文章编号:1000 - 4653(2026)04 - 0076 - 12

基于特征选择方法的船舶燃油消耗预测模型

王 舰[1], 王兴龙[1], 籍 伟[1], 曹文杰[2], 肖仲明[2], 王新建[2*]
(1. 浙江富兴海运有限公司, 浙江 杭州 310007;
2. 大连海事大学 航海学院, 辽宁 大连 116026)

摘 要:随着全球航运业对节能减排要求的不断提高,精确预测船舶燃油消耗已成为优化航运效率与降低环境影响的关键。然而,多源数据中存在的特征冗余与噪声干扰问题,制约了燃油消耗预测模型的精度与稳定性。针对上述问题,提出一种精确预测船舶燃油消耗的综合框架。整合船舶正午报告数据、欧洲中期天气预报中心(ERA5)分析数据、全球海洋物理分析与预测数据以及自动识别系统数据,构建了多维特征空间数据;针对特征冗余与噪声干扰问题,引入遗传算法进行特征选择,将 32 个原始特征精简至 15 个;然后,系统评估九种机器学习回归模型,结果显示随机森林(RF)在捕捉燃油消耗非线性特性方面具有显著优势;基于 RF 模型的特征重要性分析揭示了时间维度变量(如日航时、总航时)在燃油消耗预测中的主导作用,为能效管理提供了定量决策依据。结果表明:遗传算法特征选择有效降低了约 53.1% 的计算复杂度,随机森林模型在捕捉燃油消耗非线性特性方面表现最佳,显著提升了预测精度。研究成果为船舶航运业的燃油优化和减排策略制定提供了理论支撑,对推动航运业绿色转型具有重要意义。

关键词:绿色航运;船舶燃油消耗;多源数据融合;机器学习;遗传算法;特征选择
中图分类号:U676.3 **文献标志码**:A **DOI**:10.3969/j.issn.1000 - 4653.2026.04.009

Ship fuel consumption prediction model based on feature selection method

Wang Jian[1], *Wang Xinglong*[1], *Ji Wei*[1], *Cao Wenjie*[2], *Xiao Zhongming*[2], *Wang Xinjian*[2*]
(1. Zhejiang Fuxing shipping Co., Ltd., Hangzhou 310007, China;
2. Navigation College, Dalian Maritime University, Dalian 116026, China)

Abstract: With the continuous enhancement of global shipping's requirements for energy conservation and emission reduction, accurate prediction of ship fuel consumption has become crucial for optimizing shipping efficiency and mitigating environmental impacts. However, the issues of feature redundancy and noise interference in multi-source data restrict the accuracy and stability of fuel consumption prediction models. To address these problems, a comprehensive framework for precise fuel consumption prediction is proposed. First, ship noon report data, ERA5 reanalysis data, global ocean physical analysis and prediction data, and Automatic Identification System (AIS) data are integrated, and a multi-dimensional feature space dataset is constructed. Second, to address feature redundancy and noise interference, a genetic algorithm is introduced for feature selection, and the original 32 features are reduced to 15. Third, nine machine learning regression models are systematically evaluated, and the results indicate that the Random Forest model exhibits significant advantages in capturing the nonlinear characteristics of fuel consumption. Finally, feature importance analysis based on the RF model reveals that time-related variables (such as daily operating hours and total operating hours) play a dominant role in fuel consumption prediction, providing quantitative decision support for energy efficiency management. The results show that the genetic algorithm-based feature selection effectively reduces computational complexity by approximately 53.1%, and the

收稿日期:2025-09-25
基金项目:国家自然科学基金项目(52571403;52101399)资助
第一作者:王 舰(1977—),男,工程师,研究方向为船舶能效管理,(E-mail)wangxinjian@dlmu.edu.cn
*通信作者**:王新建(1988—),男,副教授,博士,研究方向为交通信息工程及控制,(E-mail)wangxinjian@dlmu.edu.cn
引用格式:王舰,王兴龙,籍伟,等. 基于特征选择方法的船舶燃油消耗预测模型[J]. 中国航海,2026,49(4):76 - 87.
Wang J, Wang X L, Ji W, et al. Ship fuel consumption prediction model based on feature selection method[J]. Navigation of China, 2026, 49(4): 76 - 87. (in Chinese)

Random Forest model performs best in capturing the nonlinear characteristics of fuel consumption, significantly improving prediction accuracy. The research findings provide theoretical support for fuel optimization and emission reduction strategies in the shipping industry, and are of great significance in promoting the green transformation of maritime transport.

Key words: green shipping; ship fuel consumption; multi-source data fusion; machine learning; genetic algorithm; feature selection

随着全球航运业的迅猛发展,船舶能效管理已成为航运业可持续发展的关键议题。据国际海事组织(International Maritime Organization,IMO)第四次温室气体研究报告,2018 年全球航运 CO_2 排放量约为 10.56 亿 t,占当年全球人为排放量的 2.89%[1]。此外,IMO 燃油消耗数据收集系统对 5 000 总吨及以上船舶的年度汇总显示,2022 年和 2023 年全球航运分别消耗约 2.13 亿 t 与 2.11 亿 t 燃油[2-3]。在“碳达峰”与“碳中和”的双重目标下,航运业面临着严峻的减排压力。精准预测船舶燃油消耗是实施能效管理、优化航行策略和评估减排措施有效性的基础。因此,构建高精度预测模型已成为推动航运业绿色转型的关键技术支撑和核心任务之一[4]。

作为船舶能效管理的关键环节,船舶燃油消耗预测方面的研究经历从经验公式到数据驱动模型的变化过程。传统方法主要依赖于船舶阻力理论和推进效率进行燃油消耗计算,如 Holtrop-Mennen 方法[5]和 ISO 15016 标准[6]。此类“白盒”模型虽具备清晰的物理基础,但其假设条件较为理想化,在面对复杂多变的真实航行环境时,其预测精度往往受到限制。近年来,随着船载传感器技术和数据采集系统的普及,基于机器学习的数据驱动方法已成为该领域的研究热点,展现出更强的环境适应性与预测精度。张志政等[7]利用支持向量机(Support Vector Machine, SVM)构建船舶油耗预测模型,通过考虑航速、载重等因素,取得较传统方法更高的预测精度。Yan 等[8]则采用随机森林(Random Forest,RF)算法分析海况、航速和装载状态对燃油消耗的综合影响,成功将预测误差控制在 5% 以内。此外,刘博[9]进一步引入深度学习方法,结合长短期记忆网络(Long Short-Term Memory,LSTM)捕捉燃油消耗的时序特性,实现多时间尺度的精确预测。

尽管上述研究取得了重要成果,但通过对现有文献的系统性梳理,本研究识别出当前船舶燃油消耗预测领域仍面临四大关键挑战,制约着预测模型的精度、泛化能力和可解释性。

1)多源异构数据的融合难题。现有研究的数据源相对单一,或仅进行简单拼接。例如,部分研究依赖低频的船舶正午报告[10],难以捕捉航行动态;一些研究虽引入船舶自动识别系统(Automatic Identification System,AIS)或气象数据[11],但忽略了海洋水文等关键环境因素。

2)高维特征的冗余与噪声干扰。全面的数据融合不可避免地导致特征维度急剧增加,其中必然存在大量共线性、冗余和噪声特征[12]。这些无关特征不仅会增加模型的计算负担,更容易引发“维度灾难”,导致模型过拟合,降低其泛化能力[13]。

3)模型适用性与性能缺乏系统性评估。学术界已将多种机器学习算法应用于该领域,并各自取得良好效果。然而,这些研究往往是“单点式”的,即聚焦于某一种特定算法的优化与应用[14]。

