# **abstract**

在2024年巴黎夏季奥运会期间，观众不仅关注单项比赛，还关注各国的奖牌数量排名。尽管排名靠前的国家总是备受关注，但其他国家在奖牌数量上的突破也同样受到重视。每届奥运会前，都会有“虚拟奖牌榜”来预测各国的表现，但是这种预测通常不仅仅基于历史奖牌数据，也在很大程度上受到参赛运动员的影响，本文基于各国的参赛项目数量和类别、项目成绩、赛事主办方、特色项目等信息，进行建模。

针对问题一、二，在数据预处理阶段，我们使用**皮尔逊相关系数**计算国家与不同项目之间的相关性，并将相关系数在[0.8, 1]范围内的项目视为该国家的特色项目。我们把获得过奖项的国家用**K-Means聚类**分成三类国家。针对这三类国家，我们采用**LSTM模型**进行奖牌数的训练和预测 。在权重分配上，类别1国家（参赛次数多且成绩稳定的老牌强国 ）根据参赛人数和项目数的欧式距离排序后归一化，采用**最小二乘法回归**调整东道国对奖牌数的影响，其他历史依据权重均分，特色项目权重为0；类别2国家（参赛次数多但成绩不稳定 ）则综合考虑历史数据和特色项目的影响，权重通过**皮尔逊相关系数**计算；类别3国家（参赛次数少且成绩不稳定的国家 ）则只关注特色项目，历史数据权重为0。基于本模型，我们得出以 为代表的 个国家奖牌数会上升，以 为代表的 个国家奖牌数会下降。同时，模型认定奥运会项目的数量和类型会显著影响不同国家的奖牌数，以 为例，模型认为选择更多的特色项目会提高 的奖牌获奖率。

针对问题三,我们基于**随机森林**算法,构建了一个预测首次获奖国家的模型.~~以识别影响国家奖牌数转变的关键因素。~~对首次获得奖牌的国家数据集,我们选择“参赛人数”、“参赛次数”和“参赛项目数”三个特征变量，以是否获得奖牌（Will\_Earn\_Medal）作为目标变量，利用**随机森林回归模型**进行训练,并使用**GridSearchCV**选择最佳超参数，得到最优模型配置,计算出每个国家获得奖牌的概率.最后,利用**混淆矩阵**和**ROC曲线**评估了模型性能**,**发~~现其准确率高达94%，~~

~~AUC值为0.87,表明模型具有较强的区分能力和较低的误报率。~~经研究发现，体育基础设施完善的新兴市场国家具有较高的获奖概率，表明持续的体育投入和项目专业化对奖牌数量有显著影响。

针对问题四,我们通过**图论**和**网络流理论**,量化评估"伟大教练"对国家奖牌数的影响力.我们建立了教练与国家之间的有向网络图，将教练对奖牌数的影响转化为节点和边的流量关系,通过权重计算~~w=3\*△~~*~~G+2\*~~*~~△S+1\*△B~~,进一步通过总流量和瓶颈流量分析教练对奖牌的贡献,并为国家在教练选择上提供优化路径，帮助各国制定更加精准的体育发展战略。

针对问题五，我们在针对以上问题建模过程中，发现了一些有助于国家奥委会提升奖牌数的战略，例如：随女性参与度的提高，国家奥委会关注男女混合项目可以提高获奖机会；若作为东道国，可通过申请增加本国优势项目、优化基础设施和提升财政支持，借助主场优势提高奖牌数；对于未被垄断的项目，强国应分析这些项目，投资培养新一代运动员，而新兴体育国家可以通过精确的项目定位和优秀教练的引入，在未被垄断的项目中取得突破。

**keyword:**

# **introduction**

Faster, Higher, Stronger - Together.

## **1.1 background**

巴黎奥运会的召开吸引了全球的关注，各项赛事备受瞩目，尤其是各国选手的奖牌成绩，更是成为焦点。各国运动员在赛场上拼尽全力，力求在奖牌榜上占据一席之地。除了传统奥运强国以及东道主奖牌角逐备受关注，一些排名靠后的国家，例如阿尔巴尼亚、佛得角、多米尼克和圣卢西亚，在巴黎奥运会上赢得了历史上的首枚奖牌也引发了广泛讨论。但目前，仍有60多个国家未能收获奥运奖牌。

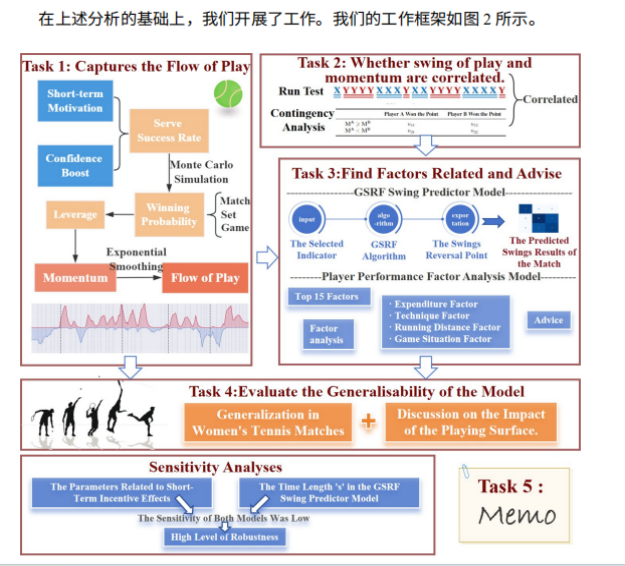
回顾历史，各国在奥运会中的奖牌表现呈现出一定规律。每届奥运会前，都会有“虚拟奖牌榜”来预测各国的表现。例如，在巴黎奥运会开幕前，尼尔森Gracenote就发布了2024年奥运会虚拟奖牌榜（VMT）最终预测。那么，这样的预测究竟依赖于哪些具体的因素呢？事实上，奖牌预测通常是在奥运会开幕临近时，结合已知的运动员参赛计划并分析过往的金牌数和总奖牌数，通过建立数学模型，从而对未来的奖牌排名进行预测。此类预测不仅对体育分析师、研究人员和政策制定者具有重要价值，也能帮助各国更好地把握影响奥运表现的趋势。

## **1.2 Restatement of the Problem**

鉴于问题的背景信息和限制条件，我们必须完成以下任务：

* 任务1：预测 2028 年美国洛杉矶夏季奥运会的奖牌榜,输出表现变好以及表现变差的国家
* 任务2：分析奥运会项目数量、类型与奖牌数之间的关系
* 任务3：对于尚未赢得奖牌的国家，预测他们在下一届奥运会上赢得首枚奖牌的概率，并提供这种预测的概率估计
* 任务4：考虑伟大教练效应对团队运动的影响，并估计其对奖牌结果的潜在影响
* 任务5：提供关于奥运奖牌数的原创性见解，并解释这些见解如何为奥委会的决策提供参考

## **1.3 Our Work**



总流程图替换

# **2.Assumptions and Justification**

为了简化问题，我们做了如下假设，每个假设都是合理的。

* **假设1**:历史奖牌数能够部分反应未来奖牌分布的规律
* **假设2**:

# **3.Notations**

|  |  |
| --- | --- |
| Symbols | Description |
| N | Number of countries |
| Y | Year of award |
| $M\_i$ | 国家$i$的奖牌总数 |
| $G\_i$ | 国家$i$的金牌数 |
| $S\_i$ | 国家$i$的银牌数 |
| $B\_i$ | 国家$i$的铜牌数 |
|  | 皮尔逊相关系数 |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

