数据驱动模型进行能源价格预测的研究回顾与未来展望

王雨虓，邱瑞祥，李梦祎

（中国海油集团能源经济研究院，北京 100010）

摘要：能源价格是重要的市场信号，其波动对国民经济发展和企业生产经营具有深远影响。为准确预测未来价格走势，掌握前沿能源价格预测模型，对相关研究进行系统回顾。首先，明确样本内拟合与样本外预测之间的差异；其次，归纳传统的计量经济预测模型；最后，从输入变量、预测方法和预测框架3个方向梳理先进的人工智能模型。同时，在进行研究回顾结合当前热点的基础上进一步提出未来值得关注的研究方向，一是使用集成特征选择方法构建稳定特征子集，二是引入混频方法提高实时预测精度。

关键词：数据驱动；能源价格预测；研究回顾；未来展望

中图分类号：TP183；F416.22 **文献标志码：**A **文章编号：**1671－1807(2025)09－0016－08

能源是人类社会生存与发展的基础，在国民经济中占据重要的战略地位。作为全球一次能源的主要消费对象，原油和天然气在社会、经济和环境的可持续发展中发挥着关键作用。然而，当今世界危机事件频发，导致油气价格大幅波动。准确预测油气价格对于稳定国家经济、保障金融资产安全，以及制定合理的油气进口监管政策具有重要的现实意义。此外，作为关键的市场信号，油气价格的波动对能源企业的生产经营产生了深远影响。通过准确预测能源价格，企业能够更好地优化决策，规避潜在风险。

与此同时，当今世界已进入大数据时代，创新的数据驱动模型已成为新质生产力的重要表现形式[1]。将数据驱动模型应用于能源价格预测领域，能够构建微观数据模型以分析能源经济发展模式，并为制定科学的宏观经济政策和优化企业生产规划提供坚实基础。本文旨在深入探究数据驱动的能源价格预测模型及其最新发展趋势，并为此进行系统性的综合回顾。

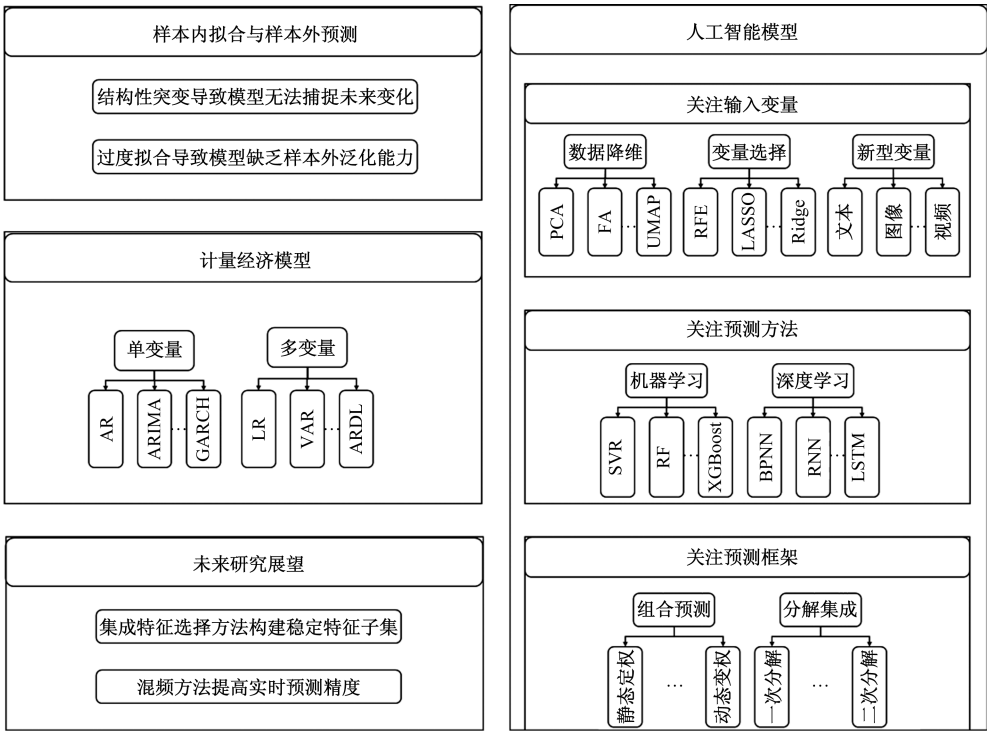
首先，明确样本内拟合与样本外预测的联系与区别至关重要，这是防止数据泄露、指导实际业务工作的首要环节。分析表明，良好的样本内拟合并不一定能保证样本外的准确预测，针对这些问题提出相应的解决方案，最终得出结论：良好的样本内拟合是有效参数估计和准确样本外预测的必要不充分条件。其次，对传统计量经济学模型进行介绍，依据为是否包含外部变量。此类模型可划分为基于时间序列分析的单变量预测模型和考虑多重影响因素的多变量预测模型。不过，传统计量经济模型在捕捉能源价格的非线性特征以及其与外部因素的复杂关系方面存在局限性。基于此，进一步阐述前沿的人工智能方法，将价格预测方法抽象为“目标-方法与框架-变量”的过程，按照这一逻辑，依次讨论关注输入变量的数据降维、变量选择和构造新型变量这3类人工智能模型，聚焦预测方法的机器学习和深度学习模型，以及组合预测和分解集成这两大预测框架。最后，结合当前研究热点与未来发展趋势，提出值得深入研究的方向。此外，还开展了实证分析，通过构建原油价格预测模型，对理论回顾内容的真实性和合理性进行验证。整体研究框架如图1所示。通过理论回顾与实证分析相结合，期望为能源领域的研究人员和从业人员理解和应用前沿方法提供有益参考。