4)预测模型的“黑盒”问题与可解释性不足。多数高性能机器学习模型本质上是“黑盒”或“灰盒”模型,其决策过程不透明。现有研究大多集中于提升预测精度,而对各影响因素如何驱动燃油消耗变化、特征之间是否存在复杂的非线性相互作用等深层机理的定量解析能力有限[15]。

为应对上述挑战,本研究旨在构建一个从数据融合、特征选择到模型评估和机理解释的综合性预测方法。其核心贡献不仅在于系统性地对比 9 种主流机器学习模型的性能,更在于该方法确保了所有模型在一个经过多源融合和全局特征优化的统一数据基础上进行公平评估。这超越了以往研究中常见的“单点式”方法,旨在为船舶燃油消耗预测这一复杂非线性问题,提供一个更具泛化能力和可解释性的建模范式,并为行业提供更为可靠的技术选型依据。

1　数据与方法论

1.1　数据来源及预处理

本研究综合利用多源异构数据,以实现高精度船舶燃油消耗预测。本研究以一艘远洋散货船为对象,主尺度为船长 224.90 m、船宽 32.23 m、设计吃水 19.10 m,主机额定功率 10 959 kW。主数据集由一家国际航运公司提供,覆盖 2021 年 1 月 14 日—2024 年 7 月 30 日共 479 d 的运营记录。主要数据来源包括:船舶正午报告数据、欧洲中期天气预报中心(European Centre for Medium-Range Weather Forecasts,ECMWF)的 ERA5 再分析数据集、全球海洋物理分析与预测数据和 AIS 数据,用于预测船舶燃油

消耗的特征来源及其特征名称见表1。

表1 每个数据源的特征

Tab.1 Features from each data source

数据来源	特征名称
船舶正午报告数据	日航程 总航程 ⋮ 主机油耗
ERA5再分析数据集	海面上10 m风速 海面上2 m温度 ⋮ 相对风浪角度
全球海洋物理分析与预测数据	总海洋速度 总潮汐速度 总波浪速度 总海面水流速度
AIS数据	经度 纬度 吃水

注:船舶正午报告数据共包括16个变量:日航程、总航程、距目的港剩余距离、日航时、总航时、淡水、天气、大气温度、总平均航速、日航速、航向、相对风向、气压、海况、主机转速和主机油耗。ERA5再分析数据集共包含12个变量:海面上10 m风速、海面上2 m温度、预测表面粗糙度、综合风浪与涌浪的显著波高、波浪影响阻力系数、总涌浪平均周期、风浪平均周期、涌浪显著波高、风浪显著波高、相对波向角度、相对总涌浪角度和相对风浪角度。此外,AIS数据用于提供船舶高频位置信息,但经纬度数据不作为模型输入。

为确保输入数据的质量与一致性,本研究设计了系统性的数据预处理和融合流程,主要包括数据清洗、时空对齐和特征归一化等3个关键步骤。

1.1.1 数据清洗

1)基于领域知识设定物理约束边界,剔除明显不合理的记录。这些不合理记录主要包括:主机油耗值小于零、航速超过该船型设计最大航速和船舶经纬度坐标出现瞬时大幅跳变等在物理或逻辑上不可能发生的异常数据点。

2)针对高频的AIS和气象数据,采用拉依达准则识别并剔除了统计学上的离群点[16]。

3)对于数据记录中的缺失值,采用线性插值法进行填补。

1.1.2 时空对齐

本研究融合的数据源具有不同的时间和空间分辨率。为实现有效对齐,本研究以低频但信息丰富的船舶正午报告作为基准时间戳。对于每一个记录,本研究通过时间和空间上的最近邻匹配原则,关联对应时间点的AIS位置信息以及该位置和时间下ECMWF与全球海洋物理分析数据中的气象海况参数,形成一个时间截面上的宽表。

1.1.3 特征归一化

为消除不同特征间量纲和数量级差异对模型训练的影响,本研究采用最小-最大归一化方法,将所有数值型特征线性映射至[0,1]区间,有

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

式(1)中:X_{norm}为归一化后的特征值;X为原始值;$X_{\min}$和$X_{\max}$分别为该特征在数据集中的最小值和最大值。

1.2 特征选择方法

为应对船舶燃油消耗预测中高维、多源异构数据的特征冗余和噪声干扰问题,本研究引入遗传算法(Genetic Algorithm,GA)以实现全局最优特征子集的搜索[17]。GA通过模拟生物进化过程,在种群内并行探索特征组合空间,能有效平衡模型精度与特征数量,从而提高模型的泛化能力和稳定性[18]。GA的相关步骤如下:

1.2.1 染色体编码与种群初始化

在本方法中,每个染色体以长度等于原始特征数量的二进制向量表示。并且为确保每个个体至少包含一个特征,初始种群在生成时将进行约束。相关数学定义为

$$C = [g_1, g_2, \cdots, g_n] \tag{2}$$

$$g_i \in \{0,1\} \tag{3}$$

式(2)和式(3)中:n为特征总数;每个染色体C为一个候选特征子集;$g_i \in \{0,1\}$为第i个特征是否被选择(1为选取,0为舍弃)。

1.2.2 适应度评估

适应度函数是*GA*中评估个体优劣的关键指标。本研究采用交叉验证下的决定系数(R^2)作为适应度指标,相关数学定义为

$$\text{fitness}(C) = \frac{1}{k}\sum_{i=1}^{k} R_i^2 \tag{4}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{l}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{l}(y_i - \bar{y})^2} \tag{5}$$

式(4)和式(5)中:fitness( )为适应度;k为交叉验证的折数;R_i^2为第i折验证集上的决定系数;y_i为真实燃油消耗值;$\hat{y}_i$为预测值;$\bar{y}$为真实值的均值;l为样

本数量。

1.2.3　选择操作

选择操作模拟自然选择过程,优先保留适应度较高的个体。本研究采用轮盘赌选择法,按照个体适应度比例分配选择概率。相关数学定义为

$$P(C_i) = \begin{cases} \dfrac{\text{fitness}(C_i)}{\sum_{j=1}^{\text{pop_size}} \text{fitness}(C_j)}, & \text{当} \sum_{j=1}^{\text{pop_size}} \text{fitness}(C_j) \neq 0 \\ \dfrac{1}{\text{pop_size}}, & \text{其他} \end{cases} \tag{6}$$

$$P_{\text{cum}}(C_i) = \sum_{j=1}^{i} P(C_j) \tag{7}$$

式(6)和式(7)中:$P(C_i)$为个体 C_i 被选中的概率;pop_size 为种群大小;P_{cum}为累计概率。

1.2.4　交叉操作

交叉操作通过交换不同染色体的基因片段产生新的后代,增加种群多样性。本研究采用单点交叉策略,以概率 P_c 选择染色体对进行交叉。相关数学定义为

$$P_1 = [g_1^1, g_2^1, \cdots, g_n^1] \tag{8}$$

$$P_2 = [g_1^2, g_2^2, \cdots, g_n^2] \tag{9}$$

$$O_1 = [g_1^2, g_2^2, \cdots, g_c^2, g_{c+1}^1, \cdots, g_n^1] \tag{10}$$

$$O_2 = [g_1^1, g_2^1, \cdots, g_c^1, g_{c+1}^2, \cdots, g_n^2] \tag{11}$$

式(8)~(11)中:P_1 和 P_2 分别为单点交叉中不同的父代染色体;O_1 和 O_2 为对应产生的后代;交叉点为 $c \in [1, n-1]$。

1.2.5　变异操作

变异操作通过随机改变个体的基因值,维持种群的遗传多样性,防止算法过早收敛于局部最优解。本研究对每个染色体以 p_m 进行变异操作,随机选择一个基因位置,将其值从 0 变为 1,1 变为 0。对于染色体 $C = [g_1, g_2, \cdots, g_n]$,随机选择位置 $m \in [0, n-1]$,变异后的染色体 C'的相关数学定义为

$$C' = [g_1, g_2, \cdots, g_{m-1}, 1 - g_m, g_{m+1}, \cdots, g_n] \tag{12}$$

1.2.6　算法迭代与最优解提取

GA 通过多代迭代,不断进化种群,逐步收敛到最优解。在每一代迭代中,记录当前代的最佳个体及其适应度,若高于历史最佳值则更新全局最优解。

为确保本研究中所采用的 GA 具有科学的严谨性和可复现性,其具体的实现与超参数配置如下:

