

【统计应用研究】

宏观经济指标、技术指标与国债期货价格预测

——基于随机森林机器学习的实证检验

陈标金^a, 王 锋^b

(华南农业大学 a. 经济管理学院; b. 集成期货研究所, 广东 广州 510642)

摘要: 分别以筛选的 4 种技术指标和 6 个宏观经济指标作为国债期货指数预测变量, 利用随机森林算法构建 4 种机器学习预测模型; 依据价格波动集聚性设计跟踪交易规则, 通过比较 4 种模型的预测精度和跟踪交易收益率, 检验宏观经济指标、技术指标和随机森林算法对国债期货指数的预测能力。研究结果发现: 用主成分精选技术指标构建的预测模型, 对国债期货指数的跟踪交易收益率虽然明显优于市场收益率, 但不如遵循单个技术指标经验交易规则的跟踪交易收益率; 用主成分精选技术指标和宏观经济指标构建的模型能够取得很好的预测精度和跟踪交易收益率, 这表明宏观经济指标与技术指标都对国债期货价格具有预测意义, 可以利用随机森林机器学习算法构建有效的国债期货量化投资模型。

关键词: 国债期货; 价格预测; 随机森林; 机器学习

中图分类号: F832.5

文献标志码: A

文章编号: 1007-3116(2019)06-0029-07

一、引言

国债期货价格预测可以为管理利率风险和调控市场利率水平提供依据, 因而受到投资者和宏观经济管理部门的重视。国债期货价格预测效果取决于预测变量的选择和预测模型的构建。

在预测变量选择上, 众多实证研究表明宏观经济指标对国债期货价格具有预测意义, 如 Piazzesi 和 Swanson 发现宏观经济和金融市场变量均会影响债券收益率, 非农工资增长率可用于预测美国联邦基金期货合约的超额收益率^[1]; Ludvigson 和 Ng 发现股票指数、利率和物价水平等变量可以用于债券风险溢价预测^[2]; Duffee 发现经济增长率和通货膨胀率会影响债券的收益率, 可用于预测债券收益率^[3]; Altavilla 等发现宏观经济意外会对债券收益率产生持续影响, 它虽仅能解释债券收益率日变化的 1/10, 却又能解释债券收益率季度变化的 1/3^[4];

尚玉皇等人发现 GDP 和通货膨胀对中国中长期国债收益率影响较大^[5]; 丁志国等人发现消费者信心指数变化率、M2 增长率的自然对数、生产者价格指数变化率 (PPI) 以及城镇居民失业率等变量可以提高国债收益率的拟合和预测效果^[6]。技术指标对金融资产价格具有预测意义也得到大量实证结论的支持, 如 Park 和 Irwin 的文献综述发现, 95 篇实证文献中有 56 篇的结论支持技术分析法可以获得超额收益, 仅 20 篇文献认为技术分析法不能获得超额收益^[7]。筛选宏观经济指标和技术指标作为预测变量, 是国债期货价格预测要解决的首要问题。

在预测模型的构建上, 金融资产价格预测模型大致可分为两类: 一类是传统统计预测模型, 如自回归滑动平均模型 (ARMA)、自回归异方差模型 (ARCH) 及其衍生扩展模型; 另一类可归纳为机器学习算法, 包括 K 近邻分类算法、神经网络算法 (ANNs)、支持向量机 (SVM) 和随机森林 (RF) 等。

收稿日期: 2018-07-21; 修稿日期: 2019-03-10

基金项目: 国家社会科学基金一般项目《基于投机视角的农产品期货定价机制研究》(13BJL065); 广东省教育改革项目《金融科技嵌入“金融工程”课程教育的教学方法研究》(2018JR009)

作者简介: 陈标金, 男, 江西石城人, 管理学博士, 教授, 硕士生导师, 研究方向: 期货市场, 金融工程;

王 锋, 男, 广东潮州人, 硕士生, 研究方向: 金融工程。

Breiman 认为随机森林算法不需要考虑多元回归模型中的多重共线性问题,可以加入数千个输入变量,具有明显的优异性^[8];Kampichler 等比较了 5 种机器学习算法的预测效果后发现,随机森林的预测效果通常都是最好的^[9];方匡南等在对基金裕隆收益率涨跌的预测中发现,随机森林预测准确率远远高于支持向量机、自回归移动平均和随机游走等预测方法^[10];方匡南等又在对保险客户利润贡献度预测中发现,随机森林方法要优于传统的类神经网络、CART 和 SVC 等模型^[11];涂艳等发现随机森林机器学习预测借款人违约行为的准确率更高^[12];王淑燕等使用随机森林算法预测股票涨跌也取得了较好效果^[13]。随机森林不需要考虑共线性问题,能够处理高维度数据,对异常值和噪声有很好的容忍度,不容易出现过拟合,预测准确率高,在高维度金融数据预测模型的构建中正在被越来越多的人所采用^[14]。