1 样本内拟合和样本外预测

首先，需要明确样本内拟合和样本外预测之间的差异。样本内拟合是指利用全部数据对模型进

收稿日期：2024-10-16

作者简介：王雨虓(1998—)，女，山东潍坊人，硕士，经济师，研究方向为计量经济、能源价格预测；邱瑞祥(2000—)，男，山东烟台人，硕士，经济师，研究方向为能源与气候经济；李梦祎(1997—)，女，山东菏泽人，硕士，经济师，研究方向为计量经济。



AR 为自回归模型;ARIMA 为差分自回归移动平均模型;GARCH 为广义自回归条件异方差模型;LR 为线性回归模型;VAR 为向量自回归模型;ARDL 为自回归分布滞后模型;PCA 为主成分分析;FA 为因子分析;UMAP 为一致流形逼近与投影;RFE 为递归特征消除;LASSO 为最小绝对收缩和选择算子回归;Ridge 为岭回归;SVR 为支持向量回归;RF 为随机森林回归;XGBoost 为极端梯度提升回归树;BPNN 为反向传播神经网络;RNN 为循环神经网络;LSTM 为长短期记忆网络

图 1　研究框架

行拟合,并基于此对样本内数据进行分析;而样本外预测则是在保持时间序列数据顺序的前提下,将完整数据集划分为训练集和测试集,利用训练集拟合模型,并在测试集上进行预测和评估,以避免信息泄露的问题。

在实际应用中,仅获得良好的样本内拟合结果是远远不够的。研究表明,良好的样本内拟合并不能保证样本外预测的准确性。其主要原因有两点:第一,市场结构的突变可能使模型难以捕捉未来市场的变化,新的预测任务可能超出模型的能力范围,从而导致预测精度下降。例如,Tian 等[2]评估了收缩和组合两种机器学习方法在新冠肺炎疫情前后对原油价格的预测有效性,结果表明疫情增加了经济不确定性,削弱了多种模型的预测能力。在这种情况下,便需要结合专家意见来辅助判断。由于专家通常对能源市场的短期动态调整和长期结构演变有更深入的理解,结合专家知识能够有效校正定量模型的预测偏差。Zhou 等[3]研究表明,在预测未来一年的原油和天然气价格时,专家调查是对定量预测模型的有力补充。第二,过度拟合是另一个影响样本外预测准确性的关键因素。过度拟合指的是模型在训练数据上表现优异,但在测试数据上的泛化能力不足。复杂的模型架构可能会记住训练数据中的细节和噪声,忽视数据中的普遍模式,从而导致在样本外预测时表现不佳。为避免这一问题,Wang 等[4]建议在模型中引入正则化参数,并采用交叉验证等方法。在利用机器学习模型预测页岩气产量时,在目标函数中添加了正则化惩罚项,以控制模型复杂度,减轻过度拟合问题。综上所述,尽管良好的样本内拟合并不能完全保证样本外预测的准确性,但它是实现有效参数估计和准确样本外预测的必要条件。

2　计量经济模型

传统计量经济模型在早期的能源价格预测中得到了广泛应用。根据是否包含外部变量,这些模型可分为单变量预测模型和多变量预测模型。

2.1　单变量预测模型

单变量预测模型主要侧重于时间序列的建模,典型模型包括自回归模型(AR)、差分自回归移动平均模型(ARIMA)和广义自回归条件异方差模型(GARCH)等。AR 模型仅利用能源价格的历史数据进行线性回归预测;ARIMA 模型在考虑数据平

稳性的基础上，将AR模型与移动平均模型（MA）相结合；GARCH模型则在AR模型的基础上，进一步捕捉了能源价格收益时间序列中常见的波动聚集现象。然而，此类模型未能考虑外部影响因素，因而其预测性能相对有限。Xu等[5]对比了ARIMA、GARCH及一系列机器学习模型对原油价格的预测效果，结果显示ARIMA模型的预测效果最差，GARCH模型略优，但仍然不理想。Čeperić等[6]对天然气价格的不同预测模型进行了比较，结果表明传统计量经济模型如AR和ARIMA的预测效果不如神经网络和支持向量回归（SVR）等人工智能方法。

