|  |  |  |
| --- | --- | --- |
| **Problem Chosen** C | **2024 MCM/ICM Summary Sheet** | **Team Control Number** 2409948 |

**Momentum in Tennis**

**Summary**

**Keywords:** Momentum; Logistic Regression; Model; Predict

Contents最后记得更新整个目录

[1 Introduction 3](#_Toc58786693)

[1.1 Problem Background 3](#_Toc58786694)

[1.2 Restatement of the Problem 3](#_Toc58786695)

[1.3 Literature Review 3](#_Toc58786696)

[1.4 Our Work 3](#_Toc58786697)

[2 Assumptions and Justifications 4](#_Toc58786698)

[3 Notations 5](#_Toc58786699)

[4 The name of model 1 6](#_Toc58786700)

[4.1 Data Description 6](#_Toc58786701)

[4.2 The Establishment of Model 1 6](#_Toc58786702)

[4.3 The Solution of Model 1 7](#_Toc58786703)

[5 The name of model 2 7](#_Toc58786704)

[6 The name of model 3 7](#_Toc58786705)

[7 Sensitivity Analysis 7](#_Toc58786706)

[8 Model Evaluation and Further Discussion 8](#_Toc58786707)

[8.1 Strengths 8](#_Toc58786708)

[8.2 Weaknesses 8](#_Toc58786709)

[8.3 Further Discussion 8](#_Toc58786710)

[9 Conclusion 8](#_Toc58786711)

[References 9](#_Toc58786712)

[Appendices 10](#_Toc58786713)

# Introduction

## Problem Background

在2023年温布尔登男子单打决赛上，网球世界目睹了一场引人入胜的比赛，年轻的西班牙新星卡洛斯·阿尔卡拉斯以惊人的表现战胜了36岁的传奇选手诺瓦克·德约科维奇。这场比赛不仅结束了德约科维奇在温网自2013年以来的统治，而且彰显了网球比赛中动量变化的关键性。动量，作为体育比赛中的一种模糊概念，通常被认为是球员或球队在比赛中的强势表现，但究竟什么因素导致了这种变化，以及如何量化和理解这一现象，仍是一个值得深入研究的问题。

在网球比赛中，动量的变化常常是在几分甚至几局内发生的，而这种瞬息万变的情况给教练、球员和观众带来了巨大的挑战。尽管体育科学家一直在努力解开动量变化的奥秘，但目前的研究仍然相对有限。本研究的主要挑战之一是开发一个全面的模型，以捕捉比赛中动量变化的关键因素，并在此基础上提供实用的战术建议。

我们的研究旨在通过深入分析温网2023男子比赛的数据，建立一个可靠的模型，揭示比赛中动量变化的模式和影响因素。通过这一模型，我们将能够识别哪位球员在比赛中表现更好、更稳定，以及他们的优势程度。这不仅将为教练提供更深刻的洞察，而且有望为运动员提供更有效的应对策略，从而提高比赛的竞争力。在这一领域的研究对于推动体育科学和运动员训练的发展至关重要。

## Restatement of the Problem

Task 1: 比赛流程捕捉模型的开发

开发一个模型，通过捕捉比赛得分时的动态变化，识别在比赛中哪位球员表现更佳以及表现水平如何。模型将应用于一个或多个网球比赛，利用可视化手段呈现比赛的动态流程。

Task 2: 对“动量”在比赛中的评估

评估“动量”在比赛中的作用，以确认或反驳一位怀疑“动量”作用的网球教练的看法。利用开发的模型和相关指标来解析比赛中球员的波动和成功是否具有随机性。

Task 3: 波动预测模型的开发

基于至少一场比赛的数据，开发一个模型，能够预测比赛中发生的波动。确定最相关的因素，并提供建议，指导球员在新比赛中与其他球员竞争时如何应对可能的波动。

Task 4: 模型在其他比赛上的测试

测试开发的模型在其他比赛中的效果，评估其对比赛波动的预测能力。识别模型表现不佳的情况，并分析可能需要纳入未来模型的因素。

Task 5: 模型泛化和最终报告

检查开发的模型在其他比赛、锦标赛、球场表面和不同运动上的泛化能力。撰写最终报告，总结研究发现，提供不超过25页的完整解决方案，并为教练提供建议，使球员能够有效应对比赛中影响流程的事件。

## Literature Review

Literature Review：文献综述就是把关于当前问题的现有研究成果做个概述。首先需要阅读大量解决该问题的论文，其次得用自己的话总结出来。

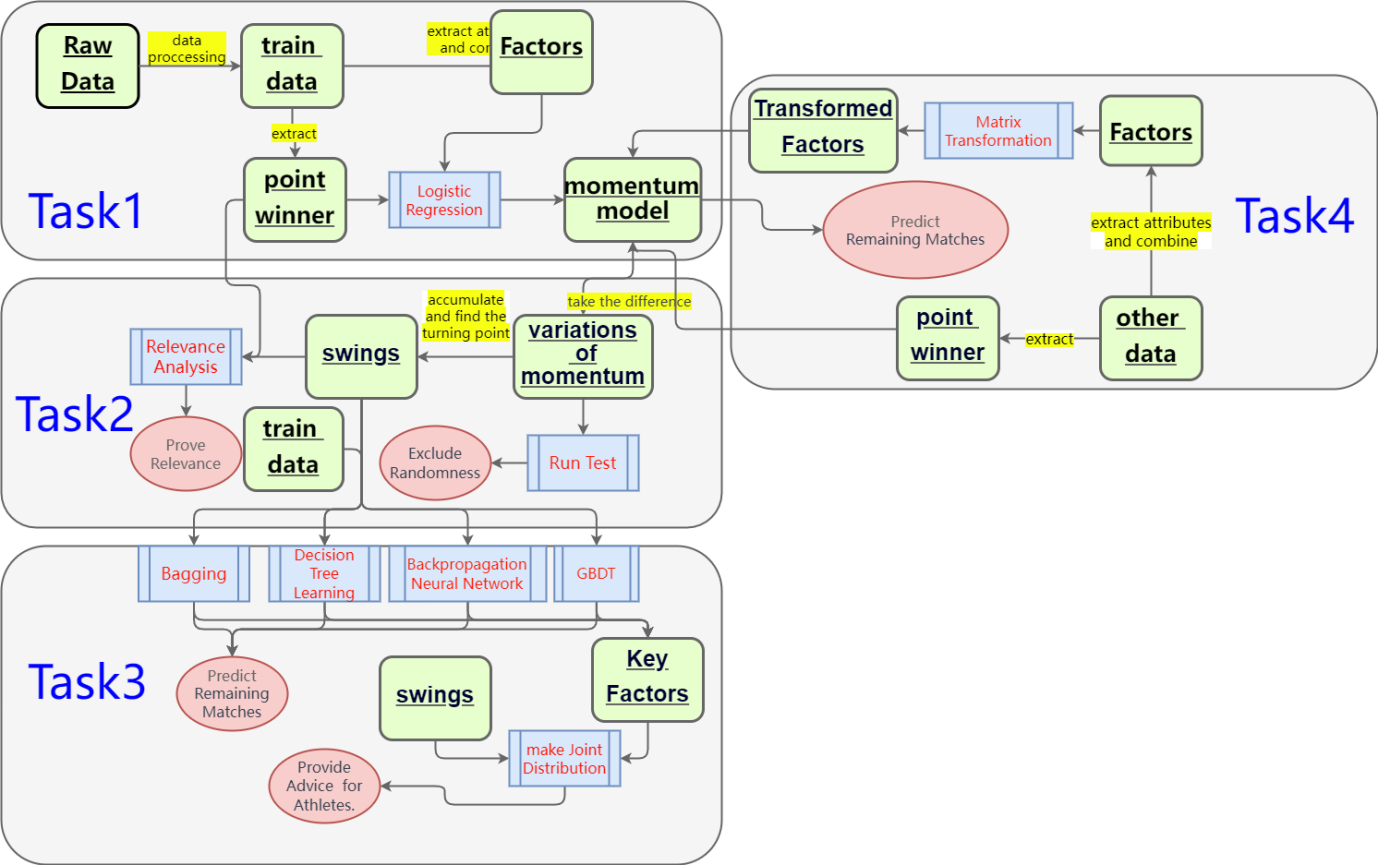
除非想冲O奖，否则别写这部分。一来竞赛时间有限，不可能去阅读大量论文；二来能力有限，不一定能写好总结。

小技巧：去搜相关论文，一般发表的论文都会有文献综述部分，照着别人的综述用自己的话描述一遍即可。

## Our Work

Our Work：这个小部分主要介绍论文的分析思路和建模的框架， 有点像国赛论文中的问题分析部分，可以画个思维导图（可用软件亿图图示），大家可以学习。

例1：



Our Work

# Assumptions and Justifications

为了使我们的模型更加的清晰和利于建立，我们设置了一些基本的假设信息。在基于现有的条件下来说，这些假设非常重要，因为它能够让我们更加专注于题目给出的客观数据，而不是远动员的主观因素。

**假设一：运动员的势头只与比赛中的因素相关，而和自身的训练、实力、教练团队等一系列在比赛时间之外的因素无关。**这是由于势头是一个无法轻易衡量的量，他是一个短时间作用于运动员心理上的虚拟的东西，因此着重关注于比赛过程中的事件发生来衡量势头是合理的。而由于每一个运动员的球龄，性别，比赛经验，身体素质等长期或者天生的因素都会随着人变化而变化，为了保证建模的鲁棒性，我们选择忽略这些与一场比赛无关的因素。

**假设二：我们认为在比赛开始的时候双方运动员的势头是一样的。**我们承认赛前的准备和运动员的心理建设是经常能够左右一场比赛的关键，但是为了消除不同运动员个人原因给势头分析带来影响的可能，将双方的势头放在同一起跑线更为严谨。

**假设三：我们认为比赛中运动员势头的改变会很快的体现在比赛的得分上。**运动员的势头会反映出他当前状态的好坏，进而影响到该局比赛得分。这一点在模型一的建立和求解中会有很显著的体现。

# Notations

Notations是对模型中使用的重要变量进行说明，表格形式三线表，表头分别是Symbol（符号）、Description （含义）、Unit（单位）（可不写），一般排版时尽量放到一页中。

The key mathematical notations used in this paper are listed in Table 1.

