专题报告

利用 XGBoost 预测规模因子收益方向

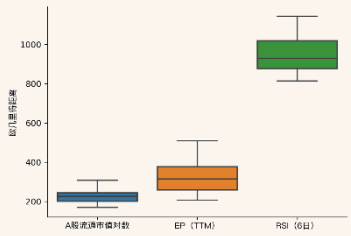
2019 年 01 月 10 日

因子模型系列之十二

在本系列的上一篇报告中我们提到：因子模型是解释性模型，因子收益的方向需要外生变量和外部模型进行预测。本文就试图用 XGBoost 模型结合宏观特征变量与技术指标特征变量对规模类因子的收益方向进行预测。根据规模类因子自身特点出发，对因子的第二类变化进行预测，提取宏观特征变量指标，用整群抽样的方式划分训练集和测试集，调参并构建模型，最后在测试集中进行模型评估。XGBoost 对比动量预测方法，其预测准确性有较为明显的提升。

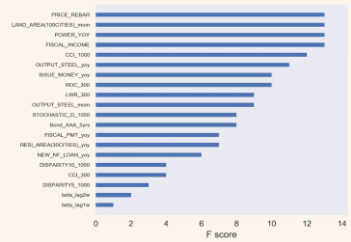
- 尽管在经济学解释上仍存在较大的争议，但是在实证研究和实践中，规模类因子的确不论在对超额收益的解释力上还是对超额收益的贡献上，相较其他类型的因子都有较大的优势。换而言之，它对组合收益的影响，要比其他因子显著很多。
- 相比于其他类型的因子，规模类因子在第一类变化（因子暴露度排序的变化）上的程度是最低的。因而对规模类因子进行预测，当预测其第二类变化（投资者对该因子情绪的变化）。
- 对于第二类变化的预测须从宏观数据入手更为合情合理。
- 我们挑选众多宏观特征变量和技术指标特征变量，根据变量之间的相关系数水平进行适当筛选。用整群抽样的方式来划分训练集和测试集。
- 将特征变量放入 XGBoost 模型进行训练，使用了一些常用的手段去防止模型出现过拟合，而后去预测规模类因子收益的方向。
- 最后评估发现，XGBoost 模型对规模类因子下一周的收益方向预测能有 65.8%的逻辑预测准确率，较基准预测方法（动量预测方法）预测的准确性提升近 20%。
- 在特征变量的重要性排序上，排名前四的分别是：螺纹钢市场价、100 大中城市:成交土地占（环比）、发电量当月同比、公共财政收入当月值。前四均为宏观特征变量，从实证数据说明，宏观变量在预测规模类因子收益方向上，还是有较强的区分力的。
- 两个被解释变量（标签）滞后项区分能力最弱，动量预测在周频数据上乏善可陈。

各代表因子第一类变化比较图



资料来源：招商证券、Wind 资讯

高区分度特征变量重要性排序



资料来源：招商证券、Wind 资讯

相关报告

《因子筛选与投资组合构建》2018-10

《各大类单因子有效性汇总比较分析》2018-04

叶涛
021-68407343
yetao@cmschina.com.cn
S1090514040002

研究助理
崔浩瀚
cuihaohan@cmschina.com.cn

正文目录

图表目录

前期报告提要

在前期的因子系列报告中，我们已经完成了单因子的测试，计算了因子在各期的收益，并根据因子超额暴露对于超额收益的解释能力挑选了适当的因子来构建多因子模型。由于计算后发现，因子数量的增加对于超额收益解释能力的提升有较大的边际递减现象，因而我们的模型中并没有添加过多的因子。所以模型对选股所要考察的维度进行有效降维，并同时最大化对超额收益的解释能力。

因子模型最大的作用在于提取个股之间的共性，对冲个股之间特殊风险，从而对选股信息进行有效降维，只在那些解释力强、且有一定预测和控制能力的因子上有所暴露。同时，在收益实现后对业绩归因进行有效划分。但是因子模型本身不具备预测能力，对因子收益方向的预测需要借助外部信息，使用其他方法进行预测。本报告则试图用宏观数据与技术指标数据对规模类因子（具体指 A 股流通市值对数这个因子）的收益方向进行预测。

尽管在经济学解释上仍存在较大的争议，但是在实证研究和实践中，规模类因子确实不论在对超额收益的解释力上还是对超额收益的贡献上，相较其他类型的因子都有较大的优势。换而言之，它对组合收益的影响，要比其他因子显著很多，而且其他因子的表现也与规模类因子会有较强联系，因而每次市场大小盘风格的切换，都会引起投资者的关注和思考。

因子的两类变化

从个股层面上来看，个股在 t 期到 t+1 期之间的超额收益的变化是由两种变化引起的：

一种变化是个股从 t 期到 t+1 期因子暴露值的变化，这种变化会引起个股暴露值在股票池中的排名发生改变，从而使个股在 t 期和 t+1 期的超额收益发生了变化，我们将这种变化称之为因子的第一类变化。

另一种变化则是市场对于因子看法的变化，或者说是投资者对于特定因子情绪上的变化，我们将这种变化称之为因子的第二类变化。

在一些传统的多因子选股策略中，一般会依据历史数据或者经济学意义锁定因子的方向，将因子规定为正向因子（因子降序）或负向因子（因子升序），在很长的时间窗口期内一直延续事先规定的因子方向进行选股。这种做法实际上是否定了因子的第二类变化。

常见的不同类型的因子都具备这两类变化，然而不同类型的因子在这两种因子上的侧重各有不同。

估值类因子有较强的经济学和会计学意义——投资者都愿意去买价值被低估的股票——因而在第二类变化上不明显，虽然从实证数据上看估值类因子同样也存在第二类变化，市场也存在不理性的时候，但是总体来说，个股在估值类因子上的收益变化主要还是来源于第一类变化。