通过上述假设和符号，明确模型的边界和变量关系，为接下来的建模和求解提供基础。

# **4.Data Preprocessing**

## **4.1 Data Cleaning**

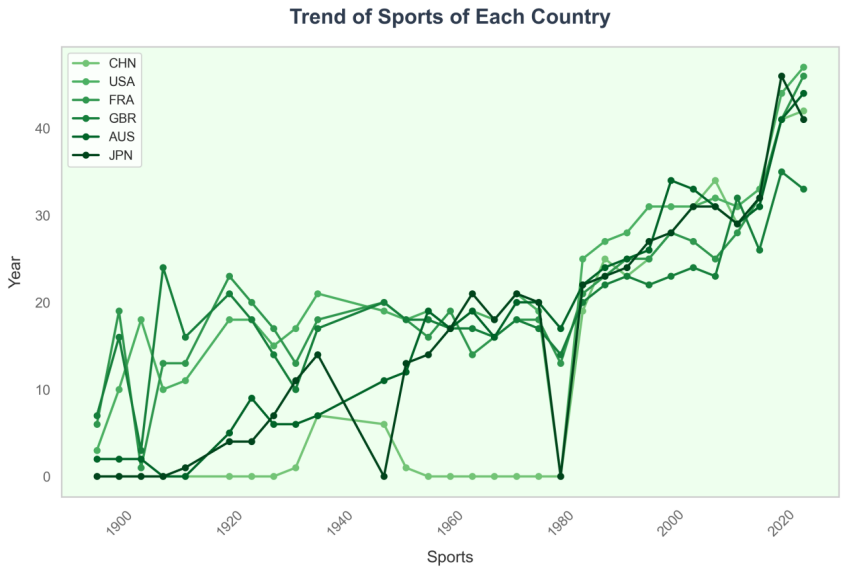
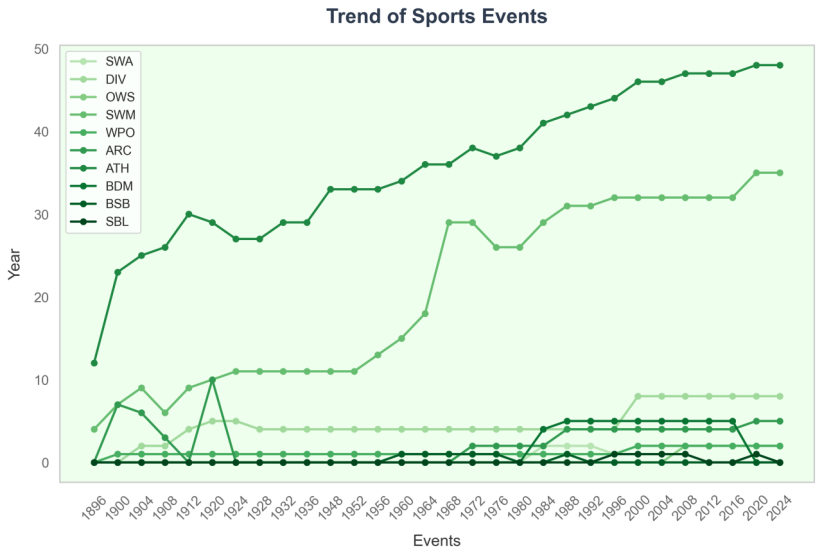
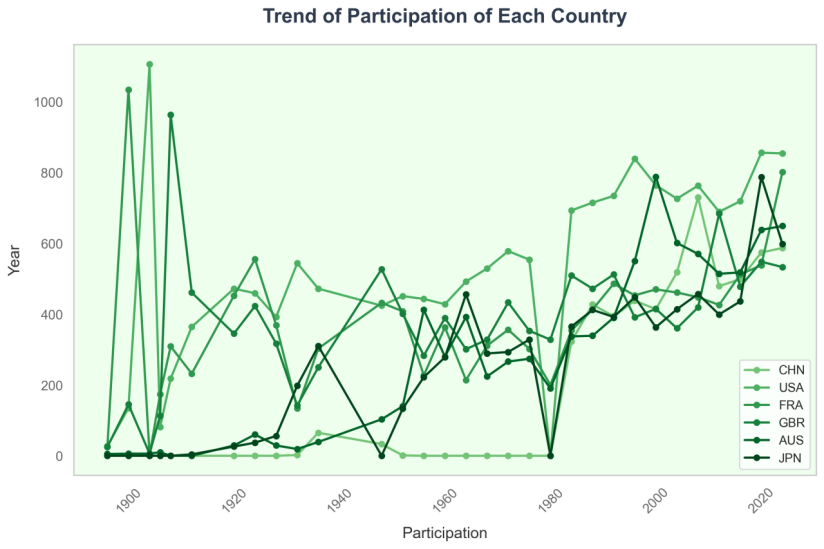
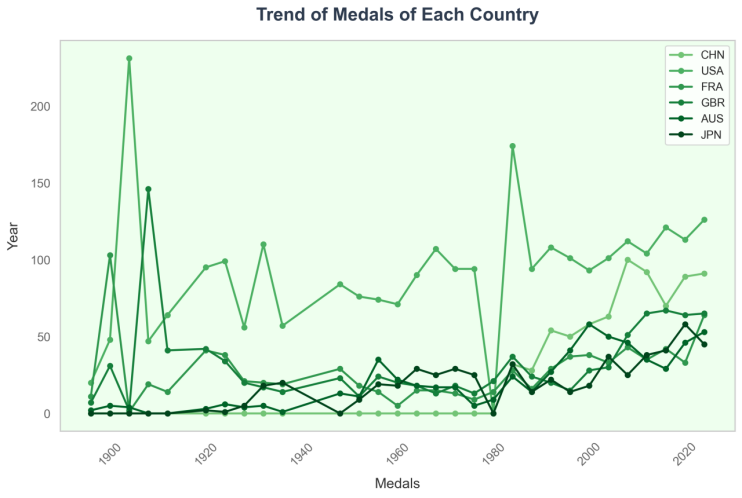
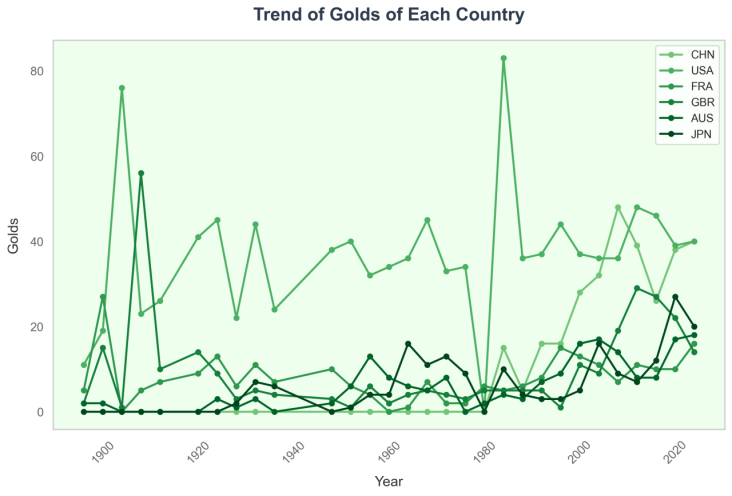
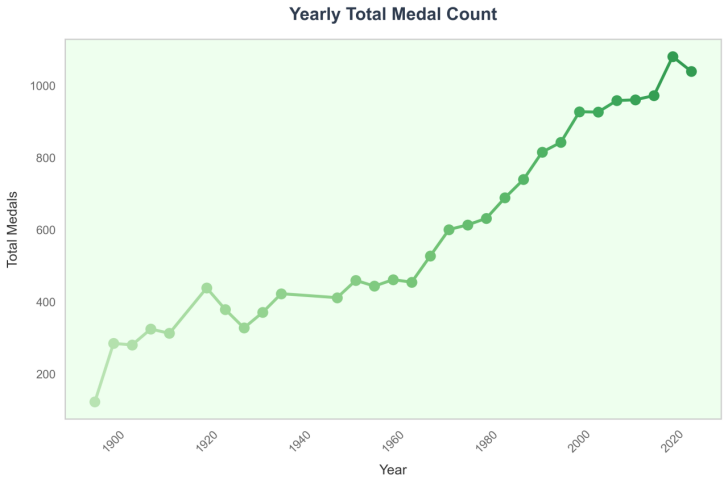
研究首先对原始数据进行了系统性的清洗和预处理。

在运动员数据处理方面，对包含运动员基本信息（姓名、团队、NOC、运动项目、具体项目、奖牌等）的summerOly\_athletes.csv数据集进行了删除重复条目、缺失值处理、标准化处理和数据类型转换等操作。此外,我们发现字符串文字中存在空格影响后续数据分析，使用str.strip()删除前导和尾随空格，以及使用str.replace()删除内部空格。对数据集summerOly\_programs.csv进行无效信息删除和反转处理，方便后续使用。

在NOC（国家奥委会）验证环节，我们以athletes数据集作为有效NOC的主要参考源，将其与hosts和medals数据集进行了交叉验证。为确保数据的一致性，我们删除了hosts和medals数据集中不存在于athletes数据集的NOC条目。这一步骤为后续的数据分析奠定了可靠的基础。

为确保数据质量，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，验证年份范围和时间连续性，以及检查诸如金牌数不超过总奖牌数等逻辑约束。在数据集合并过程中，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补充转换，创建了历史表现的聚合视图，统一了各数据集的数据格式，为后续的分析和建模准备了规范的数据结构。通过这一系列严谨的数据清洗过程，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。这些处理步骤不仅提高了数据质量，也增强了后续分析结果的可靠性。



