



华泰期货
HUATAI FUTURES

期货研究报告|量化专题报告 2024-08-06

基于多因子体系的基差预测模型

研究院 量化组

研究员

高天越

☎ 0755-23887993

✉ gaotianyue@htfc.com

从业资格号: F3055799

投资咨询号: Z0016156

联系人

李光庭

☎ 0755-23887993

✉ liguangting@htfc.com

从业资格号: F03108562

李逸资

☎ 0755-23887993

✉ liyizi@htfc.com

从业资格号: F03105861

黄煦然

☎ 0755-23887993

✉ huangxuran@htfc.com

从业资格号: F03130959

麦锐聪

☎ 0755-23887993

✉ mairuicong@htfc.com

从业资格号: F03130381

投资咨询业务资格:

证监许可【2011】1289号

摘要

本篇报告在《华泰期货量化策略专题报告 20240712: 转融通暂停影响简述》基础上展开对股指期货年化基差率相关因子的量化分析。首先用 Pearson 相关系数以及 Distance 相关系数来衡量因子的线性与非线性关系, 探讨因子在不同预测周期下的表现, 并对因子进行筛选; 其次构建了基于线性回归模型 OLS+Ridge 以及非线性的机器学习模型 Random Forest 和 Xgboost 的年化基差率预测模型, 展示了模型在不同预测周期下的预测效果。在预测周度年化基差率时, Xgboost 对目标变量的预测精度表现较好, MSE 平均 0.044%, 涨跌准确率平均 57.70%, 最高达 62.13%。

核心观点

- 1) 公募指增超额和指数正负波动率类因子对年化基差率影响较大。
- 2) 公募指增超额类因子随预测周期变长相关性下降幅度较大, 管理人的超额能力对短期内的基差水平影响较大。指数相关因子则无明显下降趋势, 现货市场的波动对未来基差走势影响的延续性较强。
- 3) 预测周度年化基差率时, Xgboost 对目标变量的预测精度方面表现较好。

目录

摘要.....

核心观点.....

基于多因子体系的基差预测模型.....

■ 股指期货基差.....

■ 因子相关性.....

■ 多因子模型构建.....

■ 模型预测效果展示.....

■ 总结.....

■ 风险提示.....

1

1

4

4

4

9

11

15

15

图表

图 1：IF 因子有效性与预测周期 | 单位：无.....7

图 2：IH 因子有效性与预测周期 | 单位：无.....7

图 3：IC 因子有效性与预测周期 | 单位：无.....7

图 4：IM 因子有效性与预测周期 | 单位：无.....7

图 5：每个预测日 T 的训练集与测试集示意图 | 单位：无.....9

图 6：模型建立与预测流程图 | 单位：无.....10

图 7：不同预测周期的准确性—OLS+RIDGE | 单位：无.....11

图 8：不同预测周期的准确性—RANDOM FOREST | 单位：无.....11

图 9：不同预测周期的准确性—XGBOOST | 单位：无.....11

表 1：不同预测周期下因子线性相关统计|单位：无.....5

表 2：不同预测周期下因子非线性相关统计|单位：无.....8

表 3：模型训练及预测效果.....12

表 4：T+5 年化基差率模型训练及预测效果 | 单位：无.....12

表 5：T+5 年化基差率涨跌方向预测效果—XGBOOST | 单位：无.....13

表 6：T+5 年化基差率入选因子前 10—XGBOOST | 单位：无.....14

基于多因子体系的基差预测模型

■ 股指期货基差

基差是股指期货研究中重点关注的指标之一，是许多对冲、套利策略的构建基础。因此，对基差的预测与判断具有重要的意义。构建一个基差预测模型首先需要结合主观的逻辑，找出与基差相关的因素并用合适的因子去量化这个影响因素，并利用数学模型作为工具，从数据的层面验证它们的相关性，再将有效的因子通过不同方式组合成为预测模型，最后选择适合的标度去衡量模型预测的准确性。在《华泰期货量化策略专题报告 20240712：转融通暂停影响简述》中，我们介绍了期现市场中包括融券在内的影响股指期货基差的不同因素与代表因子，这篇我们将介绍因子的筛选以及模型的构建。

