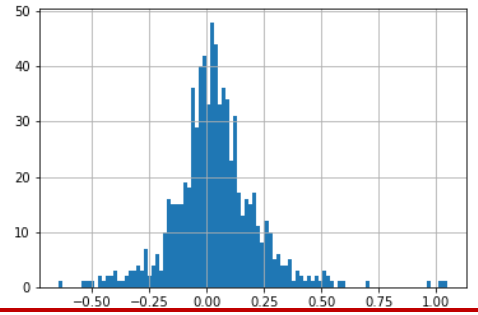
资料来源：东海证券

图 42 (RF_5 未来 20 交易日回报分布)



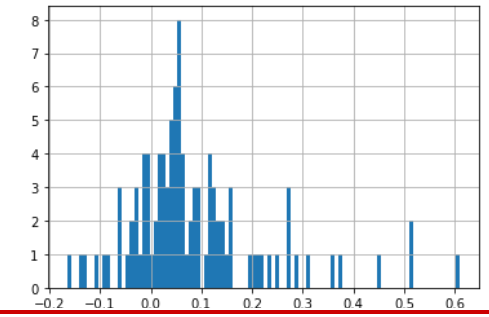
资料来源：东海证券

图 43 (RF_5 未来 60 交易日回报分布)



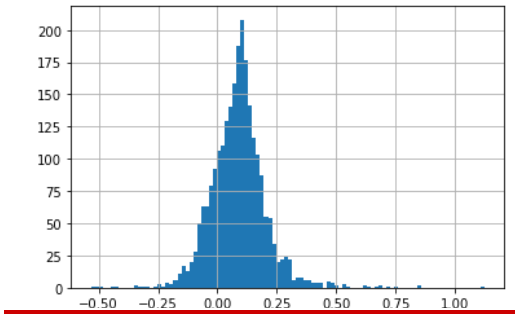
资料来源：东海证券

图 44 (RF_6 未来 5 交易日回报分布)



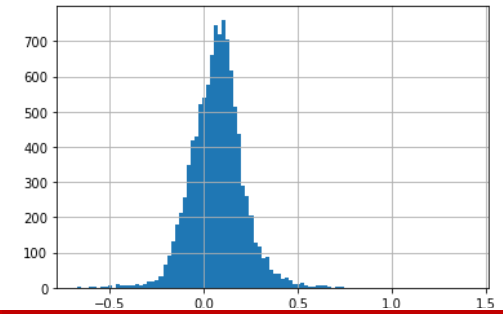
资料来源：东海证券

图 45 (RF_6 未来 10 交易日回报分布)



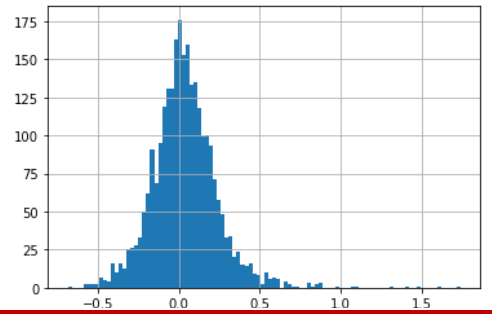
资料来源：东海证券

图 46 (RF_6 未来 20 交易日回报分布)



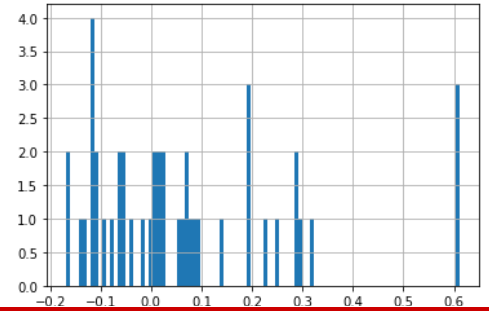
资料来源：东海证券

图 47 (RF_6 未来 60 交易日回报分布)



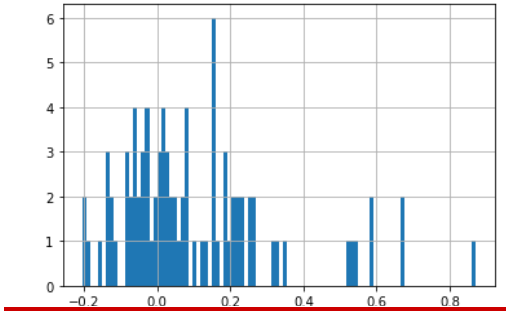
资料来源：东海证券

图 48 (RF_7 未来 5 交易日回报分布)



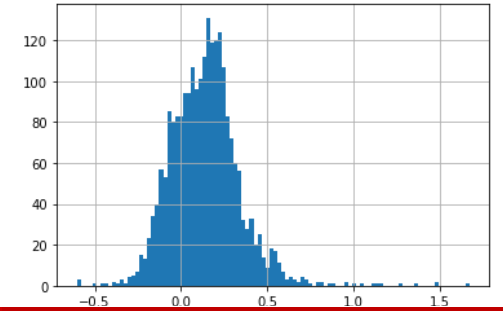
资料来源：东海证券

图 49 (RF_7 未来 10 交易日回报分布)



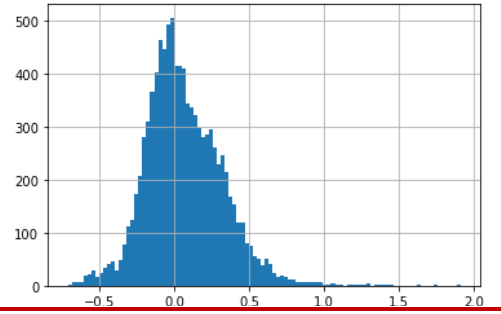
资料来源：东海证券

图 50 (RF_7 未来 20 交易日回报分布)



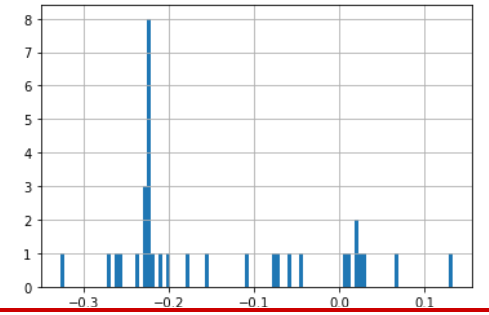
资料来源：东海证券

图 51 (RF_7 未来 60 交易日回报分布)



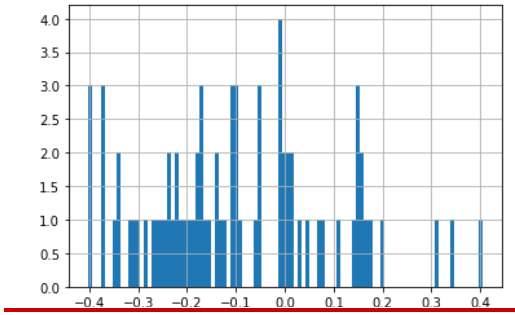
资料来源：东海证券

图 52 (MLP_0 未来 5 交易日回报分布)



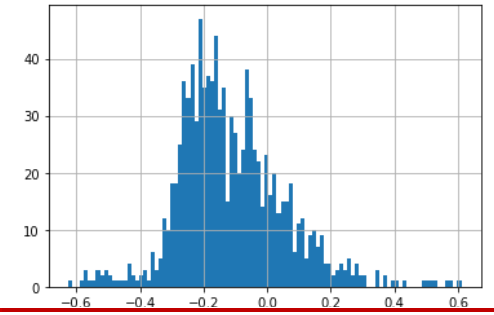
资料来源：东海证券

图 53 (MLP_0 未来 10 交易日回报分布)



