选定的问题	2025 MCM/ICM 摘要表	团队控制编号
C		2507817

奥运奖牌揭秘：

成就趋势的数学探索

摘要

在体育竞技的巅峰盛会——奥运会上，奖牌争夺不仅体现运动员的体能实力，更折射出国家整体的体育实力。随着竞技体育的发展，奖牌榜排名历经变迁。本文从多维度剖析影响奥运表现的关键因素，并展望奖牌竞争格局的未来趋势。

任务1中，我们基于预处理数据构建了**网格搜索随机森林（GSRF）**预测模型。通过**交叉**验证，获得预测金牌数与总奖牌数的关联系数分别为0.710和0.784，验证了模型有效性。 基于该模型，我们预测了2028年奥运会的金牌及总奖牌榜排名，结论显示中国与美国有望继续领跑两大榜单。同时**采用t分布**计算预测区间（见图9），并识别出最可能实现跃升与退步的国家（见图10）。

对于任务2，我们建立**了一个逻辑回归模型**，将尚未在2028年奥运会上夺得奖牌的国家分为两类：夺牌国与非夺牌国。这将夺牌概率转化为**二元分类问题**。 我们选取了合适的特征变量，并采用**最大似然估计法计算**每个特征向量的权重。通过应用sigmoid函数**和L1正则化**，我们从尚未夺牌的国家中识别出更可能夺牌的国家，如表11所示。

任务3基于任务1结果，计算了各国在各项目中的夺牌数量，并通过计算该项目奖牌数占总奖牌数的比例来量化其贡献度。我们注意到部分国家（如韩国在柔道项目）仅通过特定项目获得奖牌，这些项目对其至关重要。 总体而言，各国倾向于优先发展兼具优势与潜力的运动项目，以最大化夺牌机会。

在任务4中，我们创建了**"伟大教练"模型**。通过量化奖牌权重并计算各国得分，我们发现罗马尼亚和美国女子体操项目存在"伟大教练"现象，其得分与两国得分的**斯皮尔曼相关系数**达0.874。 在运用**Lasso回归**构建模型后，我们选取三个国家并确定具体项目进行应用。引入"伟大教练"概念后，各国2028年预测得分如表15所示，其中罗马尼亚得分从3分跃升至36.65分，增幅显著。

最后，我们为国际奥委会提供了关于主办国效应和优秀运动员影响力的原创性见解。此外，我们进行了**敏感性分析**，证明该模型在扰动后仍具有稳定性和鲁棒性。

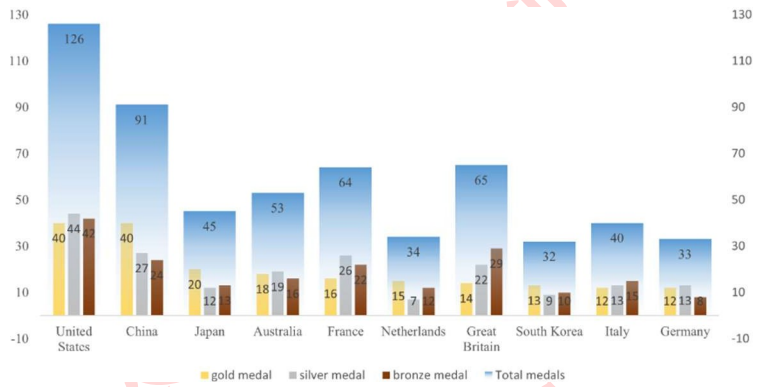
关键词：奥运会；**GSRF**；**逻辑回归**；**Lasso回归**

目录

1 引言

1.1 背景

当观众关注夏季奥运会时，他们既欣赏运动员的表现，也关注奖牌榜的排名。自1896年以来，美国等体育强国始终位居榜首，彰显其雄厚实力。赢得众多金牌的国家获得全球认可，而首次摘得奥运奖牌的国家则备受瞩目，标志着历史性的里程碑。以下是2024年夏季奥运会奖牌榜前十名国家：



1.2 澄清与 重述

本题要求基于赛事数据与运动员数据构建奖牌榜预测模型。分析将结合历届夏季奥运会奖牌榜数据、主办国信息、赛事项目数量及参赛人数等要素，解决以下问题：

任务1：基于模型预测2028年夏季奥运会奖牌排名，给出包含各项统计数据的预测区间，指出最可能进步与退步的国家。

- ✓ 影响奖牌数量的特征值经过统计后整合至数据集，用于模型训练与测试。模型验证完成后，基于2024年数据预测各国在2028年奥运会上的金牌数与总奖牌数，计算预测区间并进行可视化呈现。