鉴于此,本文利用 2013 年 9 月 9 日—2017 年 10 月 31 日 5 年期国债期货指数序列,运用相关分析、跟踪交易回测和主成分分析筛选宏观经济指标和技术指标,作为国债期货指数预测变量;分别以选取的技术指标、技术指标和宏观经济指标、主成分压缩后的技术指标、主成分压缩后的技术指标和宏观经济指标作为预测变量,利用随机森林算法构建 4 种国债期货指数预测模型;依据国债期货指数波动集聚性设计跟踪交易规则,通过比较 4 种 RF 预测模型的预测精度和跟踪交易收益率,检验宏观经济指标、技术指标和随机森林算法对国债期货指数的预测能力。

二、国债期货指数预测变量的筛选方法

(一)宏观经济指标的筛选

影响国债价格或收益率的宏观经济因素主要包括通货膨胀和经济增长两类。投资者常用消费者物价指数和生产者物价指数度量通货膨胀,常用采购经理人指数度量经济增长动力。同时,考虑到铁路货运量、用电量和银行已放贷款量组成的“克强指数”能够准确衡量中国经济发展状况^[15],并已为投资者广泛采用,本文基于数据可得性用 M_2 代替银行已放贷款量,将消费者物价指数(CPI)、生产者物价指数(PPI)、采购经理人指数(PMI)、铁路货运量(RFV)、用电量(EU)、 M_1 、 M_2 等 7 个指标月同比变化率作为宏观经济指标备选预测变量。

由于被预测变量国债期货指数(P)和预测变量

宏观经济指标(X)不一定具有正态分布和等距特性,两者的相关也不一定是线性的。因此,同时计算国债期货指数月收益率与各个备选宏观经济指标月同比变化率之间的 Pearson 相关系数 ρ_p 、Spearman 秩相关系数 ρ_s 和 Kendall 秩相关系数 ρ_k ,并用相关系数的 t 显著性检验,筛选宏观经济指标作为国债期货指数的预测变量。

(二)技术指标的筛选

技术指标种类繁多,本文选取投资者常用的 MACD、KDJ、RSI、WR、PSY、OBV 共 6 种技术指标,从中筛选国债期货指数的技术指标类预测变量。各备选技术指标的计算方法、备选参数和跟踪交易规则设定见表 1。

首先,利用国债期货指数历史交易数据序列,分别针对各备选技术指标的各备选参数,以符合跟踪交易规则当日国债期货价格指数收盘价,跟踪交易 1 个单位的国债期货指数(不考虑交易费用),回测各备选技术指标各备选参数的跟踪交易预测准确率、年化收益率、年化收益率标准差、年化夏普比率和最大回撤率。设回测实际交易日为 N ,将回测中每次开仓到平仓定义为一次跟踪交易,跟踪交易次数为 n ,取得正收益的跟踪交易次数为 n_1 ,各次跟踪交易收益率为 $r_i (i=1,2,\dots,n)$,跟踪交易算术平均收益率为 \bar{r} ,国债期货指数年化市场收益率为 r_m ,每年按 250 个交易日计算。回测跟踪交易预测准确率(f)、年化收益率(r_y)、年化收益率标准差(σ_y)、年化夏普比率(S_p)和最大回撤率(L_r)的计算方法如下:

$$f = \frac{n_1}{n} \times 100\% \quad (1)$$

$$r_y = \frac{250}{N} \sum_{i=1}^n r_i \quad \text{技术指标的测试方法} \quad (2)$$

$$\sigma_y = \sqrt{\frac{250}{N} \sum_{i=1}^n (r_i - \bar{r})^2} \quad (3)$$

$$S_p = \frac{r_y - r_m}{\sigma_y} \quad (4)$$

$$L_r = \min(\sum_{i=1}^j r_i, j=1,2,\dots,n) \quad (5)$$

其次,针对每一个备选技术指标,筛选各备选参数中跟踪交易年化收益率最高且预测准确率超过 50%、年化夏普比率大于 1 的技术指标,作为国债期货指数的预测变量,即将预测准确率超过 50%、跟踪收益率超过市场收益率一个标准差的技术指标,视为对国债期货指数具有预测意义,并纳入预测变量。

