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

Momentum in Tennis

Summary

**Keywords:** Momentum; Logistic Regression; keyword3; keyword4

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

Introduction可以翻译成引言部分。下面给出了常见的四个部分，一般写background（背景）和Restatement of the Problem（问题重述）即可，大家可以根据自己的需要保留要写的部分。

## 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。

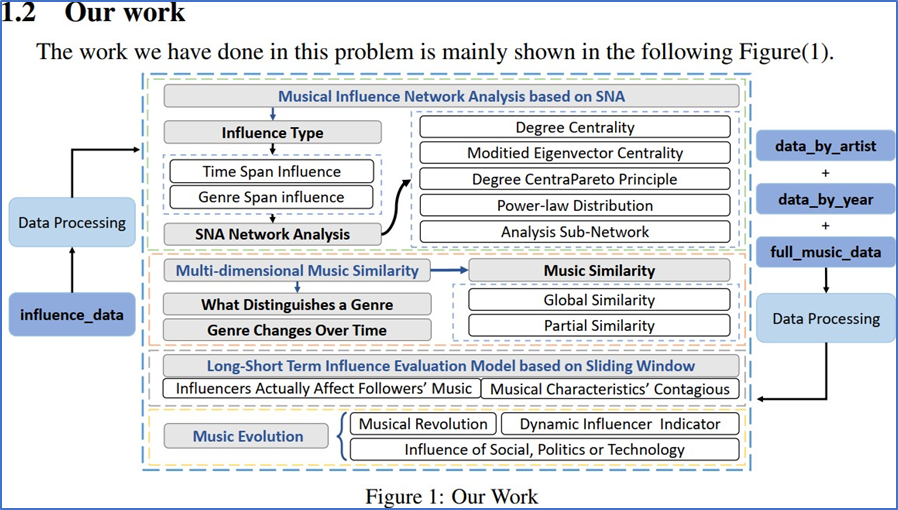
除非想冲O奖，否则别写这部分。一来竞赛时间有限，不可能去阅读大量论文；二来能力有限，不一定能写好总结。

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

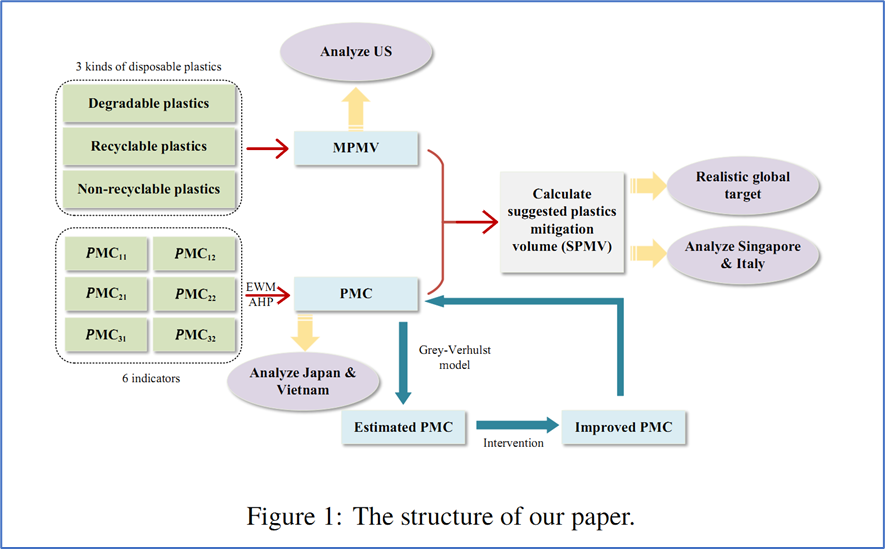
Our Work：这个小部分主要介绍论文的分析思路和建模的框架， 有点像国赛论文中的问题分析部分，可以画个思维导图（可用软件亿图图示），大家可以学习。

