机器学习模型在因子选股上的比较分析

人工智能研究报告

报告摘要:

- **研究内容**

本报告采用机器学习方法从历史数据中学习股票因子和收益率的关系，建立股票收益预测模型。本报告研究的机器学习方法包括多类别逻辑回归（MLR）、支持向量机（SVM）、随机森林（RF）、极限梯度提升树（XGBoost）、深层神经网络（DNN）等 5 类模型。

- **机器学习模型介绍**

本报告考察的 5 种机器学习模型中，MLR 和线性 SVM 属于线性分类器，但优化目标不同。RF、XGBoost 和 DNN 属于非线性分类器。其中，RF 和 XGBoost 是以决策树为基学习器的集成学习方法，但模型集成的方式不一样。DNN 是深度学习方法。这 5 种模型在机器学习领域具有很强的代表性。

- **策略表现**

从实证结果来看，5 种机器学习模型都取得了显著的超额收益，而且收益曲线相似。由于机器学习模型都是从历史数据建立起股票因子和收益率的关系，不同模型的表现有较大的相关性，模型打分相关性和模型 IC 相关性都比较高。其中，同为线性分类器的 MLR 和线性 SVM 模型的相关性最高。

日频样本训练模式平均每次用 48 万个样本训练模型。在 5 种不同的模型中，DNN 模型表现最佳，具有最高的 IC、ICIR、年化对冲收益和夏普比率。但是 DNN 模型的训练耗时，平均每个模型训练需要 5 个多小时。

半月频样本训练模式平均每次用 4.8 万个样本训练模型。在 5 种不同的模型中，XGBoost 模型表现最好。而且 XGBoost 的训练时间短，和线性分类模型的训练时间差别不大。

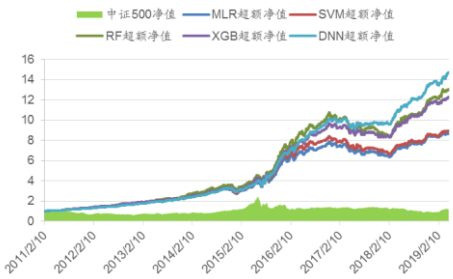
总体来看，日频样本模式训练的模型表现优于半月频样本模式训练的模型。尤其是 DNN 和 XGBoost 模型，日频样本模式明显优于半月频样本模式。因为这两种模型训练时更依赖于大量的训练样本。

从机器学习模型打分来看模型的风格暴露情况，DNN 和 XGBoost 在风格因子上的暴露相对较少，而 RF 在风格因子上的暴露最大。

- **风险提示**

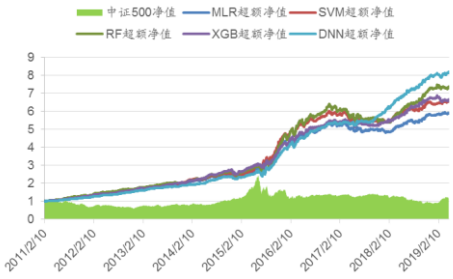
策略模型并非百分百有效，市场结构及交易行为的改变以及类似交易参与者的增多有可能使得策略失效。

图：机器学习选股表现（等权）



数据来源：Wind，广发证券发展研究中心

图：机器学习选股表现（行业中性）



数据来源：Wind，广发证券发展研究中心

分析师： 文巧钧
SAC 执证号：S0260517070001
SFC CE No. BNI358
0755-88286935
wenqiaojun@gf.com.cn

分析师： 安宁宁
SAC 执证号：S0260512020003
SFC CE No. BNW179
0755-23948352
anningning@gf.com.cn

分析师： 罗军
SAC 执证号：S0260511010004
020-66335128
luojun@gf.com.cn

请注意，罗军并非香港证券及期货事务监察委员会的注册持牌人，不可在香港从事受监管活动。

相关研究：

深度学习在指数增强策略上的应用:深度学习研究报告之六	2019-04-03

目录索引

一、问题背景 5
二、机器学习模型介绍 5
2.1 机器学习因子选股框架 5
2.2 多类别逻辑回归 7
2.3 支持向量机 8
2.4 随机森林 10
2.5 极限梯度提升树 12
2.6 深层神经网络 13
2.7 不同机器学习模型的特点分析 14
三、选股策略描述 15
3.1 模型训练方法 15
3.2 机器学习训练平台和模型超参数 16
3.3 策略回测设置 17
四、实证分析 18
4.1 机器学习模型预测性能比较 18
4.2 机器学习模型打分相关性分析 20
4.3 机器学习模型选股表现 23
4.4 机器学习模型的风格分析 28
五、总结与展望 30

图表索引

图 1：机器学习选股框架.....6
图 2：MLR 示意图.....7
图 3：SVM 二分类示意图.....8
图 4：软间隔 SVM 二分类示意图.....9
图 5：SVM 用于多分类问题示意图.....10
图 6：集成学习示意图.....11
图 7：Bagging 和 Boosting 模型训练示意图.....11
图 8：RF 示意图.....12
图 9：XGBoost 模型求解示意图.....13
图 10：DNN 示意图.....14
图 11：日频样本和半月频样本采样示意图.....15
图 12：模型滚动更新示意图.....16
图 13：时间分组交叉验证示意图.....17
图 14：日频样本模型与半月频样本模型测试集准确率对比.....19
图 15：日频样本模型与半月频样本模型训练时间对比.....20
图 16：日频样本模型 IC 序列.....21
图 17：日频机器学习模型选股表现（等权）.....24
图 18：日频样本机器学习模型选股对冲收益（等权）.....25
图 19：日频机器学习模型选股表现（行业中性）.....26
图 20：日频样本机器学习模型选股对冲收益（行业中性）.....27
图 21：不同机器学习模型选股打分与风格因子相关性雷达图（日频样本）.....29
图 22：不同机器学习模型选股打分与风格因子相关性雷达图（半月频样本）.....30

表 1：日频样本模型测试集预测准确率.....18
表 2：半月频样本模型测试集预测准确率.....19
表 3：机器学习模型 IC.....20
表 4：日频样本机器学习模型打分相关性.....21
表 5：半月频样本机器学习模型打分相关性.....22
表 6：日频样本机器学习模型 IC 相关性.....22
表 7：半月频样本机器学习模型 IC 相关性.....22
表 8：不同机器学习模型等权选股策略对冲表现（日频样本）.....23
表 9：不同机器学习模型等权选股策略分年度对冲收益（日频样本）.....25
表 10：不同机器学习模型行业中性选股策略对冲表现（日频样本）.....25
表 11：不同机器学习模型行业中性选股策略分年度对冲收益（日频样本）.....27
表 12：机器学习模型选股性能比较（等权组合）.....28
表 13：机器学习模型选股性能比较（行业中性组合）.....28
表 14：风格因子列表.....28
表 15：不同机器学习模型选股打分与风格因子相关性（日频样本）.....29

表 16：不同机器学习模型选股打分与风格因子相关性（半月频样本）.....30

一、问题背景

近年来，随着机器学习在计算机视觉、自然语言处理、专家系统等领域的巨大成功，海内外越来越多的量化基金在研究将机器学习技术引入投资策略中，并且已经出现了众多成功案例。从2016年以来，海外知名投行和对冲基金从人工智能领域引入专业人才和成立人工智能研究小组的报道时有发生，也产生了AI Powered Equity ETF（AIEQ）等公开产品。国内市场来看，以私募基金为主的众多量化投资机构在投资策略中引入了机器学习技术。自2017年以来，A股市场的量化选股基金整体表现不佳，这和市场的“一九行情”、股指期货大幅贴水、市场风格切换加快等因素有关。在此期间，有一批量化私募异军突起，在量化选股策略整体表现不佳的时期取得了优异的业绩。与这些机构的交流中，我们发现机器学习技术在其中起到了重要的作用。

广发金融工程团队在此前的一系列研究中，实证了以深度学习为代表的机器学习方法在因子选股、市场短线择时等方面具有不错的表现和发展前景。详情可以参考《深度学习研究报告之六：深度学习在指数增强策略上的应用》、《机器学习多因子动态调仓策略——多因子Alpha系列报告之（三十六）》和《基于涨跌模式识别的指数和行业择时策略》等研究报告。

本报告以因子选股为背景，比较分析典型的机器学习方法，包括逻辑回归、支持向量机、随机森林、极限梯度提升树、深层神经网络等模型。主要内容有以下几点：

1）本报告把股票收益率的预测问题构建成一个分类问题，通过机器学习方法从历史数据中学习股票因子和收益率的关系，建立股票收益预测模型。

2）采用不同的机器学习分类模型对股票分类问题进行建模，比较模型的预测性能。

3）对不同机器学习模型表现的相关性进行分析。

4）研究不同机器学习模型选股的表现，以及机器学习因子和传统风格因子的相关性。

二、机器学习模型介绍

2.1 机器学习因子选股框架

因子选股是通过对股票的收益率进行预测，寻找能够产生超额收益的股票。机器学习通过对股票历史数据的学习，建立股票收益率的预测模型。

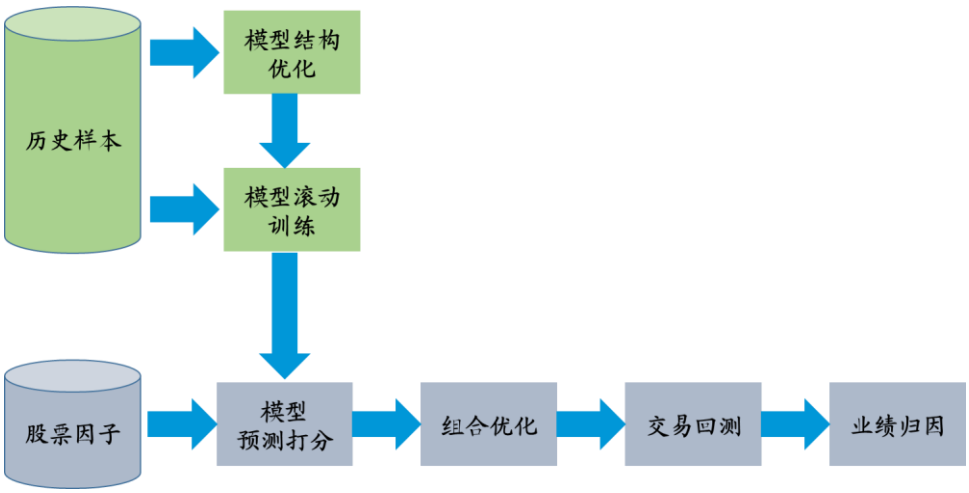
机器学习因子选股的框架如图 1 所示。首先是历史数据的处理：建立股票因子数据库并将股票因子数据标准化，构建包括输入变量和输出变量的股票样本。本报告中机器学习模型采用的特征一共有 156 个，包括估值因子、规模因子、反转因子、流动性因子、波动性因子、技术指标和行业属性。因子标准化流程包括异常值和缺失值处理、因子极值处理、时间方向的因子标准化、截面方向的因子标准化以及因子分布的调整。具体可以参考广发金融工程团队此前的报告《深度学习新进展：Alpha