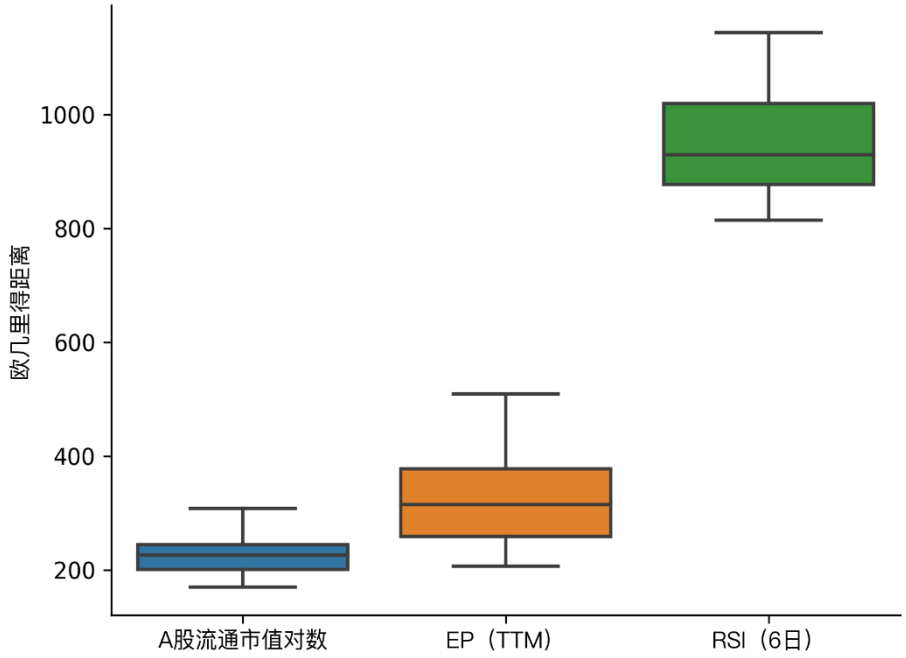
不同于估值类因子，个股在规模类因子上的收益变化基本是源于第二类变化，直观来看，股票池中个股在规模类因子上的暴露（排名）在 t 期和 t+1 期之间（较短时间内）并不

会发生重大变化，是市场上的投资者对于大小市值股票的偏好发生改变，才引起个股在规模类因子上超额收益发生改变。

为了说明不同因子在第一类变化和第二类变化之间侧重的差异，我们选择 A 股流通市值（典型规模类因子），EP(典型估值类因子)，6 日 RSI（典型技术指标类因子）进行数据测算演示，以从实证的角度来说明不同类型的因子在第一类变化上的差异。

这里我们使用最近 5 年的数据，取每月最后一交易日为横截面选取点，在截面选取点提取股票的因子暴露值的排序序列。分别计算同一个因子 t 期和 t+1 期排序序列之间的欧几里得距离，取 5 年均值进行比较。

图 1 不同类型因子第一类变化程度比较



资料来源：招商证券、Wind 资讯

相比于其他类型的因子，规模类因子在第一类变化上的程度是最低的。因而对规模类因子进行预测，当预测其第二类变化。

我们认为对于第二类变化的预测，更应该从宏观层面入手。原因主要有两个：首先，宏观经济的变化能影响投资者情绪，投资者对于市场乐观还是悲观，经常会受到对未来经济状况预测的影响。小市值的股票相对于大市值的股票更具有波动性，当投资者对未来经济预测不乐观的时候，可能会持有相对保守的大市值公司的股票。第二，从规模类风格切换的频率来看，此频率与宏观经济周期相位的切换频率更为吻合。

XGBoost 算法介绍

算法概述

XGBoost（Extreme Gradient Boosting）是一种非常有效的机器学习方法，也是时下比较前沿的机器学习算法，已经多次 Kaggle 的比赛中获得优异成绩，在机器学习领域，XGBoost 和深度学习是目前最有效的算法，在诸多领域大放异彩。本章将对 XGBoost 的原理和优势进行概述。

XGBoost 成功背后的最重要因素是其在很多应用场景下的可扩展性。我们认为它或许可以应用于对因子下期方向的判断。相较于其他机器学习算法，XGBoost 对于金融数据分析的优势主要有三点：

首先，该算法处理数据非常高效，在单台机器上的运行速度比现有流行解决方案快十倍以上，这对于处理海量的行情数据尤其重要。

再者，该算法擅长处理稀疏矩阵（稀疏数据），即存在大量缺失值的数据。我国众多宏观数据在前几年存在众多缺失，在我们处理数据的时候，发现不少较高频（周频、日频）的宏观指标直到最近几年才开始稳定公布。因而 XGBoost 的这种特性恰好可以克服早年部分宏观数据缺失的情况。

第三，XGBoost 可以学习特征之间更高级别的相互关系。经济数据和规模风格之间的关系并不是线性的，而是比线性关系更复杂的存在。这里便需要用 XGBoost 对更高级别关系进行探索和学习，往往可以更加拟合特征变量之间更高层次的关系。

XGBoost 原理简述

XGBoost 是一种树集成模型，对于集成模型来说，数据的最后结果是多个可叠加形式的树的和。对于n个样本，m个特征的数据集来说，模型会用K个可加的函数来输出最后的预测结果：

$$\hat{y}_i = \emptyset(x_i) = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}$$

其中$f(x) = w_{q(x)}$为回归树模型，这些回归树的结构和权重都是通过学习来进行确定。在整个学习的过程中，需要对如下正则化目标函数进行不断优化：

$$\mathcal{L}(\emptyset) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \qquad \text{公式（1）}$$

$$\text{其中} \quad \Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2$$

其中l是代表预测值和实际观测值之间差异的损失函数，Ω为正则项，用于对过于复杂的

模型形式实施惩罚，防止模型出现过拟合（“过拟合”一直是机器学习中重点关注的问题）。因而正则化目标函数有两个十分明确的作用，一个是使得模型的误差尽可能减小，另一个是对过拟合情况加以防范。