2.2 多变量预测模型

为纳入能源市场中的多重影响因素，并更好地捕捉这些因素与能源价格之间的关系，学者们提出了多变量预测模型，代表性模型包括线性回归模型（LR）、向量自回归模型（VAR）以及自回归分布滞后模型（ARDL）等。VAR模型通过构建一组线性回归方程，将被解释变量与其自身的滞后项以及解释变量的滞后项进行回归预测；ARDL模型不仅考虑了被解释变量和解释变量的滞后项，还纳入了解释变量的当期影响，将其同阶信息整合进预测模型中。刘存异[7]采用VAR模型和误差修正模型研究国际证券和外汇指数对国际原油价格的预测作用。其发现标准普尔500指数与原油价格的相关性较高，美元指数对于原油价格的预测作用也十分明显。Yin和Yang[8]则利用LR模型，比较了技术指标和宏观经济变量对原油价格的预测能力，发现技术指标的预测效果更为显著。

总体而言，传统计量经济模型在捕捉能源价格的非线性特征及其与外部因素的复杂关系方面存在局限。此外，这些模型通常要求原始数据满足平稳性假设，限制了其应用的广泛性。因此，近年来仅依赖单一计量经济模型进行能源价格预测的研究相对较少。

3 人工智能模型

大数据和人工智能技术的快速发展为预测领域带来了新的机遇。凭借出色的特征提取和非线性拟合能力，人工智能模型弥补了传统计量经济模型的不足。近年来，基于机器学习和深度学习的预测方法在能源领域受到了广泛关注。这类研究主要可分为3类：关注输入变量、关注预测方法和关注预测框架。

3.1 关注输入变量

在关注输入变量的研究中，可以进一步细分为3种类型。首先是数据降维方法。能源价格的影响因素繁多，Cheng等[9]从供给、需求、库存、金融市场、宏观经济、市场不确定性和技术指标等多个方面提取了原油价格的92个影响因素。虽然大量影响因素提供了丰富的信息，但也增加了预测模型的负担，可能导致多重共线性等问题，进而影响预测精度。降维方法旨在从高维数据中提取有效信息，以提高模型的预测能力。常见的降维方法包括主成分分析（PCA）、因子分析（FA）、稀疏主成分分析（sPCA）、偏最小二乘回归（PLS）以及一致流形逼近与投影（UMAP）等。He等[10]比较了PCA、sPCA和PLS在提取原油市场技术指标信息方面的能力，结果表明sPCA在样本内和样本外的表现均优于其他模型，能够提取更多有用信息来预测原油价格，并且该结果在不同频率的数据下均具稳健性。Zhang等[11]将深度学习中的注意力机制引入PCA方法中，通过加权分配预测变量的重要性，改进了传统PCA等权重分配的局限，从而过滤掉大部分噪声信息。

其次是变量选择方法。虽然数据降维方法能够提取关键信息以提高预测精度，但降维后的影响因素通常被混合，导致预测结果的可解释性较差。相比之下，变量选择方法专注于在多变量环境中选择与预测目标最相关的变量，既提高了预测精度，又确保了结果的可解释性。常见的变量选择方法包括最小绝对收缩和选择算子回归（LASSO）、岭回归（Ridge）和弹性网络回归（ElasticNet）等。LASSO通过在普通最小二乘法的目标函数中添加L1范数作为惩罚项，压缩部分变量的回归系数至零，实现变量选择；Ridge回归采用L2范数作为惩罚项；而ElasticNet则是L1和L2范数的加权组合。Miao等[12]利用LASSO考察了原油价格的影响因素，发现LASSO显著提升了原油价格的预测准确性，预测精度优于逐步回归等基准模型以及美国能源署（EIA）的专家预测。Zhang等[13]系统对比了LASSO、Ridge、ElasticNet与PCA和PLS在原油收益预测中的表现，发现LASSO和ElasticNet的预测效果优于其他方法，并强调了动态调整影响因素的重要性。除了基于惩罚项的嵌入式收缩方法（Shrink），基于预测模型性能的包装方法（Wrapper）也受到了能源价格预测领域的关注，此类方法通过评估影响因素子集的预测性能来选择最优子