因子再挖掘》。

选定机器学习模型之后，需要优化模型结构，也就是确定模型的“超参数”。模型超参数可以调节机器学习模型的拟合能力和泛化能力，减少模型的过拟合。其中，随机森林中需要事先确定的超参数包括决策树数量、单棵决策树最大深度、样本采样比例和特征采样比例等。深层神经网络中，需要事先确定的模型超参数包括网络层数、每个隐层的节点个数、激活函数种类等。支持向量机中需要事先确定的模型超参数包括核函数、核函数参数、惩罚系数等。本报告中，我们采用交叉验证方法确定模型的超参数。

确定模型的超参数之后，可以定期训练模型，在随后一段时间采用训练好的模型进行股票的收益率预测和选股交易，并对策略收益表现进行分析。

图 1：机器学习选股框架



数据来源：广发证券发展研究中心

在股票样本构建时，本报告提取每一天股票池内的全体股票，剔除涨停、跌停股票和ST股票之后，根据未来10个交易日后的股票涨跌幅给不同的股票样本贴“标签”：“上涨”、“下跌”和“平盘”。其中，标记为上涨的股票是指相对强势的股票，即未来10个交易日收益率前10%的股票；标记为下跌的股票是指相对弱势的股票，即未来10个交易日收益率后10%的股票；标记为平盘的股票是指未来10个交易日收益率处于中间的10%的股票（分位数45%至55%区间内）。

模型的目标是寻找能够产生超额收益的股票。因此，机器学习模型在训练时需要建立起股票因子和未来涨跌属性之间的关系，构建一个以股票因子为输入，股票收益率涨跌标签为输出的模型。股票因子标记为

$$\boldsymbol{x} = [x_1 \quad \cdots \quad x_m]^T$$

其中，m表示股票的因子个数，包括选股因子和行业0-1属性，在本报告中，m=156。股票涨跌属性为y，y=1（上涨），2（下跌），3（平盘），表示三种不同的股票涨跌类别。也可以将其转换为独热编码（one-hot code）

$$\boldsymbol{y} = [y^{(1)} \quad \cdots \quad y^{(K)}]^T$$

其中，K=3，通过$[1 \quad 0 \quad 0]^T$、$[0 \quad 1 \quad 0]^T$和$[0 \quad 0 \quad 1]^T$表示三种不同的类别。

本报告中采用的机器学习模型是分类模型，将股票样本$\boldsymbol{x}_i$分类为第k类（k=1，2，3）的预测概率为

$$\mathrm{p}(y_i = k|\boldsymbol{x}_i) = \mathrm{f}(\boldsymbol{x}_i; \boldsymbol{w})$$

其中，$\boldsymbol{w}$表示模型参数。$\mathrm{p}(y_i = k|\boldsymbol{x}_i)$越大，表示股票样本$\boldsymbol{x}_i$属于第$k$类的概率越大。

选股策略选择最有可能产生超额收益的股票构建多头组合，因此，机器学习的目标也就是识别出$\mathrm{p}(y_i = 1|\boldsymbol{x}_i)$最大的股票，本报告将$\mathrm{p}(y_i = 1|\boldsymbol{x}_i)$定义为机器学习模型打分或机器学习选股因子。

本报告选择了多类别逻辑回归、支持向量机、随机森林、极限梯度提升树和深层神经网络5种不同的分类模型来进行比较分析。

2.2 多类别逻辑回归

多类别逻辑回归（Multinomial Logistic Regression，MLR）又称 Softmax Regression，是逻辑回归算法的推广，用于处理类标签的数量大于或等于 2 的分类问题。与逻辑回归类似，MLR 是一种广义线性模型，用于处理线性可分的分类问题。

对于 K 分类问题，MLR 通过线性预测函数$h(k,i)$获取样本i属于类别k（k=1，2，……，K）的概率

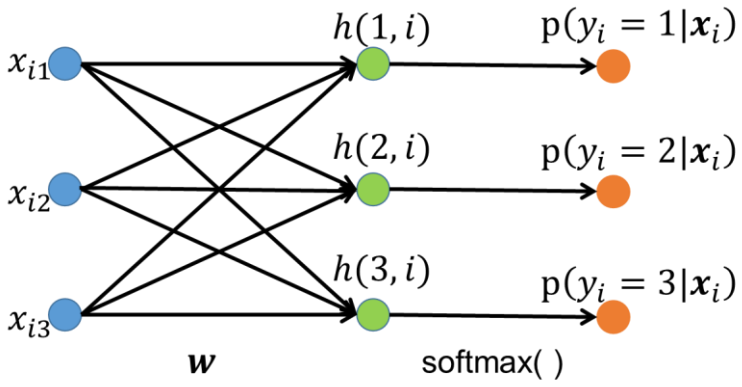
$$h(k,i) = w_{k0} + w_{k1}x_{i1} + w_{k2}x_{i2} + \cdots + w_{km}x_{im} = \boldsymbol{w}_k^T\boldsymbol{x}_i$$

最终的预测概率为

$$\mathrm{p}(y_i = k|\boldsymbol{x}_i) = \frac{e^{h(k,i)}}{\sum_k e^{h(k,i)}}$$

其中，$\mathrm{p}(y_i = k|\boldsymbol{x}_i)$关于$h(k,i)$的函数称为 Softmax 函数。通过预测概率$\mathrm{p}(k|\boldsymbol{x}_i)$，可以完成对样本的分类。

图 2：MLR示意图



数据来源：广发证券发展研究中心

一般通过极大似然法估计 MLR 模型的参数$\boldsymbol{w}$，对于训练集$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$，对数似然函数为

$$\ell(\boldsymbol{w}) = \sum_{i=1}^{N} \ln \mathrm{p}(y_i|\boldsymbol{x}_i, \boldsymbol{w})$$

可以采用梯度下降方法更新参数，获取最优的参数值。

2.3 支持向量机

支持向量机（Support Vector Machine，SVM）的雏形最早于 20 世纪 60 年代提出。在 90 年代，由于 SVM 在手写数字识别、文本分类任务等分类问题中展示出优良的性能，很快成为机器学习的主流技术，在各领域得到了广泛的应用。

在二分类 SVM 中，分类超平面可以用如下线性方程来描述

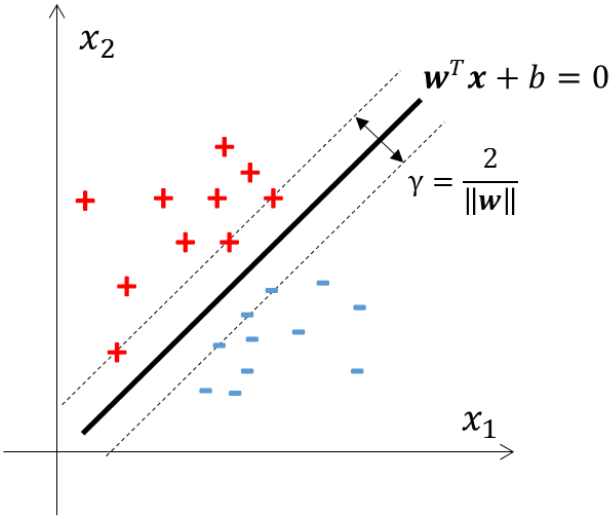
$$\boldsymbol{w}^T\boldsymbol{x} + b = 0$$

其中，$\boldsymbol{w} = (w_1, w_2, \ldots, w_m)$为法向量，决定了超平面的方向，$b$为偏置项。如下图所示，假设分类超平面可以将训练样本进行正确分类，其中，红色点为正样本，$y_i = 1$；蓝色点为负样本，$y_i = -1$。那么，可以寻找到超平面$\boldsymbol{w}^T\boldsymbol{x} + b = 0$，使得：

对于正类样本，有$\boldsymbol{w}^T\boldsymbol{x}_i + b \geq +1,\ y_i = +1$;

对于负类样本，有$\boldsymbol{w}^T\boldsymbol{x}_i + b \leq -1,\ y_i = -1$。

图 3：SVM二分类示意图



数据来源：广发证券发展研究中心

在图中虚线上的点可以使得上式的等号成立，被称为“支持向量”（support vector），不同类别支持向量到超平面的距离之和被称为“间隔”（margin），其值为

$$\gamma = \frac{2}{\|\boldsymbol{w}\|}$$

SVM 模型的目标是寻找具有最大间隔的分类超平面，也就是寻找合适的参数$\boldsymbol{w}$和b，使得间隔γ最大：

$$\max_{\boldsymbol{w},\mathrm{b}} \frac{2}{\|\boldsymbol{w}\|}$$

$$\text{s.t.}\ \ y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)\geq 1,\qquad i=1,2,\dots,N$$

当样本线性不可分时，可以通过核函数隐式地构建一个非线性映射，将样本从原始空间映射到高维特征空间，使得样本在特征空间线性可分。

对于在样本空间或者特征空间线性不可分的问题，可以引入“软间隔”（soft margin）的概念，允许支持向量机在一些样本上分类出错。如下图中圆圈标记的点所示，这些点不满足约束