* + - * 1. Notations used in this paper

|  |  |  |
| --- | --- | --- |
| **Symbol** | **Description** | **Unit** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

注意：

* 只需写主要模型用到的重点变量、全篇通用的变量
* 求解计算等过程中的局部变量不要写
* 符号要以公式的形式写；如果是物理量，可在描述里写单位
* 每个符号的描述要简短，控制在一行内

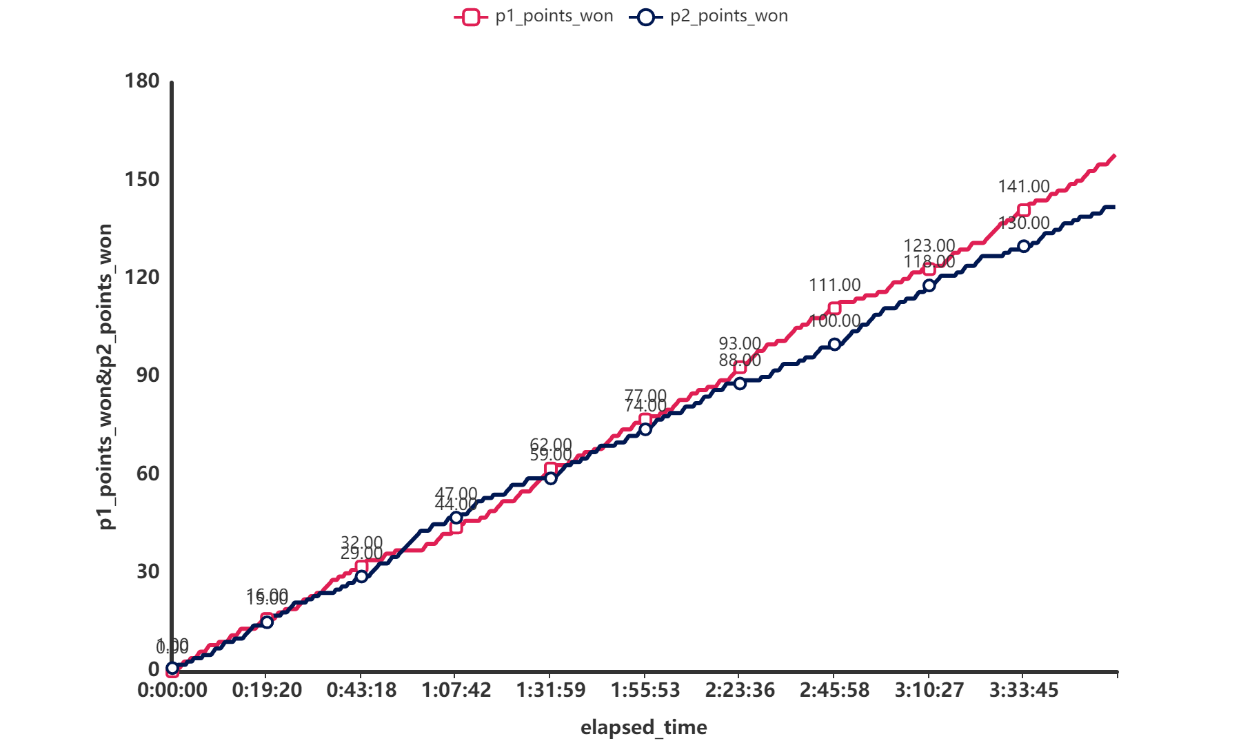
# The name of model 1

## Data Description

我们所使用的数据是由赛事组提供的2023年温布尔登网球公开赛的赛事统计数据。其中详细列出了与比赛相关的各种统计数据和信息。表格包含了两名球员的基本信息，例如他们的姓名，以及比赛进行时的详细情况，如已经进行的时间、目前的盘数、局数、分数。此外，还记录了各种特殊情况，比如无法回击的发球（ace）、赢得的得分、双误、非受迫性失误、网前得分等。这张表格还记录了运动员在比赛中的跑动距离、交换球的次数、发球速度、发球方向和深度，以及回球的深度。这样详尽的数据有助于分析球员的表现，制定策略，或是为观众提供更深入的比赛解读。

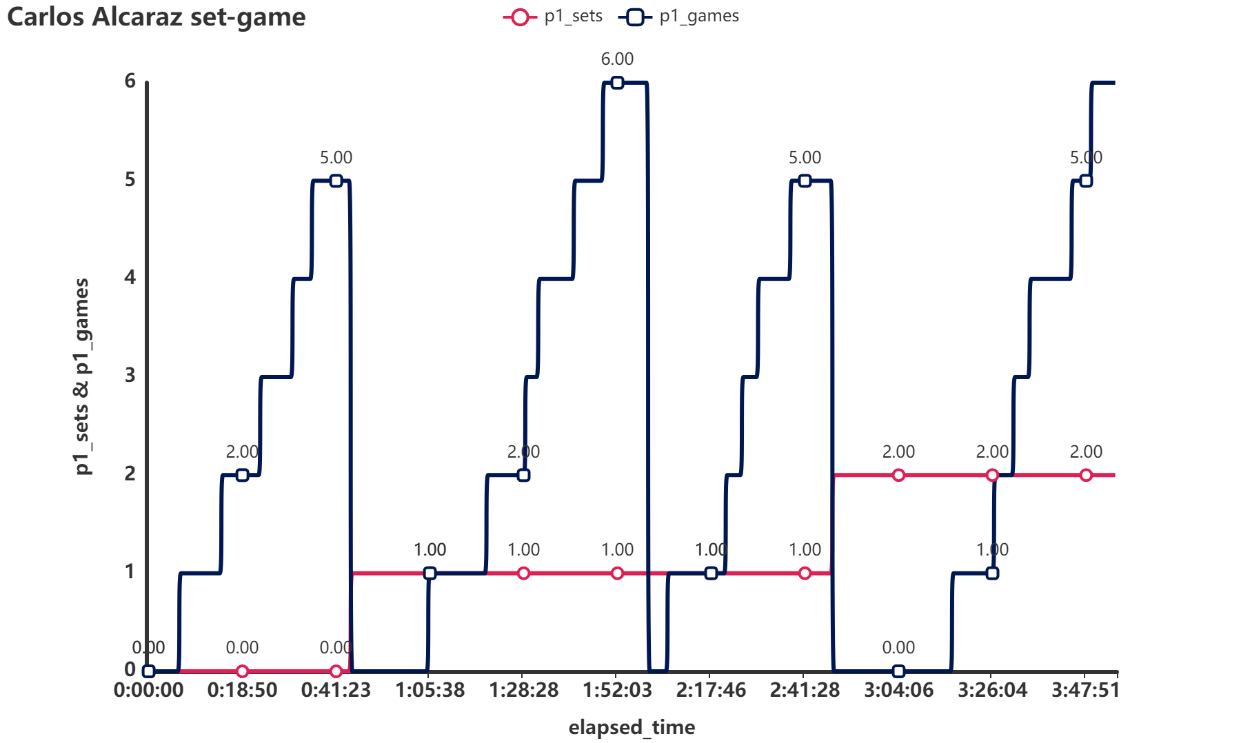
由于这是一张总表，包含了所有比赛的信息，为了能够更好的建模，我们首先提取出第一场比赛，赛事编号为2023-wimbledon-1301。这一场比赛中有本届赛事最终的冠军Carlos Alcaraz（Figure 2），因此有关于他的数据是最全面的。

