# **Summary**

Contents

[Summary 1](#_Toc30122)

[1 Introduction 3](#_Toc28753)

[1.1 Problem Background 3](#_Toc22632)

[1.2 Restatement of Problem 3](#_Toc2355)

[1.3 Our Work 4](#_Toc10255)

[2.Question Preparation 5](#_Toc12910)

[2.1 Assumptions 5](#_Toc11881)

[2.2 Notations 5](#_Toc14692)

[3 Data Preprocessing 7](#_Toc27419)

[3.1 Basic Data Preprocessing 7](#_Toc12205)

[3.2 Data Mining 8](#_Toc8133)

[3.2.1 运动员服役情况 8](#_Toc30020)

[3.2.2 国家优势项目分布 9](#_Toc9701)

[4 Task1:Many Hands Make Light Work-Olympic Medal Prediction Model Based on LightGBM 11](#_Toc22870)

[4.1 Medal Standings 11](#_Toc9462)

[4.2 Countries that Win Their First Medal 11](#_Toc11887)

[4.3 The relationship between events and the number of medals won by countries 11](#_Toc12902)

[5 Task2: 12](#_Toc21776)

[6 Task3： 13](#_Toc29266)

[7 Sensitivity Analysis 14](#_Toc6135)

[8 Model Evolution 15](#_Toc18231)

[8.1 Strengths 15](#_Toc20380)

[8.2 Weaknesses 15](#_Toc20459)

# **1 Introduction**

## **1.1 Problem Background**

在2024年巴黎夏季奥运会期间，观众们不仅关注各个单项赛事，还对各国的奖牌总数和金牌榜排名表现出浓厚的兴趣。最终，美国以126枚奖牌的总数位居奖牌榜首位，而中国和美国以40枚金牌并列金牌榜第一。东道主法国在金牌榜上排名第五，共获得16枚金牌，但在奖牌总数上排名第四。英国则以14枚金牌位列金牌榜第七，而在奖牌总数上排名第三。尽管排名靠前的国家备受瞩目，但其他国家的奖牌成绩同样受到关注。例如，阿尔巴尼亚、佛得角、多米尼克和圣卢西亚在本届奥运会上分别获得了本国历史上的首枚奥运奖牌，其中多米尼克和圣卢西亚还各赢得了一枚金牌。不过，目前仍有60多个国家尚未在奥运会上获得奖牌。对于奥运会最终奖牌数的预测是常见的，但这些预测通常不是基于历史奖牌榜，而是在即将举行的奥运会开始前，当参赛的现役运动员名单已知时进行的。

## **1.2 Restatement of Problem**

为了明确任务，下面是对问题的重述：

1. 为每一个国家的奖牌数建立一个至少包含金牌数和奖牌总数的模型，估计模型预测的不确定性/精确度，衡量模型性能。
   1. 根据模型预测2028年美国洛杉矶夏季奥运会的奖牌榜，包括所有结果的预测区间。依据模型预测判断那些国家的成绩最有可能提高，那些国家成绩会比上一届差。
   2. 建立的模型应包括尚未获得奖牌的国家，同时预测将在下一届奥运会上获得第一枚奖牌的国家数量，并对这一估计给出赔率。
   3. 建立的模型还应考虑特定运动会的项目数量和类型，探索项目与国家获得奖牌数之间的关系。对于各个国家，判断哪些项目是最重要的及其重要的原因，以及本国选择的项目对结果的影响。
2. 相比较于运动员需要改变公民身份才能代表不同国家参赛，教练不需要以公民身份执教，所以可以更轻易地转到其它国家。因此，可能产生“伟大教练”效应。例如郎平曾执教中美两国的排球队并均获得冠军，贝拉-卡洛丽曾执教罗马尼亚和美国女子体操对并获得巨大成就。请从数据中寻找可能由“伟大教练”效应引起变化的证据，估计这种效应对奖牌数的影响。选择三个国家，确定它们应考虑投资“优秀”教练的体育项目，并估计其影响。
3. 解释建立的模型中还包含了哪些有关奥运奖牌数的独到见解以及这些见解如何为各国奥委会提供信息。