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)\geq 1$$

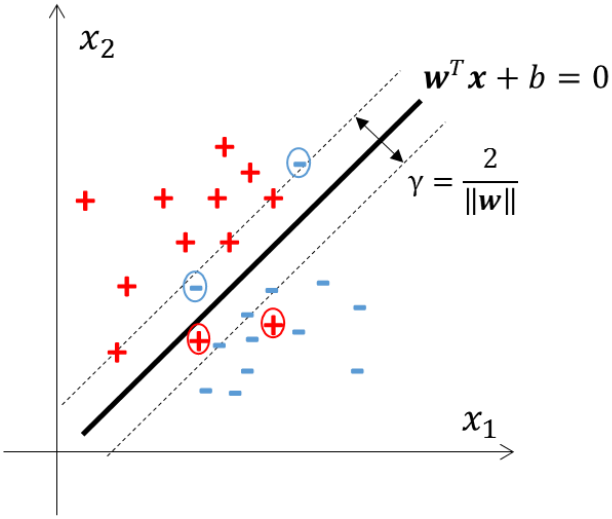
在软间隔 SVM 中，在最大化间隔的同时，要使得不满足约束的样本尽可能少，通过引入松弛变量ξ_i，优化问题可以写成

$$\min_{\boldsymbol{w},b,\xi_i}\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{N}\xi_i$$

$$\text{s.t.}\ \ y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)\geq 1-\xi_i$$

$$\xi_i\geq 0,\qquad i=1,2,\dots,N$$

图 4：软间隔SVM二分类示意图

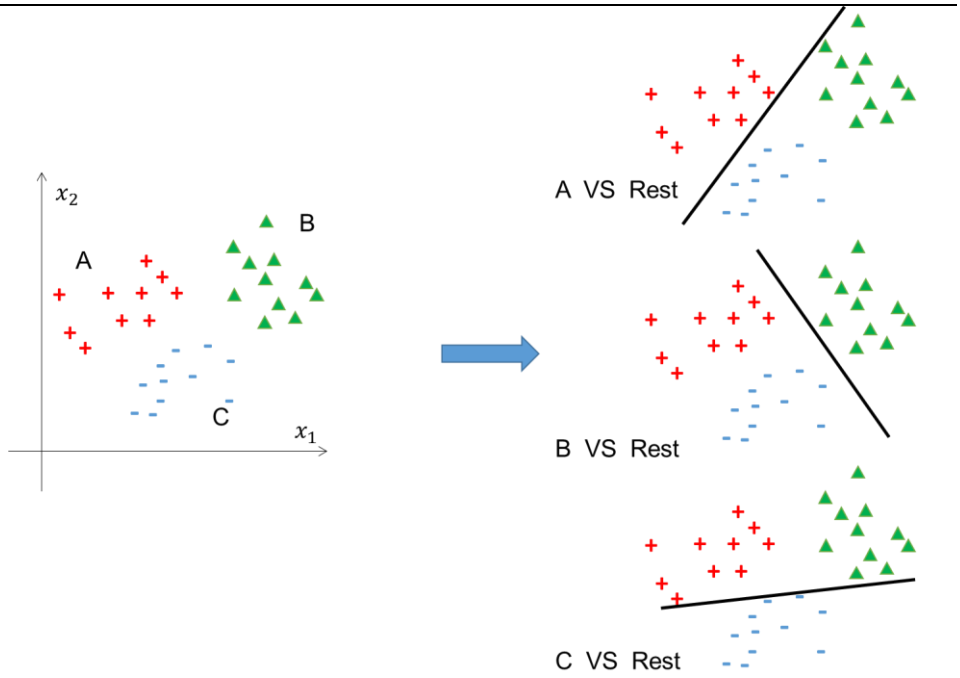


数据来源：广发证券发展研究中心

为了求解 SVM 优化问题，一般通过拉格朗日乘子法构建其对偶问题，通过求解对偶问题，获取 SVM 模型的参数。

SVM 模型一般用于对二分类问题进行分类，不直接处理多分类问题。对于多分类问题，有一对多和一对一两种处理办法。本报告中采用一对多方法处理三分类问题，也就是同时构建 3 个子分类问题，如下图所示。分别求解 3 个子分类问题，并将结果整合，获得最终的分类结果。

图 5：SVM用于多分类问题示意图



数据来源：广发证券发展研究中心

SVM 模型不能直接获得分类概率，本报告将分类时样本距分类超平面的有向距离作为分类打分，打分越高，则属于该类的“概率”越大。如果样本距K个分类超平面的距离依次为s_1，s_2，……，s_K，则“分类预测概率”

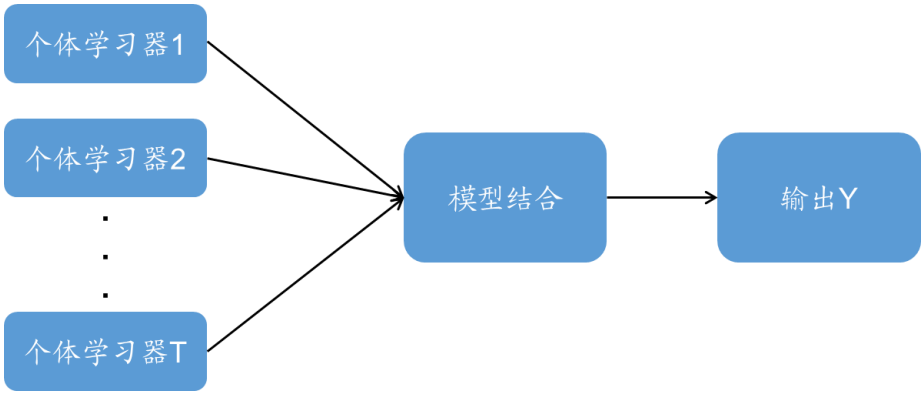
$$\mathrm{p}(y_i = k|\boldsymbol{x}_i) = \frac{e^{s_i}}{\sum_k e^{s_i}}$$

2.4 随机森林

集成学习是一大类机器学习方法，通过构建多个学习器并结合起来，完成监督学习任务。其一般思路是：先用现有的机器学习算法从训练数据中构建一组个体学习器（一般称为基学习器或者弱学习器），再用某种策略将个体学习器结合起来。如图6所示。

一般来说，集成学习通过将多个学习器进行结合，可以获得比单个学习器显著优越的泛化能力。在二分类问题中，如果通过简单的投票法进行集成，假设单个学习器的分类错误率为ϵ且不同学习器的分类错误率相互独立，则可以证明，T个学习器集成模型的分类错误率低于$\exp(-0.5T(1-2\epsilon)^2)$。随着集成学习中学习器数量的增大，集成学习的分类错误率将指数级下滑。

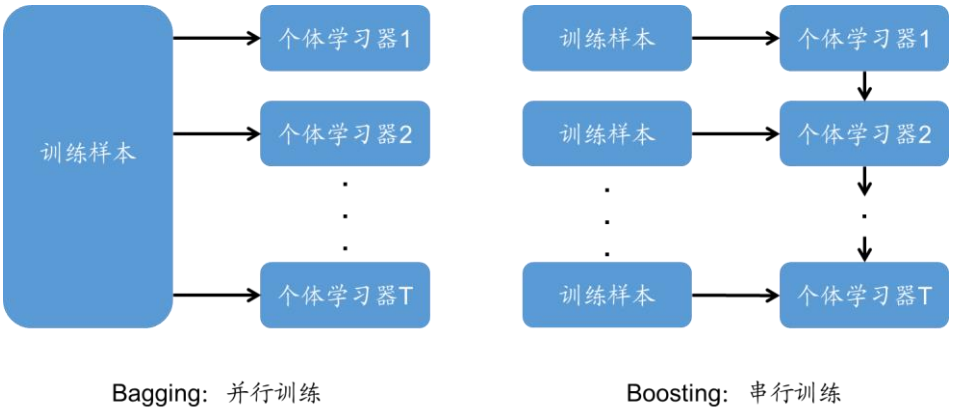
图 6：集成学习示意图



数据来源：广发证券发展研究中心

根据个体学习器生成方式的不同，集成学习方法可以分为两大类，Bagging和Boosting。其中，Bagging方法中个体学习器不存在强依赖关系，可以同时构建不同的个体学习器。而Boosting方法中个体学习器之间存在强依赖关系，按照先后次序一一生成不同的学习器。

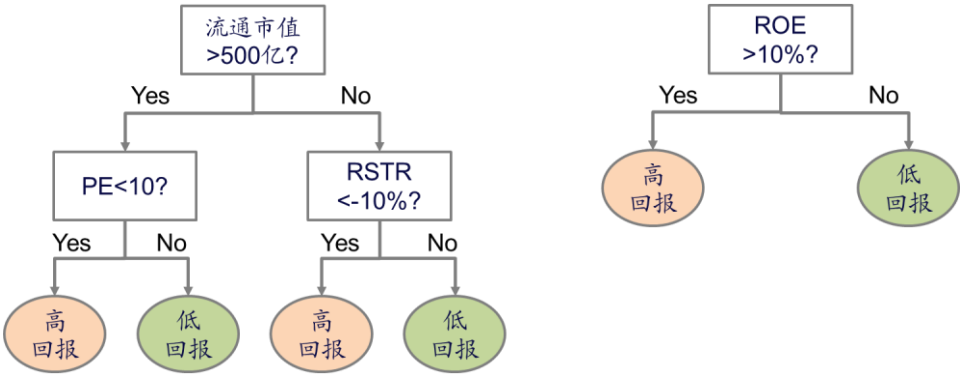
图 7：Bagging和Boosting模型训练示意图



数据来源：广发证券发展研究中心

Bagging模型的代表是随机森林（Random Forest，RF）。RF是一种以决策树为基学习器的集成学习方法，由Breiman在2001年提出。下图展示了一个简单的RF模型，由两棵用于预测股价回报的决策树构成。第一棵决策树根据流通市值、市盈率和股价的相对强弱预测股票未来是高回报或者低回报。第二棵决策树根据ROE预测股票是高回报或者低回报。RF模型将两个决策树模型的分类结果进行简单的平均（集成）。

图 8：RF示意图



数据来源：广发证券发展研究中心

如果某RF模型由T棵决策树构成，其中第t棵决策树预测样本$\boldsymbol{x}$属于类别c的概率为

$$p_t(y=c|\boldsymbol{x})$$

那么随机森林模型预测样本$\boldsymbol{x}$属于类别c的概率为

$$p(y=c|\boldsymbol{x})=\frac{1}{T}\sum_{t=1}^{T}p_t(y=c|\boldsymbol{x})$$

要获得好的集成效果，作为基学习器的决策树应该具有一定的准确度，并且具有多样性。随机森林从样本采样和特征采样两方面提高个体学习器的多样性。

样本采样又叫行采样，是指每次训练一棵新的决策树时，从训练样本总体中采样，获得训练单一决策树的样本。

特征采样又叫列采样，是指每次训练一棵新的决策树时，从全部特征中随机挑选部分特征。

通过样本采样和特征采样方法，每次用于训练一棵决策树时所采用的样本和特征与训练其他决策树都会有所差别，从而增加了决策树之间的多样性，可以使得随机森林模型具有较好的性能。

