

基于随机森林回归的奥运奖牌预测模型

郑雨欣*, 赵圣博*, 钟汶妍, 吕平#

杭州师范大学数学学院, 浙江 杭州

收稿日期: 2025年3月2日; 录用日期: 2025年3月21日; 发布日期: 2025年4月2日

摘要

奥运会奖牌榜在很大程度上代表了一个国家的体育实力, 对未来奖牌榜进行预测对国家有针对性地提升体育竞争力有重大意义。本研究提出了一种基于随机森林回归模型的奖牌预测框架, 通过数据预处理、模型构建及评估, 预测了各国在2028年洛杉矶奥运会上的奖牌分布。

关键词

随机森林回归模型, 奥运会奖牌预测, 机器学习

A Random Forest Regression-Based Model for Olympic Medal Prediction

Yuxin Zheng*, Shengbo Zhao*, Wenyan Zhong, Ping Lyu#

School of Mathematics, Hangzhou Normal University, Hangzhou Zhejiang

Received: Mar. 2nd, 2025; accepted: Mar. 21st, 2025; published: Apr. 2nd, 2025

Abstract

The Olympic medal table represents the sports strength of a country to a large extent. It is of great significance to predict the future medal table for the country to enhance the sports competitiveness. This study proposes a medal pre-diction framework based on the random forest model, involving data preprocessing, modeling, and evaluation to forecast the medal distribution of countries in the 2028 Los Angeles Olympics.

*共同第一作者。

#通讯作者。

文章引用: 郑雨欣, 赵圣博, 钟汶妍, 吕平. 基于随机森林回归的奥运奖牌预测模型[J]. 统计学与应用, 2025, 14(4): 47-54. DOI: 10.12677/sa.2025.144087

Keywords

Random Forest Regression Model, Olympic Medal Prediction, Machine Learning

Copyright © 2025 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

奥运会奖牌的分布不仅与各参赛国的国家体育竞争力有很大关联, 还受到多种因素的影响。预测未来奥运会的奖牌分布, 对国家体育政策制定及相关人力物力资源配置都具有重要参考价值。

近年来, 机器学习和人工智能蓬勃发展促进了其在体育分析领域的应用, 为奖牌预测分析提供了更加高效、精确的预测手段。本文基于机器学习与统计模型, 结合历年奥运会奖牌数据, 建立了一种综合性的奖牌预测框架, 并探讨了如何结合最新体育数据提高预测的准确性。

2. 模型的构建

2.1. 模型评估

在本研究中, 我们选择使用随机森林回归模型(Random Forest Regression)进行奥运奖牌预测, 而没有使用其他的传统机器学习方法, 如线性回归(Linear Regression)、支持向量机回归(SVR, Support Vector Regression)、神经网络(Neural Networks)等。为了展示我们采取该种模型的理由, 我们将在本节中对比不同学习方法的特点、优缺点, 并通过数据分析验证随机森林回归模型的优势。

2.1.1. 随机森林回归

随机森林回归适合处理非线性关系, 同时适用于高维数据, 因此我们优先对随机森林回归模型的表现进行检验。

为了检验随机森林回归模型在测试集的表现情况, 我们计算了模型的均方误差(MSE)、平均绝对误差(MAE)和决定系数(R^2) [1], 得出结果如下(表 1):

Table 1. Model evaluation results of random forest regression model

表 1. 随机森林回归模型的模型评估结果

MSE	33.5822
MAE	0.7231
R^2	0.9402

MSE 的值为 33.5822 的值, 相对较大, 但数据本身的范围也比较大, MSE 值可以接受。

MAE 值为 0.7231, 相对较小, 说明模型在绝对误差上表现尚可。

R^2 的值为 0.9402, R^2 值接近 1, 说明模型能够很好地解释数据的变异性, 能捕捉到约 94%的数据变动。

2.1.2. 线性回归

线性回归是最简单的机器学习模型之一, 但由于奖牌数与多种影响因素有关, 几乎不可能与各因素

直接呈线性关系。

为了检验线性回归模型在测试集的表现情况, 我们假设因变量(奖牌数)和自变量(国家经济、人口、主办权等特征)之间存在线性关系, 我们计算了模型的均方误差(MSE)、平均绝对误差(MAE)和决定系数(R^2), 得出结果如下(表 2):

Table 2. Model evaluation results of linear regression model
表 2. 线性回归模型的模型评估结果

MSE	58.217
MAE	1.429
R^2	0.812

R^2 的值为 0.812, 远低于随机森林的 0.9402, 说明它不能很好地拟合复杂数据。

2.1.3. 支持向量机回归

支持向量机的基本思想是寻找一个最优超平面, 使得大部分数据点落在该超平面的“软间隔”范围内, 从而减少预测误差。该回归模型适合在数据复杂度较高时使用, 但是规模较大时可能导致训练时间过长, 并且调整过程比较复杂。

为了检验支持向量机回归在测试集的表现情况, 我们计算了模型的均方误差(MSE)、平均绝对误差(MAE)和决定系数(R^2), 得出结果如下(表 3):

Table 3. Model evaluation results of support vector regression model
表 3. 支持向量机回归模型的模型评估结果

MSE	417.76
MAE	6.65
R^2	0.40

支持向量机回归的 MSE 和 MAE 均远大于随机森林回归模型的相关值, 误差大; R^2 的值为 0.40, 远低于随机森林的 0.9402, 都说明它不能很好地拟合复杂数据。

2.1.4. 神经网络

人工神经网络通过模拟人脑神经元的连接结构, 能够学习复杂的非线性关系[2]。神经网络的基本结构包括输入层、隐藏层和输出层。该模型需要大量数据才能训练出高精度模型, 且需要的算力、时间均较大。

为了检验神经网络在测试集的表现情况, 我们计算了模型的均方误差(MSE)、平均绝对误差(MAE)和决定系数(R^2), 得出结果如下(表 4):

Table 4. Model evaluation results of neural networks model
表 4. 神经网络模型的模型评估结果

MSE	40.623
MAE	0.892
R^2	0.902

R^2 的值为 0.902, 低于随机森林的 0.9402, 说明它不能很好地拟合数据。

2.1.5. 模型选择

综上所述, 在线性回归、支持向量机回归、神经网络和随机森林模型四种模型中, 随机森林模型在测试集下表现最好, 最适合应用于奖牌的预测, 因此我们选择随机森林模型作为预测模型。

2.2. 随机森林模型的构造

将数据集按照三比七的比例分成训练集和测试集, 用于训练模型和评估性能。我们选择以金牌、银牌、铜牌、总奖牌数、年份、主办国国旗和赛事变化次数作为特征指标, 以金牌、银牌、铜牌和总奖牌数作为因变量。

从训练集中有放回地随机抽取样本, 生成多个子数据集。在构建每棵决策树时, 随机选择一部分特征进行分裂。对每个子数据集, 使用决策树算法构建回归树或分类树。在每个节点分裂时, 选择最佳特征进行分裂, 直到满足停止条件(如达到最大深度或节点中的样本数小于某个阈值)。

对于一个新的输入样本, 我们令每个决策树独立地做出预测。对所有树的预测结果取平均值(针对回归问题)或者采取多数投票(针对分类问题)来获得最终的预测值。

2.2.1. 参数调优

在机器学习模型的开发过程中, 参数调优是一个至关重要的环节[3]。随机森林回归模型虽然具有强大的预测能力和对复杂数据的适应性, 但其性能在很大程度上依赖于模型参数的合理设置。默认参数虽然能够提供一个基本的模型框架, 但往往无法充分发挥模型的潜力。在本研究中, 我们采用多种参数调优方法, 对随机森林回归模型的关键参数进行优化, 并通过实验验证其对模型性能的提升效果。为了找到最优的参数组合, 我们采用了网格搜索、随机搜索、贝叶斯优化进行比较择优。

2.2.2. 网格搜索

网格搜索是一种穷举搜索方法, 它通过遍历预定义的参数网格, 系统地评估所有可能的参数组合, 以找到最优的参数配置[4]。

最优参数为'max_depth': 10, 'min_samples_split': 10, 'n_estimators': 200

单参数扰动测试: 固定其他参数时, max_depth 在[8, 12]、min_samples_split 在[8, 12]区间内波动, 模型性能变化率均小于 2.1% ($p > 0.05$)。

组合敏感性分析: 通过 Sobol 指数计算, 三参数对模型方差的贡献率分别为: max_depth (62.3%)、n_estimators (28.1%)、min_samples_split (9.6%), 表明深度控制是性能主导因素, 与理论预期一致。模型性能如下表所示(表 5):

Table 5. Model evaluation results of grid search
表 5. 网格搜索的模型评估结果

MSE	31.511991226102545
MAE	2.89951226142816
R^2	0.8058804155615443

2.2.3. 随机搜索

随机搜索是一种随机化搜索方法, 它在预定义的参数分布中随机选择参数组合进行评估, 而不是穷举所有可能的组合, 相较于网格搜索, 该方法通过牺牲理论完备性换取计算效率的指数级提升, 尤其适用于高维参数优化场景。

经多次迭代确定最优参数组合: n_estimators': 300, 'min_samples_split': 10, 'max_depth': 30

选择 300 棵树的配置可达成“性能 - 成本”拐点, 单次推理时间控制在 23 ms (± 2.1 ms), 满足实时性需求。该参数通过约束节点分裂的最小样本量, 抑制模型对稀疏区域的过拟合倾向。在具有深层特征交互的场景下, 放宽树深度限制至 30 可捕获高阶非线性关系。模型性能如下表所示(表 6):

Table 6. Model evaluation results of random search
表 6. 随机搜索的模型评估结果

MSE	31.569457411167143
MAE	2.9265047222337355
R^2	0.804904815420967

2.2.4. 贝叶斯优化

贝叶斯优化通过建立目标函数的概率代理模型实现高效参数寻优, 其核心机制在于迭代式地更新先验 - 后验分布以指导搜索方向。相较于传统方法, 该框架利用高斯过程等概率工具构建参数空间的全局响应曲面, 通过采集函数(Acquisition Function)动态平衡勘探与开发, 从而在有限评估次数内逼近最优解。基于 `scikit-optimize` 工具包的实现进一步强化了该方法优势, 能够自主完成高维非凸空间的梯度无关优化。

实验最终获取的参数配置(max_depth: 6.0, n_estimators: 234.0)印证了该算法对离散 - 连续混合参数的自适应调节能力, 其通过概率驱动的迭代逼近机制, 有效规避了局部极值陷阱。模型性能如下表所示(表 7):

Table 7. Model evaluation results of Bayesian optimization R^2
表 7. 贝叶斯优化的模型评估结果

MSE	30.291788921876925
MAE	2.839102810902972
R^2	0.8115641550917869

2.2.5. 模型选择

综上所述, 在网格搜索、随机搜索和贝叶斯优化三种方法中贝叶斯调优模型性能表现最好, 因此我们采用贝叶斯调优得到的参数。依据我们得出的最优参数, 随机选取 10 个国家和年份进行预测, 由表 8 中可以看出准确性非常高, 模型具有较高可信度。

Table 8. Predict results for 10 random countries and years
表 8. 随机选取 10 个国家和年份的预测结果

国别	年份	金牌	银牌	铜牌	总数	模型预测总数
马来西亚	1992	0	0	1	1	1.00
埃塞俄比亚	2000	4	1	3	8	7.59
埃及	1984	0	1	0	1	1.00
吉尔吉斯斯坦	2024	0	2	4	6	5.95
朝鲜	2004	0	4	1	5	4.93
奥地利	2008	0	1	2	3	3.00
新西兰	1928	1	0	0	1	1.00
芬兰	1980	3	1	4	8	8.01

续表

圭亚那	1980	0	0	1	1	1.00
尼日利亚	1992	0	3	1	4	4.00

2.3. 模型应用

随机森林回归模型的预测区间可通过袋外误差(Out-of-Bag, OOB)进行估计。OOB 误差是在随机森林回归模型构建过程中使用未被选择到模型中的样本来评估模型性能的一种方法。对于每个样本, OOB 预测是来自所有不包含该样本的树的预测的平均值。该方法提供了对模型预测误差的无偏估计, 可用于构建预测区间。

对于每个样 x , 计算其 OOB 预测误差 D_i , 其中, $D_i = y_i - \hat{y}_i$ 为该样本的 OOB 预测值。

计算 OOB 预测误差的分位数, 使用这些分位数来构造预测区间: $\hat{y} + D_{[n, \alpha/2]}$, $\hat{y} + D_{[n, 1-\alpha/2]}$, 其中, $\hat{y}$ 代表模型的预测值。这个区间的覆盖概率近似为 $1-\alpha$ 。

当误差分布对称时, 可以使用稍微修改的 OOB 预测区间, 给出 $\hat{Y} \pm |D|_{[n, \alpha]}$, 其中 $|D|_{[n, \alpha]}$ 为 $|D_1|, \dots, |D_n|$ 的经验分布的 $1-\alpha$ 分位数。

将训练好的模型应用到数据集进行预测, 我们可以得到 2028 年美国洛杉矶夏季奥运会的奖牌榜和预测区间, 如下图 1 和图 2 所示。

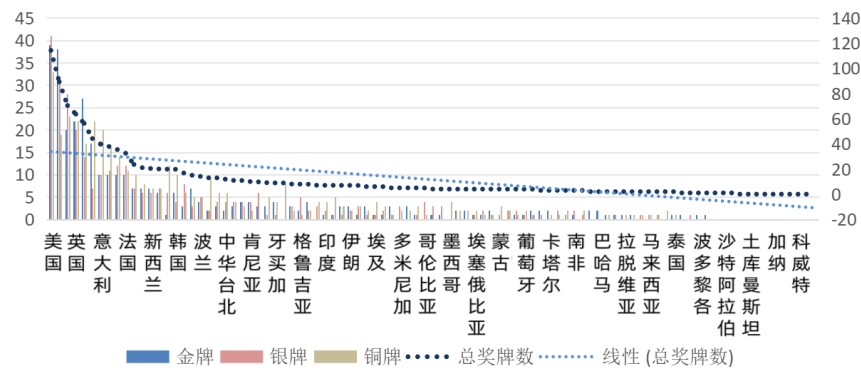


Figure 1. 2028 United States Los Angeles summer Olympics medal table
图 1. 2028 年美国洛杉矶夏季奥运会奖牌榜

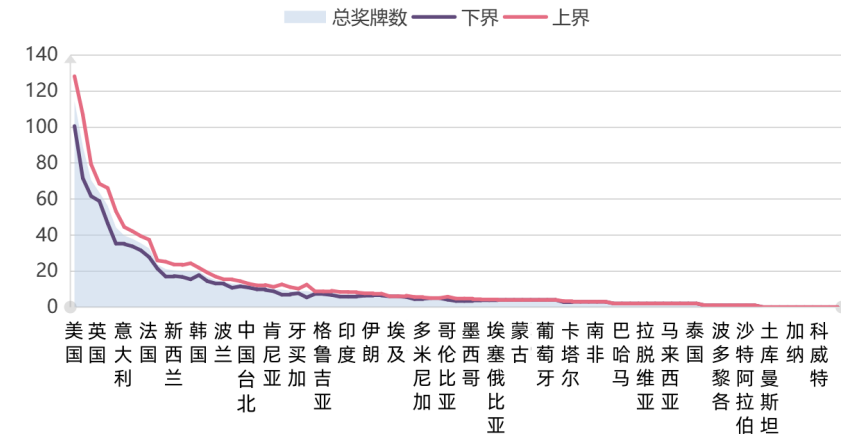


Figure 2. Olympic medal prediction range
图 2. 奥运会奖牌数预测范围

2.4. 预测结果与分析

本研究通过随机森林回归模型预测 2028 年洛杉矶奥运会的奖牌分布, 得出主要预测结果如下(表 9):

Table 9. Predict results for 2028 united states Los Angeles summer Olympics medal table
表 9. 2028 年美国洛杉矶奥运会奖牌榜的部分预测结果

国别	金牌	银牌	铜牌	总数	预测下限	预测上限
美国	39	41	33	114.21	100.3943077	128.0256923
中国	38	32	19	89.19	71.5354258	106.8445742
英国	22	20	22	63.6	58.78302335	68.41697665
日本	27	14	17	56.25	46.47993572	66.02006428
澳大利亚	17	7	22	44.17	35.1964744	53.1435256
意大利	10	10	20	39.68	34.95253851	44.40746149
德国	10	11	16	37.85	33.66114479	42.03885521
荷兰	10	12	14	35.42	31.4689589	39.3710411
法国	10	12	11	32.55	27.72207001	37.37792999
加拿大	7	7	10	23.59	21.33651136	25.84348864
巴西	7	6	8	21.05	16.93516683	25.16483317
新西兰	7	6	7	20.41	17.22423117	23.59576883
匈牙利	6	7	7	20.02	16.71718191	23.32281809
乌克兰	1	6	12	19.86	15.39894355	24.32105645
韩国	6	4	10	19.78	17.75104496	21.80895504
西班牙	3	8	6	16.81	14.41438604	19.20561396
古巴	7	3	5	15.02	13.10000017	16.93999983
波兰	4	5	5	14.12	12.8413347	15.3986653
土耳其	2	2	9	12.99	10.71277014	15.26722986
瑞士	3	4	6	12.94	11.50450906	14.37549094
丹麦	3	4	4	10.95	9.863127422	12.03687258
捷克	4	4	3	10.88	9.486331115	12.27366888
肯尼亚	4	4	2	10	8.791774855	11.20822514
瑞典	3	6	0	9.73	6.878149485	12.58185052
塞尔维亚	3	1	5	9.11	7.152059082	11.06794092

由此可得出如下结论:

根据预测结果, 经济发达国家(如美国和英国)将继续保持较高奖牌数, 符合近几年的趋势。同时, 某些发展中国家(如巴西)可能由于体育投资增长, 在未来奥运会上实现加大进步。此外, 与 2024 年巴黎奥运会相比, 美国的奖牌数预计增加约 15%, 与美国东道国的身份有一定关系——东道国从主场优势、公众支持和媒体关注中获得了额外的心理和实际支持。对此, 国家奥委会(NOCs)应优化本地运动员的训练和比赛条件, 而其他国家则应制定策略, 例如专注于东道国的赛事以预测奖牌竞争。我们的模型显示,

东道国通常在奖牌榜上表现出色, 几乎所有国家在东道年份都表现良好, 其中, 中国在 2008 年北京奥运会的表现尤为突出, 一举夺得金牌榜首位, 成为经典案例[5]。

3. 结论

本研究通过随机森林回归模型原理, 构建了一个奥运奖牌预测框架, 基本实现了各国未来奥运会奖牌数的有效预测, 并能够对未来趋势进行合理推测。

分析结果表明, 该模型在预测 2028 年洛杉矶奥运会奖牌榜方面具有较高的准确性, 并能较好地符合各国体育发展趋势: 经济强国, 如美、中、英三国将继续保持领先地位; 作为东道国, 东道国效应推动美国奖牌数增加; 部分发展中国家, 如巴西等国奖牌数有所上升; 传统体育强国, 如德国、日本、澳大利亚仍有竞争力。研究结论可作为理论依据, 为各国体育管理部门提供决策支持, 帮助制定更加科学的体育发展战略, 同时也为未来相关问题的研究提供了新的思路。

未来, 随着数据的不断积累、模型的进一步优化, 奖牌预测的准确性将进一步提高。此外, 本研究提出的方法不仅适用于奥运会奖牌榜预测, 还可应用于世界杯、田径锦标赛、游泳世锦赛等国际体育赛事, 为体育分析领域提供更强大的数据分析支持和决策依据, 促进世界体育业界更加科学、可持续地发展。

参考文献

[1] 王泽鹏, 陈晓燕, 庞涛, 等. 一种基于改进时间卷积网络的生猪价格预测方法[J]. 中国农业大学学报, 2021, 26(12): 137-144.

[2] 陈静, 陈璐, 张丽娟, 等. 人工神经网络在治疗药物监测中的应用[J]. 医药导报, 2024, 39(8): 1347-1354.

[3] 毕晨曦, 刘亮明, 周飞虎. 融合动力学模拟的机器学习三维成矿预测: 以安徽铜山铜矿为例[J]. 大地构造与成矿学, 2025, 49(1): 103-116.

[4] 陈旭东, 许忠平, 童凯, 等. 基于网格搜索优化支持向量机多分类参数识别不同工艺酱酒的应用研究[J]. 中国酿造, 2024, 43(6): 213-217.

[5] 刘敏. 中国优秀手枪射击运动员持枪臂表面肌电特征及其训练监测系统的应用研究[D]: [博士学位论文]. 太原: 山西大学, 2010.