例1：



Our Work

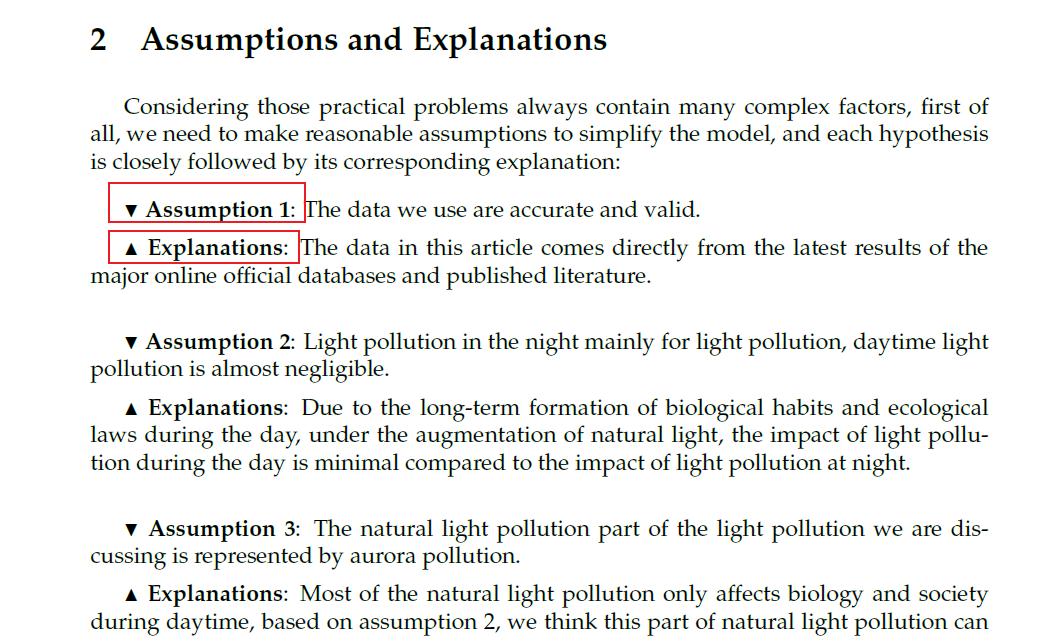
例2：



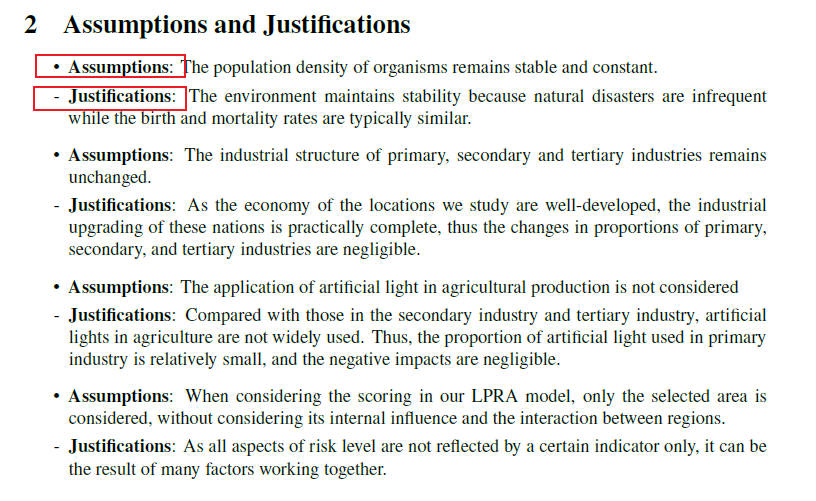
# Assumptions and Justifications

Justification翻译过来是正当理由的意思。这一部分要写模型假设，并且要对论证假设的合理性，这一点比国赛的要求要高，请大家引起足够的注意。

例1：



例2：



# 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

正文结构：

标题分级与命名

标题最多3级（模板已设置好字体格式）

命名格式多样（推荐以模型或小问来命名）

1）以模型命名

2）以具体问题命名

3）以问题1，2，3命名

4）一个标题命名

这一部分内容一定要全面，包括必要的分析，建模过程，结果展示，结果分析等。

## Data Description

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

这里可以写第一个模型的建立，模型建立是将原问题抽象成用数学语言的表达式，它一定是在先前的问题分析和模型假设的基础上得来的。因为比赛时间很紧，大多时候我们都是使用别人已经建立好的模型。这部分一定要将题目问的问题和模型紧密结合起来，切忌随意套用模型。我们还可以对已有模型的某一方面进行改进或者优化，或者建立不同的模型解决同一个问题，这样就是论文的创新和亮点。

