

奥运预测，把握奖牌趋势

摘要

在2024年巴黎夏季奥运会上，公众对个人项目和各国奖牌排名的兴趣激增。每届奥运会前，都会创建一个“虚拟奖牌榜”来预测不同国家的表现，但这些预测通常不仅基于历史奖牌数据，还受到参赛运动员的显著影响。本文利用了诸如一个国家参与的项目数量和类型、项目成绩、主办国以及专项赛事等信息来构建模型。

对于问题一和二，我们使用K-均值聚类法将获奖国家分为三类。我们通过皮尔逊相关系数计算了每个国家与不同项目的关联度，以识别该国的标志性赛事。基于这三类国家，我们应用了LSTM奖牌数训练和预测模型。利用该模型，我们预测了2028年洛杉矶夏季奥运会各国家的奖牌数量。该模型还发现，奥运会的项目数量和类型对一个国家的奖牌数有显著影响。

对于第三期，我们使用随机森林算法构建了一个模型，以预测哪些国家将首次获得奖牌。目的是识别影响一个国家奖牌数量变化的关键因素。我们选择了三个特征变量：“参赛人数”、“参赛次数”和“比赛项目数”，并将“是否会获奖牌”作为目标变量。我们使用随机森林回归模型训练了该模型，并通过GridSearchCV选择了最佳超参数，从而得到了最优的模型配置。随后，我们计算了每个国家获得奖牌的概率。模型的性能通过混淆矩阵和ROC曲线进行了评估，结果显示预测准确性很高。

在第四期中，我们利用图论和网络流理论定量评估了“优秀教练”对一个国家奖牌数的影响。我们建立了教练与国家之间的有向网络图，将教练对奖牌的影响转化为节点和边的流动关系。使用以下公式计算体重： $W = 3 \times G + 2 \times S + 1 \times B$ 。然后，我们通过总流量和瓶颈流量分析了教练对奖牌的贡献，为各国提供了选择教练的优化路径，并帮助他们制定更精确的体育发展战略。

对于第五期，在上述问题的建模过程中，我们确定了一些策略，可以帮助国家奥委会增加奖牌数量。例如，随着女性参与度的提高，国家奥委会可以专注于混合性别项目，以增加获胜的机会。如果一个国家是东道主，它可以申请增加国内优势项目，优化基础设施，并增加财政支持，以利用主场优势并提高奖牌数。对于未被少数几个国家垄断的项目，强国应分析这些项目并投资培养下一代运动员，而新兴体育国家则可以通过精准定位项目和吸引优秀教练，在这些非垄断项目中取得突破。

关键词：预测、LSTM、随机森林、奥运会、性能建模、图论

目录

1 介绍	3
1.1 背景	3
1.2 问题重述	3
1.3 我们的工作	4
2 假设和依据	4
3 评分	4
4 数据清理	5
5 任务1&2：基于LSTM的奖牌数预测	5
5.1 数据分析	5
5.1.1 基于皮尔逊相关系数的相关矩阵分析	5
5.1.2 基于K均值的奖牌国家分类模型	7
5.2 模型选择	9
5.3 长短期记忆网络	9
5.4 LSTM的实施	9
5.4.1 量化东道国的影响	10
5.4.2 构建LSTM模型	10
5.5 从结果中预测的区间	11
5.6 LSTM的优势	14
6 任务3：基于随机森林的首次获奖国家预测	15
6.1 随机森林模型	15
6.2 数据准备和预处理	15
6.2.1 数据表构建	15
6.2.2 数据集创建	15
6.3 模型构建	16
6.3.1 变量分析	16
6.3.2 模型训练	16
6.3.3 模型性能可视化	17
6.3.4 模型预测	18
6.4 模型测试	18
6.4.1 测试集构建	18
6.4.2 测试	18
7 任务4：基于网络流的“大教练效应”的定量图论模型	19
7.1 模型背景	19
7.2 模型构建	19
7.2.1 基本有向图的构造	19
7.2.2 重量计算	20
7.3 模型结果和分析	22
7.4 验证和策略建议	24
8 信	24
9 模型评估	25
9.1 优势	25
9.2 弱点	25

1 介绍

更快、更高、更强——一起。

——国际奥林匹克委员会

1.1 背景

巴黎奥运会吸引了全球的目光，赛事本身备受关注，尤其是各国运动员的奖牌成绩。来自世界各地的运动员奋力拼搏，争取在奖牌榜上占有一席之地。除了传统奥运强国和东道主的奖牌争夺战备受瞩目外，一些排名较低的国家，如阿尔巴尼亚、佛得角、多米尼克和圣卢西亚，在巴黎奥运会上首次获得奖牌，也引起了广泛关注。然而，仍有超过60个国家未能获得奥运奖牌。

回顾历史，各国在奥运会的奖牌表现呈现出一定的模式。每次奥运会前，都会有一个“虚拟奖牌榜”来预测各国的表现。例如，在巴黎奥运会之前，尼尔森·格拉森诺发布了其最终的2024年奥运会虚拟奖牌榜（VMT）预测。那么，这些预测具体依赖于哪些因素呢？事实上，奖牌预测通常在奥运会开始时通过建立数学模型来做出。这些模型考虑了已知的运动员参赛计划，并分析了过去的金牌和总奖牌数，以预测未来的奖牌排名。这样的预测不仅对体育分析师、研究人员和政策制定者有价值，还有助于各国更好地把握影响奥运表现的趋势。

1.2 问题重述

根据背景信息和问题的限制，我们必须完成以下任务：

- 任务1：预测美国洛杉矶2028年夏季奥运会的奖牌榜，输出表现更好的国家和表现更差的国家。
- 任务2：分析奥运会运动项目数量和类型与奖牌数量之间的关系，输出每个国家最重要的运动项目，并分析主办国选择的运动对表现的影响。
- 任务3：对于尚未获得奖牌的国家，预测它们在下一届奥运会上赢得第一枚奖牌的概率，并提供这一预测概率的估计值。
- 任务4：探索伟大教练效应对团队运动的影响，寻找伟大教练效应影响的证据，并估计其影响的强度。最后，选择三个国家推荐值得投资的运动项目，并估计其影响。
- 任务5：提供关于奥运奖牌数量的原始见解，并解释这些见解如何影响奥林匹克委员会的决策。

1.3 我们的工作

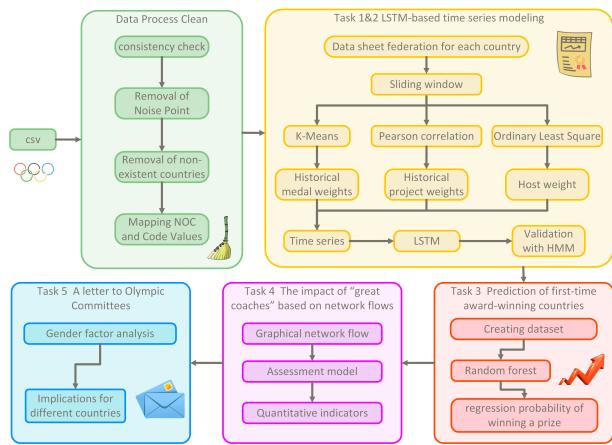


图1：CHN与程序的相关性

2 假设和依据

为了简化问题，我们做出以下假设，每个假设都是合理的。

- 假设1：假定各国每年获得的奖牌数量和类型是根据特定的选择规则分配的：这些规则可能包括运动员个人表现、国家的历史成就、项目的评分标准等。每个国家的获奖情况由一套明确的评判机制和标准决定，因此奖牌的分配可以视为基于这些规定的公式化过程，确保每个国家都有公平的机会，遵循相同的规则。
- 假设2：假定数据集中未包含的潜在影响因素在奖牌数产生中不起重要作用：在分析模型时，不考虑未给出的变量，并假定它们不会影响最终的奖牌数结果。如果这些因素有可能在分析中引入噪声，可以通过数据预处理技术消除干扰，例如去除异常值或使用标准化方法，从而使结果更准确和可靠。