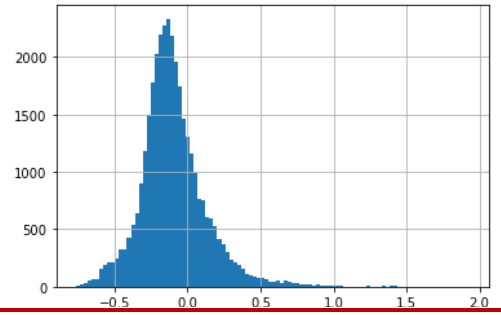
资料来源：东海证券

图 54 (MLP_0 未来 20 交易日回报分布)



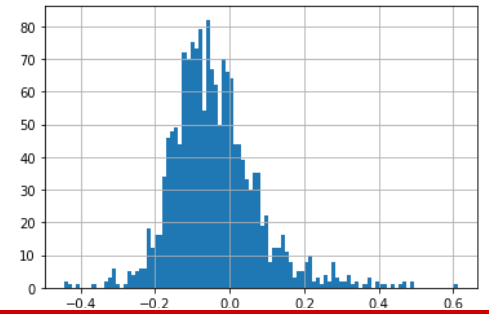
资料来源：东海证券

图 55 (MLP_0 未来 60 交易日回报分布)



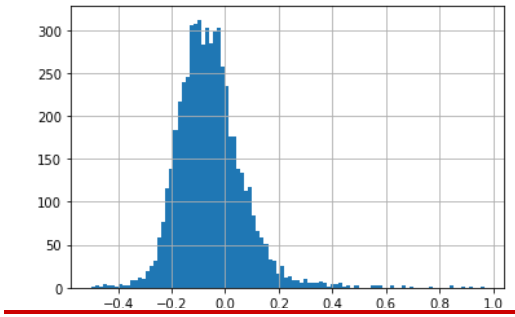
资料来源：东海证券

图 56 (MLP_1 未来 5 交易日回报分布)



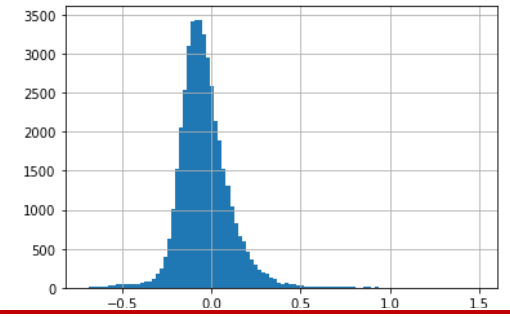
资料来源：东海证券

图 57 (MLP_1 未来 10 交易日回报分布)



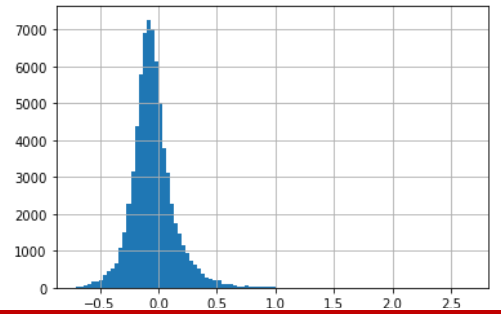
资料来源：东海证券

图 58 (MLP_1 未来 20 交易日回报分布)



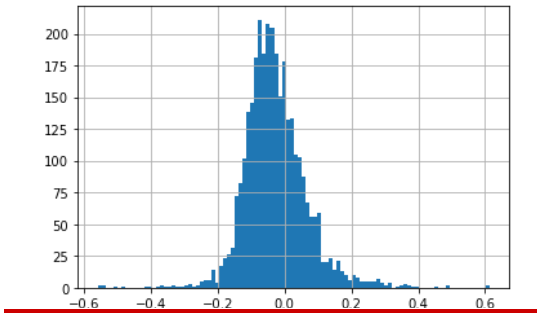
资料来源：东海证券

图 59 (MLP_1 未来 60 交易日回报分布)



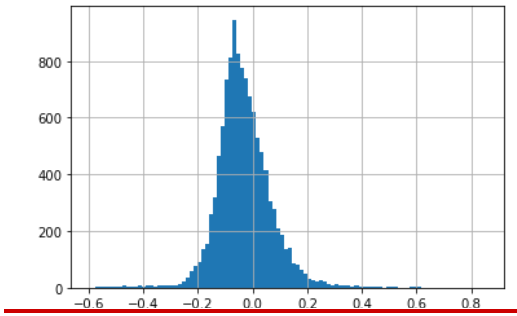
资料来源：东海证券

图 60 (MLP_2 未来 5 交易日回报分布)



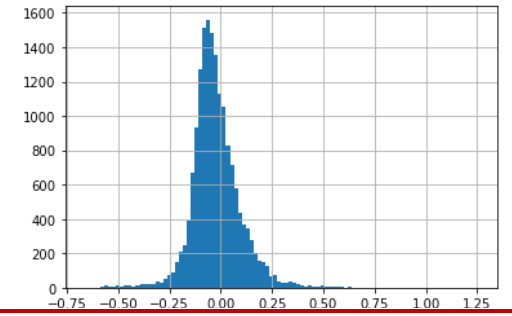
资料来源：东海证券

图 61 (MLP_2 未来 10 交易日回报分布)



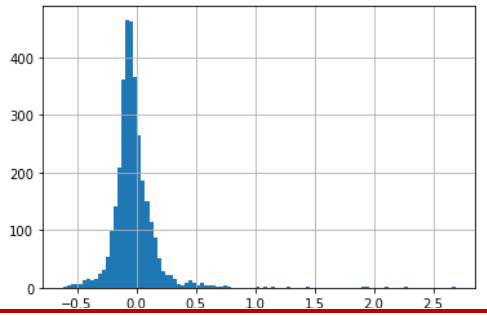
资料来源：东海证券

图 62 (MLP_2 未来 20 交易日回报分布)



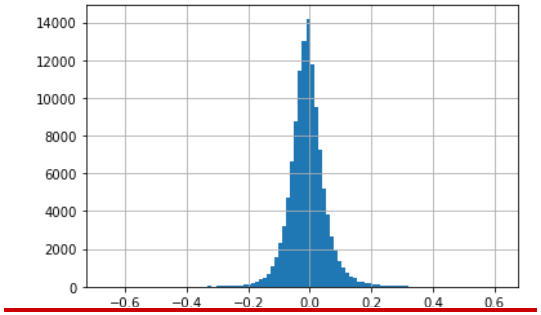
资料来源：东海证券

图 63 (MLP_2 未来 60 交易日回报分布)



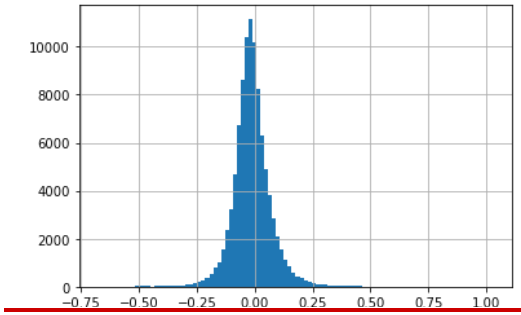
资料来源：东海证券

图 64 (MLP_3 未来 5 交易日回报分布)