Carlos Alcaraz



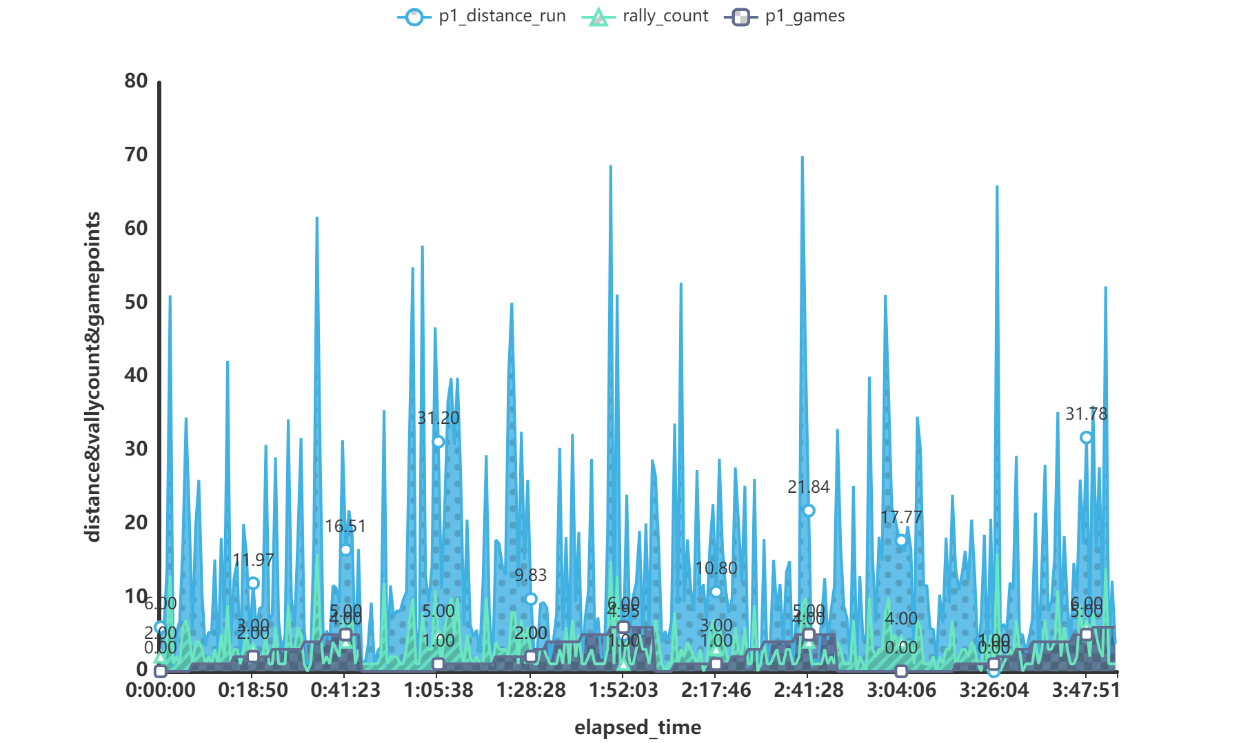
wimbledon-1301

Figure 3是本场比赛双方的得分。从图片中可以看出Carlos在前一个半小时内和对手之间的缠斗还是十分焦灼的，在这之后Carlos逐渐展现实力并且拉开和对手的分差，最终拿下这场比赛。



Carlos Alcaraz set-game

Figure 4这张图片表示了Carlos在比赛过程中拿下的game和set分。可以看到在第一个小时内Carlos拿下每一game的时间较为平均，这证明了在比赛的初期Carlos的状态比较稳定，这样帮助他成功拿下了第一set的胜利。而进入到第二盘后他拿下game的时间花费明显增长，这表明此时对手占据了比赛的主导权，因此Carlos输掉了第二盘。在这之后Carlos的状态有了明显回升，几乎仅用了第二set一半的时间就拿下了第三set，然后夺得了本场比赛的胜利。



Run distance and rally count

Figure 5 是针对Carlos每一game中的跑动距离和回合拍数。可以看到在一个game中，每当接近局点的时候，双方的单回合跑动距离和拍数都会有明显的增长，这表明在局点附近的关键阶段，双方球员都处于极度亢奋状态，此时他们的精神和注意力都达到了顶峰。而在一个game的开始阶段，双方的跑动和挥拍数又会回到一个比较低的数值，这表明在每一个game的开局阶段，双方的状态都会被缓和。根据这些数值的变化趋势，momentum也可能会有相应明显的改变，这些都是我们在之后需要去探究的问题。

这部分内容在论文中的位置也比较灵活，我们可以将这个内容放到单独一节，也可以放到模型的建立与求解中，也有部分论文放在了引言部分，还有的论文将这个内容放到“Model Preparation模型准备”这个部分。特别的，如果做的是美赛C题（C题一般是数据分析类型的题目），我们可以把这个部分单独作为一个大的部分，然后进行数据预处理和数据可视化分析。

## The Establishment of Model 1

### Momentum建模方向

我们首先需要分析“动量”的定义是什么。动量的字典定义是“通过运动或一系列事件获得的力量或力”。在体育运动中，一支球队或球员可能会觉得他们在比赛中拥有动力或“力量”，但很难衡量这种现象。因此我们要从比赛的两个角度入手，长期因素和短期因素。

关于长期因素，我们可以回想一下自己以前所参加过的一些球赛，如果已经拿到了很多的分数，那么在之后的比赛中紧张感会降低，心态会更加放松，但同样的，紧张感的降低也可能会导致注意力下降，求胜心减弱等一系列负面因素。

关于短期因素，其实也很容易进行想象，如果连续得分，那么必然会自信心增加，momentum也就会更加高涨。但如果连续失误或者被对手破掉自己的发球局，那么momentum就会受到打压。

### Momentum建模公式

下面将给出momentum的模型，由于momentum模型中包含了很多属性，因此选择引入中间变量，是特征向量，是系数向量。

|  |  |
| --- | --- |
|  | () |

特征向量是由以下8个属性组成

|  |  |
| --- | --- |
|  | () |

系数向量由八个系数标量组成，由逻辑回归进行拟合之后得出

|  |  |
| --- | --- |
|  | () |

长期因素：针对全局变量直接求得。

Score Factor表示已经拿下的分数对此时momentum的影响，具体计算方法为game分数之差与六倍的set分数之差的加和。

|  |  |
| --- | --- |
|  | () |

Serve Advantage表示当是当前的发球局归于哪个运动员，因为拥有发球局的一方获胜概率会大幅度增加，因此拥有发球局，我们直接赋值为1。没有发球局的赋值为-1。

|  |  |
| --- | --- |
|  | () |

短期因素：为了能够更好的体现短期因素的局部性影响，我们构建了长度为3时间窗口，它表示要将当前行和相邻的前三行作为一个子表格来进行求值。其中表示对子表格内的数据进行求和。

Break Factor表示在这个短期窗口内存在的破发因素，分子表示在这个窗口内双方之间的破发次数之差。分母则表示整个破发过程所需要的回合数，尝试破发的回合数越多，破发成功对momentum的正向收益就会越被稀释。

|  |  |
| --- | --- |
|  | () |

Rally Factor表示的是长回合球因素，在给定的窗口内，如果能够在较少的拍数下就能够拿下分数会对运动员的momentum有更多的正向激励。

|  |  |
| --- | --- |
|  | () |

Points Advantages表示在时间窗口内的净胜球

|  |  |
| --- | --- |
|  | () |

Service Factor表示在时间窗口内运动员的发球质量

|  |  |
| --- | --- |
|  | () |

Unforced Errors表示在时间窗口内运动员遭遇的非受迫性失误次数，对于momentum是一个负向激励。

|  |  |
| --- | --- |
|  | () |

Winners表示在时间窗口内运动员完成的致胜球数量，对于momentum是正向激励。

|  |  |
| --- | --- |
|  | () |

## The Solution of Model 1

### 逻辑回归求解系数

根据模型假设三，运动员的momentum会及时反映在当前比赛的得分情况上，因此我们以每一分的得分者为目标，用逻辑回归来拟合momentum和得分之间的关系，从而求出momentum表达式中的8个子因素各自对应的系数。

* + - * 1. 逻辑回归结果

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 实验组=1.0 | 回归系数 | 标准误差 | Wald | P | OR | OR值95%置信区间 | |
| 上限 | 下限 |
| 常数 | 0.332 |  |  | NaN | 1.394 |  |  |
| Score Factor | 0.068 | 0.071 | 0.919 | 0.338 | 0.935 | 0.814 | 1.073 |
| Rally Factor | 0.217 |  |  | NaN | 1.242 |  |  |
| PointsAdvantage | 0.802 | 0.127 | 39.662 | 0.000\*\*\* | 2.23 | 1.737 | 2.862 |
| Serve Advantage | 0.671 | 0.171 | 15.431 | 0.000\*\*\* | 1.955 | 1.399 | 2.732 |
| Unforced Errors | 0.693 | 0.306 | 5.144 | 0.023\*\* | 1.999 | 1.099 | 3.639 |
| Winners | 0.008 | 0.29 | 0.001 | 0.979 | 1.008 | 0.571 | 1.778 |
| 因变量：p1\_victor | | | | | | | |