3 评分

符号	描述
h	对流换热系数
N	国家数目
Y	获奖年份
M_i	国家 i 的奖牌总数
S_i	国家 i 的银牌数
B_i	国家 i 的铜牌数
ρ	皮尔逊相关系数

4 数据清理

研究开始时，对原始数据进行了系统性清理和预处理。

在运动员数据处理方面，我们对summeroly_athletes.csv数据集进行了多项操作，包括删除重复项、处理缺失值、标准化和转换数据类型。该数据集包含基本的运动员信息（姓名、团队、国家奥委会、运动项目、赛事、奖牌等）。此外，我们发现字符串文本中的空格影响了后续的数据分析，因此使用了str.strip()来去除前后空格，以及str.replace()来移除内部空格。我们还删除了无效信息，并对summeroly_programs.csv数据集进行了反向处理，以便于未来的使用。

在国家奥委会（NOC）验证环节中，我们使用运动员数据集作为有效国家奥委会的主要参考来源，并与主办国和奖牌数据集进行了交叉验证。为了确保数据一致性，我们从主办国和奖牌数据集中移除了未出现在运动员数据集中的国家奥委会条目。这一步骤为后续的数据分析提供了可靠的基础。

为了确保数据质量，我们进行了严格的验证和质量检查，包括确保所有数据集中NOC代码的一致性、核实年份范围和时间连续性，以及检查逻辑约束，例如金牌数量不超过总奖牌数。在整个数据集整合过程中，我们始终注意维护数据的完整性。

最后，我们对数据进行了必要的额外转换，创建了历史表现的汇总视图，统一了每个数据集的数据格式，并准备了一个标准化的数据结构以供后续分析和建模。通过这一系列严格的清洗过程，我们确保了数据集的一致性、完整性和规范性，为预测建模阶段奠定了坚实的基础。这些步骤不仅提高了数据质量，还增强了后续分析结果的可靠性。

5 任务1和2：基于LSTM的奖牌数量预测

5.1 数据分析

首先，我们为以前获得过奖牌的国家构建了一个特征相关矩阵，以确定一些国家的优势运动。随后，我们为每个国家创建了应急表，包括获奖年份、总奖牌数、金牌数、主办国状态、参赛人数、参与赛事数量、主办方设立的赛事总数以及往届赛事等特征。最后，我们使用K-均值聚类将以往获得奖牌的国家分为三类，这为不同类型的奥运国家建立独立的预测模型奠定了基础。

5.1.1 基于皮尔逊相关系数的相关矩阵分析

- 第一步：数据准备：将奖牌数数据集与事件数据集对齐，以确保每一行代表一个国家在特定事件中获得的奖牌数量。
- 第二步：皮尔逊相关系数：皮尔逊相关系数适用于分析服从正态连续分布的数据，计算公式如下：

$$\rho(x, y) = \frac{\text{cov}(x, y)}{\sigma(x) \cdot \sigma(y)} = \frac{E[(x - \mu_x)(y - \mu_y)]}{\sigma(x) \cdot \sigma(y)} \quad (1)$$

- 第三步：相关性计算：对于每个国家，我们将该国在不同项目中获得的奖牌数与不同奥运项目颁发的总奖牌数相结合，计算皮尔逊相关系数。皮尔逊相关系数量化了变量之间关系的强度。变量相关性的强度如图1所示。我们可视化了所得的相关矩阵，并以中国为例，生成了热图，如表1所示。

表1：变量相关强度

相关系数 强度	非常强 相关系数	强的 相关系数	中等 相关系数	弱 相关系数	非常弱 或否 相关系数
绝对值 价值 相关系数 系数	0.8–1	0.6–0.8	0.4–0.6	0.2–0.4	0–0.2

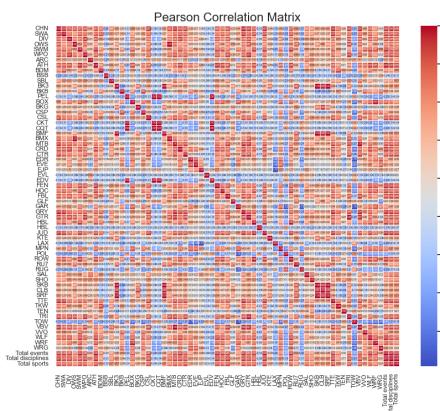


图2：中国热图

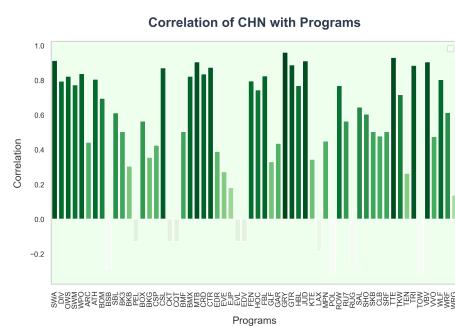


图3：CHN与程序的相关性

- 第四步：分析优势体育项目：返回包含三个独立变量的相关系数最高的系列。这些代表了对一个国家总奖牌数影响最大的三项赛事，即该国的优势体育项目。以中国为例，上述结果和热图显示，中国的优势体育项目是体操、乒乓球和跳水。
- 第五步：构建列联表：为每个国家创建一个列联表，包括获奖年份、主办国地位、总奖牌数、金牌数、主办国设立的赛事总数、参赛赛事数量、优势运动和平均获奖牌数。

5.1.2 基于K均值的奖牌国家分类模型

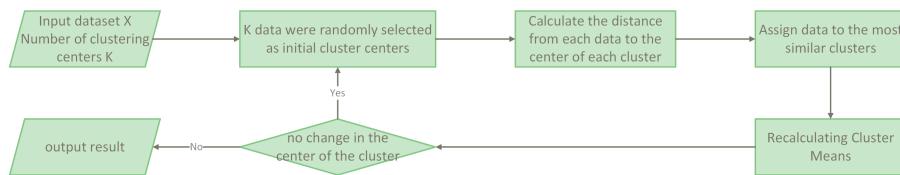


图4：K-Means流程图

为了准确预测奖牌，科学地对参赛国家进行分类至关重要。在根据奖牌得失稳定性对国家队进行分类之前，我们首先需要定义稳定性。为此，我们使用两个关键维度：NOC参与频率和运动项目多样性。基于这些维度，我们应用K-Means聚类算法对国家进行分类。

1. 特征构造

NOC参与频率：`fparticipation = count(NOC)`

体育多样性指数：`fdiversity=unique_count(Sports)`

特征标准化： $X' = \text{MinMaxScaler}$

2. 通过最小化簇内平方和 (WCSS) 目标函数实现聚类。

$$d(x_i, C_j^{(r)}) = \sqrt{\sum_{l=1}^d (x_{il} - C_{jl}^{(r)})^2} \quad (2)$$

$$C_j^{(r+1)} = \frac{1}{|S_j^{(r)}|} \sum_{x_i \in S_j^{(r)}} x_i \quad (3)$$

首先，我们计算了数据集athletes_with_gold_medal中每个NOC出现的频率以及与每个NOC相关的独特运动的数量。使用MinMaxScaler对NOC和运动的数据进行了归一化处理，并将这两个特征合并到一个数据集中进行K-Means聚类。

在聚类过程中，观察到美国的数据点表现出显著的异常特征。这主要是因为美国远远超过其他国家。

各国在NOC参与频率和运动项目数量上都存在显著差异，导致数据分布独特。这种独特性不仅反映了美国在奥运会上的全面主导地位，同时也给聚类分析带来了挑战。为了确保聚类结果的合理性，需要对聚类方法进行适当的调整。