2.5 极限梯度提升树

极限梯度提升树（eXtreme Gradient Boosting，XGBoost或XGB）是近年来Boosting方法中最热门的一种算法，由华盛顿大学的陈天奇博士于发起，是梯度提升树的一种高效实现，曾经横扫Kaggle大赛。

XGBoost模型可以以决策树为基学习器，也可以以线性分类器为基学习器，可表示为基学习器的加法模型。本报告中，采用决策树为XGBoost模型的基学习器。RF中每个决策树都可以单独进行分类，最后将分类结果平均。而在XGBoost中，不同基学习器汇总起来进行输出预测。T个基学习器构建的XGBoost模型的输出为：

$$\hat{y}_i = \sum_{t=1}^{T} f_t(x_i)$$

XGBoost模型训练时，采用前向分步法，每一步只学习一个基学习器，逐步逼近优化目标，如下图所示。首先确定初始提升树$\hat{y}_i^0 = f_0(x_i) = 0$，然后依次求解获得$f_1(x_i)$，$f_2(x_i)$，……，和$f_T(x_i)$。第t步求解的模型是：

$$\hat{y}_i^t = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{t-1} + f_t(x_i)$$

其中，$\hat{y}_i^t$表示总共t棵决策树组成的模型，$f_t(x_i)$表示第t棵决策树的输出。实际上，$f_t(x_i)$用来拟合前$t-1$棵决策树的输出$\hat{y}_i^{t-1}$与真实值之间的残差。通过不断生成新的决策树模型，XGBoost可以逐步逼近拟合目标。

图 9：XGBoost模型求解示意图

初始设置：$\hat{y}_i^0 = 0$

第1步：$\hat{y}_i^1 = f_1(x_i) = \hat{y}_i^0 + f_1(x_i)$ ⟹ $f_1(x_i)$

第2步：$\hat{y}_i^2 = f_1(x_i) + f_2(x_i) = \hat{y}_i^1 + f_2(x_i)$ ⟹ $f_2(x_i)$

……

第t步：$\hat{y}_i^t = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{t-1} + f_t(x_i)$ ⟹ $f_t(x_i)$

数据来源：广发证券发展研究中心

求解第t棵决策树时，最小化以下目标函数：

$$Obj^t = \sum_{i=1}^{N} l(y_i, \hat{y}_i^t) + \sum_{i=1}^{t} \Omega(f_i)$$

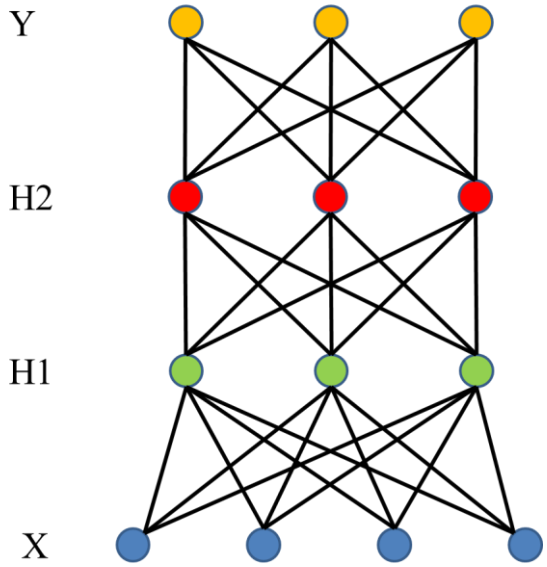
$$= \sum_{i=1}^{N} l\left(y_i, \hat{y}_i^{t-1} + f_t(x_i)\right) + \Omega(f_t) + constant$$

其中，$\Omega(f_i)$表示决策树的复杂度。通过对上述目标函数在$\hat{y}_i^{t-1}$附近作二阶泰勒展开，以获得决策树模型的最优参数。

2.6 深层神经网络

深层神经网络（Deep Neural Networks，DNN）是深度学习主流的实现方法，通过多层神经网络，建立起输入和输出的关系。深层神经网络一般含有多个隐层，如下图所示。本报告采用包含 5 个隐层的全连接深层神经网络模型。

图 10：DNN示意图



数据来源：广发证券发展研究中心

其中，第l个隐含层（l=1，2，3，4，5）的节点 j 可表示为

$$h_j^{(l)} = \sigma\left\{\boldsymbol{w}_j^{(l-1)}\boldsymbol{h}^{(l-1)}\right\}$$

即第$l-1$个隐含层的节点$\boldsymbol{h}^{(l-1)}$经过线性加权，再经过非线性激活函数σ变换之后的值。输出层的节点k可表示为

$$\hat{y}_k = \sigma_o\left\{\boldsymbol{w}_k^{(5)}\boldsymbol{h}^{(5)}\right\}$$

其中，分类模型中输出层节点的激活函数σ_o为 softmax 激活函数。本报告的分类模型中采用交叉熵（cross entropy）损失函数作为分类神经网络模型优化的目标函数:

$$E(\boldsymbol{w}) = -\sum_{n=1}^{N}\sum_{k=1}^{K}[y_{nk}\log\hat{y}_{nk} + (1-y_{nk})\log(1-\hat{y}_{nk})]$$

DNN 模型训练时，一般采用误差反向传播的方式求取梯度，优化参数。

为了提高模型的泛化能力，本报告采用 Dropout 方法，每次参数迭代更新时随机选择丢弃不同的隐层节点，这驱使每个隐层节点去学习更加有用的、不依赖于其他节点的特征。同时，本报告采用 Batch Normalization 技术提高模型的训练效率。

在模型训练中，由于深度学习模型的训练速度很慢，本报告采用 GPU 进行加速。

2.7 不同机器学习模型的特点分析

在以上 5 个分类模型中，MLR 和 SVM 都是线性分类器，但优化目标不同。MLR 以极大化对数似然函数为优化目标，而 SVM 模型目标是最大化分类间隔。优化目标的不同导致 MLR 和 SVM 对噪声信息的处理有差别。

RF 和 XGBoost 模型都是集成学习的典型模型，而且都采用决策树为基学习器。但 RF 中不同的决策树是彼此独立的，最终分类结果是不同决策树分类结果的平均，模型训练时，可以同时训练不同的决策树。XGBoost 中，不同决策树不是彼此独立

的，每次新加入的决策树是为了拟合此前决策树的预测残差，最终分类结果是不同决策树分类结果的加总，从模型训练流程而言，决策树是依次训练出来的。

从模型的线性和非线性来看， MLR 和线性 SVM 属于线性分类器，而 RF、XGBoost 和 DNN 模型属于非线性分类器。

从模型的输出数据来看，MLR、RF、XGBoost 和 DNN 模型都可以直接获得分类概率，而 SVM 模型不能够直接获得分类概率，本报告中根据样本与分类超平面的有向距离构建了一个类似概率的分类结果。

从输入特征来看，RF、XGBoost 模型是基于决策树构建的模型，可以处理连续特征和标签特征。而 MLR、SVM 和 DNN 模型不能直接处理标签特征，一种可行的方案是将标签特征预处理成独热编码然后进行处理，例如将 28 个申万一级行业属性转化成 28 个 0-1 哑变量。

此外，RF 和 XGBoost 模型对数据的分布没有要求，而 MLR、SVM 和 DNN 模型需要对数据进行标准化，使得不同特征的均值和方差可比。

三、选股策略描述

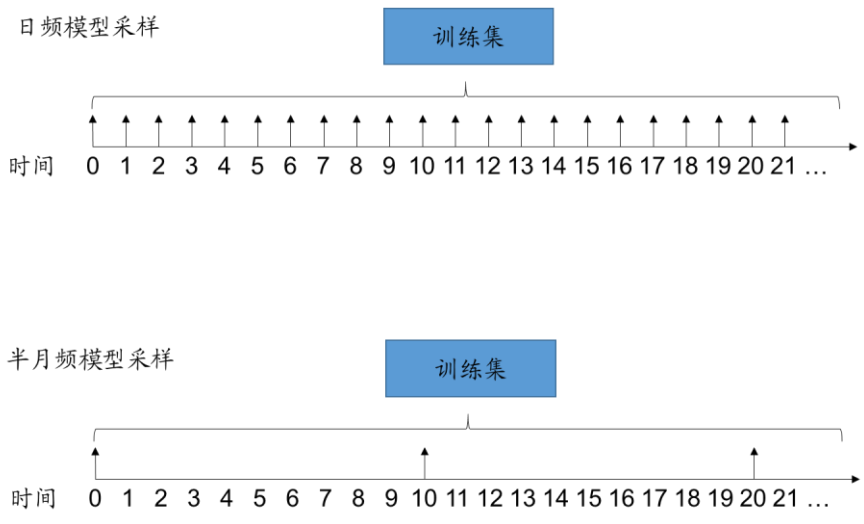
3.1 模型训练方法

本报告考察在不同训练样本量下，机器学习策略的表现。因此采用日频样本和半月频样本两种采样模式进行比较。

日频样本模式在获取样本时，将训练集时间区间内每个交易日的“上涨”、“平盘”和“下跌”样本全部作为历史样本，通过该样本进行机器学习建模。半月频样本模式在获取样本时，每 10 个交易日进行一次采样。如图 11 所示。

因此，日频样本模式下，训练集的样本量（平均为 48 万）大约为半月频样本模式的训练样本量（平均为 4.8 万）的 10 倍。但是，在日频样本模式下，样本之间的独立性比半月频样本模式下要差。在日频样本模式下，t 日标记为“上涨”类别的股票样本在 t+1 日也更有可能属于“上涨”类别。

图 11：日频样本和半月频样本采样示意图



数据来源：广发证券发展研究中心

考虑到市场在不断演化，市场数据中包含的信息也在不断更新，因此，本报告每半年重新训练模型，每次训练时，采用过去 4 年的数据训练模型。如下图所示。模型 2017A 表示最新训练样本是 2017 年上半年的数据，模型 2017B 表示最新训练样本是 2017 年下半年的数据。从 2017 年 7 月 1 日之后使用模型 2017A 进行选股（实际上，如果训练集样本输出中包含了 2017 年 7 月前 10 个交易日的信息，那么模型首次使用时间是 2017 年 7 月份的第 11 个交易日，下同），从 2018 年 1 月 1 日之后使用模型 2017B 进行选股。

图 12：模型滚动更新示意图

数据来源：广发证券发展研究中心

3.2 机器学习训练平台和模型超参数

本报告中的模型是在 Python 3.6 平台训练和预测的。其中，MLR、SVM 和 RF 模型通过 scikit learn 模块（sklearn）实现，XGBoost 模型通过 xgboost 模块实现，DNN 模型通过 tensorflow 和其高级封装 keras 模块实现。xgboost 和 tensorflow 模块支持 GPU 运算和 CPU 运算，sklearn 模块只支持 CPU 运算。本报告中，仅在 DNN 模型训练时采用 GPU 运算。