集,常用方法包括全局搜索、向前向后搜索和递归特征消除(RFE)等。Guo 等[14]将 RFE 用于原油价格预测的影响因素选择中,取得了良好的效果。

第三种研究方向是构造新型变量,即通过数据创新获得更准确的预测结果。借助网络数据、自然语言处理、图像识别和语音识别等技术,学者们从新闻文本、投资者情绪、图像、音频和视频等非结构化数据中提取有用信息。例如,Wu 等[15]提出了一种基于谷歌趋势和在线媒体文本挖掘的原油价格预测方法,利用卷积神经网络(CNN)从在线原油新闻中提取文本特征,并验证了其对原油价格预测的有效性。Obaid 和 Pukthuanthong[16]则通过 CNN 模型从新闻照片中提取市场信息,构建投资者情绪指数,并利用该指数成功预测了股票市场收益的反转情况。Yu 等[17]使用基于卫星的集装箱信息预测股票收益,结果表明集装箱图像数据能够实时捕捉经济活动,从而显著提升了多国股指收益率的预测精度。此外,何勇等[18]从财经新闻和短视频中学习交易信号,通过将短视频得分、新闻报道文本得分和图像得分作为投资组合的依据,制定了有效的投资策略,研究结果表明,财经新闻和短视频数据中包含了与股价相关的重要信息,能够有效预测市场波动,为投资者带来超额收益。尽管此类研究主要集中在股票市场预测领域,但作为重要的金融市场之一,能源市场同样有可能从这些文本、图像、音频和视频等多源异构数据中提取出对价格预测有用的信息。

3.2 关注预测方法

近年来,受多重外部因素的影响,能源价格波动剧烈,呈现出显著的非线性、突变性和不稳定性。因此,传统计量经济方法由于其线性拟合的局限性,难以在价格预测中取得理想效果。相比之下,人工智能中的机器学习和深度学习方法因其在提取时间序列非线性特征方面的优势,受到越来越多学者的关注,并被广泛应用于能源价格预测领域。

机器学习作为人工智能的核心之一,旨在使计算机通过数据学习并总结规律,以生成能够解决特定问题的模型。在价格预测中,支持向量机(SVM)、随机森林(RF)和极端梯度提升回归树(XGBoost)等传统机器学习模型被广泛应用。Henriques 和 Sadorsky[19]对比了机器学习方法与计量经济模型在稀土股票价格预测中的表现,结果表明,SVM 和 RF 等机器学习模型相比 ARIMA 和朴素贝叶斯等计量经济模型,预测精度更高,准确率超过 90%。Tiwari 等[20]研究了 11 种不同机器学习模型在不同时间尺度下对天然气和原油价格波动的预测效果,结果表明,随着预测时间尺度的变化,模型的性能也显著变化。其中,RF 和 XGBoost 在日度、周度和月度尺度上对能源价格波动的预测表现最为优异。

与机器学习相比,深度学习神经网络在处理复杂输入数据时具备潜在优势。深度学习能够直接从原始数据中学习知识,通过设计合理的神经网络架构,模型能够捕捉特定问题的相关特征,从而在预测性能上超越传统机器学习模型。反向传播神经网络(BPNN)作为基础神经网络模型,能够学习输入特征与输出结果之间的非线性关系,并对新数据做出相应预测。Zhang 等[21]开发了一种基于 BPNN 的汽车销售量预测模型,结合了 PCA 降维、投资者情绪分析和果蝇优化算法,展示了良好的预测精度和鲁棒性。极限学习机(ELM)作为 BPNN 的改进版本,以更快的训练速度和更强的泛化能力广泛应用于能源价格预测。潘凯等[22]考虑供需基本面因素和非基本面因素,将典型相关性分析和 ELM 模型结合,用于国产 LNG 出厂价格的预测。其表明 ELM 充分发挥了神经网络在拟合非线性序列方面的优势,有效地提高了预测精度。王珏等[23]基于奇异谱分析方法,以玉米、原油、黄金分别作为大宗商品农产品类、能源类、金属类代表对象的期货价格进行分解,结合 K-means 动态聚类技术将分解量聚合成不同特征的价格序列,再采用具有优良特性的 ELM 进行训练,得到精准的大宗商品期货价格预测结果。

随着深度学习的进一步发展,循环神经网络(RNN)因其在处理时间序列数据中的记忆能力,在价格预测领域展现了独特优势。RNN 通过网络层间的权重连接,能够表达前后神经元之间的关系,适合处理依赖历史数据的时间序列预测任务。然而,RNN 存在梯度消失和梯度爆炸等问题,限制了其处理长序列的能力。长短期记忆网络(LSTM)和门控循环单元(GRU)通过引入门控机制,增强了对长期依赖关系的描述能力,在时间序列预测中表现出卓越的性能。Nasirtafreshi[24]通过 LSTM 对加密货币价格进行预测,而江雨燕等[25]将 EEMD 与 LSTM 结合,并使用麻雀搜索算法优化网络参数,构建了用于预测上证指数的混合模型。Guo 等[26]对比了 SVR、多层感知机(MLP)、CNN、BPNN、

RNN、LSTM 和 GRU 在原油价格预测中的表现，结果显示 GRU 模型的预测精度比 LSTM 提高了 2.84%，较传统 SVR 模型提高了 6.79%。

3.3 关注预测框架

在能源价格预测任务中，研究人员通常根据数据特征选择相应的预测方法。然而，预测任务的复杂性使得单一模型考虑的因素通常存在局限性，在信息利用方面不够全面，从而难以应对复杂的预测情境。正如前文所述，不同预测模型在时间频度、预测步长和预测区间等方面的表现存在差异，因此很难找到适用于所有预测任务的单一最佳方法。并且各种模型的信息提取能力具有偏好性，促使研究人员从信息互补的角度开发出增强预测模型性能的组合预测框架。该框架可以选择内部结构存在差异的多个预测模型对同一时间序列进行预测，并根据各个模型的预测误差赋予模型相应的权重，将预测结果进行加权集成，从而最大化不同模型的优势。