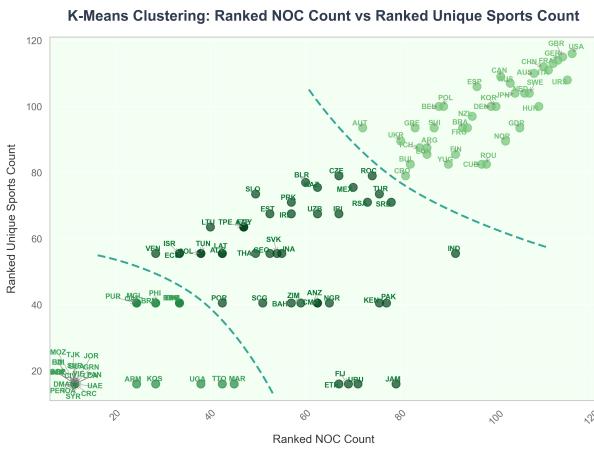


图5：K-均值聚类图

我们决定以NOC排名和运动项目出现频率作为聚类的基础，将各国划分为不同的组别。此外，在模型选择阶段，我们尝试了K均值和DBSCAN两种方法来聚类运动员奖牌数据集和运动员金牌数据集。结果表明，使用K均值模型对运动员金牌数据集进行聚类表现最佳。

根据图4，我们得到了最终的聚类结果：类别1包括37个国家获得金牌，类别2包括36个国家，类别3包括43个国家。

表2：组和特征描述

组	功能描述
集群1（奥运强队）	参与频率高 赛事分布均衡，夺牌能力稳定
集群2（发展体育国家）	适度参与频率 赛事分布相对集中，奖牌成绩起伏不定
集群3（不稳定的国家）	参与频率低，奖牌成绩不稳定，优势项目依赖性强

5.2 模型选择

考虑到我们处理的数据与特定的时间点相关，并以均匀的间隔（每四年一次）进行测量，我们决定使用时间序列分析来预测奖牌数量。

传统的时序预测方法，如常用的自回归积分滑动平均法（ARIMA），仅依赖于单一的时间序列数据源。然而，预测奖牌数量通常需要结合多个变量，而不仅仅是历史上的奖牌数量。

随着机器学习和人工智能的发展，各种深度学习算法已被应用于时间序列预测。例如，长短期记忆网络（LSTM）能够解决循环神经网络（RNN）在捕捉长期周期和季节性模式方面的局限性。在每个时间步长中，LSTM都有一个记忆单元，为网络提供选择性的记忆功能。这使得LSTM能够决定在每个时间步长保留哪些内容。

此外，LSTM可以处理多变量分析，其分析多个变量的能力与基于K-均值聚类的国家三分类分组非常契合。这使得不同类型的国家可以采用不同的变量权重策略。

5.3 长短期记忆网络

该研究采用LSTM网络作为核心预测模型，由三个关键组件构成：遗忘门、输入门和输出门。遗忘门控制历史信息的保留，输入门管理新信息的更新，输出门决定信息的输出。这三个组件共同构成了一个完整的记忆机制：

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (4)$$

$$i_t = \text{sigmoid}(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (5)$$

$$o_t = \text{sigmoid}(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (6)$$

这里， h_t 表示时间t的隐藏状态， x_t 是输入向量。

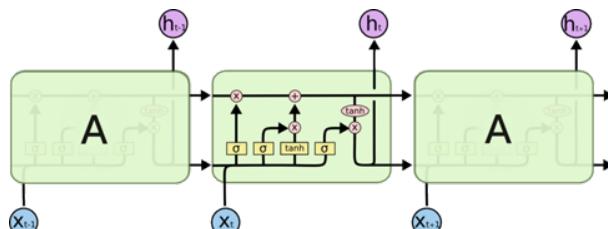


图6：LSTM机制的工作流程图

5.4 LSTM的实施

我们对数据进行了归一化处理，以减少特征之间尺度差异的影响，并确保训练过程中的稳定性。随后，我们创建了输入。

采用滑动窗口法对原始奖牌数据进行时间窗分割，时间步长设置为20年，即每年的奖牌数与前20年的数据相关。

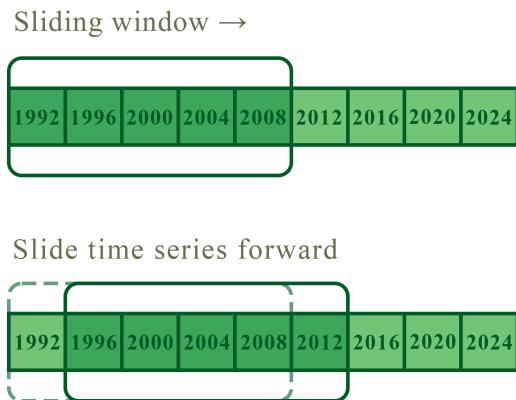


图7：滑动窗口

5.4.1 量化东道国的影响

我们采用了普通最小二乘法（OLS）方法，以总奖牌数为因变量，主办国和赛事为自变量。模型得出的相对风险比为0.758，调整后的相对风险比为0.756，两者均相对较高，表明该模型能够解释超过75.8%的总奖牌数变化，显示出良好的拟合度。自变量“主办国”的p值小于0.05，表明其具有高度显著性和统计有效性，对总奖牌数有显著影响。

表3：普通最小二乘法的结果

a	coef	std err	t	P> t
常量	10.4972	5.090	2.062	0.041
主人	26.9574	4.975	5.419	0.000
运动	1.3783	0.226	6.103	0.000

5.4.2 构建LSTM模型

如前所述，我们使用K-Means聚类将获奖国家分为三类：

- 第1类：经常参加奥运会并举办多种项目的老牌奥运强国，如美国和中国。
- 第2类：经常参与但表现不一致的国家，如卢森堡和新加坡。
- 第3类：印度和牙买加等经常参与但表现不稳定。

对于每个国家，我们构建了一个时间序列，其中包括年度奖牌数、金牌数、主办国地位、参赛人数、参与赛事数量、主办国设立的赛事数量以及历届奥运会各项目奖项分布等变量。其中，奖牌数、金牌数、主办国地位、参赛人数、赛事数量和主办国设立的赛事数量被视为该国历史上的奖牌表现预测因素。各项目奖项分布与皮尔逊相关系数的乘积被用来表示该国的优势项目。

按类别划分的权重策略：

- 第1类：这些国家通常实力强劲，在多个项目中都获得了奖牌，而不仅仅依赖于优势项目。对于这类国家，由于作为东道主的影响显著，我们将这一因素乘以OLS回归得出的影响，均匀分配其他历史预测变量的权重，并将优势项目的权重设为零。
- 第2类：对于这些国家，奖牌预测取决于过去的参与和表现，以及某些优势运动。对于这些国家，所有变量都包括在内，并且优势运动的权重计算为 $tweight = (1 - w) \times$ 皮尔逊相关系数
- 第3类：这些国家的参赛次数很少，历史奖牌数也很低，预测就像估计获得奖牌的概率一样。由于过去的参赛数据不太相关，只考虑优势项目，历史数据的权重被设为零。

模型架构设计：

我们采用了包含50个存储单元的LSTM层作为主要特征提取器，能够有效捕捉奥运会周期内的短期和长期依赖关系。在LSTM层之后，使用了一个全连接层进行特征整合。最后，输出层由两个神经元组成，分别用于预测总奖牌数和金牌数。

考虑到最近的奥运表现更能反映未来的结果，我们在模型中引入了时间衰减加权机制。这使得模型能够保留历史信息，同时更多地关注最近的数据趋势。

5.5 从结果中预测的区间

我们通过多次训练和测试迭代计算了每个国家模型的预测区间。每个国家的模型都进行了10次训练和预测，以获得最终结果的分布。去除异常值后，取最大值和最小值作为奖牌数或获胜概率的范围。计算了每个国家模型的平均损失，并通过平均所有国家模型的损失来确定整体模型性能。