表 1 备选技术指标、备选参数及跟踪交易规则表

指标	计算公式	备选参数	跟踪交易规则
MACD	$EMA(i)_t = \frac{i-1}{i+1} \times EMA(i)_{t-1} + \frac{2}{i+1} \times P_t$ $EMA(j)_t = \frac{j-1}{j+1} \times EMA(j)_{t-1} + \frac{2}{j+1} \times P_t$ $DIF_t = EMA(i)_t - EMA(j)_t$ $DEA_t = \frac{k-1}{k+1} \times DEA_{t-1} + \frac{2}{k+1} \times DIF_t$	$i = 8, 10, 12, 14, 16, 18$ $j = 17, 21, 25, 29, 33, 37$ $k = 9, 12, 15, 18, 21$	$DIF_t > 0, DEA_t > 0, DIF_t$ 向上交叉 DEA_t , 平前空做多 $DIF_t < 0, DEA_t < 0, DIF_t$ 向下交叉 DEA_t , 平前多做空
KDJ	$RSV_t = \frac{P_t - L_{it}}{H_{it} - L_{it}} \times 100$ $K_t = (j-1)/j \times K_{t-1} + 1/j \times RSV_t$ $D_t = (k-1)/k \times D_{t-1} + 1/k \times K_t$ $J_t = 3 \times D_t - 2 \times K_t$	$i = 9, 15, 21, 27$ $j = 3, 6, 9$ $k = 3, 6, 9$	$K_t < 20$, 且向上交叉 D_t , 平前空做多 $K_t > 80$, 且向下交叉 D_t , 平前多做空 $J_t > 100$, 平前多做空 $J_t < 0$, 平前空做多
RSI	$RSI_t = \frac{A_{it}}{A_{it} + B_{it}} \times 100$ <p>(A_{it} 为上涨天数涨幅平均值, B_{it} 为下跌天数跌幅平均值)</p>	$i = 6, 8, 10, 12, 14, 16, 18, 20, 22, 24$	$RSI_t > 80$, 为超买, 平前多做空 $RSI_t < 20$, 为超卖, 平前空做多
WR	$WR_t = \frac{H_{it} - P_t}{H_{it} - L_{it}} \times 100$	$i = 6, 8, 10, 12, 14, 16, 18, 20, 22, 24$	$WR_t > 80$, 平前空做多 $WR_t < 20$, 平前多做空
PSY	$PSY_t = Z_{it}/i \times 100$ <p>(Z_{it} 为 i 日内上涨天数)</p>	$i = 6, 8, 10, 12, 14, 16, 18, 20, 22, 24$	$PSY_t > 75$, 平前多做空 $PSY_t < 25$, 平前空做多
OBV	$OBV_t = OBV_{t-1} + \text{sgn}(P_t - P_{t-1}) \times VOL_t$ <p>(OBV 的初始值取第 1 日成交量)</p>	$i = 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22$	$P_t > P_{t-i+1}$, 且 $OBV_t < OBV_{t-i+1}$, 平前 多做空 $P_t < P_{t-i+1}$, 且 $OBV_t > OBV_{t-i+1}$, 平前 空做多

注: P_t 和 VOL_t 分别为第 t 日国债期货指数收盘价和成交量, H_{it} 和 L_{it} 分别为第 t 日前 i 日内国债期货指数的最高价和最低价。

(三) 预测变量的主成分分析与压缩

用上述方法筛选出的宏观经济指标和技术指标数量较多, 如果将其全部纳入国债期货指数的预测变量构建预测模型, 各个预测变量包含的信息可能存在重叠交叉, 因共线性等问题影响预测模型的精度。主成分分析可以在确保原始信息损失最小的前提下, 从多个相关联的变量系统中提取最具价值的少数变量, 将多变量系统转化为互不相关但包含了系统主要信息的少数变量系统, 在一定程度上能消除变量之间的共线性问题, 提高模型的精度。因此, 本文对筛选出的宏观经济指标和技术指标预测变量进行主成分分析, 以累计贡献率 $\geq 90\%$ 为标准; 再进一步压缩预测变量的个数, 并对主成分分析压缩前和压缩后的预测变量, 分别构建国债期货指数预测模型, 检验预测效果。