目前，组合预测框架主要分为静态（定权）组合框架和动态（变权）组合框架。静态组合框架基于模型的预测精度分配固定权重。当某个模型的预测精度较高时，其分配的权重较大，反之则权重较小。静态组合框架能部分利用单一模型的优势，从而靠近理想的综合预测效果。Jose 和 Winkler[27] 提出了两种改进的加权方法，通过缩尾加权降低了高误差的风险。然而，静态组合框架为各模型分配相同或固定权重，无法反映模型在不同时期的性能差异。针对这一局限，Tsang 等[28] 引入了一种时间衰减机制，为临近的精度更高的预测结果赋予更高的权重[28]。Hao 等[29] 结合滑窗机制与优化模型，提出了一种动态多目标优化集成框架，用于原油价格预测。该框架集成了计量经济模型、机器学习模型和深度学习模型，以应对原油价格的结构变化，从而显著提升预测性能。

除了权重分配的组合框架外，分解集成的思想也被广泛应用于预测框架设计。基于“分而治之，逐个击破”的理念，汪寿阳等[30] 提出了 TEI@I 方法论，并构建了一个外汇汇率预测模型。该方法论首先将复杂系统进行分解，利用计量经济模型分析主要趋势，使用人工智能技术捕捉系统的非线性，并结合文本挖掘等技术分析系统的突发性。最终，通过集成各部分的预测结果，形成对复杂系统的全面认知，从而实现对复杂系统的有效预测。Zhang 等[31] 将变分模态分解（VMD）与 GRU 模型结合，提出了一种原油价格预测框架，研究表明该框架在精度上显著优于基准模型，证明了其在原油价格预测中的有效性。Dong 等[32] 人在分解集成的基础上引入重构，避免了价格分量多次预测中的误差累积问题。他们结合 VMD 分解、相空间重构（PSR）、CNN 和 LSTM，提出了一种高精度的原油价格多步预测模型，其平均绝对百分比误差（MAPE）最低不足 1%。Xie 等[33] 将搜索引擎的非结构化数据纳入分解集成预测框架，并对高频分量进行二次分解，进一步降低预测复杂度，研究结果表明，搜索引擎的非结构化数据为天然气价格预测提供了有价值的信息，二次分解方法比一次分解取得了更高的预测精度。上述研究表明，分解集成框架的预测效果均优于单一模型，体现了数据分解处理在能源市场价格预测中的必要性。

4 实证研究

从实际应用角度出发，基于实证建模对理论回顾内容的真实性与合理性予以验证。以上海国际能源交易中心的数据为依据，旨在为新兴的中国原油期货价格提供精准预测，并选取上海原油期货价格作为被解释变量。

在解释变量的选取上，充分考虑到原油兼具的商品和金融双重属性。就商品属性而言，供需等基本面因素会影响原油价格，然而此类数据缺乏日度信息。考虑到不确定性冲击引发的经济状况改变会直接作用于原油的供需关系，进而影响原油期货价格。据此，选用美国经济政策不确定性指数来反映不确定性冲击。在金融属性方面，原油与其他金融市场联系紧密，故会受到包括货币市场、商品市场、股票市场在内的金融市场风险的影响。在货币市场中，考虑了汇率、利率和国债利率 3 个因素。具体而言，汇率以美元与人民币中间价衡量，利率以上海银行间同业拆借市场的隔夜利率计算，国债利率以 3 个月期国债发行利率表示。商品市场的情况则依据华南商品指数来反映。对于股票市场，选取了两个市场，即整体股票市场和能源行业股票市场。其中，整体股票市场以上证综指衡量，能源板块股票市场以能源指数代表。所有数据均采用滞后五阶信息作为输入，以此预测下一日的原油价格。

为防止因信息泄露而产生不可信的高预测精度问题，首先按照 8∶2的比例将数据划分为训练集和测试集。模型预测精度以平均绝对百分比误差（MAPE）衡量，其计算公式为

$$\mathrm{MAPE}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-y'_i}{y_i}\right| \tag{1}$$

式中:y_i 为真实值;y'_i 为预测值。

在预测模型方面,选取 RF 机器学习模型作为预测方法相关研究中的典型代表,同时将 LASSO 降维收缩模型作为关注输入变量研究的典型代表。此外,将 CEEMDAN 分解技术与 LASSO 预测模型相结合,通过该方式从中提取高频和低频分量,以此凸显分解集成框架的作用。受篇幅所限,加之核心目标并非阐述模型方法的实现原理,故在此不详细介绍各模型方法的实现过程。在构建模型时,针对各个模型,运用网格搜索法进行超参数寻优,网格搜索的超参数及其相应范围见表 1。

