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

**Momentum in Tennis**

**Summary**

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

Contents

[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 Momentum Model 6](#_Toc58786700)

[4.1 Data Description 6](#_Toc58786701)

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

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

[5 Proof of Correlation and Exclusion of Randomness 7](#_Toc58786704)

[5.1 Data Description 6](#_Toc58786701)

[5.2 Proof of Correlation 6](#_Toc58786702)

[5.3 Exclusion of Randomness 7](#_Toc58786703)

[6 Prediction of Other Matches and Generalization Evaluation 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

The tennis world witnessed a riveting match at the 2023 Wimbledon Men's Singles Final as young Spanish star Carlos Alcaraz produced a stunning performance to defeat 36-year-old legend Novak Djokovic strange. The match not only ended Djokovic's dominance at Wimbledon dating back to 2013, but also highlighted the critical importance of momentum changes in tennis. Momentum, as a vague concept in sports competitions, is usually considered to be the strong performance of a player or team in a game, but what exactly causes this change and how to quantify and understand this phenomenon is still a worthwhile question. Questions to study in depth.

In tennis matches, momentum changes often occur within a few minutes or even a few games, and this rapidly changing situation brings huge challenges to coaches, players and spectators. Although sports scientists have been working to unravel the mysteries of momentum changes, current research remains relatively limited. One of the main challenges of this research is to develop a comprehensive model that captures the key factors of momentum changes during a match and provides practical tactical advice on this basis.

Our research aims to build a reliable model through in-depth analysis of data from the Wimbledon 2023 men's match to reveal the patterns and influencing factors of momentum changes during the match. Through this model, we will be able to identify which players perform better and more consistently during the game, and how dominant they are. This will not only provide coaches with deeper insights, but will also hopefully provide athletes with more effective coping strategies, thereby increasing competition. Research in this area is critical to advancing the development of sports science and athlete training.

## Restatement of the Problem

**Task 1:** Develop a model that identifies which player performs better during a game and at what level by capturing the dynamics of scoring during a game. The model will be applied to one or more tennis matches, using visualization methods to present the dynamic process of the match.

**Task 2:** Evaluate the role of "momentum" in the game to confirm or refute the beliefs of a tennis coach who is skeptical about the role of "momentum." Utilize the developed model and associated metrics to parse out whether there is randomness in player fluctuations and success during games.

**Task 3:** Based on data from at least one match, develop a model capable of predicting fluctuations that occur during the match. Identify the most relevant factors and provide advice on how players should respond to possible fluctuations when competing against other players in new competitions.

**Task 4:** Test the developed model on other competitions and evaluate its ability to predict competition fluctuations. Identify situations where the model is underperforming and analyze factors that may need to be incorporated into future models.

**Task 5:** Check the generalization ability of the developed model to other games, tournaments, court surfaces and different sports. Write a final report summarizing the findings, providing a complete solution of no more than 25 pages, and providing recommendations to coaches to enable players to effectively respond to in-game events that impact flow.

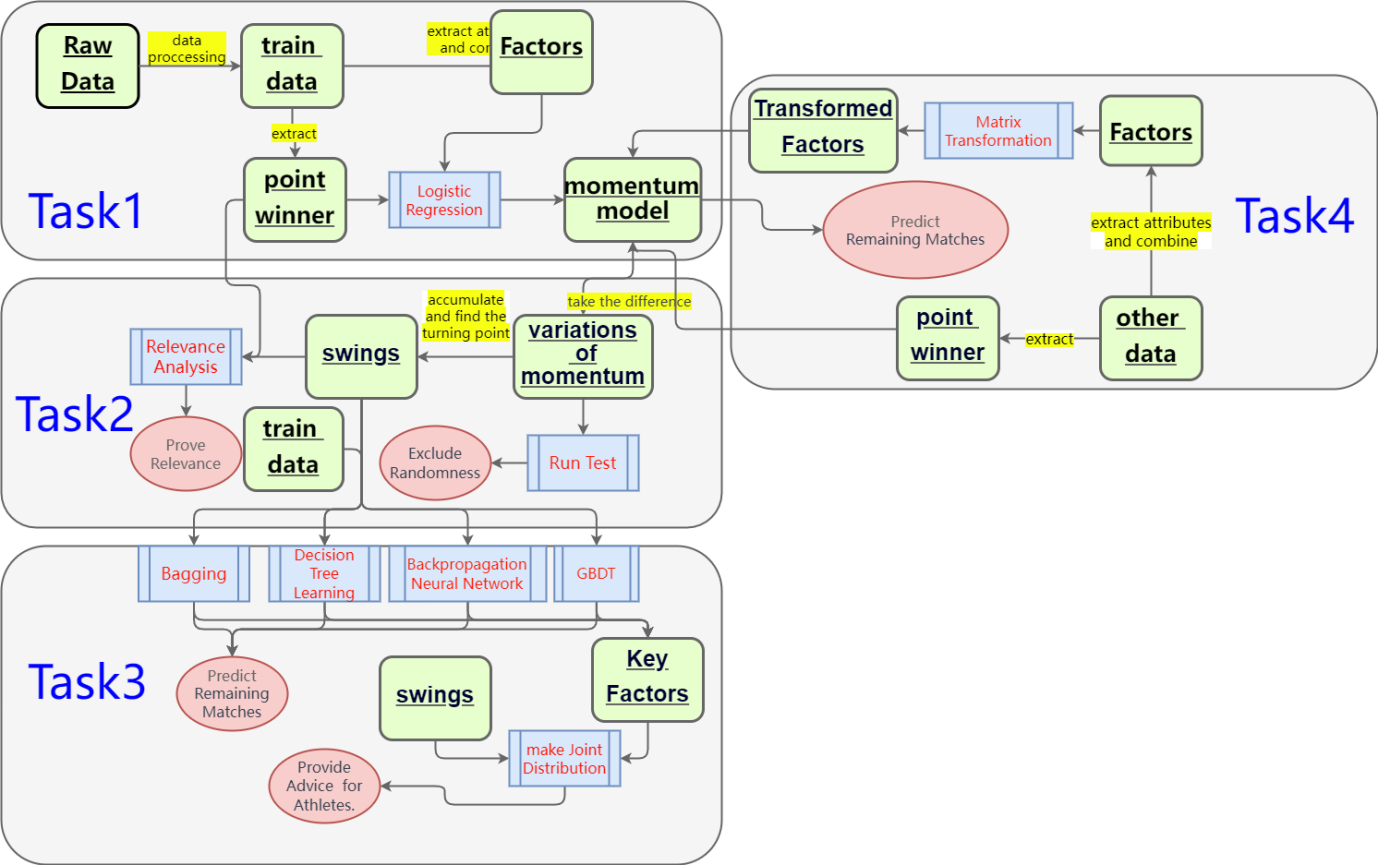
## Literature Review

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

# Assumptions and Justifications

In order to make our model clearer and easier to build, we set some basic assumption information. Based on the existing conditions, these assumptions are very important because it allows us to focus more on the objective data given in the question rather than the subjective factors of the remote players.

Assumptions 1: An athlete's momentum is only related to factors during the game, and has nothing to do with a series of factors outside the game time such as his or her own training, strength, coaching team, etc. This is because momentum is a quantity that cannot be easily measured. It is a virtual thing that affects the psychology of athletes in a short period of time. Therefore, it is reasonable to focus on the events that occur during the game to measure momentum. Since long-term or natural factors such as each player's age, gender, competition experience, physical fitness, etc. will change with the person, in order to ensure the robustness of the modeling, we choose to ignore these factors that have nothing to do with a game. .

Assumptions 2: We believe that the momentum of both athletes is the same at the beginning of the game. We admit that pre-game preparation and athletes' psychological construction are the keys that can often influence a game. However, in order to eliminate the possibility of different athletes' personal reasons affecting momentum analysis, it is more rigorous to put the momentum of both sides on the same starting line.

Assumptions 3: We believe that changes in the momentum of athletes during the game will be quickly reflected in the score of the game. A player's momentum will reflect the quality of his current state, which will then affect the score of the game. This will be significantly reflected in the establishment and solution of model one.