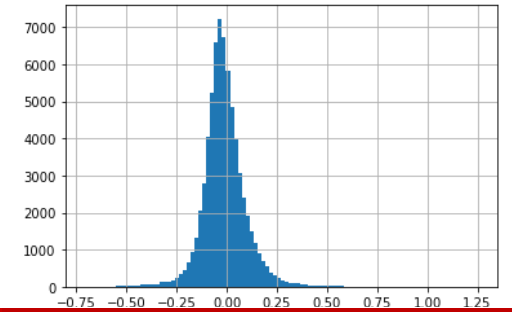
资料来源：东海证券

图 65 (MLP_3 未来 10 交易日回报分布)



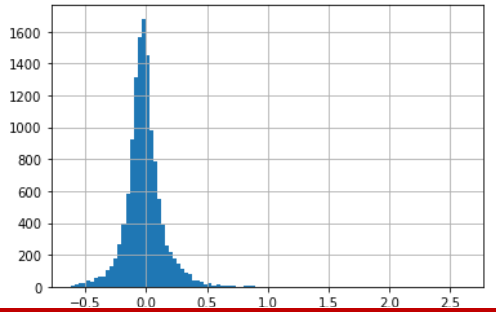
资料来源：东海证券

图 66 (MLP_3 未来 20 交易日回报分布)



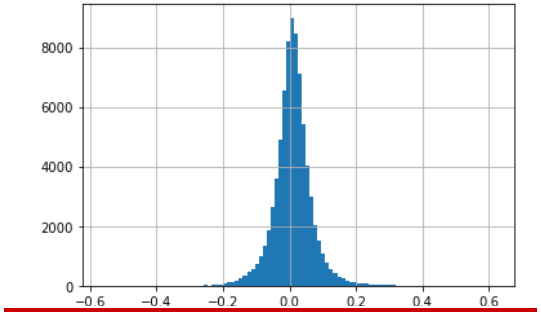
资料来源：东海证券

图 67 (MLP_3 未来 60 交易日回报分布)



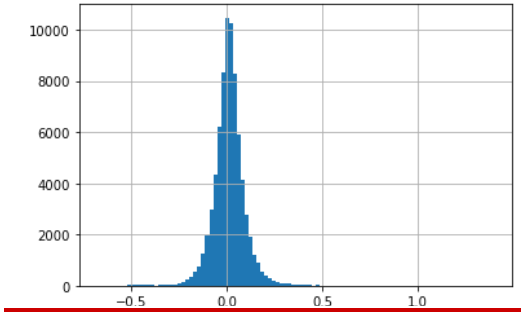
资料来源：东海证券

图 68 (MLP_4 未来 5 交易日回报分布)



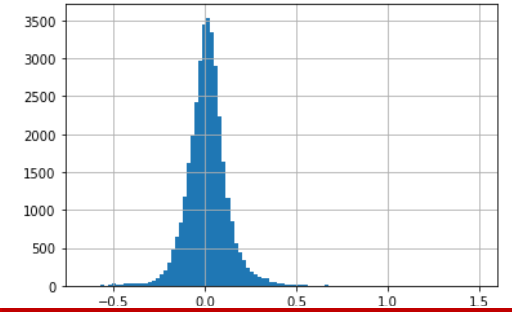
资料来源：东海证券

图 69 (MLP_4 未来 10 交易日回报分布)



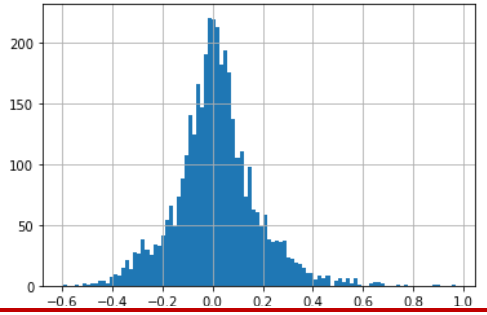
资料来源：东海证券

图 70 (MLP_4 未来 20 交易日回报分布)



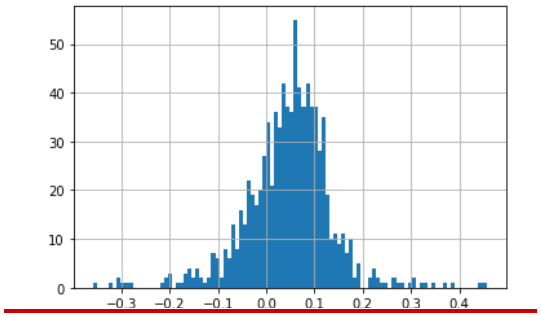
资料来源：东海证券

图 71 (MLP_4 未来 60 交易日回报分布)



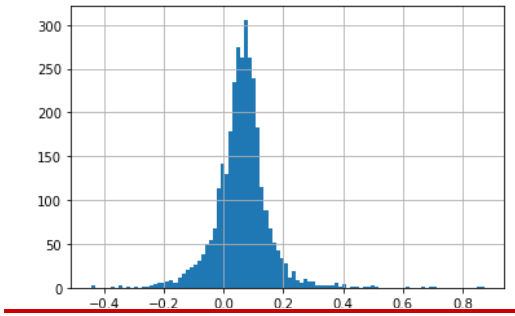
资料来源：东海证券

图 72 (MLP_5 未来 5 交易日回报分布)



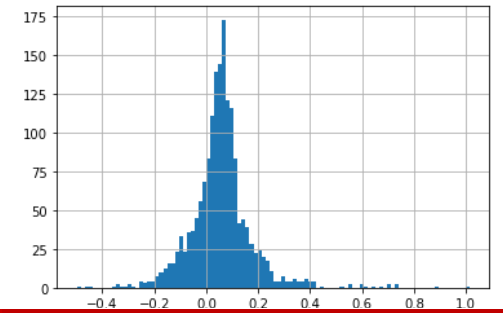
资料来源：东海证券

图 73 (MLP_5 未来 10 交易日回报分布)



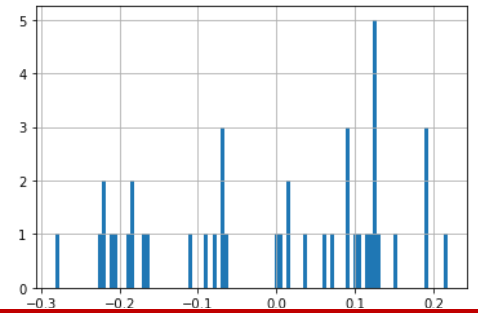
资料来源：东海证券

图 74 (MLP_5 未来 20 交易日回报分布)



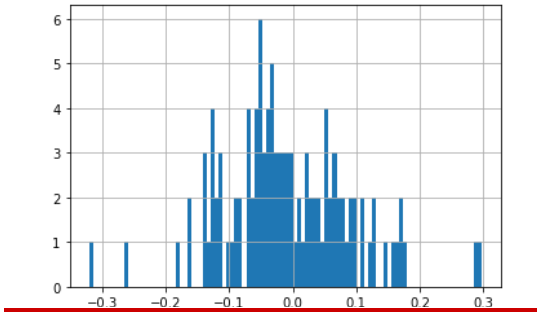
资料来源：东海证券

图 75 (MLP_5 未来 60 交易日回报分布)