在实际计算中，公式（1）难以进行直接优化，因而将公式（1）转化成叠加的形式，便于优化求解。

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \qquad \text{公式（2）}$$

又对公式（2）泰勒展开，保留二次近似以便快速求解，同时剔除常数项以后得到：

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t) \qquad \text{公式（3）}$$

其中，$g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)$和$h_i = \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)$分别是损失函数中的一阶、二阶梯度统计量。对正则项$\Omega$展开后可得：

$$\begin{aligned}\tilde{\mathcal{L}}^{(t)} &= \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i \in I_j} h_i + \lambda\right) w_j^2\right] + \gamma T\end{aligned} \qquad \text{公式（4）}$$

对公式（4）（二次函数）进行最小值的求解变得相对容易。可以直接写出最优解为：

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \qquad \text{公式（5）}$$

得出的正则目标最小值为：

$$\tilde{\mathcal{L}}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \qquad \text{公式（6）}$$

公式（6）可以作为模型评估的衡量标准，也是在产生的模型当中择优的依据；一般来说，公式（6）数值越小则代表模型越好。一般情况下，对所有可能的模型进行枚举遍历是不可能的，因而一般采用贪心算法来产生模型，即考虑树模型进一步分裂带来的额外解释力是否能进一步减小损失函数的值。

除了刚才说到的正则项之外，XGBoost 还用另两种技术来提防过拟合的发生，分别是剪枝技术和列抽样技术。

以上就是 XGBoost 算法的原理和推演过程。但是在实际使用的时候，远没有那么复杂。XGBoost 有比较完善的 Python 工具包，并提供 Scikit-Learn API，方便使用者调用。

数据处理

我们认为，经济周期会影响投资者的情绪和资产配置的需求，这可能会引起投资者在大

市值股票和小市值股票之间进行投资切换，因而我们试图用 XGBoost 来探究宏观经济数据和规模风格轮动之间的关系。

模型特征变量

一些宏观数据被认为是经济周期运行的先行指标。比如 PMI（采购经理指数），制造业及非制造业 PMI 商业报告分别于每月 1 号和 3 号发布，时间上大大超前于政府其他部门的统计报告，所选的指标又具有先导性，PMI 已成为监测经济运行的及时、可靠的先行指标。

据此，我们从 wind 终端导出以下宏观数据，并配合大小盘指数的技术指标，作为 XGBoost 模型的特征变量。对于机器学习来说，样本量是决定模型是否有效的关键，月频数据从数据量上来说并不能达到模型数据量的要求。而宏观数据中日频数据类型十分有限，而且日频因子方向存在很大的随机性，并不能反映市场的真实情况。综合考虑，我们优先按周频来提取数据，有些宏观数据只有月频的，我们将月频数据按周平均来代表其周频数据。

表 1：宏观变量数据表 1（周频数据）

特征变量名称	特征变量代码	特征变量名称	特征变量代码
日均耗煤量:6 大发电集团:合计	CCD_6P	中债企业债到期收益率(AAA):5 年	Bond_AAA_5yrs
日均产量:粗钢:国内	OUTPUT_STEEL	中债企业债到期收益率(AAA):10 年	Bond_AAA_10yrs
市场价:螺纹钢:HRB400 Φ16-25mm:全国	PRICE_REBAR	中债企业债到期收益率(AA):1 年	Bond_AA_1yrs
开工率:焦化企业(100 家):产能>200 万吨	COKEENTERP_OR	中债企业债到期收益率(AA):3 年	Bond_AA_3yrs
开工率:汽车轮胎:半钢胎	SEMISTEEL_OR	中债企业债到期收益率(AA):5 年	Bond_AA_5yrs
当周日均销量:乘用车:厂家零售	PCV_DAILY_SELLS	中债企业债到期收益率(AA):10 年	Bond_AA_10yrs
30 大中城市:商品房成交面积	RESI_AREA(30CITIES)	同业存单:发行利率:3 个月	IBNCD_3m
100 大中城市:成交土地占地面积:当周值	LAND_AREA(100CITIES)	公开市场操作:货币净投放	ISSUE_MONEY

特征变量名称	特征变量代码	特征变量名称	特征变量代码
农产品批发价格 200 指数	ARGI_PRICE_200_INDEX	在岸人民币日均成交量(百万美元)	VOL_USD_CNY(M)
OPEC:一揽子原油价格	OPEC	人民币兑美元汇率	USD_CNY
期货收盘价(活跃合约):螺纹钢	FUC_RB	人民币 CFETS 汇率指数	CFETS_IND
期货收盘价(活跃合约):阴极铜	FUC_NEGA_CU	信用债净发行量 AAA	ISSSUE_DEBENTURE_AAA
期货收盘价(活跃合约):动力煤	FUC_PC	信用债净发行量 AA+	ISSSUE_DEBENTURE_AA+
波罗的海干散货指数(BDI)	BDI	信用债净发行量 AA	ISSSUE_DEBENTURE_AA
中债企业债到期收益率(AAA):1 年	Bond_AAA_1yr	信用债净发行量 AA 以下或无评级	ISSSUE_DEBENTURE_OTHER
中债企业债到期收益率(AAA):3 年	Bond_AAA_3yrs	同业存单净发行量	IBNCD

资料来源：招商证券、Wind 资讯

表 2：宏观变量数据表 2（月频数据）

特征变量名称	特征变量代码	特征变量名称	特征变量代码
社会融资规模:当月值	SOCAL_FINANCE	出口金额:当月值	EX
商品房销售面积:累计同比	RE_ARAEA_CUM_YOY	产量:发电量:当月同比	POWER_YOY
商品房销售额:累计同比	RE_AMT_CUM_YOY	产量:粗钢:当月值	RAW_STEEL
贸易差额:当月值	DIFF_EX_IM	PPI:生产资料:环比	PPI_MOM
贸易差额:当月同比	DIFF_EX_IM_YOY	PPI:生产资料:当月同比	PPI_YOY