和国赛不同，在美赛论文中，很少有论文直接用“模型建立”作为这部分的标题，一般使用模型要解决的问题作为标题。

如果需要用公式的话，可以复制下面这个隐藏的表格，并粘贴到你需要的地方，这个公式的编号会同步更新。

|  |  |
| --- | --- |
| 这里插入公式 | () |

## The Solution of Model 1

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 2

和上一个部分类似的写法。

# 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：

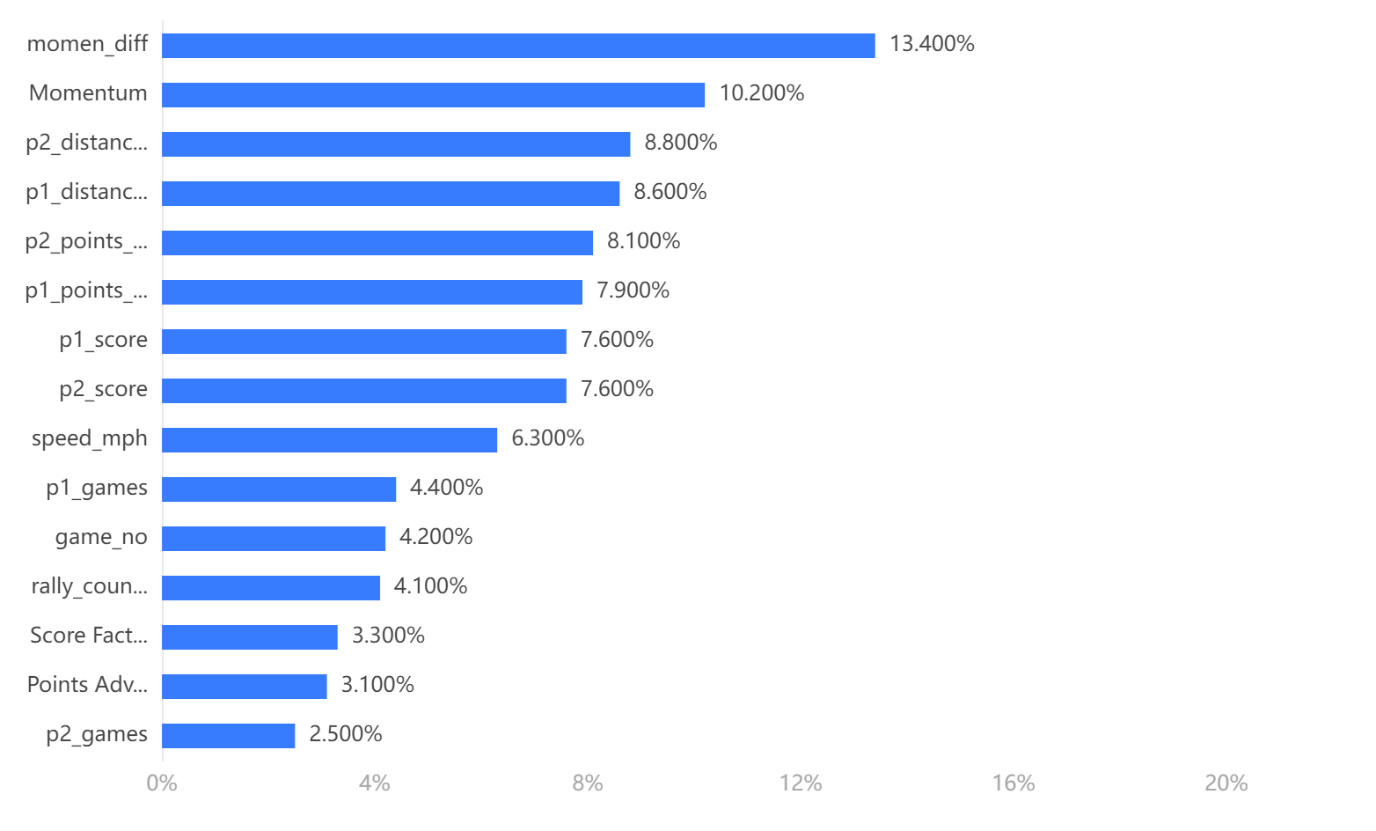


Figure x

Table x 模型评估

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 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：