任务2：针对尚未获得奖牌的国家，预测将在下一届奥运会首次夺牌的国家数量及其概率

- ✓ 本质上属于分类问题，需将2028年未获奖国家的预测结果划分为"获奖"与"未获奖"两类。判断过程需计算该国进入获奖类别的概率。

任务3 分析历届奥运会参赛项目数量与类型对各国夺牌数量的影响。确定每个国家最重要的项目并说明理由，同时考察主办国项目选择对结果的影响。

- ✓ 我们梳理各国奖牌累计数量并分析其关联性。同时引入可量化项目对国家重要性的概念。运用数据进行判断并阐述依据。

任务4 预测"伟大教练"效应对于奖牌数量的贡献。选取

三个国家，分析应引入"伟大教练"的项目，并评估其影响。

✓ 首先识别受影响的国家及项目，随后计算包含伟大教练影响变量与最终预测值之间的回归方程。接着筛选出具备预测表现变化潜力的国家及项目。

任务5 基于对奥运奖牌榜的原创性洞见，分析这些洞见能为国家奥委会提供的信息价值。

✓ 找出先前任务中对金牌数量和奖牌总数产生影响的特征值，并对其进行研究

1.3 我们的 工作

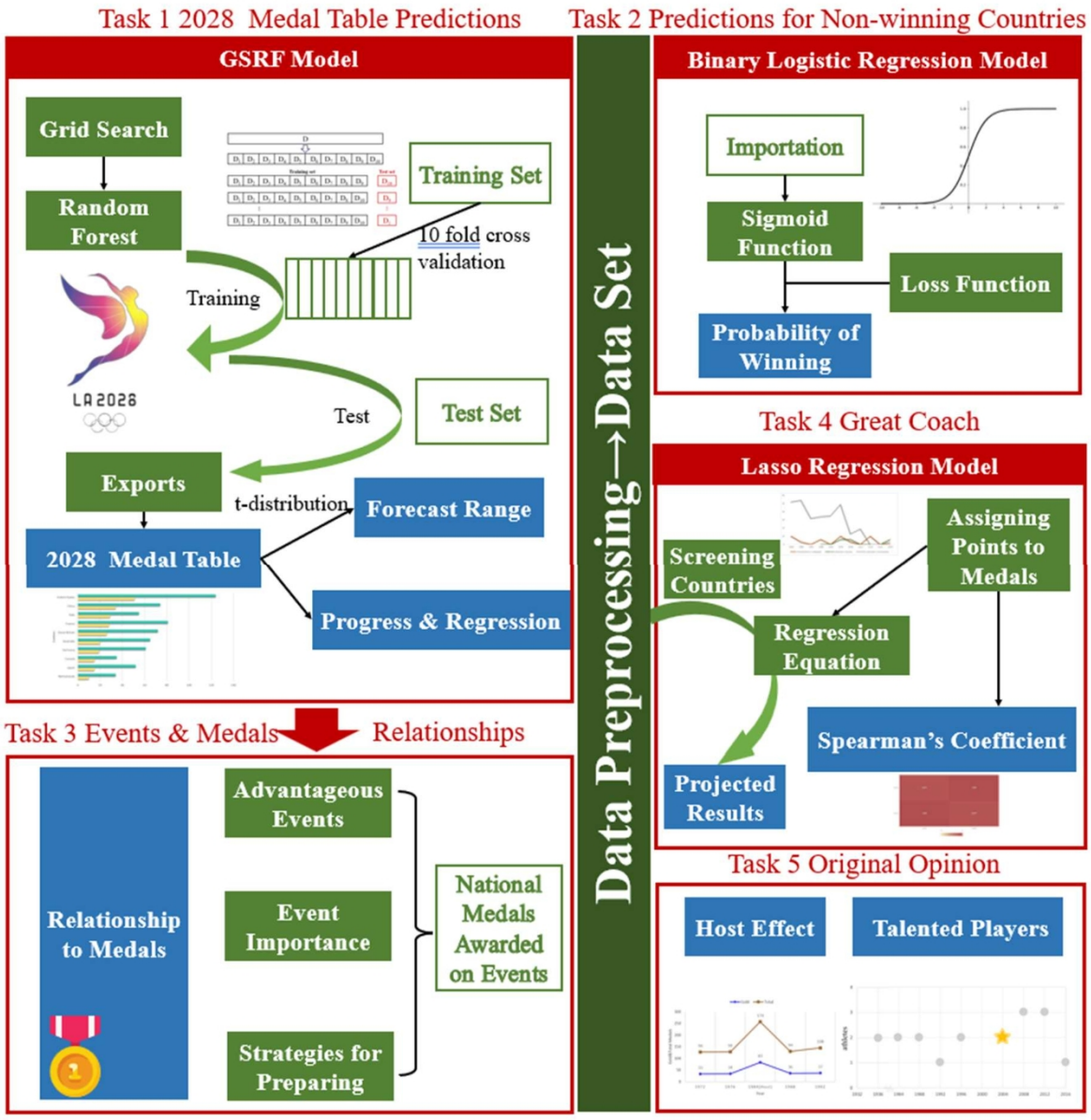


图2 我们的工作

2 基本 假设

➢ **假设1：假设参加2028年奥运会的运动员人数、参赛国家数量和举办的赛事数量与2024年相同。**

合理性：数据审查显示，近期奥运会的参赛国家数量和举办项目数量变化不大。为简化模型计算，假设该数值保持不变。

➢ **假设2：这些变量彼此独立，仅影响结果。**

论证：尽管这些变量存在交互作用，但在模型中纳入它们会增加复杂性和不可预测性。为进行深入分析，我们更侧重于可衡量且稳定的因素。

➢ **假设3：排除投资优秀教练的成本**

理由：旨在聚焦教练质量对运动员表现、团队绩效或整体体育体系发展的直接影响。

3 符号

符号	定义
Aij	第i个国家在第j届奥运会中的运动员总数
G_{ij}	第i个国家在第j届奥运会之前获得的金牌总数
T_{ij}	第j届奥运会之前第i个国家获得的奖牌总数
Eij	第 j 届奥运会赛事总数
Y_g	
Y_t	预测金牌数预测奖牌总数第k个国家的运动员人
a_k	
e_k	数
p_k	第k个国家参与的项目数量
	第k个国家历史参赛次数

4 数据 预处理

■ **异常值处理：**同一项目下各国可能存在不同队伍，这些队伍名称包含代表队伍顺序的标记符，且国家名称后存在乱码。我们已完成对这些异常值的清理。

■ **数据标准化：**在预测过程中，为确保各特征值在数据中具有相对均衡的权重，需采用数据标准化处理，其中需计算每个特征的均值($\bar{x}$)与标准差(SD)

$$x = \frac{(x - \bar{x})}{SD} \tag{0}$$

- **国家与地区名称转换**：建模过程中使用了两个文件："summerOly_medal_counts.csv"和"summerOly_athletes.csv"。由于文件中国家名称格式不一，在数据整合前采用ISO映射表将所有国家及地区全称转换为标准化代码。 已解散或遭禁赛的国家数据（如苏联与俄罗斯）均被排除。
- **运动员与赛事统计**：按国家及年份分类统计运动员与赛事数据并计算总值。独立中立运动员（AIN）的赛事成绩不代表任何国家，在预测中也应予以剔除。

5 任务1：基于GSRF模型预测金牌及总奖牌数

5.1 GSRF预测模型

5.1.1 ARIMA模型初始预测

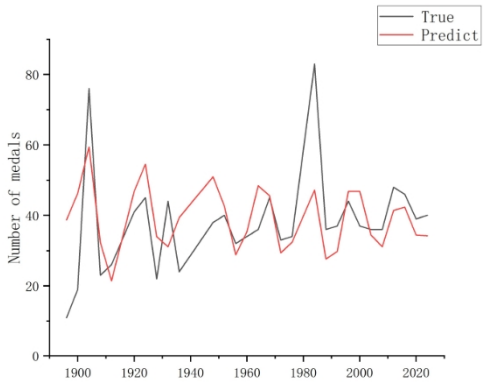
ARIMA模型因其灵活性和强大的预测能力，在时间序列分析中被广泛应用，能够适应多种数据特征并作出精准预测。该模型融合自回归（AR）与移动平均（MA）方法，并引入差分（I）操作以增强数据平滑效果。通过分析历史数据的自相关性，模型假设未来趋势将延续历史模式，从而预测未来数据点。ARIMA(p,d,q)模型可表示为方程(2)：

$$X_t = c + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_p X_{t-p} + \vartheta_1 s_{t-1} + \vartheta_2 s_{t-2} + \cdots + \vartheta_q s_{t-q} + s_t \tag{1}$$

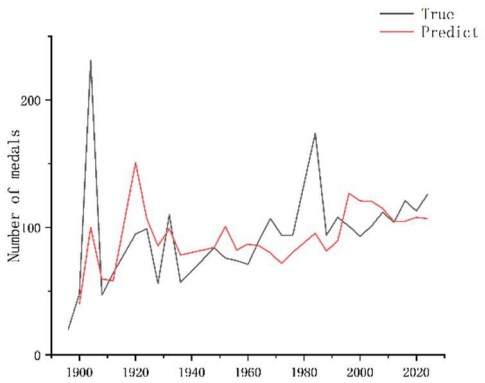
其中：

- X_t 表示我们正在考虑的时间序列数据；
- c 为常数项；
- $\varphi_1, \varphi_2, \ldots, \varphi_p$ 是AR模型的参数，用于描述当前值与过去p个时间点的值之间的关系；
- $\vartheta_1, \vartheta_2, \ldots, \vartheta_q$ 是MA模型的参数，用于描述当前值与过去q个时间点的误差之间的关系；
- s_t是时间点 t 的误差项。

