Advances in Applied Mathematics 应用数学进展, 2023, 12(9), 4006-4012
Published Online September 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.129392

Hans汉斯

基于多模型融合的二手帆船价格评估模型

曹子昂，杨玉琦，阳鹏程，王明春*

湖南农业大学信息与智能科学技术学院，湖南 长沙

收稿日期：2023年8月12日；录用日期：2023年9月6日；发布日期：2023年9月14日

摘　要

本文搜集了不同类型帆船的制造商、年份、长度、横梁尺寸、吃水深度、排水量、满载水线、船帆面积、平均货物吞吐量、出售区域，单双体的数据，同时搜集了帆船出售地的人均GDP、平均气温、平均降水量的数据。检查并剔除了数据集中的异常值，对于二手帆船挂牌价格做对数变换的处理，使其更符合正态的分布。基于随机森林、XGBoost、GBDT、LightGBM模型，模型优化阶段以网格搜索算法进行超参数优化，构建了四种可以对二手帆船对数价格进行估计的模型，其中随机森林的估计精度达到96.61%，其余三种模型的估计精度也可达到95%以上。在此基础上本文以Stacking模型融合对四种单体模型进行综合融合，最终得到拟合度更高的融合模型，其对数价格估计的均方误差接近于0.01，此融合模型具有较高的准确性。利用该模型，可以对二手帆船进行合理的定价。

关键词

随机森林，XGBoost，GBDT，LightGBM，Stacking模型融合

Second-Hand Sailing Price Evaluation Model Based on Multi-Model Fusion

Ziang Cao, Yuqi Yang, Pengcheng Yang, Mingchun Wang*

College of Information and Intelligence, Hunan Agricultural University, Changsha Hunan

Received: Aug. 12th, 2023; accepted: Sep. 6th, 2023; published: Sep. 14th, 2023

Abstract

This paper collects the data of manufacturers of different types, year, length, beam size, draft depth, displacement, full load waterline, sail area, average cargo throughput, sale area, and single pair of sailing land, and collects the data of per capita GDP, average temperature and average precipitation of sailing land. Check and eliminate the outliers in the data set, and log-transform the listing price

*通讯作者。

文章引用：曹子昂，杨玉琦，阳鹏程，王明春. 基于多模型融合的二手帆船价格评估模型[J]. 应用数学进展，2023, 12(9): 4006-4012. DOI: 10.12677/aam.2023.129392

of the second-hand sailboat to make it more in line with the normal distribution. Based on the random forest, XGBoost, GBDT and LightGBM models, the model optimization stage uses the grid search algorithm to build four models that can estimate the log price of second-hand sailboats. Among them, the estimation accuracy of random forest reaches 96.61%, and the estimation accuracy of the other three models can also reach more than 95%. On this basis, this paper comprehensively integrates the four monomer models with Stacking model fusion, and finally obtains the fusion model with higher fit. The mean square error of the log price estimation is close to 0.01. This fusion model has high accuracy. Using this model, second-hand sailboats can be reasonably priced.

Keywords

Random Forest, XGBoost, GBDT, LightGBM, Stacking Model Fusion

1. 引言

然而相对于新船交易，二手帆船因为其“一船一况”的特性在交易方面比一般的新船交易要复杂得多，这主要是因为二手帆船的价格难以准确估计和设定。随着市场的发展，一套客观的二手帆船价格评估方法是很重要的，在[1] [2] [3] [4]中，各作者建立了多元回归模型来预测价格、基于随机森林算法的二手帆船定价估计模型、基于 LightGBM 模型的二手帆船价格分析模型、人工蜂群-BP 神经网络的二手帆船价格估算。这些模型都较为单一，各自集中在特定的算法或方法上，而未充分考虑其他关键方面。在模型建立和估价问题中，模型的多样性和综合性能对于获得更准确的结果至关重要。特别是，在现实世界的复杂问题中，单一模型可能无法捕捉到所有的变化和模式，因此需要更综合的方法来应对挑战。本文将收集一些二手帆船数据(见附录)，基于这些数据，本文将建立多模型融合的二手帆船价格评估模型。

2. 数据的预处理和描述性分析

经过前期的数据搜集，我们得到数据集，包括 14 个变量，有帆船的制造商、年份、长度、横梁尺寸、吃水深度、排水量、满载水线、船帆面积、平均货物吞吐量、出售区域，单双体、出售地的人均、平均气温、平均降水量，然后对缺失值进行了补充或者删除，对异常值予以了剔除，进一步得到了更加完备的数据集。