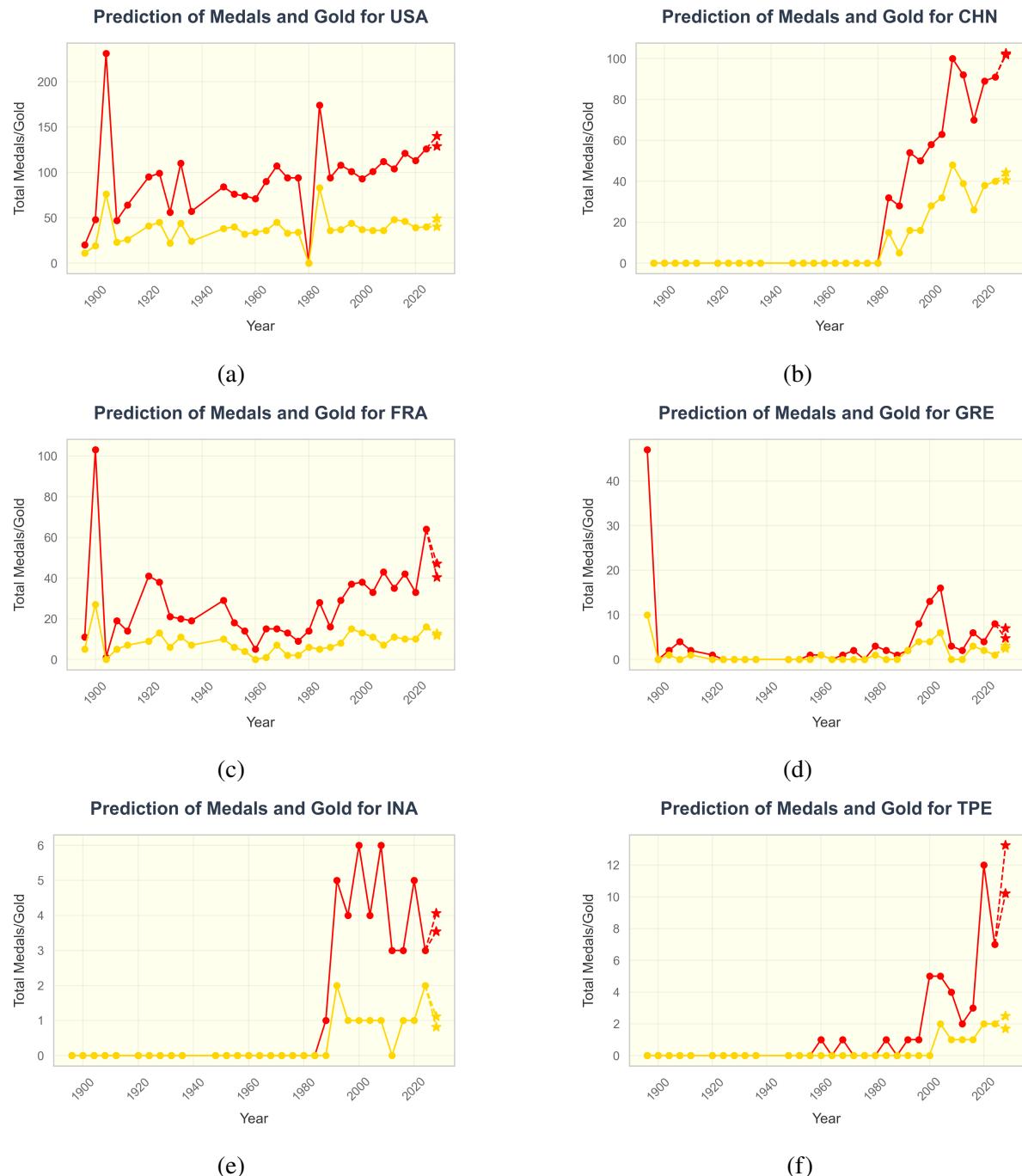
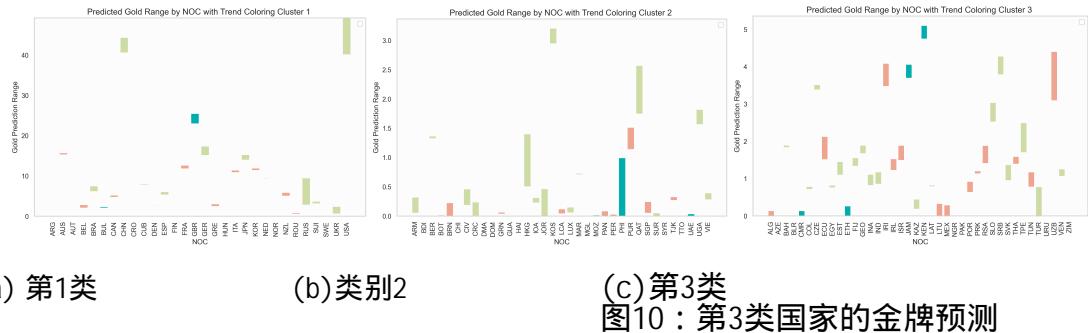
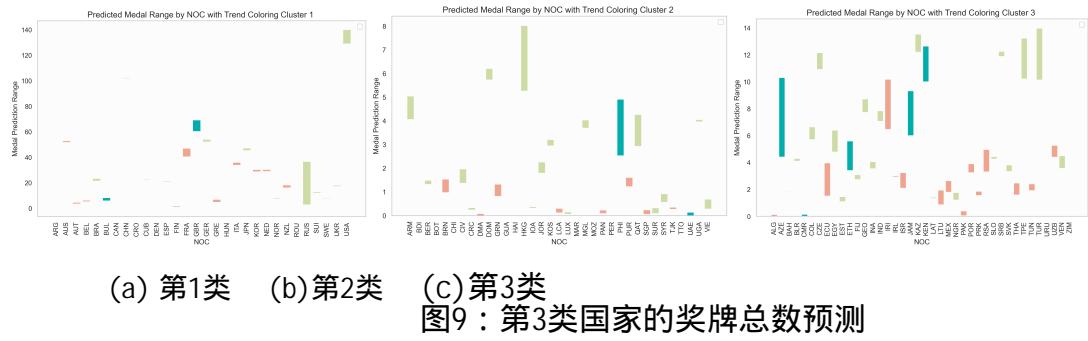


图8：不同国家的总奖牌数和金牌数预测区间

使用训练好的模型，对三个类别的国家进行了预测。归一化的预测结果被反归一化回原始值，结果如图8所示。绿色表示奖牌数量增加，红色表示减少，蓝色表示趋势稳定。条形图代表了数值范围。



如上图所示，美国和中国代表的73个国家预计将获得更多的奖牌，而澳大利亚和加拿大代表的47个国家预计将获得更多的奖牌。

我们分析了以往奥运会新增项目的情况，并通过将主办国在这些新项目中的历史总奖牌数与其总体奖牌数相结合，进行了皮尔逊相关性分析。结果显示平均皮尔逊相关系数为0.4489306，表明主办国选择新项目对其总体奖牌数有显著影响。

