doi: 10.19920/j.cnki.jmsc.2025.08.001

基于宏观大数据的 CPI 预测及方法比较①

郑挺国[1,2,3]，范馨月[4*]，靳 炜[5]

(1. 厦门大学宏观经济研究中心，厦门 361005；2. 厦门大学福建统计科学重点实验室，厦门 361005；
3. 厦门大学经济学院，厦门 361005；4. 上海财经大学经济学院，上海 200433；
5. 鹏华基金管理有限公司，深圳 518055)

摘要：大数据时代的到来为 CPI 的预测带来了前所未有机遇和挑战. 充分利用高维数据信息，发展可解释的机器学习预测模型，对于理论发展和现实实践均具有重要意义. 为此，本研究构建了包含 9 个类别 239 个变量的中国月度宏观经济数据库，并对比了包含传统时间序列模型、正则化回归、因子模型和集成算法等在内的 13 个模型在大型数据集下对 CPI 的预测能力. 进一步地，基于控制变量的思想构建了机器学习衍生算法，对相关的结果进行解释和机制分析. 结果表明，随机森林和 XGBoost 具有良好的预测效果，尤其是在中长期预测中表现出了较大优势. 通过进一步的分析发现它们的优势在于非线性的模型设定和非稀疏的变量处理，前者使得模型中的变量关系更加符合实际，而后者能够充分地利用大数据信息. 同时，这两个模型也筛选出了自回归项、价格、就业等在 CPI 预测中更加合理且重要的变量类别.

关键词：宏观大数据；CPI 预测；机器学习；非线性；衍生算法；变量选择

中图分类号：F064.1 **文献标识码**：A **文章编号**：1007－9807(2025)08－0001－16

0 引 言

居民消费价格指数(CPI)作为反映通货膨胀水平的一个重要指标，是连接名义变量和实际变量的桥梁，能够为国家的政策制定、宏观调控和国民经济核算提供参考依据，因此对 CPI 进行准确的预测一直都是政府部门、学者乃至市场参与者都十分关心的问题.

传统的 CPI 预测主要有以下几种方法：时序计量模型、灰色预测和组合预测等. 其中，时序计量模型主要使用时间序列数据来进行参数估计和变量预测，近年来该类模型得到了较快发展，例如混频数据建模技术能够同时使用不同频率的数据信息来对宏观经济变量进行预测[1－4]. 灰色预测主要针对含有不确定性因素的系统，它通过进行关联性分析来寻找系统因素之间的相异程度，然后结合系统规律来生成数据系列并建立微分方程进行预测，该方法已经从单变量模型发展出多变量模型，并逐步应用于经济变量的预测[5,6]. 组合预测是对多个预测模型结果的综合，该方法普遍采用加权平均的方式，将来自不同模型的预测结果进行加总处理，已有很多研究致力于寻找恰当的权重选取方法来提高预测精度[7－10]. 然而，传统的预测模型虽然能够通过一定的优化和改进来提高预测精度，但始终面临着可用数据不足、解释力度有限等问题.

① 收稿日期：2021－04－22；修订日期：2023－07－06.
基金项目：国家社会科学基金资助重大项目(23&ZD074)；教育部人文社会科学重点研究基地资助重大项目(22JJD790050)；全国统计科学研究项目资助重点项目(2022LZ37)；上海市晨光计划资助项目(2024111015)；人工智能促进科研范式改革赋能学科跃升计划资助一般项目(2024111008).
通讯作者：范馨月(1996—)，女，云南弥勒人，博士，讲师. Email: xinyue.96@qq.com

大数据时代的到来,为宏观经济预测带来了新的机遇,同时也带来了前所未有的挑战. 随着数据获取速度的加快和可用数据量的扩大,传统模型出现了维数灾难、精度下降等问题,甚至由于估计方法的限制,部分模型已无法适用于大型数据集. 因此,如何恰当地构建预测模型来充分使用大数据信息,并以此提高预测精度,已成为亟待解决的一个重要问题. 一些国内学者已在大数据预测领域进行了许多有益的探索. 如刘涛雄和徐晓飞[11]运用经济变量、政府统计指标和互联网数据三种信息构建了六个多元线性模型对中国的宏观经济总量进行预测,他们发现在选择适当的模型后,可以通过大数据信息来提高预测效果; 王娜[12]提出网络结构自回归分布滞后(ADL) 模型,并运用百度指数和媒体指数进行碳价预测,发现该模型适用于大数据环境并能够得到较好的预测结果; 董莉等[13]使用百度指数并运用弹性网(elastic net) 方法对 CPI 进行实时预测,一定程度上克服了传统模型的维度灾难问题. 但这些方法都有一定的局限性,具体而言包括如下两个方面: 一是模型的线性设定不符合实际,例如根据菲利普斯曲线,通货膨胀率与就业的关系是非线性的,此外,伴随着经济不确定性的增强和变量个数的增多,CPI 与其他变量间的关系将更加复杂,不能仅通过线性模型来进行刻画; 二是稀疏性的约束将损失大量数据信息,为了实现数据的降维,此前有多种方法通过将部分变量的系数进行压缩来构造稀疏的自变量集,这种做法不仅在解释力上存在争议,而且容易损失样本信息[14].

为改进此前方法在模型设定和数据利用上的不足,国外学者开展了丰富的研究,其主要思想是将机器学习方法用于经济变量的预测. 例如,Milunovich[15]对比了传统时间序列模型和机器学习的样本外预测能力,发现传统模型仅在短期表现良好,而机器学习在中长期预测中有突出优势. Rodriguez-Vargas [16]运用随机森林、长 - 短期记忆神经网络(long-short term memory network, LSTM) 等模型对 CPI 进行预测,发现这些机器学习算法的预测性能优于此前的基础模型,并且能通过组合平均的方式进一步提升预测效果. Kohei 和 Motot-sugu[17]运用因子模型和机器学习方法对日本的时间序列数据进行预测,得出了这两类模型在多数情况下优于 AR 模型的结论,并再次验证了机器学习在长期预测中的优越性. 这些研究均指出,机器学习方法的优势主要来源于变量间的非线性设定,但他们并未与大数据背景紧密结合,也因此未能揭示非稀疏变量集在预测中的重要性.

相比而言,国内仅有较少研究讨论了机器学习算法在经济预测中的应用. 例如,Du 等[18]采用 BP 神经网络 (BPNN) 、灰色预测 (GM(1,1)) 和自回归综合移动平均 (ARIMA) 三种模型对 CPI 的分类指标进行预测然后加总得到总 CPI 指数,他们证实了非线性设定有助于提高 CPI 的预测精度. 薛晔等[19]采用决策树算法对预测指标进行筛选和优化,然后借助 BP 神经网络模型对通货膨胀的风险等级进行预测,得出了优于 ARIMA 模型的结果. 然而,这些研究仍存在着几个方面的不足: 首先,他们虽然部分解决了模型设定问题,但并未与大数据的背景相结合,仅在较小数据集下进行了实证分析,不能够将结论推广到大数据的情景. 其次,他们采用了不同于参数压缩的方法进行变量选择,但所用的算法仅能部分缓解稀疏性问题,仍然会有数据信息的损失. 同时,他们的研究把变量选择和模型估计作为两个独立的过程,会造成变量选择不一致和结果缺乏解释力的问题. 最后,他们缺乏各个模型在不同预测期限下的对比,因此仅能说明其所用模型在短期内具有一定优势,但在长期预测中这一结果并不具有可信度.

基于此,本研究结合宏观大数据背景及实际预测需要,运用多种方法对我国月度 CPI 同比增长率进行预测,并对结果进行详细地分析和比较,为我国现实经济中的政府部门、学者和市场参与者提供预测工具和参考依据. 本研究的贡献主要体现在以下三点: 首先,类比于美国 FRED-MD 数据库,构建了中国月度宏观经济数据库(CMED-MD) ,其中包括了 9 个类别共 239 个数据序列. 该数据库能够为宏观经济的预测提供丰富的数据信息,不仅适用于本研究,也能为此后的相关研究提

供帮助; 其次,根据机器学习中著名的“没有免费的午餐定理(no free lunch theorem) ”②,在进行实证研究前难以知晓何种方法能够在大数据环境下对我国的 CPI 走势进行更加准确的预测,因此本研究系统地考察了多种预测方法在大数据集下的预测效果,进行了严谨的样本外预测对比分析,并得出了适用于大数据集的预测模型,这具有重要的实证意义; 第三,参照 Goulet-Coulombe 等[22] 控制变量的思想构造了衍生模型,并揭示了非线性设定和非稀疏的变量集在 CPI 预测中的重要性,在一定程度上克服了机器学习的黑箱难题,增强了模型的解释力; 最后,本研究通过对 CPI 预测中的变量重要性进行测度,对结果展开了详细讨论和机制分析,弥补了此前相关预测文献在解释分析方面的不足. 此外,对比 Medeiros 等[23] 的研究,本研究在其基础上增加了极端梯度提升算法(XGBoost) 及相关模型,并由此得出了更为精确的预测结果.

1　预测模型及估计方法

为了寻找适用于大型数据集的模型,并说明非线性设定和非稀疏变量集在预测中的重要性. 本研究对此前常用的预测方法进行了梳理和对比,在此基础上使用机器学习及其衍生算法对月度 CPI 增长率进行预测研究. 本节首先对常用预测模型进行分类,然后阐明机器学习的原理和衍生算法思路.