本文首先对数据进行描述性统计分析，以发现各变量对于二手帆船的挂牌价格是否存在某些直观的可见影响，同时为后期的模型建立提供一定的帮助。对帆船的挂牌价格进行描述统计，绘制单体帆船挂牌价格分布的直方图，如图 1(左)所示。二手单体帆船价格大部分位于(150,000$, 400,000$)区间，而二手双体帆船价格大部分位于(400,000$, 800,000$)区间，同时可以发现，挂牌价格均呈现左偏的分布。所以为后期模型构建方便，对挂牌价格做取对数值的变换处理，并对其分布做正态 QQ 图检验其正态性，绘制的正态 QQ 图，如图 1(右)所示。

由正态 QQ 图可见，对数变换后的挂牌价格可以大致上看作近似的正态分布了，这对后期的模型精度的提高十分重要。本文发现数据集中大部分变量对于二手帆船的挂牌价格存在一定的影响，但如果仅仅只考察某一单一变量对挂牌价格的影响，其效果是比较模糊的，所以二手帆船的挂牌价格不仅仅只由某一单一变量起主导作用，而是同时受各个变量的影响。其次各变量之间或许存在某种交互作用，即二手帆船的不同配置对挂牌价格具有影响。

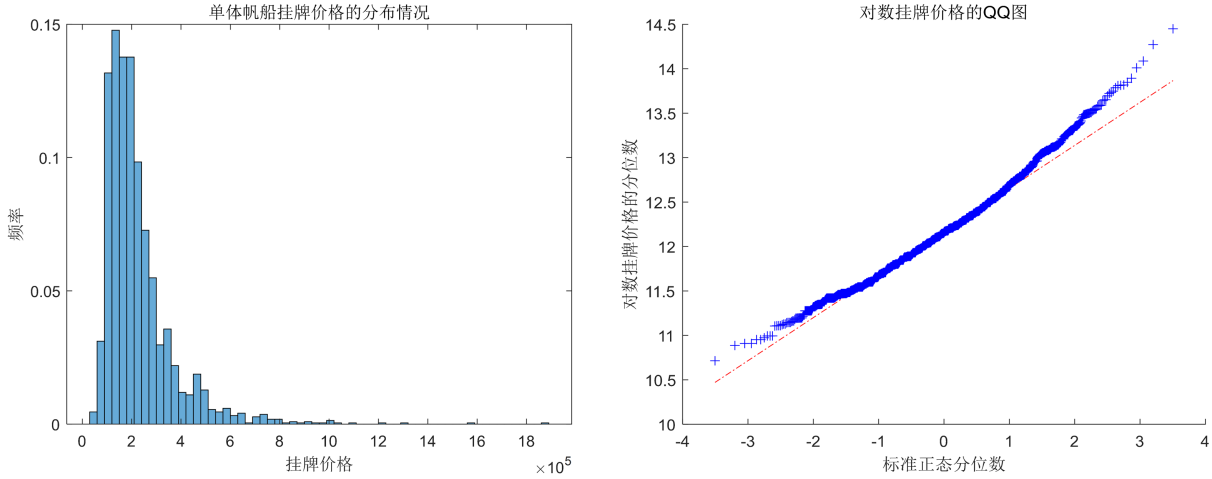

Figure 1. Second-hand single sailing boat listing price distribution (left) and log price QQ chart (right)
图 1. 二手单体帆船挂牌价分布(左)和对数价格 QQ 图(右)

3. 二手帆船价格评估模型的建立

3.1. 准备工作

经过数据的收集和简单的预处理，得到的数据集还需要进一步处理，主要是对各变量进行标准化处理，后续模型构建均基于此标准化后的数据。同时对变量制造商和区域分别进行 K-means 聚类，本文将所有制造商和出售区域依据肘部法则，聚类划分为 20 个类别和 6 个区域，在一定程度上可以削弱数量少的制造商和出售区域给模型带来的误差，后期模型构建均基于此。本文选取随机森林[5]、XG Boost [6]、GBDT [7]、Light GBM [8]四种策略，对二手帆船的挂牌价格进行模型构建，并分别比较各模型的优缺点，最后以 Stacking [9]模型融合的方式对四种模型进行融合，保证最终的“二手帆船定价模型”具有各模型的优点，同时削弱各模型的缺点。

3.2. 模型假设

1) 重点考虑已搜集的变量对挂牌价格的影响，忽略其它因素对价格的影响。

理由：在现实中，二手帆船的定价由多个因素共同决定，我们考虑对其价格影响较大的因素。对于其他影响较小的因素，同时结合他们之间的正负相互效益，可以忽略他们对价格的影响。