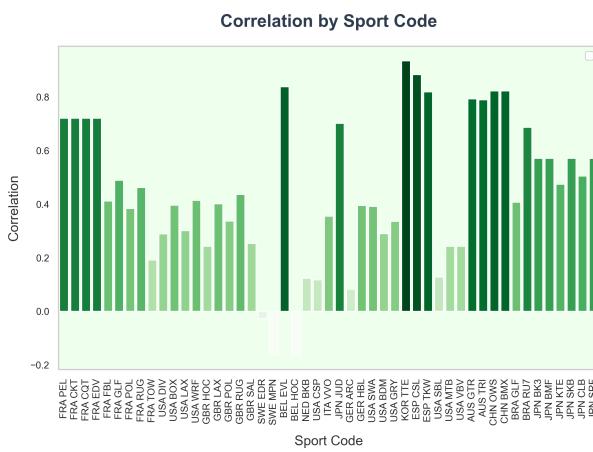
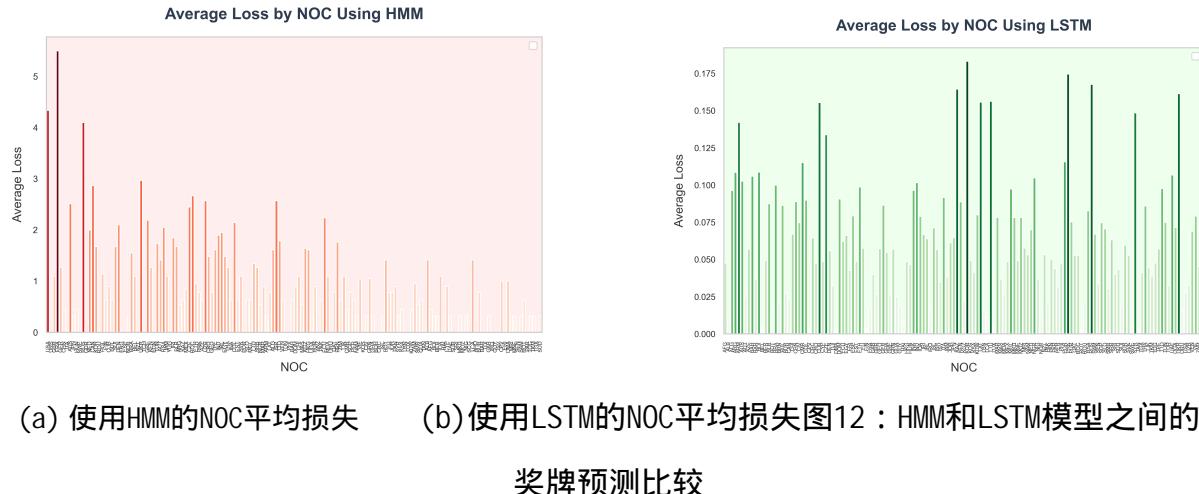


图11：按运动代码划分的相关性

5.6 LSTM的优势

使用LSTM模型预测最终奖牌数通常不仅仅基于历史奖牌数的时间序列，而是综合考虑了多个因素，从而能够分析历史因素和优势项目的影响力。为了评估LSTM模型在奥运会奖牌预测任务中的优势，我们与传统的隐马尔可夫模型（HMM）进行了比较实验。通过对两种模型在测试集上的预测结果与实际值之间的均方误差（MSE），我们观察到LSTM模型表现出显著的优势，其预测误差明显低于HMM模型，如图11所示。



从图中可以看出，与传统的隐马尔可夫模型（HMM）相比，LSTM模型在预测奥运奖牌方面表现明显更好。我们将其归因于以下原因：

- 1. 模型结构：LSTM的门控机制（包括遗忘门、输入门和输出门）使模型能够自适应地调节历史信息的保留。这使得LSTM模型能够捕捉到奥运强国奖牌数量的长期稳定增长趋势。相比之下，HMM受其马尔可夫假设的限制，只能建模相邻时间步之间的依赖关系，难以捕捉长期演变模式。此外，HMM在准确分析历史奖牌数极少的国家再次获胜的概率或预测高奖牌波动性或依赖优势项目的国家结果方面存在困难。如图[损失比较]所示，LSTM预测曲线与实际值的吻合度远高于HMM模型。
- 2. 特征表示能力：LSTM通过其存储单元，可以同时整合多个维度的信息，包括历史奖牌数、赛事参与度、运动员规模和其他关键因素。这种动态多维特征融合机制使得模型能够更准确地预测东道国效应和新兴体育强国崛起等复杂现象，而HMM的状态空间表示能力相对有限。
- 3. 优化目标：LSTM采用端到端梯度下降训练方法，直接优化预测误差。相比之下，HMM依赖于最大似然估计进行参数学习，这导致了预测误差与实际值之间的不匹配。

训练目标和实际预测任务之间的差异是LSTM在实验中表现出更强鲁棒性的主要原因。

6 任务3：基于随机森林的首次获奖预测 获胜国家

6.1 随机森林模型

要确定影响一个国家获得奖牌数量转变的关键因素，需要两个步骤：

- 首先，建立一个预测模型来预测奖牌数量从零到一的转折点。
- 第二，根据模型结果推断关键指标。

6.2 数据准备和预处理

6.2.1 数据表构建

构建三个数据表，分别为participation_by_year_country、participation_by_year_country_count和sport_count_pivot，分别记录每年的参赛人数、每年累计参赛次数和每年参加的项目数。

为了排除没有获得任何奖牌的国家，从参与国家/年表中删除了与奖牌统计表中NOC列所列国家相对应的条目。

在medals_by_year和medals_by_year_gold数据集中，遍历了每个国家的历史数据，以确定它们首次获得奖牌的年份。如果某一列的所有值均为零，则返回None。结果随后存储在first_medal_filtered中供预览。该数据集仅包含首次获得奖牌或金牌的年份大于或等于1920年的国家。

6.2.2 数据集创建

第一次迭代：从参与国家/地区数量、参与国家/地区和运动项目数量数据框中提取每个国家/地区的NOC相关信息：

- participation_count：该国首次获得奖牌的那一年的参赛次数。
- 事件参与度：该国在首次获得奖牌的那一年所参加的事件数量。
- sports_count：该国在第一年参加的运动项目数量获得了一枚奖牌。这些提取的值随后被添加到，这将作为后续数据集的基础。

然后将这些提取的值添加到new_table_data中，这将作为后续数据集的基础。

第二次迭代：为了从第一个奖牌年之前的四年中获得数据，从第一个奖牌年之后的四年中减去数据。

第三次迭代：对于每个从未获得过奖牌的国家，将该国的NOC、参与人数、赛事参与和运动项目数量添加到new_table_data中作为列表。

最后，将new_table_data中存储的所有数据转换为Pandas的DataFrame。在填补缺失值和清理无效数据后，形成了最终的数据集tree_dataset。

6.3 模型构建

6.3.1 变量分析

经过对有限数据的仔细筛选和考虑，我们选择了三个特征变量来训练模型，具体变量列表如下：

表4：变量表

指示器	说明
参与人数	来自该国的与会者人数
活动参与人数	该国参与的次数
运动项目	该国每次参加的运动项目数量

同时，我们的目标变量是Will_Earn_Medal，它表示是否获得奖牌（1或0）。

6.3.2 模型训练

使用StandardScaler对数据进行标准化后，将数据集分为训练集和测试集，其中80%用于训练，20%用于测试。

我们利用上述特征作为随机森林模型的输入参数，在训练集上进行训练。随机森林模型的参数网格如下：

表5：森林模型的输入参数

指示器	说明
估计数的数目	[50, 100, 150, 200]
最大深度	[无, 10, 20, 30]
最小样品片数	[2, 5, 10]
最小样品分割	[1, 2, 4]
最大功能	['auto', 'sqrt', 'log2']

我们使用了来自scikit-learn (sklearn) 机器学习库的RandomForestRegressor算法，以及GridSearchCV方法进行超参数调整。从GridSearchCV获得的最佳超参数组合如下：