# 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

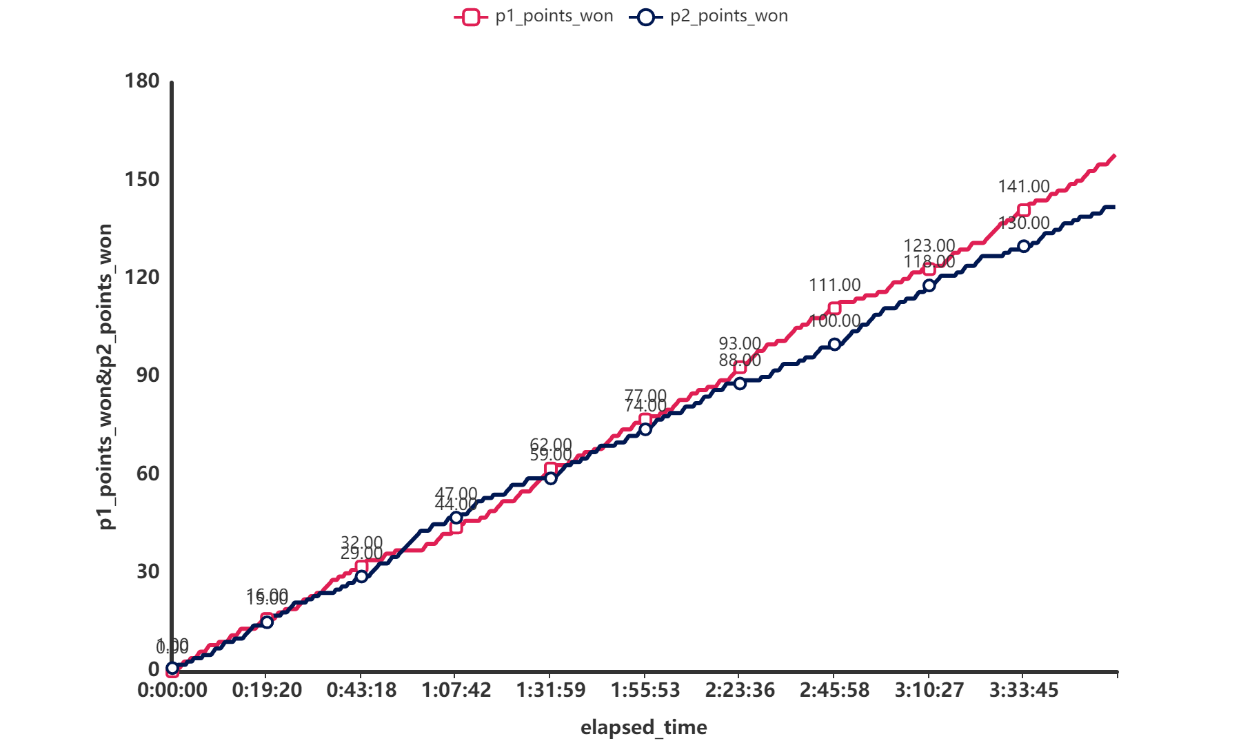
# Momentum Model

## Data Description

The data we use is the 2023 Wimbledon tournament statistics provided by the tournament group. It details various statistics and information related to the match. The table contains basic information about the two players, such as their names, and details of the game, such as the elapsed time, current number of sets, games, and scores. In addition, various special situations are recorded, such as unreturnable serves (aces), points won, double faults, unforced errors, points scored at the net, etc. This table also records the distance traveled by the players during the game, the number of ball exchanges, the speed of the serve, the direction and depth of the serve, and the depth of the returned ball. Such detailed data can help analyze player performance, formulate strategies, or provide viewers with a deeper understanding of the game.

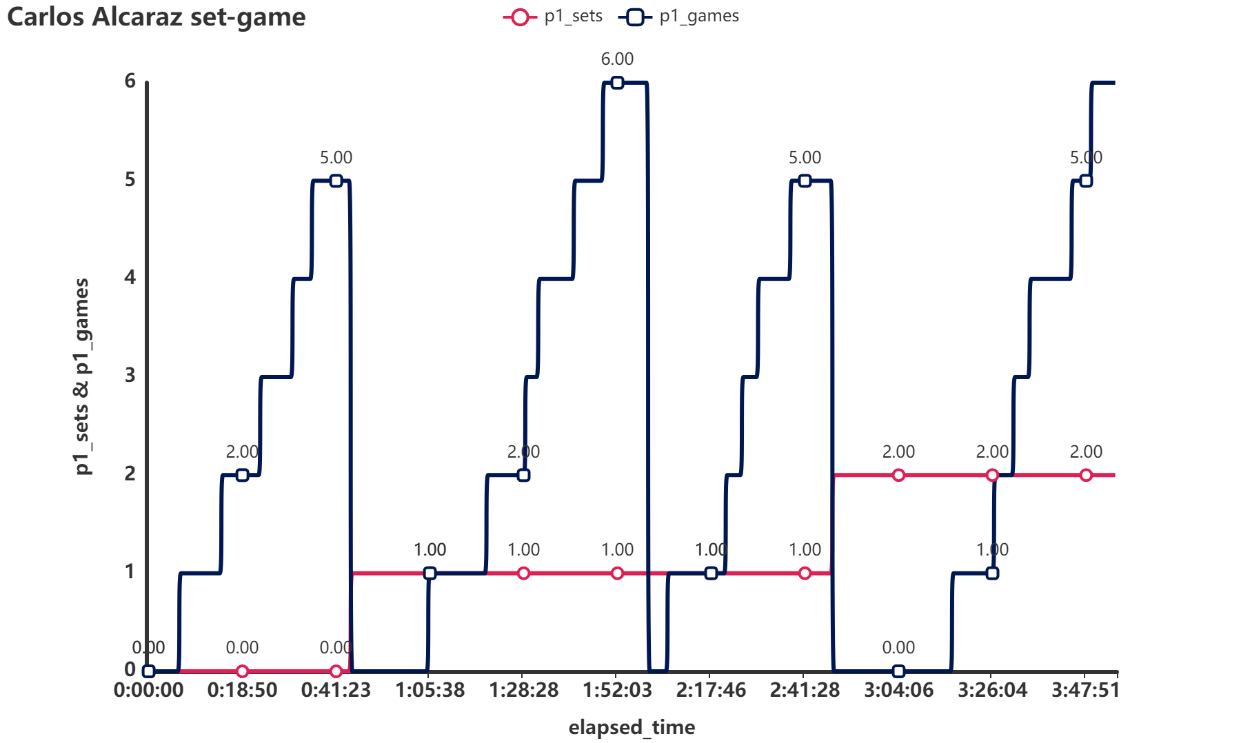
Since this is a general table that contains information about all games, in order to better model, we first extract the first game, the event number is 2023-wimbledon-1301. This game includes the final champion of this event, Carlos Alcaraz (Figure 2), so the data about him is the most comprehensive.

Carlos Alcaraz

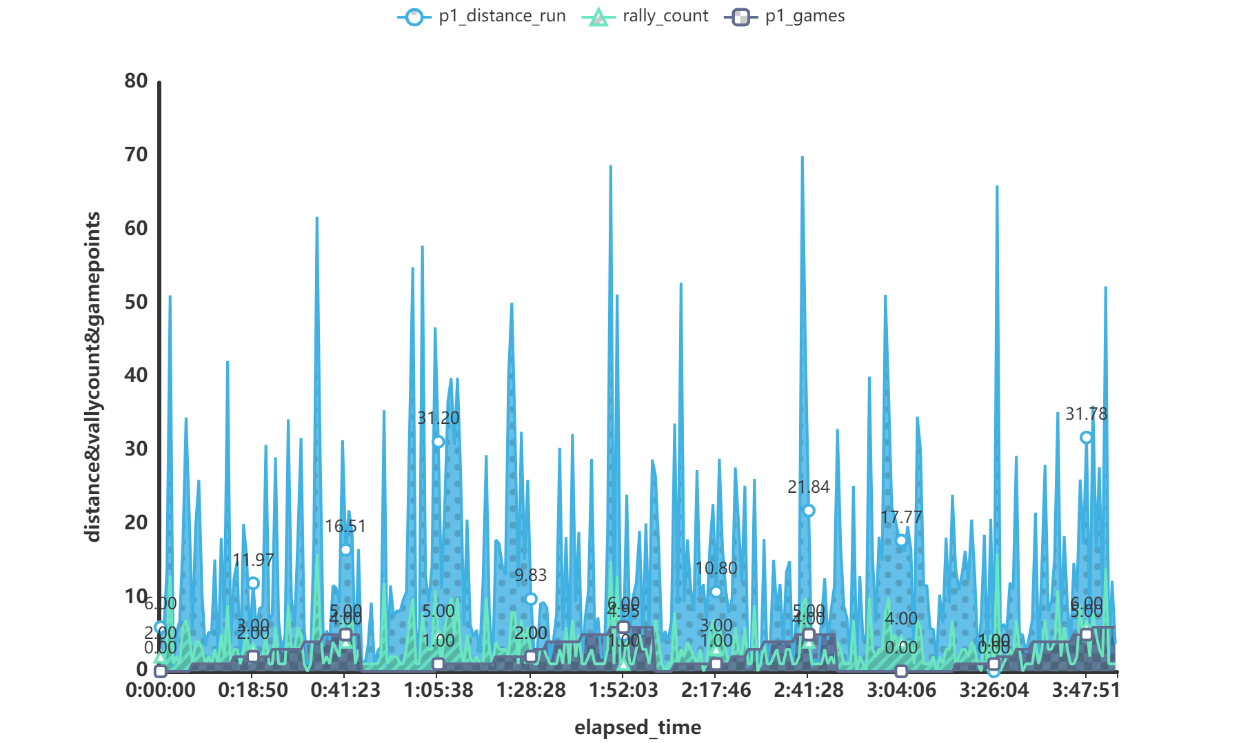


wimbledon-1301

Figure 3 shows the scores of both sides in this game. It can be seen from the picture that Carlos was still very anxious in the fight with his opponent in the first hour and a half. After that, Carlos gradually showed his strength and widened the point difference with his opponent, and finally won the game.