特征变量名称	特征变量代码	特征变量名称	特征变量代码
进口金额:季调:环比	IMP_MOM	PPI:全部工业品:环比	PPI_IND_MOM
进口金额:季调:当月同比	IMP_YOY	PPI:全部工业品:当月同比	PPI_IND_YOY
进口金额:当月值	IMP	PMI:主要原材料购进价格	PPI_RAW
进出口金额:季调:环比	EX_IM_MOM	PMI:新订单	PPI_NEW_ORDER
进出口金额:季调:当月同比	EX_IM_YOY	PMI:生产	PPI_PROD
金融机构:新增人民币贷款:居民户:当月值	NEW_RES_LOAN	PMI:产成品库存	PMI_INV
金融机构:新增人民币贷款:非金融性公司及其他部门:当月值	NEW_NF_LOAN	PMI	PMI
金融机构:新增人民币贷款:当月值	NEW_LOAN	M2:同比	M2_YOY
金融机构:各项贷款余额:同比	OTHER_LOAN_YOY	M1:同比	m2_YOY
固定资产投资完成额:累计同比	FA_AMT_CUM_YOY	CPI:食品:环比	CPI_FOOD
公共财政支出:当月值	FISCAL_PMT	CPI:食品:当月同比	CPI_FOOD_YOY
公共财政收入:当月值	FISCAL_INCOME	CPI:环比	CPI_MOM
工业增加值:当月同比	IND_VAL_ADDED_YOY	CPI:非食品:环比	CPI_NON_FOOD_MOM
房地产开发投资完成额:累计同比	RE_INV_CUM_YOY	CPI:非食品:当月同比	CPI_NON_FOOD_YOY
出口金额:季调:环比	EX_MOM	CPI:当月同比	CPI_YOY

特征变量名称	特征变量代码	特征变量名称	特征变量代码
出口金额:季调:当月同比	EX_YOY		

资料来源：招商证券、Wind 资讯模型标签

XGBoost 是监督学习的一种，因而输入的数据是标签数据，这里的标签我们定为 A 股流通市值对数因子下一期的收益方向。本文需要用 XGBoost 算法进行 logistic 分类预测（逻辑分类预测），这里需要先对 A 股流通市值对数因子下一期的收益进行二值化处理：若下一期该因子的收益为正，则取 1，否则取 0。

关于因子收益的计算，我们在系列报告的前几篇已经进行详细的介绍，并在我们的周报中有持续跟踪汇报 16 个重要因子的每周收益。大致做法是以个股下一期的超额收益 $\Delta r^{B,M}_{i,[t_0,t_1]}$ 为被解释变量，$\Delta r^{B,M}_{i,[t_0,t_1]} = \Delta r^{B}_{i,[t_0,t_1]} - \Delta \tilde{\beta}^{M}_{i,t\le t_0} \cdot r_{M,[t_0,t_1]}$，以因子在 t_0 时刻的超额暴露的标准化赋值为解释变量，进行 WLS 回归计算，得到的待估参数即为因子的收益（具体做法详见因子系列报告（1）至因子报告（10））。

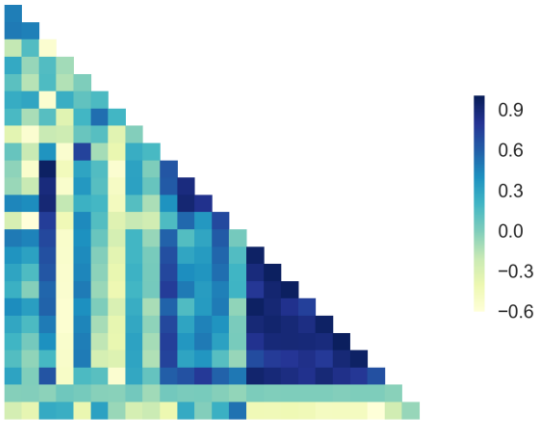
为了适合机器学习所需要的数据量和频率，我们用周频的因子收益的二值标签作为机器学习的数据标签。

数据预处理

我们收集的宏观数据中，有些特征变量之间会存在较强的相关性，若特征变量之间的相关性太强，意味着存在信息冗余，这种情况下，我们可以去掉部分特征变量。

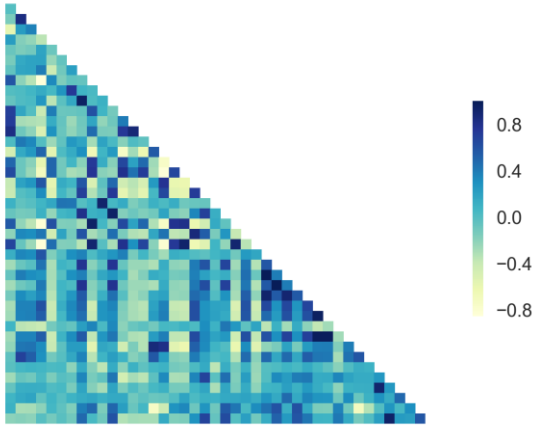
我们分别对周频、月频宏观特征变量组内两两计算相关系数，绘制成热力图，展示如下：

图 2 周频特征变量相关系数热力图



资料来源：招商证券、Wind 资讯

图 3 月频特征变量相关系数热力图



资料来源：招商证券、Wind 资讯

不难看出，部分特征变量之间存在很强的相关性，存在信息冗余。我们剔除了信息重叠度较高的变量，留下彼此相关性低且具有区分度特征变量。用于训练模型的周频特征变量 14 个、月频特征变量 23 个，列示如下：

表 3：入选周频宏观特征变量

特征变量名称	特征变量代码	特征变量名称	特征变量代码
日均耗煤量:6 大发电集团:合计	CCD_6P	100 大中城市:成交土地占地面积:当周值	LAND_AREA(100CITIES)
日均产量:粗钢:国内	OUTPUT_STEEL	农产品批发价格 200 指数	ARGI_PRICE_200_INDEX
市场价:螺纹钢:HRB400 Φ16-25mm:全国	PRICE_REBAR	OPEC:一揽子原油价格	OPEC
开工率:焦化企业(100 家):产能>200 万吨	COKEENTERP_OR	波罗的海干散货指数(BDI)	BDI
开工率:汽车轮胎:半钢胎	SEMISTEEL_OR	中债企业债到期收益率(AAA):5 年	Bond_AAA_5yrs
当周日均销量:乘用车:厂家零售	PCV_DAILY_SELLS	公开市场操作:货币净投放	ISSUE_MONEY
30 大中城市:商品房成交面积	RESI_AREA(30CITIES)	人民币兑美元中间价	USD_CNY