■ 因子相关性

我们选取了两个寻找因子相关性的度量。一个是 Pearson 相关系数，用于衡量因子的线性关系，一个是 Distance 相关系数，由 Gábor J. Székely 于 2005 年第一次提出，用于衡量因子的非线性关系。Pearson 相关系数通过评估两个变量在各自均值距离上的协变趋势来捕捉变量间的线性关系，Distance 相关系数则评估它们与其它所有点之间距离的协变趋势，从而捕捉变量间除线性关系之外的依赖关系。因此，Distance 相关系数的包容性比 Pearson 相关系数更强，筛选因子时会更多的因子判定为有效。另外，与 Pearson 相关系数不同的是，Distance 相关系数只能提供相关性的强弱。它的取值范围为 $[0, 1]$ ，越接近 1 则相关性越强，但无法提供相关性的正负方向。

我们将模型的预测目标变量 Y 定为 $t+n$, $n \in [1, 60]$ 的下季连续合约的年化基差率（经过分红调整）。当 n 取不同值时，我们分别计算 Y_{t+n} 与 t 时因子 X_t 的相关系数，旨在捕捉不同预测周期下因子可能存在的相关性差异，观察模型的预测能力是否和预测周期存在一定关系。

为找到长期有效的因子，我们的数据全样本取 2017 年至今，用 2023 年以前的数据计算 Pearson 和 Distance 相关系数，初步筛选出一部分有效的因子。所有因子已经过滞后处理，确保在预测日 t 可以获取；取值范围较大的因子已经过 z-score 标准化处理。

我们先来看看因子与目标变量在不同预测周期下整体的线性相关性统计。

表 1：不同预测周期下因子线性相关统计|单位：无

标的	因子	保留次数	Pearson corr 平均值
IF	公募指增 60 日累计超额	60	-0.495
	指数前一日收盘价	60	-0.402
	指数 90 日负向波动率	60	-0.249
	指数成分股融券余额	60	-0.237
	11 月哑变量	58	0.230
	6 月哑变量	38	-0.212
	融券对冲需求比	60	-0.201
	公募指增 10 日累计超额极端涨幅	53	-0.185
	公募指增 10 日累计超额极端跌幅	60	0.174
	指数 90 日累计收益率	45	-0.173
IH	指数前一日收盘价	60	-0.544
	融券对冲需求比	60	-0.457
	指数成分股融券余额	60	-0.428
	公募指增 60 日累计超额	60	-0.318
	指数 60 日负向波动率	51	-0.226
	公募指增 10 日累计超额极端涨幅	47	-0.216
	期货合约沉淀资金	60	-0.202
	期货合约总持仓量	60	-0.182
	多空力量	60	0.180
	指数 90 日累计收益率	38	-0.178
IC	多空力量	60	0.464
	公募指增 60 日累计超额	60	-0.365
	指数 90 日累计收益率	60	-0.355
	指数 90 日正向波动率	60	-0.300
	指数 90 日负向波动率	60	-0.267
	期货合约沉淀资金	60	0.233
	11 月哑变量	44	0.223
	指数成分股融券余额	55	0.200
	期货多头力量	60	0.182
	7 月哑变量	56	-0.180
IM	期货合约总持仓量	60	0.691
	期货合约沉淀资金	60	0.691
	融券对冲需求比	60	-0.684
	公募指增 90 日累计超额	60	-0.673
	指数 90 日负向波动率	60	-0.663
	期货多头力量	60	0.640
	指数 60 日累计收益率	60	-0.637
	指数 30 日正向波动率	60	0.588
	指数前一日收盘价	60	-0.557
	8 月哑变量	60	-0.519