Carlos Alcaraz set-game

Figure 4 This picture shows the game and set points that Carlos scored during the game. It can be seen that in the first hour, Carlos won each game at an even time, which proved that Carlos's state was relatively stable in the early stages of the game, which helped him successfully win the first set. After entering the second set, the time it took him to win the game increased significantly, which showed that his opponent took the dominant position in the game at this time, so Carlos lost the second set. After that, Carlos' condition improved significantly, and he won the third set in almost half the time of the second set, and then won the game.

Run distance and rally count

Figure 5 is for Carlos’ running distance and number of rounds in each game. It can be seen that in a game, whenever it is close to the game point, the running distance and number of shots in a single round of both sides will increase significantly. This shows that at the critical stage near the game point, both players are in an extremely excited state. At this point their energy and concentration are at their peak. At the beginning of a game, the running and swing counts of both sides will return to a relatively low value, which shows that the state of both sides will be eased at the beginning of each game. According to the changing trends of these values, momentum may also change significantly. These are issues we need to explore later.

## The Establishment of Momentum Model

### Momentum Modeling Ideas

We first need to analyze what is the definition of "momentum". The dictionary definition of momentum is "the force or force gained by motion or a series of events." In sports, a team or player may feel like they have momentum or "power" during a game, but this phenomenon can be difficult to measure. Therefore, we have to start from two perspectives of the game, long-term factors and short-term factors.

Regarding long-term factors, we can think back to some of the ball games we have participated in before. If we have scored a lot of points, then the tension will be reduced in subsequent games and the mentality will be more relaxed, but similarly, the reduction of tension will also It may lead to a series of negative factors such as decreased concentration and weakened desire to win.

Regarding short-term factors, it is actually easy to imagine. If you score continuously, your self-confidence will inevitably increase and your momentum will rise even higher. But if you make consecutive mistakes or your opponent breaks your serve, then momentum will be suppressed.

### Momentum建模公式

The momentum model will be given below. Since the momentum model contains many attributes, we choose to introduce intermediate variables. F is the feature vector and c is the coefficient vector.

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

The F feature vector is composed of the following 8 attributes:

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

The c coefficient vector consists of eight coefficient scalars, which are obtained after fitting by logistic regression.

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

**Long-term factors:** obtained directly for global variables.

Score Factor represents the impact of the scored score on the momentum at this time. The specific calculation method is the sum of the difference between the game score and six times the difference in the set score.

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

Serve Advantage indicates which player the current serving game belongs to, because the winning probability of the party who owns the serving game will greatly increase, so if the party owns the serving game, we directly assign a value of 1. A game without a serve is assigned a value of -1.

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

**Short-term factors:** In order to better reflect the local impact of short-term factors, we constructed a time window with a length of 3, which means that the current row and the first three adjacent rows are evaluated as a subtable. Among them, Σ means summing the data in the subtable.

Break Factor represents the break factor that exists in this short-term window, and the numerator represents the difference in the number of breaks between the two sides in this window. The denominator represents the number of rounds required for the entire break process. The more rounds the break is attempted, the more the positive benefits of a successful break to momentum will be diluted.

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

Rally Factor represents the long-round ball factor. Within a given window, if you can score points in fewer shots, it will have more positive incentives for the player's momentum.

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

Points Advantages represent the goal difference within the time window

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

Service Factor represents the quality of the player’s serve within the time window

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

Unforced Errors represent the number of unforced errors encountered by athletes within the time window, which is a negative incentive for momentum.

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

Winners represents the number of winning goals completed by athletes within the time window, which is a positive incentive for momentum.

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

## The Solution of Momentum Model

### Coefficient Calculation with Logistic Regression

According to the third hypothesis of the model, the athlete's momentum will be reflected in the current game's score in time. Therefore, we target the scorer of each point and use logistic regression to fit the relationship between momentum and score, thereby obtaining the momentum expression. The corresponding coefficients of each of the 8 sub-factors in the formula.

* + - * 1. Logistic Regression Results

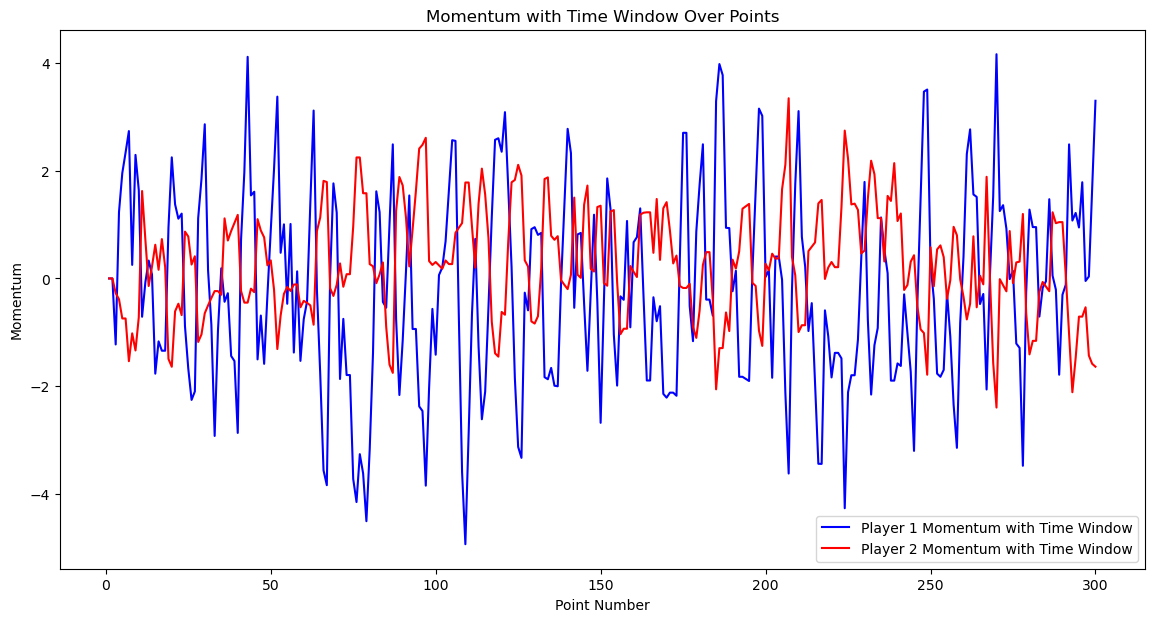
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 实验组=1.0 | 回归系数 | 标准误差 | Wald | P | OR | OR值95%置信区间 | |
| 上限 | 下限 |
| 常数 | 0.332 |  |  | NaN | 1.394 |  |  |
| Score Factor | 0.068 | 0.071 | 0.919 | 0.338 | 0.935 | 0.814 | 1.073 |
| Rally Factor | 0.217 |  |  | NaN | 1.242 |  |  |
| 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 | | | | | | | |

A series of results obtained by logistic regression have been presented in the table above. The regression coefficients corresponding to each factor have been presented in the table above. After multiplying these coefficients by the corresponding factors, the sum is the final result. The momentum corresponding to the athlete at each moment.

In particular, it can be found in the table that the significant P values ​​of Points Advantage, Serve Advantage and Unforced Errors are very low, showing significance at the level. Therefore they can have a significant impact on player wins.

* + - * 1. Logistic Regression Evaluation Metrics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Accuracy | Recall | Precision | F1 | AUC |
| 0.77 | 0.77 | 0.771 | 0.769 | 0.841 |

The accuracy of logistic regression is 77%, which is still very reliable. The comprehensive accuracy AUC is around 84%, which shows that this model can effectively determine the real-time momentum status of an athlete. By substituting these obtained coefficients into the model, you can get the momentum of both players on the field, and then you can judge who is in better condition at this moment. Take the 2023-wimbledon-1301 game as an example.