本报告中，采用交叉验证方法获取模型超参数：

MLR 中，通过交叉验证确定正则化参数 C。

非线性 SVM 的计算复杂度在$O(mN^2)$到$O(mN^3)$之间，当样本量N很大时计算量巨大，难以获得最优参数，因此本报告采用线性 SVM 构建模型。线性 SVM 中，通过交叉验证获取松弛变量惩罚系数 C。

RF 中，通过交叉验证确定决策树最大深度和决策树数量。

XGBoost 中，通过交叉验证确定决策树最大深度、决策树数量、样本采样比例和特征采样比例。

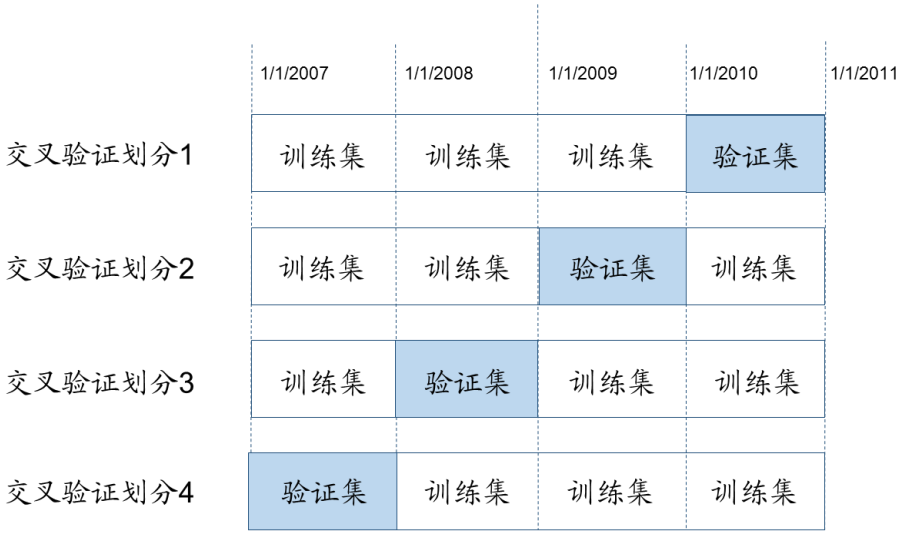
DNN 中，通过交叉验证确定神经网络的隐层节点数量。

K 折交叉验证是常见的交叉验证方法。K 折交叉验证随机将样本等分成 K 组，每次取其中一组作为测试集，其他组作为训练集，考察不同的超参数在测试集的表现，根据 K 次实验确定最优的超参数。但在机器学习因子选股模型中，将每个时刻的单个股票样本视为一个训练样本，K 折交叉验证在随机分组的时候，可能将同一个交易日的不同股票样本同时划分到训练集和测试集中，这样不利于判断模型是否在

与训练集样本不同的时刻表现依然有效。

因此，本报告中先按照时间顺序，将股票样本分组，然后采用分组 K 折交叉验证方法进行处理。本报告中采用 2007 年至 2010 年的股票数据进行分组 4 折交叉验证，首先按照时间顺序将股票样本分成 4 个不同的组，在每次交叉验证计算时，选取其中一组作为验证集，其他组作为训练集，重复 4 次计算并取平均，如下图所示。在这种情况下，同一个时刻的样本不会同时被划分到训练集和验证集中。

图 13：时间分组交叉验证示意图



数据来源：广发证券发展研究中心

对日频样本下的数据进行分组交叉验证，得到的模型超参数如下：

MLR 中，正则化参数 C=0.01。

线性 SVM 中，松弛变量惩罚系数 C=0.01。

RF 中，决策树最大深度为 14，决策树数量为 250。

XGBoost 中，决策树最大深度为 5，决策树数量为 100，样本采样比例为 0.8，特征采样比例为 0.9。

DNN 中，5 个隐层节点数量依次为 512、200、200、200、128。

本报告的后续测算按照上述超参数进行模型的测试和比较。

3.3 策略回测设置

在每个调仓日，根据机器学习模型的打分，筛选打分靠前的 10%的股票构建组合，进行策略的回测。相关参数如下：

调仓周期：10 个交易日；

股票池：全市场选股、剔除涨停、跌停的股票，停牌股票和 ST 股票；

超配组合：机器学习打分前 10%的股票；

对冲基准：中证 500 指数；

原始因子数据：估值因子、规模因子、反转因子、流动性因子、波动性因子、技术指标，共计 128 个因子，以及 28 个行业 0-1 变量；

机器学习模型训练：每半年滚动更新模型，采用最近 4 年的样本作为训练集；
组合构建：等权、行业中性两种方案；
策略回测：2011 年 1 月-2019 年 4 月 26 日；
交易成本：千分之三。

四、实证分析

4.1 机器学习模型预测性能比较

采用日频样本训练模型时，每隔半年训练的机器学习模型对测试集的预测准确率如下表所示。其中第二列展示了每个模型训练时所用的训练样本数量，日频样本训练模式下，平均每个模型训练时的训练样本为48万。对于三分类问题，随机预测的准确率为33.3%。可以看到，不同的机器学习模型的样本外预测准确率都显著超过了随机预测。其中，DNN模型的预测准确率最高，其次为XGBoost模型。

表 1：日频样本模型测试集预测准确率

模型	训练样本数量	MLR	SVM	RF	XGBoost	DNN
2010B	299800	48.7%	49.0%	54.2%	55.8%	58.6%
2011A	319900	49.0%	49.1%	54.2%	56.0%	58.5%
2011B	343500	48.6%	48.6%	53.7%	55.3%	57.9%
2012A	369300	48.6%	48.7%	53.6%	55.0%	58.6%
2012B	400100	48.4%	48.4%	52.9%	55.2%	59.3%
2013A	427800	49.1%	49.2%	53.9%	55.7%	59.2%
2013B	461300	49.5%	49.5%	53.9%	55.5%	58.9%
2014A	492100	49.8%	49.8%	54.3%	56.0%	59.9%
2014B	520600	49.5%	49.5%	54.4%	55.6%	58.5%
2015A	535500	50.1%	50.1%	54.9%	56.0%	59.7%
2015B	536200	50.6%	50.6%	54.4%	56.2%	59.7%
2016A	549900	51.0%	50.9%	54.6%	56.5%	60.2%
2016B	561700	51.5%	51.6%	54.9%	56.7%	59.8%
2017A	572600	51.1%	51.1%	55.1%	56.7%	60.1%
2017B	582900	51.2%	51.1%	55.5%	56.9%	59.3%
2018A	602800	51.0%	50.9%	55.6%	57.0%	60.4%
2018B	617600	51.4%	51.2%	55.9%	57.3%	59.6%
平均值	481976.5	49.9%	50.0%	54.5%	56.1%	59.3%

数据来源：Wind，广发证券发展研究中心

采用半月频样本训练模型时，平均每个模型训练时的训练样本为4.8万。不同的机器学习模型的样本外预测准确率也都超过了随机预测的表现。其中，XGBoost模型的预测准确率最高。

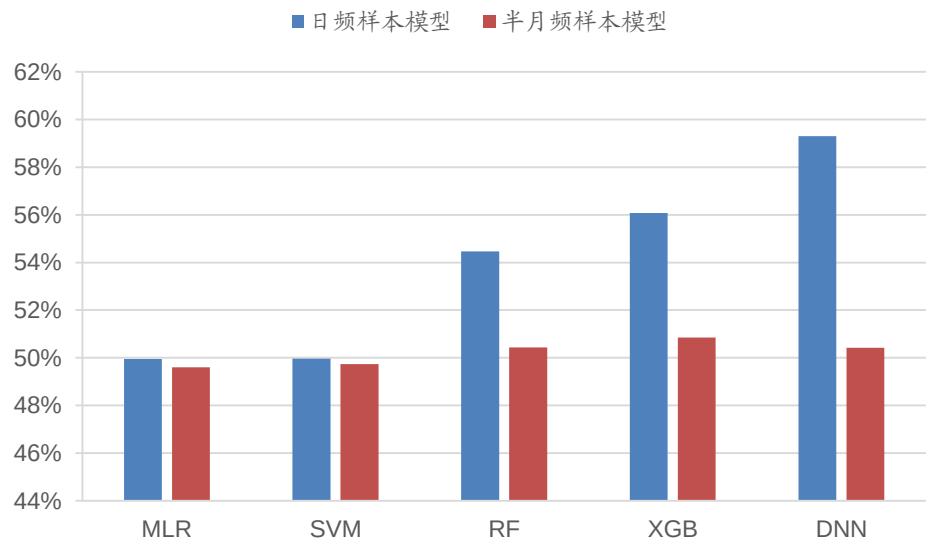
表 2：半月频样本模型测试集预测准确率

模型	样本数量	MLR	SVM	RF	XGBoost	DNN
2010B	29700	48.1%	48.7%	49.6%	49.6%	44.4%
2011A	32000	49.7%	50.0%	49.7%	48.9%	48.0%
2011B	34500	48.3%	48.5%	49.6%	49.4%	49.2%
2012A	37000	47.4%	48.1%	48.6%	49.0%	47.8%
2012B	39900	48.2%	48.5%	49.2%	49.3%	47.9%
2013A	42700	48.5%	49.5%	48.6%	49.8%	49.9%
2013B	46100	48.9%	49.1%	50.0%	50.2%	50.6%
2014A	49200	49.8%	49.5%	51.4%	50.6%	50.3%
2014B	51700	49.1%	49.3%	49.3%	49.4%	49.2%
2015A	53300	49.9%	49.3%	50.2%	50.0%	50.0%
2015B	52900	50.8%	50.9%	51.2%	52.4%	53.0%
2016A	55000	50.4%	50.3%	51.0%	51.8%	52.5%
2016B	55600	50.4%	50.2%	51.2%	51.7%	53.0%
2017A	57100	51.6%	51.7%	52.0%	52.7%	52.6%
2017B	57600	49.5%	49.8%	51.3%	51.7%	52.3%
2018A	59500	51.8%	51.3%	52.7%	54.2%	53.8%
2018B	62100	50.7%	50.7%	52.0%	53.4%	52.8%
平均值	47994.1	49.6%	49.7%	50.4%	50.8%	50.4%