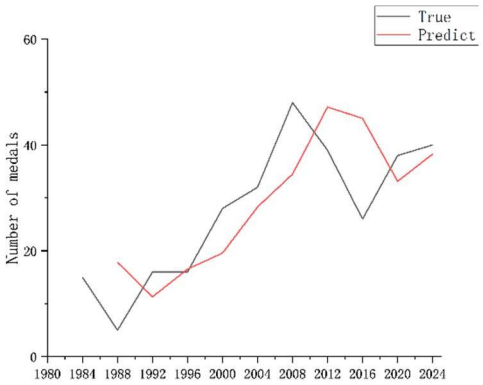
我们对中美两国金牌数和奖牌总数进行的初始ARIMA预测，得出如下预测图表：



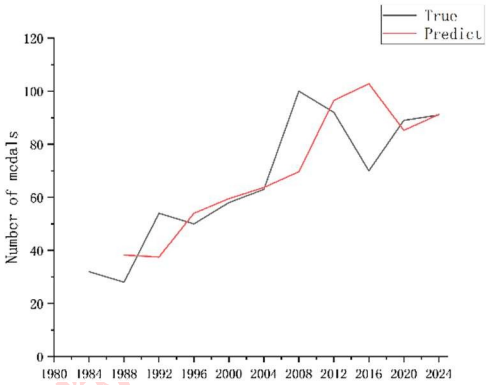
a) 美国金牌



b) 美国总预测



c)中国黄金预测



d) 中国总预测图3 美国和中国奖牌数预测图

对应上述预测值的ARIMA(p,d,q)模型及其相关系数R(2)*系数R^2* 如下表所示：

表1 预测模型相关数据

	ARIMA (p, d, q)	R^2
美国黄金预测	ARIMA (2, 0, 2)	0.218
美国总预测	ARIMA (2, 1, 1)	0.257
中国黄金预测	ARIMA (1, 1, 0)	0.430
中国总预测	ARIMA (1, 1, 0)	0.494

预测所得的相关系数较小且远离1，证明采用ARIMA模型预测奖牌数量并不理想，因此我们转而使用GSRF摆动预测模型进行预测。

5.1.2 GSRF波动预测建模

首先，我们需要建立一个模型来预测各国奖牌数的波动，并找出与这种波动最相关的因素。为此，我们使用了GSRF波动预测模型。

为了预测该国在下一届奥运会将获得的奖牌数，我们首先从往届参赛运动员的获奖情况出发，分别使用网络搜索随机森林算法预测未来将获得的金牌数和奖牌总数。

接下来，我们将描述所选指标数据并阐释算法流程。

1) 选定指标

基于往届数据集，我们选取了四个特征输入指标：

- **该国在本届奥运会参赛运动员总数：**运动员越多意味着更多参赛机会和更广泛的选拔范围，从而提高夺牌概率。我们用A_{ij}表示第j届奥运会中第i国的运动员总数：
- **该国截至本届奥运会的金牌数与总奖牌数：**历届奖牌榜可初步反映国家实力及其发展趋势。我们用G_{ij}表示国家i在第j届奥运会前的金牌总数，用$T_{(ij)}$表示其总奖牌数。
- **当前奥运会项目总数：**项目总量决定奖牌分配基础，项目越多奖牌越多。国家参与项目越多，夺冠机会越大。项目增加将导致奖牌分布更广，更多国家获得夺冠机会。我们用 E_j 表示第 j 届奥运会项目总数
- **是否为东道国：**作为主办国，该国在场馆、设施和后勤方面具有优势。主办国可增加其擅长的项目数量，减少其不擅长的项目数量，并自动获得更多参赛名额，从而提高运动员的参赛机会。若该国是第j届奥运会的主办国，则$令_{Hij}=1$；反之，$令_{Hij}=0$。

2) GSRF算法

GSRF算法通过网格搜索优化随机森林模型。随机森林是基于多棵决策树训练的机器学习算法，每棵树随机选取特征子集，通过投票整合分类结果。为避免过拟合或欠拟合，采用网格搜索优化算法。该算法遍历预定义参数网格，训练并评估每种组合，最终输出最佳参数集与模型性能。

相较于标准随机森林算法，GSRF算法采用最优超参数组合进行模型训练与预测。该改进显著提升模型性能，有效缓解过拟合或欠拟合等问题。下图展示了GSRF算法的工作流程。

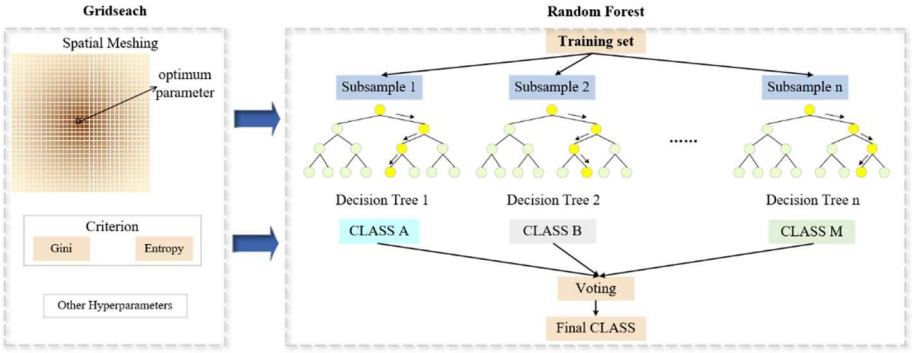


图4 GSRF算法流程图

3) 对金牌及总奖牌数进行预测

根据文献记载，运动员的实力数量、赛事种类数量等一系列数据已被证实会影响下一届奥运会的奖牌数量。因此，我们采用上述四个特征输入指标作为

随机森林模型预测奖牌数量的输入参数：

$$\begin{aligned}(A_{ij,g}, G_{ij}, E_{j,g}, H_{ij,g}) &\xrightarrow{RF} Y_g \\ (A_{ij,t}, T_{ij}, E_{j,t}, H_{ij,t}) &\xrightarrow{RF} Y_t\end{aligned} \tag{2}$$

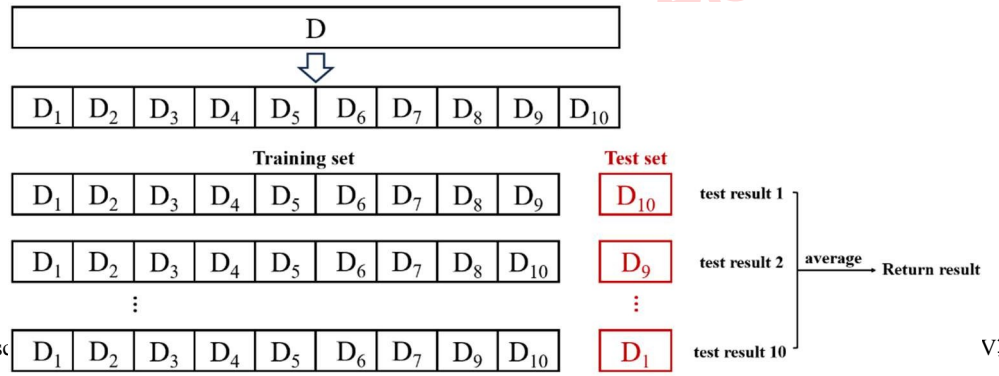
其中。

$A_{ij,g}$, $E_{j,g}$, $H_{ij,g}$ 分别表示金牌预测模型中该国运动员总数、奥运赛事总数以及该国是否为本届奥运会主办国 Y_g 表示预测获得的金牌数量。

$A_{ij,t}$, $E_{j,t}$, $H_{ij,t}$ 分别表示该国运动员在奖牌总数预测模型中的总人数、奥运赛事总数，以及该国是否为本届奥运会主办国。$_{Y(t)}$ 代表预测的金牌总数。

奥运会主办国。Y_t 表示预测的金牌数量。

我们采用2024年前的数据作为训练集，2024年数据作为测试集。采用k折交叉验证法，k值设为10。训练集被均分为10个等大小子集。每次选取其中一个子集作为验证集，其余9个子集作为训练集。 随机森林模型在训练集上进行训练，并在验证集上评估模型性能并记录评估指标。最终计算所有子集的平均评估指标，以此获得模型性能的综合估计值，作为评估模型稳定性和泛化能力的依据。示意图如下所示：

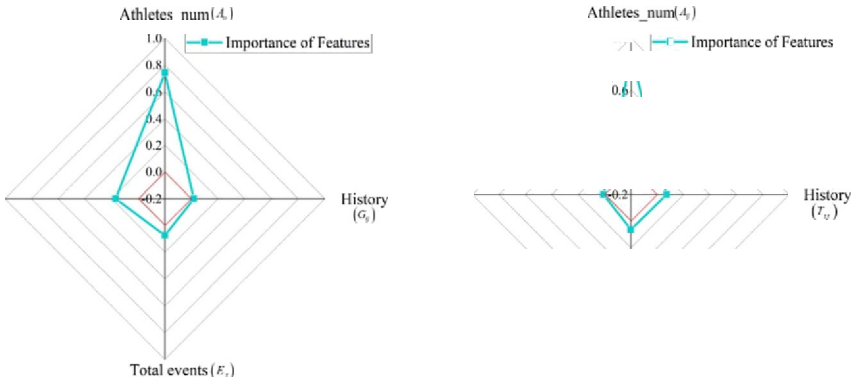


采用sc V算法进行参数搜索。GridSearchCV获得的最优参数组合如下：

表2 各类指标对预测结果的影响

黄金 奖牌				总计 奖牌			
$A_{ij,g}$	G_{ij}	$E_{j,g}$	$H_{ij,g}$	艾伊，特	T_{ij}	$E_{j,t}$	$H_{ij,t}$
74.10%	1.71%	7.40%	16.90%	85.70%	7.20%	6.30%	0.80%

我们使用雷达图可视化每个指标对预测结果的影响，如下所示：



a) 金牌预测 b) 奖牌总数预测图6 特征重要性雷达图

图表显示，运动员人数对金牌预测和总奖牌预测的影响最大。主办国身份对金牌预测有显著影响，对总奖牌预测影响较小。奥运会项目总数对奖牌数量有一定影响，但不显著。

部分测试集的预测值与实际值对比结果如下：

(a) 金牌预测 b) 奖牌总数预测图7 2024年部分国家实际值与预测值对比图

5.1.3 模型预测有效性与性能评估

我们使用前80%的数据训练随机森林模型，并用后20%的数据进行测试。基于预测值与真实值，该模型的决定系数(R^2)、平均绝对误差(MAE)及均方根误差(RMSE)计算如下：

相关系数R^2：将模型预测值与仅用均值预测的值进行比较，R^2值越接近1，模型准确性越高。

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{3}$$

平均绝对误差 MAE：绝对误差的平均值，可反映预测值误差的实际情况。数值越小，模型精度越高。

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right| \tag{4}$$

均方根误差 RMSE：均方误差的平方根，RMSE 越小，模型越准确

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

计算得出的模型评估结果如下：表3 金牌预测模型评估结果

	R^2	MAE	均方根误差
训练集 Training set	0.903	0.626	1.829
交叉验证集 Cross validation set	0.710	0.994	3.145
测试集 Test set	0.795	0.914	2.574

表4 奖牌总数预测模型的评估结果

	R^2	MAE	RMSE
训练集 Training set	0.933	1.947	4.411
交叉验证集 Cross-validation set	0.784	2.744	7.162
测试集 Test set	0.736	2.430	7.145

两个预测模型的R^2值均接近1，且平均绝对误差(MAE)与均方根误差(RMSE)均较小。这表明我们构建的预测模型准确性更高，模型性能更优。

5.2 预测2028年金牌与奖牌 表格

5.2.1 预测2028年奥运会金牌榜与奖牌榜

基于前文建立的GSRF预测模型，本文对2028年美国洛杉矶夏季奥运会的奖牌榜进行预测。下表预测了2028年洛杉矶奥运会各国家/地区获得金牌数排名前十的国家/地区、奖牌总数排名前十的国家/地区及其对应数量。

表5 2028年奥运会金牌预测前十名

国家/地区 Country	美国 USA	中国 CHN	意大利 ITA	法国 FRA	英国 GBR	澳大利亚 AUS	德国 GER	日本 JPN	加拿大 CAN
金牌数 Gold medals	51	34	29	28	26	26	20	19	15

表6 2028年奥运会奖牌预测前十名

Country 国家	USA 美国	FRA 法国	CHN 中国	GBR 英国	AUS 澳大利亚	GER 德国	ITA 意大利	JPN 日本	CAN 加拿大	NED 荷兰	BRA 巴西
Total medals 总奖牌数	124	81	74	72	65	61	55	52	35	34	34

我们使用绘图软件将结果可视化如下所示：

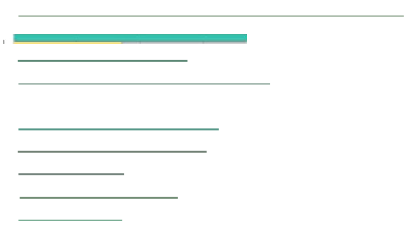


图8 2028年洛杉矶奥运会金牌及奖牌总数预测榜

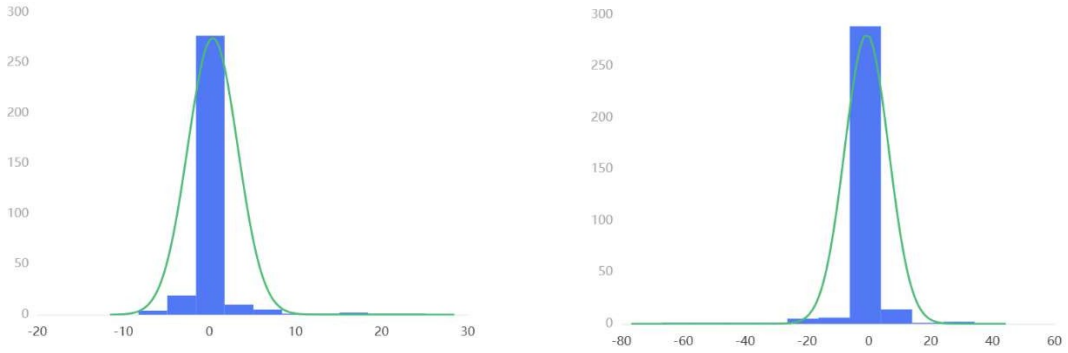
美国在金牌榜和奖牌榜上均位居榜首，法国在奖牌榜上排名第二但金牌数量不多，中国在金牌榜上排名第二，奖牌榜上位列第三。预测2028年金牌榜时，加拿大上榜而韩国落榜，其余国家排名均有变动。

上述预测仅为最可能的情景，接下来我们将求解95%置信水平下的预测区间：

5.2.2 计算预测区间

首先，我们判断两个预测模型的残差是否服从正态分布：

我们绘制了残差直方图并进行拟合，同时对数据执行了夏皮罗-威尔克检验，结果如下所示：



a) 金牌预测 b) 总奖牌数预测图9 残差直方图

表7 拟合数据与SW检验结果

	标准差	偏度	峰度	夏皮罗-威尔克检验
金牌	2.908	4.723	35.429	0.438
总奖牌数	7.123	-4.644	40.523	0.434

根据上述图表可发现，两个预测模型均存在显著峰值和偏度，且数据的夏皮罗-威尔克检验显著性P<0.05，表明两个预测模型均不符合正态分布。

因此我们采用t分布计算预测区间

$$\text{预测区间} = y \pm t_{\alpha/2,n-p} \times SE(y) \tag{6}$$

其中。

y 代表*预测值*，$t_{\alpha 2,n-p}$ 是 t 分布的分位数，n 是/

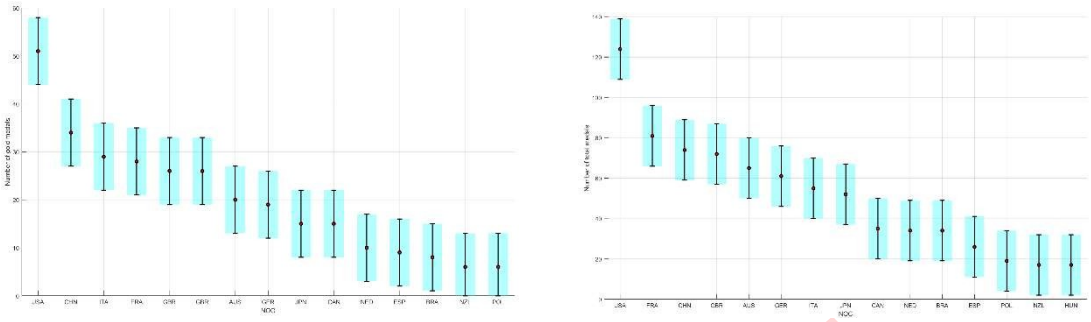
样本数，p为模型参数数，$SE(y)$为预测标准误差。

误差。相关结果计算如下：表8 预测模型T分布相关数据

	$t\alpha\ 2,n-p$ /	$SE(y)$
金牌预测	2.404	2.908
奖牌总数预测	1.993	7.123

那么，在95%置信水平下，金牌数量的预测区间为

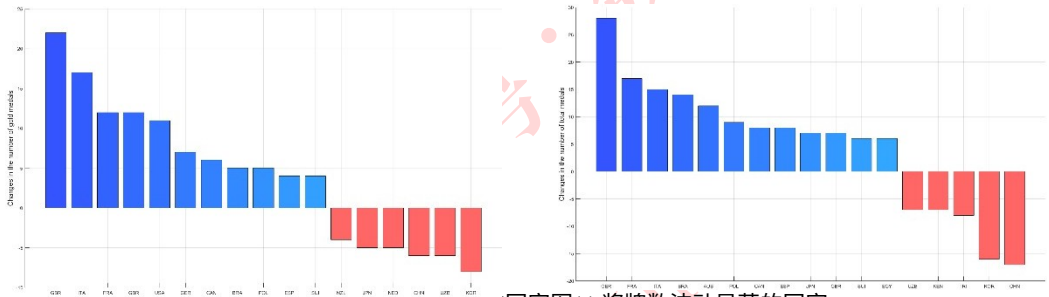
$y \pm 7$而奖牌总数预测区间为 $y \pm 15$



(a) 金牌预测 b) 总奖牌数预测图10 预测模型的误差条形图

5.2.3 预测国家表现的进步与退步

该问题要求我们预测最可能进步或退步的国家，我们计算了2024年金牌数、奖牌总数预测值及其与实际获奖情况的差异，以确定价值变化幅度，并绘制了如下直方图：



a) 金牌数波动超过2枚的国家b) 总奖牌数波动超过5枚的国家图11 奖牌数波动显著的国家

图表显示英国金牌增幅最大，德国奖牌总数增幅最大，两者均取得显著进步。韩国金牌降幅最大，中国奖牌总数降幅最大。除美国外，金牌显著增长的国家奖牌总数也呈现不同程度进步。美国金牌增长更多，但奖牌总数增长幅度较小。

6 任务2：未获奖国家 预测

该问题要求我们预测从未获得过奖牌的国家是否会在2028年夏季奥运会上赢得首枚奖牌，并预测其获胜概率。本质上这是一个二元分类问题，预测结果分为"将获胜"与"不会获胜"两类。因此我们构建二元逻辑回归模型。

6.1 双分类逻辑回归 建模

我们将无奖牌国家记录为：**若**能在2028年夏季奥运会夺牌则归类**为1类**，否则**归类为0类**。

6.1.1 Sigmoid函数（数学）

逻辑回归是一种广义线性回归，它将非线性函数与线性函数结合进行映射。我们使用Sigmoid函数将线性函数的输出值映射为0到1之间的概率值。Sigmoid函数的公式如下：

$$\sigma(z) = \frac{1}{1+e^{\wedge -z}} \tag{7}$$

其中：

$$z = x_1 a_k + x_2 e_k + x_3 p_k + b \tag{8}$$

■ a_k , e_k , p_k 为输入特征值

a_k：第k个国家在2028年奥运会的运动员人数。更多运动员意味着更多人能参加更多项目，或更多人能参加同一项目，这大大增加了获胜概率。

e_k : 第k个国家在2028年奥运会中参与的项目数量。相较于专注于单一项目，参与更多项目能在一定程度上提升夺冠概率。

p_k：第k个国家的历史参赛记录：国家参与国际赛事越多，积累的经验越丰富，这能显著提升其获胜概率。

■ x_1 ,x_2 ,□,x_n 代表权重系数，其值通过梯度下降法计算。

■ b 为偏置项

当sigmoid函数输出值接近1时，该国属于类别1的概率较高，即获奖概率较高；当输出值接近0时，该国属于类别0的概率较高，即未获奖概率较高。

6.1.2 引入损失函数

设 $p_i = \sigma(z) = \frac{1}{1+e^{-z}}$, 我们得到 $P(z=0|x) = 1-P(z=1|x) = 1-p_i$, 则

$$P(y_i|x,x) = (p)^{y_i}(1-p)^{1-y_i} \tag{9}$$

为进一步优化模型，我们引入交叉熵损失函数（对数损失）来衡量模型预测概率与真实类别之间的差距，对数损失定义如下：

$$L = \sum_{i=1}^{m} z_i \ln(p_i) + (1-z_i)\ln(1-p_i) \tag{10}$$

其中 z_i为真实类别；p_i为模型的预测概率；m 为样本量。

6.2 双分类逻辑回归 结果

模型参数结果通过MATLAB计算得出，如下表所示：

表9 二元逻辑回归模型参数结果

	回归系数	标准误差	Waid	P
b	2.425	0.098	607.134	<0.01
a_k	-0.012	0.004	10.62	<0.01
e_k	-0.068	0.008	76.036	<0.01
p_k	-0.013	0.010	1.665	0.197