Momentum in 2023-wimbledon-1301

As shown in the figure, this is the momentum state change diagram of the players on both sides of the 2023-wimbledon-1301 match. The performance of the players at any time can be represented by momentum. For such a game, Carlos (Athlete 1)'s performance changed more drastically than his opponent's, especially in the first half of the schedule. Carlos' performance was once suppressed by his opponents, but in the second half of the schedule there was improvement, and the momentum curve also improved. Gradually gain the upper hand.

# Proof of Correlation and Exclusion of Randomness

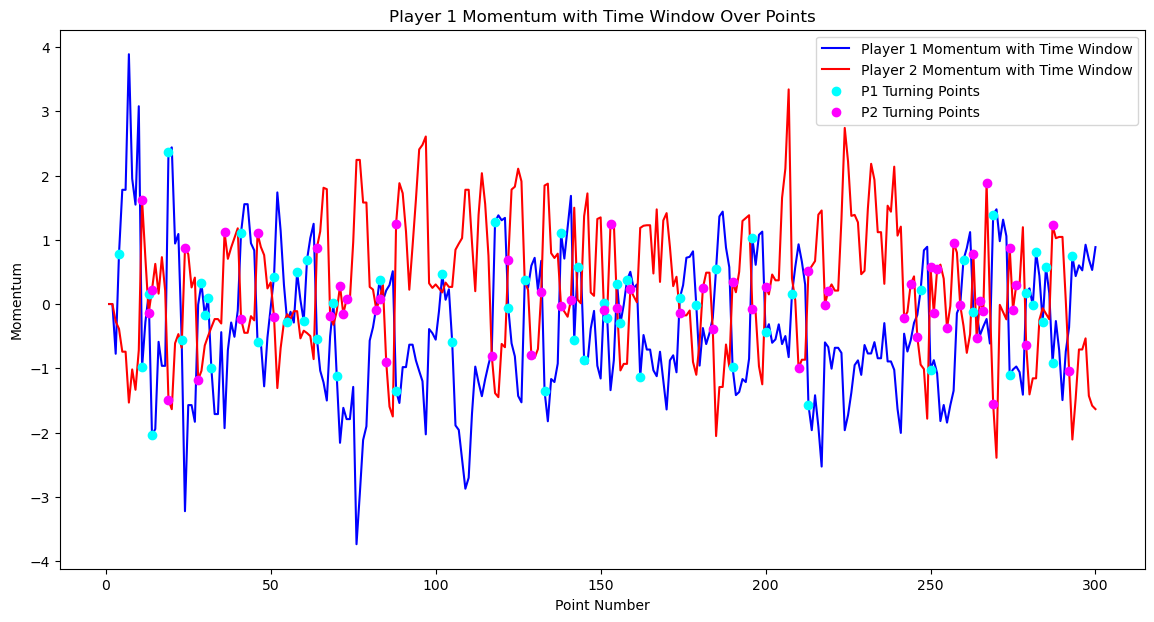
## Data Description

In this task, the data we use is based on the momentum model that has been established in the first task. We perform a differential operation on the momentum of each row of data to obtain a new column of data, which is recorded as momentum difference. Then the momentum difference value is accumulated to obtain a new value v\_i. For all i, if v\_i×v\_(i+1)<0, it means that the momentum changes from negative to positive or from positive to negative, then i is a fluctuation. point, mark k\_i=1, otherwise mark k\_i=0, and finally you can get the sequence k about the fluctuation, and its image is as shown in the figure below:



Fluctuation Plot

In order to display the fluctuation points more intuitively, we superimpose them into the momentum image and label them, as shown in the following figure:



Momentum Plot With Turning Points

## Proof of Correlation

Our goal in this section is to show that momentum is related to an athlete's score, thereby proving that momentum does play a role in the game. We used two different correlation testing methods for testing, namely calculating the Pearson correlation coefficient and the Spearman correlation coefficient. The two methods are briefly described below:

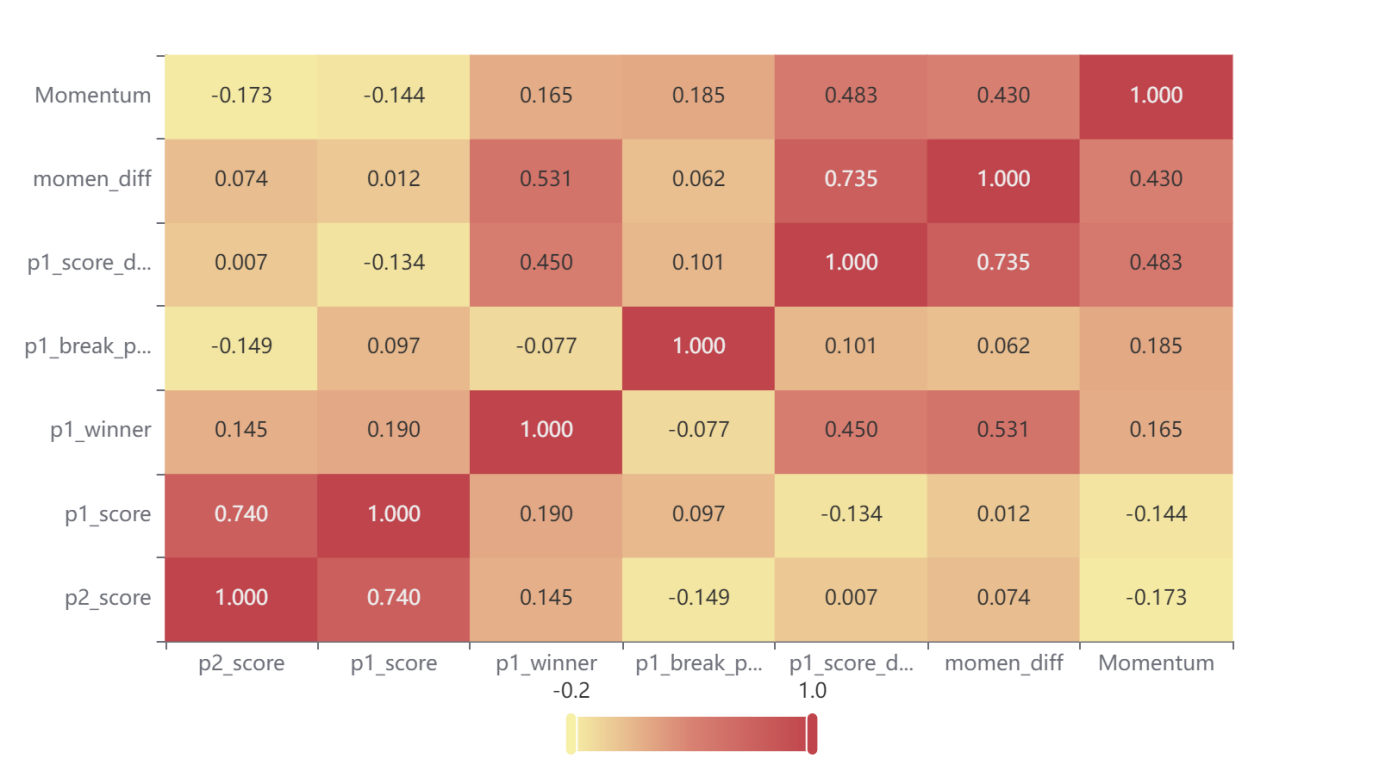
**Pearson** **Coefficient：**

Pearson coefficient is used to measure the linear relationship between two variables. For any two variables X and Y, the calculation formula of their Pearson correlation coefficient is as follows:

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

We can observe that the numerator is the value of the covariance of X and Y, and the denominator is the product of the standard deviations of X and Y. The calculated result range is [-1,1]. If the result approaches 1, it means that X and Y are positively correlated. If it approaches -1, it means that X and Y are negatively correlated. It is worth noting that if X and Y are independent, then =0. But on the contrary, if =0, we cannot judge whether X and Y are independent. For example, consider , E(XY)=E(X)=0, so cov(X,Y)=0, but obviously X and Y are not independent.