数据来源：同花顺 华泰期货研究院

为了在初步筛选时留下更多的因子，我们选择留下 Pearson 相关系数绝对值大于 0.1，同时 p-value 小于 0.1 的因子。我们按照不同预测周期下的 Pearson 相关系数平均值以及在不同预测周期中因子被保留的次数对因子进行排序。相关系数的绝对值越大，说明因子与目标变量的相关性越强，因子被保留的次数越多，说明因子在预测周期变长时有效性的延续性越强。我们在每一类因子中选取相关性最强且具有代表性的因子在上表进行展示。从表中的结果我们可以看到，不同标的下，相关性强的因子有一定差异，但有部分因子展示了它们的普适性，如公募指增超额、指数负向波动率、指数累计收益率、指数收盘价和月份哑变量。

在不同时间窗口下，60 日或 90 日的公募指增超额与年化基差率的相关性最强，且呈负相关。这验证了我们前期的推测，对冲需求随着超额的增大而上升时，体现在基差上的反应则是贴水扩大，且传导到基差上的反应需要一定时间，时间窗口越短，与基差的相关性普遍更弱。

指数 90 日负向波动率与年化基差率呈负相关，当指数的负向波动率增大时，市场的做空情绪可能会相应增加，反应到基差上则是贴水扩大。而正向波动率的相关性从数据结果来看更不稳定，IC 的年化基差率和指数正向波动率呈负相关，IF 则呈弱正相关（未被入选在表中展示）。我们可以推测，市场认为中证 500 的正向波动比起沪深 300 来说更不可持续，反转效应较强，所以在正向波动大的时候会带来一部分做空力量入场，使得贴水扩大。反转与动量效应也与选择的时间窗口有关，拿 IM 举例，5/10/20/30/60 日的正向波动率都与年化基差率呈正相关，而 90 日的正向波动率则呈负相关。

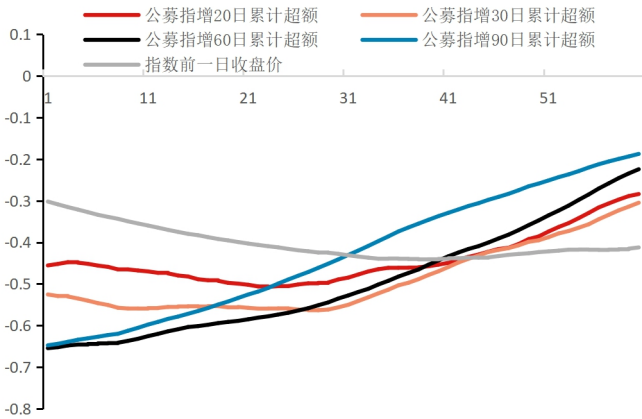
指数 60/90 日累计收益率和收盘价与年化基差率呈负相关，而在更短的时间窗口下相关性则可能呈相反的情况。现货市场短时间内的涨幅会对基差造成正向影响，而长时间的涨幅更可能引发反转效应，使得做空力量增加，从而对基差造成负向影响。