2) 忽略美元在 2005 年到 2023 年的货币通货膨胀率。

理由：2005 年到 2023 年，虽然全球经济都得到快速发展，但是对于奢侈品帆船的价格来说，其波动幅度基本维持稳定，所以可以忽略通货膨胀率带来的影响。

3) 在评估不同地区的二手帆船价格时，我们忽略当地政策对帆船价格的影响。

理由：全球出售二手帆船的地方众多，当地政策对帆船定价影响属于极个别地区，这种情况的二手帆船价格往往要视情况而定。

3.3. 四个单体模型的建立

本文这一部分将 14 个变量作为自变量，对数价格作为因变量，借助 python 的机器学习的模型算法，以二手帆船数据集为基础，利用 K-fold 算法将其随机分为训练集和测试集，分别构建“基于随机森林的二手帆船定价模型”、“基于 XGBoost 的二手帆船定价模型”、“基于 GBDT 的二手帆船定价模型”以及“基于 LightGBM 的二手帆船定价模型”。

本文将这四类单体模型对变量重要性的计算结果、对数估价与对数现实二手价的精度情况、对数估价与对数现实二手价接近程度得分，分别放置在一起进行比较来分析各模型的优势。特别说明，由于文章篇幅问题，将只给出基于随机森林的二手帆船定价模型的数据图，其余三种单体模型是类似的。

首先是各变量对于挂牌价格影响的重要性，如下图 2 是基于随机森林的二手帆船定价模型变量的重要性，横坐标为各变量因素，纵坐标为变量对于挂牌价格的影响的重要程度。可以发现，四种模型对于变量重要性的测度结果都有一定的差别，但经过对比发现，制造商、横梁尺寸、排水量、年份这 4 个变量对于挂牌价格的影响十分重要。

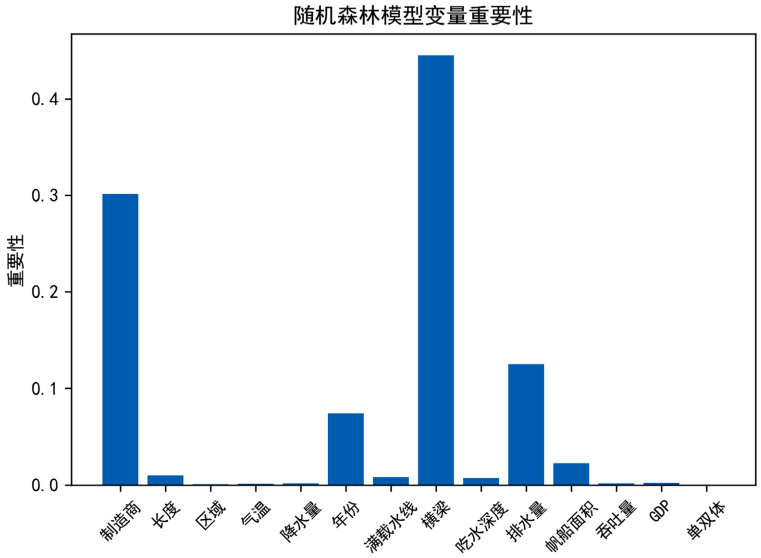

Figure 2. Importance of random forest variable
图 2. 随机森林变量的重要性

其次是对数估价与对数现实二手价的精度情况，如下图 3(左)，可以发现基于本文使用的这四种算法模型的估计值与实际值的拟合程度均高于 95% ($R^2 > 0.95$)，每一个单体模型的估计效果均表现出较高的准确性，但都有一定程度的缺陷。对四种单体模型的精度效果进行汇总，主要考察其均方误差和 R^2，如下表 1 所示：