Back to this task, in order to prove that momentum directly affects an athlete's score, we can prove it more intuitively by calculating the correlation between the change in score (difference in score) and the change in momentum (difference in momentum). To determine the magnitude of the correlation, we can also use other data such as players' break points, players' wins and losses, etc., and calculate the Pearson coefficients between each other to compare horizontally whether the correlation between momentum changes and score changes is greater. Below is the heat map of the Pearson coefficient between different variables we calculated:



Pearson Correlation Coefficient Heat Map

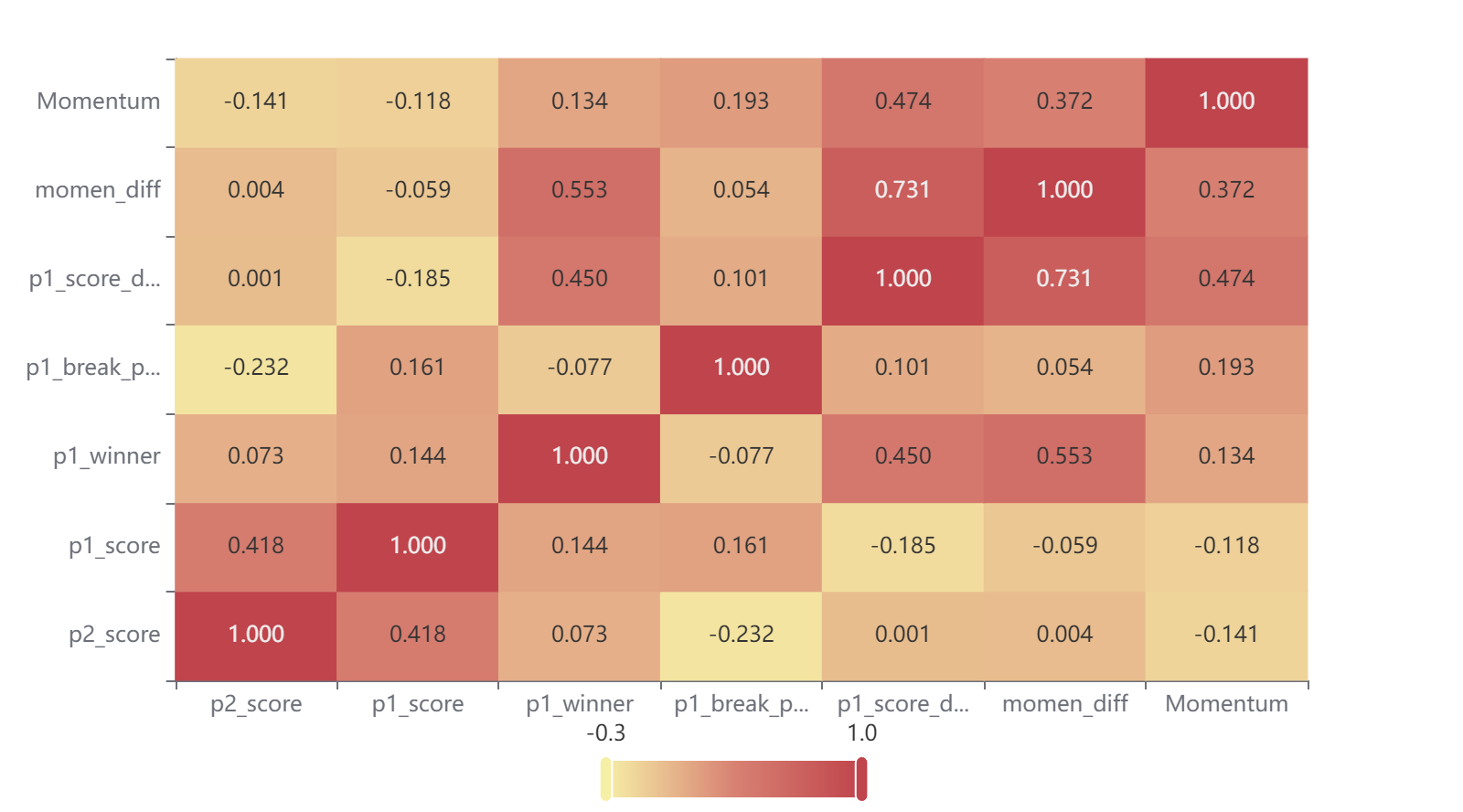
It can be seen that the Pearson coefficient of momentum change (momen\_diff) and score change (p1\_score\_diff) has reached 0.735, which is a large value. It shows that changes in momentum do directly affect changes in scores, and they have a positive impact. This further shows that momentum does play a certain role in the game.

**Spearman Coefficient：**

In addition to using the Pearson coefficient, we also used the Spearman coefficient to calculate the correlation. The Spearman correlation coefficient indicates the degree of association between the levels of two variables. For any two variables X and Y, assuming their levels are and respectively, the calculation formula of their Spearman coefficients is as follows:

|  |  |
| --- | --- |
|  | (13) |

Where is the level difference between X and Y, that is, . The rank of a number is the position of after the sequence V in which v\_i is located is arranged in order from small to large. The value range of the finally obtained coefficient ρ is also [-1,1], and like the Pearson coefficient, the closer the absolute value is to 1, the more relevant the two quantities are. Below is a heat map of Spearman coefficients between different variables:



Spearman's Rank Correlation Coefficient Heatmap

It can be seen that the Pearson coefficient of momentum change (momen\_diff) and score change (p1\_score\_diff) reached 0.731, which is still a large value.

The correlation coefficients obtained by both correlation tests are large, indicating that momentum plays an important role in the game.

## Exclusion of Randomness

In this section, our goal is to eliminate the theory that "swings in play and runs of success by one player are random". We need to explore whether the sequence S composed of the difference sequence of fluctuations (denoted as X) and the difference sequence of fractions (denoted as Y) is random. sequence. The method we use is the runs test. The specific process is as follows:

**First we make the hypothesis:**

H0: Sequence S is random.

**Then we need to calculate the test statistic:**

that is, to calculate the probability when the number of runs R=r, we need to consider the cases where r is an odd number and an even number respectively. Assume that the number of elements in the sequence X is and the number of elements in the sequence Y is .

When r is an odd number, the calculation formula is as follows:

|  |  |
| --- | --- |
|  | (14) |

When r is an even number, the calculation formula is as follows:

|  |  |
| --- | --- |
|  | (15) |

**Finally, we can make a decision:**

Assuming that the confidence level is 0.95 (95% confidence level), we can get P(R≤r)=0.05. At this time, if the number of runs in the sample is less than or equal to r, then a small probability event has occurred, and the null hypothesis can be rejected at the 5% significance level, that is, the sequence can be considered not random. On the contrary, if the number of runs is greater than r, the null hypothesis is not rejected and the two distributions are considered to be consistent.

In the case of large samples, according to the central limit theorem, the total number of runs obeys the normal distribution. At this time, we have

|  |  |  |  |
| --- | --- | --- | --- |
| |  |  | | --- | --- | |  | (16) | | (16) |

And Z~N(0,1), so

|  |  |
| --- | --- |
|  | (17) |

At this time, the rejection criterion is or

Therefore, using the above run-length test steps, we can finally get the results of the run test, as shown in the following table：

* + - * 1. Run Test Result

|  |  |  |  |
| --- | --- | --- | --- |
| name | Sample Size | z | P |
| moment\_diff | 59 | 3.024 | 0.002 |
| p1\_score\_diff | 59 | 2.082 | 0.037 |

From the table, we can see that P<0.05, indicating that the null hypothesis can be rejected at the 5% significance level, that is, the sequence S can be considered to be non-random, that is, the statement " swings in play and runs of success by one player are random" is wrong. The relationship between the two is not random.

# Swing Prediction and Factors Uncovering

Based on the model and data of Task 2 above, we obtained the potential fluctuation point sequence in the game, which has a significant impact on the trend of momentum and scoring. However, the vast number of model parameters poses a great obstacle to model calculation and data prediction, and useless parameters will damage the sensitivity and robustness of the model. To this end, we selected all the games in which Carlos Alcaraz was the first player, trained the existing model by dividing the training set and the test set (in units of games), and screened out the significance through models such as random forest bagging and BGDT. Characteristic Parameters. And test the remaining games to verify the correctness of the fluctuation prediction. At the same time, based on the differences in momentum fluctuations in previous games, significant changes in parameters are discovered at the fluctuation points, and targeted guidance and suggestions are given for games with new players. Here is our solution to the problem：

## 因素的显著重要性筛选

Before starting, we process the data, combine the parameters of the model into a matrix By inputting the data set Below are several model constructions we selected：

### 随机森林

Considering that the problem is nonlinear, locally uncorrelated, and has high dimensions, we choose random forest as the main solution. Random forest is a method based on ensemble learning that can more accurately capture complex nonlinear relationships by combining the prediction results of multiple decision trees. For problems with a large number of features, random forest can extract features that have a significant impact on Y through the combination of random selection of subsets and decision trees. In addition, random forest increases the resistance to noise (such as missing data, uneven data distribution, etc.) by sampling the data with replacement.