从月份哑变量的结果来看，经过分红调整后的年化基差率仍存在一定季节性，分红高峰期的 6/7 月相关性普遍为负，10/11 月相关性为正。

（以上提到的部分因子因相关系数较小未被入选在表中展示）

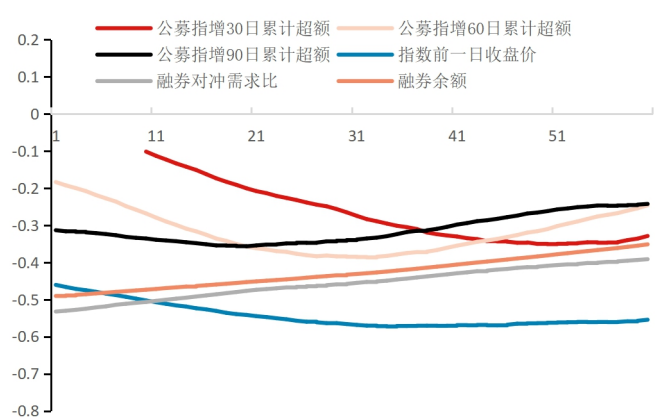
其次我们可以观察因子在不同预测周期下的具体表现。

图 1: IF 因子有效性与预测周期 | 单位: 无



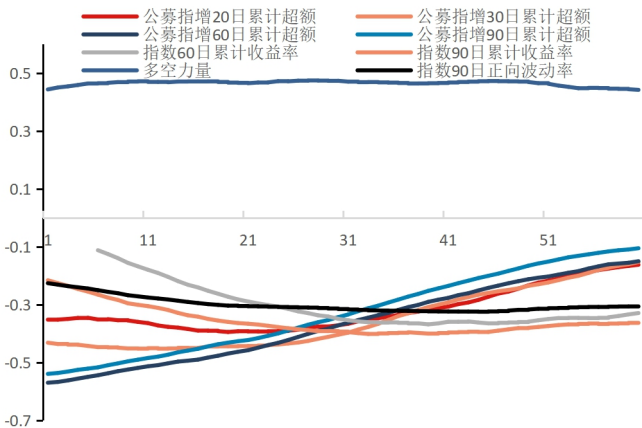
数据来源: 同花顺 华泰期货研究院

图 2: IH 因子有效性与预测周期 | 单位: 无



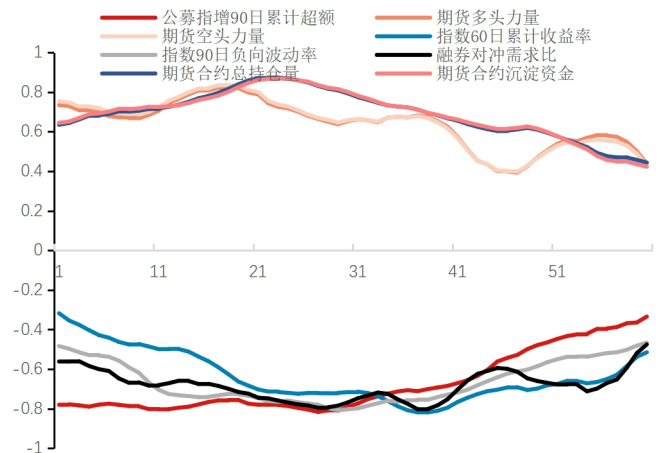
数据来源: 同花顺 华泰期货研究院

图 3: IC 因子有效性与预测周期 | 单位: 无



数据来源: 同花顺 华泰期货研究院

图 4: IM 因子有效性与预测周期 | 单位: 无



数据来源: 同花顺 华泰期货研究院

我们选取每个标的下相关系数平均值的绝对值最大的因子, 观察它们的相关性在预测周期变长时是否发生一定变化。从上图我们可以看到, 不同时间窗口的公募指增超额在 4 个标的中相关性都排名靠前。但当预测周期变长时, 除 IH 外, 其它标的的指增超额类因子相关性下降较快, 时间窗口越长, 前期的相关性越强, 但下降的速度也越快, 说明管理人的超额能力对短期内的基差水平影响较大。其它因子如指数收盘价、指数收益率、指数波动率则无明显单调下降趋势, 说明现货市场的涨跌对未来基差走势影响的延续性较强。有部分期货市场因子如沉淀资金、持仓量的相关性则呈先升后降的趋势, 说明此类因子对基差的影响有一定的滞后性。