三、国债期货指数随机森林预测模型的构建与检验

考虑到金融市场的适应性市场特性, 随着市场环境的变化, 国债期货指数预测变量和预测模型都可能具有时变特性^[16]。因此, 对样本序列采用 100 日滚动训练的方法, 以筛选出的宏观经济指标和技

术指标为预测变量 X , 以国债期货指数日收益率 R 为被预测变量, $R_t = (P_t - P_{t-1})/P_{t-1} \times 100\%$, 利用 Breiman 提出的随机森林算法动态构建国债期货指数日收益率预测模型, 用被预测日前约 5 个月 100 个交易日的历史数据滚动建模, 逐日向前滚动预测各日收益率; 然后, 计算国债期货指数日收益率预测值与实际观察值的均方根误差 (RMSE) 和平均绝对误差 (MAE) 检验模型的预测精度, 设计交易规则回测检验模型的跟踪交易效果。

(一) 随机森林预测模型的构建

随机森林 (RF) 预测模型的构建步骤与方法是: 第一步, 从实证样本 N 个交易日时间序列中, 导出第 1~100 个交易日的国债期货指数日收益率及其对应的预测变量序列作为机器学习训练集 D , 生成随机向量序列 $\{\Theta_i, i = 1, 2, \dots, 100\}$, 采用 bootstrap 重抽样法从训练集 D 中随机抽取 100 个样本生成子样本集, 这样重复 100 次生成 100 个子样本集 D_i ($i = 1, 2, \dots, 100$); 第二步, 对每个 bootstrap 子样本集 D_i ($i = 1, 2, \dots, 100$) 分别用随机森林算法建立国债期货指数日收益率树分类模型 $\{h(X, \Theta_i), i = 1, 2, \dots, 100\}$, 假设参数集 $\{\Theta_i\}$ 是独立同分布的; 第三步, 通过 100 轮训练, 得到一个决策树分类序列

$\{h_1(X), h_2(X), \dots, h_{100}(X)\}$, 再用该序列构建决策树分类模型系统; 第四步, 将第 101 日的预测变量 X 代入这个决策树分类模型系统, 以简单多数投票决定第 101 日国债期货指数日收益率最终分类结果预测值 $H(X)$:

$$H(X) = \arg \max_R \sum_{i=1}^{100} I(h_i(X) = R) \quad (6)$$

其中 $I(\cdot)$ 为示性函数。沿着实证样本时间序列向前滚动, 重复这 4 个步骤和方法, 用第 2 个交易日至第 101 个交易日作为机器学习训练集 D 构建随机森林预测模型, 预测第 102 个交易日的国债期货指数日收益率; 以此类推, 逐日向前滚动学习构建预测模型, 对下一交易日国债期货指数日收益率作出预测。

分别选取已筛选出的不同指标作为预测变量, 遵循上述方法构建 4 种 RF 预测模型: 模型 1 的预测变量为已筛选出的所有技术指标; 模型 2 的预测变量为已筛选出的所有技术指标和宏观经济指标; 模型 3 的预测变量为主成分分析压缩后的技术指标; 模型 4 的预测变量为主成分分析压缩后的技术指标和宏观经济指标。通过比较 4 种模型的预测精度和跟踪交易回测收益率, 检验技术指标、宏观经济指标和随机森林模型对国债期货指数的预测能力。

(二) 预测精度检验方法

采用国债期货指数日收益率预测值与实际观察值的均方根误差 (RMSE) 和平均绝对误差 (MAE) 作为 4 种 RF 模型预测精度的评价标准。设 N 为样本数, R'_i 和 R_i 分别表示预测值和实际观察值, 即:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^N (R'_i - R_i)^2}{N}} \quad (7)$$

$$\text{MAE} = \frac{\sum_{i=1}^N |R'_i - R_i|}{N} \quad (8)$$

RMSE 和 MAE 的数值越小, 表示预测精度越高。

(三) 跟踪交易回测效果检验方法

利用 4 种 RF 模型预测值构建跟踪交易规则, 回测跟踪交易收益率, 也可以检验预测变量和预测模型的有效性。设计跟踪交易规则的关键在于设定做多和做空的阈值。以国债期货指数日收益率序列标准差的一定比例作为阈值, 设计预测模型的跟踪交易规则, 即当预测值大于标准差一定比例时做多; 当预测值小于负的标准差一定比例时做空。通过比较 4 种 RF 模型最优阈值比例系数的跟踪交易收益率, 检验预测变量和预测模型的有效性。