资料来源：招商证券、Wind 资讯

表 4：入选月频宏观特征变量

特征变量名称	特征变量代码	特征变量名称	特征变量代码
社会融资规模:当月值	SOCAL_FINANCE	产量:发电量:当月同比	POWER_YOY
商品房销售额:累计同比	RE_AMT_CUM_YOY	PPI:全部工业品:环比	PPI_IND_MOM
期末汇率:美元兑人民币	USD_CNY	PPI:全部工业品:当月同比	PPI_IND_YOY
贸易差额:当月值	DIFF_EX_IM	PMI:产成品库存	PMI_INV
贸易差额:当月同比	DIFF_EX_IM_YOY	PMI	PMI
进口金额:季调:环比	IMP_MOM	M2:同比	M2_YOY
金融机构:新增人民币贷款:居民户:当月值	NEW_RES_LOAN	M1:同比	m2_YOY
金融机构:新增人民币贷款:非金融性公司及其他部门:当月值	NEW_NF_LOAN	CPI:环比	CPI_MOM
公共财政支出:当月值	FISCAL_PMT	CPI:非食品:环比	CPI_NON_FOOD_MOM

特征变量名称	特征变量代码	特征变量名称	特征变量代码
公共财政收入:当月值	FISCAL_INCOME	CPI:非食品:当月同比	CPI_NON_FOOD_YOY
房地产开发投资完成额:累计同比	RE_INV_CUM_YOY	CPI:当月同比	CPI_YOY
出口金额:季调:环比	EX_MOM		

资料来源：招商证券、Wind 资讯

我们参详了国外的一些关于机器学习在金融领域的应用文献，国外的文献更偏好于用技术指标来对个股未来的收益进行预测。其中一个重要的原因可能是行情数据便于获取而且能提供充足的样本量。借鉴于此，我们也引入了一些沪深 300 和中证 1000 的技术指标分别代表大盘股和小盘股的走势。示列如下：

表 5：技术指标特征变量表

特征变量名称	特征变量代码	特征变量名称	特征变量代码
随机 K 沪深 300	STOCHASTIC_K_300	随机 K 中证 1000	STOCHASTIC_K_1000
随机 D 沪深 300	STOCHASTIC_D_300	随机 D 中证 1000	STOCHASTIC_D_1000
ROC 沪深 300	ROC_300	ROC 中证 1000	ROC_1000
威廉 R 沪深 300	LWR_300	威廉 R 中证 1000	LWR_1000
5 日差异 沪深 300	DISPARITY5_300	5 日差异 中证 1000	DISPARITY5_1000
10 日差异 沪深 300	DISPARITY10_300	10 日差异 中证 1000	DISPARITY10_1000
CCI 沪深 300	CCI_300	CCI 中证 1000	CCI_1000
RSI 沪深 300	RSI_300	RSI 中证 1000	RSI_1000

资料来源：招商证券、Wind 资讯

其他数据处理细节如下：

1. 由于在过去 10 年中，我国的经济发展比较迅速，一些绝对量的特征变量数据一直处于递增的过程，因此在时间方向上会存在异方差性。机器学习中的交叉验证（Cross Validation）会打乱特征变量的次序，若在时间维度上存在异方差，会对模型产生误导，因而我们又提取了 2007 年至今的全国 GDP 季度数据，用绝对量的特征变量除以对应时间的全国 GDP 数据，实现 GDP 中性，以减轻异方差的影响。
2. XGBoost 算法在进行树分裂的时候，依据的是数据的排序，因而是否对输入的特征变量数据进行归一化基本没有影响。我们直接导入数据原值，不对其进行归一化操作。
3. 输入数据是 2007 年第 15 周数据到 2018 年第 27 周的数据，共 576 个样本。特征变量为每周最后一个交易日能获取到的最新的变量数据（月频数据并非每周更新），标签为下一个交易周的因子收益方向，0 或 1。

4. 不少宏观数据在最近几年才陆续稳定公布，在训练集和测试集的切分上，若使用前8年为训练集后2年为测试集的切分方法并不妥当，这种切分方式会使得近期才公布的特征变量得不到有效训练。为了克服上述情况，我们将样本按时间顺序，用1到5循环编号，然后将编号为1的分成一组，编号为2的分成一组……以此类推，分成5组，然后从中随机抽取一组作为测试集，剩下四组的作为训练集。用训练集来训练模型，而后用测试集数据对模型进行最终评估。

模型参数与结果展示

我们采用 XGBoost 的 Python 包，辅之以 XGBoost 的 scikit-learn 接口作为模型搭建、训练和测试的工具。XGBoost 模型在训练前有众多的参数需要调试，在训练的过程中需要不停调参。其中最主要的几个参数（超参数）罗列如下：

表 6：XGBoost 主要参数表

参数名称	释义
max_depth	**树的最大深度。树越深通常模型越复杂，更容易过拟合。默认值：3**
learning_rate	学习率或收缩因子。学习率和迭代次数 / 弱分类器数目 n_estimators 相关。 默认值: 0.1
n_estimators	**弱分类器数目。默认值:100**
gamma	节点分裂所需的最小损失函数下降值。默认值：0
min_child_weight	叶子结点需要的最小样本权重和。默认值：1
subsample	**构造每棵树的所用样本比例（样本采样比例），同 GBM。默认值：1**
colsample_bytree	构造每棵树的所用特征比例。默认值：1
colsample_bylevel	树在每层每个分裂的所用特征比例。默认值：1
reg_alpha	L1/L0 正则的惩罚系数。默认值：0
reg_lambda	L2 正则的惩罚系数。默认值：1

资料来源：招商证券、Wind 资讯

经过我们的调试比较，对训练结果影响比较大的参数是 n_estimators，max_depth 和 subsample，其中 n_estimators 我们用 early stopping 方法确认，max_depth 和 subsample 用网格搜索（GridSearchCV）的方式调优，其他参数我们按默认值处理。