接下来我们看看从 Distance 相关系数的角度下因子相关性是否发生一定变化。

表 2：不同预测周期下因子非线性相关统计|单位：无

标的	因子	出现次数	Distance corr 平均值
IF	公募指增 60 日累计超额	60	0.477
	指数前一日收盘价	60	0.378
	指数 90 日负向波动率	60	0.262
	指数成分股融券余额	60	0.259
	11 月哑变量	60	0.242
	7 月哑变量	48	0.226
	融券对冲需求比	60	0.226
	指数 60 日正向波动率	60	0.222
	期货多头力量	60	0.191
	公募指增 10 日累计超额极端涨幅	59	0.185
IH	指数前一日收盘价	60	0.537
	融券对冲需求比	60	0.488
	指数成分股融券余额	60	0.446
	期货合约总持仓量	60	0.355
	公募指增 60 日累计超额	60	0.347
	期货合约沉淀资金	60	0.318
	期货空头力量	60	0.291
	多空力量	60	0.277
	指数 60 日正向波动率	60	0.262
	指数 90 日负向波动率	60	0.245
IC	多空力量	60	0.458
	公募指增 60 日累计超额	60	0.404
	指数 90 日累计收益率	60	0.378
	指数 90 日正向波动率	60	0.341
	期货多头力量	60	0.319
	期货合约沉淀资金	60	0.301
	期货合约总持仓量	60	0.293
	指数 90 日负向波动率	60	0.287
	指数成分股融券余额	60	0.248
	11 月哑变量	50	0.210
IM	融券对冲需求比	60	0.760
	期货合约总持仓量	60	0.742
	期货合约沉淀资金	60	0.740
	指数 90 日负向波动率	60	0.708
	指数 60 日累计收益率	60	0.703
	公募指增 90 日累计超额	60	0.697
	期货空头力量	60	0.685
	指数前一日收盘价	60	0.658
	指数 90 日正向波动率	60	0.641
	8 月哑变量	60	0.566

数据来源：同花顺 华泰期货研究院

我们用和筛选线性相关性同样的筛选方法来筛选并展示因子的非线性相关性。从表中我们可以看到，Distance 相关系数筛出的强相关因子与 Pearson 相关系数筛出的强相关因子基本无异，但期货多头、空头力量类的因子在 4 个标中的相关性排名都有所上升。

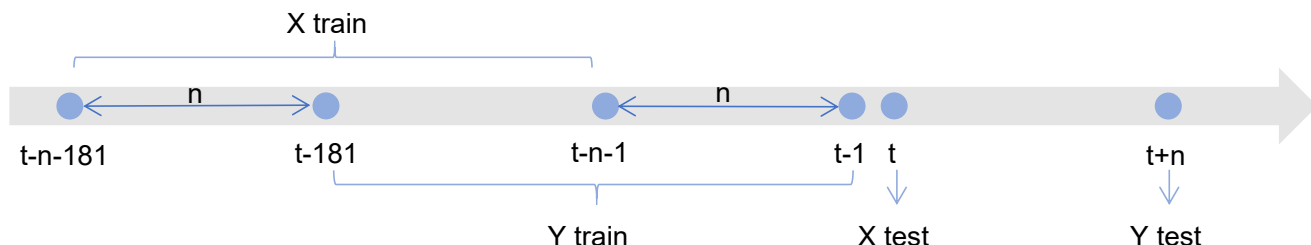
■ 多因子模型构建

上一步我们用 Pearson 相关系数和 Distance 相关系数分别筛选出了一部分因子进入模型的构建。

模型层面，我们首先选用 Pearson 相关系数选出的因子结合线性回归模型来捕捉因子的线性关系。为解决可能存在的共线性问题，我们选择 Ridge 回归，通过在损失函数中加入 L2 正则化项约束回归系数，增强回归系数的稳定性。其次，我们选用 Distance 相关系数结合两个基于决策树的集成机器学习模型，Bagging 类模型的代表 Random Forest 和 Boosting 类模型的代表 Xgboost 来捕捉因子的非线性关系。在将上一步初步筛选出的因子放入机器学习模型之前，我们用模型自带的 feature importance 对因子进行了进一步的筛选，并用 2023 年以前的数据对模型进行了调参。

预测层面，我们采用滚动窗口的方式来进行模型的训练和预测。令滚动窗口长度等于 180，每个交易日 t 取 $X_{[t-181-n, t-1-n]}$ 和 $Y_{[t-181, t-1]}$ 训练模型，再向训练好的模型输入 X_t 预测 Y_{t+n} ，即 $t+n$ 的年化基差率。