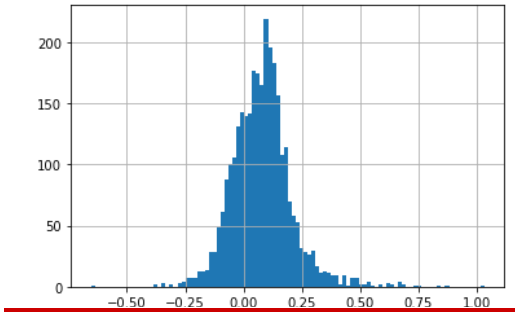
资料来源：东海证券

图 76 (MLP_6 未来 5 交易日回报分布)



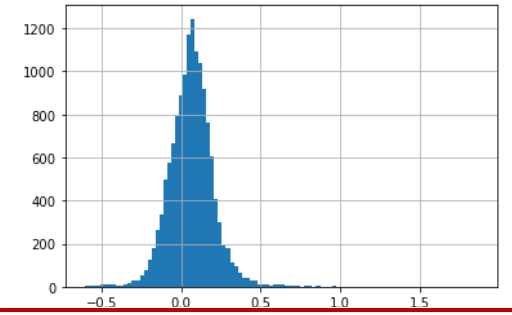
资料来源：东海证券

图 77 (MLP_6 未来 10 交易日回报分布)



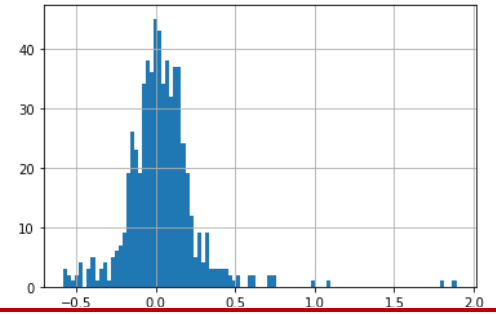
资料来源：东海证券

图 78 (MLP_6 未来 20 交易日回报分布)



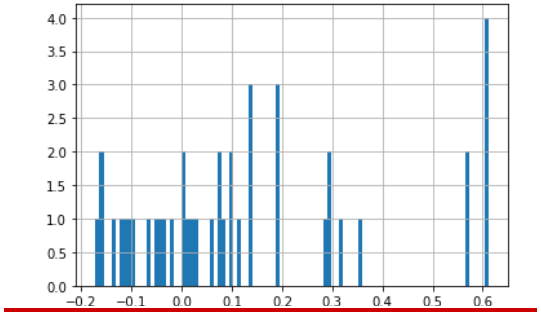
资料来源：东海证券

图 79 (MLP_6 未来 60 交易日回报分布)



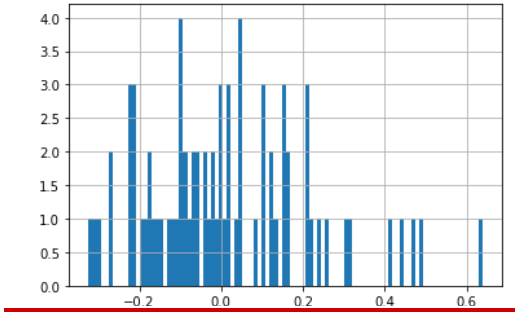
资料来源：东海证券

图 80 (MLP_7 未来 5 交易日回报分布)



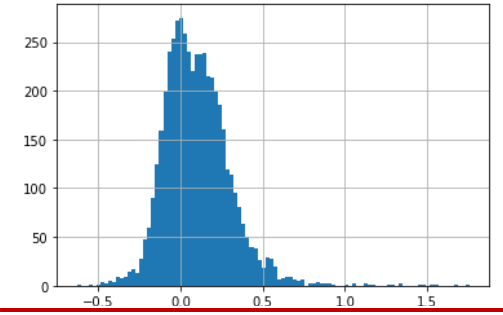
资料来源：东海证券

图 81 (MLP_7 未来 10 交易日回报分布)



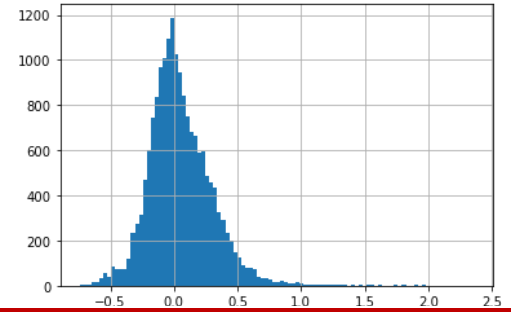
资料来源：东海证券

图 82 (MLP_7 未来 20 交易日回报分布)



资料来源：东海证券

图 83 (MLP_7 未来 60 交易日回报分布)