由于在训练集上，通过调整参数设置使估计器的性能达到了最佳状态；但在测试集上可能会出现过拟合的情况。此时，测试集上的信息反馈足以颠覆训练好的模型，评估的指标不再有效反映出模型的泛化性能。为了解决此类问题，还应该准备另一部分被称为“validation set(验证集)”的数据集，模型训练完成以后在验证集上对模型进行评估。

然而，通过将原始数据分为 n 个数据集合，我们就大大减少了可用于模型学习的样本数量，并且得到的结果依赖于集合对(训练,验证)的随机选择。这里我们用交叉验证(CV)来解决这个问题。

最基本的交叉验证方法被称之为，k-折交叉验证。k-折交叉验证将训练集划分为 k 个较小的集合（其他方法会在下面描述，主要原则基本相同）。每一个 k 折都会遵循下面的过程：

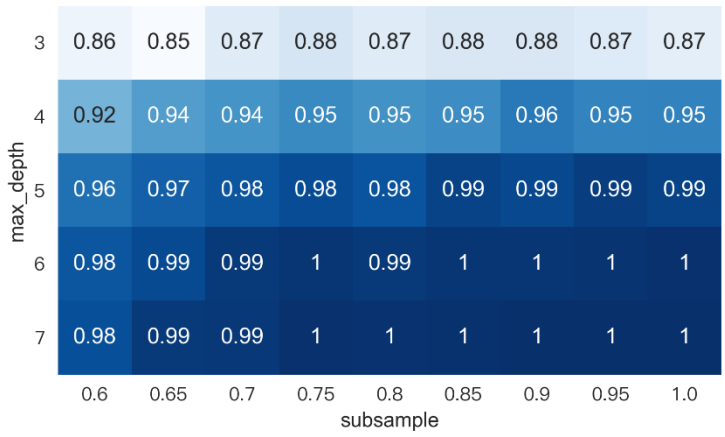
1. 将 k-1 份训练集子集作为训练集训练模型；
2. 将剩余的 1 份训练集子集作为验证集用于模型验证(也就是利用该数据集计算模型

的性能指标，例如准确率）。

3. k-折交叉验证得出的性能指标是循环计算中每个值的平均值。该方法虽然计算代价很高，但是它不会浪费太多的数据（如固定任意测试集的情况一样），在处理样本数据集较少的问题（例如，逆向推理）时比较有优势。在交叉验证的过程中，我们将 k 设为 5。

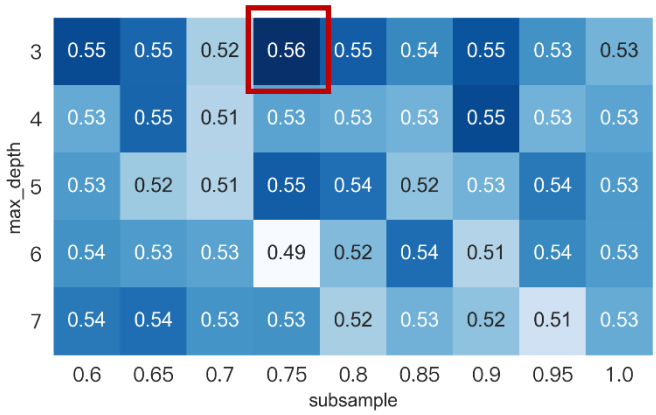
将训练集数据输入模型，用网格搜索（GridSearch）的方法来进行 max_depth 和 subsample 的参数调整，并用交叉验证的方式汇报模型的准确性，以评估模型，结果如下：

图 4 训练集 GridSearchCV 参数调优结果



资料来源：招商证券、Wind 资讯

图 5 验证集 GridSearchCV 参数调优结果



资料来源：招商证券、Wind 资讯

上图中，纵轴是 max_depth 代表 XGBoost 树的深度，横轴为 subsample 用于构造每棵树的所用样本比例，热力图中每个像素内的数字代表的是 k-折交叉验证得出的 5 种抽样组合准确性的平均值。

在训练集下，随着树深度和用于构造每棵树的所用样本比例增加，分类的结果准确率可以达到 100%，然而在验证集中，准确率并未随着模型复杂性的增加而增加。引入交叉验证可以避免盲目选用训练集训练出来的过拟合模型。

XGBoost 模型在训练集下产生过拟合，一个重要的原因可能是由于样本量过少。稍复杂的模型就能完全拟合样本标签。

选用验证集数据中准确性最好的参数设置，即 max_depth=3，subsample=0.75，在测试集上进行最后的评估。

预测结果

表 7：测试集评估结果(仅使用宏观特征变量)

XGBoost 预测准确性	动量预测方法准确性	提升比例
0.640	0.552	15.9%

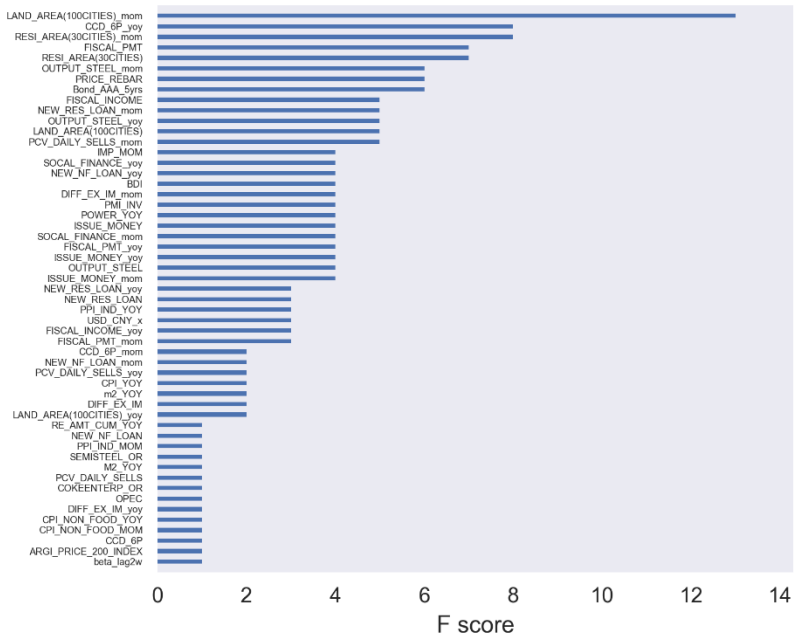
资料来源：招商证券、Wind 资讯

若我们只用宏观变量作为特征变量来训练，准确率为 64%。由于我们使用的是周频预测，因子周方向切换比之月方向会更加频繁，因而对周的预测也会更困难。若直接沿用因子上周的方向作为本周方向的预测值做法（动量预测法）的准确率是 55.2%，而使用 XGBoost 方法可以将准确率提高到 64%，提升了 15.9%。