如表所示，p_k 对应P>0.05，表明p_k 对预测结果无影响。我们规定因变量取值为1表示非获奖国能在2028年夏季奥运会夺冠，取值为0则表示无法获奖。

影响预测结果。我们规定因变量取值为1表示非获奖国能在2028年夏季奥运会夺冠，取值为0表示无法获奖。由此推导出非获奖国能否赢得2028年奥运会的回归方程：

$$z = -0.012a_k - 0.068e_k + 2.425 \tag{11}$$

6.2.1 模型评估与修正

为评估模型分类效能，首先考察其敏感度与特异度：

a) 敏感度（TPR）：实际阳性样本中被预测为阳性样本的比例。

b) 特异性（FPR）：实际为阴性样本的结果中被预测为阳性样本的比例。

c) AUC：即ROC曲线下的面积，用于衡量二元分类模型的整体性能。

该模型的ROC曲线如图所示：

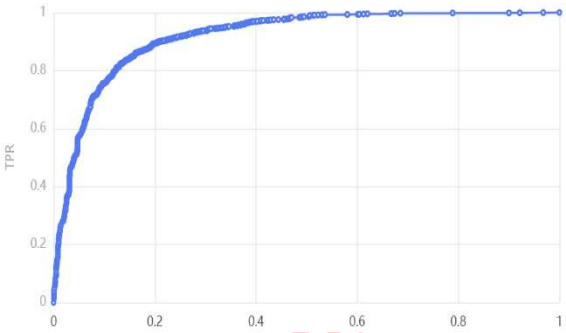


图12 ROC曲线

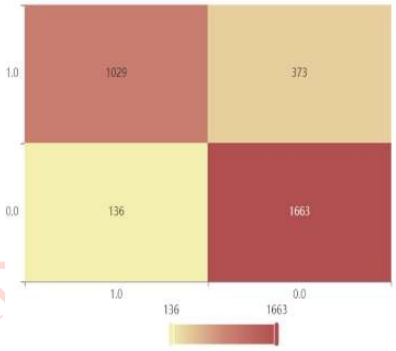


图13 混合矩阵热力学图

ROC曲线结合了敏感度（真阳性率）和特异度（假阳性率），可同时衡量二者的关系。理想情况下，真阳性率应接近1，假阳性率应接近0，AUC值应接近1。根据ROC曲线计算得出AUC=0.918，可见该模型的敏感度较好。

为进一步通过量化指标评估逻辑回归的分类效果，绘制了上图所示的混淆矩阵热力图。据此计算模型分类评估指标如下表所示

表10 分类评估指标

准确率	召回率	精确度	F1
0.841	0.841	0.848	0.839

d) 准确率：阳性样本占总样本的比例，准确率越高越好。

e) 召回率：预测阳性样本中实际阳性样本的比例，召回率越高越好。

f) 精确度：预测阳性样本中实际阳性的比例，精确度越高越好。

g) F1：精确率与召 回率的综合平均值，二者相互影响。

上表显示准确率、召回率和精确率均较高，F1值也较大，表明该分类模型在保证较高准确率的同时召回率也较高，分类效果更优。

6.2.2 分类结果如下

需求解因变量为1的概率p：

$$p = \frac{1}{1+e^{-z}}$$
$$z = -0.012a_k - 0.068e_k + 2.425 \quad (12)$$

计算结果后，我们仅选取获奖概率大于0.2的国家，如下表所示

表11 各国夺牌概率

LBN	GUM	PLE	ANG	ESA
0.290	0.230	0.220	0.206	0.201

最终可视化结果如下所示：

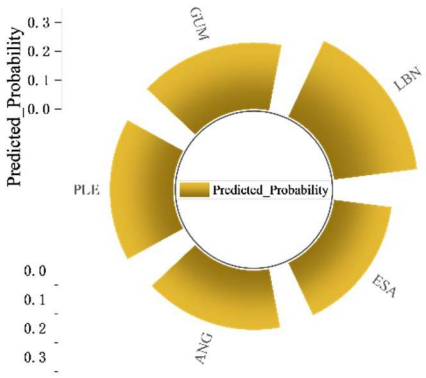


图14 各国夺牌概率分布

我们发现这些未获奖的国家主要集中在中东、亚洲东部和南部地区、撒哈拉以南非洲以及部分岛屿地区。其中部分国家因政治问题饱受战争困扰，另一些则因自然条件限制无法获得专业训练场地。 还有些国家经济落后，无力发展体育事业。然而，它们在2028年夏季奥运会上仍有机会赢得奖牌，我们期待它们的精彩表现。

7 任务3 体育与 奖牌的关系

该问题要求我们基于已建立的GSRF预测模型，统计各国在每项赛事中的奖牌数量。求解体育项目与各国奖牌数之间的关系，识别对各国最重要的体育项目并探究原因。通常主办国会在其擅长的体育项目中增加赛事，需探讨此举对其他国家奖牌数的影响。

7.1 各项目与金牌及总奖牌数的关系

基于先前绘制的雷达图（图6），我们发现参赛项目对奖牌数量具有显著影响。我们筛选出在奥运会中表现优异的国家，统计其获得的奖牌数量，并选取其夺牌较多的项目如下所示：

表12 部分国家及其擅长项目

国家奥委会	项目	奖牌数	国家奥委会	运动项目	奖牌
美国	游泳	1206	中国	游泳	120
	田径	1190		跳水	119
	赛艇	388		体操	109
	篮球	341		乒乓球	94
日本	体操	166	澳大利亚	游泳	505
	游泳	127		曲棍球	188
	柔道	102		赛艇	162
	排球	101		田径	100
KOR	手球	96	GBR	田径	393
	射箭	90		赛艇	319

从表中可以看出，这些国家将在其擅长的项目中赢得更多奖牌

7.2 国家最重要的运动项目

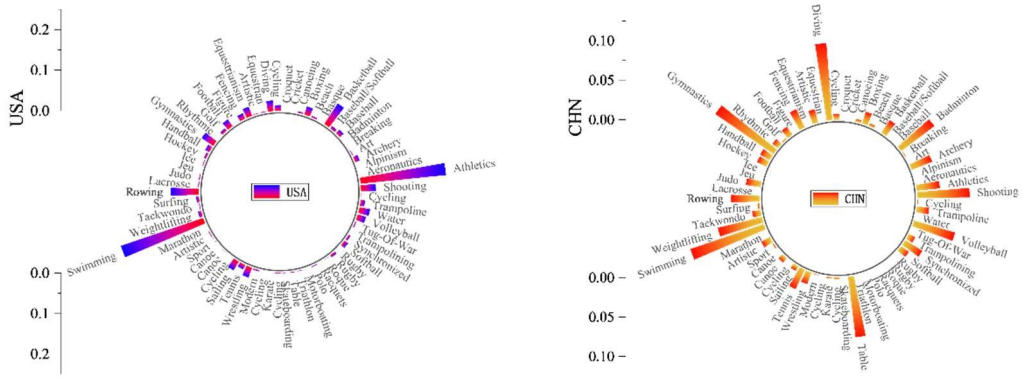
为评估项目对国家的重要性，我们引入项目重要性指标$I_{(j)}$：

$$I_j=\frac{m_k}{M_k} \tag{13}$$

其中：

mk代表该国在此项目中获得的奖牌数；Mk代表该国获得的奖牌总数

对于综合体育实力强的国家，其在多项运动中均能斩获奖牌，项目重要性并非单一数值。以美国和中国为例，计算各项运动对其重要性，绘制如下径向直方图：



a) 美国 b) 中国

图15 项目重要性径向直方图

从图中可以清晰看出，游泳对中美两国而言都是最重要的运动项目。根据相关文献可知，中国在跳水项目上的表现优于游泳[2]，但游泳的重要性仍高于跳水，这是因为游泳项目数量众多，极大提升了夺牌概率。这也恰当地反映了项目数量对结果的影响。

对于体育弱势国家而言，其可能仅在一项运动中摘得奖牌——当该项目重要性为1而其余项目均为0时，该项目便成为该国最重要的运动。本统计办公室统计的此类国家及其对应运动项目如下图所示。

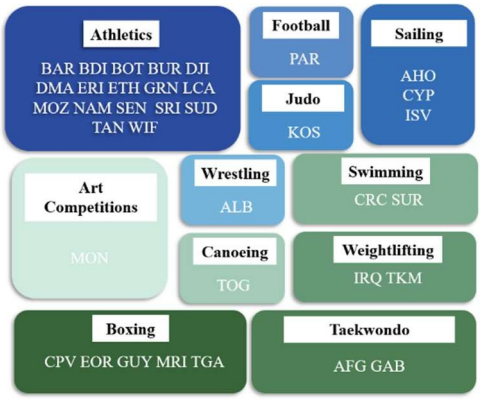


图16 仅在一项运动中获得奖牌的国家

上图显示田径运动对许多国家至关重要，这是因为田径项目赛事众多，为体育实力较弱的国家提供了更多夺奖机会。这些国家多为非洲国家，它们在田径项目上具有优势。

7.3 国家选定项目对 成绩的影响

奥运备战是持续波动与调整的漫长过程，各国将根据历届奥运会及各类国际赛事的综合成绩调整备战方案。

国家将投入更多时间和精力发展这项运动，为那些已具备优势的人创造条件，同时吸引更多人参与其中，从而为推动该项运动发展提供更优质的人才资源，以实现更卓越的

奥运会成绩。

对于潜力巨大但过去成绩平平的项目，该国将增加资金投入，并采取适合该项目的策略，从而提高奥运成绩。

对于从未获奖的项目，通常该国并不普遍开展这类项目，或者该国人民普遍不擅长此类项目。该国不会投入大量时间和资金在这些项目上。

8 任务4 顶级教练的影响力

该问题要求我们分析历届赛事中聘请"优秀教练"引发的成绩波动，评估此因素对奖牌数量的影响，并识别最需要聘请优秀教练的国家。需构建回归模型以量化该效应。

8.1 Lasso回归模型

8.1.1 模型准备

为量化各国历届奥运会表现，我们采用以下规则为不同奖牌赋值：

表13 计分表

金牌	银牌	铜牌	无奖牌
10	6	3	0

我们查阅了历史上著名且杰出的教练资料，发现女子体操教练贝拉·卡罗利曾培养出众多奥运冠军[3]。他在1976年和1980年奥运会期间执教罗马尼亚队，并在1984年至2016年奥运会期间执教美国队[4]。 我们统计了1952至2024年历届奥运会罗马尼亚与美国女子体操项目的得分情况，如下所示：

图1

我们发现该教练执教后，两国女子体操得分均显著提升。其离开罗马尼亚后，该国得分虽略有增长，但总体呈缓慢下滑趋势；而其执教美国期间，尽管得分波动较大，但整体呈现上升态势。随后采用斯皮尔曼相关系数进行分析，计算公式如下：

$$p = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{14}$$

"优秀教练"与奖牌数量的相关系数计算得出为0.874。

通过上述分析可清晰认识到，伟大教练在提升体育竞技表现方面发挥着不可忽视的作用，因此我们建立Lasso回归模型进行定量分析

8.1.2 Lasso回归建模

Lasso回归是替代最小二乘法的压缩估计方法，其主要目标是在模型简洁性与准确性之间寻求平衡。

1) 线性回归模型

线性回归模型构建如下：

$$y = \beta_0 + \beta_1 I + \beta_2 V + \beta_3 S + \beta_4 A + \beta_5 H + \beta_6 C + s \tag{15}$$

- I、V、S、H、C为回归指标

  I：项目对该国的重要性。根据任务3的分析，该国将在重要性较高的项目中采用质量策略提升表现，这些策略包括聘请优秀教练。

  V：该国近两届奥运会平均得分。该指标取最近两届赛事总分之和除以2，反映国家在最近两届奥运会的整体表现。

  S：该项目理论得分公式。$s=I\,V$ 表示项目重要性与平均得分的函数关系

  A：该国在本届奥运会中的参赛人数。参赛运动员越多，夺奖概率越大

  H：该项目该国的历史平均得分。该指标反映国家项目的整体水平，并能体现其表现波动。

  C：教练影响系数。为量化教练对成绩的影响，我们设定：执教首年C=1，每增加一个奥运周期，成绩线性提升0.1；当教练离任时，成绩呈指数级下降，t代表已开始及终止执教的奥运周期数：

$$C = \begin{cases} 1 + 0.1t & \text{开始执教 停止执教} \\ e^{-t} & \end{cases} \tag{16}$$

- y为回归分析所得得分
- β_i 是待估*系数*，ε 是误差项

2) L1正则化

在回归分析中，我们引入L1正则化项，使回归模型不会优先缩减特定参数，从而保留若干关键特征并

促进模型稀疏性：

$$\text{Lasso} = \text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j| \quad (17)$$

其中用于控制正则化程度：

R 当 — 0 时，所有特征均被纳入考虑且正则化效应消失，相当于普通线性回归

当T趋于无穷大时，所有特征均不被考虑，部分系数被精确归零，这将逐步剔除更多特征，从而有效控制模型复杂度。

由此产生的Lasso回归将系数收缩至0，其偏差随增加而增大，方差随减小而增大。

3) 目标函数

Lasso回归的目标是寻找使预测值与真实值之间平方差之和最小化的系数：

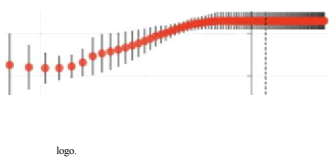
$$RSS = \sum_{i=1}^{6} \left(\hat{y}_i - y_i \right)^2 \quad (18)$$

目标函数——min(RSS+L)

8.2 Lasso回归 结果

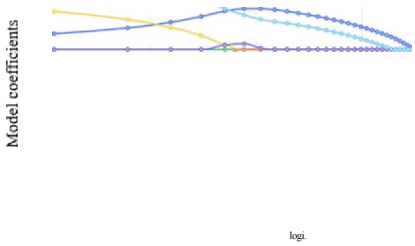
8.2.1 回归方程如下所示

我们采用交叉验证来选择参数。所选参数应使模型均方误差（MSE）最小化，下图可视化了参数选择的过程：



logo.

图18 交叉验证示意图



logi.

图19 变量与模型系数关系图图18显示：随着变量的对数值变化，各变量的模型系数亦随之变化

变量的系数随之变化，当系数趋近于零时可视为应从模型中剔除。

回归方程的系数使用Python计算得出，如下表所示：

表14 回归方程系数

Variables 变量名称	Intercept 截距	I	V	S	A	H	C
Standardized coefficient 标准化系数	2.721	0	-0.046	0.906	0.027	0.144	26.944
R^2				0.708			

当模型中标准化变量的系数为0时，意味着该变量对模型的解释性为零。

该变量被排除在模型之外。可见项目对国家的重要性系数(I)为0，证明该变量对模型结果无影响，故被剔除。回归方程如下：

$$J \quad 2.721 — 0.046U + 0.9065 + 0.027A \quad 0.144N + 26.944C \tag{19}$$

8.2.2 投资"伟大教练"的建议

经过深入分析与数据可视化处理，我们筛选出以下潜力巨大的运动项目：中国女排、巴西女足和罗马尼亚女子体操。这些项目曾有辉煌战绩，如今虽处于中等水平，但具有重要战略意义：

图20 国家在该项目得分的波动情况

基于Lasso回归模型，我们预测若这三个国家在其各自体育项目中投资"伟大教练"计划，则其2028年奥运会该项目得分如下表所示：

表15 2024年成绩与2028年成绩对比

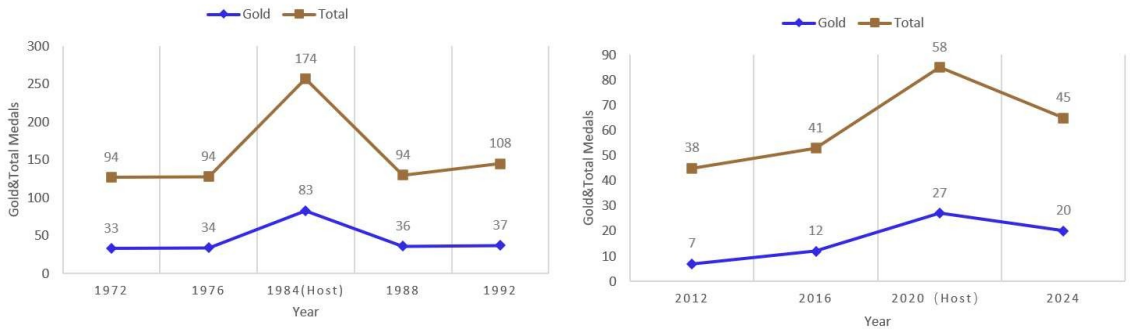
	CHN, Women's Volleyball 中国，女子排球	BRA, Women's Soccer 巴西，女子足球	ROU, Women's Gymnastics 罗马尼亚，女子体操
2024	0	6	3
2028	5.33	8.86	36.65

我们发现，在投入卓越指导后，各国在这些潜在项目的执行成效上均取得显著提升，由此印证了卓越指导对这些国家及项目的重要性。

9 任务S 原文 观点

9.1 主持人 效果

在任务一中，我们发现作为主办国在获得更多奖牌方面具有优势，因此我们选取了美国和日本两个国家，比较了它们作为主办国与非主办国时金牌数量和奖牌总数的变化：



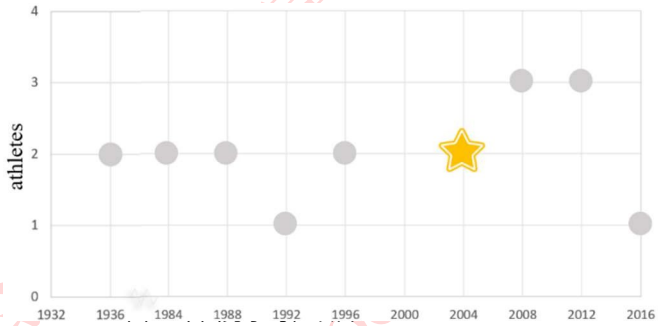
a) 美国 b) 日本

图21 奖牌数量变化

研究发现，当美国和日本担任东道主时，金牌数与总奖牌数会增加1.4至1.9倍。国际奥委会可依据此现象鼓励各国申办奥运会。

9.2 的优秀运动员

在竞技体育中，既有刻苦训练取得佳绩的运动员，也有天赋异禀的选手。 如果一个国家长期参与某项赛事却仅获一枚奖牌，该项目冠军可称为"天赋型选手"。例如男子跨栏项目，中国自1984年参赛以来，仅刘翔在2004年夺冠，其余选手均未摘牌。



因此，国际奥委会可以鼓励各国挖掘此类项目的人才，加大投资力度发掘更多天赋选手，这些选手反过来又能推动更多努力训练的选手登上奥运领奖台。

10 分析

基于首个问题建立的GSRF模型，运动员数量（athletes_num）的敏感性分析显示：

通过改变运动员数量模型A_{ij} ，总事件*期望值*E_{ij} 在

模型参数，并观察模型中对应的变动。

10.1 敏感度定义

将本次预测结果定义为y ，即该结果与先前

$\hat{y}$ ，平均敏感度定义为 s_p 如下

：

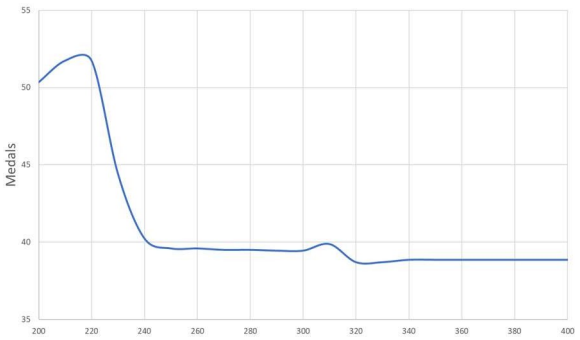
$$s_p = \hat{y} / y \tag{20}$$

该指标反映预测奖牌数量变化的幅度

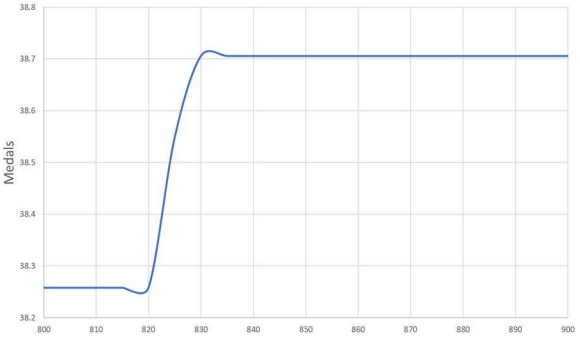
指标变化后的影响。

10.2 运动员人数、金牌及总奖牌数对 预测结果的影响

在任务1中，我们发现多个特征值影响金牌数与奖牌总数的预测。在控制其他特征值不变的前提下，分别改变运动员数量A_{ij} 、*项目总数*E_{ij} ，观察对应的预测奖牌数的变化，并计算平均敏感度：运动员人数：0.058329%；赛事总数：-1.2113%。



a) 总事件



b) 运动员人数图23 敏感

性分析

从图中可以看出，模型在受到扰动后基本不会出现太大波动，平均敏感度也很低。这表明我们的模型具有很强的稳定性和鲁棒性。

11 模型的评估

(1) 优势

- ✓ GSRF在传统随机模型基础上采用超参数最优组合进行训练与预测，显著提升模型性能，有效缓解过拟合或欠拟合等问题
- ✓ GSRF模型的均方误差(MAE)与均方根误差(RMSE)均更小，证明模型预测结果更精准可靠
- ✓ 二元逻辑回归模型具有更高的F1指标，在保证更高准确率和更优分类效果的同时，也确保了较高的召回率。
- ✓ 通过改变模型的特征值，发现模型在受到扰动后基本不会出现过大波动，表明模型具有强稳定性和鲁棒性。

(2) 改进建议

- ➢ 假设2028年奥运会的运动员人数、参赛国家和运动项目数量将与2024年相同，但实际情况可能发生变化，这将影响模型的预测结果。
- ➢ 变量间可能存在交互作用，例如运动员人数与项目数量之间可能存在正相关关系，这会导致模型预测产生偏差。

12 参考文献

[1] Christoph S、L. S S、Dominik S 等. 奥运奖牌分布预测——一种社会经济机器学习模型[J]. 技术预测与社会变革,2022,175

[2] 谭开峰, 张庆义. 《巴黎奥运会世界竞技格局与中国竞技水平分析》[J]. 《辽宁体育科技》,2025,47(01):56-62.DOI:10.13940/j.cnki.lntykj.2025.01.023.

[3] *教练团队——贝拉与玛莎·卡罗利*，美国体操协会，https://usagym.org/halloffame/inductee/coaching-team-bela-martha-karolyi/

[4] [2024 年 11 月 16 日] *美国著名体操教练贝拉·卡罗利逝世*，ESPN 新闻服务，https://www.espn.com/olympics/gymnastics/story/_/id/42433795/bela-karolyi-famed-us-gymnastics-coach-dies-82