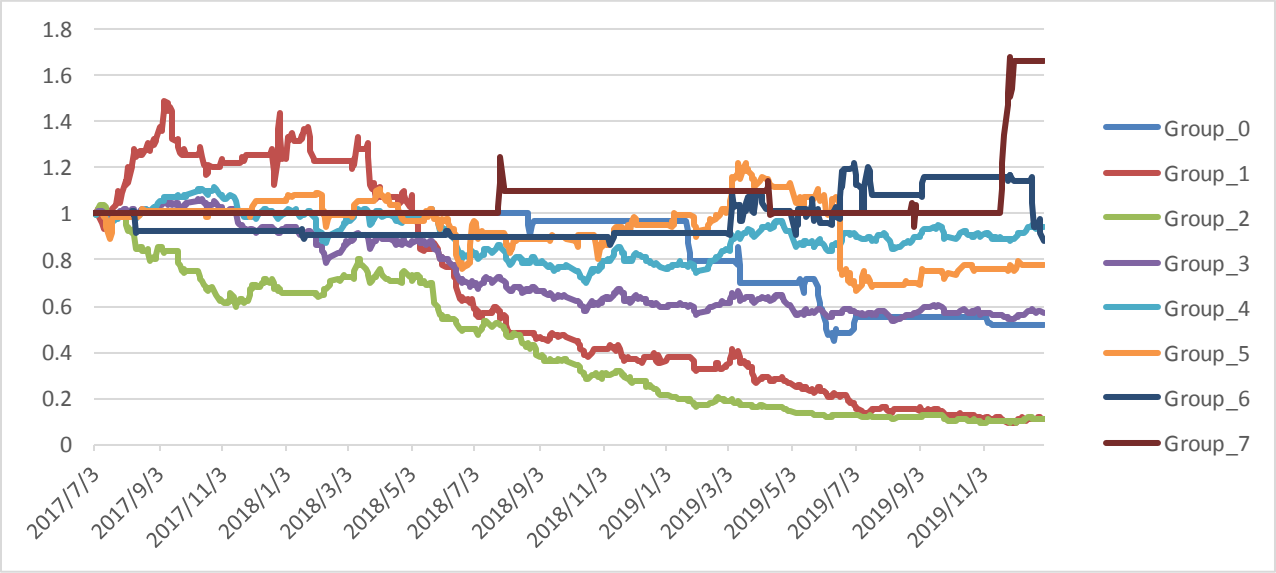


资料来源：东海证券

回归测试分析

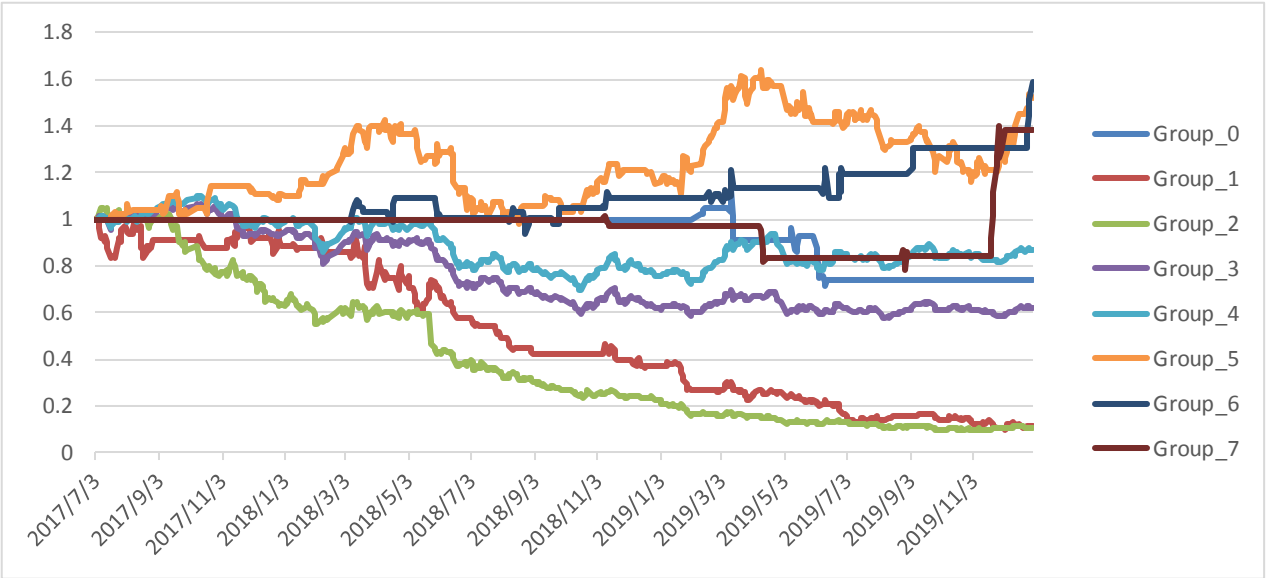
本文对模型训练结果的最后一项评估是用预测准确率最高的随机森林和多层感知器模型的预测结果作为选股依据，构建投资组合并进行回归测试。具体测试方法是，对涨跌幅预测周期为 N 个交易日（$N = 5,10,20,60$）的，以 N 个交易日为间隔建仓，并持有 N 个交易日再换仓。在每个建仓日，将全市场个股按预测级别分为 0 到 7 共计 8 组，每组个股按等权重形成投资组合。为充分检验预测模型为选股策略带来的效果，回测过程中不进行止损或其他操作。回测起止时间为 2017 年 7 月至 2019 年 12 月，换仓时的交易成本设为 0.05%。以下图 84 至图 91 展示了利用随机森林和多层感知器在各预测任务中的预测结果，所形成投资组合的收益净值曲线；表 2 至表 5 统计不同预测周期下各投资组合的表现。

图 84 (MLP 预测未来 5 个交易日的各分组组合走势)



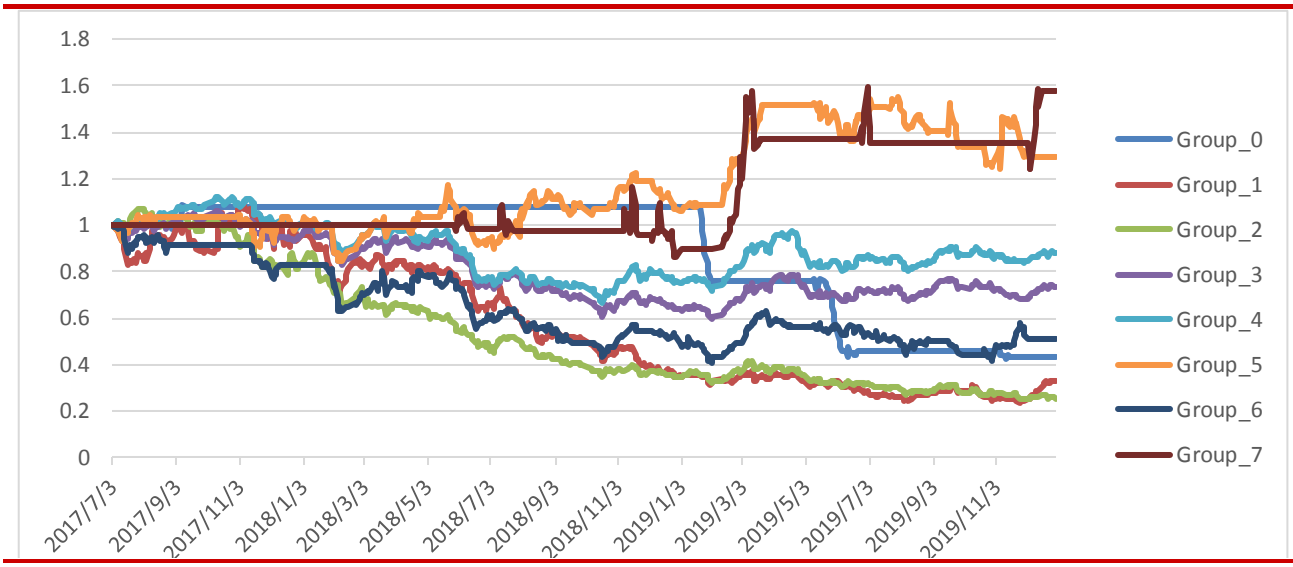
资料来源：东海证券

图 85 (随机森林预测未来 5 个交易日的各分组组合走势)



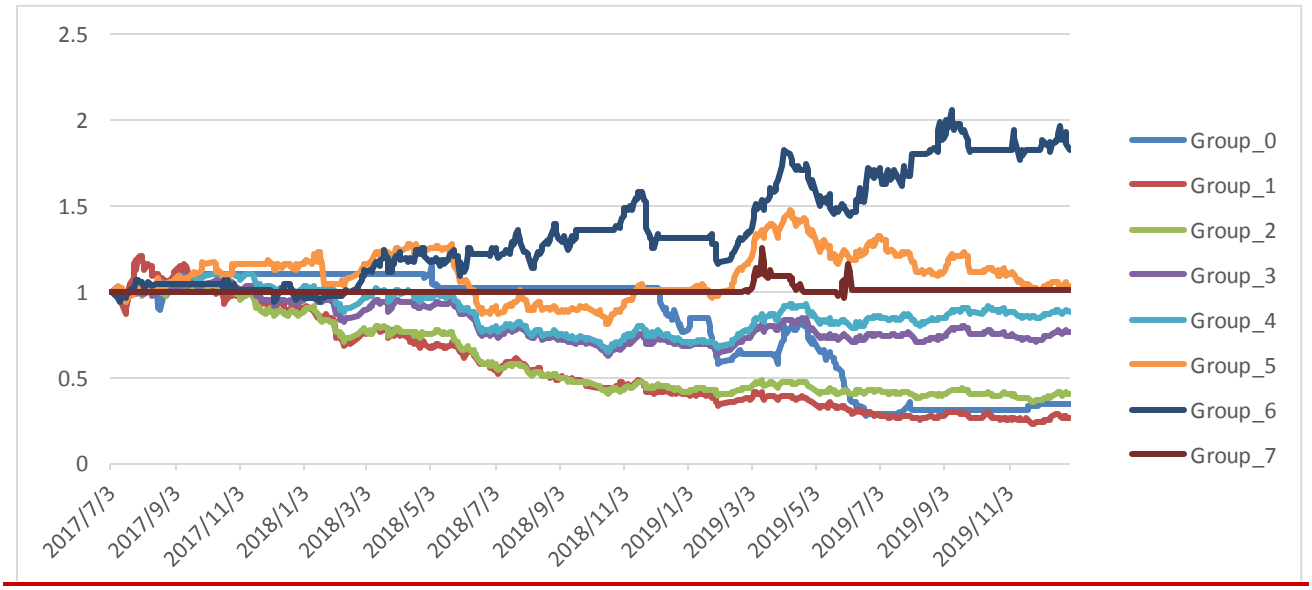
资料来源：东海证券

图 86 (MLP 预测未来 10 个交易日的各分组组合走势)