表6：最佳超参数组合

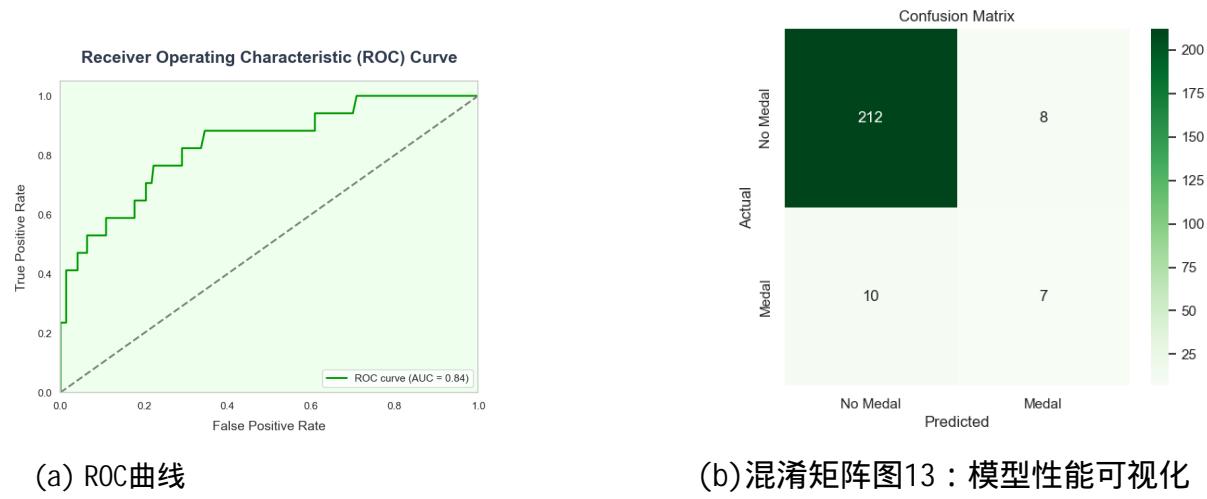
指示器	说明
估计数的数目	200
最大深度	10
最小样品片数	1
最小样品分割	2
最大功能	sqrt

6.3.3 模型性能可视化

评估模的性能

- 混淆矩阵：显示模型预测与真实标签之间的关系，有助于分析模型所犯的错误类型。
- ROC曲线：绘制接收器操作特征（ROC）曲线，该曲线测量模型在不同阈值下的性能。

混淆矩阵和ROC曲线如下所示：



混淆矩阵显示模型的预测准确率为94%，假阳性率仅为0.039 (9/233)，表明模型具有高度可靠性，大多数样本被正确分类。测试集中预测值与真实标签之间的匹配度很高。

该模型的AUC（曲线下面积）为0.87，表明具有较强的区分能力，能够有效区分正类和负类。此外，ROC曲线在低假阳性率（FPR）区域迅速上升，表明模型在保持低假阳性率的同时实现了高召回率。该模型的真阳性率值得称赞。

6.3.4 模型预测

使用训练最好的模型best_rf_model，我们预测了新标准化数据的结果。predict_proba()方法返回每个样本属于每个类别的概率，这使我们能够计算每个国家获得奖牌的概率。

6.4 模型测试

6.4.1 测试集构建

为了验证模型在实际预测任务中的表现，我们从运动员数据集中识别并选择了从未获得过奖牌的国家的历史样本。我们采用时间分组方法以确保数据的连续性和完整性。在标准化特征数据并应用质量控制措施后，最终形成了一个高质量的测试数据集。

为2024年生成了一个包含从未获得过奖牌的所有国家相关统计数据的新数据集。

6.4.2 测试

利用训练好的随机森林模型best_rf_model，我们预测了每个国家的特征，并计算了每个国家获得奖牌的概率。这些概率被存储在一个列表中，进行了排序，并确定了最有可能获得奖牌的国家。然后生成一个条形图，以可视化每个国家获得奖牌的预测概率。

通过概率分布分析，发现预测结果具有明显的分层特征，预测概率大于0.20的国家主要集中在体育基础设施发达的新兴市场国家，这些国家一般都对体育有持续的投资，并且体育项目专业化程度较高。

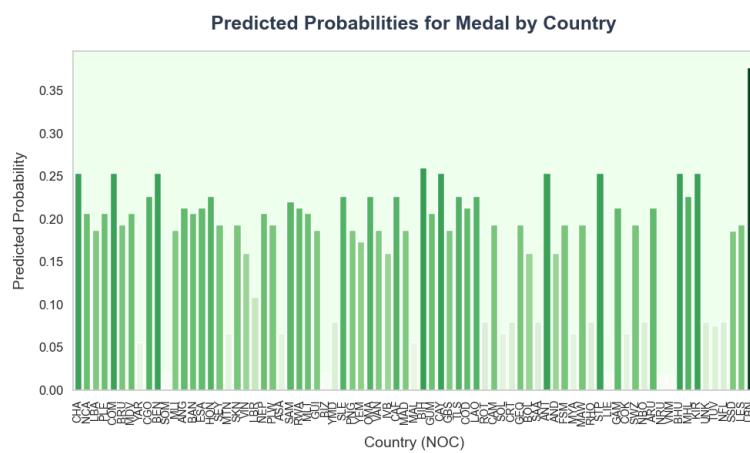


图14：按国家列出的奖牌预测概率

7 任务4：基于网络流的“大教练效应”的定量图论模型

7.1 模型背景

在国际比赛中，如奥运会，运动员的国籍通常受到限制；然而，教练不必是他们所指导国家的公民。这使得他们可以在各国之间自由流动。在某些情况下，教练更换对运动队的表现产生了显著影响，尤其是在“伟大教练效应”下。这样的教练可以帮助团队或运动员突破表现障碍，提高成绩。该模型旨在通过数据分析验证“伟大教练效应”的存在，并量化其对奖牌数的贡献。

在本节中，我们需要完成：

- 数据分析：探索教练的更换是否与不同国家和运动项目的奖牌数的变化有关。
- 模型构建：通过基于图论的网络流模型分析教练效应对奖牌数的影响。
- 影响预测：确定哪些运动适合投资“伟大的教练”，并分析他们的潜在影响

7.2 模型构建

本部分旨在基于图论中的网络流模型构建一个有向图，该模型能够有效反映教练流动如何影响各国奖牌数的变化。通过这一模型，我们可以量化教练的影响，特别是教练流动或更换后对国家体育成绩的影响。图中的节点代表不同的国家，而有向边则表示教练在各国之间的流动及其对奖牌数的贡献。利用总流和瓶颈流算法，我们可以深入分析教练的角色，评估其对奖牌数的影响，并为未来的教练选拔和资源配置提供科学建议。

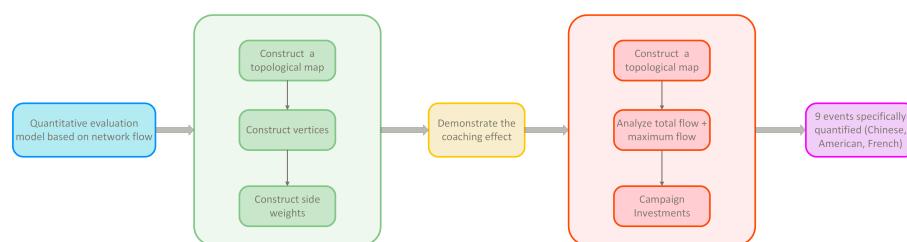


图15：基于网络流模型的定量评价框架

7.2.1 基本有向图的构造

在基于图论的网络流模型中，我们首先构建一个基本的有向图来反映教练的移动。节点代表国家，而有向边则表示教练从一个国家到另一个国家的流动路径（包括奥运会）。

游戏作为时间段)如果教练留在同一个国家并参加多次奥运会,可以使用自环边

- 节点:每个节点代表一个国家,并存储该国在特定时间段内获得的金、银、铜牌变化(?)G、?)S、?)B)。

节点集V:表示参与者和国家,V={vi|i|I},其中I是国家的索引集。

- 有向边和方向:有向边表示教练从一个国家到另一个国家的流动。每条有向边的方向由教练的转移方向决定,即从原籍国(流出国)到目标国(流入国)。有向边的容量代表教练对目标国奖牌数变化的贡献。