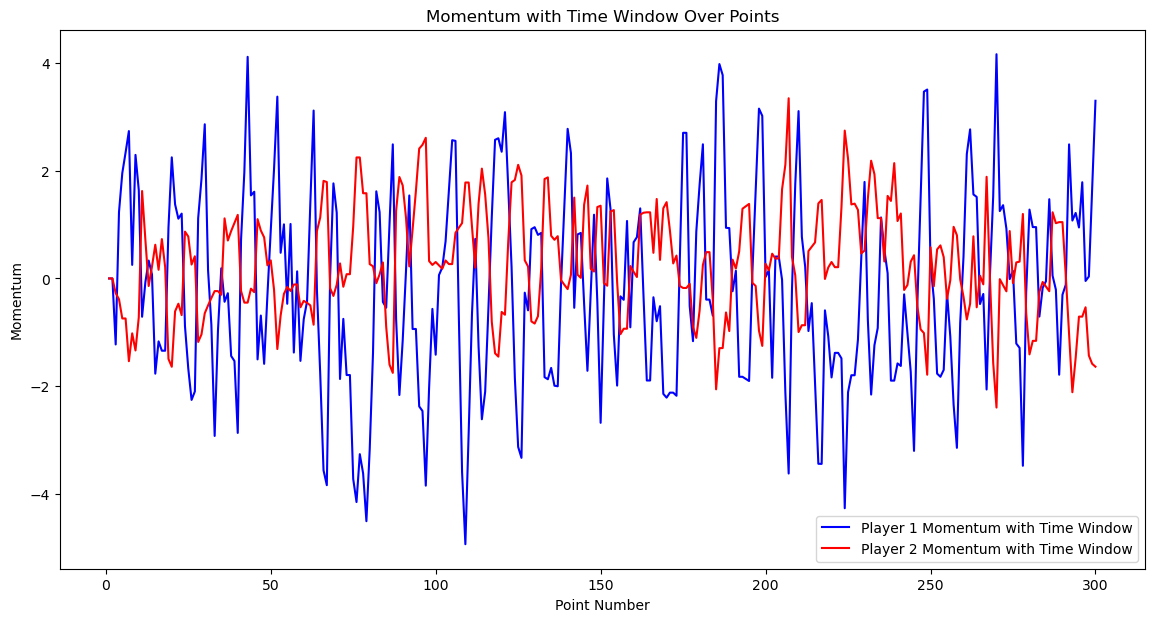
逻辑回归求出来的一系列结果都已经呈现在上面的表格中，其中每一个因素对应的回归系数都已经呈现在上面的表格内，将这些系数与对应的因素相乘后，再求和就是最终每一个时刻运动员对应的momentum。

特别的，在表格中可以发现Points Advantage, Serve Advantage和Unforced Errors的显著性P值非常低，水平上呈现显著性。因此它们会对球员获胜产生显著的影响。

* + - * 1. 逻辑回归评价指标

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 准确率 | 召回率 | 精确率 | F1 | AUC |
| 0.77 | 0.77 | 0.771 | 0.769 | 0.841 |

逻辑回归的准确率在77%，还是有不错的可靠性的。综合准确率AUC在84%左右，这表明利用这个模型可以有效地判断一个运动员实时的momentum状态。将得到的这些系数代入模型，就可以得到双方运动员在场上momentum，进而可以判断在这一刻谁的状态比较好，以2023-wimbledon-1301比赛为例。



Momentum in 2023-wimbledon-1301

如图所示，这是2023-wimbledon-1301比赛双方球员的momentum状态变化图，对于任意时刻球员的表现，均可以用momentum来进行表示。对于这样一场比赛，Carlos（运动员1）的表现变化要比他的对手更加剧烈，尤其是在前半段赛程，Carlos的表现一度被对手压制，但是在后半段赛程有了改善，momentum曲线也逐渐占据上风。

# Proof of Correlation and Exclusion of Randomness:

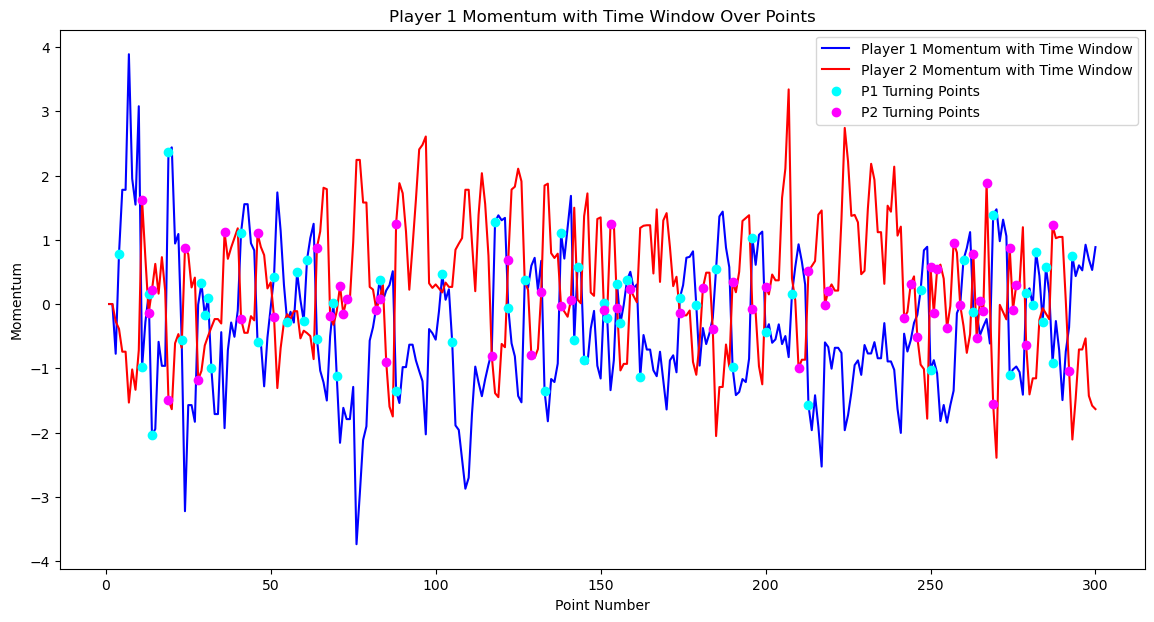
## Data Description

在本任务中，我们所用的数据是在第一个任务已经建立了momentum model的基础上，对每一行数据的momentum进行差分运算得到一列新的数据，记为momentum difference。然后再对momentum difference的值进行累加运算得到新的值，对于所有i，如果，说明势头从负变为正或者从正变为负，则i为波动点，记，否则记，最后即可得到关于波动的序列k，其图像如下图所示：



Fluctuation Plot

为了更直观地展现波动点，我们将其叠加到momentum图像中并标注出来，如下图所示：



Momentum Plot With Turning Points

## Proof of Correlation

在本小节中我们的目标是证明动量和运动员的得分是相关的，以此证明动量的确在比赛中发挥了一定的作用。我们使用了两种不同的相关性检验方法来做检验，分别是计算Pearson相关系数以及Spearman相关系数。下面分别简述两种方法：

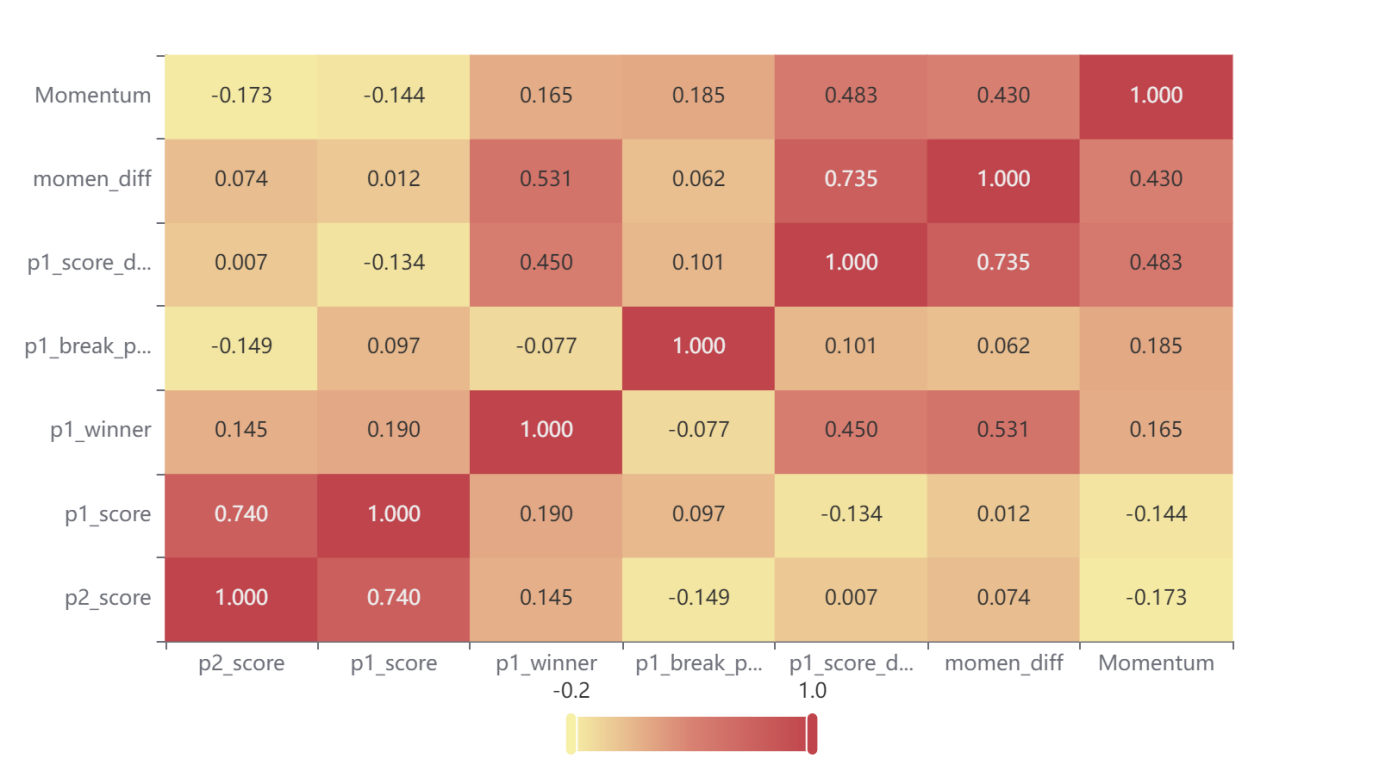
**Pearson系数：**

Pearson系数用于度量两个变量之间的线性关系。对任意两个变量X、Y，它们的Pearson相关系数的计算公式如下：

|  |  |
| --- | --- |
|  | (2) |

其中可以观察到分子为X和Y的协方差的值，分母为X和Y的标准差的乘积。计算的结果取值范围为[-1,1]，如果结果趋于1则表示X和Y正相关，趋近于-1则表示X和Y负相关。值得注意的是，若X和Y独立，则=0，但反之若=0我们则无法判断X和Y是否独立。例如考虑，E(XY)=E(X)=0,因此cov(X,Y)=0,但显然X和Y并不独立。