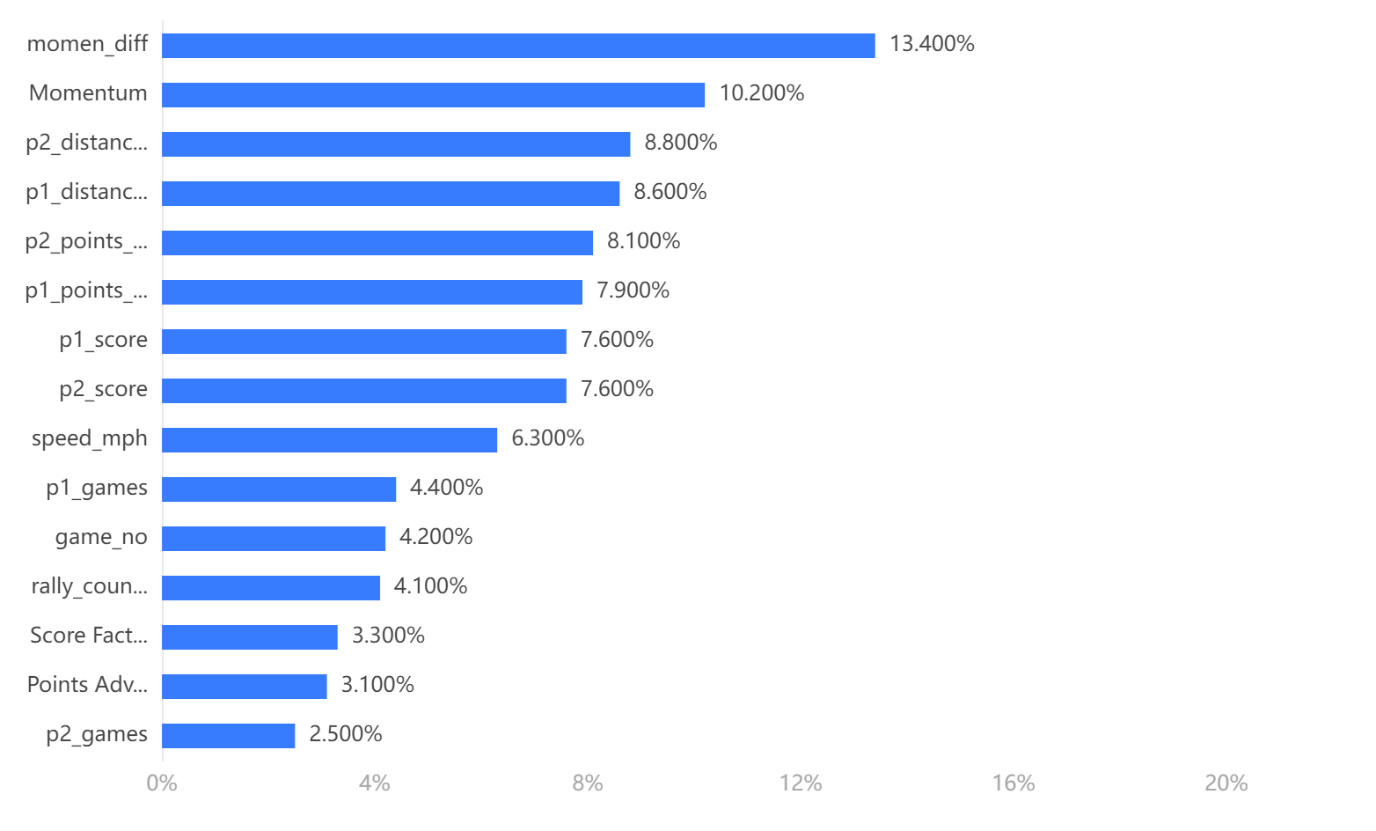
First, multiple decision tree models are generated by sampling the data with input X and label Y, and each decision tree model predicts the fluctuation point sequence. We assume that the data set under the decision tree is D, A represents the feature, and the formula of its Gini impurity is as follows:

|  |  |
| --- | --- |
|  | (18) |

The formula of Gini impurity for splitting under characteristic A is as follows：

|  |  |
| --- | --- |
|  | (19) |

After that, we select the feature with the smallest Gini coefficient as the criterion for node splitting and apply this process recursively. In the decision tree growing process of the random forest, an evaluation mechanism for the importance of the feature is built in during each tree splitting process. , the algorithm calculates the importance of each feature by measuring its splitting contribution. Therefore, the random forest algorithm continuously updates the list of importance parameters until the recursive stopping condition is met (here set to the maximum node depth of the tree). Finally, according to the obtained feature importance ranking, we select the top N features that have a significant impact on Y. These features will be used as input to the model to further optimize the accuracy of the model. The statistical ranking of feature importance and model evaluation results are as follows:：



Feature Importance Ranking

* + - * 1. Model evaluation

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

It can be seen from the statistical data that momentum difference and momentum sequence have the greatest impact on player fluctuations, followed by running distance and scoring. In addition, ball speed, number of wins, point advantage, etc. will all affect player fluctuations. Condition.

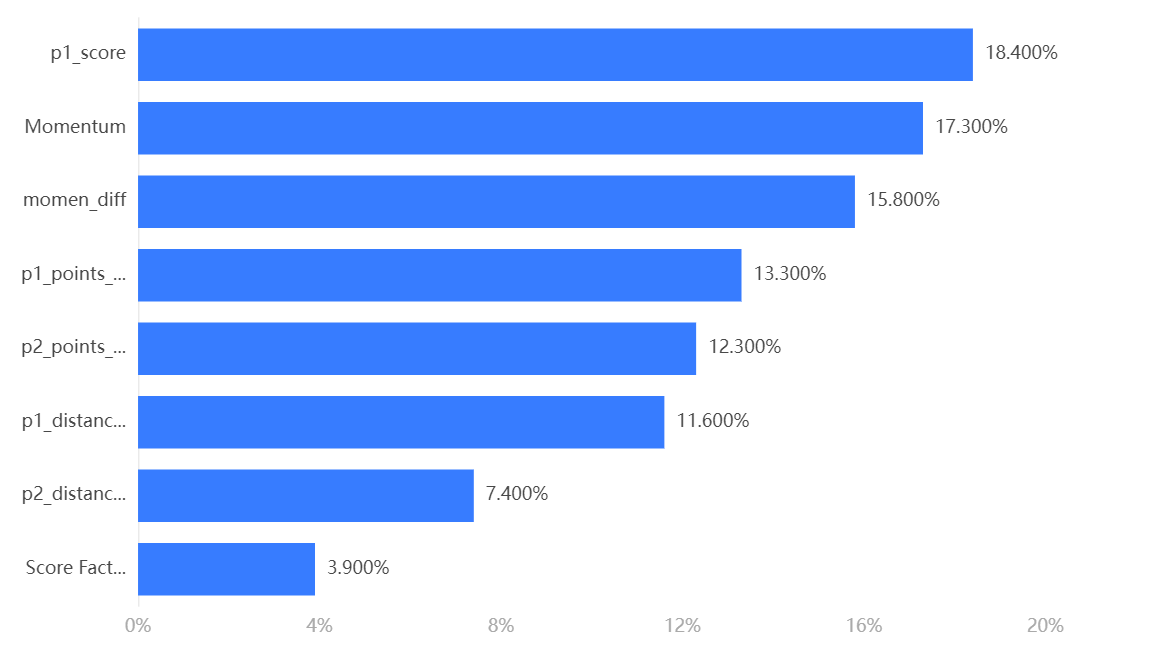
In the model evaluation results, the accuracy refers to the proportion of correctly predicted samples to the total samples. Recall rate, also known as sensitivity or true positive rate, measures the proportion of results that are actually positive samples that are correctly predicted as positive samples. The precision rate measures the proportion of results that are predicted to be positive samples and are actually positive samples. The higher the three, the better. The F1 value is the harmonic average of precision and recall, providing a balance that considers both precision and recall in evaluation. Therefore, overall, the evaluation effect of random forest is excellent, close to 90%, and has high credibility.