考虑到国债期货指数存在波动集聚特性, 日收益率序列的标准差是时变的。因此, 作为跟踪交易规则阈值的标准差, 也应该是被预测日的标准差预测值。与 RF 滚动建模相对应, 用 2 种方法测算标准差 σ_t 的预测值: 一是用被预测日前 100 个交易日的国债期货指数日收益率计算普通标准差, 简单外推, 作为被预测日标准差预测值 σ_t :

$$\sigma_t = \sqrt{\frac{1}{100} \sum_{i=1}^{100} (R_{t-i+1} - \bar{R}_t)^2} \quad (9)$$

二是考虑到 GARCH(1,1) 能较好地识别方差的时变性, 用被预测日前 100 个交易日的国债期货指数日收益率序列构建 GARCH(1,1) 模型, 获得被预测日标准差的预测值 σ_t :

$$\sigma_t^2 = \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2 \quad (10)$$

不考虑交易费用, 分别选取阈值比例系数 $\gamma = 0.05, 0.10, 0.25, 0.5, 0.75, 1.0$ 的 6 个比例, 按照国债期货指数日收益率预测值 $R'_i \geq \gamma \sigma_t$ 当天以收盘价做多 1 单位国债期货指数或继续持有 1 单位已做多仓位, 直至 $R'_i < \gamma \sigma_t$ 当天以收盘价平仓; $R'_i \leq -\gamma \sigma_t$ 当天以收盘价做空 1 单位国债期货指数或继续持有 1 单位已做空仓位, 直至 $R'_i > -\gamma \sigma_t$ 当天以收盘价平仓。同样, 可以用式 (1) ~ (5) 计算跟踪交易的预测准确率 (f)、年化收益率 (r_y)、年化收益率标准差 (σ_y)、年化夏普比率 (S_p) 和最大回撤率 (L_r) 筛选各个 RF 模型的最优阈值比例系数, 比较 4 种 RF 模型的跟踪交易效果。

由于年化夏普比率是评估跟踪交易回测效果的综合指标, 这里只以 S_p 为依据筛选各 RF 模型的最优阈值比例系数, 比较 4 种 RF 模型跟踪交易的最优回测效果; $S_p > 0$ 表示 RF 模型的跟踪交易收益率好于市场收益率, $S_p > 1$ 表示 RF 模型的跟踪交易收益率超过市场收益率一个标准差; S_p 越大, RF 模型的预测效果越好。

四、国债期货指数预测与跟踪交易回测实证分析

(一) 变量样本数据说明

本文选取 5 年期国债期货上市当日 2013 年 9 月 9 日—2017 年 10 月 31 日为实证样本时间区间。第一, 被预测变量为 5 年期国债期货加权指数日收益率, 预测变量为宏观经济指标与技术指标, 其中备选宏观经济指标为消费者物价指数 (CPI)、生产者物价指数 (PPI)、采购经理人指数 (PMI)、铁路货运量 (RFV)、用电量 (EU)、 M_1 、 M_2 共 7 个指标的月同

比变化率,备选技术指标为 $MACD(i, j, k)$ 、 $KDJ(i, j, k)$ 、 $RSI(i)$ 、 $WR(i)$ 、 $PSY(i)$ 、 $OBV(i)$ 共 6 种技术指标;第二,国债期货指数各交易日的开盘价、收盘价、最高价、最低价和交易量数据来源于大智慧行情系统,国债期货指数日收益率和各备选技术指标根据这些市场交易数据计算得到;第三,宏观经济数据来源于东方财富网,由于各备选宏观经济指标是每月公布上个月的月度数据,本文按公布当日更新数的方式构造宏观经济指标日序列,交易日无新宏观经济数据公布则沿用上一日的数据;第四,本文将选取的实证样本分为前后两个时段,用 2013 年 9 月 9 日—2015 年 12 月 31 日的样本筛选具有预测意义的宏观经济指标和技术指标,然后以选出宏观经济指标和技术指标作为预测变量,用 2016 年 1 月 1 日—2017 年 10 月 31 日的样本按 100 日滚动机器学习构建 RF 预测模型,检验宏观经济指标、技术指标、RF 模型的预测和跟踪交易回测效果。

(二) 预测变量筛选结果

1. 宏观经济指标筛选结果。2013 年 9 月—2015 年 12 月,国债期货指数月收益率与 7 个备选宏观经济指标前一个月同比变化率的 Pearson 相关系数 ρ_p 、Spearman 秩相关系数 ρ_s 、kendall 秩相关系数 ρ_k 的计算结果如表 2。在 5% 的显著性水平下,国债期货指数月收益率除了与 M_1 的月同比变化率不相关外,与 CPI、PPI、PMI、铁路运量(RFV)、用电量(EU)、 M_2 的月同比变化率都显著负相关。因此,本文选取 CPI、PPI、PMI、RFV、EU、 M_2 的月同比变化率 6 个宏观经济指标作为国债期货指数的预测变量。

表 2 国债期货指数月收益率
与宏观经济指标月同比变化率相关系数表

相关系数	CPI**	PPI**	PMI**	RFV**	EU**	M_1	M_2^*
ρ_p	-0.790	-0.881	-0.747	-0.844	-0.764	0.241	-0.380
[P 值]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.217]	[0.046]
ρ_s	-0.767	-0.818	-0.801	-0.844	-0.812	0.075	-0.386
[P 值]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.704]	[0.042]
ρ_k	-0.571	-0.623	-0.591	-0.656	-0.594	-0.037	0.262
[P 值]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.782]	[0.052]

注: *、** 分别为 5%、1% 显著性水平下相关显著。

2. 技术指标筛选结果。遵循表 1 和式(1)~(5)的方法,用 2013 年 9 月 9 日—2015 年 12 月 31 日国债期货指数日交易数据,分别针对 $MACD(i, j, k)$ 、 $KDJ(i, j, k)$ 、 $RSI(i)$ 、 $WR(i)$ 、 $PSY(i)$ 、 $OBV(i)$ 6 种技术指标的各备选参数进行跟踪交易回测。以年化收益率最高且预测准确率超过 50%、年化夏普比率

大于 1 的标准,对每一技术指标技术各备选参数进行筛选,结果如表 3。表 3 中, $MACD(8, 17, 9)$ 、 $KDJ(9, 6, 3)$ 、 $WR(6)$ 、 $PSY(6)$ 的预测准确率都在 60% 以上,跟踪交易年化夏普比率均大于 1.8,对国债期货指数具有很强的预测能力,本文将这 4 种技术指标选为国债期货指数的预测变量。

表 3 技术指标最优参数筛选结果表

指标	(i, j, k)	预测准确率	年化收益率	年化标准差	夏普比率	最大回撤率
MACD	(8, 17, 9)	60.432	8.320	2.720	1.811	-1.210
KDJ	(9, 6, 3)	81.159	17.730	2.510	5.586	-0.510
RSI	—	—	—	—	—	—
WR	(6)	77.586	24.850	2.640	7.934	-0.210
PSY	(6)	85.385	7.300	1.700	2.316	-0.480
OBV	—	—	—	—	—	—

注: — 为没有符合筛选标准的参数。

(三) 4 种 RF 模型的预测精度比较

用 2016 年 1 月 1 日—2017 年 10 月 31 日的样本数据,按照式(6)以 100 日滚动机器学习构建 RF 模型。针对不同的预测变量构建 4 种 RF 模型:模型 1 的预测变量是筛选出的 4 种技术指标;模型 2 的预测变量是筛选出的 4 种技术指标和 6 个宏观经济指标;模型 3 的预测变量是取累计贡献率 90% 以上主成分压缩后的技术指标;模型 4 的预测变量是取累计贡献率 90% 以上主成分压缩后的技术指标和宏观经济指标。根据式(7)~(8),4 种 RF 模型国债期货指数日收益率预测值与实际观察值的均方根误差(RMSE)和平均绝对误差(MAE)的计算结果如表 4。

表 4 4 种 RF 模型预测误差比较表 单位: %

误差	模型 1	模型 2	模型 3	模型 4
RMSE	0.313	0.303	0.305	0.297
MAE	0.176	0.173	0.192	0.190

比较 4 种 RF 模型的预测精度可以看出:首先,模型 2 的 RMSE 和 MAE 都小于模型 1,模型 4 的 RMSE 和 MAE 都小于模型 3,即同时用技术指标和宏观经济指标作为预测变量构建的 RF 模型,预测精度总要优于只用技术指标作为预测变量构建的 RF 模型,这表明宏观经济指标对国债期货价格具有预测意义,基本分析法有助于提升国债期货价格预测的准确性;其次,模型 3 的 RMSE 小于模型 1,但 MAE 大于模型 1;模型 4 的 RMSE 小于模型 2,但 MAE 大于模型 2,这表明用主成分分析压缩预测变量个数并不一定能改善 RF 模型的预测精度。

(四) 4 种 RF 模型的跟踪交易回测效果比较

分别用式(9)式(10)的预测标准差作为阈值设计基础,选取阈值比例系数 $\gamma=0.05、0.10、0.25、0.5、0.75、1.0$ 的 6 个比例,按前文设计的跟踪交易规则对 4 种 RF 模型 2016 年 1 月 1 日—2017 年 10 月 31 日间预测值的跟踪交易回测效果如表 5。

表 5 4 种 RF 模型跟踪交易回测效果年化夏普比率表

标准差	阈值系数	模型 1	模型 2	模型 3	模型 4
σ_t	0.05	-0.410	0.916	1.762*	1.796*
	0.10	-0.565	0.811	1.713	1.575
	0.25	-0.793	0.968	1.008	1.271
	0.50	-0.729	0.724	0.010	0.443
	0.75	-0.392	1.267	0.311	0.179
	1.00	-0.365	2.785*	0.012	0.047
GARCH σ_t	0.05	-0.442	1.011	1.274*	1.819
	0.10	-0.594	0.825	1.206	2.632
	0.25	-0.028	1.721	0.695	5.556*
	0.50	-0.370	0.914	0.291	—
	0.75	-5.942	2.951	0.597	—
	1.00	-11.198	6.350*	0.446	—

注: * 为跟踪交易的最优回测年化夏普比率, — 为回测中没有交易机会。

比较 4 种 RF 模型的跟踪交易回测效果可以看出:

首先,以主成分分析压缩后的技术指标为预测变量构建的 RF 模型具有预测能力,但以多种技术指标构建的 RF 模型,跟踪交易收益率不如单个技术指标的经验交易规则;模型 3 的最优回测年化夏普比率大于 1,模型 1 的回测年化夏普比率却小于 0,而且模型 3 和模型 1 的回测年化夏普比率都远远小于表 3 中单个技术指标经验交易规则,这显示出只经过 100 个交易日短期学习训练的 RF 模型,不能同时有效识别多个技术指标的预测规则;主成分分析压缩技术指标个数可以提高 RF 模型的预测效果,但利用技术指标构建 RF 模型预测效果总体不如经验丰富的交易者采用单个技术指标的预测效果。

其次,在技术指标基础上再将宏观经济指标引入预测变量构建 RF 预测模型,可以明显提高模型跟踪交易收益率;各阈值的回测年化夏普比率模型 2 普遍大于模型 1,模型 4 普遍大于模型 3,无论以普通标准差还是 GARCH 标准差作为阈值基础,最优回测年化夏普比率也是模型 2 大于模型 1,模型 4 大于模型 3,这进一步验证了宏观经济指标对国债期货指数具有预测意义。

再次,用主成分分析精选宏观经济指标和技术指标构建 RF 模型,兼顾国债期货价格的波动集聚特性构建跟踪交易规则,可以取得很好的预测精度和跟踪交易收益率。模型 4 和模型 2 的最优回测年化夏普比率分别高达 5.556 和 6.350,表明精选宏观经济指标和技术指标构建 RF 模型预测国债期货指数是有效的。

五、结论与启示

首先,本文通过相关分析和跟踪交易回测,筛选出 CPI、PPI、PMI、铁路运量(RFV)、用电量(EU)、 M_2 的月同比变化率 6 个宏观经济指标和 MACD(8,17,9)、KDJ(9,6,3)、WR(6)、PSY(6) 4 种技术指标,作为国债期货指数的预测变量;其次,考虑国债期货指数预测变量和预测模型都可能存在时变性,分别以技术指标、技术指标和宏观经济指标、累计贡献率 90% 以上主成分压缩后的技术指标、累计贡献率 90% 以上主成分压缩后的技术指标和宏观经济指标为预测变量,利用随机森林算法按 100 日滚动训练,动态构建了 4 种国债期货指数日收益率 RF 预测模型;最后,依据国债期货指数波动集聚性设计跟踪交易规则,通过比较 4 种 RF 预测模型的预测精度和跟踪交易收益率,检验了宏观经济指标、技术指标和随机森林算法对国债期货指数的预测能力,结果发现:第一,精选越少的技术指标作为预测变量对国债期货指数的预测效果反而越好,以多种技术指标构建的 RF 模型,跟踪交易收益率不如单个技术指标的经验交易规则;第二,在技术指标基础上再引入宏观经济指标构建 RF 模型,能明显提高 RF 模型对国债期货指数的预测精度和跟踪交易收益率;第三,精选宏观经济指标和技术指标,利用随机森林算法构建动态国债期货指数 RF 预测模型,依据价格波动集聚性设计跟踪交易规则,能够取得很高的跟踪交易收益率。