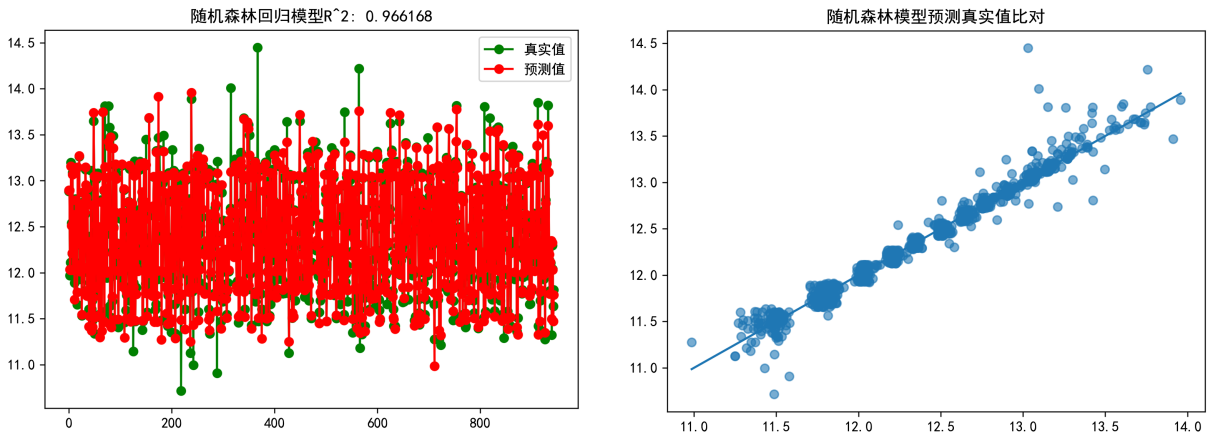

Figure 3. Random forest regression (left); Random forest prediction (right)
图 3. 随机森林回归(左)；随机森林预测(右)

Table 1. Comparison of the effect of four models
表 1. 四种模型的效果比较

	平均绝对误差	平均平方误差	均方根误差	R^2	得分
GBDT	0.07058	0.01780	0.13344	0.95039	0.93788
LightGBM	0.08089	0.01655	0.12867	0.95387	0.93788
XGBoot	0.07118	0.01356	0.11648	0.96220	0.93979
随机森林	0.06079	0.01214	0.11020	0.96616	0.96633

最后，对于每一单体模型的对数估价与对数现实二手价接近程度得分进行比较，如上图 3(右)所示。每一个散点表示预测值与真实值之间的关系，其分布越靠近蓝色实线的周围即表示其对数估价数值与实际对数挂牌价格的误差越小。可以发现，在单体模型中，基于随机森林的二手帆船定价模型的得分是最高的。

基于以上对四种单体模型的求解以及结果可视化展示，可以发现基于随机森林的二手帆船对数定价模型的估价与实际对数挂牌价是最接近的，已经具有较好的估计效果，但本文尝试以这四种单体模型作为 Stacking 的第一层基础模型，以其他模型作为第二层融合模型，构建新的融合模型，以期待对挂牌价格估计得更加准确。

3.4. 模型的融合

Stacking 融合算法可以将众多模型设定为第一层基础模型，综合各个模型结果的优点，以每一单体模型的结果重新作为一个新的数据集。选择第一层中性能表现最好的模型作为第二层的模型或者选择其它模型作为第二层的模型，经过多次尝试，本文选取线性回归作为融合的第二层模型，并以新数据集重新划分训练集和测试集进行模型的训练，得到最终结果。过程可见图 4。