## **1.3 Our Work**

# **2.Question Preparation**

## **2.1 Assumptions**

1. 每个国家每年获得的奖牌数量通常波动较小，不会出现突增或骤降的情况。一个国家的奖牌数量更多地反映了过去多年在体育领域的投入与培养成果，而不是短期内能大幅改变的结果。此外，影响奖牌数量的外部因素（如国际赛事的组织、规则变化、裁判标准等）在短期内通常保持相对稳定。

2. 每年都会有一部分新的运动员加入奥运会，他们可能是首次参赛。具体比例将通过我们的模拟进行估算。体育发展的更新换代是一个必然趋势，每年都会有一批年轻运动员崭露头角，逐步取代退役或不再参赛的老运动员。通过模拟计算新增运动员的比例，可以为模型提供更加贴合现实的输入，从而提升模型的准确性和适用性。

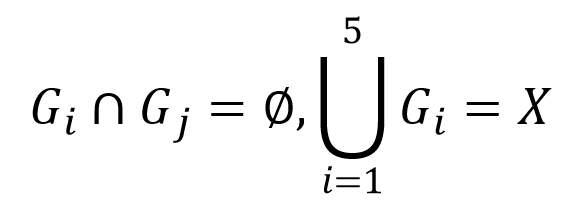
## **2.2 Notations**

|  |  |
| --- | --- |
| 符号 | 意义 |
| Ci | 某个国家 |
| Si | 某项运动项目 |
| G | 金牌数 |
| V | 银牌数 |
| B | 铜牌数 |
| M | 三类奖牌总数 |
| M\_c | 某国家奖牌总数 |
|  | 某国家某运动奖牌总数 |
| N | 参赛人数 |
| A | 新增的运动员人数 |
| D | 减少的运动员人数 |
| P | 项目优势占比 |
| E | 运动员参加奥运会的次数 |
| Gi | 奖牌数聚类类别 |
| X | 历届各国金、银、铜奖牌总数的集合 |