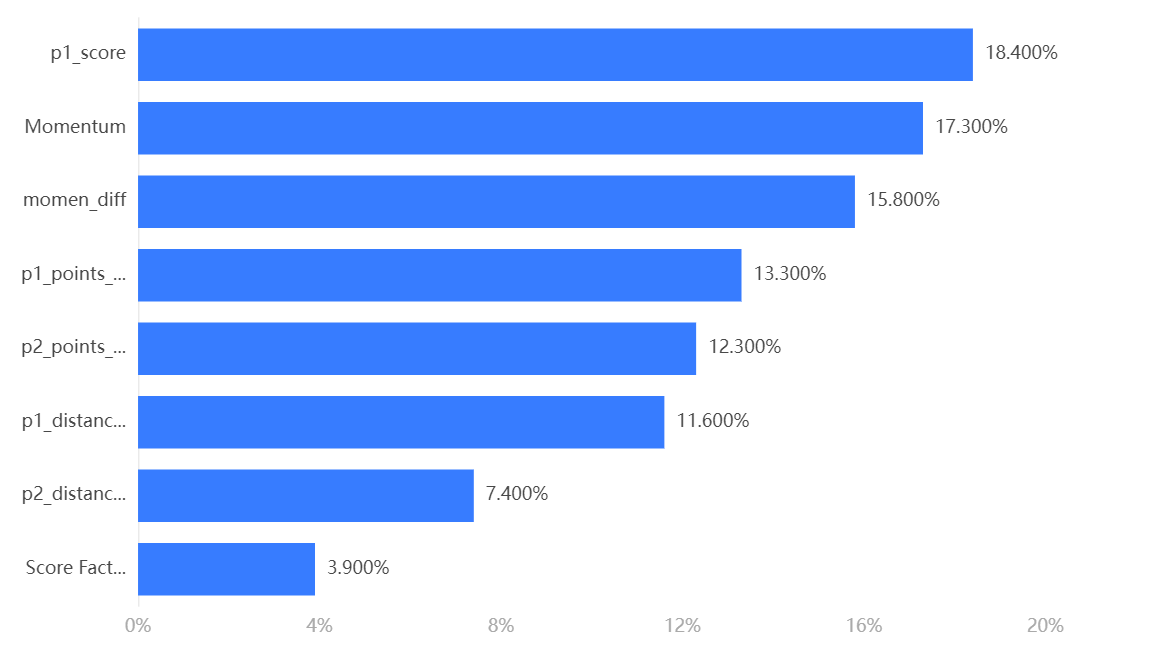


Figure x

Table x

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 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：

Table x

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 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：

Table x

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable Name | Sample Size | Median | 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：

# Sensitivity Analysis

模型的分析 ：在建模比赛中模型分析主要有两种，一个是灵敏度(性)分析，另一个是误差分析。灵敏度分析是研究与分析一个系统（或模型）的状态或输出变化对系统参数或周围条件变化的敏感程度的方法。其通用的步骤是：控制其他参数不变的情况下，改变模型中某个重要参数的值，然后观察模型的结果的变化情况。误差分析是指分析模型中的误差来源，或者估算模型中存在的误差，一般用于预测问题或者数值计算类问题。

模型的检验：模型检验可以分为两种，一种是使用模型之前应该进行的检验，例如层次分析法中一致性检验，灰色预测中的准指数规律的检验，这部分内容应该放在模型的建立部分；另一种是使用了模型后对模型的结果进行检验，数模中最常见的是稳定性检验，实际上这里的稳定性检验和前面的灵敏度分析非常类似，等会大家看到例子就明白了。