回到本次任务，我们为了证明动量直接影响到运动员的得分，可以通过计算得分的变化（得分的差分）和动量的变化（动量的差分）之间的相关性来证明，为了更直观地体现相关性的大小，我们还可以拿来其他数据例如运动员的破发得分、运动员的胜负等等并计算出彼此之间的Pearson系数来横向比较动量变化和得分变化的相关性是否较大。下面是我们计算得到的不同变量之间的Pearson系数热力图：

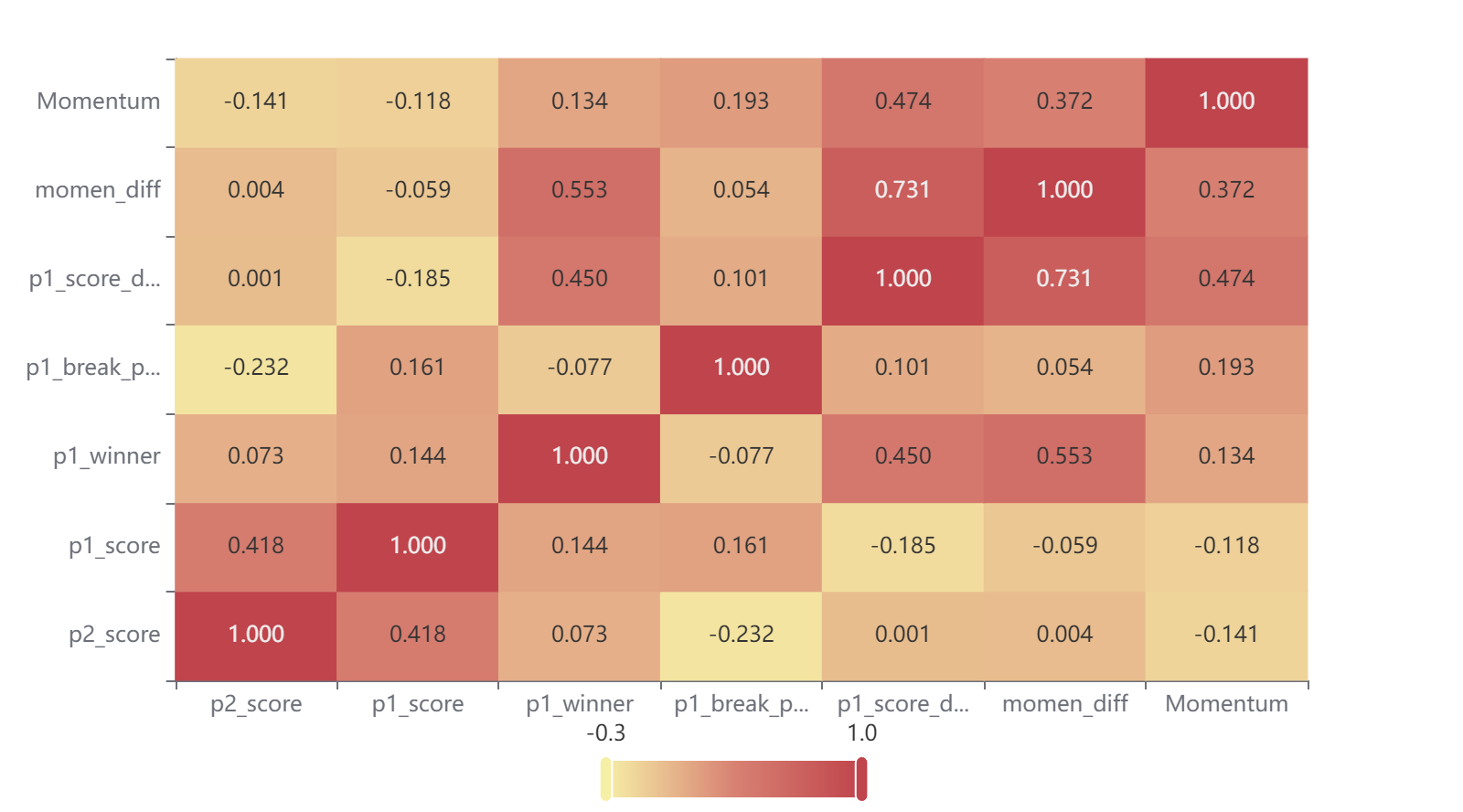


Pearson Correlation Coefficient Heatmap

可以看到动量变化(momen\_diff)和得分变化(p1\_score\_diff)Pearson系数达到了0.735，是一个较大值。说明动量的变化确实直接影响到了得分的变化，并且属于正向影响。进而也就说明了动量确实在比赛中发挥了一定作用。

**Spearman系数：**

除了使用Pearson系数以外，我们还用了Spearman系数来计算相关性。



## Exclusion of Randomness

这里可以写第一个模型的求解，把实际问题归结为一定的数学模型后，就要利用数学模型求解所提出的实际问题了。一般需要借助计算机软件进行求解，例如常用的软件有Matlab, Spss, Lingo, Excel, Stata, Python等。求解完成后，得到的求解结果应该规范准确并且醒目，若求解结果过长，最好编入附录里。（注意：如果使用智能优化算法或者数值计算方法求解的话，需要简要阐明算法的计算步骤）

同样的，很少有论文直接用“模型求解”作为这部分的标题，大家可以根据得到的结论来合理设计这里的标题。

另外，很多美赛论文对于模型的建立和求解没有区分开，这里沿用的是类似于国赛的形式，这样可以让论文框架清晰点。

|  |  |
| --- | --- |
| 这里插入公式 | () |

|  |  |
| --- | --- |
| 这里插入公式 | () |

# The name of model 3

基于上面Task 2的模型以及数据，我们获得了比赛中潜在的波动点序列，对于势头的走向和得分情况具有显著性的影响。然而，浩如烟海的模型参数对于模型计算和数据预测造成了很大的阻碍，并且无用的参数会损伤模型的灵敏度和鲁棒性。为此，我们选取了选手Carlos Alcaraz作为选手一的所有比赛，通过划分训练集和测试集（以比赛为单位）对既有的模型进行训练，通过随机森林bagging、BGDT等模型筛选出显著性特征参数。并对剩余的比赛进行测试，以验证对于波动预测的正确性。同时，综合以往比赛的势头波动差异，在波动点处发掘显著变化的参数，针对性地对于与新球员的比赛给出指导方法和建议。下面是我们对问题的求解：

## 因素的显著重要性筛选

在开始前，我们对数据进行处理，将模型的参数组合成矩阵 X，将模型中标记的波动点序列记为 Y，其中 X 来自2023-Wimbledon-1301比赛数据集，作为输入，Y 作为标签。通过将数据集X输入到预测模型中，并且通过预测值与真实波动序列（Y）进行比较得出预测正确率，进而评估模型的优劣与重要性参数。下面是我们选取的几种模型建构：