本研究采用二进制编码方案,其中每个染色体(个体)为一个长度等于原始特征总数(32)的二进制向量,向量中的每一位(基因)代表对应特征是否被选择(1 为选择,0 为不选择)。为避免生成无意义的解,在种群初始化阶段即施加约束,确保每个染色体至少包含一个被选择的特征。算法的适应度函数被定义为在特定特征子集下,采用五折交叉验证的随机森林回归模型的平均决定系数。这种基于交叉验证的适应度评估机制,旨在降低特征选择过程中的过拟合风险,提升所选特征子集的泛化能力[19]。

在遗传算子的具体配置上,选择操作采用经典的轮盘赌选择机制,使适应度更高的个体有更大的概率被选中进入下一代。交叉操作采用单点交叉,其交叉概率设定为 0.25。变异操作则采用单点变异,其变异概率设定为 0.01。为防止在迭代过程中丢失已发现的最优解,本研究还引入精英主义策略,即每一代都无条件地将当前种群中的最优个体复制到下一代种群中。

算法的终止条件设定为达到最大迭代代数(100 代)。这些超参数包括种群大小、交叉和变异概率等,均是通过一系列探索性的预试验,并在参考进化计算领域相关文献的基础上确定的,其核心目标是在算法的全局探索能力与局部利用能力之间取得稳健的平衡[19]。

1.3　船舶燃油消耗预测模型

为构建高精度的船舶燃油消耗预测模型,本研究设计并实施一个系统化的多模型评估与选择框架,其核心旨在超越单一算法应用的局限性,通过对不同机器学习范式的系统性考察,识别出与船舶燃油消耗预测任务内在特性方面最为匹配的理想模型。在此框架下,本研究选取 9 种在学术界和工业界均被广泛验证的主流回归算法[20]。这些算法涵盖集成学习、核方法、实例学习和正则化线性模型等多种技术路径,从而能从多视角对问题进行探索。涉及的各机器学习模型与评价指标的说明见表 2。

为对上述各模型的预测性能进行全面、多维度的量化评估,本研究构建一套包含多种标准评价指标的综合评估体系。该体系旨在从不同维度衡量模型的预测能力,不仅考察预测值和真实值之间的偏差大小,也评估模型的拟合优度与对数据方差的解释能力。本研究选取的关键指标包括决定系数(R^2)、均方根误差(Root Mean Square Error,S_{RME})、平均绝对误差(Mean Absolute Error,A_{ME})、平均绝对百分比误差(Mean Absolute Percentage Error,P_{MAE})和可释方差(Explained Variance,E_{V})。这些指标将从不同角度量化模型的拟合能力和预测精度[21-22]。

表 2 机器学习模型与评价指标说明

Tab. 2 Descriptions of machine learning models and evaluation metrics

英文名称	说明
Random Forest	基于决策树集成思想的非线性回归模型
Gradient Boosting	利用梯度下降方法构建决策树集成的回归模型
XGBoost	高效、可扩展的梯度提升决策树模型
LightGBM	高性能、低内存消耗的梯度提升框架,适合大规模数据
SVR	基于支持向量机原理的回归算法,适用于小样本学习
KNN	基于实例学习,通过距离度量实现预测的回归算法
Lasso	引入 L1 正则化约束的线性回归模型
AdaBoost	采用加权集成思想提升弱学习器性能的回归方法
Decision Tree	以树形结构构建的回归模型,易于解释和实现
R^2	反映模型拟合优度,取值范围通常在(0,1)
S_{RME}	衡量预测值与真实值之间偏差的均方根
A_{ME}	测量预测值与真实值之间绝对差值的平均
P_{MAE}	衡量预测误差占真实值百分比的平均值
E_{V}	反映模型解释数据总方差比例的指标

1) R^2 用于衡量模型对数据变异性的解释程度,其值越接近 1,表明模型的拟合优度越高。

2) S_{RME}则对预测误差较大(即离群点)的样本更为敏感,是衡量模型预测精度的常用标准。

3) A_{ME}直接反映预测误差的平均绝对大小,对离群点不敏感,因此在评估模型整体性能时更具鲁棒性。

4) 为便于在不同尺度的数据集上进行模型性能的横向比较,本研究还引入无量纲的 P_{MAE}。

5) E_{V} 被用以衡量模型在多大程度上解释了数据集的方差,其值越接近 1,表明模型对数据波动的捕捉能力越强。

这些指标的计算公式分别为

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{13}$$

$$S_{\mathrm{RME}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{14}$$

$$A_{\mathrm{ME}} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{15}$$

$$P_{\mathrm{MAE}} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{16}$$

$$E_{\mathrm{V}} = 1 - \frac{\mathrm{Var}[y - \hat{y}]}{\mathrm{Var}[y]} \tag{17}$$

式(13)~式(17)中:n 为测试集样本的数量;y_i 为第 i 个样本的真实燃油消耗值;$\hat{y}_i$ 为模型对应的预测值;$\bar{y}$ 为所有真实值的平均值;Var[ ]为方差。

为对模型进行严格的训练与评估,本研究将构建的数据集按照 80∶20 的比例,通过分层随机抽样划分为训练集和测试集。为实现各模型超参数的系统性寻优,本研究采用一种结合五折交叉验证和贝叶斯优化的先进策略。

2 结果与讨论

2.1 基于 GA 的特征选择结果

船舶燃油消耗预测的准确性在很大程度上取决于输入特征的质量与相关性。本研究采用 GA 进行特征选择,以优化预测模型的性能。

为直观展示 GA 搜索过程中不同特征在高适应度染色体中的保留情况,绘制特征选择的染色体基因分布热力图见图 1。图 1 中:深色区域表示高度相关性,而浅色区域则表示较低相关性;横轴为 32 个原始特征的基因位点,而纵轴则为按适应度从高到低排序的 100 个优秀个体(染色体);每一个像素点的颜色状态(深色为 1,浅色为 0)揭示了特定特征是否在某个高性能的特征子集方案中被激活。

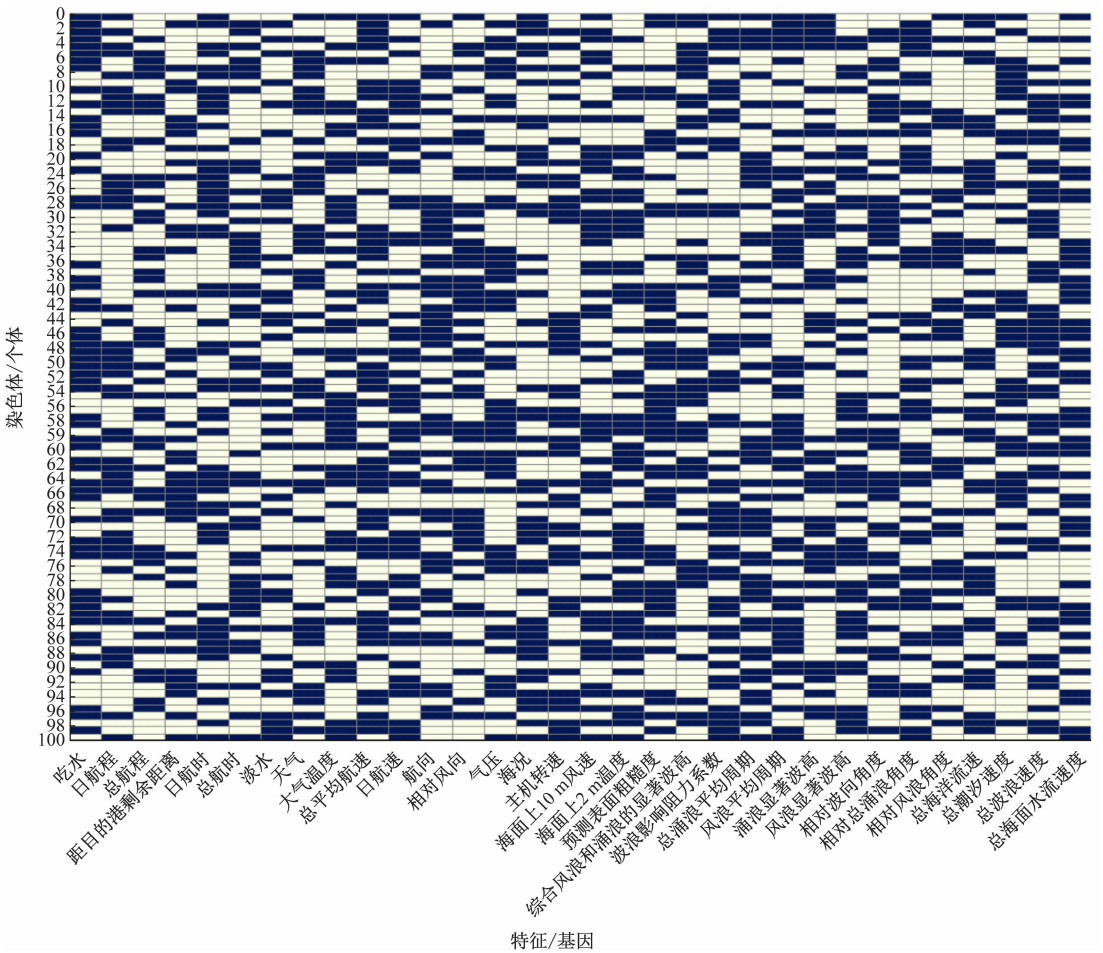


图 1　特征选择的染色体基因分布热力图

Fig. 1　Heat map of chromosomal gene distribution for feature selection

通过 GA 选出的特征集合在很大程度上降低了特征间的共线性,从而提升各特征集的独立贡献度。由图 1 可知:经过充分的进化搜索,所筛选出的优秀个体在基因构成上表现出高度的收敛性,尤其是在某些关键特征位点上。例如,日航时、主机转速等特征对应的列呈现出近乎一致的深色,这表明他们是构成高适应度解的优势基因,在进化过程中被稳定地保留下来。基于该选择策略,本研究将 32 个原始特征降至 15 个,构成最终特征子集,见表 3。该措施不仅削减了约 53.1% 的计算复杂度,还为后续模型的快速迭代与训练提供更为精简而有效的输入空间。这一结果充分表明,基于 GA 的特征选择方法在船舶燃油消耗预测领域具有显著优势。该方法能在平衡预测精度和模型复杂度的基础上,挖掘出对燃油消耗最具解释力的关键特征。

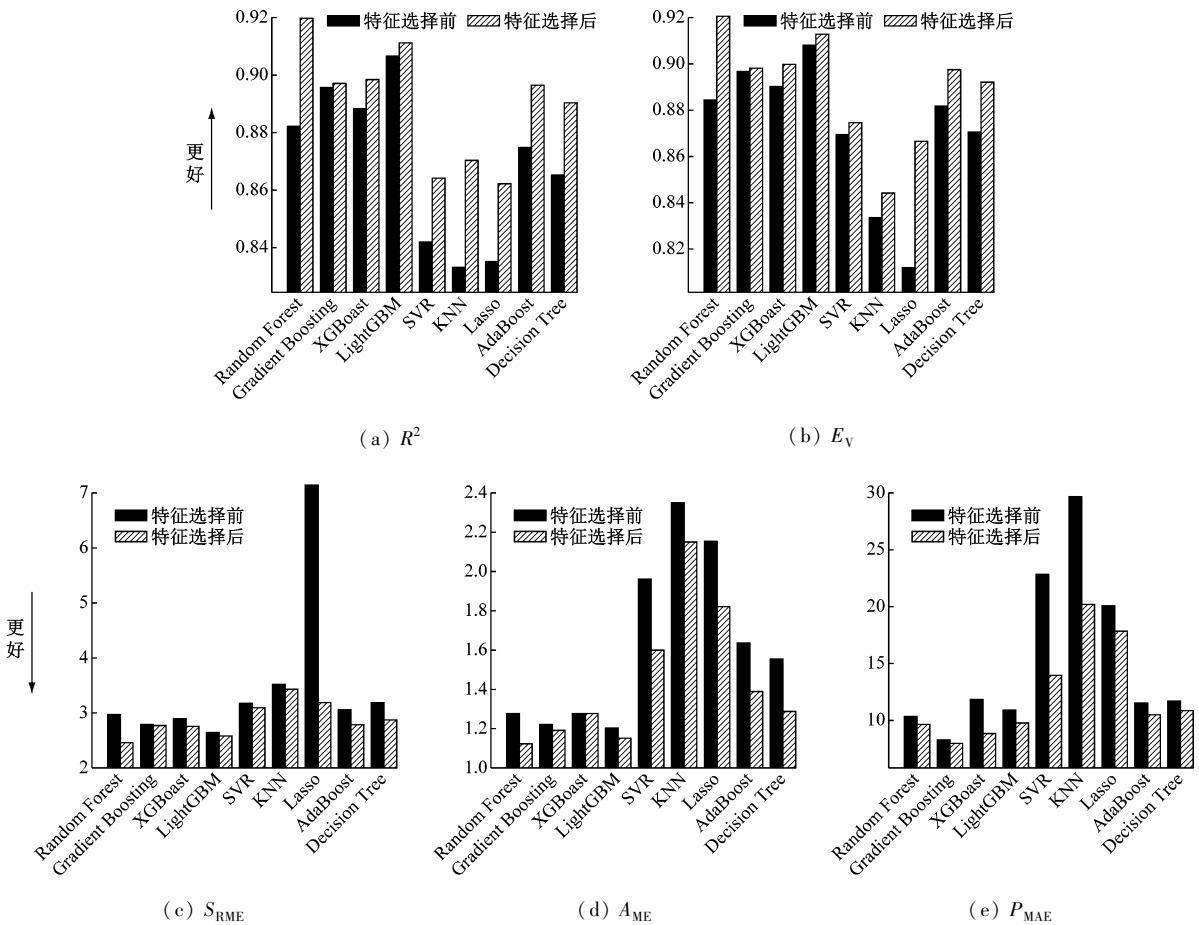
各特征对模型预测性能的贡献存在显著差异。其中,日航时、主机转速和大气温度等特征在 GA 筛选过程中被高频保留。此过程不仅降低模型的维度和计算成本,还有效减少特征冗余,提升模型对燃油消耗的预测稳定性和解释性。

表 3　经 GA 筛选后的最终特征子集

Tab. 3　Final feature subset selected by the Genetic Algorithm

序号	特征名称
1	吃水
2	日航程
3	距目的港剩余距离
4	日航时
5	日航速
6	气压
7	主机转速
8	海面上 10 m 风速
9	综合风浪和涌浪的显著波高
10	总涌浪平均周期
11	相对波向角度
12	相对风浪角度
13	总海洋流速
14	总潮汐速度
15	总波浪速度

为全面评估 GA 在特征选择中的有效性,本研究采用多种机器学习模型,对未进行特征选择和基于 GA 的特征选择两种方案在测试集上的预测性能进行了系统性对比分析见图 2。从 R^2、E_V、S_{RME}、A_{ME} 和 P_{MAE} 等多角度评估模型性能。采用 GA 进行特征选择后,各模型的 R^2 和 E_V 均呈现显著提升,说明精简后的特征集更能有效地捕捉燃油消耗的内在机理。同时,S_{RME} 和 A_{ME} 等衡量绝对误差的指标也显著降低,表明这一选择过程有效剔除了噪声与冗余信息,进而减少了预测偏差与不确定性。此外,在 P_{MAE} 这一相对误差指标上,GA 所选取的特征集也表现出更低的误差水平,进一步印证了该方法在刻画实际燃油消耗量方面的精确度。



(a) R^2 (b) E_V

(c) S_{RME} (d) A_{ME} (e) P_{MAE}

图 2 特征选择前后性能对比

Fig. 2 Performance comparison before and after feature selection

综上所述,基于 GA 的特征选择在本研究中取得良好的效果,能显著提升船舶燃油消耗预测模型的性能,尤其在处理融合高维复杂的船舶数据、气象和海浪数据时展现出优异的性能,为船舶运营优化和燃油消耗管理提供更为精确的理论支持。

2.2 船舶燃油消耗预测性能的比较分析

本研究采用 9 种不同的机器学习回归模型,对测试集上的预测表现进行系统对比,以评估各模型在船舶燃油消耗预测中的适用性和优势。9 种机器学习回归模型的预测值与实际值散点对比见图 3。图 3 中:拟合线及 95% 置信区间直观地反映模型的预测精度和稳定性。从散点分布情况分析,Random Forest、AdaBoost 和 Gradient Boosting 模型表现出较为优异的预测能力,其预测值与实际观测值分布较为接近理想的 45°参考线,充分说明这些集成学习方法能有效地捕捉船舶燃油消耗的复杂非线性特征。尤其是 Random Forest 模型,其预测结果如图 3a 所示,其预测点几乎全部紧密分布于拟合线周围,呈现出高度的一致性。相比之下,KNN 和 SVR 模型在燃油消耗较高的区域则出现一定程度的预测偏差,部分样本点明显偏离拟合线,这可能与其在捕捉极端工况下燃油消耗特性方面存在局限性有关。

基于泰勒图的模型性能评估结果见图 4。图 4 中:所有模型的相关系数均超过 0.90,表明各模型均能较好地捕捉燃油消耗的变化趋势。然而,在标准差指标上,各模型都表现出明显的差异。以 RF

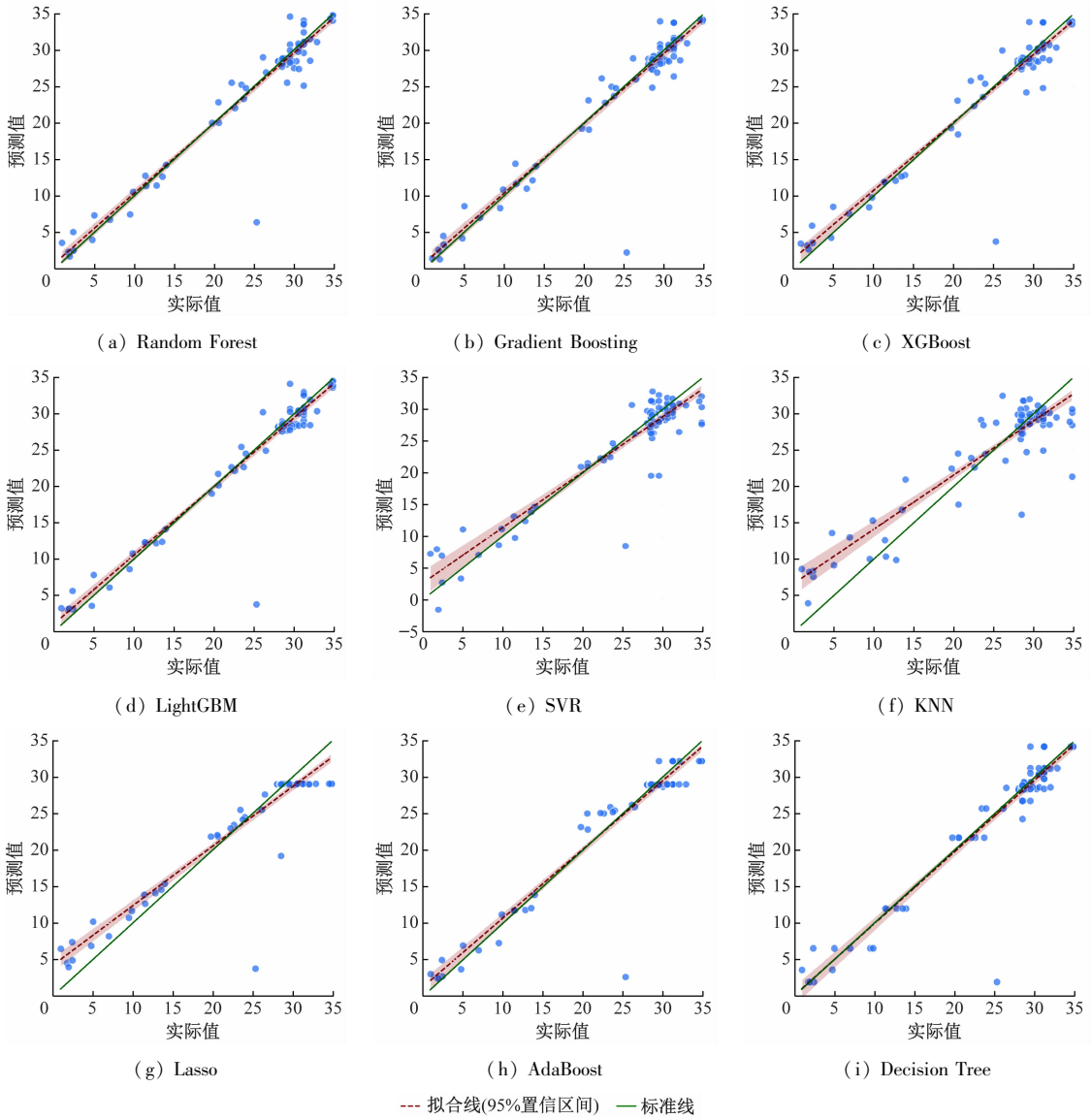


图 3　9 种机器学习回归模型的预测值与实际值散点对比

Fig. 3　Scatter comparison of predicted and observed values for nine machine learning regression models

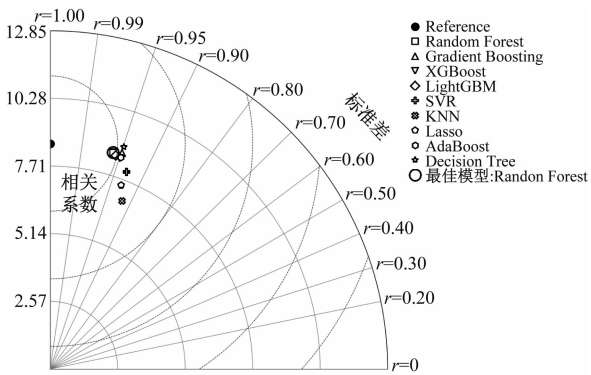


图 4　基于泰勒图的模型性能评估结果

Fig. 4　Results of model performance evaluation based on a Taylor diagram

模型为例,其标准差接近参考值,相关系数最大,显示出该模型不仅能准确预测燃油消耗的均值趋势,而且能有效地反映数据的波动特性。Gradient Boosting 和 AdaBoost 模型紧随其后,三者构成性能最优的模型集群,充分验证集成学习方法在复杂非线性系统建模中的优势。相比之下,KNN 和 Lasso 模型的标准差明显偏离参考值,反映出其在捕捉数据变异性方面存在不足,这可能与 KNN 模型对局部数据点过于敏感以及 Lasso 模型正则化过程中对部分重要特征的过度惩罚有关。

从实际应用角度考虑,RF 模型凭借其优异的预测精度和数据拟合能力,成为船舶燃油消耗预测的首选方法。该模型能有效地处理船舶运行环境中多变量的复杂交互作用,如气象条件、航速变化和吃水等因素对燃油消耗的综合影响。尽管 XGBoost 模型在许多预测任务中表现卓越,但在本研究的特定应用场景中,其性能略逊于 RF 模型,可能归因于 XGBoost 模型在参数调优上较高的复杂性与数据特性

之间匹配度不足。因此,模型选择应基于具体应用场景及数据特征,而非单纯追求算法理论上的先进性。

除预测精度和拟合稳定性外,模型的计算效率也是评估其工程适用性的重要维度,尤其是在未来可能需要进行在线更新或实时预测的应用场景中。为此,本研究进一步对 9 种模型的计算成本进行量化评估,主要考量反映模型离线构建复杂度的训练时间和决定模型在线响应能力的单样本平均预测时间两个关键指标。所有试验均在同一硬件平台(CPU:Intel® Core™ i7-14650HX;GPU:NVIDIA GeForce RTX 4060 Laptop GPU;RAM:16 GB)上执行,以确保比较的公平性。

不同模型在计算成本上表现出显著的异质性,见表 4。表 4 中:以 SVR、Lasso 和 DecisionTree 为代表的非集成或简单线性模型,其训练时间极低;以 LightGBM 和 XGBoost 为代表的先进梯度提升树模型,则凭借其高效的并行计算优化,在保持较高精度的同时,展现出卓越的训练速度。

表 4 不同模型的计算效率对比

Tab. 4 Comparison of computational efficiency for different models

模型名称	训练时间/s	单样本平均预测延迟/ms
RF	0.741	119.701
Gradient Boosting	0.744	3.077
XGBoost	0.415	13.459
LightGBM	0.225	3.308
SVR	0.036	2.787
KNN	0.005	4.339
Lasso	0.007	2.524
AdaBoost	1.216	53.362
Decision Tree	0.013	2.559

在进行精度与效率的权衡分析时,模型的综合效用变得更为清晰。值得注意的是,本研究最终选用的 RF 模型,其推理延迟相对较高,这归因于其需要对森林中每一棵独立的树进行预测并聚合结果。尽管如此,考虑到船舶燃油消耗预测这一特定应用场景,其百毫秒级的延迟对于航次级别的决策支持而言是完全可以接受的。更为重要的是,相较于推理速度更快的 GradientBoosting、LightGBM 等模型,RF 模型在预测精度和拟合稳定性上表现出更优的性能,这对于高经济价值的燃油消耗预测任务而言是首要考量。虽然 KNN 模型训练阶段的训练时间极短,但在预测阶段需要计算与所有训练样本的距离,导致其预测延迟相对较高,且精度表现不佳。

综上所述,综合考虑预测精度、拟合稳定性和实际应用需求,RF 模型为船舶燃油消耗预测提供了最为可靠的解决方案,其优异表现不仅为船舶能效管理提供了科学依据,也为未来开发更为精确的燃油消耗预测系统奠定了坚实的方法论基础。

2.3 船舶燃油消耗预测重要性及其驱动机制解析

2.3.1 燃油消耗预测的关键特征识别

基于 RF 模型对船舶燃油消耗预测进行深入剖析,采用特征重要性分析方法量化各变量对燃油消耗的贡献,并揭示模型决策过程中的核心驱动因素。具体而言,本研究采用基于基尼不纯度平均减少量(Mean Decrease Impurity, MDI)的度量标准。该方法通过计算每个特征在森林中所有决策树的构建过程中,作为分裂节点时平均带来的节点不纯度减少量来评估其重要性。一个特征的 MDI 值越大,表明其在划分数据、构建决策树模型时的贡献越大,因此被认为是更重要的预测因数。

RF 模型识别出 9 个关键特征及其相对重要性权重,见图 5。为解析船舶燃油消耗的影响机制提供了数据支撑。结果表明,时间维度变量在燃油消耗预测中占据主导地位。其中,日航时作为反映单日连续航行时长的关键变量,其重要性权重高达约 0.7,这充分印证了发动机持续功率输出与燃油消耗之间的线性关系。这一结果与第 2.1 节中通过 GA 筛选过程中日航时被高频保留的发现高度一致,进一步验证不同特征评估方法在识别关键预测因数上的内在一致性。此外,作为累积航行时间的度量,总航时的重要性权重约 0.2,表明燃油消耗受航行时间累积效应的显著影响,可能与发动机运行效率随时间变化的特征相关。尽管日航程和主机转速在物理直觉上应与燃油消耗密切相关,但其重要性权重显著低于时间相关变量。这一结果挑战了传统认

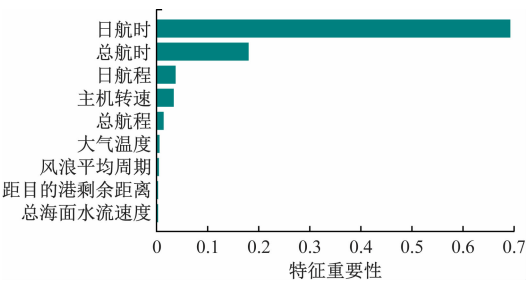


图 5 RF 模型的特征重要性

Fig. 5 Importance of features in Random Forest model

知,表明在实际航行环境中,时间变量可能间接融合了航行模式、航速变化和动力输出等多重因素的影响。该发现不仅与船舶动力学基本原理高度契合,还为燃油消耗预测提供了明确的变量选择依据。

2.3.2　特征重要性的层级结构分析

通过对特征重要性的定量评估可知,RF 模型呈现出明显的层级结构。总体上,可将所有特征划分为核心预测因数(如日航时、总航时)、次级贡献因数(如日航程、主机转速)和辅助影响因数(包括大气温度、风浪平均周期等)等 3 个梯度。

1)核心预测因数累计贡献了近 90% 的预测能力,表明在复杂航行环境中,发动机运行时长作为能量转换的关键变量起到决定性作用。

2)次级贡献因数虽贡献度较低,但仍在捕捉船舶空间位移与动力输出强度对燃油消耗的调节效应方面发挥着不可或缺的作用。

3)辅助影响因数虽然在单一指标上的权重偏低,但通过与主要特征的协同作用,间接反映环境因素对燃油消耗的累积影响。

2.3.3　特征重要性分析对船舶能效管理的启示

RF 模型的特征重要性分析不仅为船舶燃油消耗的预测提供了量化依据,还对实际能效管理提出了深刻启示。

1)鉴于日航时和总航时在模型中的主导地位,优化航行时间应成为降低燃油消耗的首要策略,这可以通过合理规划航线、优化航行路径以及避开不利海况来实现。

2)尽管主机转速的重要性相对较低,但作为一种直接可控的操作参数,其动态优化(例如在不同航段采用变速策略)同样具有显著的燃油节约潜力。

3)虽然大气温度、风和浪条件等环境因素的单项重要性不高,但其累积效应不可忽视,提示在长航程规划中,通过气象导航优化进一步降低燃油消耗的方式仍具备实际意义。

综上所述,RF 模型的特征重要性分析不仅揭示船舶燃油消耗的内在驱动机制,还为船舶能效管理提供定量决策依据。这些发现为构建智能船舶能效管理系统,制定精准减排策略以及推动航运业绿色转型奠定了坚实的理论基础。

2.4　模型决策过程与特征交互的可解释性分析

为超越静态特征重要性排序的局限,深入探究所构建的 RF 模型如何将输入特征映射至燃油消耗预测,并揭示特征间的非线性交互机制,本文引入偏依赖图(Partial Dependence Plot, PDP)和个体条件期望图(Individual Conditional Expectation,ICE)进行深度可解释性分析[23-24]。PDP 通过对协变量空间进行边缘化处理,能揭示目标特征对模型预测的平均边际效应,从而阐明模型的宏观决策趋势[23]。然而,PDP 本质上是一种全局平均解释,可能会掩盖由特征交互引起的个体预测行为的异质性。为此,本研究进一步引入 ICE,其通过可视化每个独立数据实例的预测响应轨迹,为审视模型微观决策行为提供了窗口[24]。两者的协同分析,能够系统性地解构模型的复杂响应曲面。

模型中最重要的 9 个预测因数的 PDP 与 ICE 分析结果,见图 6。不同特征的边际效应模式呈现出显著差异。例如:对于日航时如图 6a 所示,其 PDP 呈现出清晰的非线性递增趋势,且绝大多数 ICE 紧密地簇拥于 PDP 曲线周围,表现出高度的同质性。这表明日航时对燃油消耗存在一个强大的、稳定的主效应,其影响在不同数据实例中表现出高度的一致性,受其他特征的调节作用相对有限;在主机转速的分析中,揭示了截然不同的、更为复杂的现象如图 6c 所示。尽管其 PDP 曲线揭示一个复杂的平均响应模式,但其 ICE 曲线表现出极大的异质性,形态各异且在预测空间中广泛散布。这种平均效应(PDP)与个体效应(ICE)的显著背离,是预测因数间存在强烈交互作用的典型信号。这一发现科学地表明,主机转速的边际效应并非一个固定函数,而是被其他一个或多个特征(如海况、吃水等)所动态调节;综合风浪和涌浪的显著波高以及相对波向角度等环境因素的分析也呈现出类似的模式如图 6d 和图 6f 所示,即相对平坦的 PDP 曲线与广泛分散的 ICE 曲线并存。这进一步印证环境因素对燃油消耗的影响并非孤立的线性叠加,而是通过与船舶操作参数的复杂非线性耦合来共同驱动最终的能耗表现。

综上所述,PDP 与 ICE 的协同分析不仅直观地打开了模型的“黑箱”,揭示其内部的决策逻辑,更成功地识别并可视化了关键驱动因素间的复杂交互作用。这一发现不仅从理论上验证本研究采用能捕捉高阶交互效应的非线性模型的必要性和优越性,也为航运实践中制定超越“单一变量控制”的,考虑多因素耦合的精细化,动态化能效管理策略,提供坚实的定量决策依据。

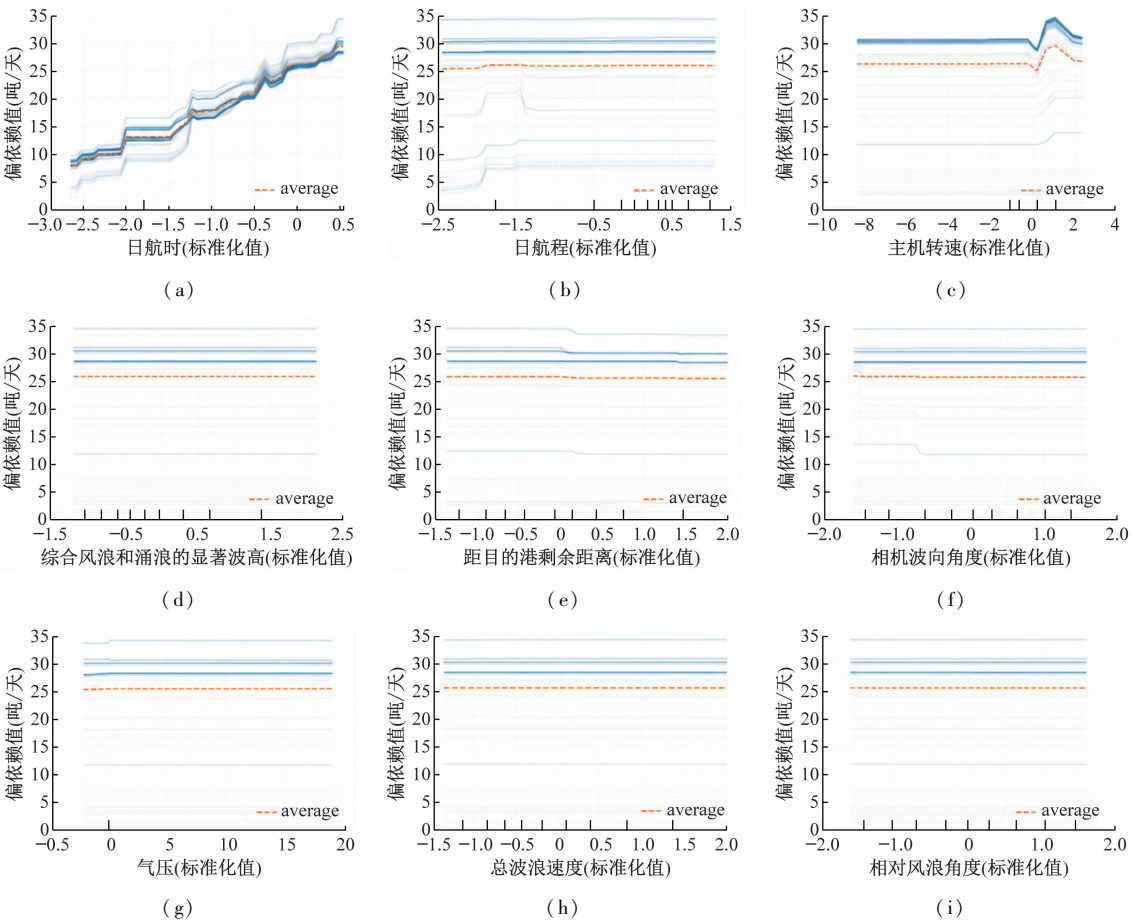


图 6 关键预测因数的边际效应与个体条件期望分析

Fig. 6 Marginal effects and individual conditional expectations analysis for key predictors

3 结 语

本研究针对船舶燃油消耗预测问题，构建一个基于多源异构数据融合与机器学习方法的综合预测框架，并围绕特征筛选、模型比较和关键特征分析展开系统研究。研究结论如下。

1）本研究整合船舶正午报告数据、ECMWF 再分析数据、全球海洋物理分析与预测数据和 AIS 数据，构建一个数据准确性高、覆盖面广的综合数据集，为燃油消耗预测提供坚实的数据基础。

2）针对 32 个原始特征，本研究采用 GA 进行特征选择，有效降低特征冗余与共线性问题，将输入维度精简至 15 个。试验结果证明，所筛选出的特征集不仅大幅降低了计算复杂度（降低约 53.1%），而且显著提升了模型的预测能力。

3）通过对 9 种机器学习模型的系统比较，试验结果表明，RF 模型在燃油消耗预测任务中表现最为优异，具备良好的非线性映射能力和特征交互捕捉能力，预测精度和数据拟合稳定性均优于其他模型。

4）基于 RF 模型的特征重要性分析结果显示，日航时和总航时等时间维度变量在燃油消耗预测中占据主导地位，揭示时间变量在燃油消耗机理中的核心作用，为船舶能效管理和燃油优化提供了定量决策依据。

本研究在单一散货船数据集上构建并验证了预测模型，虽然在该案例中取得良好效果，但其在不同船型、航线和极端海况下的泛化能力仍待进一步验证。此外，经 GA 筛选的特征子集主要针对当前数据集，是否具备跨场景的普适性仍需在更大规模、多样化的数据上加以检验。未来研究可通过扩展多源异构数据库及引入迁移学习等方法，提升模型的鲁棒性与适用性。

参 考 文 献

[1] International Maritime Organization. Fourth IMO GHG Study 2020[R]. London: IMO, 2021.

[2] International Maritime Organization. IMO data collection system[R]. London: IMO, 2024.

[3] International Maritime Organization. Report of fuel oil consumption data submitted to the IMO Ship Fuel Oil Consumption Database in GISIS (Reporting year: 2023)[R]. London: IMO, 2024.

[4] 包书景，熊浩，何昌福，等. 笔谈：新质生产力发展与

“双碳”目标互融共赢[J]. 华中师范大学学报, 2025, 59(1): 1-8.
Bao S J, Xiong H, He C F, et al. Symposium: Integrated development of new quality productive forces and the "Dual Carbon" goals[J]. Journal of Central China Normal University, 2025, 59(1): 1-8. (in Chinese)

[5] Sarris E. Naval ship propulsion and electric power systems selection for optimal fuel consumption[D]. Massachusetts: Massachusetts Institute of Technology, 2011.

[6] Kim K S, Roh M I, Engineering. ISO 15016: 2015-based method for estimating the fuel oil consumption of a ship[J]. Journal of Marine Science and Engineering, 2020, 8(10): 791.

[7] 张志政, 王冬捷, 张勇亮. 基于PSO改进KPCA-SVM的故障监测和诊断方法研究[J]. 现代制造工程, 2020(9): 101-107.
Zhan Z Z, Wang D J, Zhang Y L. Fault monitoring and diagnosis method based on PSO-improved KPCA-SVM [J]. Modern Manufacturing Engineering, 2020(9): 101-107. (in Chinese)

[8] Yan R, Wang S, Du Y, et al. Development of a two-stage ship fuel consumption prediction and reduction model for a dry bulk ship[J]. Transportation Research Part E: Logistics and Transportation Review, 2020, 138: 101930.

[9] 刘博. 基于LSTM神经网络-灰箱模型的船舶油耗估算研究[D]. 大连: 大连海事大学, 2023.
Liu B. Research on ship fuel consumption estimation based on LSTM neural network and grey-box model[D]. Dalian: Dalian Maritime University, 2023. (in Chinese)

[10] Safae A A, Hassan G, and Ghiasi M. VLCC′s fuel consumption prediction modeling based on noon report and automatic identification system [J]. Cogent Engineering, 2019, 6(1): 1595292.

[11] Fan A, Yang J, Yang L, et al. A review of ship fuel consumption models [J]. Ocean Engineering, 2022, 264: 112405.

[12] Feng Y, Wang X, Chen Q, et al. Prediction of the severity of marine accidents using improved machine learning[J]. Transportation Research Part E: Logistics and Transportation Review, 2024, 188: 103647.

[13] Soner O, Akyuz E, Celik M. Use of tree based methods in ship performance monitoring under operating conditions [J]. Ocean Engineering, 2018, 166: 302-310.

[14] Tarelko W, Rudzki K. Applying artificial neural networks for modelling ship speed and fuel consumption [J]. Neural Computing and Applications, 2020, 32(23): 17379-17395.

[15] UYANıK T, Karatuğ Ç, Arslanoğlu Y. Machine learning approach to ship fuel consumption: a case of container vessel[J]. Transportation Research Part D: Transport and Environment, 2020, 84: 102389.

[16] Trodden D G, Murphy A J, Pazouki K, et al. Fuel usage data analysis for efficient shipping operations[J]. Ocean Engineering, 2015, 110: 75-84.

[17] Che J, Xia W, Xu Y, et al. Multivariate wind speed forecasting with genetic algorithm-based feature selection and oppositional learning sparrow search [J]. Information Sciences, 2025, 695: 121736.

[18] Rostami M, Berahmand K, Forouzandeh S. A novel community detection based genetic algorithm for feature selection[J]. Journal of Big Data, 2021, 8(1): 2.

[19] 郭慧, 刘明艳. 基于遗传算法和随机森林的入侵检测方法研究[J]. 计算机应用与软件, 2024, 41(1): 304-309.
Guo H, Liu M Y. Research on intrusion detection method based on genetic algorithm and random forest [J]. Computer Applications and Software, 2024, 41(1): 304-309. (in Chinese)

[20] Nguyen V N, Chung N, Balaji G N, et al. Internet of things-driven approach integrated with explainable machine learning models for ship fuel consumption prediction[J]. Alexandria Engineering Journal, 2025, 118: 664-680.

[21] 张大恒, 张英俊, 张闯. 基于BP神经网络的船舶气象航线决策系统[J]. 中国舰船研究, 2022, 17(4): 98-106.
Zhang D H, Zhang Y J, Zhang C. Ship weather routing decision-making system based on BP neural network [J]. Chinese Journal of Ship Research, 2022, 17(4): 98-106. (in Chinese)

[22] 周翔宇, 吉哲, 王凤武. 基于EEMD-PE-LSTM的短时船舶交通流量预测与航道交通状态可视化[J]. 大连海事大学学报, 2023, 49(2): 58-68.
Zhou X Y, JI Z, Wang F W. Short-term ship traffic flow prediction and channel traffic state visualization based on EEMD-PE-LSTM [J]. Journal of Dalian Maritime University, 2023, 49(2): 58-68. (in Chinese)

[23] 李聪颖, 袁锴璐, 李静怡, 等. 基于分层时空框架的共享单车需求预测:LP-TFT模型与SHAP可解释分析[J]. 中国公路学报, 2025, 38(10): 305-323.
Li C Y, Yuan K L, Li J Y, et al. Shared bicycle demand prediction based on hierarchical spatiotemporal framework: LP-TFT model and SHAP interpretable analysis[J]. China Journal of Highway and Transport, 2025, 38(10): 305-323. (in Chinese)

[24] 尹力, 魏伟, 李泓锐, 等. 欧洲“三区”空间转型中的冲突演化特征、模拟预测与政策启示[J]. 经济地理, 2025,45(6): 11-25.
Yin L, Wei W, Li H R, et al. Conflict evolution characteristics, simulation prediction and policy implications in the spatial transformation of the “three regions” in Europe[J]. Economic Geography, 2025, 45(6): 11-25. (in Chinese)