实证结果有力地证明了理论回顾内容的真实性与合理性。首先,图 2(a)中 RF 模型的绘图结果清晰地呈现出欠拟合状况。具体而言,训练集的散点沿“真实值=预测值”这条线分布,且训练集预测的 MAPE 仅约为 1%。然而,当模型拓展至样本外时,预测性能显著下降,表现为测试集散点不再沿“真实值=预测值”的线分布,同时 MAPE 升高了近 5 倍。这一结果强调了在预测研究中,明确样本内拟合和样本外预测这一任务的首要性。其次,针对过度拟合问题,依据第 1 节内容,结合专注于输入变量的变量选择降维收缩算法,进一步运用 LASSO 对原油价格进行预测。如图 2(b)所示,引入惩罚因子的 LASSO 算法在确保良好的样本内拟合基础上,能够实现准确的样本外预测,样本外测试集 MAPE 约为 2%,展现出了卓越的性能。最后,如图 2(c)所示,将分解集成思路与LASSO相结合后,预测性能并未得到大幅提升。这主要是因为 LASSO 算法本身已能实现较好的预测效果,仅引入分解思路对于实现更高精度的预测而言是较为困难的。

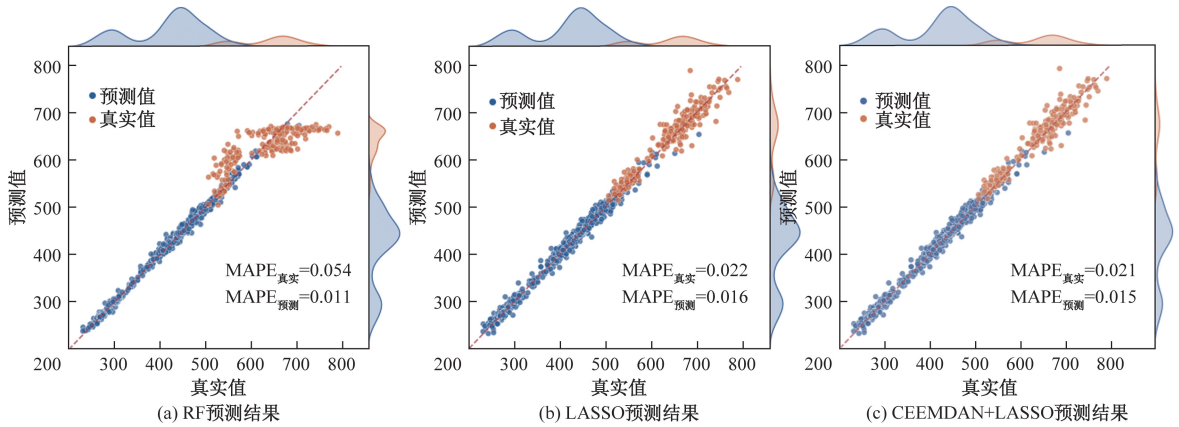
5 未来研究展望

5.1 集成特征选择方法构建稳定特征子集

数据驱动模型进行精准的能源价格预测依赖于高质量的输入数据。然而,由于能源价格的影响机制复杂,如何有效提取影响因素并整合不同的特征选择方法,已成为提升预测精度的关键。鉴于现有特征选择方法繁多,且逻辑各异,依赖单一方法可能导致特征子集不稳定。因此,需要开发集成特征选择方法,提高特征选择的稳定性和预测模型的精度。

表 1 网格搜索超参数范围

模型	超参数	寻优范围
LASSO 最小绝对收缩和选择算子	alpha(正则化强度)	[0.01,0.1,1,10,100,1 000]
	selection(系数更新策略)	['cyclic(循环更新)', 'random(随机更新)']
RF 随机森林	n_estimators(决策树数量)	[5,10,15]
	max_depth(最大深度)	[1,3,5,7,9,15]
	min_samples_split(拆分内部节点所需的最小样本数)	[2,5,10]
	min_samples_leaf(叶子节点所需的最小样本数)	[1,2,4]
	bootstrap(是否使用自助采样)	[True(是), False(否)]
	max_features(最佳分割时要考虑的特征数量)	[0.5, 'sqrt(平方根)', 'log$_2$(以 2 为底的对数)']



(a) RF预测结果　(b) LASSO预测结果　(c) CEEMDAN+LASSO预测结果

图 2 模型预测结果对比

目前，只有少数研究在特定领域中应用了这一理念。例如，Eskandari等[34]通过投票机制集成了F检验、RF等4种变量选择方法，成功构建了一个稳定且有效的能源消耗时间序列预测特征子集。Xiang等[35]则采用加权与投票相结合的方式，集成了RF、XGBoost等变量选择方法，构建了适用于机械设备时间序列数据预测的稳定变量子集。然而，能源价格预测领域在这一方面的研究仍较为薄弱。