数据来源：Wind，广发证券发展研究中心

半月频样本训练模式与日频样本训练模式比较的结果如图14所示。所有的日频样本模型的预测准确率都高于半月频样本模型，尤其是RF、XGB和DNN模型，日频样本模型的表现明显优于半月频样本模型。

图 14：日频样本模型与半月频样本模型测试集准确率对比

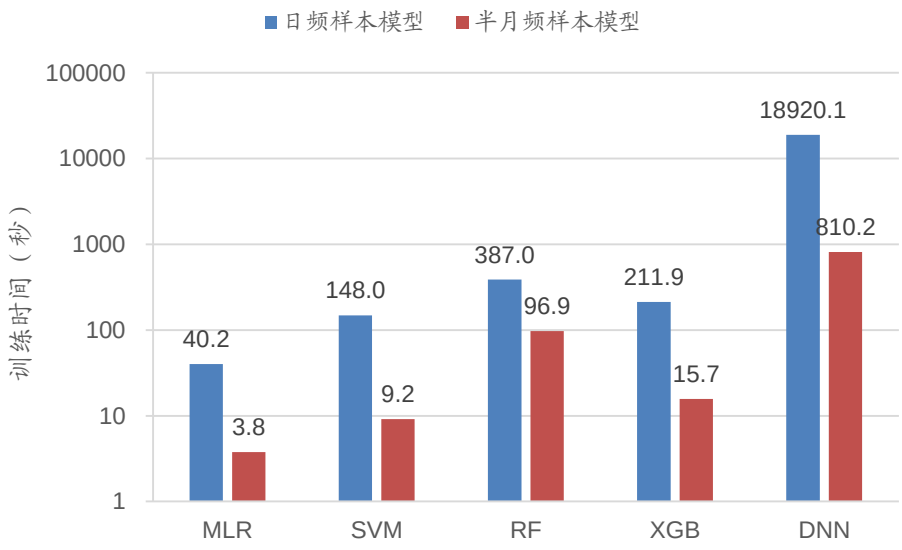


数据来源：Wind，广发证券发展研究中心

从训练时间来看，MLR模型训练时间最短，采用日频样本训练模型的平均训练时间为40.2秒，采用半月频样本训练模型的平均训练时间为3.8秒。DNN模型训练时间最长，采用半月频样本训练模型的平均训练时间为810.2秒，采用日频样本训练模型的平均训练时间为18920.1秒（5.3小时）。

总体来看，日频样本下，DNN模型的表现最好，但模型训练耗时最长。半月频样本下，XGBoost模型表现最好，而且模型训练耗时和线性模型差别不大。

图 15：日频样本模型与半月频样本模型训练时间对比



数据来源：Wind，广发证券发展研究中心

4.2 机器学习模型打分相关性分析

日频样本模型的IC和半月频样本模型的IC分别如表3所示。可以看到，与半月频样本模型相比，日频样本模型具有更高的IC和更高的ICIR。在日频样本模型中，DNN模型的IC和ICIR最高，其次是XGBoost模型，MLR、SVM和RF模型的IC差不多。在半月频样本模型中，不同模型的差异比日频样本模型小。

表 3：机器学习模型IC

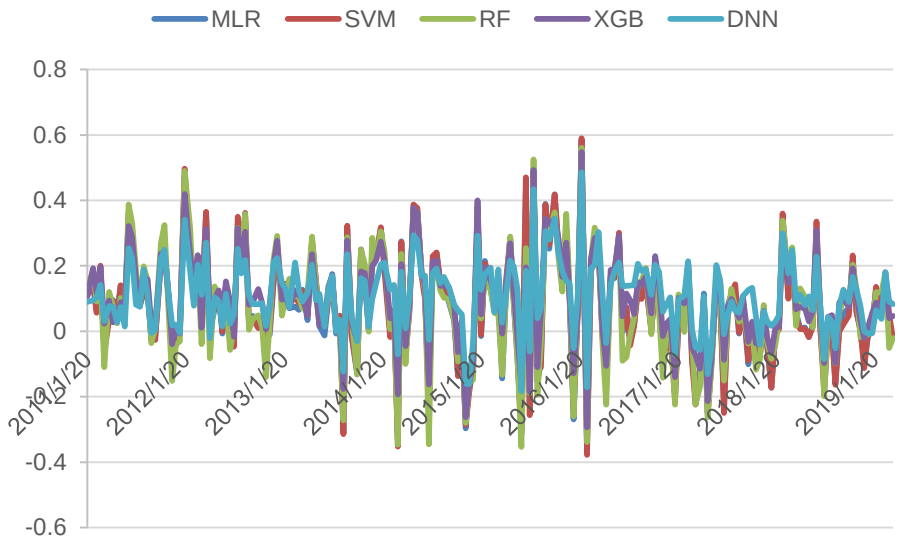
模型	指标	MLR	SVM	RF	XGBoost	DNN
日频样本模型	IC 平均值	7.4%	7.5%	7.8%	9.9%	10.7%
	IC 标准差	16.2%	16.6%	16.7%	13.5%	10.5%
	ICIR	0.46	0.46	0.47	0.73	1.02
半月频样本模型	IC 平均值	6.5%	6.7%	6.2%	7.8%	8.4%
	IC 标准差	16.0%	15.8%	15.3%	12.0%	13.7%
	ICIR	0.41	0.42	0.41	0.65	0.61

数据来源：Wind，广发证券发展研究中心

日频样本训练的模型中，不同机器学习模型的IC序列如下图所示。可以看到，

不同机器学习模型的IC相关性很高。这意味着不同机器学习模型的表现有很强的相关性。

图 16：日频样本模型IC序列



数据来源：Wind，广发证券发展研究中心

为了分析不同机器学习模型表现的相关性，本报告从模型打分的相关性和模型IC的相关性两方面进行比较。模型打分的相关性是指在同一个时间截面，不同机器学习模型对股票打分的相关性。模型打分的相关性体现了不同机器学习模型对股票收益率的预测结果的相似程度。

下表展示了日频样本训练时，不同的机器学习模型打分的相关系数均值。可以看到，MLR和SVM模型的相关性非常高，相关系数超过0.97，这是因为MLR和线性SVM都是线性模型。DNN模型和其他机器学习模型的相关性相对较弱。RF和XGBoost虽然都是以决策树模型为基学习器的集成学习方法，但由于集成方式差别很大，打分相关性并不特别高。

表 4：日频样本机器学习模型打分相关性

模型	MLR	SVM	RF	XGBoost	DNN
MLR	1	0.976	0.712	0.683	0.571
SVM		1	0.692	0.674	0.569
RF			1	0.718	0.564
XGBoost				1	0.606
DNN					1

数据来源：Wind，广发证券发展研究中心

下表展示了半月频样本训练时，不同的机器学习模型打分的相关系数均值。与日频样本模型相似， MLR模型和SVM模型的相关性非常高，相关系数超过0.97。

XGBoost模型和其他机器学习模型的相关性相对较弱。可以看到，采用半月频样本训练模型时，DNN模型和其他模型的相关性比日频样本训练时要高，意味着DNN模型的表现和其他模型差别不大。

表 5：半月频样本机器学习模型打分相关性

模型	MLR	SVM	RF	XGBoost	DNN
MLR	1	0.972	0.702	0.638	0.764
SVM		1	0.656	0.624	0.760
RF			1	0.657	0.647
XGBoost				1	0.632
DNN					1

数据来源：Wind，广发证券发展研究中心

模型IC的相关性是指不同机器学习模型选股的IC时间序列的相关性。模型IC的相关性体现了不同机器学习模型选股表现的相似程度。下表展示了日频训练样本下，不同机器学习模型IC的相关性。模型IC的相关性明显高于模型打分的相关性。其中，DNN和RF模型IC的相关性最小，为0.848。MLR和SVM模型IC的相关性最强，高达0.994。这意味着不同机器学习模型的选股表现具有很强的相关性，当其中一个模型表现好时，其他模型也更有可能表现好；当其中一个模型表现不佳时，其他模型也更有可能表现不佳。

表 6：日频样本机器学习模型IC相关性

模型	MLR	SVM	RF	XGBoost	DNN
MLR	1	0.994	0.946	0.961	0.860
SVM		1	0.938	0.957	0.862
RF			1	0.943	0.848
XGBoost				1	0.894
DNN					1

数据来源：Wind，广发证券发展研究中心

下表展示了半月频训练样本下，不同机器学习模型IC的相关性。与日频样本训练的模型相比，半月频样本模型中，模型IC之间的相关性更高，其中相关性最低的是DNN模型和RF模型的相关系数，为0.908。

表 7：半月频样本机器学习模型IC相关性

模型	MLR	SVM	RF	XGBoost	DNN
MLR	1	0.993	0.949	0.957	0.943
SVM		1	0.926	0.950	0.938
RF			1	0.950	0.908
XGBoost				1	0.939
DNN					1

数据来源：Wind，广发证券发展研究中心

4.3 机器学习模型选股表现

日频样本训练时， 不同机器学习模型等权选股策略的表现如下表所示。可以看到，不同策略都有不错的表现。MLR和SVM收益稍差，也有30%左右的年化对冲收益，DNN表现最好，年化对冲收益高达38.51%。但是MLR、SVM和RF模型的回撤稍大。DNN模型具有最高的年化收益和夏普比率。

表 8：不同机器学习模型等权选股策略对冲表现（日频样本）

年份	MLR	SVM	RF	XGBoost	DNN
累积收益率	765.75%	796.19%	1203.96%	1127.61%	1369.75%
年化收益率	29.90%	30.45%	36.52%	35.52%	38.51%
最大回撤	-18.36%	-20.20%	-21.98%	-13.64%	-10.96%
夏普比率	2.65	2.67	3.04	3.43	4.20

数据来源：Wind，广发证券发展研究中心

不同机器学习模型等权选股策略的净值表现如图17所示。如上文所述，不同机器学习模型选股的收益曲线有较强的相关性。5种机器学习选股策略在2015年都收益很高，而在2017年有一定的回撤。2016年之前表现最好的是RF模型，但2017年RF模型回撤最大，而DNN模型相对更稳定。