图 5：每个预测日 t 的训练集与测试集示意图 | 单位：无

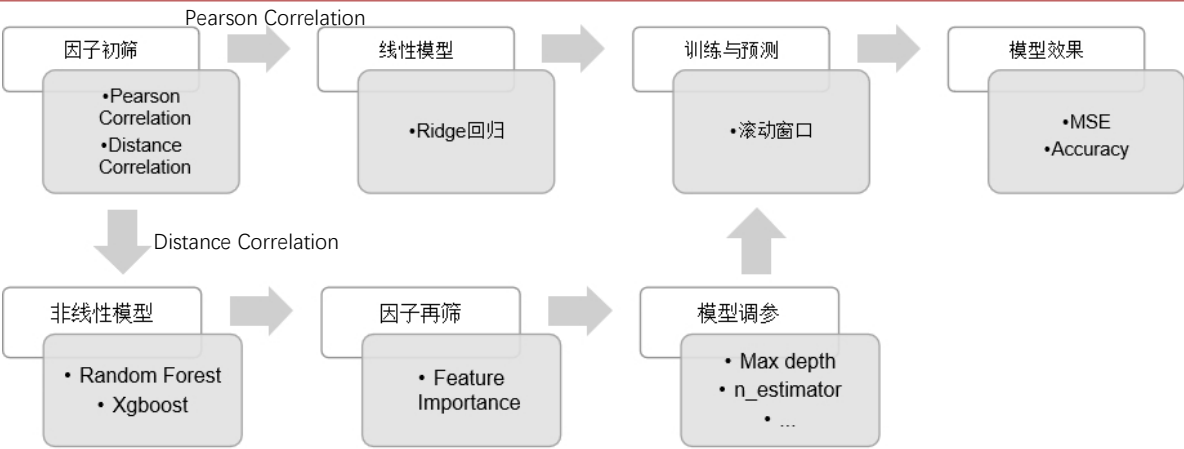


数据来源：华泰期货研究院

模型效果层面，我们选用 MSE 和年化基差率的涨跌方向准确率（Accuracy）作为模型的评判标准。

我们对涨跌方向预测准确与否的判断标准为，当 $\hat{Y}_{t+n} < Y_t$ & $Y_{t+n} < Y_t$ ，或 $\hat{Y}_{t+n} > Y_t$ & $Y_{t+n} > Y_t$ ，或 $\hat{Y}_{t+n} = Y_t$ & $Y_{t+n} = Y_t$ 时记为预测正确，其中 \hat{Y}_{t+n} 为预测值， Y_{t+n} ， Y_t 为真实值。