5.2 混频方法提高实时预测精度

标准回归预测模型要求被解释变量和解释变量的数据频率一致，这在进行中长期能源价格预测时存在一定的局限性。传统方法通常采用同频率的低频数据进行建模。例如，在季度价格预测中，通常只能使用截至上一季度的相关数据，无法利用当前季度内已公布的月度数据，这限制了预测的准确性和时效性。为了解决这一问题，学者们提出了混频数据抽样回归(MIDAS)方法。该方法通过将高频数据以灵活且简化的方式整合到低频回归模型中，有效解决了不同频率数据下的模型估计和预测难题。因此，MIDAS模型能够纳入较季度数据更具时效性的月度数据，从而更实时地反映能源市场的动态变化。

然而，针对能源价格的实时预测研究相对匮乏，现有研究主要集中在宏观经济指标的实时预测领域。例如，刘汉和刘金全[36]利用MIDAS模型对中国季度国内生产总值(GDP)增速进行了实时预测；王金明[37]则基于MIDAS模型，结合高频数据，对中国GDP增速进行了短期实时预测，预测误差低于0.1%。随着MIDAS模型的理论不断完善，其应用已扩展至多个领域，包括土地价格[38]、旅游需求[39]以及二氧化碳排放量[40]的预测分析。

总的来说，针对中长期能源价格的实时预测具有重要的现实意义，但目前研究仍较为不足。基于MIDAS模型的实时预测方法为能源价格预测提供了有益的借鉴，通过参考已有的相关研究，有望实现更及时、精准的能源价格预测。

参考文献

[1] 任保平. 生产力现代化转型形成新质生产力的逻辑[J]. 经济研究，2024(3)：12-19.

[2] TIAN G，PENG Y，MENG Y. Forecasting crude oil prices in the COVID-19 era：can machine learn better? [J]. Energy Economics，2023，125：106788.

[3] ZHOU F，PAGE L，PERRONS R K，et al. Long-term forecasts for energy commodities price：what the experts think[J]. Energy Economics，2019，84：104484.

[4] WANG M，HUI G，PANG Y，et al. Optimization of machine learning approaches for shale gas production forecast [J]. Geoenergy Science and Engineering，2023，226：211719.

[5] XU Z，MOHSIN M，ULLAH K，et al. Using econometric and machine learning models to forecast crude oil prices：Insights from economic history[J]. Resources Policy，2023，83：103614.

[6] ČEPERIC E，ŽIKOVIC S，ŽEPERIC V. Short-term forecasting of natural gas prices using machine learning and feature selection algorithms[J]. Energy，2017，140：893-900.

[7] 刘存异. 金融市场对国际原油价格预测作用研究[J]. 新西部，2018(2)：58-59，50.

[8] YIN L，YANG Q. Predicting the oil prices：do technical indicators help? [J]. Energy Economics，2016，56：338-350.

[9] CHENG X，WU P，LIAO S S，et al. An integrated model for crude oil forecasting：Causality assessment and technical efficiency [J]. Energy Economics，2023，117：106467.

[10] HE M，ZHANG Y，WEN D，et al. Forecasting crude oil prices：a scaled PCA approach[J]. Energy Economics，2021，97：105189.

[11] ZHANG X，CHENG S，ZHANG Y，et al. An attention-PCA based forecast combination approach to crude oil price[J]. Expert Systems with Applications，2024，240：122463.

[12] MIAO H，RAMCHANDER S，WANG T，et al. Influential factors in crude oil price forecasting[J]. Energy Economics，2017，68：77-88.

[13] ZHANG Y，MA F，WANG Y. Forecasting crude oil prices with a large set of predictors：can LASSO select powerful predictors? [J]. Journal of Empirical Finance，2019，54：97-117.

[14] GUO J，ZHAO Z，SUN J，et al. Multi-perspective crude oil price forecasting with a new decomposition-ensemble framework [J]. Resources Policy，2022，77：102737.

[15] WU B，WANG L，LÜ S X，et al. Effective crude oil price forecasting using new text-based and big-data-driven model[J]. Measurement，2021，168：108468.

[16] OBAID K，PUKTHUANTHONG K. A picture is worth a thousand words：measuring investor sentiment by combining machine learning and photos from news [J]. Journal of Financial Economics，2022，144(1)：273-297.