在美赛的写作中，写的最多的就是灵敏度分析（Sensitivity Analysis），因此这里我们的标题就直接取得是灵敏度分析；如果你既要写灵敏度分析，又要写误差分析（Error Analysis），那么你可以把标题改成： Sensitivity Analysis and Error Analysis

# 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。

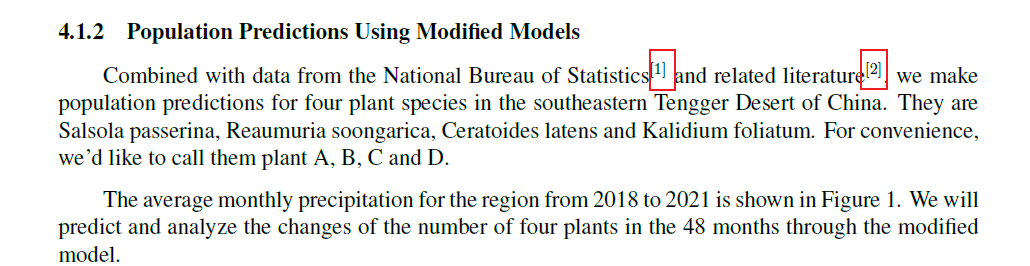
这个部分可以是论文中心思想的重申、研究结果或主要观点的归纳，也可以是某些启示性的解释或考虑。

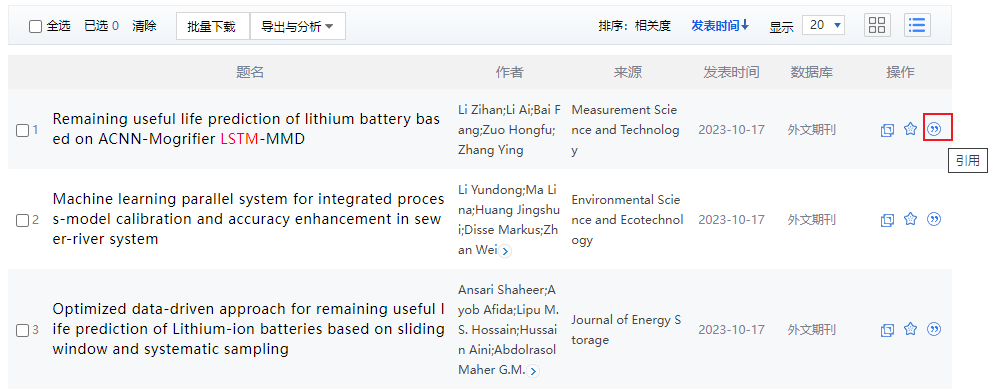
# References

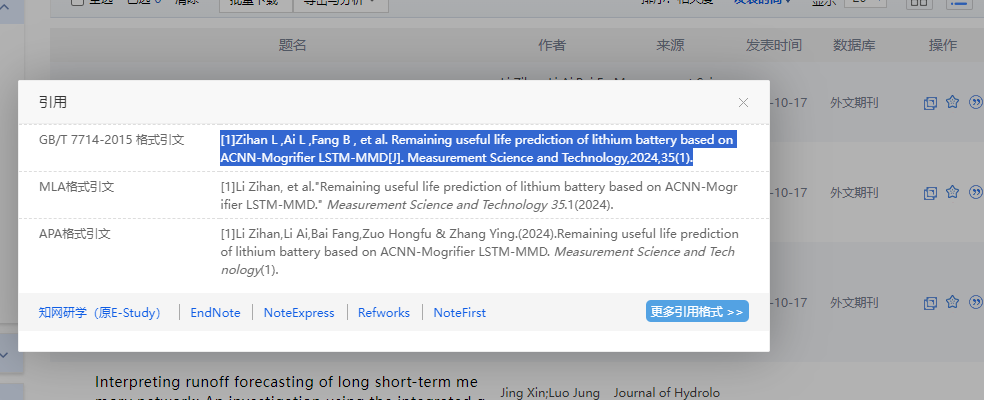
参考文献至少五六篇，引用中文文献记得翻译成英文。

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

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