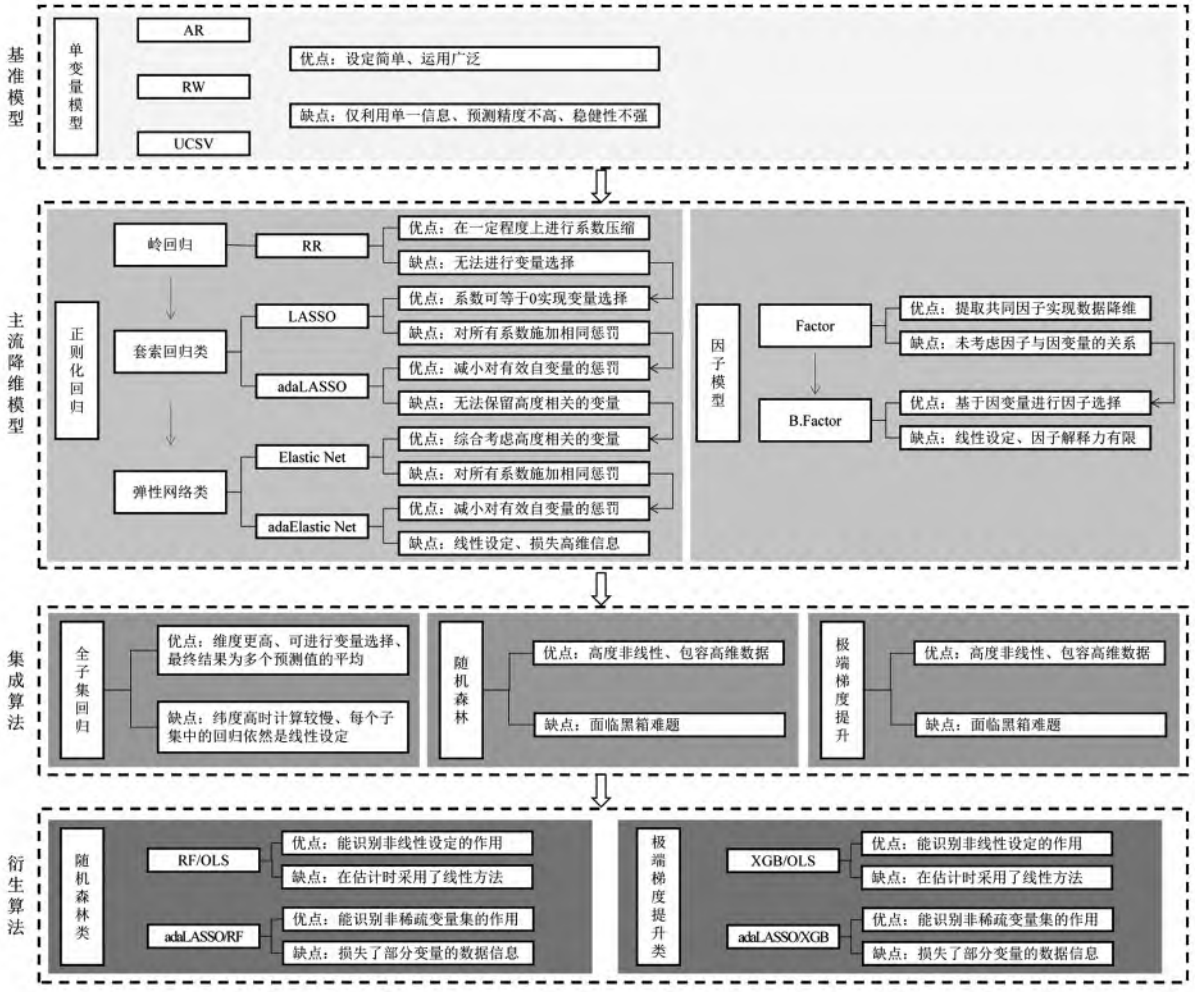

图 1　预测方法分类及概要

Fig. 1 Classification and summary of forecasting methods

② Wolpert[20] 提出了没有免费的午餐定理(no free lunch theorem) ,周志华[21] 给出了一个更为通俗的解释: 模型 A 在特定数据集表现优于模型 B,一定伴随着模型 A 在另一个数据集上表现不如模型 B.

1.1 预测模型分类

对既有方法进行梳理,有助于掌握各种模型的原理及优劣,从而为后续的算法创新和预测研究奠定坚实基础. 因此,基于不同的建模视角,本研究将常用的预测模型分为以下四个类别,并分别考察它们对 CPI 的预测能力: 一是使用单变量的时间序列模型,考察 CPI 序列的动态性质并进行预测,以此作为后续比较的基准; 二是以传统计量经济学的方法为基础,结合机器学习中的降维技术,构建正则化回归模型进行估计和预测; 三是从自变量集中提取共同因子,然后再将其加入回归模型,这种做法隐含着一个假设,即众多经济变量中存在着共同的信息且能够用于预测和分析; 四是采用集成算法,先通过一定的规则生成多个对 CPI 的预测,然后依据相应的集成策略将多个预测值进行组合. 方法的分类概要及其在大数据预测中的优缺点如图 1 所示.

一般地,用于对 CPI 进行向前 h 步预测的模型可以表示为

$$CPI_{t+h} = G_h(\boldsymbol{x}_t) + u_{t+h},\ h = 1,\ldots,H \quad t = 1,\ldots,T \tag{1}$$

其中 CPI_{t+h} 表示第 $t+h$ 期的 CPI 同比增长率; $\boldsymbol{x}_t = (x_{1t}, x_{2t}, \ldots, x_{nt})'$ 是用于预测 CPI 的 n 维向量,它包括 CPI 的滞后项、与 CPI 相关的自变量及其共同因子; $G_h(\boldsymbol{x}_t)$ 为特定模型下的预测函数.

表 1　模型分类及参数选择概要

Table 1 Model classification and parameter selection

模型类别	模型名称	参数及其选择依据
基准模型	随机游走(RW)	
	自回归(AR)	(1) 滞后阶数 p: BIC 准则
	不可观测成分随机波动模型(UCSV)	(1) 先验分布方差 $V\tau$, Vh: 参考 Stock 和 Watson[24] 设为 0. 12
正则化回归	岭回归(RR)	(1) 惩罚参数 λ: BIC 准则
	套索回归(LASSO)	(1) 惩罚参数 λ: BIC 准则
	自适应套索回归(adaLASSO)	(1) 惩罚参数 λ: BIC 准则; (2) 初始估计值 $\beta^*_{h,i}$: RR 估计值
	神经网络(Elastic Net)	(1) 惩罚参数 λ: BIC 准则; (2) L_1 正则化权重 α: 参考 Medeiros[23] 取 0. 5
	自适应神经网络(adaElastic Net)	(1) (2) 与 Elastic Net 保持一致; (3) 初始估计值 $\beta^*_{h,i}$: RR 估计值
因子模型	基础因子模型(Factor)	(1) 因子个数: 参考 Medeiros[23] 取 4; (2) 滞后阶数: BIC 准则
	提升因子模型(B. Factor)	(1) 更新步长 v: 参考 Bai 和 Ng[25] 取 0. 2; (2) 滞后阶数: 参考 Medeiros[23] 取 4
集成算法	全子集回归(CSR)	(1) 自变量集维数 $\tilde{K}$: 20; (2) 子集变量数 q: 4. $\tilde{K}$ 和 q 均参考 Medeiros[23]
	随机森林(RF)	(1) 决策树数量 N_{tree}: 500; (2) 每棵树变量数 m_{try}: K /3. 均为 RF 常用设定
	极端梯度提升(XGBoost)	(1) 决策树数量 B: 1 000: (2) 每棵树最大深度 $deep$: 4. 均为 XGBoost 常用设定

注: 1) RF 类、XGBoost 类衍生算法的参数及选择依据分别参照 RF 和 XGBoost; 2) 本研究将重点介绍 RF 和 XGBoost 的原理及其衍生算法思路,其余模型的介绍详见附录 B. 6.

按照以上分类,本研究选取了随机游走、弹性网络、因子模型和极端梯度提升等 13 个模型对 CPI 进行预测,并将这些方法归分为以下四个类别: 基准模型、正则项回归、因子模型和集成算法. 对于所选的模型,本研究都将采用滚窗和递归两种基本的框架进行估计和向前预测,以考察它们在不同期限下的预测表现③. 此外,本研究的主要模型方法及其所涉及的参数选择依据如表 1 所示.

1.2 机器学习及衍生算法

集成算法是机器学习的延伸,它首先训练出若干个个体学习器,然后将其按照一定的策略结合在一起,形成一个强学习器,以达到提升机器学习效果的目的. 以随机森林和极端梯度提升为代表,本研究将集成算法与宏观经济预测相结合,并基于控制变量的思想构造出衍生算法对 CPI 进行预测研究,下面将对它们的原理进行简要介绍.

③ 本研究在预测中使用了滚窗和递归两种估计框架,具体细节可见本文正文所提到的附录,有兴趣的读者可以和作者索要.

1.2.1　随机森林(random forest, RF)

随机森林是 Breiman[26] 在决策树的基础上提出的一种机器学习方法,最初主要作为一种分类技术用于统计推断,后面逐步运用于回归问题. RF 是基于装袋算法(bootstrap aggregating, bagging) 的一个经典模型,其思想是: 从原始的数据训练集中有放回地抽取 N_{tree} 个样本作为子训练集,并在每一个子训练集中随机抽取 m 个变量构造决策树,然后将每棵决策树结果的平均值作为最终预测. 这种在自助抽样(bootstrapping) 得到的不同子样本上训练单个学习器,然后对预测进行平均的过程就称为 bagging. 事实上,这里的每一棵决策树就相当于一个非参数模型,它使用递归的方式对变量空间进行划分,然后在每个空间内用局部预测来对非线性函数进行逼近. 而由若干决策树构成的随机森林,其随机性体现在两个方面: 一是在构建决策树时对训练集的数据点进行了随机抽样; 二是在进行节点分裂时考虑了变量的随机子集. RF 的具体算法及参数选择依据参见附录 B.3.