图 17：日频机器学习模型选股表现（等权）

数据来源：Wind，广发证券发展研究中心

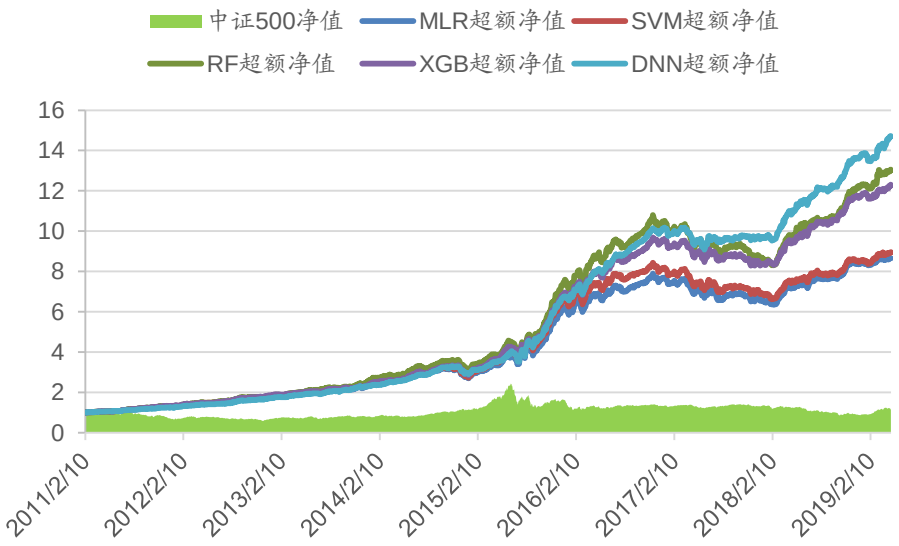
分年度的策略表现如下表所示。在2017年，不同机器学习模型策略的对冲收益都为负，其中表现最好的是DNN。在其他年份，所有机器学习模型的对冲收益都为正。对冲策略的净值比较如图18所示。

表 9：不同机器学习模型等权选股策略分年度对冲收益（日频样本）

年份	MLR	SVM	RF	XGBoost	DNN
2011	31.17%	29.89%	31.17%	31.08%	23.80%
2012	34.86%	37.05%	38.21%	41.20%	38.11%
2013	32.70%	34.92%	41.67%	27.82%	35.58%
2014	17.49%	16.69%	23.37%	23.02%	28.07%
2015	126.01%	137.08%	139.64%	139.10%	127.83%
2016	22.92%	22.66%	37.50%	37.16%	49.46%
2017	-13.98%	-14.81%	-16.67%	-12.06%	-4.01%
2018	26.91%	22.19%	39.99%	39.08%	40.32%
2019	3.50%	5.64%	7.07%	5.17%	7.93%

数据来源：Wind，广发证券发展研究中心

图 18：日频样本机器学习模型选股对冲收益（等权）



数据来源：Wind，广发证券发展研究中心

行业中性组合下， 不同机器学习模型选股策略的对冲表现如下表所示。其中，MLR模型收益稍差，年化对冲收益为24.08%，DNN模型表现最好，年化对冲收益为29.06%。行业中性下，SVM和RF模型仍然有较大的回撤。DNN模型具有最高的年化收益和夏普比率。

表 10：不同机器学习模型行业中性选股策略对冲表现（日频样本）

年份	MLR	SVM	RF	XGBoost	DNN
累积收益率	493.11%	562.42%	635.45%	567.26%	720.51%
年化收益率	24.08%	25.76%	27.36%	25.87%	29.06%
最大回撤	-9.81%	-11.24%	-15.49%	-8.62%	-6.53%
夏普比率	2.56	2.72	2.68	3.22	3.82

数据来源：Wind，广发证券发展研究中心

不同机器学习模型行业中性选股策略的净值表现如下图所示。不同机器学习模型选股的收益曲线有较强的相关性。与等权策略相比，行业中性策略的回撤较小。其中，XGBoost和DNN模型的回撤最小。2017年，只有DNN模型获得了正的对冲收益。

图 19：日频机器学习模型选股表现（行业中性）

数据来源：Wind，广发证券发展研究中心

分年度的策略表现如下表所示。DNN模型在所有年份都获得了正的对冲收益。

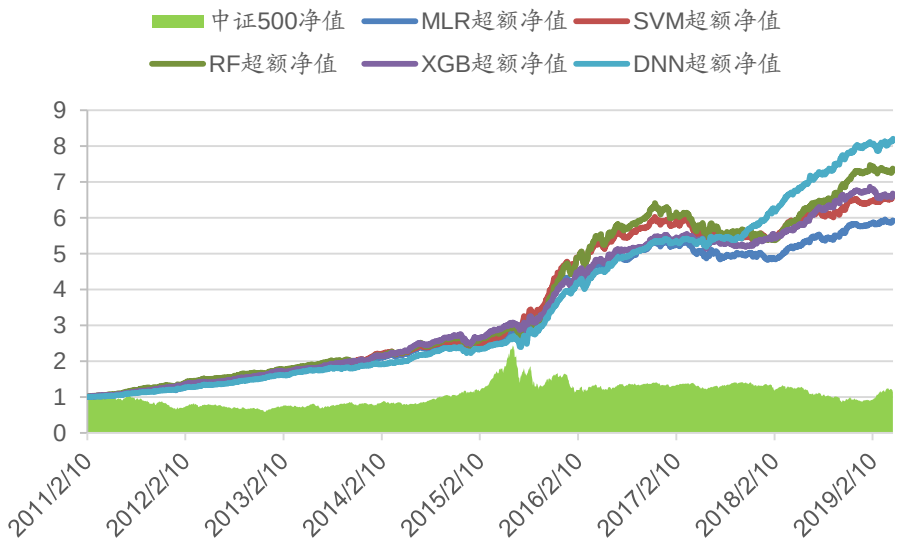
表 11：不同机器学习模型行业中性选股策略分年度对冲收益（日频样本）

年份	MLR	SVM	RF	XGBoost	DNN
2011	23.41%	26.89%	32.23%	26.12%	18.96%
2012	32.00%	32.89%	30.33%	34.30%	32.47%
2013	20.71%	24.26%	18.98%	20.70%	19.82%
2014	16.88%	12.31%	20.57%	23.06%	20.38%
2015	88.58%	103.00%	91.38%	68.80%	74.38%
2016	22.62%	23.81%	31.95%	29.31%	35.97%
2017	-6.70%	-6.65%	-11.98%	-2.02%	10.06%
2018	16.41%	16.32%	32.06%	24.96%	34.83%
2019	2.75%	3.15%	1.34%	-0.76%	2.59%

数据来源：Wind，广发证券发展研究中心

对冲策略的净值比较如下图所示。2016年之前，DNN相比其他机器学习模型稍差，但整个回测区间的表现更好一些。

图 20：日频样本机器学习模型选股对冲收益（行业中性）



数据来源：广发证券发展研究中心

类似的，可以计算半月频样本训练的模型的表现。表12展示了等权组合策略中，半月频样本模型与日频样本模型的表现对比。可以看到，不同机器学习模型在日频样本训练时的表现都优于半月频样本训练时的表现。其中，MLR在日频样本下训练的选股模型的夏普比率为2.65，而半月频样本下训练的选股模型的夏普比率为2.24。在半月频样本下，表现最好的模型是XGBoost，夏普比率为2.98。而在日频样本下，表现最好的模型是DNN，夏普比率为4.20。在日频样本下和半月频样本下，表现差异最大的是DNN模型。

表 12：机器学习模型选股性能比较（等权组合）

模型	指标	MLR	SVM	RF	XGBoost	DNN
日频样本模型	年化收益率	29.90%	30.45%	36.52%	35.52%	38.51%
	最大回撤	-18.36%	-20.20%	-21.98%	-13.64%	-10.96%
	夏普比率	2.65	2.67	3.04	3.43	4.20
半月频样本模型	年化收益率	25.52%	24.87%	30.19%	28.92%	27.09%
	最大回撤	-20.55%	-20.60%	-21.66%	-16.37%	-15.72%
	夏普比率	2.24	2.21	2.59	2.98	2.60

数据来源：Wind，广发证券发展研究中心

行业中性组合策略中，总体来看，日频样本下训练的模型表现更佳。但MLR、SVM和RF模型在两种频率样本下的行业中性选股表现差别不大。在日频样本下和半月频样本下，表现差异最大的是DNN模型。

表 13：机器学习模型选股性能比较（行业中性组合）

模型	指标	MLR	SVM	RF	XGBoost	DNN
日频样本模型	年化收益率	24.08%	25.76%	27.36%	25.87%	29.06%
	最大回撤	-9.81%	-11.24%	-15.49%	-8.62%	-6.53%
	夏普比率	2.56	2.72	2.68	3.22	3.82
半月频样本模型	年化收益率	26.56%	26.66%	24.29%	22.78%	23.09%
	最大回撤	-16.35%	-15.21%	-17.83%	-10.64%	-8.10%
	夏普比率	2.95	2.98	2.64	2.98	2.89

数据来源：Wind，广发证券发展研究中心

4.4 机器学习模型的风格分析

本报告通过考察机器学习打分和风格因子的相关性，对机器学习模型的风格暴露进行分析。

本报告采用了7个风格因子进行策略的风格分析。分别表示股票的规模因子、Beta因子、反转因子、波动性因子、流动性因子、估值因子和杠杆因子等7类不同的风格特征。

表 14：风格因子列表

风格因子	因子说明	因子方向
规模因子	流通市值	负向
Beta 因子	股票的 120 日 beta	负向
反转因子	月度反转	负向
波动性因子	月度波动率	负向
流动性因子	月均换手率	负向
估值因子	盈市率	正向
杠杆因子	资产负债率	负向

数据来源：Wind，广发证券发展研究中心

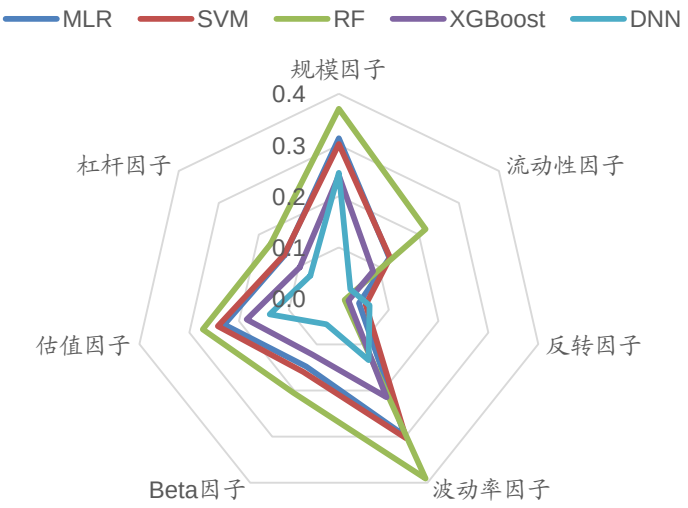
分别计算不同的机器学习模型的选股打分与风格因子的相关性。在日频样本训练的模型中，不同机器学习模型选股打分与风格因子的截面相关系数均值如下表所示。总体来看，DNN模型与不同风格因子的相关性相对较弱，RF模型与不同风格因子的相关性相对较强。为了便于比较，将相关系数取绝对值之后，画成雷达图。雷达图清楚的展示了不同机器学习模型选股的风格暴露情况。风格暴露最小的是DNN模型，其次是XGBoost模型，而RF模型的风格暴露最大。