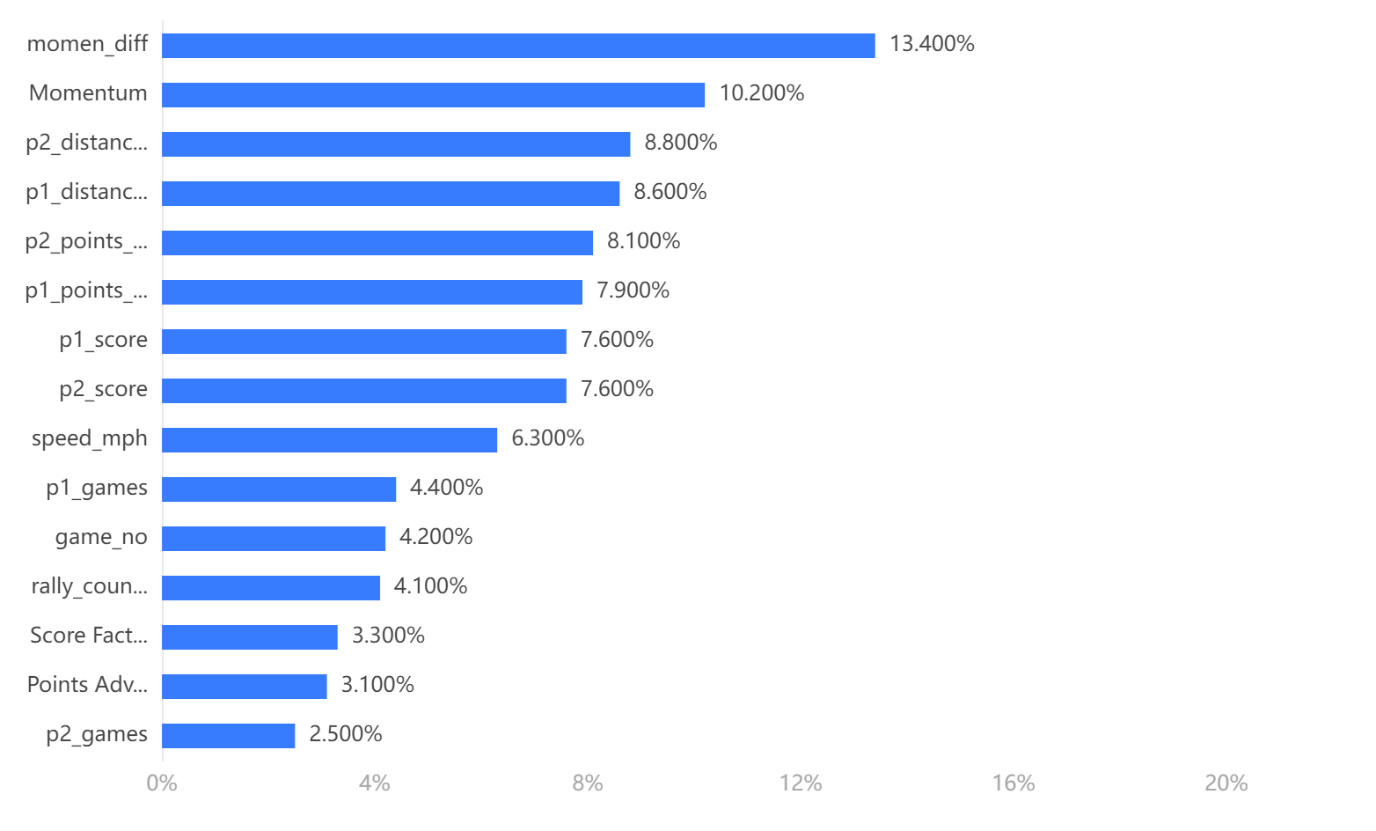
### 随机森林

考虑到问题具有非线性、局部无关联的特点，且维度较高，我们选择随机森林作为主要解决方案。随机森林是一种基于集成学习的方法，通过组合多个决策树的预测结果，能够更准确地捕捉复杂的非线性关系。对于大量特征的问题，随机森林通过子集的随机选择和决策树的组合，可以提取出对 Y 影响显著的特征。此外，随机森林通过对数据的有放回抽样，增加了对噪声（例如数据缺失、数据分布不均匀等）的抗干扰性。

首先，通过输入X和标签Y进行数据的又放回抽样生成多个决策树模型，其中每个决策树模型都会对波动点序列进行预测。我们假设在决策树下的数据集为D，A表示特征，其基尼不纯度的公式如下：

而在特征A下进行分裂的基尼不纯度的公式如下：

之后，我们选择具有最小基尼系数的特征作为节点分裂的标准，并递归地应用该过程，在随机森林的决策树生长过程中，内置了对特征重要性的评估机制，在每次树的分裂过程中，该算法通过测量每个特征的分裂贡献，计算出各个特征的重要性。因此，随机森林算法不断更新重要性参数列表，直到满足递归停止条件结束（这里设置为树的最大节点深度）。最后，根据得到的特征重要性排序，选择对 Y 影响显著的前 N 个特征。这些特征将作为模型的输入，进一步优化模型的精确性。统计出来的特征重要性排序与模型评估结果如下所示：



Random Forest Classification

* + - * 1. Random Forest Prediction Accuracy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Recall | Precision | F1 |
| Training set | 1 | 1 | 1 | 1 |
| Test set | 0.889 | 0.889 | 0.879 | 0.876 |

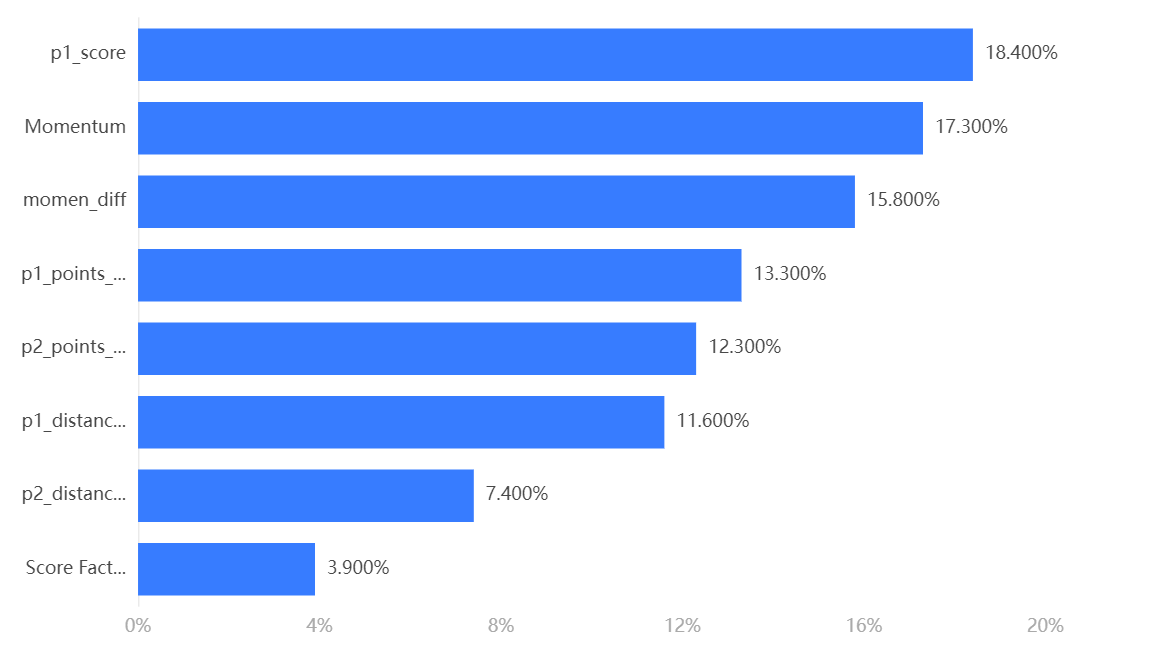
通过统计得出的数据可以看出，势头的差分和势头序列对于选手的波动影响最大，其次则是跑动距离和得分情况，另外还有球速、胜利局数、点数优势等都会影响球员的波动情况。

而在模型评估结果中，准确率是指正确预测的样本占总样本的比例。召回率也称为敏感性或真正例率，衡量的是实际为正样本的结果中，被正确预测为正样本的比例。精确率衡量的是被预测为正样本的结果中，实际为正样本的比例，三者越高越好。而F1值是精确率和召回率的调和平均，提供了在评估中同时考虑精确率和召回率的平衡。因此，综合来看，随机森林的评估效果优良，接近90%，具有很高的可信度

### GBDT梯度提升树

考虑到数据量的庞大和随机性分布，为了结果的严谨性，我们采用了另外一些方法进行数据特征重要性的筛选与排序。其中，GBDT梯度提升方法与随机森林同样作为集成学习，具有良好的抗噪声和迭代拟合效果。通过不断拟合当前模型的残差（预测值与实际值之差）来逐步提升模型性能，在第t轮时，模型的预测结果可以表示为当前模型的累计预测和前t-1轮模型残差之和：

每一棵决策树都尝试修正前一轮模型的预测错误。并且在迭代的过程中重要的特征会被频繁的选择，从而影响模型的走向。因此，类似于随机森林的特征确定方法，GBDT同样可以根据Gini系数决定决策树的分裂，并且记录对算法迭代具有显著贡献的特征参数，从而筛选出相应的Top N个重要特征，其运行结果与模型评估如下所示：



GBDT Classification

* + - * 1. GBDT Prediction Accuracy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Recall | Precision | F1 |
| Training set | 1 | 1 | 1 | 1 |
| Test set | 0.811 | 0.811 | 0.817 | 0.814 |

从分析结果可以看出，位于前面的重要参数与随机森林得到的结果保持一致，即势头和势头的差分、选手的得分情况与跑动的距离，说明这些参数对于模型的构建与优化具有显著的重要程度。

而在模型评估中，显然各项参数都不如随机森林优秀，因此其对于模型参数重要性的筛选的可信度相对低于随机森林的结果

### BP神经网络

虽然GBDT和随机森林在非线性问题上具有明显的优势，但是BP神经网络可能具有更灵活的拟合方式，并且具有较强的抽象数据挖掘能力，因此，可以作为对于上面两个模型的补充。

在训练的过程中，BP神经网络通过反向传播算法计算损失函数对每个权重的梯度，然后利用梯度下降法调整权重，以降低模型的预测误差。在第t轮训练中，其计算公式如下所示：

梯度是损失函数对于权重的变化率，这个梯度公式反映了损失函数对于权重的敏感程度，从而可以判断该权重的相对重要性。较大的绝对值梯度表示模型对该参数敏感，故该参数可能具有较大的贡献值，从而筛选出重要的特征属性。其评估结果如下所示：

* + - * 1. BP Neural Network Prediction Accuracy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Recall | Precision | F1 |
| Training set | 0.79 | 0.79 | 0.757 | 0.759 |
| Test set | 0.711 | 0.711 | 0.807 | 0.745 |

可以看出其评估性能明显差于前面的两个模型，故不能参考其重要性分析的结果

### 综述

综上所述，我们采用了三种方法对模型的参数重要性进行了评估，分别是随机森林、GBDT和BP神经网络。通过模型预测评估的结果我们可以看到随机森林得到的结果最好，因此我们采用随机森林提供的重要性参数列表，即势头和势头差分、跑动距离、得分情况、球速、胜利局数、点数优势等。我们将这些具有良好的重要性参数作为模型的参数列表，从而过滤掉无用的信息，提高模型的鲁棒性和预测精准度，能够更具有针对性地做出预测。