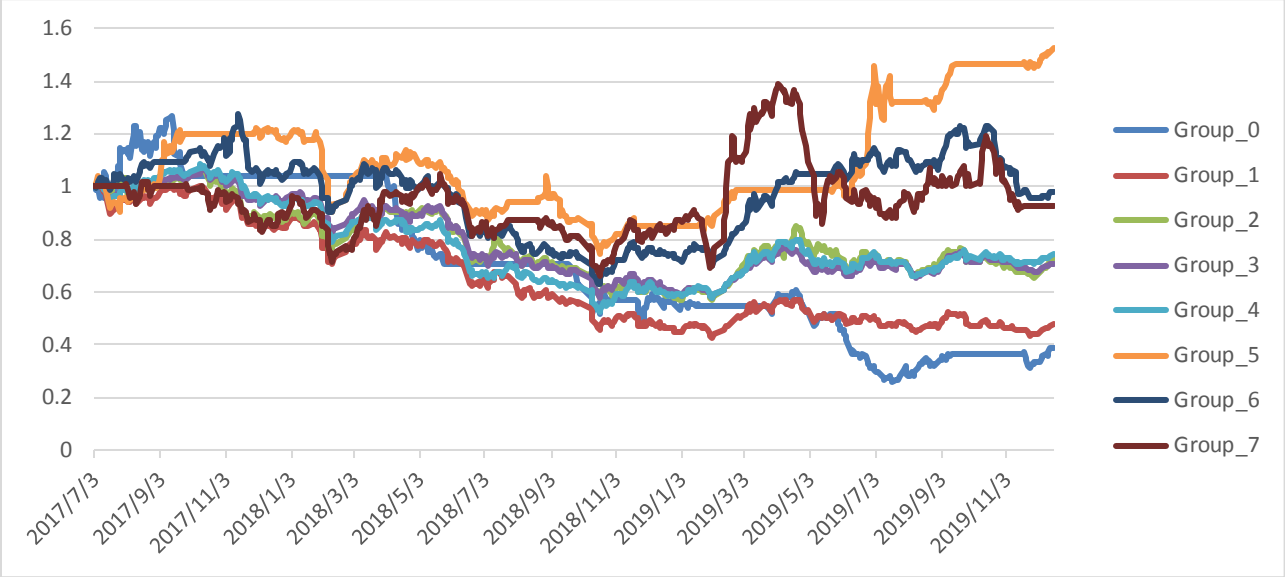
资料来源：东海证券

图 87 (随机森林预测未来 10 个交易日的各分组组合走势)



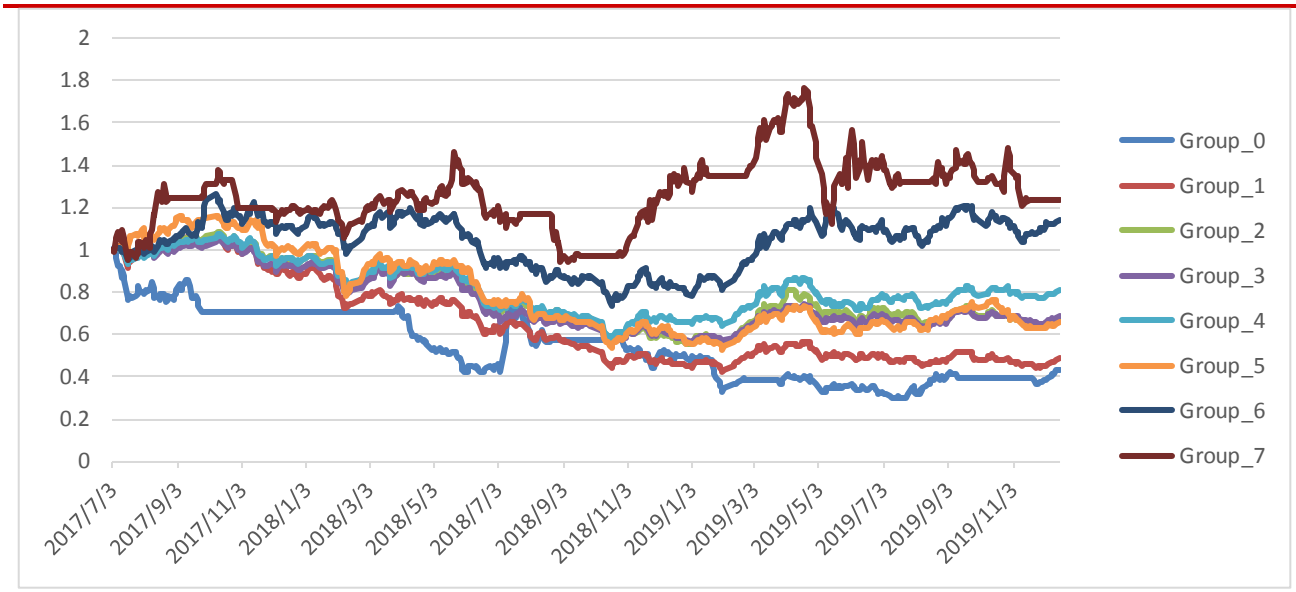
资料来源：东海证券

图 88 (MLP 预测未来 20 个交易日的各分组组合走势)



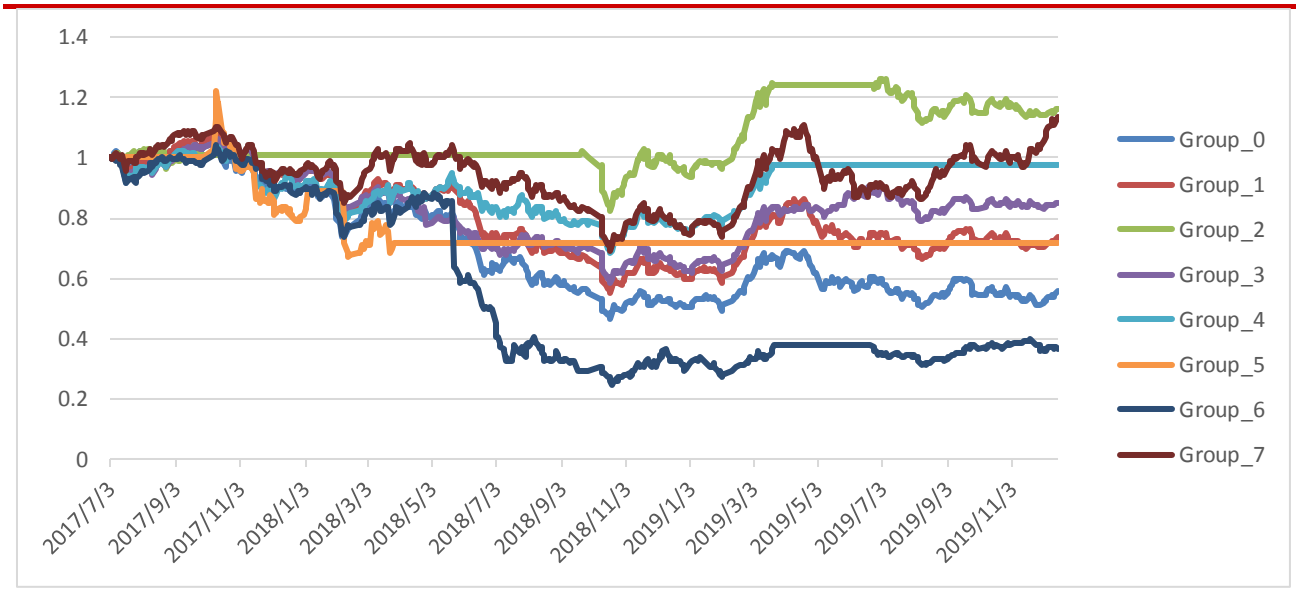
资料来源：东海证券

图 89 (随机森林预测未来 20 个交易日的各分组组合走势)