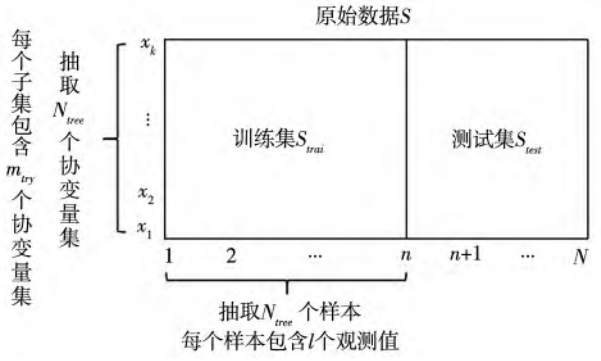

图 2　两个方向的随机抽样示意

Fig. 2 Schematic diagram of random sampling

1.2.2　极端梯度提升(XGBoost)

Chen 和 Guestrin[27] 在梯度提升算法的基础上提出了极端梯度提升算法(extreme gradient boosting, XGBoost) ,该算法主要运用于分类和回归问题. 利用滚窗将时间序列数据转换为适用于监督学习的形式,XGBoost 就可以用于对时间序列数据进行分析和预测. 其主要思想是: 运用训练集的数据进行学习,通过自变量的节点分裂来生成回归树,然后不断地生成下一棵树去拟合上一棵树的预测误差,最终得到 B 棵回归树. 对于一个需要进行预测的样本,依据其变量特征,它会在每一棵树上归属于一个叶片节点,每一个节点对应一个分数,将该样本在每棵树上得到的拟合值相加即得到最终预测值. XGBoost 的具体算法及参数选择依据参见附录 B.5.

1.2.3　衍生算法

随机森林和极端梯度提升主要在两个方面区别于其他的预测模型: 非线性的设定和非稀疏的变量集. 前者表示因变量与自变量以及不同自变量之间的复杂关系; 后者表示包含所有变量的数据集. 通过 LASSO、弹性网等正则化回归模型能够得到稀疏的系数矩阵,将剔除系数为 0 的变量后得到的数据集合称为“稀疏变量集”. 与之相对,未进行变量选择,保留所有原有变量的数据集合称为“非稀疏变量集”. 为了识别这两个特性对预测效果的影响,可以进一步使用衍生算法对 CPI 进行预测.

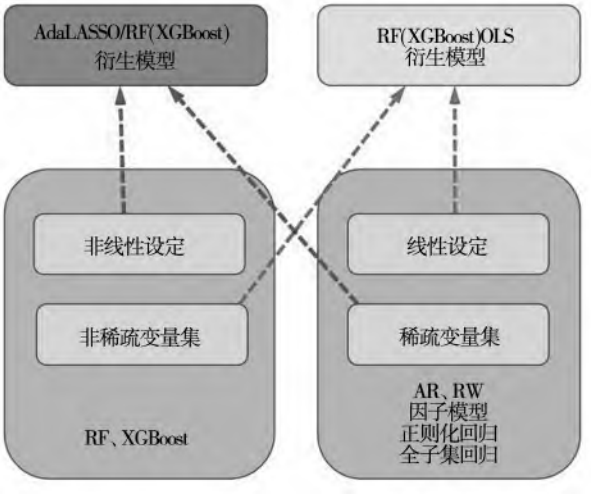

图 3　衍生算法思路示意

Fig. 3 Derivative algorithm ideas

以随机森林算法为例,本研究遵循控制变量的思想构造了 RF/OLS 和 adaLASSO/RF 两种衍生模型,分别用于分离非线性设定和非稀疏变量集这两个特性. 其中,RF/OLS 方法保持随机森林中的非稀疏变量集,然后使用 OLS 代替非线性设定,其基本思想为: 先利用随机森林的原理进行变量选择,然后在线性模型的框架下进行估计和预测. 通过对比随机森林和 RF/OLS 两种方法的预

测效果,可以识别出非线性设定在 CPI 预测中的重要性. 例如,如果 RF/OLS 模型得出的结果与 RF 差异较小,则可以认为非线性设定对 CPI 预测帮助不大,RF 的优势主要体现在非稀疏的变量集. 而如果 RF/OLS 的预测效果与 RF 相比较差,则可以认为非线性设定有重要作用. RF/OLS 的估计步骤参见附录 B.4.

另一方面,adaLASSO/RF 保留随机森林中的非线性设定,使用 adaLASSO 所得变量替代非稀疏变量集,以此识别非稀疏变量集在 CPI 预测中的重要性,其基本思想为: 首先运用 adaLASSO 的原理进行变量选择,然后把所有选择出的变量加入 RF 模型,即不再进行自变量的抽取,而是直接使用所有变量构建决策树. 同理,如果 adaLASSO/RF 的预测结果与 RF 差异较小,则可以认为非稀疏变量集对 CPI 预测帮助不大.

与 RF 类似,XGBoost 也可以通过使用 XGBoost/OLS 和 adaLASSO/XGBoost 来识别非线性设定和非稀疏变量集对 CPI 预测效果的影响. 其中,XGBoost/OLS 方法首先运用 XGBoost 构建决策树,然后选择重要性排名前 20 的自变量构建多元线性回归模型,最后使用 OLS 进行估计. 而 adaLASSO/XGBoost 则是使用 adaLASSO 算法构建稀疏的变量集,然后将这些选择出的变量加入 XGBoost 模型. 算法步骤和分析原理均与前文类似.

2　数据选取与描述

参照 McCracken 和 Ng[28] 搭建的 FRED-ME 数据库,本研究综合利用国家统计局、Choice 和 CEIC 数据库中的可得数据构建了中国月度宏观经济数据库 (CMED-MD). 如表 2 所示,该数据库选择了 239 个月度宏观经济指标,并将其划分为劳动力市场、房地产、消费等 9 个类别,总样本区间为 1991 年 1 月—2022 年 6 月. 此外,本研究在进行建模估计时,考虑了从所有变量中提取的 4 个主成分因子,所有变量和共同因子的 4 阶滞后项,以及 CPI 的 4 个自回归项. 因此,最终用于预测 CPI_{t+h} 的自变量集 x_t 实际包含了 $(239+4)\times(1+4)+4=1\,219$ 个潜在的预测变量.

表 2　变量分类概要

Table 2 Summary of variable classification

类别	编号	代表变量	类别	编号	代表变量
劳动力市场	A1 ~ A36	失业再就业人数、各行业从业人数	利率及汇率	F1 ~ F63	同业拆借利率、美元兑人民币汇率
房地产	B1 ~ B19	国房景气指数、房屋新开工面积	价格	G1 ~ G17	工业生产价格指数、出口价格指数
产出及收入	C1 ~ C29	克强指数、工业增加值	金融市场	H1 ~ H16	沪深 300 指数、平均市盈率
消费	D1 ~ D16	消费品零售总额、消费者信心指数	其他	I1 ~ I18	黄金储备、财政预算收入
货币市场	E1 ~ E25	货币供应、存款准备金率	—	—	—

决策树可以包容缺失值的存在,而随机森林本身也可以用于处理缺失值,故 RF、XGBoost 等方法可以对前端缺失值进行处理,从而允许不同样本长度的变量混合在一起进行建模分析; 但对于其他以计量方法为基础的模型,如 AR、LASSO 等,当时间序列数据前端存在缺失值时,将无法使用该序列进行估计. 由于部分变量起始时间较晚,本研究设置两个不同长度的样本以适应不同模型,具体如下: 长样本包含 239 个自变量和 1 个目标变量 CPI,区间为 1991 年 1 月—2022 年 6 月,其中选取 1991 年 1 月—2017 年 6 月的数据进行样本内拟合,长样本主要用于 RF、XGBoost 等方法; 短样本区间为 2002 年 1 月—2022 年 6 月,其中截取 2002 年 1 月—2017 年 6 月的数据进行样本内拟合,该样本删减了起始日期在 2002 年 1 月以后的变量,剩余 146 个自变量和 1 个目标变量

CPI,适用于本研究所有模型.

参考郑挺国等[29]的做法将所有变量均转换为平稳序列,并构造出长、短两个样本. 由此既可利用短样本来保证各个模型的正确估计和比较,又可利用长样本凸显出 RF、XGBoost 等方法对于缺失值的处理优势. 此外,本研究针对插值方法和数据平稳性进行了稳健性检验,结果详见附录 A.2和附录 A.3.

3　实证分析

本节将上述长短两个样本分别运用于各个模型进行建模和估计,从误差度量和预测能力对比检验两个方面考察了各个模型在不同预测期限下的表现; 然后利用 RF 和 XGBoost 的衍生算法,识别了非线性设定和非稀疏变量集在 CPI 预测中的作用; 最后,对自变量的重要性进行了测度和加总,由此对不同的变量类别进行了重要性排序.