从这些实证结论中至少可以得到三点启示:一是宏观经济指标对国债期货价格具有预测意义,基本分析法有助于提高国债期货价格的预测效果;二是技术分析法对国债期货价格也具有预测意义,国债期货投资中遵循精选的单个技术指标跟踪交易收益率更高;三是可以利用随机森林算法,依据价格波动集聚性设计跟踪交易规则,构建有效的国债期货人工智能量化投资模型。

参考文献:

- [1] Piazzesi M, Swanson E. Futures Prices as Risk-Adjusted Forecasts of Monetary Policy[J]. Journal of Monetary Economics, 2008, 55(4).
- [2] Ludvigson S, Ng S. Macro Factors in Bond Risk Premia[J]. Review of Financial Studies, 2009, 22(12).
- [3] Duffee G. Information in (and not in) the Term Structure[J]. Review of Financial Studies, 2011, 24(9).
- [4] Altavilla C, Giannone D, Modugno M. Low Frequency Effects of Macroeconomic News on Government Bond Yields[J]. Journal of Monetary Economics, 2017, 92(12).
- [5] 尚玉皇, 郑挺国, 夏凯. 宏观因子与利率期限结构: 基于混频 Nelson-Siegel 模型[J]. 金融研究, 2015(6).
- [6] 丁志国, 耿迎涛, 覃朝晖. 中国市场国债利率期限结构的动态特征研究[J]. 统计研究, 2016(1).
- [7] Park Cheol-H, Irwin S. What Do We Know About the Profitability of Technical Analysis? [J]. Journal of Economic Surveys, 2007, 21(4).
- [8] Breiman L. Random Forests[J]. Machine Learning, 2001, 45(1).
- [9] Kampichler C, Wieland R, Calmé S, et al. Classification in Conservation Biology: A Comparison of Five Machine-learning Methods[J]. Ecological Informatics, 2010, 5(6).
- [10] 方匡南, 朱建平, 谢邦昌. 基于随机森林方法的基金收益率方向预测与交易策略研究[J]. 经济经纬, 2010(2).
- [11] 方匡南, 吴见彬, 谢邦昌. 基于随机森林的保险客户利润贡献度研究[J]. 数理统计与管理, 2014(6).
- [12] 涂艳, 王翔宇. 基于机器学习的 P2P 网络借贷违约风险预警研究——来自“拍拍贷”的借贷交易证据[J]. 统计与信息论坛, 2018(6).
- [13] 王淑燕, 曹正凤, 陈铭芷. 随机森林在量化选股中的应用研究[J]. 运筹与管理, 2016(3).
- [14] 方匡南, 吴见彬, 朱建平, 等. 随机森林方法研究综述[J]. 统计与信息论坛, 2011(3).
- [15] 董雨, 张泽宇. “克强指数”的应用效果检验及改进[J]. 首都经济贸易大学学报, 2016(4).
- [16] 周孝华, 宋庆阳, 刘星. 适应性市场假说及其在中国资本市场的实证[J]. 管理科学学报, 2017(6).

**Macroeconomic Indicators, Technical Indicators and Treasury Bond Futures Price Prediction:
Empirical Test on Random Forest Machine Learning**

CHEN Biao-jin, WANG Feng

(a. College of Economics and Management; b. Jicheng Futures Markets Research Institute,
South China Agricultural University, Guangzhou 510642, China)

Abstract: Four technical indicators and six macroeconomic indicators are selected as the predictor variables of the Treasury bond futures price. Four Treasury bond futures price prediction machine learning models are constructed by using the random forest algorithm. Designing the track trading rules in accordance with price fluctuation aggregation, the predict ability of macroeconomic indicators, technical indicators and random forest algorithm has been tested by comparing prediction accuracy and track trading return of the four RF models. The empirical results show that, the track trading return of the RF model constructed with the principal component selection technical indicators is much better than market return, but obviously worse than the track trading return of a single technical indicators empirical trading rules. The RF model constructed with the principal component selection technical indicators and macroeconomic indicators can achieve good prediction accuracy and track trading return. It means that both macroeconomic indicators and technical indicators have ability to predict the Treasury bond futures price, and random forest machine learning algorithm can be used to construct Treasury bond futures quantitative investment model effectively.

Key words: treasury bond futures; price prediction; random forest; machine learning

(责任编辑:郭诗梦)