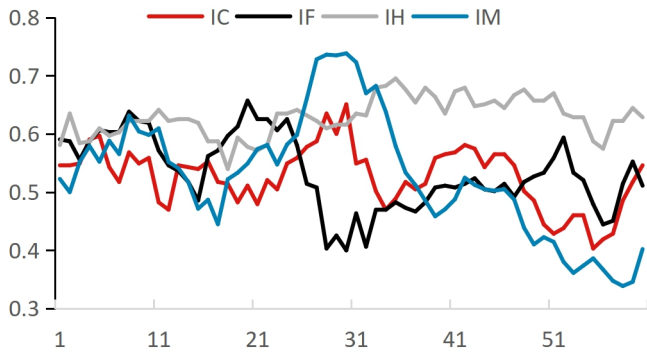
图 6：模型建立与预测流程图 | 单位：无



数据来源：华泰期货研究院

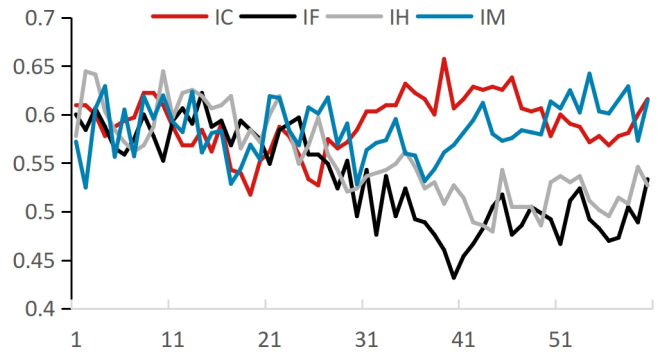
■ 模型预测效果展示

图 7: 不同预测周期的准确性—OLS+Ridge | 单位: 无



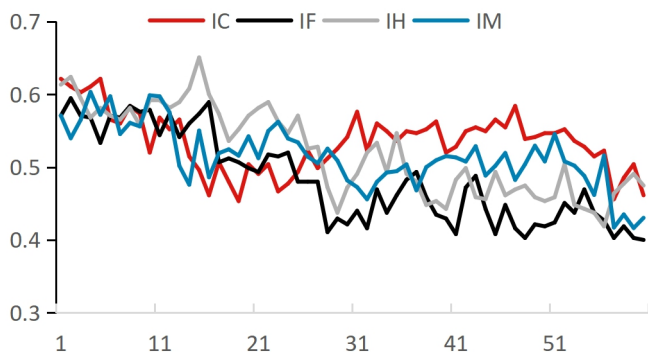
数据来源: 同花顺 华泰期货研究院

图 8: 不同预测周期的准确性—Random Forest | 单位: 无



数据来源: 同花顺 华泰期货研究院

图 9: 不同预测周期的准确性—Xgboost | 单位: 无



数据来源: 同花顺 华泰期货研究院

图 7、8、9 展示的是模型的测试集 2023 年后 Accuracy 随预测周期变化的趋势图。分模型来看, 在 Xgboost 中, 4 个品种的预测准确率随预测周期变长呈震荡下降的趋势, 在 OLS 和 Random Forest 中则无统一趋势。其中 Random Forest 从 30 天以后开始出现 50% 以下的准确率, 其它 2 个模型则在 10 天左右开始出现。

表 3：模型预测效果 | 单位：无

模型	标的	MSE	Accuracy
OLS+Ridge	IF	0.00050	53.35%
	IH	0.00027	62.87%
	IC	0.00149	52.62%
	IM	0.00238	52.78%
	平均	0.00116	55.41%
Random Forest	IF	0.00040	53.39%
	IH	0.00029	55.47%
	IC	0.00089	59.11%
	IM	0.00151	58.52%
	平均	0.00077	56.62%
Xgboost	IF	0.00044	48.31%
	IH	0.00030	52.01%
	IC	0.00097	53.49%
	IM	0.00160	51.54%
	平均	0.00083	51.34%

数据来源：同花顺 华泰期货研究院

表格中展示的是 2023 年后的模型效果。我们将 t+1 到 t+60 预测周期的结果计算平均值来衡量模型在 3 个月内的总体预测效果。可以看到 Random Forest 在模型预测精度方面表现较好, MSE 平均在 0.077%; OLS 在 IH 的涨跌准确度达到 62.87%, Random Forest 在 4 个标的总体准确度表现更好, 平均达到 56.62%。

下面我们 以 t+5（周度预测）为例，展示模型在 2023 年以后的预测效果。

表 4：t+5 年化基差率模型预测效果 | 单位：无

模型	标的	MSE	Accuracy
OLS+Ridge	IF	0.00020	60.63%
	IH	0.00019	60.95%
	IC	0.00046	59.68%
	IM	0.00093	55.23%
	平均	0.00045	59.12%
Random Forest	IF	0.00020	56.51%
	IH	0.00017	58.41%
	IC	0.00046	58.73%
	IM	0.00087	55.65%

Xgboost	平均	0.00043	57.33%
	IF	0.00024	53.33%
	IH	0.00018	58.13%
	IC	0.00045	62.13%
	IM	0.00090	57.19%
	平均	0.00044	57.70%

数据来源：同花顺 华泰期货研究院

Xgboost 在预测 t+5 年化基差率时 MSE 平均 0.044%，涨跌准确率平均 57.70%。

下表展示的是 Xgboost 在预测 t+5 年化基差率时每年的涨跌预测准确率。

表 5：t+5 年化基差率涨跌方向预测效果—Xgboost | 单位：无

		年份	Accuracy
IF		2018	61.02%
		2019	51.64%
		2020	58.85%
		2021	61.32%
		2022	50.83%
		2023	55.37%
		2024	49.62%
IH		2018	60.45%
		2019	58.61%
		2020	53.91%
		2021	53.91%
		2022	59.92%
		2023	59.50%
		2024	55.64%
IC		2018	58.76%
		2019	54.51%
		2020	65.84%
		2021	59.26%
		2022	57.85%
		2023	63.64%
		2024	59.40%
IM		2023	51.81%
		2024	63.91%