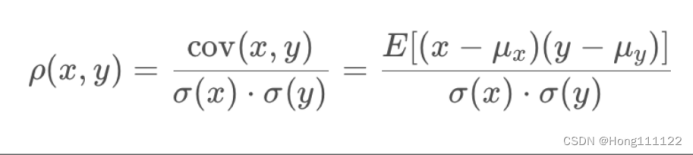
# **5.TASK1&2:基于LSTM的时间序列模型**

## **5.1 Data Analysis**

首先，我们对曾经获得过奖牌的国家构造特征相关性矩阵，得出部分国家的特色项目（即在单项比赛项目中具有垄断地位或主要依靠某些项目拿奖）。然后为每一个国家构造联表，其中包含获奖年份、总奖牌数、金牌数、是否为东道主、参赛人数、参加项目的数量、主办方设立的项目总数、历届项目。最后用K-Means把曾经获得过奖牌的国家划分为三类，用以后面给不同类型的奥运国家分别建立预测模型。

### **5.1.1基于Pearson correlation的相关性矩阵分析**

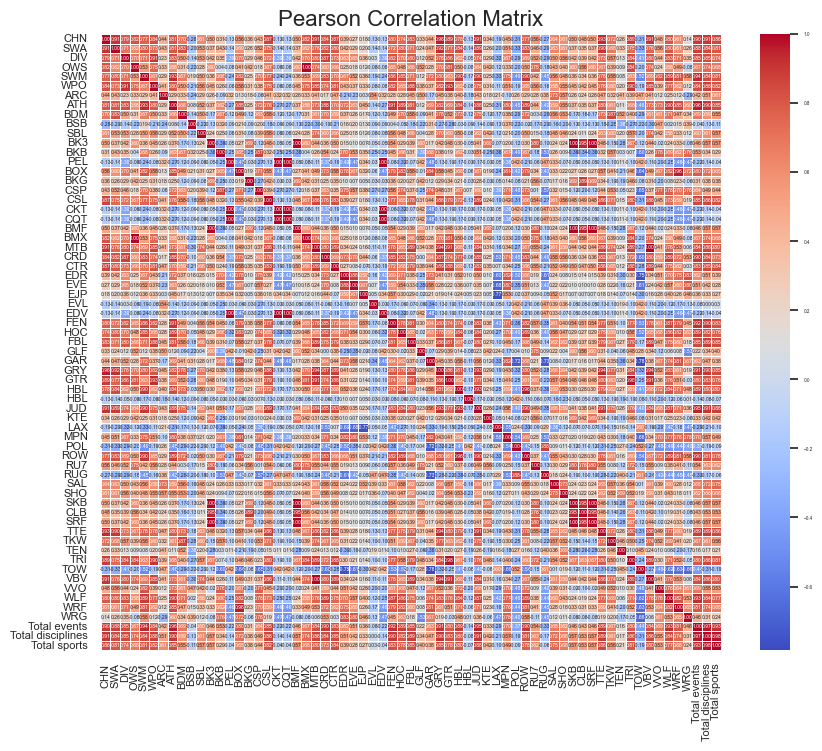
* **Step1数据准备**：将储存奖牌数的数据集与项目数据集进行数据对齐，确保每一行代表一个国家在某一项目的奖牌数量。
* **Step2 Pearson correlation：**皮尔逊相关系数适用于对连续正态分布的数据。计算公式如下：

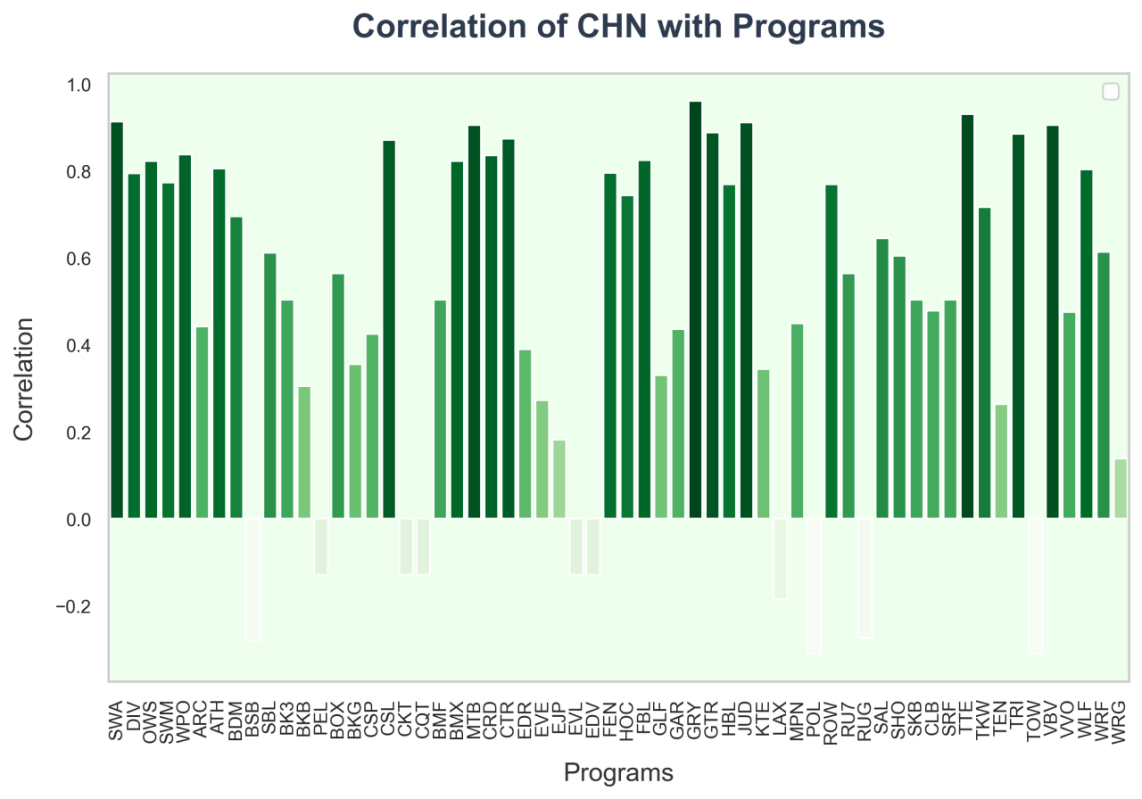


* **Step3计算相关性**：对于每一个国家，我们将各个项目中该国家获得的奖牌数和奥运会设立的奖牌数相结合，计算皮尔逊相关系数。皮尔逊相关系数可以量化奖牌数。变量相关强度如图【】所示，我们对生成的相关系数矩阵进行可视化操作，以中国为例，生成热力图如图【】所示



图：变量相关强度





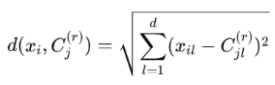
* **Step4分析特色项目**：返回一个包含前三个最大相关性的自变量机器相关系数的系列，得到最能影响该国家奖牌总数的三个项目作为该国家的特色项目。以中国为例，从返回结果和上图可以看出中国的特色项目为体操、乒乓、跳水。
* **Step5构造联表**：为每一个国家构造联表，其中包含获奖年份、是否为东道主、奖牌数、金牌数、主办方设立的项目总数、参加项目的数量、特色项目、平均得奖

### **5.1.2基于K-Means 的获奖国家分类模型（kmeans步骤简化）**IMG_267

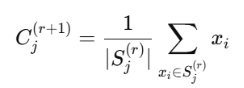
在曾获奖国家的数据集下，我们在为每一个国家建立奖牌预测模型的过程中，需要首先对参赛代表队进行分类，而在按照各国代表队获奖稳定性分类之前，需要先定义稳定性。对此，我们决定根据NOC和Sports属性定义稳定性划分，将相似国家归类进行统一预测模型的搭建。

基于距离的聚类通常使用 K-means、K-medoids 等测量方法。其中，目前最流行的启发式方法是 K-means 算法，**K-means** 的核心是通过迭代优化聚类的质量，直到收敛。因此，我们对获得过金牌的atheletes\_with\_gola\_medal数据集，利用NOC和Sports属性作为聚类基础，运行K-means聚类模型，K-means 算法的一般实现步骤如下：

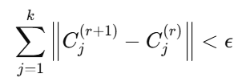
* 步骤 1：给定一个大小为n的样本数据集{x1,x2,...,xn}，设迭代次数为R，确定聚类最终簇数为 k ，并随机选择 K 个数据点作为初始聚类中心，标记C1(0)[上标],C2(0),...,Ck(0)​[肘部法则/轮廓系数][为什么k=3]
* 步骤 2：计算每个数据点与 k 个初始聚类中心之间的欧氏距离，并根据最小距离进行聚类划分，最终形成 k 个区域。



* 步骤 3：对每个簇中的所有数据点，计算其均值作为新的聚类中心，并将其作为下一次迭代的聚类中心。其中，Sj（r）表示第 j 个簇中所有数据点的集合，|Sj（r）|是该簇的大小。

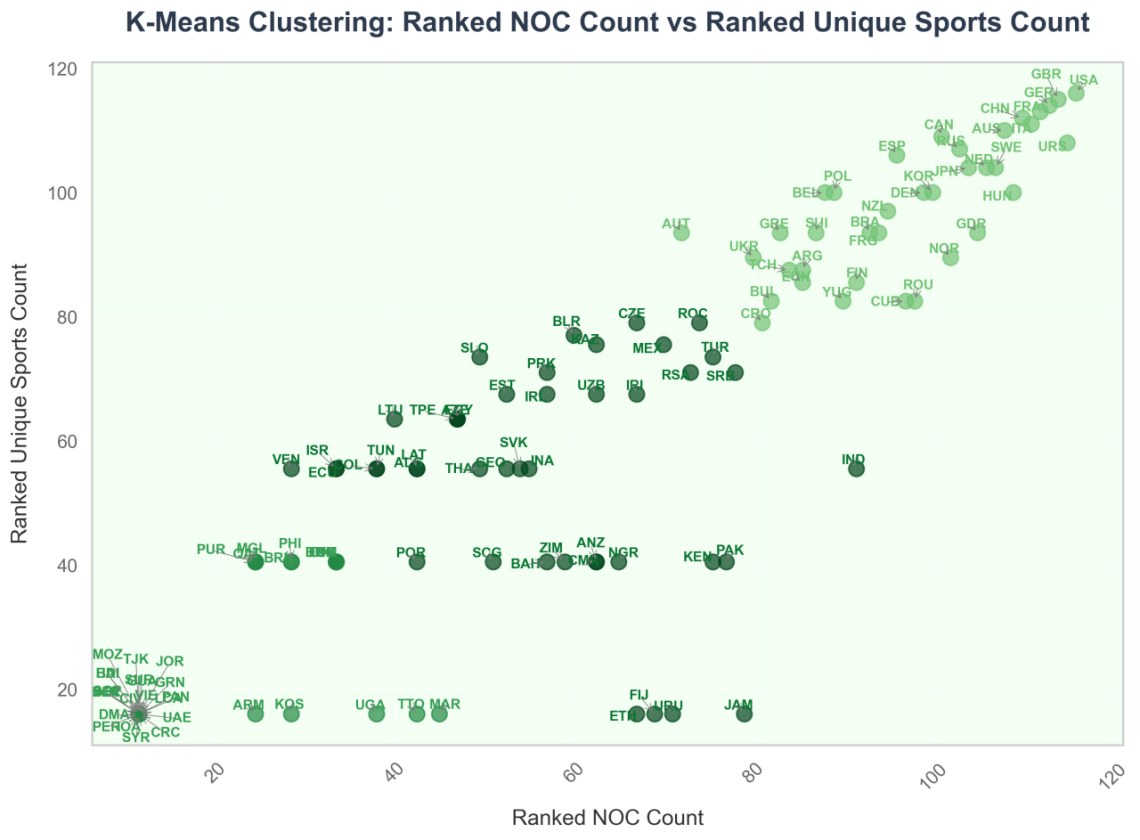


* 步骤 4：重复上述步骤，直到最后两次聚类结果之间的变化小于预设阈值ϵ。



首先，我们计算出每个NOC在数据集atheletes\_with\_gola\_medal中出现的次数，并计算每个NOC对应的不同Sport数量，分别使用MinMaxScaler对数据进行归一化，将NOC和Sport数据合并成一个数据集进行 KMeans 聚类。

在聚类过程中，发现美国队的数据点表现出显著的离群特征。这主要体现在其在NOC出现频次和运动项目数量两个维度上都远超其他国家，形成了独特的数据分布。这种特殊性反映了美国在奥运会上的全面优势，同时也为聚类分析带来了挑战。为确保聚类结果的合理性，需要对聚类方法进行适当的修改。



因此，我们决定以NOC与Sport出现次数的排名作为聚类基础进行不同国家的聚类划分。此外，在模型选择阶段，我们尝试了Kmeans与DBSCAN对atheletes\_with\_meda和atheletes\_with\_gola\_medal数据集进行聚类，结果证明Kmeans模型下对atheletes\_with\_gola\_medal数据集进行聚类的效果最优。

根据图1，得到我们最终的聚类结果：

**聚类 1 包含 XXX 个曾获金牌的国家，聚类 2 包含 XXX 个国家，聚类 3 包含 XXX 个国家。**

聚类1内国家参加比赛次数较多且参加项目种类多，属于老牌奥运强队，稳定性较高，对特色项目与强势项目的依赖性小，可以大致从历届获奖数据对奖牌数进行预测。

聚类 2 内国家参赛次数居中，可能由于政治、战争原因间隔多年未参赛，或是连续参加但成绩不稳定的国家，类2较不稳定，需要考虑多种因素对奖牌预测的影响。

聚类3内国家参加比赛次数较多且参加项目种类多，属于老牌奥运强队，稳定性较高，可以大致从历届获奖数据对奖牌数进行预测。

|  |  |
| --- | --- |
| 表 聚类结果 | |
| 群组 | 特征描述 |
| Cluster 1 (奥运强国) | 高参赛频次 多样化的项目分布 稳定的奖牌获得能力 |
| Cluster 2 (发展中体育国家) | 中等参赛频次 项目分布相对集中 奖牌获得呈波动性 |
| Cluster 3 (XXX国家) | XXX |

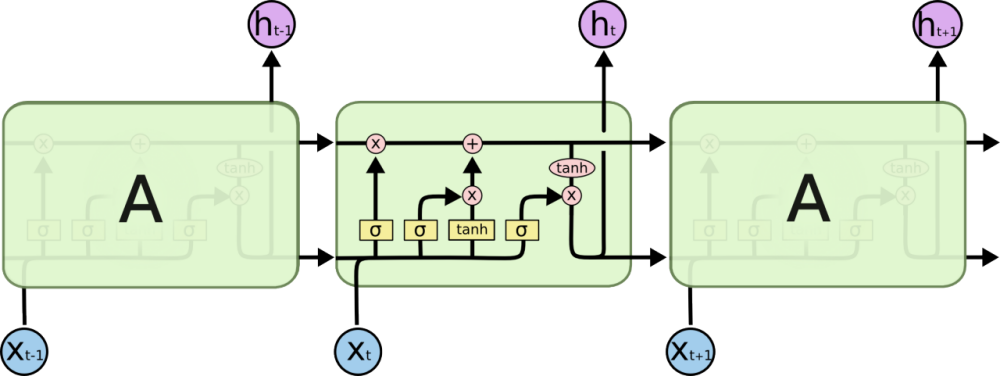
## **4.1 Model Selection**

经过【4.2】的数据分析，我们已经把数据集分为Cluster\_0，Cluseter\_1，Cluster\_2，可以对各国的奖牌数进行预测了。考虑到我们的处理的数据每个数据点都与一个特定的时间点相关联，并且这些数据点按照均匀的时间间隔（4年）进行测量和记录，我们决定选用时间序列来预测奖牌数。

传统经典时间序列预测方法，如最常用的Auto-Regressive Integrated oving Averages（ARIMA），但是ARIMA模型只依赖于一个单一的时间序列数据源，然而，通常奖牌数预测不只基于历史奖牌数这一个变量。随着机器学习和人工智能的发展，大量深度学习算法被用于时间序列的预测，如**Long and Short Term Memory Network（LSTM）**可以解决RNN不能捕获较长时间的周期季节等信息。在LSTM的每个时间步里面，都有一个记忆cell，这个东西给予了LSTM选择记忆功能，使得LSTM有能力自由选择每个时间步里面记忆的内容。与此同时，LSTM可以用于多变量分析，LSTM的多变量分析能力与基于K-Means聚类的三类国家分组相契合，能够针对不同类型的国家采用差异化的变量权重策略。

## **4.2Long and Short Term Memory Network（理论解释简化）**

~~Long and Short Term Memory Network（LSTM）是一种特殊的RNN，能够学习依赖关系，通过在隐层的每个神经单元中增加记忆单元，解决了RNN在长序列训练过程中的梯度爆炸和梯度消失的问题。【引用文献blog】~~

~~~~

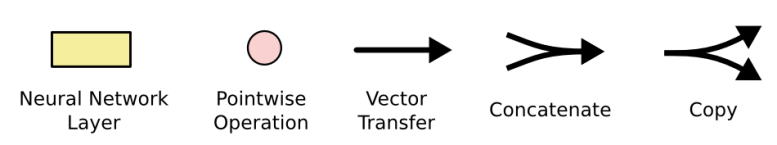
~~~~

图 ：LSTM工作机制流程图【需要重新画可能】

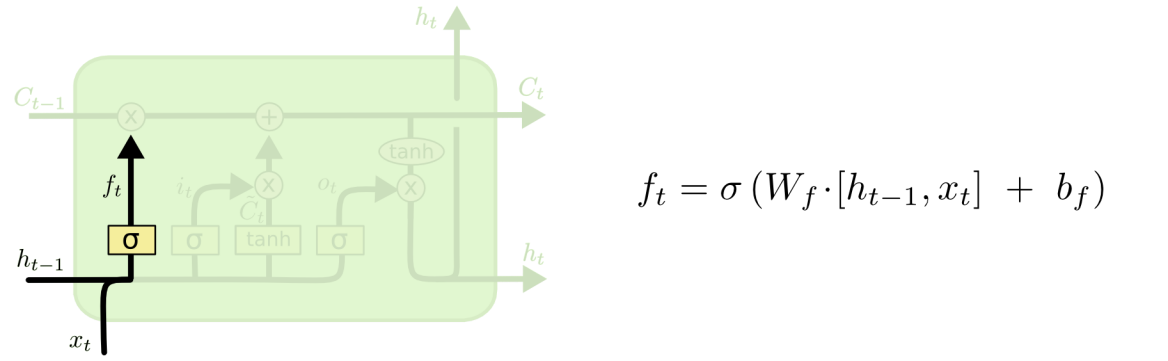
~~记忆单元的核心是门控机制，包括三个门：遗忘门、输入门、输出门。这些门通过Sigmoid函数输出介于0和1之间的数字来控制信息的流动，即控制时间序列的记忆信息，下面我将介绍各个门的作用。【引用文献An ideal point method for the design of compromise experiments to simultaneously estimate the parameters rival mathematical models.】~~

研究采用LSTM网络作为核心预测模型，其结构包含三个关键组件，分别是Forget Gate、Input Gate和Output Gate。遗忘门控制历史信息的保留程度，输入门管理新信息的更新，输出门决定信息的输出，三者共同构成完整的记忆机制：

其中，表示t时刻的隐藏状态，为输入向量。

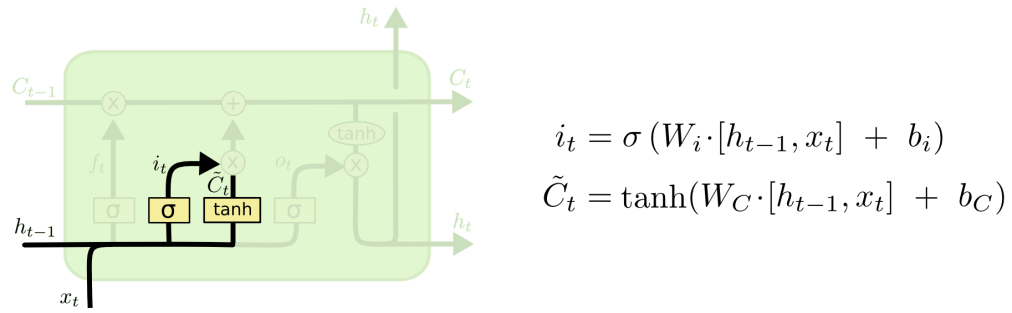
* Forget Gate

输入数据$x\_i$和上一个时间步的隐藏状态$h\_{t-1}$被送入LSTM。遗忘门$f\_t$用来选择哪些信息从细胞状态$C\_{t-1}$中丢弃。



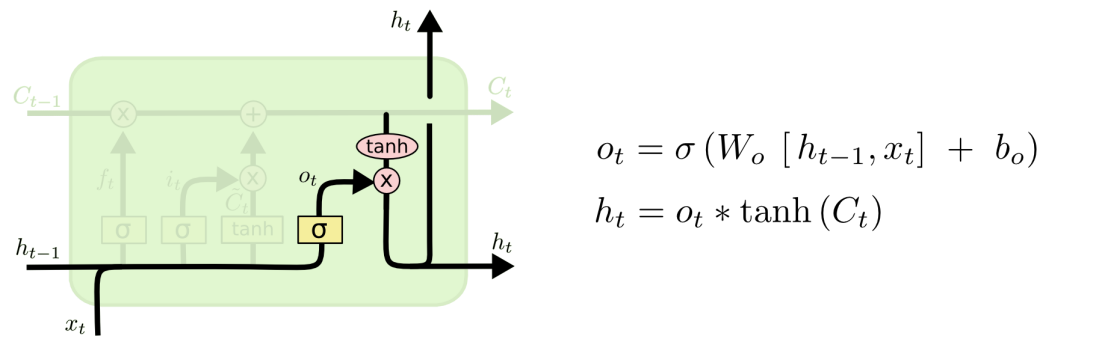
* Input Gate

输入门$i\_t$决定哪些新信息被存储到细胞状态$C\_t$中。这里分为两部分，第一部分是输入层，通过sigmoid函数确定那些值需要更新，第二部分是tanh层，生成新的候选细胞状态$【latex中重新打】$，把这两者结合起来，得到这一部分的细胞状态，并摒弃不需要的值。



* Output Gate

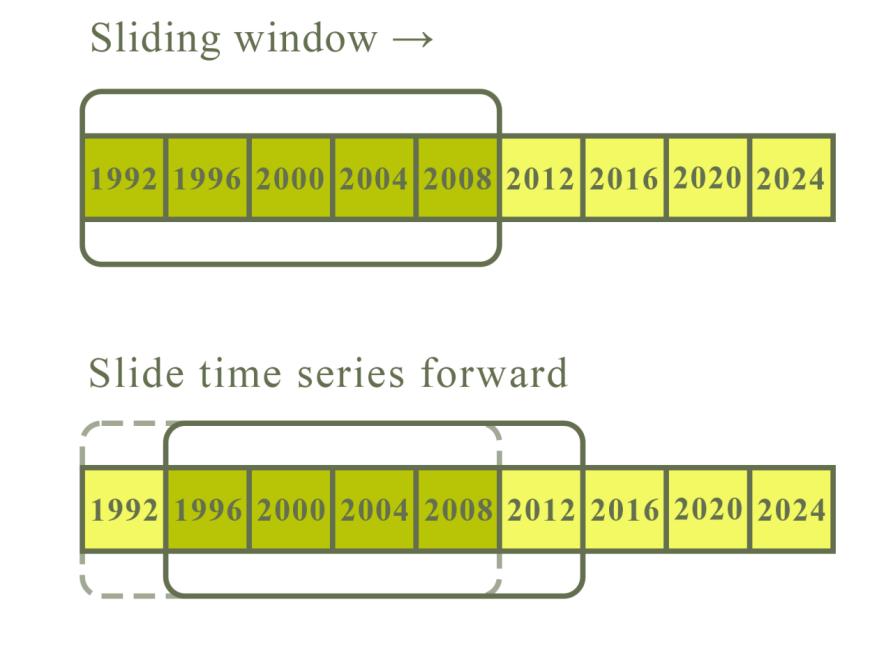
输出门$o\_t$决定哪些信息从细胞状态$C\_t$中输出。先通过tanh函数将数值缩放至-1和1之间，再将输出与sigmoid函数相乘，得到模型最终输出$h\_t$。隐藏状态和细胞状态被传递到下一个时间步。



## **4.3Implementation of LSTM**

4.3.1标准化处理

我们对数据进行了归一化处理，减少不同特征之间尺度的影响，保证训练过程的稳定性，然后通过**滑动窗口**创建输入数据集，将原始的奖牌数据按时间窗口切割，将time step设为20，即每一年的奖牌数都与它前20年的数据有关系。



4.3.2 量化东道国影响力

我们用最小二乘法，以总奖牌作为因变量，东道国、项目作为自变量，拟合得到R^2为0.758，调整后的R^2为0.756都较高，表示模型可以解释越75.8%的总奖牌的变化，说明模型拟合较好；自变量东道主的p值小于0.05，显著性较高，具有统计意义，对总奖牌有较大影响。结果如表所示

表 最小二乘法结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | coef | std err | t | P>|t| |
| const | 10.4972 | 5.090 | 2.062 | 0.041 |
| host | 26.9574 | 4.975 | 5.419 | 0.000 |
| sport | 1.3783 | 0.226 | 6.103 | 0.000 |

4.3.3构建LSTM模型

在上文中提到，我们用K-Means聚类把获得过奖牌的国家分成三类，分别是

* 类别1：参赛次数多、参赛项目多的老牌奥运强国，如美国、中国等国家；
* 类别2：参赛次数多、成绩不稳定的国家，如卢森堡、新加坡等国家；
* 类别3：参赛次数少、成绩不稳定的国家，如印度、牙买加等国家。

权重w的获取的依据是 计算参赛人数、参赛项目数的对于原点欧式距离排序后归一化(+公式 Distance = p^2 + s^2)

我们给每个国家构造一个时间序列，包含该国历年的奖牌数、金牌数、东道主、参赛人数、参赛项目、主办方设立的项目数和历届各项目设立奖项的情况。我们将奖牌数、金牌数、东道主、参赛人数、参赛项目、主办方设立的项目数视为该国获奖的历史依据，历届各项目设立奖项的情况\*皮尔逊相关系数作为该国的特色项目情况，对于类别1，这类国家通常实力较强劲，在多种竞赛项目上获过奖，不依赖于特色项目。因此，对于这类国家，由于东道国对奖牌数的影响是巨大的，所以将其数值乘以最小二乘法回归出的影响力，其他历史依据权重均分的方式，而将特色项目的权重置为0；对于类别2，这类国家的奖牌预测既依赖于往届的参赛情况和获奖情况，又依赖于某些特色项目。因此，对于这类国家，我们把所有变量都纳入，并且$weight=（1-w）\*皮尔逊相关系数$；对于类别3，这类国家参赛次数少，在历史时间内获奖次数非常少，可以近似于获得预测获奖的概率，参考往届的参赛情况没有意义，所以仅仅只关注特色项目的情况，历史的数据的权重设为0。

在模型架构设计中，我们采用了一个包含50个记忆单元的LSTM层作为主要特征提取器，通过这些记忆单元可以有效捕获奥运周期内的长短期依赖关系。LSTM层之后连接一个全连接层进行特征整合，最终通过包含两个神经元的输出层分别预测总奖牌数和金牌数。考虑到近期奥运会的表现对未来预测具有更强的指示性，我们在模型中引入了时间衰减权重机制，使模型在保留历史信息的同时，能够更多地关注近期数据的变化趋势.

## **4.4Prediction Interval from the Result**

我们通过对每个国家的模型的多次训练和测试来进行预测区间的计算，我们将每个国家的模型分别训练和预测10次，来得到最终结果的分布。去除结果中的离群点后，获取最大最小的数值来作为获奖的范围或获奖概率的范围，并计算平均的损失来作为一个国家模型的损失，再通过每个国家损失求平均值，来计算整个模型的性能。

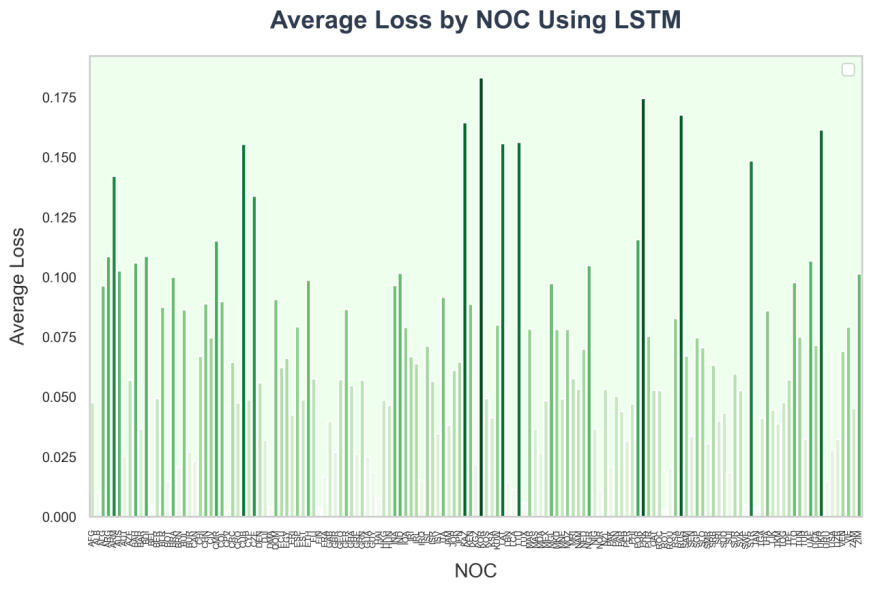
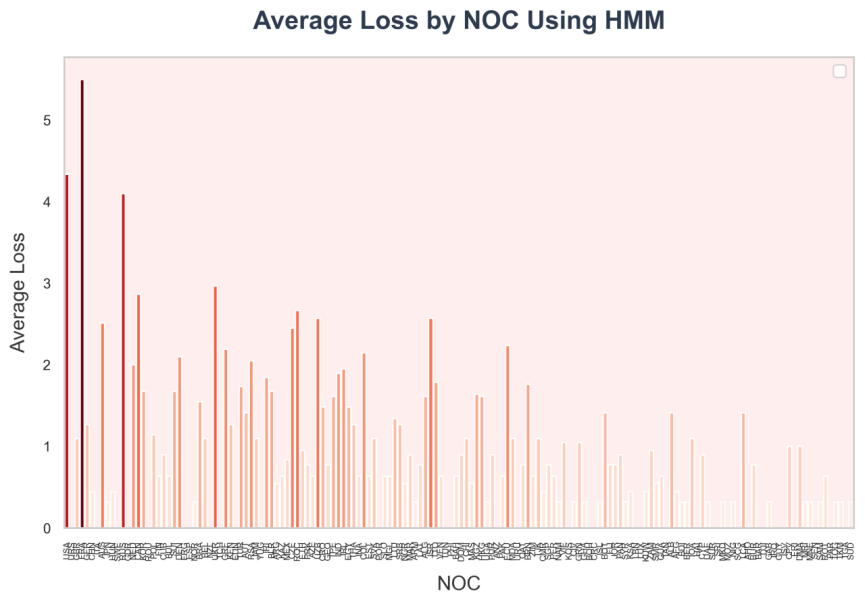
（公式（单个LSTM的损失公式求平均）再全部国家求平均）

用训练好的模型对三类国家（分别以美国、 、为例）进行预测，将归一化的预测结果反标准化回原始值，得到结果如图【】所示

从上图可以看出，以美国为代表的 个国家奖牌数会上升，以 为代表的 个国家奖牌数会减少；以 为代表的 个国家金牌数会上升，以 为代表的 个国家金牌数会下降。

## **4.5the Advantage of LSTM**

LSTM模型对最终奖牌数的预测通常不仅仅基于历史奖牌数的时间序列而是多元因素的综合考虑，能够分析出包含历史因素和特色项目的影响。为了评估LSTM模型在奥运奖牌预测任务上的优势，我们将其与传统的隐马尔科夫模型HMM)进行了对比实验。通过比较两个模型在测试集上的预测结果和实际值的均方误差(MSE)，我们观察到LSTM模型展现出显著优势，其预测误差明显低于HMM模型，如图【】所示。



从图中可以看出，与传统的隐马尔科夫模型相比，LSTM在奖牌预测的问题上表现与传统机器学习模型相比十分突出，我们分析了以下几点原因:

1. **模型结构**：LSTM的门控机制(包括遗忘门、输入门和输出门)能够自适应地调节历史信息的保留程度，这使得模型可以捕捉奥运强国奖牌数长期稳定增长的趋势特征。而HMM受限于其马尔可夫假设，仅能建模相邻时间步之间的依赖关系，难以刻画长期演化规律,并且无法准确从特色项目的角度分析出一些获奖数极少的国家再次获奖的概率和波动较大或依赖特色项目的国家的获奖情况。从图【loss对比图】中可以清晰地观察到，LSTM的预测曲线与实际值的拟合度明显高于HMM模型。
2. **特征表达能力：**LSTM通过其记忆单元(memory cell)可以同时整合多个维度的信息，包括历史奖牌数、参赛项目、运动员规模等关键因素。这种多维度特征的动态融合机制，使得模型能够更准确地预测如东道主效应、新兴体育强国崛起等复杂现象。相比之下，HMM的状态空间表达能力相对有限。
3. **优化目标：**LSTM采用端到端的梯度下降训练方式，可以直接优化预测误差。而HMM基于最大似然估计的参数学习方法，与实际预测任务存在一定的目标偏差。这也是LSTM在实验中展现出更强鲁棒性的重要原因。

# **6.TASK3：基于随机森林的首次获奖国家预测**

## **6.1随机森林模型**

要确定影响某国家奖牌数转变的关键指标，我们需要完成两个步骤：

* 首先，开发一个预测模型，预测奖牌数从0到1的转折点。
* 其次，根据模型结果推断关键指标。

随机森林对非线性关系和特征间交互作用的建模能力较强，这与奥运会获奖的复杂影响因素相匹配。

## **6.2数据准备与预处理**

### **6.2.1 数据表构造**

构造participation\_by\_year\_country、participation\_by\_year\_country\_count、sport\_count\_pivot三张数据表，分别记录国家每年的参赛人数、按年累计的参赛次数和每年参赛的项目数。

从 participation\_by\_year\_country 中删除包含在 medal\_counts 中 NOC 列的国家数据，从而去掉那些没有获得奖牌的国家的数据。

在medals\_by\_year和medals\_by\_year\_gold数据集中遍历每个国家的历史数据，通过索引获得第一次获得奖牌的年份，若某个列的所有值都为0，则返回 None，最终结果存入first\_medal\_filtered中预览，该数据集仅包含那些首次获得奖牌或金牌年份大于等于1920年的国家数据。

### **6.2.2 数据集建立**

第一次遍历：根据每个国家的NOC，从participation\_by\_year\_country\_count, participation\_by\_year\_country, 和 sport\_count\_pivot 数据框中获取对应的信息：

·participation\_count：获取该国家在首次获奖年份的参与次数。

·events\_participation：获取该国家在首次获奖年份的赛事参与次数。

·sport\_count：获取该国家在首次获奖年份的运动种类数。

提取后将这些值添加入new\_table\_data中，作为后续数据集使用。

第二次遍历：为了减去首次获得奖牌年份后的4年数据，得到 first\_medal\_year 之前4年的数据。

第三次遍历：对每一个没有获得过奖牌的国家进行遍历，将每个国家（noc）、articipation\_count、events\_participation、sport\_count以列表的形式添加到 new\_table\_data 中。

最后，将前面存储的所有数据列表 new\_table\_data 转换为 Pandas DataFrame，在缺失值填充和无效数据清理后，形成最终表格tree\_dataset。

## **6.3模型搭建**

### **6.3.1 变量分析**

经过对现有有限数据的仔细筛选和考量，我们选择了3个特征变量来训练模型，变量的详细列表如下：

|  |  |
| --- | --- |
| 指标 | 说明 |
| participation\_count | 国家的参赛人数 |
| Participation\_count | 国家的参赛次数 |
| sport\_count | 国家每届参赛的项目数 |

同时，我们的目标变量为Will\_Earn\_Medal，表示是否会获得奖牌（1 或 0）

### **6.3.2 模型训练**

对数据进行StandardScaler标准化后，将其拆分为训练集和测试集，训练集占80%，测试集占20%。

我们利用上述指标作为随机森林模型的输入参数，对训练集进行模型训练。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| n 个估计值 | 最大深度 | 最小样本叶片 | 最小样本分割 | 最大特征数 |
| [50,100,150, 200] | [None,10,20, 30] | [2, 5, 10] | [1, 2, 4] | ['auto','sqrt', 'log2'] |

我们使用了scikit-learn (sklearn)机器学习库集合模块的 RandomForestRegressor 算法和 sklearn 的 GridSearchCV 算法GridSearchCV 得出的最优超参数组合如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| n 个估计值 | 最大深度 | 最小样本叶片 | 最小样本分割 | 最大特征数 |
| 200 | 10 | 1 | 2 | sqrt |

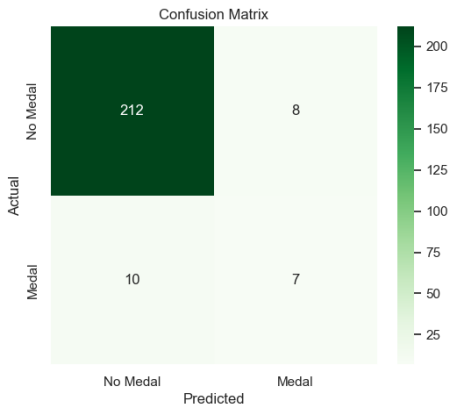
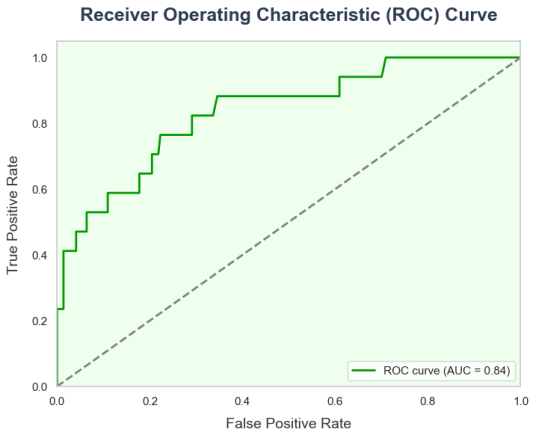
### **6.3.3 模型效果可视化**

针对模型的效果，我们通过混淆矩阵和ROC曲线对模型性能进行一个初步评估。

**混淆矩阵：**显示模型预测结果与真实标签的关系，帮助分析模型的错误类型。

**ROC 曲线**：绘制 Receiver Operating Characteristic (ROC) 曲线，衡量模型在不同阈值下的表现。

混淆矩阵图和ROC曲线如下【】所示：



混淆矩阵显示,模型的预测准确率高94%，false positive rate仅为0.039(9/233),表明模型预测具有较高可靠性,大部分样本都被正确归类。测试集预测值与真实标签的匹配度较高。

模型的AUC为0.87，证明模型有较强的区分能力，能够较好地将正负类别区分开来。并且，ROC曲线在低FPR区间快速上升,模型在保持低误报率的同时实现了较高的召回率,，模型的真正率表现较好，属于一个良好的模型。

### **6.3.4 模型预测**

用训练好的最佳模型 best\_rf\_model 对标准化后的新数据进行预测。predict\_proba()方法返回每个样本属于每个类别的概率，从而得出每个国家获得奖牌的概率。

## **6.4模型测试**

### **6.4.1 测试集构造**

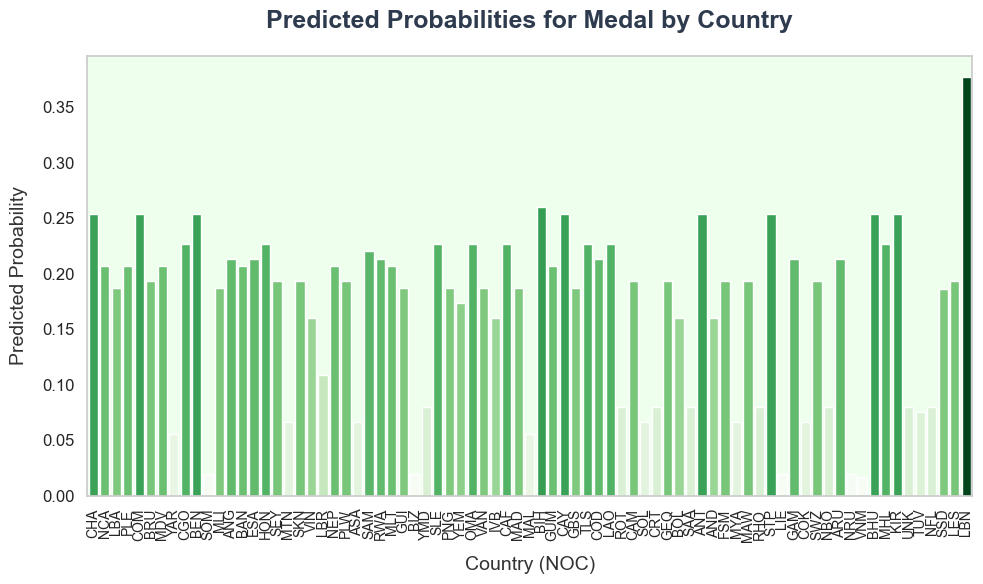
为验证模型在实际预测任务中的表现,我们从athletes数据集中识别并筛选出历史上从未获得奖牌的国家样本,采用时序分组方法确保数据的时间连续性和完整性,接着通过对特征数据的标准化处理和质量控制,最终形成了高质量的测试数据集。

最终输出一个新的数据表，包含所有没有获得奖牌的国家在2024年的相关统计数据。

### **6.4.2 测试**

使用训练好的随机森林模型 best\_rf\_model 对每个国家的特征进行预测，计算每个国家获得奖牌的概率存储入列表中，对预测概率进行排序，找出最可能获得奖牌的国家，绘制出柱状图，展示每个国家获得奖牌的预测概率。

通过概率分布分析发现,预测结果呈现明显的分层特征。预测概率超过0.20的国家主要集中在体育基础设施完善的新兴市场国家,这些国家普遍表现出持续的体育投入和项目专业化特征。



# **TASK4:基于网络流的伟大教练影响力量化图论模型**

## **7.1模型背景**

~~为量化评价伟大教练对各个国家获奖的影响，我们可以通过图论和网络流的理论，用不同国家节点之间的流量传输情况表示教练对国家获奖的影响情况，由此我们可以通过对每条边总流量和最大流的量化评价，进而分析出伟大教练的影响力及国家在教练选择策略上的优化路径。~~

## **7.2模型构建**

#### **7.2.1构建网络**

基于图论和网络流的理论，针对每个教练，我们建立一个有向网络图表示教练在各个国家的执教流动情况。

* 节点：网络中的各个点。模型中体现为国家。
* 边：连接两个节点之间的通道，每条边具有对应的流量。模型中体现为某一段时间内能够获得的加权奖牌总数，这个流量反应项目本身的潜力和教练的贡献

本模型的网络结构由以下要素构成：

节点集合V：表示参与国家，，其中I为国家索引集

边集合E：表示教练流动路径，

权重函数W：E→R，定义为，反映奖牌增益效应

#### IMG_283IMG_284

#### **7.2.2权重计算**

对于留教教练（以排球教练郎平为例）：

当教练的流出国家与流入国家相同（即教练在同一国家留任）时，在该国节点上绘制自环边表示，如郎平2004年与2008年均于美国执教，随后根据设定权重【优化】计算美国女子排球奖牌数的奖牌数增减，做差，记为得分。

对于转队教练（以排球教练郎平为例）：

当教练的流出国家与流入国家不同时，绘制单项边表示教练的转队。郎平2014于美国执教，随后在下一届奥运会中作为中国教练执教，我们根据设定权重【优化】计算中国2016的奖牌数与美国2008的奖牌数，然后对中国的正向增长和对美国的逆向增长做差，记为得分。

w=3\*▲G+2\*▲DS+1\*▲DB

最后绘制如下的流量分析表如下【】。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 运动 | 教练 | 年份 | 奖牌数量 | 奖牌变化 | 得分 |
| 排球 | 郎平 | 2004→2008（美国自环边）： | 金牌  银牌+1  铜牌 | +0  +1  +0 | 2 |
| 2012→2016（美国→中国） | 金牌+1  银牌+1  铜牌 | USA+0 CHN+1  USA-1 CHN+0  USA+1 CHN+0 | 4 |

#### **7.2.3 流量计算**

**总流量：**总流量表示节点之间的最大流量，即总奖牌数，反应教练和其他因素对项目总奖牌数的贡献。

**瓶颈流量：**每条路径上的最小流量，瓶颈流量表示在教练作用下，其他因素（如运动员素质、赛事技术要求等）对奖牌数的限制。瓶颈流量低意味着项目的提升潜力受限，而瓶颈流量高则表示教练对该项目的贡献很大。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 运动 | 教练 | 年份 | 奖牌数量 | 奖牌变化 | 总流量 | 最大流量(预测流量) |
|  |  | 2004→2008（美国自环边）： | 金牌  银牌+1  铜牌 |  |  |  |
| 2012→2016（美国→中国） | 金牌+1  银牌+1  铜牌 |  |  |  |

## **7.2模型结果**

问1：在该模型中，我们仅考虑教练在不同国家之间的流通对该国的奖牌指数（w=3\*DG+2\*DS+1\*DB），通过分析奖牌指数我们可以得到该教练对该国该项目的影响关系。

问2：基于上述对总流量和瓶颈流量的解释，我们利用瓶颈流量和总流量的计算大小评估“伟大教练”效应对奖牌数的影响.

1)瓶颈流量大，总流量大：教练影响力强，奖牌数显著增加，项目潜力较大。

2)瓶颈流量大，总流量小：教练影响有限，外部因素限制较多，奖牌数提升空间有限。

3)瓶颈流量小，总流量大：教练影响力大，奖牌数稳定增加，项目有较强的上升潜力。

4)瓶颈流量小，总流量小：教练作用有限，其他因素并不限制，整体成绩提升空间较小。

问3：

# **TASK5：**

1. 【描述性分析】历史上总体男女运动员的比例为男：女=7:3。

【画一个扇形图和一个历年男女运动员数量和比例变化趋势】

从上图中可以看出早期基本都是男性参加，直到2012年这个比例才接近于1。这一现象可以为各国提高表现提供一下思路：

* + **国家奥委会**（country Olympic Committees）**可以通过增加性别平等的项目来提高表现**：随着女性运动员在奥运会中的参与度逐渐提高，出现了越来越多的男女混合项目和1:1设置的项目，某些国家可以通过更广泛的性别平等参与，提升整体的奖牌收获
  + **不同项目设置的性别差异对奖牌总数的影响**：在一些历史上设立时间较晚的女性项目，虽然近年来有所增加，但从小项数量来看，女性项目仍处于相对不利地位，这种性别不平衡的设置可能会影响某些国家在特色项目中的奖牌表现。国家奥委会应注意如何调整备战策略，最大化在不同性别项目上的优势。

1. 我们在观察每个国家历年获得奖牌数的时候发现，当该国作为东道国时，对奖牌数的提升有很大影响。于是在上文，我们用最小二乘法验证了东道国对国家奖牌数的影响。而这种影响主要来由于运动员旅行的疲劳、对场地的熟悉程度以及主场观众对运动员的影响。因此作为东道国的国家可以通过以下方式提升奖牌数：
   * **可以根据历史经验和主场优势制定更加周密的备战计划**：东道国可以与奥委会协商，增加本国优势项目，减少弱势项目，提高本国奖牌数。例如2028年洛杉矶奥运会，美国奥委会增加了棒球/垒球（Baseball/Softball ）、板球（Flag Football）、腰旗橄榄球（Cricket(Twenty20)）、棍网球（Lacrosse (Sixes) ）以及壁球（ Squash）五个竞赛项目。其中的棒球/垒球虽然在全球范围内有大量爱好者，但是作为美国的“国球”，在美国本土拥有广泛的参与和观众基础。而腰旗橄榄球这项在全球范围内普及率远不如举重的运动，几乎完全为美国人所熟知。
   * **东道国可以在基础设施、财政支持等方面给予更多保障**：国家奥委会可以借助主办方的身份，在赛前投入更多的资金和资源来提升运动员的经济水平和比赛适应性。此外，作为东道国的运动员收到国民的期待，这种“为国争光”的心里驱动能激发运动员的潜力。
2. 我们统计分析了每个项目中各个国家的奖牌占比，并取每个项目中最高占比进行升序排序，找到没有被垄断的项目，如【】【】【】。这个结果揭示了当项目的奖牌分布越均匀，意味着更多国家有机会争夺奖牌，这为不同了类型的国家提供了有价值的参考。
   * **对奥运强国的启示**：强国奥委会应深入分析这些“未被垄断”的项目，评估是否可以在这些领域加大投资，培养新一代运动员，增强其在奥运会中某项目的表现。
   * **对新兴体育国家的启示**：对于新兴体育国家来说，奥运强国在传统大项，如田径、游泳、体操等项目的优势长期存在，但上文中验证的“伟大教练”效应也为新兴体育国家提供一个新的突破点，比如肯尼亚的长跑具有优势，但是缺乏教练，聘请“伟大领袖”来打破奥运强国的长期垄断。同时这个结果提供了一个潜在的奖牌增长点，通过精确的项目定位和科学的训练计划，可以快速提升在“未被垄断”项目中的竞争力，取得显著成绩 。

# **8 Model Assessment**

## **8.1 Strengths**

1. 我们对模型进行数据分析时，综合考虑了参赛人数、参赛项目、主办方、优秀教练等多个关键因素，因此进行了大量的数据验证工作，最后我们的预测得以不仅仅依赖于历届的奖牌数。

2. 我们在搭建模型时，使用了大量其他模型，如隐马尔科夫模型、多元logistic回归等进行对比实验。

3. 运用多种数学建模手段，包括机械学习、神经网络、图论，灵活调整变量和权重，以适应不同国家的奖牌预测。

## **8.2 Weaknesses**

1. 在数据清洗的时候，由于时间有限，我们以athletes数据集作为有效NOC的参考源，删除了hosts和medals数据集中不存在于athletes数据集的NOC条目，没有更加细致的考虑国家由于历史因素的变化。

2.缺乏数据，无法进行更多维的验证模型，舍弃了GDP等影响因素，多元线性回归可以更好验证“伟大教练”。