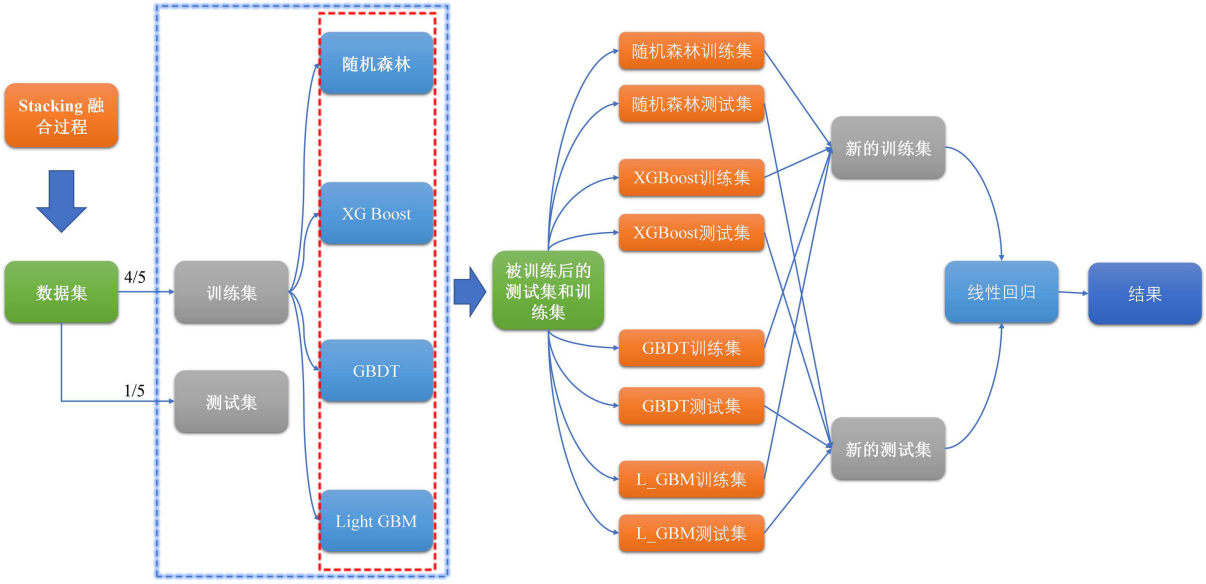

Figure 4. Stacking model fusion process
图 4. Stacking 模型融合过程

根据融合模型的分析以及构建思路，本文借助 Python 构建其算法模型，经过 Stacking 融合后的模型对于二手帆船的对数价格估计值的平均绝对误差是最接近 0.01 的，这也正好证明融合模型比四个单体模

型更具优良性。其中模型优化过程为：1) 利用网格搜索算法进行超参数优化；2) 利用 Stacking 融合模型算法对上述四种模型进行融合。

经过融合的模型表现出良好的估计效果，其对数价格的估计精度均高于每一单体模型，同时在测试集上的精度可达 96%以上。

4. 灵敏度分析

由于未来形势的不确定性，将气温、降水和人均 GDP 作为一定范围内变化的指标，其他参数保持不变，得到预测价格。然后观察到价格变化的程度。敏感性分析结果如图 5 所示，说明由于气温、降水和人均 GDP 的变化，它们对预测价格的影响不显著。这说明，当气候、降水和人均 GDP 的值在合理范围内时，它们对结果的影响不大。分析结果表明，该模型具有良好的稳定性和通用性。