## 与过去比赛的势头差异比较

在上一问中，我们利用了球员Carlos Alcaraz的2023-Wimbledon-1301场次比赛

数据对模型进行了训练。而为了得到更为泛化的结论和数据，我们对于该选手的以往的五场比赛的momen\_diff（即势头的差分）进行了差异性分析，由于样本量相对较小且整体数据与正态分布契合程度不够高，我们采用了多配对样本Friedman检验方法，其检验结果如下所示：

* + - * 1. Results of Friedman Test Analysis

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable Name | Sample Size | Me dian | Standard Deviation | Statistic | P | Cohen's f Value | |
| momen\_diff\_1 | 334 | 0 | 1.46 | 0.65 | 0.957 | 0.002 | |
| momen\_diff \_2 | 334 | 0 | 1.473 |
| momen\_diff \_3 | 334 | 0 | 1.628 |
| momen\_diff \_4 | 334 | 0 | 0.988 |
| momen\_diff \_5 | 334 | 0 | 1.513 |
|  | | | | | | |

通过Friedman检验分析结果表可知，显著性P值为0.957，因此统计结果不显著，说明momen\_diff\_ \_1、momen\_diff\_ \_2、momen\_diff\_ \_3、momen\_diff\_ \_4、momen\_diff\_ \_5之间不存在显著差异；其差异幅度Cohen's f值为：0.002，极小程度差异。因此，可以说明该选手过往比赛呈现的波动性变化并不存在显著差异，说明过往的比赛可以作为联合参考并给出意见。

## 给出的建议与方法指导

# The name of model 4

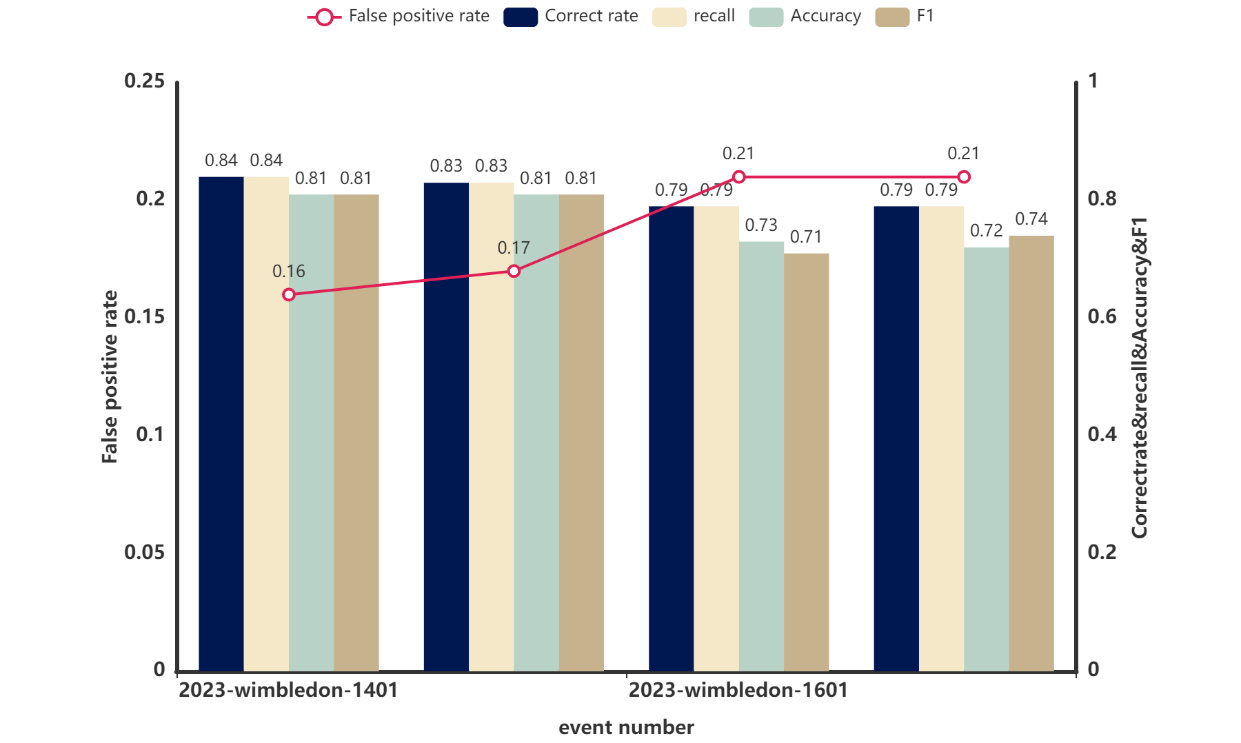
## 对不同赛程的多场比赛进行预测

### 数据提取

在这一部分要着重去测试模型对于同一运动员的其他场次的比赛的预测效果，因此我们提取了Carlos在这届赛事中剩下的四场比赛作为测试对象，其中就包括了和德约科维奇的决赛。我们要从结果中评估模型是否学习到了这个运动员本身的行为模式，同时也要考虑到在锦标赛不同赛段的情况下，运动员的momentum的变化情况。

### 模型预测

模型依旧延续了上一问中所使用的随机森林模型进行波动点的预测。下面的表格集成了这些比赛的预测准确率等结果。



后续比赛预测

* + - * 1. 后续比赛预测结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 赛事编号 | 正确率 | 召回率 | 精确率 | F1 | 误判率 |
| 2023-wimbledon-1401 | 0.8356 | 0.8356 | 0.8106 | 0.8083 | 0.1644 |
| 2023-wimbledon-1501 | 0.8307 | 0.8307 | 0.8135 | 0.8131 | 0.1693 |
| 2023-wimbledon-1601 | 0.7862 | 0.7862 | 0.7308 | 0.7132 | 0.2138 |
| 2023-wimbledon-1701 | 0.7934 | 0.7934 | 0.7225 | 0.7425 | 0.2066 |

我们可以很明显的发现距离2023-wimbledon-1301时间越长的比赛整个预测的准确度都会慢慢下降这是由于锦标赛的赛制原因，由于每晋级下一轮所遇到的对手的实力都会逐渐上升以及球员在不同的赛段给自己的心理压力的不同，都会导致自身momentum的波动模式有所改变，但是在整体的准确率上还是保持在了80%附近，证明这个模型的在同类运动中的泛化能力还是不错的。

## 对羽毛球的泛化程度评估

衡量一个模型的好坏，不能仅从模型的鲁棒性和预测精准度去考量，还要考察模型对于不同命题的泛化程度。因此，为了测量模型对于其他不同运动的势头及波动程度的泛化程度，我们选择了羽毛球这一项运动进行泛化度量。

### 数据处理

对于羽毛球比赛的数据，我们从开放数据平台Kaggle上获取，网址如下：<https://www.kaggle.com/datasets/sanderp/badminton-bwf-world-tour>。对于搜集到的数据集，我们首先需要进行数据清洗，将所有的空值（NAN）进行填充，默认填充为0。并且由于下载下来的数据集格式存在些许损坏和不匹配，需要对单个属性列中的所有数据的格式进行统一。同时，由于数据庞大会引入较多的噪声，因此我们基于赛程选取了具有代表性的阶段性比赛，如常规赛、淘汰赛、半决赛、决赛作为泛化测试数据集。在格式方面，我们也对羽毛球比赛数据集进行了与网球比赛数据集方面的匹配，对数据的格式进行了规范化。

### 参数映射矩阵变换

获得了预处理后的数据集后，由于输入参数矩阵的维数与模型的既定维数不匹配，需要进行矩阵变换将羽毛球比赛中的比赛参数与模型输入相匹配。假设羽毛球数据集中参数个数为m，组成大小为(1,m)的矩阵A，网球中参数为n，注意这里的参数并非是一开始建立模型所用到的，而是通过分析其特征重要性筛选出的Top N个参数，组成大小为(1,n)的矩阵B。因此，我们的目的是求出变换矩阵C(m,n)，使得羽毛球参数矩阵A可以通过右乘矩阵C得到变换后的维度匹配的矩阵A’，进而参与模型的输入与评估。

上面的模型建构可以由以下的公式表出，并且通过最小二乘法最小化误差进行矩阵求解。设A和B的关系为：

其中是待求变换矩阵的转置，E是误差表征。我们可以通过最小化误差，即利用最小二乘法去迭代求解，即：

表示L2范数即欧式距离，对于上面的问题的解可以表示为：

上面的公式成立的前提是B与B的转置相乘存在逆矩阵，如果不可逆，则需要采用正则化方法如岭回归进行求解。最终，我们求得了矩阵C的转置，只需要再进行一次转置即可得到变换矩阵C。之后对于得到的羽毛球参数矩阵A，则可以通过变换即AC得到A’从而与网球模型参数维度相匹配。

### 模型的输入与泛化评估

通过上面求得的变换矩阵，我们将羽毛球的参数输入到我们在第二问已经建立好的模型。而在输入之前，我们需要优化模型架构，根据第三个Task求得的具有良好的特征重要性参数列表，我们将模型参数列表置换为筛选出的具有良好特征重要性的参数，进一步过滤无用的信息并强化模型的预测能力。