边集E:表示节点之间的连接E={(vi,vi)|i,j|I}

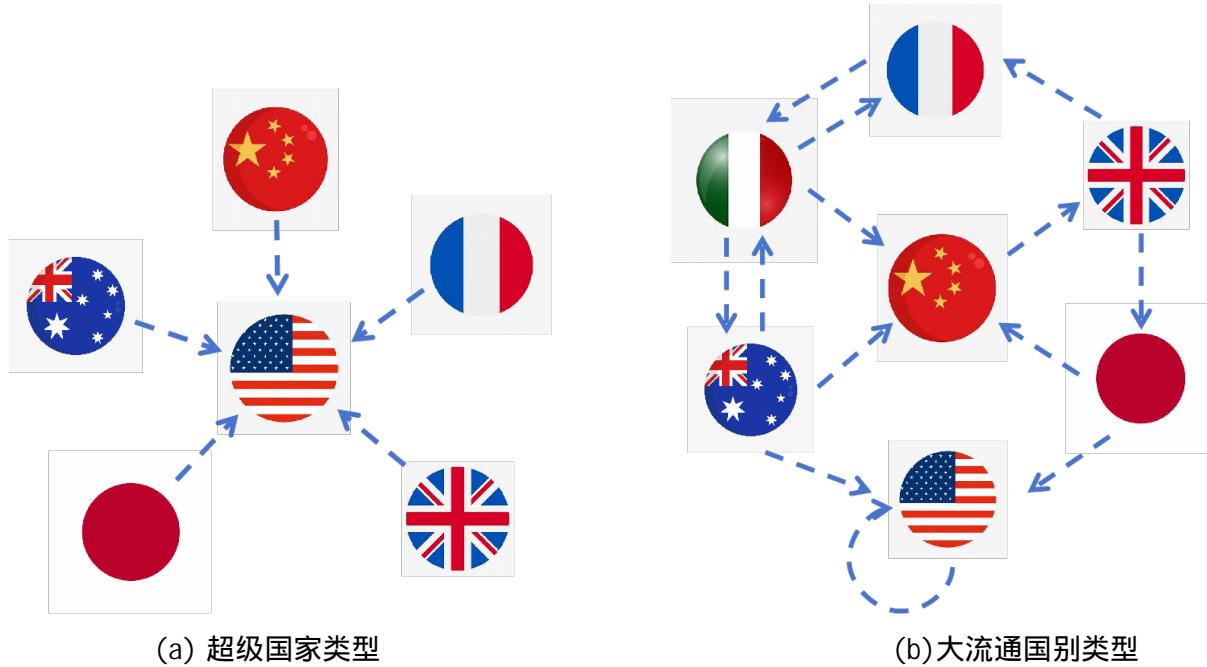


图16:图论模型

7.2.2 重量计算

在这个模型中,每条有向边的权重w代表教练调动前后奖牌数的变化,通过奖牌数的增减来量化,权重计算公式如下:

$$w = w_G \Delta G + w_S \Delta S + w_B \Delta B \quad (7)$$

由于缺乏详细的教练数据来量化奖牌变化对每个国家表现的具体影响,我们为每种奖牌类型分配固定的权重系数:wG=4,wS=2,wB=1。

数据提取我们使用函数filter_and_count_medals来筛选和统计奥运会数据，以便在特定条件下分析不同国家在各种运动项目中的表现。该函数根据输入条件（如国家、年份、性别和运动项目）过滤数据，并计算不同类型奖牌（金牌、银牌、铜牌）的数量。

同一国家内的循环边对于教练在同一国家内的移动，我们计算该国连续两届奥运会奖牌数的变化，并应用奖牌权重系数来获得得分。自环边的计算公式如下：

$$w = 4 \times \Delta G + 2 \times \Delta S + 1 \times \Delta B \quad (8)$$

教练在不同国家之间的流动对于在不同国家之间流动的教练，我们计算两国奖牌数的变化，乘以权重系数，然后计算目标国家正增长与原国家负增长之间的差异，以获得有向边的权重ww。

$$W = W_A - W_B \quad (9)$$

- A是流入国；
- B是流出国；
- wA是国家A奖牌数变化的加权得分；
- wB是B国奖牌数变化的加权得分。

郎平教练在中国和美国之间的运动。

在这个模型中，郎平的执教运动被看作是国家间的影响转移，从美国到中国或反之亦然。我们需要根据奖牌数的变化来计算郎平从美国到中国的转移权重。

例如，郎平从2005年到2008年执教美国女排，2008年美国女排获得银牌。2012年，她回到中国，带领中国女排在2016年里约奥运会上夺得金牌。

通过利用数据提取方法，我们计算了美国女子排球队在2004年、2008年、2012年和2016年的奖牌数，以及中国在2012年和2016年的奖牌数。权重计算方法包括计算奖牌数的变化，应用权重系数，并确定郎平执教变动对两国的影响。

在2016年期间，郎平从美国搬到了中国。奖牌数量

运动	教练	流路	共计流	瓶颈流量
排球	郎平	2004 2008 (美国自转)：金牌+0，银牌+1，铜牌+0，得分=2	7	2

运动	教练	流路	共计流	瓶颈流量
		2012 2016 (美国 中国) : 美国金牌+0 , 银牌-1 , 铜牌+1 ; 中国金牌+1 , 银牌+0 , 铜牌+0 , 得分=5		

7.3 模型结果和分析

本节分析了中国、美国和法国奖牌数的变化，重点关注“名帅效应”的影响，利用流动网络模型量化教练变动对奖牌数的影响，并根据模型结果提出投资策略。

(1) 中国：乒乓球、跳水、体操

运动	教练	交通路径	总流量	瓶颈流量
乒乓球	刘国良	2004年：金牌+1 , 银牌+1 , 铜牌+0 , 得分= -2 2004 2008：金牌+2 , 银牌+0 , 铜牌+0 , 得分=8	6	-2
潜水	周继红	2004 2008：金牌+1 , 银牌-1 , 铜牌+2 , 总分=4 2008 2012：金牌-1 , 银牌+2 , 铜牌-2 , 总分= -2 2012 2016：金牌+1 , 银牌-1 , 铜牌+0 , 总分=2	4	-2
体操	黄玉斌	1996年：金牌+1 , 银牌-1 , 铜牌+1 , 得分=3 1996年和2000年：金牌-1 , 银牌+0 , 铜牌-1 , 得分= -5	-2	-5

表8：体育统计

? 乒乓球：在刘国梁教练的指导下，乒乓球表现出相对较大的总流量(6)，但瓶颈流量为-2，说明虽然教练的贡献是显而易见的，但某些限制因素限制了改进的潜力。

? 潜水：周继红的执教时期相对稳定，总流量为4，瓶颈流量为-2，说明尽管教练做出了贡献，但项目经历了显著的波动，特别是在2008年到2012年之间。

? 体操：黄玉斌教练组的体操队表现不佳，总流量为-2，瓶颈流量为-5，说明教练贡献有限，外部对项目的影响较大。

乒乓球是很有前途的项目，尽管存在瓶颈，教练的作用还是很大的，这表明有必要进一步投资“伟大教练”效应

在这一领域，虽然跳水技术比较稳定，但由于存在瓶颈流，需要在其他方面进行改进。体操需要更多的关注和提高，因为教练的贡献较小，而总流量和瓶颈流量都显示出很大的改进空间。