### GBDT梯度提升树

Considering the vast amount of data and its randomly distributed nature, we employed alternative methods for the screening and ranking of data feature importance to ensure the rigor of the results. Among these, Gradient Boosting Decision Trees (GBDT) and Random Forest, both belonging to ensemble learning, exhibit excellent noise resistance and iterative fitting effects. By continuously fitting the residuals (the difference between predicted and actual values) of the current model, the model performance is gradually improved. In the t-th round, the model's prediction can be expressed as the cumulative prediction of the current model and the sum of residuals from the previous t-1 rounds：

|  |  |
| --- | --- |
|  | (20) |

Each decision tree attempts to correct the prediction errors of the previous round of models. And during the iteration process, important features will be frequently selected, thus affecting the direction of the model. Therefore, similar to the random forest feature determination method, GBDT can also determine the split of the decision tree based on the Gini coefficient, and record the feature parameters that contribute significantly to the algorithm iteration, thereby screening out the corresponding Top N important features, and its running results with model evaluation as shown below：



Feature Importance Ranking

* + - * 1. Model evaluation

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

It can be seen from the analysis results that the important parameters in the front are consistent with the results obtained by random forest, that is, the momentum and momentum difference, the player's scoring situation and the running distance, indicating that these parameters are significant for the construction and optimization of the model. importance.

In model evaluation, it is obvious that all parameters are not as good as random forest, so its reliability in screening the importance of model parameters is relatively lower than the results of random forest.

### BP神经网络

Although GBDT and random forest have obvious advantages in nonlinear problems, BP neural network may have a more flexible fitting method and strong abstract data mining capabilities, so it can be used as a supplement to the above two models.

During the training process, the BP neural network calculates the gradient of the loss function for each weight through the back propagation algorithm, and then uses the gradient descent method to adjust the weight to reduce the prediction error of the model. In the tth round of training, the calculation formula is as follows：

|  |  |
| --- | --- |
|  | (21) |

The gradient is the rate of change of the loss function to the weight. This gradient formula reflects the sensitivity of the loss function to the weight, so that the relative importance of the weight can be judged. A larger absolute value gradient indicates that the model is sensitive to this parameter, so this parameter may have a larger contribution value, thereby filtering out important feature attributes. The evaluation results are as follows：

* + - * 1. Model evaluation

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

It can be seen that its evaluation performance is significantly worse than the previous two models, so the results of its importance analysis cannot be referred to.

### Overview

To sum up, we used three methods to evaluate the parameter importance of the model, namely random forest, GBDT and BP neural network. From the results of the model prediction evaluation, we can see that random forest has the best results, so we use the list of important parameters provided by random forest, namely momentum and momentum difference, running distance, scoring situation, ball speed, number of wins, points Advantages etc. We use these parameters with good importance as the parameter list of the model to filter out useless information, improve the robustness and prediction accuracy of the model, and make more targeted predictions.。

## Comparison of momentum difference with past games

In the previous question, we used the player Carlos Alcaraz’s 2023-Wimbledon-1301 games

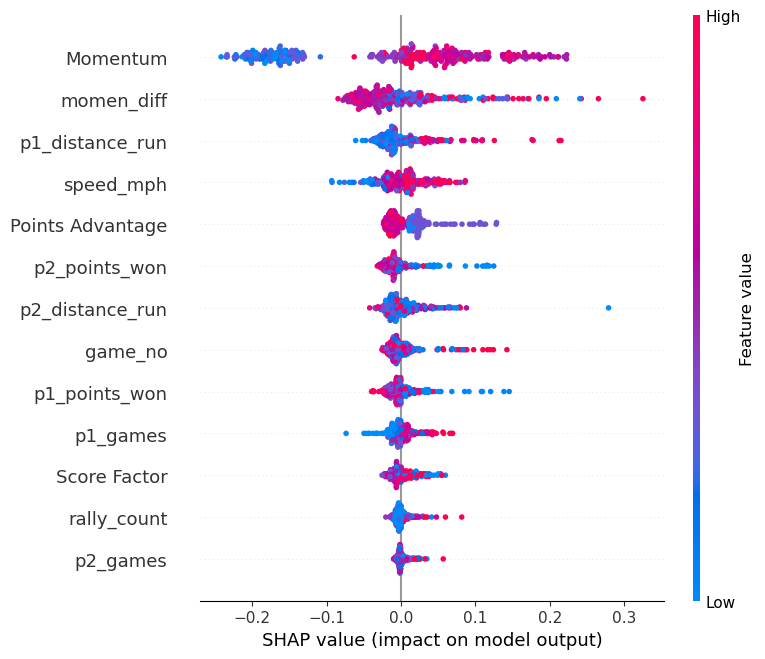
The model was trained on the data. In order to obtain more general conclusions and data, we conducted a difference analysis on the moment\_diff (i.e., the momentum difference) of the player's past five games. Since the sample size is relatively small and the overall data is consistent with the normal distribution The degree is not high enough, we used the multi-paired sample Friedman test method, and the test results are as follows：

* + - * 1. Friedman Test Result

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

From the Friedman test analysis results table, it can be seen that the significant P value is 0.957, so the statistical result is not significant, indicating that there is no significant difference between momen\_diff\_ \_1, momen\_diff\_ \_2, momen\_diff\_ \_3, momen\_diff\_ \_4, and momen\_diff\_ \_5; the magnitude of the difference is Cohen's The f value is: 0.002, a very small difference. Therefore, it can be shown that there is no significant difference in the volatility changes of the player's past games, indicating that past games can be used as a joint reference and opinions can be given.

## Suggestions and method given



Shap value for features

We will provide guiding recommendations for some of the attributes:- Rally\_count: - It can be observed from the graph that Carlos has an advantage in long rally counts. Considering Wimbledon's fast-paced nature, with over 50% of points occurring in the first 4 shots [1], it is advisable for the athlete to enhance their ability in short rallies.- Speed\_mph: - The graph indicates that higher ball speed correlates with a stronger ability for the athlete to change their momentum and gain an advantage in matches. Research suggests that the quality of shots in tennis is significantly influenced by the duration of recovery time[2]. Therefore, coaches are recommended to devise efficient physical recovery plans during game breaks to ensure the quality of shots in the next game.- P1\_distance\_run: - It is evident from the graph that a longer distance run in a single round can assist the athlete in changing their momentum. Particularly in their service games, reducing movement distance is associated with a decline in forehand proficiency [3]. Therefore, coaches are advised to encourage athletes to be more active in running during their service games to maintain higher forehand proficiency and establish a scoring advantage.- Score factor: - The graph indicates that if an athlete already has a scoring advantage, it positively influences their momentum. Research suggests that, among athletes with similar skill levels, the one who scores first has an advantage in the final victory[4]. Professional athletes understand how to leverage this advantage. It is recommended that coaches formulate strict strategies to capitalize on the advantage of scoring first, utilizing effective serving games. Alternatively, counter-strategies should be devised to address situations when the athlete is at a disadvantage.

# Prediction of Other Matches and Generalization Evaluation

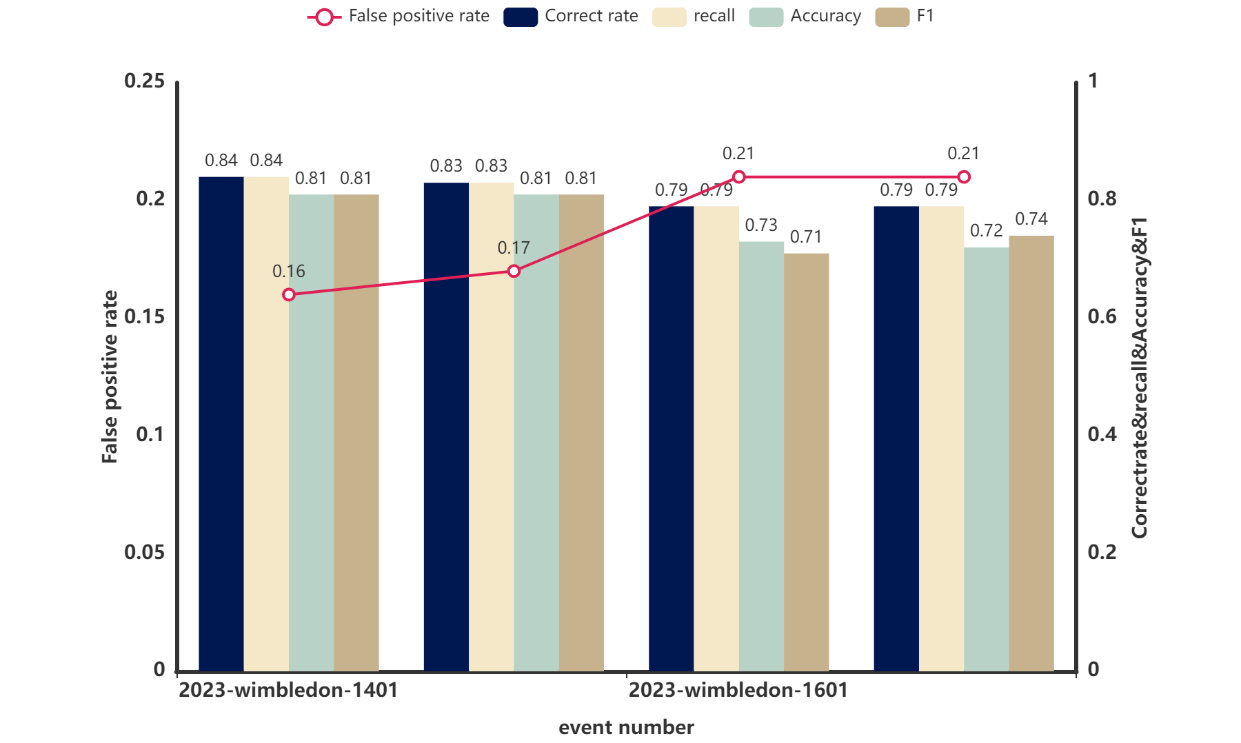
## Predictions for multiple games on different schedules