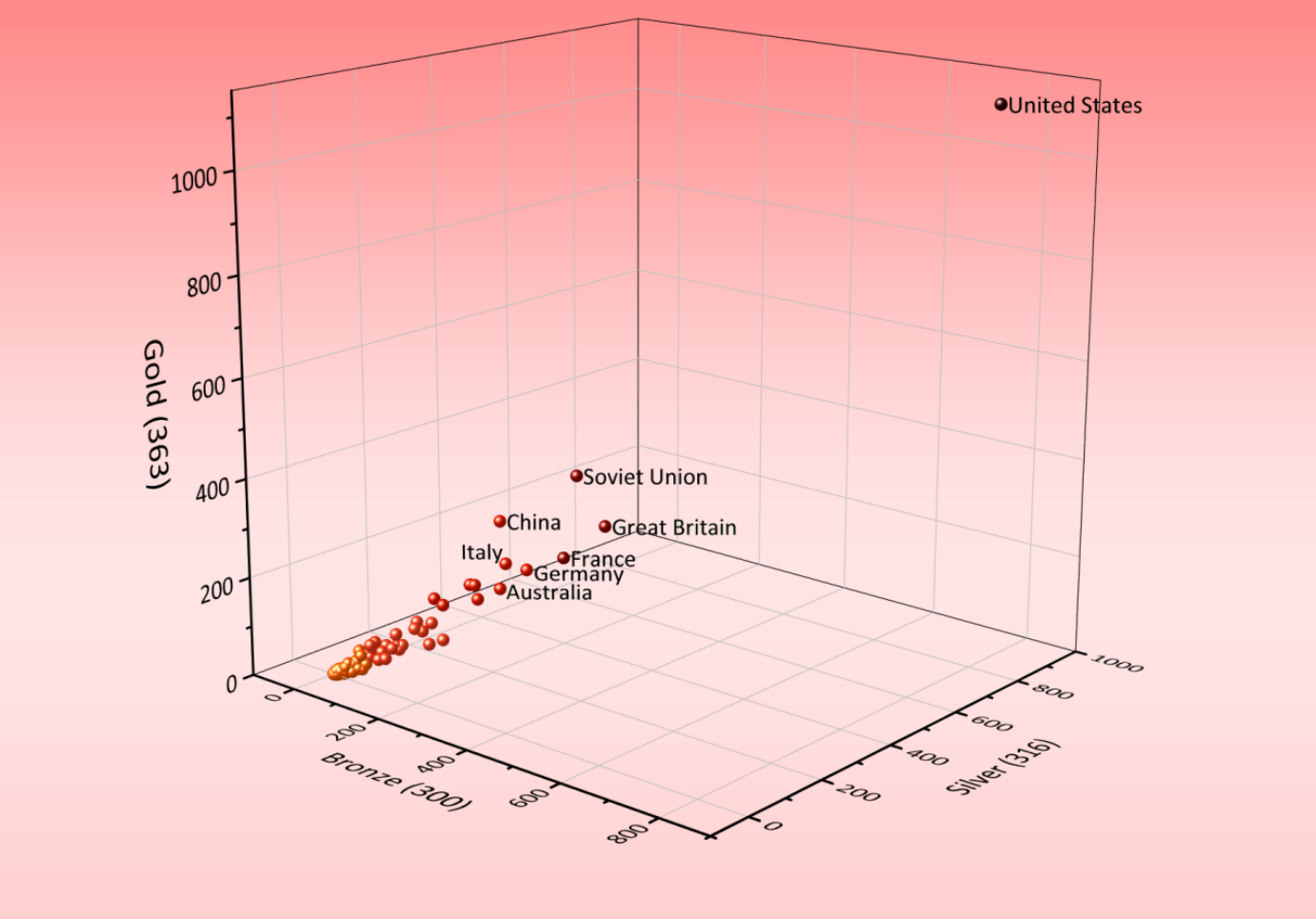
# **3 Data Preprocessing**

## **3.1 Basic Data Preprocessing**

由于各种因素的影响，各国在奥运会中的竞赛水平存在较大差异，这些差异在历届比赛中得到了体现，具体表现为各国过去获得的金、银、铜奖牌的总数。为了展示各国在历届奥运会中水平的差异（即国家水平），我们采用了 k-means++ 聚类算法，根据各国在历届奥运会中的金、银、铜牌数以及奖牌总数，将其划分为五个等级。这五个等级G1,G2,G3,G4,G5形成对样本集合X（历届各国金、银、铜奖牌的总数）的划分：



划分完后，五个等级的国家分布如图：



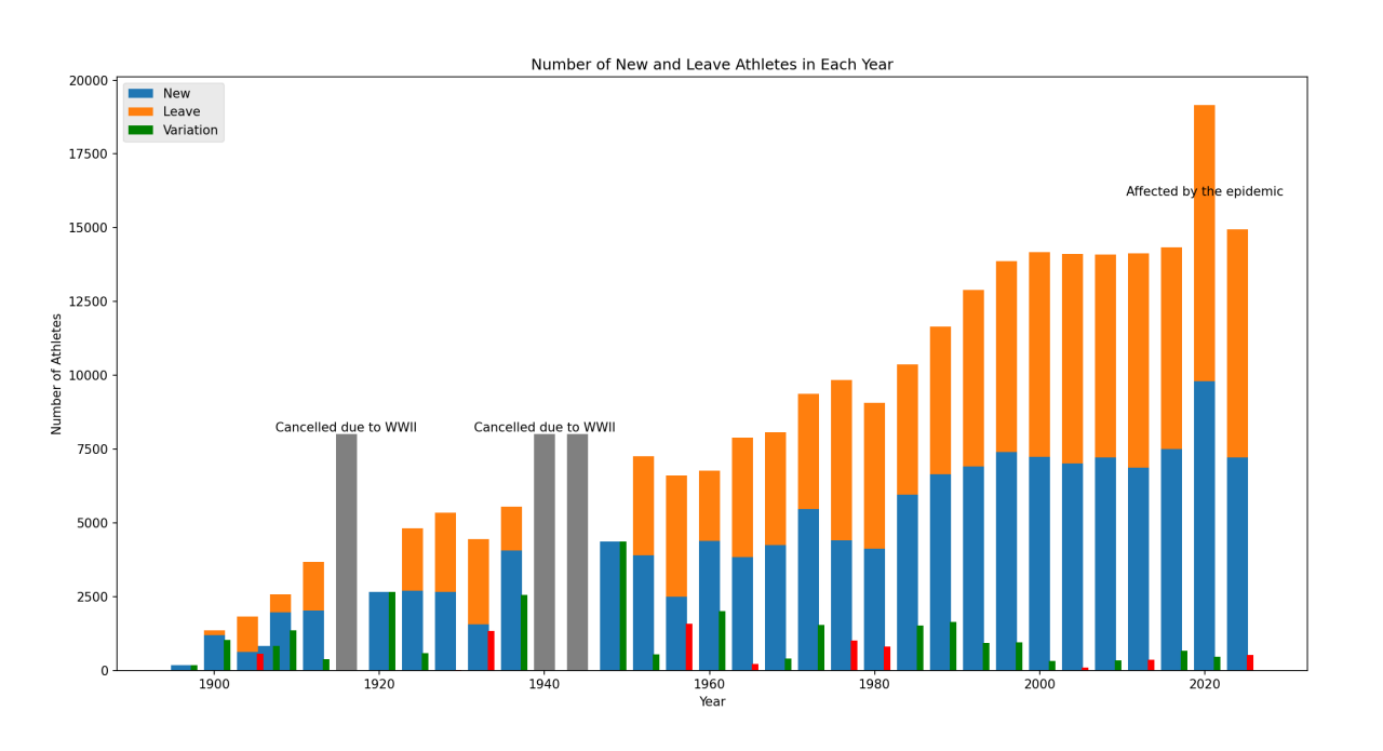
## **3.2 Data Mining**

### **3.2.1 运动员服役情况**

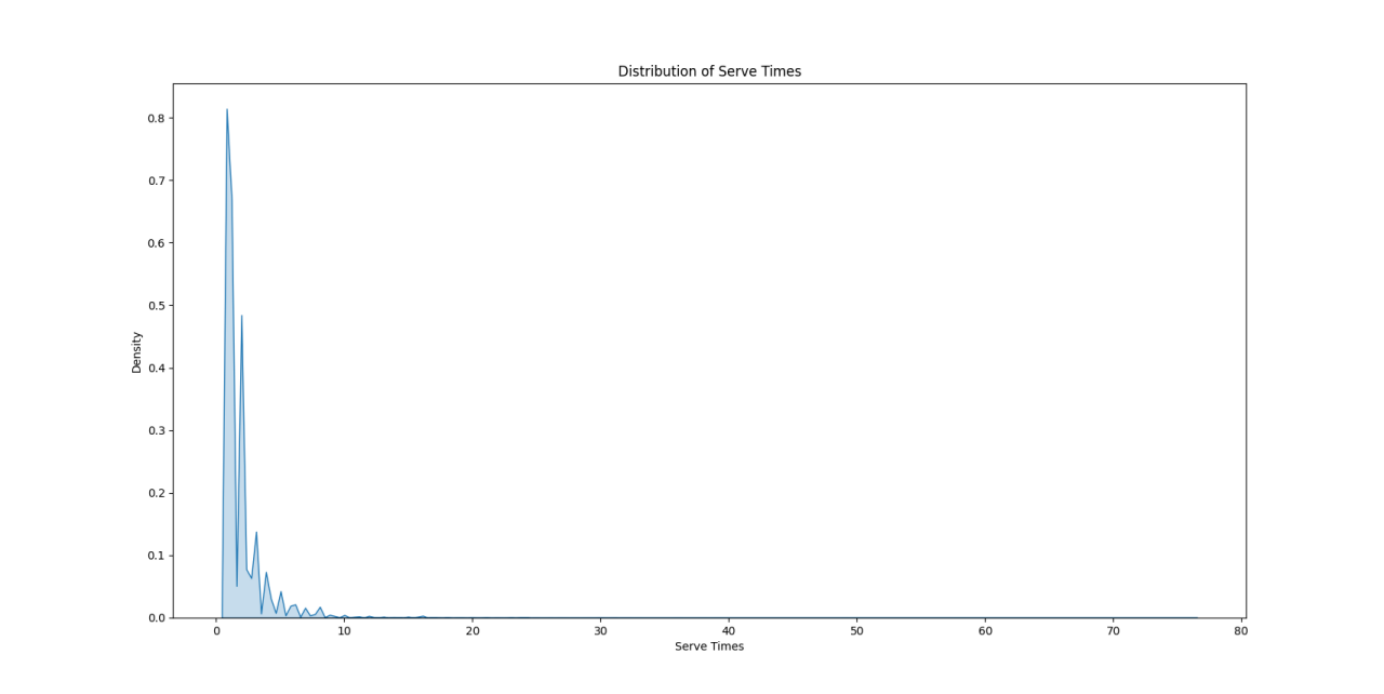
我们同时考虑了国家水平（Nation level）、是否为东道主（Whether host or not）以及该国在特定项目中的强项（Strength in this sport），将这三者统一归纳为国家背景因素（National background factors）。

Expected athlete number

为了对未来情况进行预测，我们分析各个国家的各个项目，各年份的人数变化走势，利用线性回归进行拟合。



对于具体的参赛人员，由于每年都有一些新兴的年轻运动员崭露头角，逐步取代退役或不再参赛的老运动员。首先，我们根据往年数据整理了不同运动员参赛次数的分布情况（英文），以及运动员年限分布的图像（英文），如图所示。

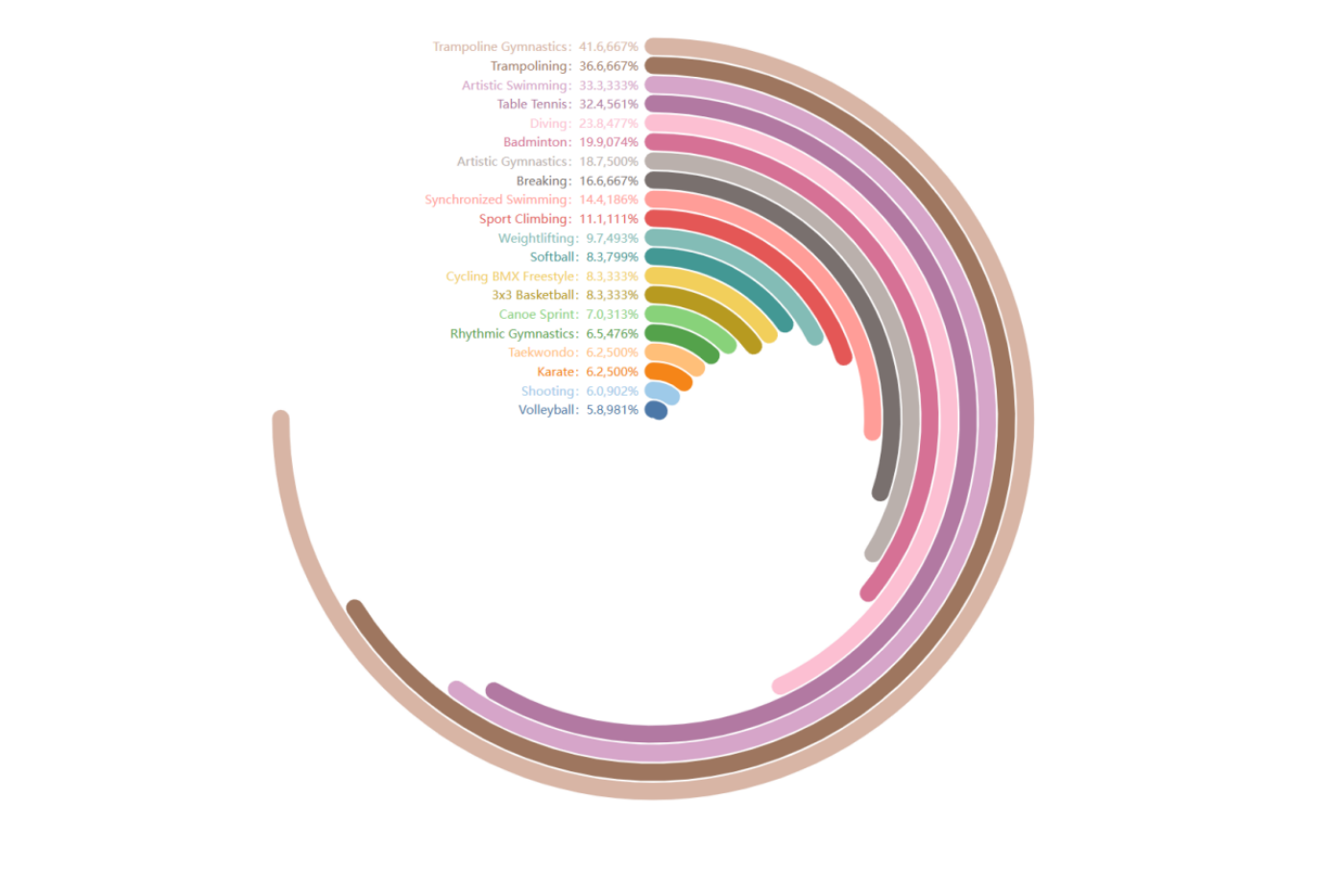


从分布图可以看出，接近80%的运动员仅参加了一到两次奥运会，且平均参赛次数为1.94次，即每位运动员大约参与两届奥运会。基于此假设，我们认为每个参赛队员平均参加两届比赛。另外，约80%的运动员只参加了一届奥运会，接近20%的运动员参加了两届。因此，我们可以假定有20%的老队员会继续参加下一届奥运会，这为我们确定哪些运动员将继续参赛提供了依据。

### **3.2.2 国家优势项目分布**

由于人文和地理环境等因素的影响，各国在不同运动项目上各具优势。例如，肯尼亚和埃塞俄比亚在长跑，尤其是马拉松和田径项目中表现突出，主要得益于高海拔地区的训练条件。巴西足球文化深厚，培养了众多世界级球员，这得益于温暖的气候和全年适宜进行户外运动的环境。

通过分析历届奥运会各国在各项目上的获奖情况，我们可以判断出这些国家在某些项目中的优势。因此，我们在数据集中挖掘出各国对不同项目的擅长程度，比如中国在各项运动中的表现，如下图所示：



# **4 Task1:Many Hands Make Light Work-Olympic Medal Prediction Model Based on LightGBM**

## **4.1 Medal Standings**

## **4.2 Countries that Win Their First Medal**

## **4.3 The relationship between events and the number of medals won by countries**

# **5 Task2:**

# **6 Task3：**

# **7 Sensitivity Analysis**

# **8 Model Evolution**

## **8.1 Strengths**

## **8.2 Weaknesses**

由于训练的数据没有时间关系，容易受到久远的历史数据影响，导致预测结果出现误差