(2)美国：体操、篮球、排球

运动	教练	交通路径	总流量	瓶颈流量
体操	贝拉·卡罗利	2004年：金牌+1，银牌+4，铜牌+0，得分=12 2004—2008年：金牌+1，银牌+1，铜牌+0，得分=6 2008—2012年：金牌+1，银牌-4，铜牌+0，得分=-4	14	-4
篮球	迈克·布热津斯基	2004—2008：金牌+1，银牌-1，铜牌+0，得分=2 2008—2012：金牌+0，银牌+0，铜牌+0，得分=0 2012—2016：金牌+0，银牌+0，铜牌+0，得分=0	2	0
排球	卡奇·基拉利	2008—2012：金牌+0，银牌+0，铜牌+0，得分=0 2012—2016：金牌+0，银牌-1，铜牌+1，得分=-1 2016—2020：金牌+1，银牌+0，铜牌-1，得分=3	2	-1

表9：体育统计

- 体操：Bela Karolyi 的总流量相对较高（14），表明教练贡献显著，但瓶颈流量为负（-4），表明表现改善受到外部因素的限制。
- 篮球：在格雷格·波波维奇的执教下，球队总流量为2，瓶颈流量为0，表明球队表现稳定，波动小。球队多次获得奥运会金牌，教练的能力也很突出。
- 排球：在Karch Kiraly教练的指导下，总流量为2，瓶颈流量为-1，表明教练贡献有限，由于外部限制导致表现波动。

篮球表现出稳定的性能，但教练的贡献相对较小，项目表现受限。排球则显示出显著的波动性，教练的影响有限。体操虽然教练的贡献很大，但由于外部限制需要进一步改进。体操项目最适合进一步投资于“优秀教练”效应。

(3)法国：击剑、篮球、足球

运动	教练	交通路径	共计 车流	瓶颈流量
击剑	南斯拉夫·盖·乌布利	2016年：金牌+1，银牌+1，铜牌+1，得分=7 2020年（法国 中国）：中国金牌+1，银牌-1，铜牌-1；法国金牌+1，银牌+1，铜牌+0，总分=-5	2	-5
篮球	文森特·科莱	2012 2016：金牌+0，银牌+0，铜牌+0，得分=0 2016 2020：金牌+0，银牌+1，铜牌+0，得分=2	2	0
足球	蒂里·亨利	2016 2020：金牌+0，银牌+0，铜牌+0，得分=0 2020 2024：金牌+0，银牌+1，铜牌+0，得分=2	2	0

表10：体育统计

- 在伊夫·吉拉德的执教下，教练的影响是复杂的。尽管法国队的表现相对较高，但转战中国后的负面影响表明存在显著的外部限制。瓶颈效应是负面的，这表明成绩提升受到了限制。
- 篮球：文森特·科莱的执教使法国篮球队的表现略有提高，瓶颈流量为0，表明表现稳定，没有明显的外部限制。
- 足球：蒂埃里·亨利的执教对法国足球产生了积极的影响，尤其是在银牌方面。瓶颈流量为0，表明教练对项目绩效有直接影响。

7.4 验证和策略建议

分析了中国、美国和法国及其教练的表现后，本节探讨了“伟大教练”效应对奖牌数量的影响，特别是在教练流动的背景下。尽管运动员可能因国籍要求而难以更换国家，但教练可以自由地在各国之间调动，并对表现产生重大影响。根据分析，我们可以得出结论，“伟大教练”效应在某些运动项目中贡献显著，尤其是在体操、乒乓球和足球领域，教练的专业知识和经验带来了显著的进步。

8 信

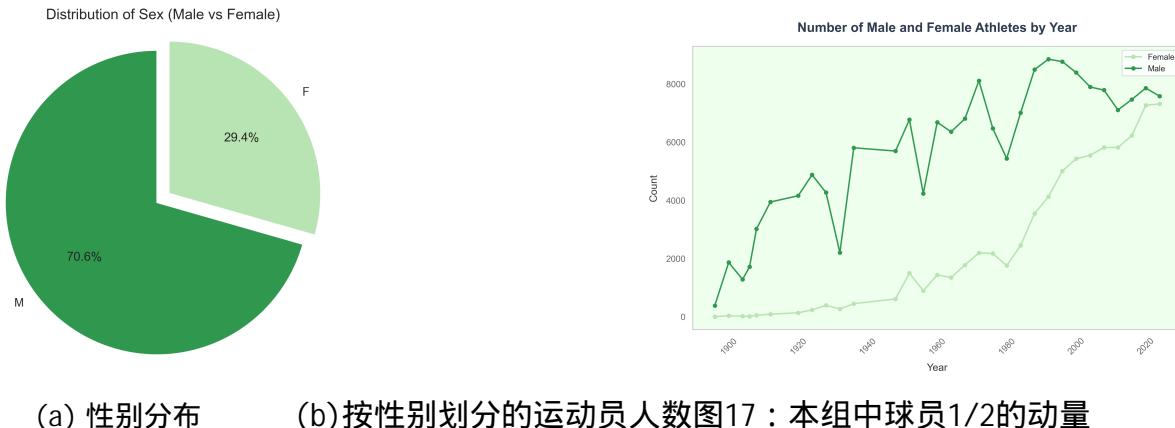
尊敬的先生或女士：感谢您百忙之中抽出时间阅读我们的分析结果，分享我们对奥运会奖牌预测的研究成果。作为全球竞技的舞台，奥运会不仅展示了体育卓越的巅峰，还促进了国际团结。在我们对奥运会奖牌预测的数学建模中，我们确定了几个关键点。

研究结果，包括性别参与的趋势，东道国对奖牌数量的影响，以及研究不足的运动项目的潜力。

这些见解不仅反映了历史和当前的趋势，而且还为国家奥林匹克委员会在筹备未来的奥林匹克赛事方面提供了切实可行的战略建议。

以下是这些发现的详细分析以及对未来规划的影响：

1. 通过对奥运会参与数据的统计分析，发现性别参与具有显著的进化特征。在早期的奥运会上，男性运动员占主导地位（约占70%）



(a) 性别分布

(b)按性别划分的运动员人数图17：本组中球员1/2的动量

9 模型评估

9.1 优势

1. 我们综合考虑了参赛人数、赛事数量、主办方和杰出教练等关键因素，因此我们进行了广泛的数据验证，我们的预测并不完全依赖于历史奖牌数。
2. 在构建模型时，我们通过实验验证将其与其他几个模型如隐马尔可夫模型和多元逻辑回归模型进行了比较。
3. 我们应用了各种数学建模技术，包括机器学习、神经网络和图论，并灵活调整变量和权重，以适应不同国家的奖牌预测。

9.2 弱点

1. 在数据清理过程中，由于时间限制，我们主要参考了运动员数据集来确定有效的国家奥委会，并从主办国和奖牌数据集中移除了未出现在运动员数据集中的国家奥委会。我们没有详细考虑各国历史上的变化。
2. 数据的缺乏限制了多维模型验证的能力。我们排除了GDP等因素，而多元线性回归可以更好地验证“伟大教练”效应。

参考文献

- [1]Koc C K. 幂运算滑动窗口技术分析[J]. 计算机与应用数学 , 1995, 30 (10) : 17-24.
- [2]Hartigan J A , Wong M A. A k-means clustering algorithm[J]. Applied statistics , 1979, 28(1) : 100-108.
- [3]Cohen I , Huang Y, Chen J, 等. 皮尔逊相关系数[J]. 语音处理中的噪声抑制, 2009: 1-4.
- [4]Dismuke C , Lindrooth R. 普通最小二乘法[J]. 方法与设计在结果研究中的应用 , 2006, 93(1) : 93-104.
- [5]Hochreiter S. 《长短期记忆》[J]. 神经计算, 麻省理工学院出版社, 1997年.
- [6]Eddy S R. 隐藏马尔可夫模型[J]. 当代结构生物学观点 , 1996, 6(3) : 361-365.
- [7]Breiman L. 随机森林[J]. 机器学习, 2001, 45: 5-32.
- [8]Scarselli F, Gorini M, Tsoi AC, 等. 图神经网络模型 [J]. IEEE神经网络汇刊 , 2008, 20(1): 61-80.