3.1　预测效果对比

本研究以平方根均方误差(*RMSE*) 作为判别预测效果的主要指标,它能够衡量各模型在样本外窗口上的整体预测效果. 假设样本期为 $t = T_0, T_0+1, \ldots, T$,每一期的 CPI 同比增长率的真实值为 CPI_t,第 m 个模型的向前 h 步预测值为 $\widehat{CPI}_{m,t+h|t}$,则预测的 *RMSE* 可以表示为

$$RMSE_{m,h} = \sqrt{\frac{1}{T - T_0 + 1}\sum_{t=T_0}^{T}\left(CPI_{t+h} - \widehat{CPI}_{m,t+h|t}\right)^2} \tag{2}$$

表 3 给出了各模型对所有 h 步向前预测的 *RMSE* 结果. 如表 3 所示,由基准模型、正则化回归以及 CSR 估计所得到的 *RMSE* 基本呈现出随着预测期限 h 的增大而上升的趋势,这说明对大部分模型而言,长期预测比短期预测更为困难,但对于 RF 和 XGBoost 而言,它们的预测能力在不同预测期限中都较为稳定. 就预测准确性来看,正则化回归和集成算法整体上优于基准模型和因子模型,其中 RF 和 XGBoost 在大多数期限中的预测效果明显占优于其他预测方法,并且随着 h 的增大,其优势越明显; 而因子模型的表现相对最差,这表明该类模型并不适用于高维数据集下的变量预测,后文通过计算因子类变量的重要性亦可佐证该结论. 最后,各个模型在滚窗和递归两种估计框架下的预测效果差异不大,即估计框架并不影响模型的优劣排序,由此表明本研究的结论具有较强的稳健性.

上述预测效果的比较亦可通过相对 *RMSE* 更直观地进行表示. 以 RW 为比较基准,选取 AR、UCSV、adaLASSO、adaElastic Net、Factor、RF 和 XGBoost 这 7 个在各自类别中表现较优的模型,计算其在短样本滚窗预测中的相对 *RMSE*. 具体公式如下

$$\text{相对 } RMSE(m \mid h) = \frac{RMSE(m \mid h) - RMSE_{\text{基准}}(m \mid h)}{RMSE_{\text{基准}}(m \mid h)} \tag{3}$$

如图 4 所示,在短期预测($h = 1, 2$) 中,不同的模型的预测效果并未展现出明显差异,这可能是因为通货膨胀具有较强的惯性,当预测期限较短时,预测期的 CPI 会显著地依赖于其前几期的自回归项,其他变量在 CPI 预测中的作用相对较弱,无法凸显出 RF 和 XGBoost 中非线性设定和非稀疏变量集的优势. 随着预测期限的增加($h \geqslant 3$) ,RF 和 XGBoost 的相对 *RMSE* 不断下降,相比于其他模型的优势逐渐凸显,这可能是由于 CPI 对于自回归项的依赖性减弱,从而使得其他自变量在预测中的相对作用不断增强,此时非线性关系和非稀疏变量集将发挥出重要作用. 这也与此前的相关研究的结果一致[15, 17].

进一步地,可以通过 Quaedvlieg[30]提出的多步长预测能力占优检验(multi-horizon superior predictive ability test, MH-SPA) 比较上述模型在各个预测步长下的综合预测能力,结果如表 4 所示. 综合考虑 1 步 ~ 12 步向前预测的结果,XGBoost 整体上优于其他模型,RF 次之,因子类模型则相对较差.

表 3　向前 1 步至 12 步预测的 *RMSE*

Table 3 *RMSE* for step 1-step to 12-step forward forecasting

样本	模型	滚窗预测												递归预测											
		1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12
短样本	RW	0.53	0.82	1.04	1.18	1.28	1.39	1.49	1.62	1.74	1.83	1.90	1.95	0.53	0.82	1.04	1.18	1.28	1.38	1.49	1.62	1.74	1.83	1.90	1.95
	AR	0.55	0.86	1.08	1.19	1.24	1.28	1.34	1.40	1.46	1.52	1.55	1.55	0.53	0.83	1.03	1.15	1.18	1.22	1.27	1.33	1.38	1.44	1.46	1.46
	UCSV	0.53	0.83	1.04	1.18	1.28	1.38	1.49	1.62	1.74	1.83	1.89	1.94	0.53	0.82	1.04	1.18	1.28	1.38	1.49	1.62	1.74	1.83	1.89	1.94
	RR	1.30	1.30	1.30	1.29	1.29	1.29	1.30	1.30	1.31	1.32	1.32	1.33	1.25	1.25	1.25	1.24	1.24	1.24	1.25	1.26	1.27	1.28	1.29	1.29
	LASSO	0.53	0.78	0.88	1.01	1.04	1.02	1.02	1.06	1.15	1.20	1.26	1.30	0.53	0.77	0.87	0.94	1.01	1.04	1.04	1.05	1.20	1.24	1.31	1.22
	adaLASSO	0.53	0.77	0.90	0.99	1.03	0.94	1.01	0.93	0.91	0.90	0.89	0.93	**0.52**	0.75	0.86	0.98	1.07	1.04	1.02	0.97	0.94	0.90	0.84	0.86
	Elastic Net	0.57	0.81	0.89	0.99	1.02	1.00	0.97	104	1.10	1.06	1.10	1.12	0.53	0.74	0.86	0.94	0.99	1.06	0.98	1.00	1.09	1.10	1.09	1.07
	adaElastic Net	0.53	0.81	0.90	1.01	1.02	0.97	1.02	1.00	0.93	0.90	0.84	0.87	0.54	0.74	0.84	0.95	1.03	1.02	1.01	0.97	0.97	0.88	0.86	0.86
	Factor	0.52	0.79	0.98	1.08	1.15	1.24	1.28	1.32	1.39	1.33	1.37	1.31	0.52	0.79	0.97	1.07	1.12	1.18	1.23	1.29	1.27	1.31	1.29	1.28
	B. Factor	1.39	1.42	1.36	1.42	1.45	1.44	1.45	1.48	1.55	1.55	1.57	1.51	1.27	1.27	1.29	1.24	1.23	1.23	1.25	1.26	1.30	1.37	1.36	1.35
	CSR	0.56	0.84	1.02	1.13	1.17	1.20	1.25	1.32	1.49	1.62	1.71	1.75	0.55	0.81	1.01	1.11	1.16	1.20	1.25	1.32	1.37	1.48	1.51	1.47
	RF	0.68	0.79	0.84	0.92	0.94	0.97	0.97	0.95	0.93	0.90	0.87	0.78	0.65	0.78	0.84	0.90	0.92	0.92	0.91	0.91	0.89	0.87	0.83	0.77
	XGBoost	0.60	**0.70**	**0.76**	**0.80**	**0.80**	0.91	**0.86**	**0.85**	0.88	**0.88**	0.84	**0.73**	0.59	**0.69**	0.82	**0.81**	**0.84**	0.87	**0.82**	**0.79**	**0.78**	**0.77**	**0.68**	**0.64**
	adaLASSO/RF	0.64	0.79	0.85	0.92	0.97	0.95	0.99	0.98	0.96	0.94	0.88	0.81	0.62	0.78	0.83	0.97	0.97	0.93	0.95	0.98	1.00	1.02	0.92	0.82
	RF/OLS	0.53	0.77	0.95	1.08	1.15	1.17	1.15	1.10	1.00	0.93	0.96	0.95	0.53	0.80	0.97	1.08	1.13	1.19	1.21	1.19	1.09	0.99	0.95	0.92
	adaLASSO/XGB	0.71	0.84	0.83	0.84	0.91	**0.90**	0.87	0.91	**0.86**	0.89	**0.79**	0.77	0.74	0.81	0.79	0.97	0.93	0.85	0.83	0.92	0.92	0.94	0.83	0.75
	XGB/OLS	0.68	0.87	1.12	1.23	1.23	1.42	1.32	1.27	1.45	1.24	1.20	1.29	0.66	0.84	0.89	1.01	1.24	1.53	1.59	1.56	1.53	1.47	1.09	1.03
长样本	RW	0.53	0.82	1.04	1.18	1.28	1.38	1.49	1.62	1.74	1.83	1.90	1.95	0.53	0.82	1.04	1.18	1.28	1.38	1.49	1.62	1.74	1.83	1.90	1.95
	AR	0.52	0.82	1.04	1.18	1.25	1.31	1.37	1.47	1.55	1.60	1.64	1.64	0.53	0.85	1.10	1.27	1.38	1.50	1.64	1.79	1.95	2.07	2.17	2.25
	UCSV	0.54	0.83	1.04	1.18	1.28	1.39	1.50	1.63	1.74	1.83	1.90	1.95	0.53	0.82	1.04	1.18	1.28	1.38	1.49	1.62	1.74	1.83	1.90	1.95
	RF	0.70	0.81	0.87	0.90	0.91	0.92	0.91	0.91	0.89	0.88	0.85	0.80	0.74	0.84	0.90	0.92	0.93	0.91	0.89	0.87	0.88	0.87	0.83	0.79
	XGBoost	0.56	**0.76**	**0.78**	**0.85**	**0.84**	**0.87**	**0.83**	**0.81**	**0.81**	**0.74**	**0.65**	**0.59**	0.56	**0.74**	**0.78**	**0.88**	0.89	0.84	0.86	**0.77**	**0.72**	**0.81**	**0.72**	**0.66**
	adaLASSO/RF	0.59	1.14	1.34	1.39	1.00	0.97	0.99	1.06	0.99	0.93	0.87	0.82	0.57	0.85	0.99	0.91	**0.85**	0.85	0.89	0.92	0.86	0.87	0.81	0.73
	RF/OLS	**0.51**	0.80	1.02	1.13	1.19	1.24	1.25	1.31	1.33	1.36	1.36	1.41	**0.50**	0.81	1.03	1.16	1.23	1.29	1.36	1.46	1.54	1.58	1.56	1.56
	adaLASSO/XGB	0.61	0.86	1.05	1.06	1.07	0.95	0.92	1.01	0.98	1.03	0.84	0.66	0.63	0.81	1.04	0.97	0.94	**0.82**	**0.80**	0.87	0.79	0.86	0.83	0.69
	XGB/OLS	0.53	0.77	1.18	1.30	1.42	1.36	1.44	1.73	1.73	1.87	1.89	2.07	0.52	0.87	1.18	1.42	1.97	1.78	1.85	1.87	2.13	2.21	2.19	2.74