表 15：不同机器学习模型选股打分与风格因子相关性（日频样本）

相关系数	MLR	SVM	RF	XGBoost	DNN
规模因子	-0.313	-0.302	-0.371	-0.241	-0.246
流动性因子	0.127	0.128	0.217	0.086	0.029
反转因子	-0.040	-0.054	-0.012	-0.020	-0.062
波动率因子	0.298	0.303	0.391	0.215	0.134
Beta 因子	0.148	0.160	0.204	0.122	0.056
估值因子	-0.230	-0.242	-0.273	-0.185	-0.139
杠杆因子	-0.136	-0.136	-0.171	-0.098	-0.071

数据来源：Wind，广发证券发展研究中心

图 21：不同机器学习模型选股打分与风格因子相关性雷达图（日频样本）



数据来源：Wind，广发证券发展研究中心

在半月频样本训练的模型中，不同机器学习模型选股打分与风格因子的截面相关系数均值如下表所示。总体来看，XGBoost和DNN模型与传统风格因子的相关性最小。

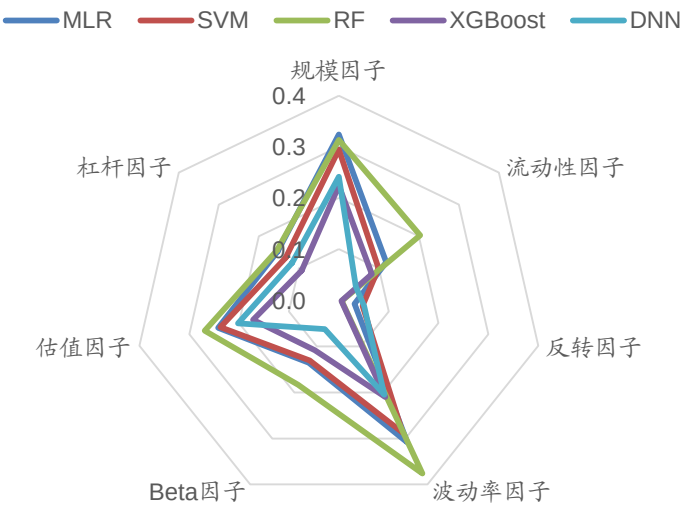
类似的，从雷达图可以看到，风格暴露最小的是XGBoost模型，其次是DNN模型，而RF模型的风格暴露最大。

表 16：不同机器学习模型选股打分与风格因子相关性（半月频样本）

相关系数	MLR	SVM	RF	XGBoost	DNN
规模因子	-0.324	-0.295	-0.313	-0.224	-0.241
流动性因子	0.119	0.100	0.203	0.082	0.043
反转因子	-0.032	-0.047	0.006	-0.005	-0.051
波动率因子	0.308	0.287	0.376	0.210	0.205
Beta 因子	0.136	0.131	0.184	0.109	0.063
估值因子	-0.242	-0.237	-0.269	-0.172	-0.202
杠杆因子	-0.153	-0.133	-0.155	-0.093	-0.117

数据来源：Wind，广发证券发展研究中心

图 22：不同机器学习模型选股打分与风格因子相关性雷达图（半月频样本）



数据来源：Wind，广发证券发展研究中心

五、总结与展望

本报告通过实证研究，展示了不同机器学习模型在因子选股上的可行性。5种机器学习模型都取得了显著的超额收益。由于模型都是从历史数据建立起股票因子和收益率的关系，不同机器学习模型表现有较大的相关性，模型打分相关性和模型IC相关性都比较高。

日频样本平均每次用48万个样本训练模型。在5种不同的模型中，DNN模型表现最佳，具有最高的IC、ICIR、年化对冲收益和夏普比率。但是DNN模型的训练耗时，平均每个模型训练需要5个多小时。

半月频样本平均每次用4.8万个样本训练模型。在5种不同的模型中，XGBoost模型表现最好，而且训练时间和线性分类模型差别不大。

总体来看，日频样本模式训练的模型表现优于半月频样本模式训练的模型。尤

其是DNN和XGBoost模型，日频样本模式明显优于半月频样本模式。

从机器学习模型打分来看模型的风格暴露情况，DNN模型和XGBoost模型在风格因子上的暴露相对较少，而RF模型在风格因子上的暴露最大。从回测结果来看，RF模型选股策略的回撤较大，可能与风格因子暴露较多有一定关系。

风险提示

策略模型并非百分百有效，市场结构及交易行为的改变以及类似交易参与者的增多有可能使得策略失效。

广发证券—行业投资评级说明

买入：预期未来 12 个月内，股价表现强于大盘 10%以上。
持有：预期未来 12 个月内，股价相对大盘的变动幅度介于-10%～+10%。
卖出：预期未来 12 个月内，股价表现弱于大盘 10%以上。

广发证券—公司投资评级说明

买入：预期未来 12 个月内，股价表现强于大盘 15%以上。
增持：预期未来 12 个月内，股价表现强于大盘 5%-15%。
持有：预期未来 12 个月内，股价相对大盘的变动幅度介于-5%～+5%。
卖出：预期未来 12 个月内，股价表现弱于大盘 5%以上。

联系我们

	广州市	深圳市	北京市	上海市	香港
地址	广州市天河区马场路 26 号广发证券大厦 35 楼	深圳市福田区益田路 6001 号太平金融大厦 31 层	北京市西城区月坛北街 2 号月坛大厦 18 层	上海市浦东新区世纪大道 8 号国金中心一期 16 楼	香港中环干诺道中 111 号永安中心 14 楼 1401-1410 室
邮政编码	510627	518026	100045	200120	
客服邮箱	gfyf@gf.com.cn				

法律主体声明

本报告由广发证券股份有限公司或其关联机构制作，广发证券股份有限公司及其关联机构以下统称为“广发证券”。本报告的分销依据不同国家、地区的法律、法规和监管要求由广发证券于该国家或地区的具有相关合法合规经营资质的子公司/经营机构完成。

广发证券股份有限公司具备中国证监会批复的证券投资咨询业务资格，接受中国证监会监管，负责本报告于中国（港澳台地区除外）的分销。

广发证券（香港）经纪有限公司具备香港证监会批复的就证券提供意见（4 号牌照）的牌照，接受香港证监会监管，负责本报告于中国香港地区的分销。

本报告署名研究人员所持中国证券业协会注册分析师资质信息和香港证监会批复的牌照信息已于署名研究人员姓名处披露。

重要声明

广发证券股份有限公司及其关联机构可能与本报告中提及的公司寻求或正在建立业务关系，因此，投资者应当考虑广发证券股份有限公司及其关联机构因可能存在的潜在利益冲突而对本报告的独立性产生影响。投资者不应仅依据本报告内容作出任何投资决策。

本报告署名研究人员、联系人（以下均简称“研究人员”）针对本报告中相关公司或证券的研究分析内容，在此声明：（1）本报告的全部分析结论、研究观点均精确反映研究人员于本报告发出当日的关于相关公司或证券的所有个人观点，并不代表广发证券的立场；（2）研究人员的部分或全部的报酬无论在过去、现在还是将来均不会与本报告所述特定分析结论、研究观点具有直接或间接的联系。

研究人员制作本报告的报酬标准依据研究质量、客户评价、工作量等多种因素确定，其影响因素亦包括广发证券的整体经营收入，该等经营收入部分来源于广发证券的投资银行类业务。

本报告仅面向经广发证券授权使用的客户/特定合作机构发送，不对外公开发布，只有接收人才可以使用，且对于接收人而言具有保密义务。广发证券并不因相关人员通过其他途径收到或阅读本报告而视其为广发证券的客户。在特定国家或地区传播或者发布本报告可能违反当地法律，广发证券并未采取任何行动以允许于该等国家或地区传播或者分销本报告。

本报告所提及证券可能不被允许在某些国家或地区内出售。请注意，投资涉及风险，证券价格可能会波动，因此投资回报可能会有所变化，过

去的业绩并不保证未来的表现。本报告的内容、观点或建议并未考虑任何个别客户的具体投资目标、财务状况和特殊需求，不应被视为对特定客户关于特定证券或金融工具的投资建议。本报告发送给某客户是基于该客户被认为有能力独立评估投资风险、独立行使投资决策并独立承担相应风险。

本报告所载资料的来源及观点的出处皆被广发证券认为可靠，但广发证券不对其准确性、完整性做出任何保证。报告内容仅供参考，报告中的信息或所表达观点不构成所涉证券买卖的出价或询价。广发证券不对因使用本报告的内容而引致的损失承担任何责任，除非法律法规有明确规定。客户不应以本报告取代其独立判断或仅根据本报告做出决策，如有需要，应先咨询专业意见。

广发证券可发出其它与本报告所载信息不一致及有不同结论的报告。本报告反映研究人员的不同观点、见解及分析方法，并不代表广发证券的立场。广发证券的销售人员、交易员或其他专业人士可能以书面或口头形式，向其客户或自营交易部门提供与本报告观点相反的市场评论或交易策略，广发证券的自营交易部门亦可能会有与本报告观点不一致，甚至相反的投资策略。报告所载资料、意见及推测仅反映研究人员于发出本报告当日的判断，可随时更改且无需另行通告。广发证券或其证券研究报告业务的相关董事、高级职员、分析师和员工可能拥有本报告所提及证券的权益。在阅读本报告时，收件人应了解相关的权益披露（若有）。

本研究报告可能包括和/或描述/呈列期货合约价格的事实历史信息（“信息”）。请注意此信息仅供用作组成我们的研究方法/分析中的部分论点/依据/证据，以支持我们对所述相关行业/公司的观点的结论。在任何情况下，它并不（明示或暗示）与香港证监会第 5 类受规管活动（就期货合约提供意见）有关联或构成此活动。

权益披露

(1)广发证券（香港）跟本研究报告所述公司在过去 12 个月内并没有任何投资银行业务的关系。

版权声明