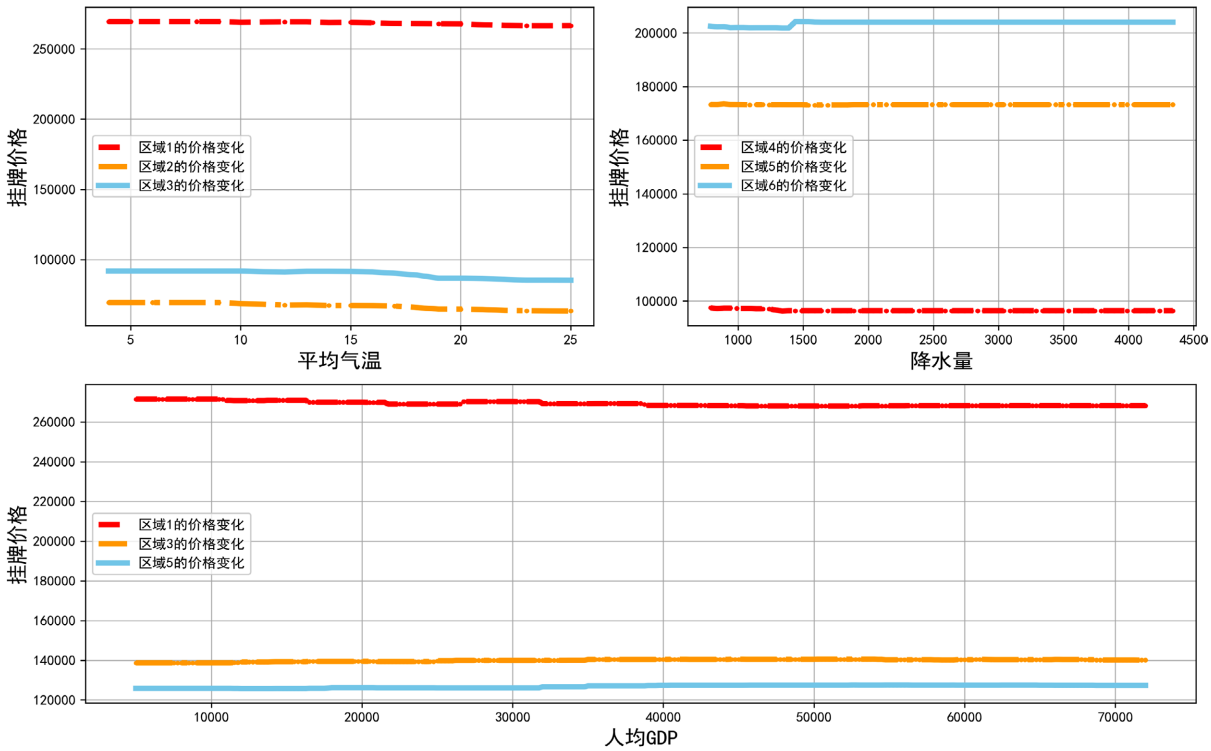

Figure 5. Sensitivity analysis
图 5. 灵敏度分析

5. 模型的优缺点和改进

5.1. 模型的优点

1) 本文模型的泛化性能较好，可以适用于不同地区的情况。

2) 本文采取了 Stacking 模型融合的方法，将各个模型的效果取长补短，提高模型的稳定性和表现。

3) 本文构建的融合模型不是单纯地对二手帆船的挂牌价格进行估计，而是对二手帆船价格的对数进行估计，精度相较原来有一定的提升。

5.2. 模型的缺点

1) 由于数据的搜集过程十分繁琐冗余，一些涉及商业机密的数据难以爬取，所以本文模型对于可能

影响二手帆船价格的一些变量并没有考虑到，诸如：引擎马力、帆船电子产品、帆船内部休息空间。考虑的变量有限，这对模型的精度有一定的影响。

2) 对一些特殊的二手帆船，该模型难以估计，比如具有极强的收藏价值的，它们的性能可能没有其他二手帆船好，甚至年代有久远，模型的评估难免会有较大误差。

5.3. 模型的改进

可以进一步收集其他因素的数据，使模型精度更高。我们可以将保值率代替对数价格，可以更加精准地对二手帆船的价格进行评估，保值率计算公式为：$\log\left(\frac{\text{原价}}{\text{原价}-\text{当前价格}}\right)$。

参考文献

[1] Zhang, J.H. and Zhang, Z.S. (2023) Modeling and Solving Used Sailboat Market Strategy and Pricing Method. *Highlights in Business, Economics and Management*, **16**, 612-620. https://doi.org/10.54097/hbem.v16i.10674

[2] Yang, C.Y., Tang, S.N. and Chen, J.H. (2023) An Estimation of the Pricing of Second-Hand Sailboats Based on the Random Forest Algorithm. *Proceedings of the 2nd International Conference on Mathematical Statistics and Economic Analysis, MSEA* 2023, Nanjing, 26-28 May 2023, 9 p. https://doi.org/10.4108/eai.26-5-2023.2334481

[3] Gao, X., Zhu, J. and Yang, R. (2023) Price Analysis of Used Sailboats Based on LightGBM Model. *Highlights in Science, Engineering and Technology*, **53**, 168-176. https://doi.org/10.54097/hset.v53i.9709

[4] Wang, M., Lu, B. and Wang, H. (2023) Exploring the Market: Used Sailboat Price Estimates Based on Artificial Bee Colony-BP Neural Network. *Highlights in Business, Economics and Management*, **16**, 72-79. https://doi.org/10.54097/hbem.v16i.10539

[5] 宋玉华, 王子晓, 李焕群, 王珺. 一种基于随机森林模型的消防监督检查频率预测方法[J]. 中国人民警察大学学报, 2023, 39(2): 51-56.

[6] 李威, 刘检生, 施增强, 郭万里. 基于 XGBoost 算法的堆石料南水模型参数反演及应用[J]. 水利水运工程学报, 2023(3): 111-120.

[7] 皮理想, 崔桂梅. 进化算法优化 GBDT 的带钢卷取温度预测[J]. 华南师范大学学报(自然科学版), 2022, 54(1): 122-127.

[8] 柯于锭, 陈可. 基于 LightGBM 的心血管疾病预测模型研究[J]. 信息与电脑(理论版), 2022, 34(13): 71-73+78.

[9] 孙昭, 李云, 江毓武, 等. 基于 Stacking 机器学习模型的南海北部海温预报[J]. 海洋预报, 2023, 40(1): 39-45.

附　录

数据	网址
2023 年美国大学生数学建模竞赛春季赛试题	https://www.contest.comap.com/undergraduate/contests/mcm/contests/2023/problems/
各个型号帆船的信息	https://www.sailboatlistings.com/sailboats_for_sale/ https://www.yachtworld.com/boats-for-sale/make-bali/model-4.5/
地区气候数据	https://www.usclimatedata.com/climate/maine/united-states/3189 https://www.climatestotravel.com/climate/nigeria/lagos
地区经纬度数据	https://www.latlong.net/
世界各地区人均 GDP 数据	https://data.worldbank.org.cn/