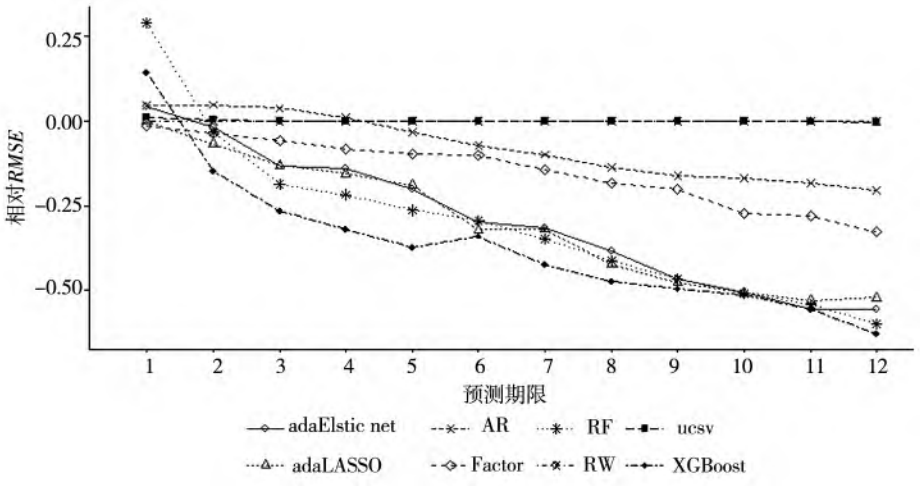

图 4　短样本滚窗预测中的相对 *RMSE*

Fig. 4 Relative *RMSE* for rolling window forecasting in short sample

表 4　MH-SPA 检验结果

Table 4 Test results of MH-SPA

方法	RW	AR	UCSV	adaLASSO	adaElastic Net	Factor	RF	XGBoost
RW		0.05	0.29	0.01	0.02	0.02	0.02	0.01
AR	0.94		0.94	0.01	0.01	0.06	0.00	0.00
UCSV	0.69	0.05		0.01	0.02	0.03	0.01	0.01
adaLASSO	0.99	0.99	0.98		0.71	0.98	0.43	0.04
adaElastic Net	0.98	0.99	0.98	0.31		0.98	0.39	0.08
Factor	0.98	0.94	0.97	0.02	0.03		0.01	0.01
RF	0.99	1.00	0.99	0.58	0.66	0.99		0.03
XGBoost	0.99	1.00	0.99	0.95	0.92	0.99	0.97	

注：MH-SPA 检验是对 1 步～12 步向前预测误差的联合检验，表格中的数字表示检验的 p 值，目标模型相对于第一列中比较模型的 p 值小于 0.01，0.05 和 0.10 时，分别表明在 1%，5% 和 10% 置信水平下目标模型的预测效果优于基准比较模型.

3.2　非线性/非稀疏与预测能力

为进一步探究非线性设定和非稀疏变量集在 CPI 预测中的重要性，本研究将 RF 和 XGBoost 两种算法分解为变量选择和模型估计两个步骤，并在这两个步骤中分别引入 adaLASSO 和 OLS 两种方法构造出相应的衍生模型.

本研究考察了 CPI 对变量二次项、交互项的回归系数，据此说明利用大数据对 CPI 进行预测时自变量间的复杂关系. 图 5 所示为回归系数的 p 值分布直方图，可见 CPI 与大量的变量交互项及二次项之间具有显著的关系，即在高维数据集中，变量间存在着复杂的非线性关系，RF 和 XGBoost 通过灵活的模型设定可能捕捉到了这些关系，并对 CPI 进行了更加有效的预测.

为验证上述结论，表 5 给出了 RF、XGBoost 及其两类衍生模型的样本外预测 *RMSE*. 如表所示，对比 RF、XGBoost 及其所对应的 adaLASSO 衍生模型，可以识别出非稀疏变量集对预测效果的影响. 在短样本中，XGBoost 显著占优于 adaLASSO/XGB，而 RF 与 adaLASSO/RF 未表现出明显差异，这可能是由于在构建短样本的过程中删减了大量存在前端缺失值的变量，使得数据集维度大幅降低，RF 算法中的非稀疏变量集不容易展现出优势. 但在维度较高的长样本中，XGBoost 和 RF 的预测效果都明显优于其对应的 adaLASSO 衍生模型，这表明在大型数据集

下,非稀疏性能够充分地利用数据信息,从而提高 CPI 的预测精度. 另一方面,对比 RF、XGBoost 及其所对应的 OLS 线性衍生模型,可以识别出非线性设定对预测效果的影响. 无论是在短样本还是长样本中,RF 和 XGBoost 都分别显著地占优于 RF/OLS 和 XGBoost/OLS,这表明模型的非线性设定有助于提高 CPI 的预测精度. 因此,非线性设定和非稀疏变量集是 RF 和 XGBoost 区别于其他预测模型两个特征,也是他们的优势所在.

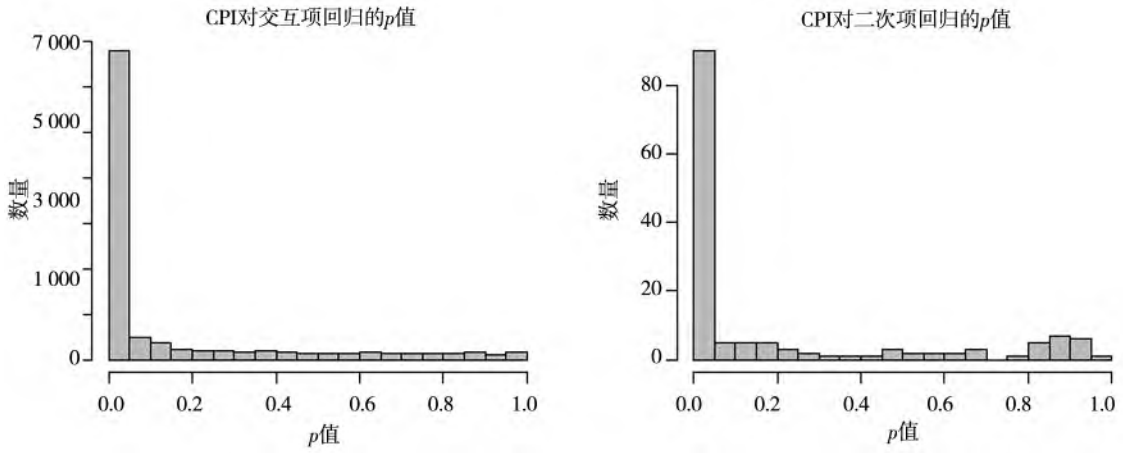

图 5 变量间的非线性关系检验

Fig. 5 Nonlinear relationship testing between variables

表 5 RF、XGBoost 及其衍生模型的预测效果对比

Table 5 Comparison of forecasting effects of RF, XGBoost, and their derived models

样本	模型	滚窗估计 *RMSE*												MH-SPA 检验		
		1	2	3	4	5	6	7	8	9	10	11	12	M	adaL/M	M/OLS
短	RF	0.68	0.79	**0.84**	**0.92**	**0.94**	**0.97**	**0.97**	**0.95**	**0.93**	**0.90**	**0.87**	**0.78**		0.68	0.94
	adaLASSO/RF	0.64	0.79	0.85	0.92	0.97	0.95	0.99	0.98	0.96	0.94	0.88	0.81	0.33		0.99
	RF/OLS	**0.53**	**0.77**	0.95	1.08	1.15	1.17	1.15	1.10	1.00	0.93	0.96	0.95	0.04	0.02	
	XGBoost	**0.60**	**0.70**	**0.76**	**0.80**	**0.80**	0.91	**0.86**	**0.85**	0.88	**0.88**	0.84	**0.73**		0.99	1.00
	adaLASSO/XGB	0.71	0.84	0.83	0.84	0.91	**0.90**	0.87	0.91	**0.86**	0.89	**0.79**	0.77	0.01		0.98
	XGB/OLS	0.68	0.87	1.12	1.23	1.23	1.42	1.32	1.27	1.45	1.24	1.20	1.29	0.00	0.02	
长	RF	0.70	0.81	**0.87**	**0.90**	**0.91**	**0.92**	**0.91**	**0.91**	**0.89**	**0.88**	**0.85**	**0.80**		0.94	1.00
	adaLASSO/RF	0.59	1.14	1.34	1.39	1.00	0.97	0.99	1.06	0.99	0.93	0.87	0.82	0.05		1.00
	RF/OLS	**0.51**	**0.80**	1.02	1.13	1.19	1.24	1.25	1.31	1.33	1.36	1.36	1.41	0.00	0.00	
	XGBoost	0.56	**0.76**	**0.78**	**0.85**	**0.84**	**0.87**	**0.83**	**0.81**	**0.81**	**0.74**	**0.65**	**0.59**		0.98	1.00
	adaLASSO/XGB	0.61	0.86	1.05	1.06	1.07	0.95	0.92	1.01	0.98	1.03	0.84	0.66	0.01		1.00
	XGB/OLS	**0.53**	0.77	1.18	1.30	1.42	1.36	1.44	1.73	1.73	1.87	1.89	2.07	0.00	0.01	

图 6 更直观地给出了 RF、XGB 与衍生模型的预测能力比较. 除表 5 所得结论外,可以进一步总结得出以下两点: adaLASSO 衍生模型的估计效果优于 OLS 衍生模型,因此相对于非稀疏变量集,非线性设定在 CPI 预测中可能更为重要; 随着预测期限 h 的增大,RF 和 XGBoost 相对于衍生模型的优势越明显,这表明在中长期预测中,非线性设定和非稀疏变量集这两个特性将变得更为重要.

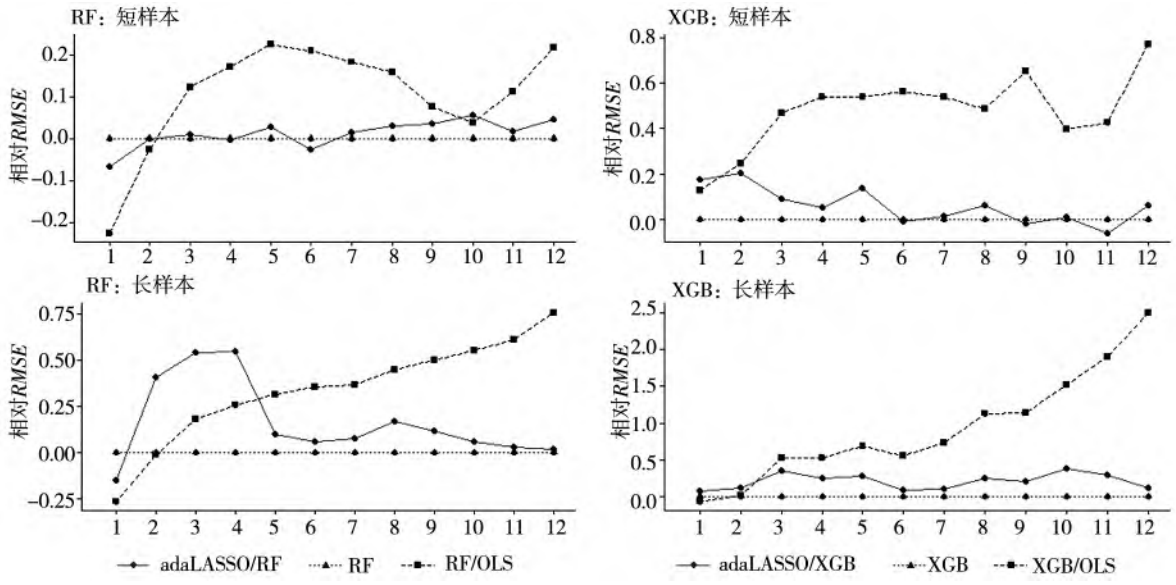

图 6　RF、XGBoost 及其衍生模型的相对 *RMSE*

Fig. 6 Relative *RMSE* of RF, XGBoost and their derived models

3.3　变量重要性

高度非线性的设定和较高维度的数据集使得模型中的变量关系更加复杂，机器学习虽然能够在多个层面进行抽象推断，从而处理因变量和自变量之间这种复杂的关系并得到较高的预测精度，但却面临着黑箱问题，即无法得知模型的机制和变量间的真实关系，这一点使得机器学习饱受诟病. 在 RF 和 XGBoost 算法下，可以从变量重要性的角度增强模型的透明度和解释力. 基于此，本节选取 Ridge、adaLASSO、adaElastic Net、RF 和 XGBoost 五个预测模型，测度了预测 CPI 所用的 11 类变量在这些模型中的重要性，并就自变量对因变量 CPI 的潜在影响机制进行了探索和分析. 选择这几个模型的原因是：Ridge、adaLASSO、adaElastic Net 分别代表了正则化模型中的三种不同惩罚函数，且在同类模型中表现较优. 各模型中重要性的测度方法见附录 B.7.

本研究将 CPI 的自回归项和共同因子作为两个单独的类别，加上表 2 所示的 9 个类别，可得到 11 个类别的变量集，图 7 展示了各个类别的变量在向前 1 步 ~ 12 步预测中的重要性. 如图所示，在不同期限中，adaLASSO 和 adaElastic Net 的变量重要性展现出了较大差异，在这两个模型的短期预测（$h=1, 2$）中，CPI 的自回归项发挥着重要作用，随着预测期限的增加，其重要性不断减弱，而其他变量的重要性逐渐凸显，这与前文的分析一致. 而 Ridge、RF 和 XGBoost 的变量重要性则较为稳定，这意味着在这三个模型中，变量间的关系结构不会随着预测期限的增加而发生变化. 此外，共同因子在所有 5 个模型中都未呈现出显著的重要性，这可以对前文因子类模型预测效果不佳的结果作出解释. 在大数据环境下，将变量进行高度聚合的做法是不合理的. 因为随着变量个数的增加，变量集中的特异性成分占比上升，所能提取出的共同信息减少，使用这样的共同因子将难以进行有效预测.

在 RF 和 XGBoost 两个模型中，各类别变量的重要性排序相对一致. 其中，CPI 自回归项、价格、利率汇率、就业和货币位列前 5. 为说明该排序的合理性，下面将进一步分析这 5 类变量对 CPI 的潜在影响机制.

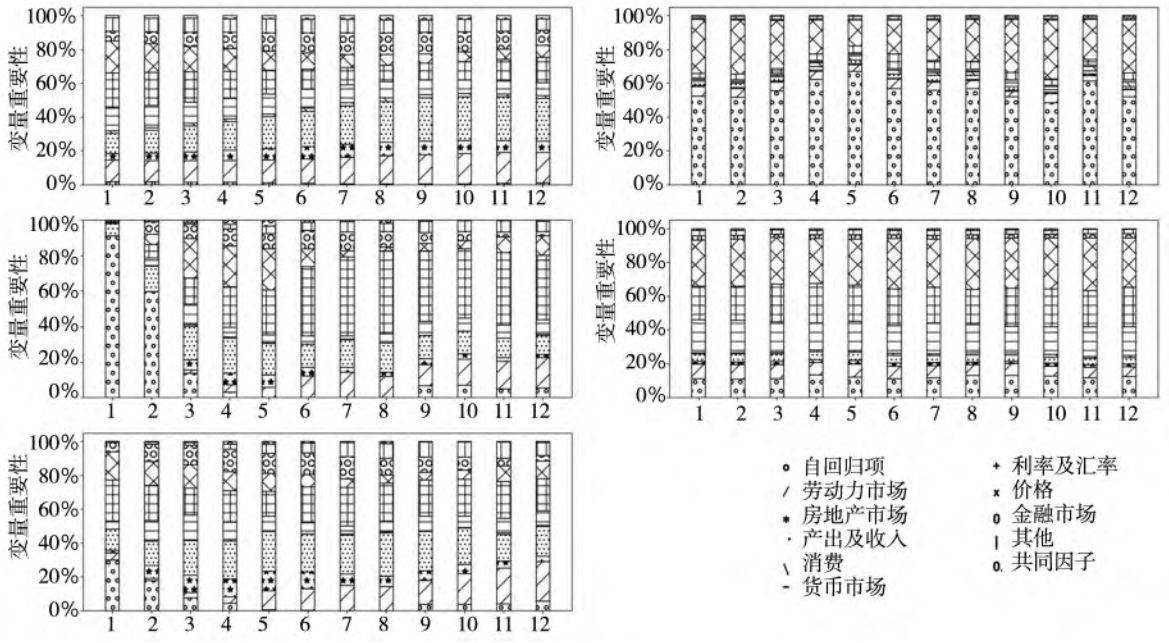

图 7　变量重要性

Fig. 7 Variable importance

注: 该图的彩色版本见本文附录.

首先,关于 CPI 自回归项. 凯恩斯主义提出了“价格粘性”和“工资粘性”的重要概念,它们认为价格和工资会较强地依赖于前期的价格及工资水平,且在短期内不会发生急剧的变化,而由于这两种粘性的存在,CPI 也会较强地依赖于其滞后值,这种现象被称为“通胀惯性”或“通胀粘性”. 此前的实证研究表明,我国的通胀水平具有较强的惯性特征[31, 32],这种惯性的存在是由于工资及价格的粘性和人们对于通胀的预期,导致 CPI 较强地依赖于其滞后值. 因此,当对 CPI 进行预测时,其自回归项能够发挥重要作用,这一结论符合相关的经济理论与实证经验.

其次,对于价格类变量,CPI 是反映居民消费品及服务价格水平的重要宏观经济指标,是关于食品烟酒、衣着、居住等多种商品和服务价格的加权平均值. 从 CPI 的统计性质上看,经济中价格的变动会对 CPI 产生直接影响. 从实证经验上看,价格类变量中包含的 *PPI*、*PPIRM*、*RPI* 等指标已被证实与 CPI 紧密相关且对 CPI 有较强的预测能力[33 - 35].

第三,关于汇率类变量. 根据汇率传递(exchange rate pass-through) 理论,汇率变动会对一国进出口商品价格和国内价格产生影响. 传统的汇率传递理论以“一价定律”为基础,认为在完全竞争市场上,厂商存在着套利行为,因而汇率变动会引起进出口价格乃至国内商品价格的变动. Krugman 在其基础上提出了汇率的不完全传递,他指出由于厂商存在市场定价行为,汇率变动时厂商能够调整产品成本从而稳定出口价格,因此汇率变动仅部分地传递至进出口价格和国内商品价格中. 这一类有关汇率传递的理论均认为,汇率变动与国内价格水平之间存在着对应关系. 针对我国实际情况的实证研究也表明,人民币汇率确实会对国内价格水平产生传递效应[36 - 38],这是因为在开放经济中利率及汇率的变动会影响人们的生产、生活成本,进而影响贸易价格、商品价格和工资水平等.

第四,对于就业市场的相关变量,Phillips[39] 提出的菲利普斯曲线表明失业率与工资变动之间存在着反向关系. 随后,Samuelson 和 Solow[40] 提出根据成本推动的通货膨胀理论,货币工资的提高是通货膨胀的主要诱发因素,即“工资 - 物价螺旋”. 因此菲利普斯曲线成为了描述失业率和通货膨胀之间反向关系的曲线. 基于中国数据的实证研究表明,菲利普斯曲线在我国经济中确实存在[41, 42],因此,我国就业部门的数据对于 CPI

的预测能够发挥重要作用.

最后,针对货币类变量. Friedman[43] 提出了现代货币数量论,并指出“通货膨胀无论何时何地都是一种货币现象”,即通货膨胀与货币供给需求之间存在着紧密的关系. 一方面,当货币流通速度和实际产出水平外生给定时,货币供给的变动将直接导致价格的变动,从而影响名义变量的数值. 另一方面,通货膨胀作为货币政策分析中的重要变量,其变动情况也将成为货币政策制定、宏观调控的核心参考要素之一. 事实上,有关我国的实证经验表明,央行的货币政策、货币供给情况与通货膨胀率密切相关[44-46]. 因此,有效地利用货币类的数据能够提高对通货膨胀的预测能力.

4　结束语

CPI 与通货膨胀紧密相关,是反映一国物价水平的关键指标,对 CPI 进行准确预测能够为国家的政策制定、宏观调控和国民经济核算提供有效的帮助. 大数据的发展带来了更快的样本更新速度和更加多样化的样本信息,恰当地选择预测模型能够充分地发挥大数据的优势,提高预测精度,为政策制定者和市场参与者提供更加可靠的预测信息. 基于此,本研究构建了一个覆盖 9 个类别,包括 239 个变量的中国月度宏观经济数据库,并将其用于 13 个不同的模型来对 CPI 进行预测研究. 通过对样本外预测效果进行对比,分析了各个模型的预测能力,同时基于控制变量的思想构造了衍生算法,揭示了非线性设定和非稀疏变量集在 CPI 预测中的重要作用,并对各类别的变量在 CPI 预测中的重要性进行了排序.

本研究的主要结论可以概括如下: 首先,对于大部分模型而言,长期预测比短期预测更为困难,但 RF 和 XGBoost 的预测能力在不同预测期限下较为稳定. 具体而言,在短期预测中,除因子模型外的其他模型均未展现出明显差异,而在中长期预测中,RF 和 XGBoost 在向前 3 步 ~ 12 步预测中明显占优,且随着预测期限的增加,其优势越明显. 综合短期和中长期预测,正则化回归和集成算法在整体上优于基准模型和因子模型,其中 XGBoost 整体上最优,RF 次之,因子模型最差. 其次,通过构造衍生算法进行深入分析,可以发现 RF 和 XGBoost 的预测能力优势体现在非线性的模型设定和非稀疏变量集两个方面,且随着预测期限增加,这两个特征的重要性将不断增强. 第三,adaLASSO 和 adaElastic Net 的变量重要性在不同期限中有着较大差异. 在短期预测中,CPI 的自回归项发挥着重要作用,而随着预测期限的增加,其重要性不断减弱,其他类别变量的重要性逐渐凸显. 最后,RF 和 XGBoost 的变量重要性在不同期限中较为稳定,CPI 自回归项、价格、利率汇率、就业和货币在 CPI 的预测中发挥着重要作用,而共同因子在其中贡献最小.

相比于以往研究,本研究具有以下四个方面的优势: 一是结合大数据背景,所用的数据集覆盖了宏观经济的各个方面,充分地利用了数据信息; 二是综合考虑了从短期到长期的预测,对比了各个模型在不同期限中的预测表现,使得结果更具有可信度; 三是考虑了模型的非线性设定和非稀疏变量集,使得模型更加符合实际,由此得出了更加精确的预测值; 四是从衍生算法和变量重要性两个方面增强了机器学习的解释性,揭示了在 CPI 预测中起到重要作用的模型特征和变量类别,克服了机器学习的黑箱难题,为作用机制和相关理论的分析提供了一定的参考. 因此,本研究对于现实的政策制定和经济分析,以及未来的大数据运用和预测模型发展均具有重要的理论和实践意义.

参 考 文 献:

[1]刘　汉, 刘金全. 中国宏观经济总量的实时预报与短期预测——基于混频数据预测模型的实证研究[J]. 经济研究, 2011, 46(3): 4-17.
Liu Han, Liu Jinquan. Nowcasting and short-term forecasting of Chinese macroeconomic aggregates: Based on the empirical study of MIIDAS model[J]. Economic Research Journal, 2011, 46(3): 4-17. (in Chinese)

[2]郑挺国, 尚玉皇. 基于金融指标对中国 GDP 的混频预测分析[J]. 金融研究, 2013, (9): 16-29.

Zheng Tingguo, Shang Yuhuang. Mixed frequency forecasting analysis of China's GDP based on financial indicators [J]. Financial Research, 2013, (9): 16 – 29. (in Chinese)

[3] 龚玉婷, 陈 强, 郑 旭. 基于混频模型的 CPI 短期预测研究 [J]. 统计研究, 2014, 31(12): 25 – 31.
Gong Yuting, Chen Qiang, Zheng Xu. Short-term forecasting of CPI based on MIDAS models [J]. Statistical Research, 2014, 31(12): 25 – 31. (in Chinese)

[4] 许启发, 卓杏轩, 蒋翠侠. 反向有约束混频数据模型的市场化利率预测 [J]. 管理科学学报, 2019, 22(10): 55 – 71.
Xu Qifa, Zhuo Xingxuan, Jiang Cuixia. Predicting market interest rates via reverse restricted MIDAS model [J]. Journal of Management Sciences in China, 2019, 22(10): 55 – 71. (in Chinese)

[5] 曾 波, 刘思峰, 方志耕, 等. 灰色组合预测模型及其应用 [J]. 中国管理科学, 2009, 17(5): 150 – 155.
Zeng Bo, Liu Sifeng, Fang Zhigeng, et al. Grey combined forecast models and its application [J]. Chinese Journal of Management Science, 2009, 17(5): 150 – 155. (in Chinese)

[6] Norouzi N, Fani M. Black gold falls, black plague arise: An OPEC crude oil price forecast using a gray prediction model [J]. Upstream Oil and Gas Technology, 2020, (5): 100015.

[7] Elliott G, Timmermann A. Optimal forecast combinations under general loss functions and forecast error distributions [J]. Journal of Econometrics, 2004, 122(1): 47 – 49.

[8] 张新雨, 邹国华. 模型平均方法及其在预测中的应用 [J]. 统计研究, 2011, 28(6): 97 – 102.
Zhang Xinyu, Zou Guohua. Model averaging method and its application in forecast [J]. Statistical Research, 2011, 28(6): 97 – 102. (in Chinese)

[9] 徐小峰, 余乐安, 林姿汝, 等. 基于特征融合的生鲜商品短期销量组合预测 [J]. 管理科学学报, 2022, 25(12): 102 – 123.
Xu Xiaofeng, Yu Le'an, Lin Ziru, et al. Combination forecasting of short-term sales for fresh products based on feature fusion [J]. Journal of Management Sciences in China, 2022, 25(12): 102 – 123. (in Chinese)

[10] Chen B, Maung K. Time-varying forecast combination for high-dimensional data [J]. Journal of Econometrics, 2023, 237(2): 105693.

[11] 刘涛雄, 徐晓飞. 互联网搜索行为能帮助我们预测宏观经济吗? [J]. 经济研究, 2015, 50(12): 68 – 83.
Liu Taoxiong, Xu Xiaofei. Can Internet search behavior help to forecast the macro economy [J]. Economic Research Journal, 2015, 50(12): 63 – 83. (in Chinese)

[12] 王 娜. 基于大数据的碳价预测 [J]. 统计研究, 2016, 33(11): 56 – 62.
Wang Na. Forecasting of carbon price based on big data [J]. Statistical Research, 2016, 33(11): 56 – 62. (in Chinese)

[13] 董 莉, 彭凯越, 唐晓彬. 大数据背景下的 CPI 实时预测研究 [J]. 调研世界, 2017, (8): 51 – 54.
Dong Li, Peng Kaiyue, Tang Xiaobin. Research on CPI nowcasting in the context of big data [J]. The World of Survey and Research, 2017, (8): 51 – 54. (in Chinese)

[14] Giannone D, Lenza M, Primiceri G E. Economic predictions with big data: The illusion of sparsity [J]. Econometrica, 2021, 89(5): 2409 – 2437.