[17] YU H，HAO X，WU L，et al. Eye in outer space：satellite imageries of container ports can predict world stock returns[J]. Humanities and Social Sciences Com-

munications, 2023, 10(1): 01891-9.
[18] 何勇, 李琪琪, 焦丽, 等. 另类数据在中国股票市场投资中有用吗? 基于财经短视频、图像、文本数据的探究[J]. 计量经济学报, 2023, 3(4): 1008-1031.
[19] HENRIQUES I, SADORSKY P. Forecasting rare earth stock prices with machine learning[J]. Resources Policy, 2023, 86: 104248.
[20] TIWARI A K, SHARMA G D, RAO A, et al. Unraveling the crystal ball: machine learning models for crude oil and natural gas volatility forecasting[J]. Energy Economics, 2024, 134: 107608.
[21] ZHANG C, TIAN Y X, FAN Z P. Forecasting sales using online review and search engine data: a method based on PCA-DSFOA-BPNN[J]. International Journal of Forecasting, 2022, 38(3): 1005-1024.
[22] 潘凯, 谢翔, 张曦, 等. 基于 CCA-ELM 模型的国产 LNG 出厂价格中短期预测研究: 以陕西省为例[J]. 国际石油经济, 2024, 32(7): 99-106.
[23] 王珏, 齐琛, 李明芳. 基于 SSA-ELM 的大宗商品价格预测研究[J]. 系统工程理论与实践, 2017, 37(8): 2004-2014.
[24] NASIRTAFRESHI I. Forecasting cryptocurrency prices using recurrent neural network and long short-term memory[J]. Data & Knowledge Engineering, 2022, 139: 102009.
[25] 江雨燕, 邵金, 陈梦凯, 等. 融合分解集成和深度学习的金融时间序列预测模型[J]. 统计与决策, 2023, 39(24): 152-156.
[26] GUO L, HUANG X, LI Y, et al. Forecasting crude oil futures price using machine learning methods: evidence from China[J]. Energy Economics, 2023, 127: 107089.
[27] JOSE V R R, WINKLER R L. Simple robust averages of forecasts: some empirical results[J]. International Journal of Forecasting, 2008, 24(1): 163-169.
[28] TSANG T K, DU Q, COWLING B J, et al. An adaptive weight ensemble approach to forecast influenza activity in an irregular seasonality context[J]. Nature Communications, 2024, 15(1): 8625.
[29] HAO J, FENG Q, YUAN J, et al. A dynamic ensemble learning with multi-objective optimization for oil prices prediction[J]. Resources Policy, 2022, 79: 102956.
[30] 汪寿阳, 余乐安, 黎建强. TEI@I 方法论及其在外汇汇率预测中的应用[J]. 管理学报, 2007, 23(1): 21-27.
[31] ZHANG S, LUO J, WANG S, et al. Oil price forecasting: a hybrid GRU neural network based on decomposition-reconstruction methods[J]. Expert Systems with Applications, 2023, 218: 119617.
[32] DONG Y, JIANG H, GUO Y, et al. A novel crude oil price forecasting model using decomposition and deep learning networks[J]. Engineering Applications of Artificial Intelligence, 2024, 133: 108111.
[33] XIE G, JIANG F, ZHANG C. A secondary decomposition-ensemble methodology for forecasting natural gas prices using multisource data[J]. Resources Policy, 2023, 85: 104059.
[34] ESKANDARI H, SAADATMAND H, RAMZAN M, et al. Innovative framework for accurate and transparent forecasting of energy consumption: a fusion of feature selection and interpretable machine learning[J]. Applied Energy, 2024, 366: 123314.
[35] XIANG F, ZHAO Y, ZHANG M, et al. Ensemble learning-based stability improvement method for feature selection towards performance prediction[J]. Journal of Manufacturing Systems, 2024, 74: 55-67.
[36] 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.
[37] 王金明. 基于混频回归方法对经济总量的实时预测[J]. 统计与决策, 2021, 37(24): 42-45.
[38] 秦梦, 祁新洲, 宋玉茹, 等. 基于混频误差修正模型的土地价格预测[J]. 统计与决策, 2022, 38(3): 178-183.
[39] 秦梦, 刘汉. 百度指数、混频模型与三亚旅游需求[J]. 旅游学刊, 2019, 34(10): 116-126.
[40] 秦华英, 韩梦. 基于 MIDAS 模型中国碳排放量的实时预报与短期预测[J]. 环境科学学报, 2018, 38(5): 2099-2107.

Research Review and Future Prospect of Data-driven Model for Energy Price Forecasting

WANG Yuxiao, QIU Ruixiang, LI Mengyi

(CNOOC Energy Economics Institute, Beijing 100010, China)

Abstract: Energy price is a key market signal, and their fluctuations have profound impact on national economic development and business operations. To accurately predict future price trends and master cutting-edge energy price forecasting models, a systematic review of relevant research was conducted. First, the distinction between in-sample fitting and out-of-sample prediction was clarified. Second, traditional econometric forecasting models was summarized. Finally, advanced artificial intelligence models was organized from the perspectives of input variables, forecasting methods, and forecasting frameworks. Based on the research review and current trends, there are two future directions worth attention: first, the use of integrated feature selection methods to construct stable feature subsets, and second, the introduction of mixed-frequency methods to improve the accuracy of real-time predictions.

Keywords: data-driven; energy price forecasting; research review; future outlook