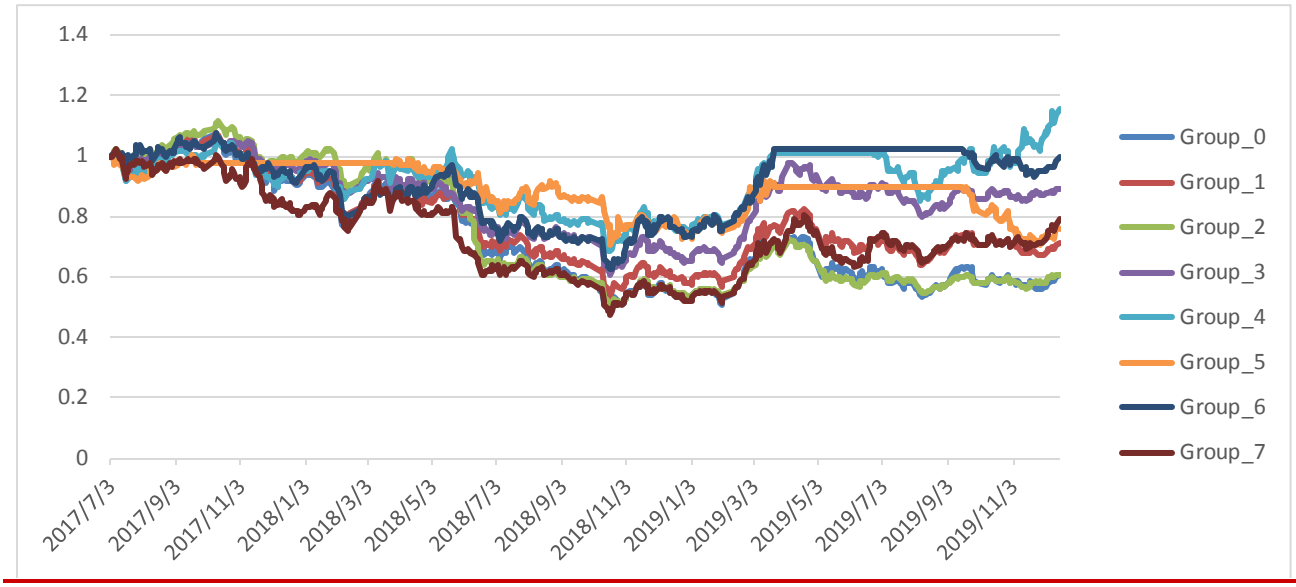
资料来源：东海证券

图 90 (MLP 预测未来 60 个交易日的各分组组合走势)



资料来源：东海证券

图 91 (随机森林预测未来 60 个交易日的各分组组合走势)



资料来源：东海证券

表 2 预测未来 5 个交易日的各分组表现

	分组	收益率	日平均收益率	年化收益率	最大回撤	夏普比率
MLP	0	-47.93%	-0.11%	-23.63%	-54.82%	-2.4287
	1	-89.11%	-0.36%	-59.98%	-94.04%	-1.9826
	2	-89.01%	-0.36%	-59.83%	-91.14%	-3.1678
	3	-42.68%	-0.09%	-20.54%	-49.65%	-2.4805
	4	-5.70%	-0.01%	-2.40%	-36.96%	-0.6140
	5	-21.85%	-0.04%	-9.68%	-45.43%	-1.7033
	6	-12.01%	-0.02%	-5.15%	-27.68%	-1.3092
	7	65.91%	0.08%	23.26%	-24.49%	5.0144
随机森林	0	-26.27%	-0.05%	-11.83%	-36.99%	-2.3670
	1	-89.00%	-0.36%	-59.82%	-90.23%	-2.9582
	2	-89.91%	-0.38%	-61.23%	-91.05%	-3.0143
	3	-38.06%	-0.08%	-17.95%	-46.02%	-2.3776
	4	-13.31%	-0.02%	-5.73%	-36.75%	-1.3675
	5	52.23%	0.07%	18.96%	-31.34%	3.1806
	6	58.84%	0.08%	21.07%	-14.06%	5.5508
	7	38.38%	0.05%	14.36%	-22.67%	3.5539

资料来源：东海证券

表 3 预测未来 10 个交易日的各分组表现

	分组	收益率	日平均收益率	年化收益率	最大回撤	夏普比率
MLP	0	-56.79%	-0.14%	-29.29%	-60.86%	-2.1663
	1	-66.91%	-0.18%	-36.68%	-77.97%	-2.4350
	2	-74.88%	-0.23%	-43.49%	-76.58%	-2.9118
	3	-26.77%	-0.05%	-12.08%	-43.77%	-2.0548
	4	-12.00%	-0.02%	-5.15%	-40.77%	-1.0996
	5	29.20%	0.04%	11.16%	-23.20%	1.4715
	6	-48.74%	-0.11%	-24.12%	-58.96%	-2.9935
	7	57.90%	0.07%	20.77%	-25.76%	3.0150
随机森林	0	-64.89%	-0.17%	-35.11%	-75.80%	-2.0962
	1	-73.82%	-0.22%	-42.52%	-80.71%	-2.6371
	2	-59.76%	-0.15%	-31.35%	-65.65%	-2.6291
	3	-23.45%	-0.04%	-10.45%	-41.58%	-1.9909
	4	-11.35%	-0.02%	-4.86%	-41.78%	-0.9678
	5	3.77%	0.01%	1.54%	-37.00%	0.2791
	6	82.46%	0.10%	28.20%	-27.07%	2.7069
	7	1.39%	0.00%	0.57%	-23.25%	0.5493

资料来源：东海证券

表 4 预测未来 20 个交易日的各分组表现

	分组	收益率	日平均收益率	年化收益率	最大回撤	夏普比率
MLP	0	-60.93%	-0.16%	-32.61%	-79.32%	-2.1748
	1	-51.76%	-0.12%	-26.37%	-57.90%	-2.7760
	2	-26.78%	-0.05%	-12.27%	-46.78%	-2.0540
	3	-28.99%	-0.06%	-13.39%	-45.99%	-2.0760
	4	-25.89%	-0.05%	-11.83%	-51.56%	-1.7733
	5	52.29%	0.07%	19.32%	-38.91%	2.6714
	6	-2.36%	0.00%	-1.00%	-50.81%	-0.1587
	7	-7.25%	-0.01%	-3.11%	-38.20%	-0.5601
随机森林	0	-56.72%	-0.14%	-29.66%	-70.70%	-3.5049
	1	-51.30%	-0.12%	-26.08%	-60.82%	-2.5335
	2	-33.02%	-0.07%	-15.49%	-49.98%	-2.2115
	3	-31.47%	-0.06%	-14.68%	-48.08%	-2.1890
	4	-19.38%	-0.04%	-8.65%	-47.29%	-1.5818
	5	-33.74%	-0.07%	-15.88%	-54.71%	-1.8089
	6	13.72%	0.02%	5.55%	-41.70%	1.1400
	7	23.12%	0.03%	9.13%	-36.31%	1.4681

资料来源：东海证券

表 5 预测未来 60 个交易日的各分组表现

	分组	收益率	日平均收益率	年化收益率	最大回撤	夏普比率
MLP	0	-44.04%	-0.10%	-21.64%	-55.60%	-2.4830
	1	-26.08%	-0.05%	-11.92%	-48.69%	-1.9292
	2	16.39%	0.03%	6.58%	-20.81%	1.6404
	3	-14.64%	-0.03%	-6.43%	-44.75%	-1.2651
	4	-2.23%	0.00%	-0.94%	-34.30%	-0.2714
	5	-28.16%	-0.06%	-12.97%	-45.07%	-2.5686
	6	-63.11%	-0.17%	-34.22%	-76.11%	-2.2979
	7	13.86%	0.02%	5.60%	-37.45%	1.4798
随机森林	0	-39.28%	-0.08%	-18.91%	-54.63%	-2.2809
	1	-28.73%	-0.06%	-13.26%	-49.99%	-2.0369
	2	-39.22%	-0.08%	-18.87%	-55.51%	-1.9869
	3	-10.61%	-0.02%	-4.60%	-42.46%	-0.9449
	4	15.47%	0.02%	6.23%	-35.60%	1.6592
	5	-24.11%	-0.05%	-10.94%	-29.42%	-3.0283
	6	-0.18%	0.00%	-0.08%	-42.09%	-0.0157
	7	-20.87%	-0.04%	-9.36%	-53.32%	-1.4998

资料来源：东海证券

就以上各个预测周期的回测结果而言，随机森林和多层感知器在 5 日和 10 日的预测分组形成的投资组合都表现出了较好的区分效果。其中多层感知器在第 7 个分组的年化收益率最高，随机森林在第 6 和第 5 分组的年化收益率最高。随机森林在第 7 个分组的年化收益率水平低于第 6 和第 5 分组，这主要是因为随机森林预测级别为 7 的个股过少，在许多换仓周期为 0 持仓导致的，收益曲线呈现出平直形态。但总体而言，随机森林在 5 日和

10 日的预测分组形成的投资组合的表现仍然好于多层感知器。对于 20 日和 60 日的回测结果，随机森林和多层感知器的表现均下降，这与预测准确率随预测周期延长而下降是相关的。在预测周期为 20 日时，随机森林仍保持了较好的区分度，但各分组的收益表现均下降。在预测周期为 60 日时，随机森林和多层感知器均未实现对个股表现的区分。

5.结论

通过上述对模型训练结果的评估可以看出，机器学习可以在适当的条件下将技术指标所包含的信息转换为对个股未来表现的有效预测，并实现较好的投资收益。这为发展历史悠久的技术分析方法提供了新的思路。但也应看到，模型的预测准确率，特别是对未来涨跌幅的预测准确率仍然不尽人意，随着预测周期的延长，预测准确率和投资组合的收益水平均下降。因此，如何提高预测准确率应是我们未来重点关注的研究问题。正如前面提到的，“数据决定了模型可能达到的最佳效果，而模型只是尽可能逼近这个效果”。在我们纳入模型训练的技术因子中，能够对标的未来表现进行很好区分的因子并不多。特别是可以区分趋势强度的因子较多，但可以识别趋势方向的因子较少。因此，我们认为寻找区分能力更强的技术指标，以及能够综合描述不同技术指标相互关系的复合型指标，或者能够引入更多信息来源的指标，是提高模型训练准确度的重要方向。

单就本次模型训练结果的投资应用而言，我们认为可以将效果较好的，预测周期为未来 5 日、10 日的随机森林和多层感知网络，作为投资组合选择的初步筛选器，再结合行业、板块、市值等信息进一步完成投资组合的构建。对于完成构建的投资组合，应结合相应的择时策略、止损规则、资金管理策略进行操作，或者结合股指期货、期权等衍生品，设计满足特定投资需求的风险收益结构。

分析师简介：

丁竞渊，数学与应用数学学士，计算机应用技术博士，东海证券研究所高级研究员，证券从业13年，具有丰富的金融工程研究、金融数据分析、量化交易平台研发经验。

附注：

一、市场指数评级

看多——未来6个月内上证综指上升幅度达到或超过20%

看平——未来6个月内上证综指波动幅度在-20%—20%之间

看空——未来6个月内上证综指下跌幅度达到或超过20%

二、行业指数评级

超配——未来6个月内行业指数相对强于上证指数达到或超过10%

标配——未来6个月内行业指数相对上证指数在-10%—10%之间

低配——未来6个月内行业指数相对弱于上证指数达到或超过10%

三、公司股票评级

买入——未来6个月内股价相对强于上证指数达到或超过15%

增持——未来6个月内股价相对强于上证指数在5%—15%之间

中性——未来6个月内股价相对上证指数在-5%—5%之间

减持——未来6个月内股价相对弱于上证指数5%—15%之间

卖出——未来6个月内股价相对弱于上证指数达到或超过15%

四、风险提示

本报告所载的全部内容只提供给客户做参考之用，并不构成对客户的投资建议，并非作为买卖、认购证券或其它金融工具的邀请或保证，建议客户如有任何疑问应当咨询独立财务顾问并独自进行投资判断。

五、免责条款

本报告基于本公司研究所及研究人员认为可信的公开资料或实地调研的资料，但对这些信息的真实性、准确性和完整性不做任何保证。本报告仅反映研究员个人出具本报告当时的分析和判断，并不代表东海证券股份有限公司，或任何其附属或联营公司的立场，本公司可能发表其他与本报告所载资料不一致及有不同结论的报告。本报告可能因时间等因素的变化而变化从而导致与事实不完全一致，敬请关注本公司就同一主题所出具的相关后续研究报告及评论文章。在法律允许的情况下，本公司的关联机构可能会持有报告中涉及的公司所发行的证券并进行交易，并可能为这些公司正在提供或争取提供多种金融服务，本公司的关联机构或个人可能在本报告公开发布之间已经了解或使用其中的信息。

分析师承诺“本人及直系亲属与本报告中涉及的内容不存在利益关系”。本报告仅供“东海证券股份有限公司”客户、员工及经本公司许可的机构与个人阅读。

本报告版权归“东海证券股份有限公司”所有，未经本公司书面授权，任何人不得对本报告进行任何形式的翻版、复制、刊登、发表或者引用。

六、资格说明

东海证券股份有限公司是经中国证监会核准的合法证券经营机构，已经具备证券投资咨询业务资格。我们欢迎社会监督并提醒广大投资者，参与证券相关活动应当审慎选择具有相当资质的证券经营机构，注意防范非法证券活动。

上海 东海证券研究所

地址：上海市浦东新区东方路1928号 东海证券大厦
网址：Http://www.longone.com.cn
电话：（8621）20333619
传真：（8621）50585608
邮编：200215

北京 东海证券研究所

地址：北京市西三环北路87号国际财经中心D座15F
网址：Http://www.longone.com.cn
电话：（8610）66216231
传真：（8610）59707100
邮编：100089