[15] Milunovich G. Forecasting Australia's real house price index: A comparison of time series and machine learning methods [J]. Journal of Forecasting, 2020, 39(7): 1098 – 1118.

[16] Rodriguez-Vargas A. Forecasting Costa Rican inflation with machine learning methods [J]. Latin American Journal of Central Banking, 2020, 1(1 – 4): 100012.

[17] Kohei M. Macroeconomic forecasting using factor models and machine learning: An application to Japan [J]. Journal of the Japanese and International Economies, 2020, 58: 101104.

[18] Du Y, Cai Y, Chen M, et al. A novel Divide-and-Conquer model for CPI prediction using ARIMA, Gray model and BPNN [J]. Procedia Computer Science, 2014, 31: 842 – 851.

[19] 薛 晔, 蔺琦珠, 任 耀. 我国通货膨胀风险的预测模型——基于决策树 – BP 神经网络 [J]. 经济问题, 2016, (1): 82 – 89.
Xue Ye, Lin Qizhu, Ren Yao. Research on inflation risk forecast of China based on the big-data technology [J]. On Economic Problems, 2016, (1): 82 – 89. (in Chinese)

[20] Wolpert D H. The lack of a priori distinctions between learning algorithms [J]. Neural Computation, 1996, 8(7): 1341 –

1390.

[21]周志华. 机器学习[M]. 北京：清华大学出版社，2016.
Zhou Zhihua. Machine Learning[M]. Beijing: Tsinghua University Press, 2016. (in Chinese)

[22]Goulet Coulombe P, Leroux M, Stevanovic D, et al. How is machine learning useful for macroeconomic forecasting? [J]. Journal of Applied Econometrics, 2022, 37(5): 920 – 964.

[23]Medeiros M C, Gabriel F R V, Veiga Á, et al. Forecasting inflation in a data-rich environment: The benefits of machine learning methods[J]. Journal of Business & Economic Statistics, 2021, 39(1): 98 – 119.

[24]Stock J H, Watson M W. Why has U.S. inflation become harder to forecast[J]. Journal of Money, Credit and Banking, 2007, 39(7): 3 – 33.

[25]Bai J, Ng S. Boosting diffusion indexes[J]. Journal of Applied Econometrics, 2009, 24(4): 607 – 629.

[26]Breiman L. Random Forests[J]. Machine Learning, 2001, 245: 5 – 32.

[27]Chen T, Guestrin C. XGBoost: A scalable tree boosting system[C]. Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016.

[28]McCracken M W, Ng S. FRED-MD: A monthly database for macroeconomic research[J]. Journal of Business & Economic Statistics, 2016, 34(4): 574 – 589.

[29]郑挺国，范馨月，靳　炜，等. 通胀预期形成与信息黏性特征：基于媒体新闻视角[J]. 世界经济，2023，46(4)：60 – 82.
Zheng Tingguo, Fan Xinyue, Jin Wei, et al. Inflation expectations and information rigidity: An analysis from the media news perspective[J]. The Journal of World Economy, 2023, 46(4): 60 – 82. (in Chinese)

[30]Quaedvlieg R. Multi-Horizon forecast comparison[J]. Journal of Business & Economic Statistics, 2021, 39(1): 40 – 53.

[31]陈彦斌. 中国新凯恩斯菲利普斯曲线研究[J]. 经济研究，2008，43(12)：50 – 64.
Chen Yanbin. Research on New Keynesian Phillips Curve in China[J]. Economic Research Journal, 2008, 43(12): 50 – 64. (in Chinese)

[32]何启志，范从来. 中国通货膨胀的动态特征研究[J]. 经济研究，2011，46(7)：91 – 101.
He Qizhi, Fan Conglai. The dynamic characteristics of Chinese inflation[J]. Economic Research Journal, 2011, 46(7): 91 – 101. (in Chinese)

[33]Cushing M J, McGarvey M G. Feedback between wholesale and consumer price inflation: A reexamination of the evidence [J]. Southern Economic Journal, 1990, 56(4): 1059 – 1729.

[34]刘　敏，张艳丽，杨延斌. PPI 与 CPI 关系探析[J]. 统计研究，2005，(2)：24 – 28.
Liu Min, Zhang Yanli, Yang Yanbin. Analysis of the relationship between PPI and CPI[J]. Statistical Research, 2005, (2): 24 – 28. (in Chinese)

[35]杨　宇，陆奇岸. CPI、RPI 与 PPI 之间关系的实证研究——基于 VAR 模型的计量经济分析[J]. 价格理论与实践，2009，(5)：57 – 58.
Yang Yu, Lu Qi'an. Empirical study on the relationship between CPI, RPI, and PPI: An econometric analysis based on VAR model[J]. Price: Theory & Practice, 2009, (5): 57 – 58. (in Chinese)

[36]卜永祥. 人民币汇率变动对国内物价水平的影响[J]. 金融研究，2001，(3)：78 – 88.
Bu Yongxiang. The impact of RMB exchange rate changes on domestic price levels[J]. Journal of Financial Research, 2001, (3): 78 – 88. (in Chinese)

[37]施建淮，傅雄广，许　伟. 人民币汇率变动对我国价格水平的传递[J]. 经济研究，2008，(7)：52 – 64.
Shi Jianhuai, Fu Xiongguang, Xu Wei. Pass-through of RMB exchange rate to Chinese domestic prices[J]. Economic Research Journal, 2008, (7): 52 – 64. (in Chinese)

[38]陈六傅，刘厚俊. 人民币汇率的价格传递效应——基于 VAR 模型的实证分析[J]. 金融研究，2007，(4)：1 – 13.
Chen Liufu, Liu Houjun. The price transmission effect of RMB exchange rate: An empirical analysis based on VAR model [J]. Journal of Financial Research, 2007, (4): 1 – 13. (in Chinese)

[39]Phillips A W. The relation between unemployment and the rate of change of money wage rates in the United Kindom, 1861 – 1957[J]. Economica, 1958, 25(100): 283 – 299.

[40]Samuelson P A, Solow R M. Analytical aspects of anti-inflation policy[J]. The American Economic Review, 1960, 50 (2): 177 – 194.

[41]刘金全，金春雨，郑挺国. 中国菲利普斯曲线的动态性与通货膨胀率预期的轨迹：基于状态空间区制转移模型的研究[J]. 世界经济，2006，(6)：3－12.
Liu Jinquan，Jin Chunyu，Zheng Tingguo. The dynamics of China's Phillips curve and the track of inflation expectation: A study based on the state space regime switching model[J]. The Journal of World Economy，2006，(6)：3－12. (in Chinese)
[42]郑挺国，王 霞，苏 娜. 通货膨胀实时预测及菲利普斯曲线的适用[J]. 经济研究，2012，47(3)：88－101.
Zheng Tingguo，Wang Xia，Su Na. Real-time inflation forecasting and its applicability of Phillips curve to China [J]. Economic Research Journal，2012，47(3)：88－101. (in Chinese)
[43]Friedman M. The quantity theory of money: A restatement[J]. Studies in the Quantity Theory of Money，1956，5：3－31.
[44]易 纲. 中国的货币供求与通货膨胀[J]. 经济研究，1995，(5)：51－58.
Yi Gang. China's currency supply and demand and inflation[J]. Economic Research Journal，1995，(5)：51－58. (in Chinese)
[45]刘 霖，靳云汇. 货币供应、通货膨胀与中国经济增长——基于协整的实证分析[J]. 统计研究，2005，(3)：14－19.
Liu Lin，Jin Yunhui. Money supply，inflation，and China's economic growth: An empirical analysis based on cointegration [J]. Statistical Research，2005，(3)：14－19. (in Chinese)
[46]张成思. 中国通胀惯性特征与货币政策启示[J]. 经济研究，2008，(2)：33－43.
Zhang Chengsi. The nature of inflation inertia in China and its implications on monetary policy[J]. Economic Research Journal，2008，(2)：33－43. (in Chinese)

CPI forecast and model comparison in a data-rich environment

ZHENG Ting-guo[1,2,3]，*FAN Xin-yue*[4*]，*JIN Wei*[5]
1. Center for Macroeconomic Research，Xiamen University，Xiamen 361005，China;
2. Fujian Provincial Key Laboratory of Statistical Science，Xiamen University，Xiamen 361005，China;
3. School of Economics，Xiamen University，Xiamen 361005，China;
4. School of Economics，Shanghai University of Finance and Economics，Shanghai 200433，China;
5. Penghua Fund Management Co.，Ltd.，Shenzhen 518055，China

Abstract: The advent of the Big Data era has brought unprecedented opportunities as well as challenges to CPI forecasting. Making full use of high-dimensional data and developing interpretable machine learning methods for forecasting are of great significance both theoretically and practically. Thus，this paper constructs a large monthly macroeconomic dataset for China，which consists of 239 variables across 9 categories. Based on this large dataset，the paper evaluates the forecasting performances of 13 common methods for CPI，including traditional time series models，regularized regressions，factor models，and ensemble methods. Further，based on the idea of control variables，a derivative algorithm for machine learning is constructed to explain the results and conduct the mechanism analysis. According to the results，random forest and XGBoost exhibit superior predictive performance，especially in the medium and long-term horizons. Further investigation proves that the non-linearity and non-sparsity of ensemble methods play a vital role for better forecast precision. Meanwhile，according to the variable importance measures of the two ensemble methods，variables in the autoregressive，price，and employment categories contribute a large portion of predictive power，which is in line with economic theory and stylized facts.

Key words: Big Data; CPI forecast; machine learning; nonlinearity; derivative algorithm; variable selection