### Data extraction

In this part, we want to focus on testing the prediction effect of the model on other games of the same athlete. Therefore, we extracted the remaining four games of Carlos in this tournament as test objects, including the match with Djokovic. of the final. We need to evaluate from the results whether the model has learned the behavior pattern of the athlete itself, and we must also take into account the changes in the athlete's momentum during different stages of the championship.。

### 模型预测

The model still continues the random forest model used in the previous question to predict fluctuation points. The table below integrates the prediction accuracy and other results of these competitions.



Subsequent Match Prediction

* + - * 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 |

We can clearly find that the accuracy of the entire prediction will gradually decrease the longer the time between 2023-wimbledon-1301. This is due to the format of the tournament, because the strength of the opponents encountered will gradually increase each time they advance to the next round. And the differences in the psychological pressure that players put on themselves in different stages will lead to changes in the fluctuation pattern of their own momentum. However, the overall accuracy rate is still maintained at around 80%, which proves the effectiveness of this model in similar sports. Generalization ability is pretty good.

## 对羽毛球的泛化程度评估

To measure the quality of a model, we should not only consider the robustness and prediction accuracy of the model, but also examine the degree of generalization of the model to different propositions. Therefore, in order to measure the generalization degree of the model to the momentum and fluctuation degree of other different sports, we chose the sport of badminton for generalization measurement.

### 数据处理

For badminton match data, we obtained it from the open data platform Kaggle, the URL is as follows: https://www.kaggle.com/datasets/sanderp/badminton-bwf-world-tour. For the collected data set, we first need to clean the data and fill in all null values (NAN). The default filling is 0. And because the format of the downloaded data set is slightly damaged and mismatched, the format of all data in a single attribute column needs to be unified. At the same time, since huge data will introduce more noise, we selected representative stage competitions based on the schedule, such as the regular season, knockout rounds, semi-finals, and finals as the generalization test data set. In terms of format, we also matched the badminton match data set with the tennis match data set and standardized the format of the data.

### 参数映射矩阵变换

After obtaining the preprocessed data set, since the dimensions of the input parameter matrix do not match the established dimensions of the model, matrix transformation is required to match the game parameters in the badminton game with the model input. Assume that the number of parameters in the badminton data set is m, forming a matrix A of size (1, m), and the parameters in tennis are n. Note that the parameters here are not used to build the model at the beginning, but by analyzing the importance of their features. The selected Top N parameters form a matrix B of size (1,n). Therefore, our purpose is to find the transformation matrix C(m,n), so that the badminton parameter matrix A can be right-multiplied by the matrix C to obtain the transformed dimension matching matrix A', and then participate in the input and evaluation of the model.

The above model construction can be expressed by the following formula, and the matrix is solved by minimizing the error using the least squares method. Let the relationship between A and B be：

|  |  |
| --- | --- |
|  | (22) |

Where is the transpose of the transformation matrix to be obtained, and E is the error representation. We can iteratively solve by minimizing the error, that is, using the least squares method, that is

|  |  |
| --- | --- |
|  | (23) |

represents the L2 norm, which is the Euclidean distance. The solution to the above problem can be expressed as：

|  |  |
| --- | --- |
|  | (24) |

The premise for the above formula to be established is that there is an inverse matrix for the multiplication of the transpose of B and B. If it is irreversible, a regularization method such as ridge regression needs to be used to solve it. Finally, we obtain the transpose of matrix C, and only need to perform another transposition to obtain the transformation matrix C. Then for the obtained badminton parameter matrix A, A’ can be obtained by transforming AC to match the tennis model parameter dimension.。

### 模型的输入与泛化评估

Using the transformation matrix obtained above, we input the parameters of the badminton shuttlecock into the model established in the second question. Prior to inputting the parameters, we need to optimize the model architecture. Based on the feature importance parameter list obtained from the third task, we replace the model's parameter list with the selected parameters exhibiting good feature importance. This further filters out irrelevant information and strengthens the predictive capability of the model.

After optimizing the model, we can proceed with inputting and solving. Firstly, the transformed badminton shuttlecock parameter list is used as input parameters, and the model's output is obtained, including the momentum, momentum difference, and fluctuation points of the badminton game. The transformed parameter input serves as the X matrix, and the fluctuation points serve as the label Y. We utilize the predictive model to forecast the label Y, and evaluate the model's generalization performance through accuracy. The results are as follows.：

* + - * 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 |

From Table 9, we can observe that the accuracy, recall, and precision are consistently stable at around 74%, indicating a good level of generalization across different types of sports. The lower generalization accuracy may be attributed to precision loss and errors during the matrix transformation from m to n dimensions, increasing the probability of misclassification. Consequently, the actual accuracy is not very high. However, overall, the model maintains a prediction accuracy of over 70% when facing situations with rules having certain similarities and substantial inconsistency in the data. This suggests a reasonably good level of generalization.

# 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

⚫ 我们的模型使用准确的数据，并且选用了多场比赛进行训练与拟合，保证原始数据的严谨性。

⚫ 我们的模型在训练参数的过程中引入了机器学习中的逻辑回归方法，通过迭代训练调整最优的权重，而非用人工进行随机指定。

⚫ 模型训练出的权重系数中的负权重成功表征了选手在领先时的心理压力，兼顾了数据理论的严谨性和人文关照。

⚫ 我们在拟合模型参数的时候没有使用简单相加的方法将所有的因子线性加权，而是综合考虑了短期因素和长期因素，并通过时间窗口截取一段时间评估选手的势头表现。

⚫ 我们通过GBDT、随机森林、BP神经网络方法对模型参数重要特征进行筛选，并且通过三种方法的预测准确率选用前N个最重要特征集，从而优化模型结构，去除冗余特性

⚫ 在模型泛化过程中，我们通过矩阵变换将维度不同的参数列表进行匹配，完成了不同球类运动比赛数据处理的接口，实现了良好的泛化性和实用性

## Weaknesses

⚫ 没有过多考虑选手自身的实力、教练团队、天气、场地等影响，实际效果可能并不理想

⚫ 模型对于场内场外随机因素的考量不够全面，对于数据的偏差处理也只是简单的丢弃，可能会无法通过势头预测比赛中的较小概率事件

## Further Discussion

⚫ 模型在泛化程度上做的并不够好，相对于在网球中的预测成功率低了将近15%。因此，可以在原先矩阵变换的基础上预先对数据进行相关性匹配，通过人工标签的方式指定两种不同球类的比赛数据的相关联性，这样可以有效减少矩阵变换时数据的伸缩带来的精度损失。但是这样会耗费大量的人力物力，以我们目前的时间和人力是难以胜任的，争取未来有机会得以实现，提供更好的变换接口。

# 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

Carboch J, Siman J, Sklenarik M, et al. Match characteristics and rally pace of male tennis matches in three grand SLAM tournaments[J]. Physical Activity Review, 2019 (7): 49-56.

Ferrauti A, Pluim B M, Weber K. The effect of recovery duration on running speed and stroke quality during intermittent training drills in elite tennis players[J]. Journal of sports sciences, 2001, 19(4): 235-242.

Pereira T J C, Nakamura F Y, de Jesus M T, et al. Analysis of the distances covered and technical actions performed by professional tennis players during official matches[J]. Journal of sports sciences, 2017, 35(4): 361-368.

Apesteguia J, Palacios-Huerta I. Psychological pressure in competitive environments: Evidence from a randomized natural experiment[J]. American Economic Review, 2010, 100(5): 2548-2564.

# 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

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