另外，由于 XGBoost 是白盒算法，在分类的过程中，可以调出各特征变量在分类中起到的作用。此处我们汇报各特征变量的 F 得分(对应特征变量作为节点被拆分的总次数)，

如下：

图 6 宏观特征变量区分能力比较图



资料来源：招商证券、Wind 资讯

宏观特征变量里面排名前五的特征变量是：100 大中城市:成交土地占（环比）、日均耗煤量:6 大发电集团（同比）、30 大中城市:商品房成交面积（环比）、公共财政支出:当月值、30 大中城市:商品房成交面积。

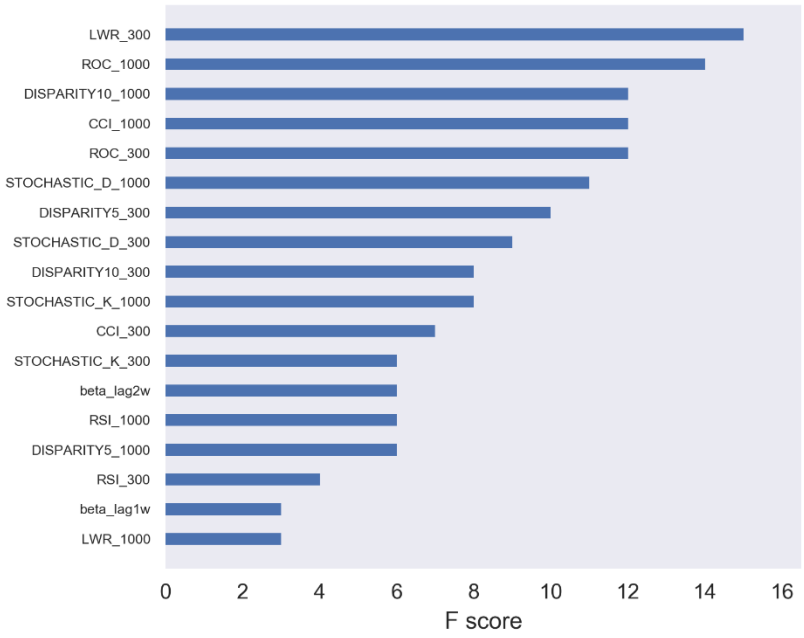
表 8：测试集评估结果(仅使用技术指标)

XGBoost 预测准确性	动量预测方法准确性	提升比例
0.640	0.552	15.9%

资料来源：招商证券、Wind 资讯

若我们只用技术指标作为特征变量来训练，比较凑巧，准确率也是 64%。若直接沿用因子上周的方向作为本周方向的预测值做法（动量预测法）的准确率是 55.2%，而使用 XGBoost 方法可以将准确率提高到 64%，也提升了 15.9%。

图 7 技术指标特征变量区分能力比较图



资料来源：招商证券、Wind 资讯

技术指标特征变量里面排名前五的特征变量是：沪深 300 威廉值、 中证 1000 的 ROC 值、沪深 1000 的 10 日差异、中证 1000 的 CCI 和沪深 300 的 ROC 值。

特征变量重要性图给了我们调整变量的依据。一般而言，将区分能力弱的特征变量剔除，不会对模型的预测能力造成太大影响。因而我们挑选宏观特征变量和技术指标特征变量中区分能力排名靠前的特征变量共 19 个。重新进行模型训练，并在测试集中进行最后评估。

表 9：高区分度特征变量表（加被解释变量滞后项）

特征变量代码	特征变量代码	特征变量代码	特征变量代码
Bond_AAA_5yrs	OUTPUT_STEEL_mom	FISCAL_PMT_yoy	DISPARITY5_1000
LAND_AREA(100CITIES)_mom	ISSUE_MONEY_yoy	CCI_1000	CCI_300
RESI_AREA(30CITIES)_yoy	NEW_NF_LOAN_yoy	DISPARITY10_1000	ROC_300
OUTPUT_STEEL_yoy	FISCAL_INCOME	STOCHASTIC_D_1000	**Beta_lag1w**
PRICE_REBAR	POWER_YOY	LWR_300	**Beta_lag2w**

资料来源：招商证券

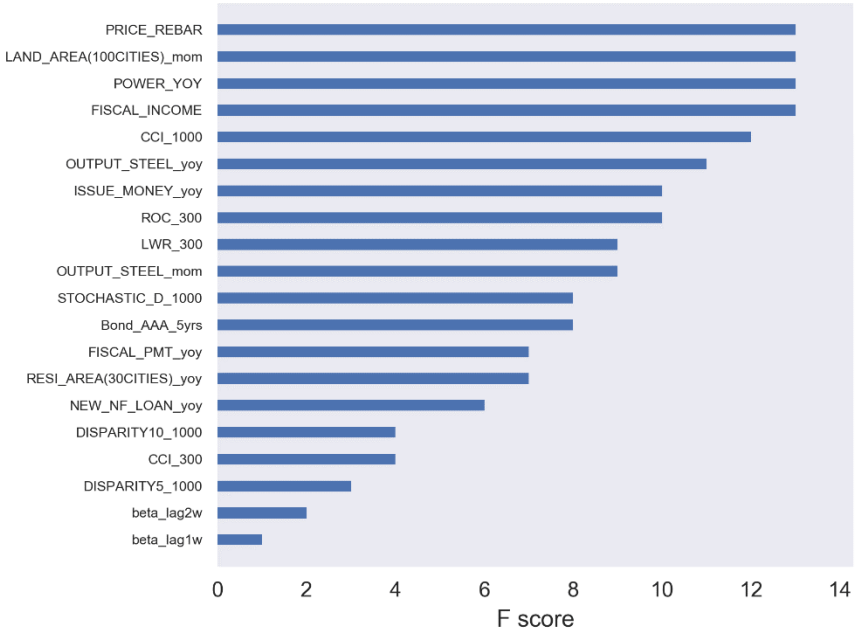
表 9：测试集最终评估结果

XGBoost 预测准确性	动量预测方法准确性	预测能力提升
0.658	0.552	19.20%

资料来源：招商证券、Wind 资讯

经过特征变量筛选后，模型的预测能力又有所提升，达到 65.8%，较基准预测方法预测的准确性提升了近 20%。以下给出特征变量重要性排序图：

图 8 高区分力特征变量重要性排序图



资料来源：招商证券、Wind 资讯

排名前四的分别是：螺纹钢市场价、：100 大中城市:成交土地占（环比）、产量:发电量:当月同比、公共财政收入当月值。前四均为宏观特征变量，从实证数据说明，宏观变量在预测规模类因子收益方向上，还是有较强的区分力的。而两个被解释变量滞后项区分能力最弱。

结论

在构建因子模型时，我们并未锁死因子的第二类变化，将所有的因子均视为风险因子，它们的方向均会随着市场情绪的变化而发生改变。不同类型的因子都具备第一类变化（因子暴露度排序的变化）和第二类变化（投资者对该因子情绪的变化），但不同的因子在这两类变化上侧重是不同的。

对于规模类因子的预测，应当着重关注其第二类变化；而对于第二类变化的预测须从宏观数据入手更为合情合理。我们挑选众多宏观特征变量和技术指标特征变量，根据变量之间的相关系数水平进行适当筛选。将特征变量放入模型去预测规模类因子收益的方向，使用了一些常用的手段去防止模型出现过拟合。最后得到 65.8%的逻辑预测准确率。

报告的最后花不多的篇幅来聊一聊我们对机器学习在金融领域应用上的两点看法，这也是最近在与投资者交流过程中，最常被投资者提及的两个问题。

第一点是机器学习或者人工智能是否是一个全新的领域。我们认为不是。目前各大量化产品使用成熟的因子模型，其本质是线性回归模型，而线性回归本就是机器学习中最常

用的算法之一。与因子模型一样，机器学习的目的也是探索自变量和因变量之间真实存在的关系。只不过线性模型在拟合的过程中，限定了模型的线性形式，用观测到的样本对模型参数进行估计；而机器学习既没有限定模型的具体形式也没有限定模型的待估参数，用样本来拟合模型形式和待估参数。因而因子模型是机器学习的一个子集。

第二点是关于机器学习的黑箱性质。相比于因子模型，机器学习输出的中间过程往往难以被人的思维所理解，这也是许多投资者不敢轻易尝试机器学习的最重要的原因。我们并不否认这种担忧，也支持投资过程中的谨慎。AlphaZero 在进行日本将棋对弈时会把王将移动到棋盘的中心，这违反了日本将棋的理论（从人类的视角），但是它仍能掌握局势，锁定胜局。也正是因为一次次的胜局让人类觉得可能我们自身的日本将棋理论还存在其他的可能性。我们相信，在未来金融领域，若有机器学习策略管理的产品取得了若干成绩，投资者的这种担忧也会日渐消弭，而用更为包容的视角来看待这个问题。

分析师承诺

负责本研究报告的每一位证券分析师，在此申明，本报告清晰、准确地反映了分析师本人的研究观点。本人薪酬的任何部分过去不曾与、现在不与，未来也将不会与本报告中的具体推荐或观点直接或间接相关。

叶涛：首席分析师。上海交通大学管理学硕士，2005 年起从事金融工程研究，曾先后任职于易方达基金机构投资部、上投摩根基金研究部、申万菱信基金投资管理总部、长江证券研究部、广发证券发展研究中心，2014 年 3 月加盟招商证券研究发展中心。

崔浩瀚：研究助理。浙江大学经济学硕士，2017 年 7 月加盟招商证券研究发展中心金融工程组。

投资评级定义

公司短期评级

以报告日起 6 个月内，公司股价相对同期市场基准（沪深 300 指数）的表现为标准：

强烈推荐：公司股价涨幅超基准指数 20%以上
审慎推荐：公司股价涨幅超基准指数 5-20%之间
中性：公司股价变动幅度相对基准指数介于±5%之间
回避：公司股价表现弱于基准指数 5%以上

公司长期评级

A：公司长期竞争力高于行业平均水平
B：公司长期竞争力与行业平均水平一致
C：公司长期竞争力低于行业平均水平

行业投资评级

以报告日起 6 个月内，行业指数相对于同期市场基准（沪深 300 指数）的表现为标准：

推荐：行业基本面向好，行业指数将跑赢基准指数
中性：行业基本面稳定，行业指数跟随基准指数
回避：行业基本面向淡，行业指数将跑输基准指数

重要声明

本报告由招商证券股份有限公司（以下简称“本公司”）编制。本公司具有中国证监会许可的证券投资咨询业务资格。本报告基于合法取得的信息，但本公司对这些信息的准确性和完整性不作任何保证。本报告所包含的分析基于各种假设，不同假设可能导致分析结果出现重大不同。报告中的内容和意见仅供参考，并不构成对所述证券买卖的出价，在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议。除法律或规则规定必须承担的责任外，本公司及其雇员不对使用本报告及其内容所引发的任何直接或间接损失负任何责任。本公司或关联机构可能会持有报告中所提到的公司所发行的证券头寸并进行交易，还可能为这些公司提供或争取提供投资银行业务服务。客户应当考虑到本公司可能存在可能影响本报告客观性的利益冲突。