做好了模型的优化工作，我们可以进行模型的输入与求解了。首先将变换后的羽毛球参数列表作为参数输入，得到模型的输出：羽毛球比赛的势头、势头差分和波动点。将上面的参数输入作为X矩阵，波动点作为标签Y，利用预测模型对标签Y进行预测，通过正确率对模型的泛化程度进行评估，其结果如下所示：

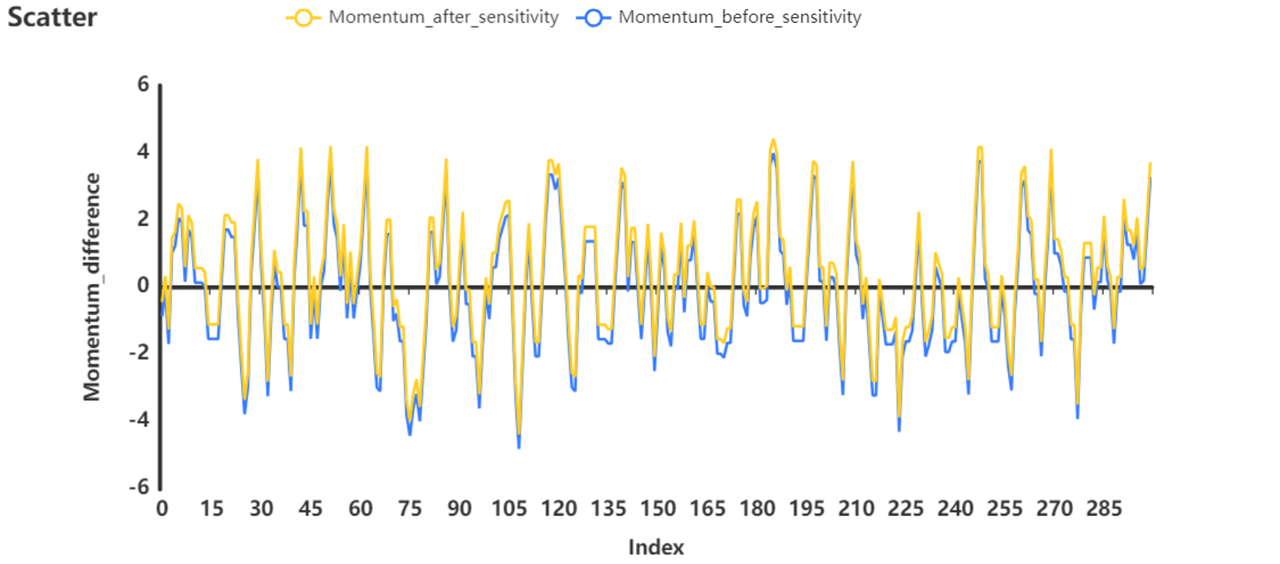
* + - * 1. Model Prediction Evaluation for Badminton

|  |  |
| --- | --- |
| evaluation metric | evaluation results on the test set |
| Accuracy | 0.7413793103448276 |
| Recall | 0.7413793103448276 |
| Precision | 0.7137818773738469 |
| F1 | 0.7214673913043479 |
| error rate | 0.2586206896551724 |

通过表格9我们可以观察到，预测正确率、召回率、精确率都稳定在74%左右，这对于不同种类的运动而言已经具有较为良好的泛化性。而泛化性预测正确率不高的原因则可能是在进行矩阵变换时参数从m到n维的变换会导致数据的精度损失和错误，因此增加了错误分类的概率大小，因而实际的准确率并不高。但是总体而言，模型在面对规则具有一定相似度且数据大量不一致的情况下，仍然保持70%以上的预测精度，已经具备了较为良好的泛化程度。

# Sensitivity Analysis

至此，我们完成了模型的建立、评估和优化，考虑参数变化的扰动，我们引入并构建了灵敏度分析模型：以第三题计算出的最重要特征参数集合为基础，从中选取参数进行整体变换，观察势头的计算曲线是否发生明显的波动与误差。这里我们选取了rally\_count参数，整体同乘1.1即增长10%，并且施加0.01绝对值范围内的随机扰动，前后的势头曲线走向如下所示：



Momentum\_sensitivity\_analysis

从上图看出，即使施加了增长性的因子扰动，模型的整体趋势并没有明显变化，相当于整体进行了提升，说明对重要因素的扰动并没有破坏模型输出的波形结构，具有良好的鲁棒性和稳定性。

# Model Evaluation and Further Discussion

注：本部分的标题需要根据你的内容进行调整，例如：如果你没有写进一步讨论的话，就直接把标题写成模型的评价。（优缺点一定要写）

## Strengths

这里写论文或者模型的优点

## Weaknesses

这里写缺点：缺点写的个数一般要比优点少

## Further Discussion

进行进一步的讨论，这里可以写模型的改进和拓展：

模型的改进：主要是针对模型中缺点有哪些可以改进的地方；

模型的拓展：将原题的要求进行扩展，进一步讨论模型的实用性和可行性。

# Conclusion

结论部分，这个部分在国赛论文很少见到，但在美赛中出现的频率很高。

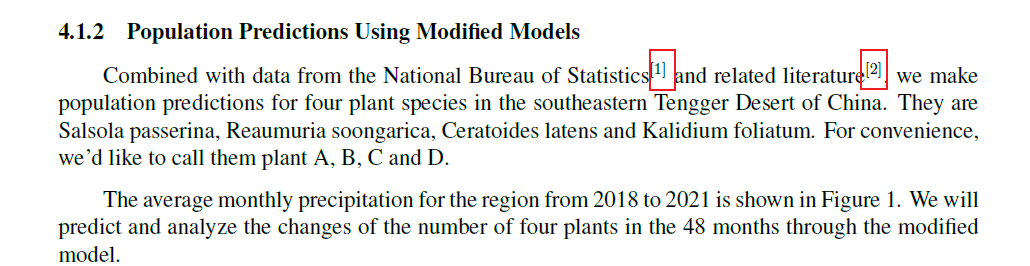
这个部分可以是论文中心思想的重申、研究结果或主要观点的归纳，也可以是某些启示性的解释或考虑。

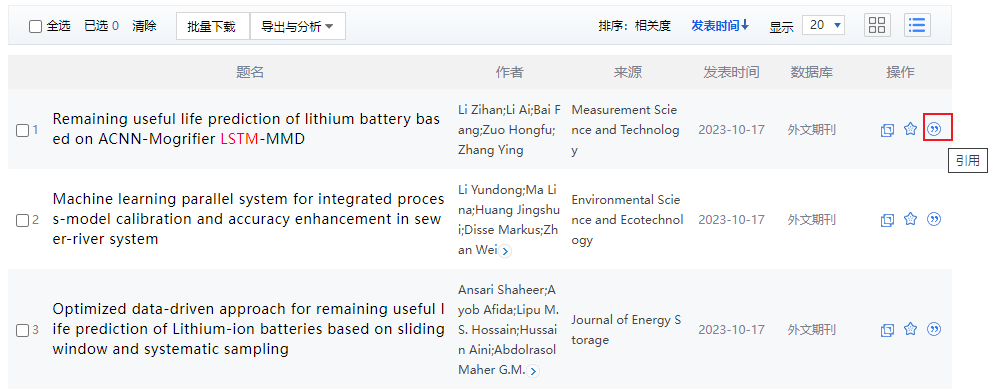
# References

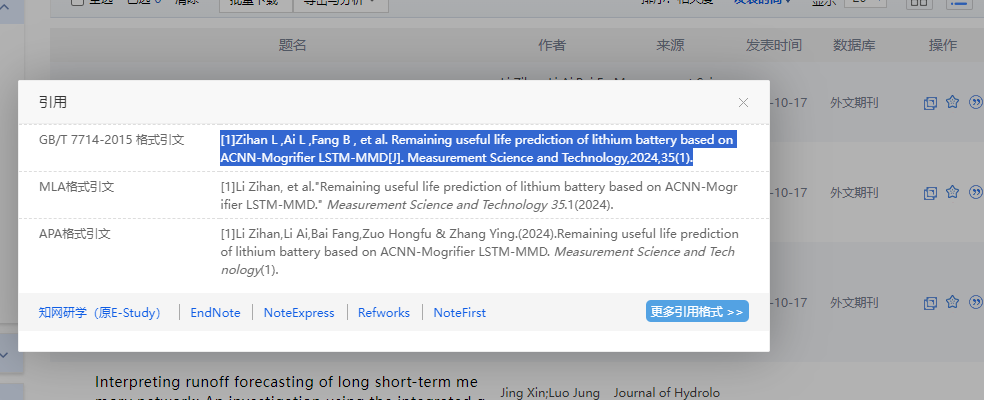
参考文献至少五六篇，引用中文文献记得翻译成英文。

在正文对应部分也设置序号

引用格式一定要正确，建议从检索网站直接导出，下图以知网为例







# Appendices

|  |
| --- |
| Appendix 1 |
| Introduce: 这里放上附录1的介绍 |
|  |

|  |
| --- |
| Appendix 2 |
| Introduce: 这里放上附录2的介绍 |
|  |

附录：可以放入重要的代码、一些中间计算过程、复杂的推导等内容

可有可无。比赛规定整个论文不能超过25页（包括附录，但不包括人工智能使用报告），所以完全可以不写附录

写的话，选重要的代码放在表格里，写清简介

* + - * 1. 123

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |