基于循环神经网络的奥运奖牌榜预测

赵芯禾，赵娜娜 *，陈齐兴，韩怡
北京石油化工学院，北京 102617
DOI:10.61369/ASDS.12183

摘　　要： 在全球体育经济中，奥运奖牌榜备受瞩目。本文以2024年巴黎奥运会为研究对象，综合历史奖牌数据、奥运项目数量和类型等因素，用XGBoost填补缺失值，整理出奖牌数、主办国标识符等特征。随后采用Seq2Seq模型，并用模拟退火算法优化，使判定系数从0.763提升至0.896。最后，通过配对样本t检验发现，执教后球队奖牌数显著变化，证实了“伟大教练效应”，为各国奥委会的资源分配和战略规划提供支持。

关 键 词： 奥运奖牌预测；XGBoost；Seq2Seq；优秀教练；资源优化

Olympic Medal List Prediction Based on Recurrent Neural Network

Zhao Xinhe, Zhao Nana*, Chen Qixing, Han Yi
Beijing Institute of Petrochemical Technology, Beijing 102617

Abstract: In the global sports economy, the Olympic medal table attracts significant attention. This paper takes the 2024 Paris Olympics as the research object, comprehensively considering historical medal data, the number and types of Olympic events, and other factors. Missing values are filled using the XGBoost method, and features such as the number of medals and host country identifiers are organized. Subsequently, the Seq2Seq model is adopted, and the simulated annealing algorithm is used for optimization, increasing the coefficient of determination from 0.763 to 0.896. Finally, through paired sample t-tests, it is found that the number of medals won by teams significantly changes after coaching, confirming the "great coach effect", and providing support for the resource allocation and strategic planning of national Olympic committees.

Keywords: Olympic medal prediction; XGBoost; Seq2Seq; outstanding coach; resource optimization

引言

作为世界上最大的体育赛事之一，奥运会不仅吸引了全球数十亿观众，也是各国体育实力的象征。奖牌榜，尤其是金牌榜，往往被视为衡量一个国家整体体育实力和运动员表现的直接标准。随着2028年洛杉矶奥运会的临近，预测各国在奥运会上的表现已成为当前的热门话题。完美的预测不仅能改善运动员的备战策略，还能促进体育运动的发展，提升国家的整体表现。因此，本文的核心是建立一个合理的模型来预测2028年奥运会奖牌榜的结果，并通过现有的奥运奖牌等数据来评估不同因素对各国成绩的影响。

一、奖牌榜预测相关模型简介

（一）XGBoost方法填补缺失值

XGBoost[1]是一种高效的机器学习算法，能够自动处理数据中的缺失值。其核心在于稀疏感知分裂发现算法（Sparsity-aware Split Finding），该算法通过为每个树节点设置默认方向来处理缺失值，从而避免了对缺失值进行手动填充。此外，XGBoost还支持通过missing参数指定缺失值的处理方式，例如将缺失值设置为数据集中不存在的值（如-999）[2]。这种处理方式在实验中被证明能够有效提升模型的性能。

XGBoost能捕捉变量间的非线性关系，高效合理地填补缺失的数值，为后续分析奠定了基础。

（二）Seq2Seq[3]预测模型

Seq2Seq模型的核心在于其端到端的学习方式，能够处理输入和输出序列长度不一致的问题，同时通过引入注意力机制，进一步提升了模型对长序列的处理能力。经过模拟退火算法[4]优化的Seq2Seq模型，能够有效捕捉奥运会奖牌数的时序变化以及国家间的复杂关系。优化超参数[5]后，该模型在测试集上表现出色，预测精度较高。我们还采用蒙特卡洛dropout方法[6]对预测结果的不确定性进行量化分析。该方法通过在模型中随机丢弃部分神经元，生成多个预测结果，从而评估预测的不确定性。

项目/基金信息：国家级大学生创新创业训练计划（2025J00124），北京市教委科研计划一般项目（KM202410017004），北京石油化工学院致远基金（(2024212）。
通讯作者简介：赵娜娜，北京石油化工学院讲师，博士，专业：概率论与数理统计。

（三）SHAP模型

我们构建了一个基于SHAP模型[7]的奥运会奖牌预测系统。该系统通过分析历史奖牌数据，并结合机器学习算法，对各国在未来奥运会中可能获得的奖牌数量进行预测。利用SHAP值分析，揭示不同特征对奖牌预测结果的贡献度，为奥运会奖牌预测提供了新的视角和方法。

二、模型求解

（一）优化模型参数[8]

在训练Seq2Seq模型时，训练集的R^2值为0.976，但测试集的R^2值仅为0.805，存在过拟合问题。我们通过TPE优化调整模型参数，最终使训练集的R^2值降至0.772，测试集的R^2值升至0.990，有效缓解了过拟合，提升了模型的泛化能力。

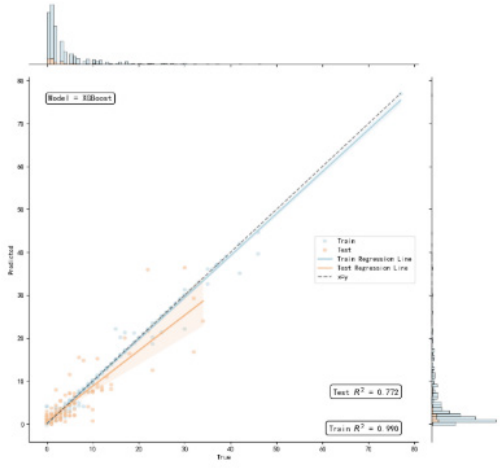


图1 模型训练

（二）数据清洗

为准确预测2028年洛杉矶奥运会奖牌榜情况，我们需要构建新的数据集，基于2024年巴黎奥运会奖牌榜以及其他数据，对于2028年新增5项项目，我们做出以下分析：

✓查找相关资料，并剔除无关数据（如：2028年俄罗斯被禁赛）

✓假设新增加的5各项目，各国新增项目奖牌榜基数全部以2枚打底：

newX[‘Baseball’]=newX[‘Baseball’]+2

newX[‘Softball’]=newX[‘Softball’]+2

newX[‘Cricket’]=newX[‘Cricket’]+2

newX[‘Sixes’]=newX[‘Sixes’]+2

newX[‘Squash’]=newX[‘Squash’]+2

newX[‘Flag football’]=newX[‘Flag football’]+2

而后，为每个奖牌计算置信区间：

	gold	Silver	Bronze	gold_mean	gold_CI_lower	gold_CI_upper	Silver_mean
0	39.752754	44.465527	42.390995	38.785717	37.497152	40.074282	38.727448
1	39.431030	26.888872	23.508928	37.948128	35.978237	39.918018	26.872393
2	19.534157	13.425087	12.662200	15.884511	12.938937	18.830085	12.217851
3	17.879230	19.071848	16.958645	14.894158	12.259883	17.528434	13.842833
4	15.174778	25.482542	20.379179	13.859722	11.555674	16.163770	21.669180
...	...	...	...	...	...	...	...
101	0.130617	0.000000	0.496482	0.000000	0.000000	0.484139	0.000000
102	0.081197	0.225503	1.456575	4.851646	0.000000	15.511469	4.441838
103	0.107504	0.103913	1.245036	0.000000	0.000000	0.701567	0.233750

> 图2 置信区间

三、预测分析

（一）预测结果

2028年预测结果显示，这些国家的成绩将显著增长，ROC和Mixed team的奖牌数量呈上升趋势，反映出这些国家在体育运动中展现出新活力。

	NOC	2024Total	2028Total	Improvement
78	ROC	0.0	17.0	17.0
74	Poland	7.0	13.0	6.0
54	Jordan	1.0	4.0	3.0
62	Mixed team	0.0	3.0	3.0
39	Greece	3.0	6.0	3.0
95	Thailand	4.0	7.0	3.0
61	Mexico	4.0	7.0	3.0
65	Morocco	1.0	3.0	2.0
57	Kosovo	1.0	3.0	2.0

> 图3 结果预测

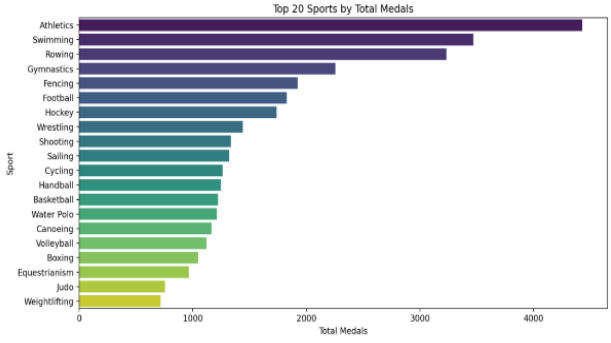
（二）未来奖牌的分布情况

模型的预测值在第t次迭代时可以近似为：

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$$

SHAP模型用于解释XGBoost模型的预测结果，SHAP值的计算公式为：

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{\| S \| !(\| F \| - \| S \| - 1)!}{\| F \| !} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) x_S)]$$



> 图4 奖牌榜前20运动

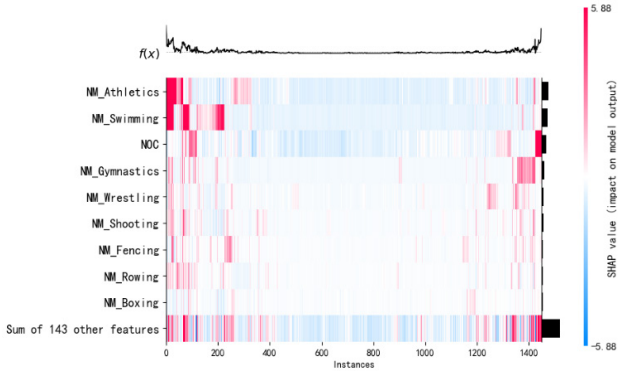
根据SHAP模型的输出，我们可以有效预测未来奖牌分布情况。比如，对一些体育强国（美国、英国、俄罗斯等），可以推断出在游泳等项目上分布人数居多，而对于一些相对来说较小的国家，一些新兴项目（比如2028年洛杉矶奥运会新增的5项项

目）对其来说更具有影响力。

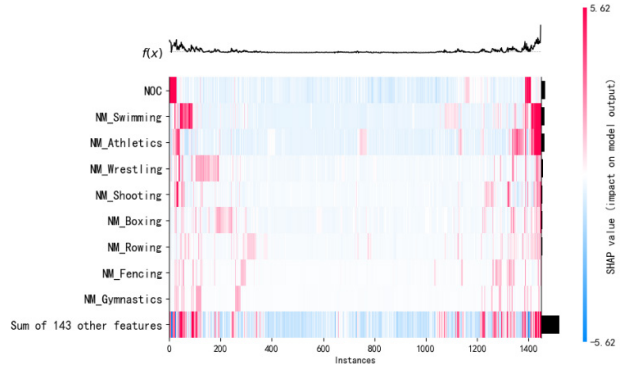
（三）数据分析

表1 模型评估结果

	MSE	RMSE	MAE	MAPE	R^2
训练	0	0.001	0.001	32.674	1
测试	0.014	0.119	0.074	64.485	0.704

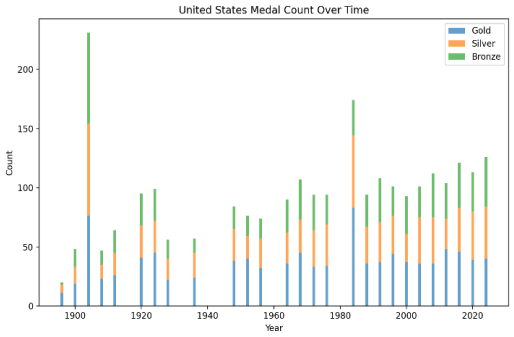


> 图5 .1预测结果



> 图5 .2预测结果

这两张图是特征重要性可视化图表，纵轴展示了如田径、滑雪等运动项目相关特征及其他143个特征的集合，横轴为特征重要性数值。颜色编码表示重要性正负，红正蓝负，颜色深浅对应重要性程度。顶部曲线与模型预测值相关。两图特征类别相同，但颜色分布和深浅有别，体现不同情境下各特征对模型结果影响的变化。



> 图6 美国不同年份奖牌分析

在最近的年份，美国的奖牌数量保持在较高水平，显示出美国在国际体育赛事中的持续竞争力。经过不断的深入分析，有助于理解美国在国际体育赛事中的地位和影响力，以及体育发展的历史脉络。

四、“伟大教练”效应模型建立与求解

（一）量化分析 [9]

根据中国女子乒乓球队在李隼教练指导下的比赛成绩数据，筛选出所需信息。为了确定李隼的执教效应，需要构建一个是否执教的变量，1997年之前，未执教。

分析的步骤：

1.根据定类变量（X）对定量字段（Y）进行分组，进行方差齐性检验，查看P值是否小于0.05，倘若P值大于0.05，使用方差分析 [10]，查看P值是否呈显著性 (P<0.05)。

2.独立样本T检验 [11] 呈现显著性，可借助效应量化分析对差异性进行量化分析。

执教前均值 ($\bar{x}_1$)：

$$\bar{x}_1 = \frac{5+4+5}{3} = \frac{14}{3} \approx 4.67$$

执教后均值 ($\bar{x}_2$)：

$$\bar{x}_2 = \frac{3+15+14+14+16+17}{6} = \frac{79}{6} \approx 13.17$$

执教前方差 (S^2_1)：

$$s_1^2 = \frac{(5-4.67)^2+(4-4.67)^2+(5-4.67)^2}{3-1} = \frac{0.11+0.44+0.11}{2} = 0.33$$

执教后方差 (S^2_2)：

$$s_2^2 = 25.063$$

执教前标准差 (S_1)：

$$s_1 = \sqrt{0.33} \approx 0.57$$

执教后标准差 (S_2)：

$$s_2 = \sqrt{25.06} \approx 5.01$$

使用公式：

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

其中 n1=3，n2=6，

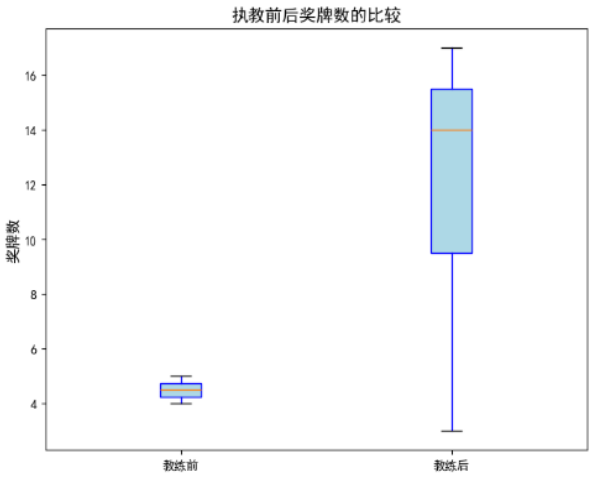
$$t = \frac{4.67-13.17}{\sqrt{\frac{0.33}{3}+\frac{25.06}{6}}} = \frac{-8.50}{\sqrt{0.11+4.18}} = \frac{-8.50}{\sqrt{4.29}} \approx \frac{-8.50}{2.07} \approx -4.11$$

$$df == \frac{(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2})^2}{\frac{(\frac{s_1^2}{n_1})^2}{n_1-1}+\frac{(\frac{s_2^2}{n_2})^2}{n_2-1}} = \frac{(0.11+4.18)^2}{\frac{0.11^2}{2}+\frac{4.18^2}{5}} = \frac{18.07}{0.0061+3.45} = \frac{18.07}{3.4561} \approx 5.23$$

通常，自由度取整数，因此取 df=5。

使用t分布表或统计软件，查找自由度为5，t统计量为 −4.11 的p值。给定的p值为0.0118，表明在5%的显著性水平下，执教前后的奖牌得分有显著差异。因此可以拒绝原假设，李隼执教中国女子乒乓球后对奖牌数量产生了影响，从而支持“伟大教练”效应的存在。

（二）对奖牌数的影响



> 图7 执教前后奖牌数影响

回归分析显示，李隼执教后，中国女子乒乓球队的奖牌得分显著增加，具体提升了7.5分。这表明教练的变动对奖牌数量有重要影响，验证了“伟大教练”效应在中国女子乒乓球队中的显著作用。基于此预测和分析，各国在选聘顶级教练时，可以提前进行战略布局，提升自身在体育赛事中的竞争力。

（三）教练投资建议

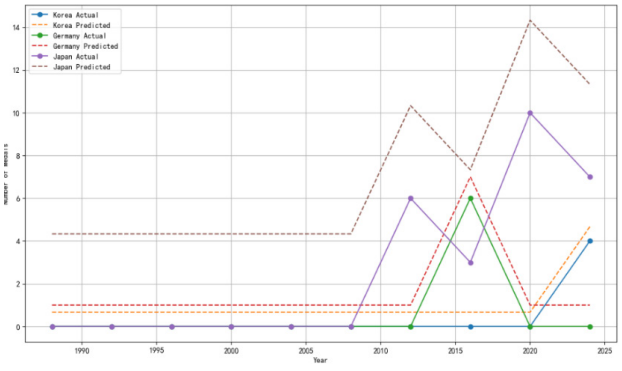
根据此问题，我们构建一个量化模型，评估各国在教练投资上的效率和效果，以确定哪些国家在教练投资方面具有更高的优先级。结合了经济学、管理学和体育科学的理论，提出了一个综合评估模型。

教练投资与体育绩效之间的关系可以通过以下公式表示：

$$P = f(I, R, E)$$

下面我们构建一个综合评估模型：

$$CP = w_1 \cdot EI + w_2$$



> 图8 教练投资与奖牌数对比

综上所述，通过对教练回报投资的预测，可以为各国运动事业添砖加瓦，帮助其在乒乓球事业上作出更大的贡献。

五、结束语

本研究通过整合 XGBoost 缺失值填补、模拟退火算法优化的 Seq2Seq 模型以及 SHAP 可解释性分析，构建了一套高效的奥运奖牌榜预测框架，为 2028 年洛杉矶奥运会的奖牌分布提供了科学依据。模型在测试集上 R^2 值提升至 0.990，验证了其捕捉时序动态与复杂国家关系的能力。研究首次将 SHAP 值引入奥运预测领域，揭示了项目参与度、东道国效应等核心驱动因素，并量化了“伟大教练效应”的显著性（如李隼执教使中国乒乓球队奖牌得分提升 7.5 分），为各国优化教练资源配置提供了数据支持。此外，针对小国提出“专项突破”策略，指出新增项目可成为其奖牌增长的关键契机。

本研究的创新在于融合时序建模 [12]、可解释性分析与多学科方法，构建了兼具预测精度与策略指导价值的综合模型。然而，模型的静态数据假设及对国际政治、运动员个体差异等动态因素的忽略，可能限制其实际应用的普适性。未来研究可引入实时数据流与社会网络分析，增强模型动态响应能力；同时探索模型在冬奥会、亚运会等赛事的迁移应用，并进一步结合运动员训练质量、伤病风险等微观变量，提升预测的全面性与精细化水平，为全球体育事业均衡发展提供更强大的理论工具。

参考文献

[1] 蒋晋豫，王海燕，张馨之，等 . 基于 XGBoost 算法的森林生物量多源遥感反演 [J]. 西北林学院学报，2025,40(2):198-206,219.

[2]Chen,T., Guestrin, C. XGBoost: A Scalable Tree Boosting System[C]. Acm Sigkdd International Conference on Knowledge Discovery & Data Mining, 2016.785-794.

[3] 常圣南 . 基于 Seq2Seq 模型的多标签文本分类研究 [D]. 辽宁：大连海事大学，2022.

[4] 许洁林，张灏龙，李静，等 . 基于模拟退火算法的装备组合优化算法与仿真 [J]. 计算机仿真，2025,42(1):13-18.

[5] 冯纬枢 . 优化超参数的 LSTM 网络频谱预测 [J]. 福建质量管理，2020(9):295-297.

[6] 门超杰，赵静，张楠 . 基于图增强和注意力机制的时间序列不确定性预测 [J]. 华东师范大学学报（自然科学版），2025(1):82-96.

[7] 陈明良，马志远，张东辉，等 . 基于 SHAP 可解释性的焊缝缺陷类型超声识别 XGBoost 模型 [J]. 无损检测，2024,46(6):36-42.

[8] 姜瑶，颜泽文，黎良辉，等 . 灌区用水优化模型参数全局敏感性分析与不确定性优化 [J]. 农业机械学报，2023,54(7):372-380.

[9] 王海军，王涛，俞慈君 . 基于递归量化分析的 CFRP 超声检测缺陷识别方法 [J]. 浙江大学学报（工学版），2024,58(8):1604-1617.

[10] 李润，钟林，郭蓓蓓，等 . 方差分析和灰色关联分析在玉米品种评价中的应用 [J]. 现代农业科技，2025(5):38-41.

[11] 高艺祥，杨民红，李兰会 . 独立样本 t 检验的 Excel 和 SPSS 分析 [J]. 畜牧与饲料科学，2018,39(10):79-82.

[12] 栗志磊，李俊，施智平，等 . 用于视频行为识别的高效二维时序建模网络 [J]. 计算机工程与应用，2023,59(3):127-134.