数据来源：同花顺 华泰期货研究院

表 6：t+5 年化基差率入选因子前 10—Xgboost | 单位：无

	IF	IH	IC	IM
1	3 月哑变量	指数成分股融券余额	1 月哑变量	公募指增 90 日累计超额
2	公募指增 60 日累计超额	3 月哑变量	指数 90 日正向波动率	指数 60 日正向波动率
3	公募指增 90 日累计超额	指数 90 日负向波动率	指数 90 日负向波动率	指数成分股融券余额
4	1 月哑变量	融券对冲需求比	多空力量	指数 90 日累计收益率
5	指数 90 日负向波动率	8 月哑变量	公募指增 90 日累计超额	指数 30 日负向波动率
6	指数 60 日负向波动率	7 月哑变量	公募指增 60 日累计超额	指数 60 日负向波动率
7	10 日累计超额收益率极端跌幅	公募指增 90 日累计超额	指数 30 日正向波动率	指数 10 日正向波动率
8	指数成分股融券余额	指数 60 日负向波动率	指数 60 日正向波动率	指数 30 日正向波动率
9	11 月哑变量	指数 60 日正向波动率	2 月哑变量	指数 60 日累计收益率
10	指数 90 日正向波动率	期货空头力量	指数 30 日负向波动率	指数 10 日负向波动率

数据来源：同花顺 华泰期货研究院

上表中展示的是用于预测 t+5（周度）年化基差率的因子，由 Distance 相关系数进行初步筛选，再由 Xgboost 中的 Feature Importance 二次筛选后排在前十名的因子。可以看到指数波动率和公募指增累计超额类因子占据大多数，在对不同品种的预测中均有出现。

■ 总结

本篇报告在《华泰期货量化策略专题报告 20240712：转融通暂停影响简述》基础上展开对股指期货年化基差率相关因子的量化分析。首先用 Pearson 相关系数以及 Distance 相关系数来衡量因子的线性与非线性关系，探讨因子在不同预测周期下的表现，并对因子进行筛选；其次构建了基于线性回归模型 OLS+Ridge 以及非线性的机器学习模型 Random Forest 和 Xgboost 的年化基差率预测模型，展示了模型在不同预测周期下的预测效果。在预测周度年化基差率时，Xgboost 对目标变量的预测精度方面表现较好，MSE 平均 0.044%，涨跌准确率平均 57.70%，最高达 62.13%。

■ 风险提示

回测结果基于历史数据得出，不排除失效的可能。

免责声明

本报告基于本公司认为可靠的、已公开的信息编制，但本公司对该等信息的准确性及完整性不作任何保证。本报告所载的意见、结论及预测仅反映报告发布当日的观点和判断。在不同时期，本公司可能会发出与本报告所载意见、评估及预测不一致的研究报告。本公司不保证本报告所含信息保持在最新状态。本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

本公司力求报告内容客观、公正，但本报告所载的观点、结论和建议仅供参考，投资者并不能依靠本报告以取代行使独立判断。对投资者依据或者使用本报告所造成的一切后果，本公司及作者均不承担任何法律责任。

本报告版权仅为本公司所有。未经本公司书面许可，任何机构或个人不得以翻版、复制、发表、引用或再次分发他人等任何形式侵犯本公司版权。如征得本公司同意进行引用、刊发的，需在允许的范围内使用，并注明出处为“华泰期货研究院”，且不得对本报告进行任何有悖原意的引用、删节和修改。本公司保留追究相关责任的权力。所有本报告中使用的商标、服务标记及标记均为本公司的商标、服务标记及标记。

华泰期货有限公司版权所有并保留一切权利。

公司总部

广州市天河区临江大道1号之一2101-2106单元 | 邮编：510000

电话：400-6280-888

网址：www.htfc.com