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

Momentum in Tennis

Summary

In the 2023 Wimbledon men's singles final, Carlos Alcaraz beat Novak Djokovic, marking a shift in Wimbledon's dynamics. The study examines the 2023 Wimbledon men's data to identify momentum change patterns and factors, offering coaches and athletes strategies to enhance competitiveness.

In Task 1, we developed a momentum model based on eight factors—score, serve, break, round, goal difference, service quality, unforced errors, and winning shots—using logistic regression to determine each factor's impact. Points Advantage, Serve Advantage, and Unforced Errors significantly affect winning. The model proved reliable for real-time momentum assessment, with a 77% accuracy rate and an 84% AUC. We also analyzed the 2023 Wimbledon matches, showcasing the model's application and explanatory power through athletes' momentum state change diagrams.

In Task 2, we use the data of the previously established momentum model to obtain the momentum difference through difference and accumulation operations. Identify swing points in swing charts that mark turning points in momentum changes. The relationship between momentum changes and score changes was tested through the Pearson and Spearman correlation coefficients, which proved that there is a significant positive correlation between the two. In addition, the randomness between volatility and success was explored through the run test, and the results showed that the relationship between volatility and success was not random.

In Task 3, we refined our analysis with Carlos Alcaraz's game data, employing models like random forest, GBDT, and BP neural network for parameter optimization to enhance model sensitivity and robustness. Random forest excelled in parameter importance evaluation, guiding our selection. Analyzing Alcaraz's momentum variance over five games with the Friedman test revealed consistent performance fluctuations, offering joint reference for model advice. Based on key parameters, we suggested improvements in short rounds, physical recovery during game breaks, and increased activity in service games. These strategies are tailored to boost competitive performance, derived from model insights.

In Task 4, using Carlos Alcaraz's data, we verified our model's predictions across different matches. We observed a slight decrease in accuracy with longer game durations, yet it remained around 80%. Extending the model to badminton, with matrix transformation, it maintained over 70% accuracy. This underscores the model's strong predictive and generalization abilities across sports, offering robust support for forecasting athlete performance trends, highlighting its practical application potential.

In conclusion, our research offers a thorough method for analyzing and predicting momentum in sports competitions, providing valuable insights and practical advice for coaches and athletes. We plan to enhance prediction accuracy and utility by optimizing the model and broadening the dataset. Our goal is to support sports science research and practical training, ultimately helping athletes excel in their competitions.

Keywords: Momentum; Logistic Regression; Model; Predict

Contents

1 Introduction

1.1 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 strangely. 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 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.

1.2 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.

1.3 Literature Review

The role of momentum in tennis matches is increasingly considered a key factor affecting match outcomes. Research shows that the effects of momentum vary among players, revealing new challenges for coaches and psychological counseling in leveraging momentum. Especially in US Open men's singles, breaking serve at crucial moments significantly increases the probability of holding serve in subsequent games, highlighting the role of momentum at critical points in the match.[1]

Further analysis has found that professional tennis players can benefit from momentum when they control the match; however, once control is lost, anti-momentum significantly reduces their chances of winning the next set, indicating a bidirectional impact of momentum and anti-momentum on match strategy.[2]

On the psychological level, early success can boost the winner's confidence in victory but also place greater psychological pressure on the leader. At the same time, it puts the loser at a psychological disadvantage, emphasizing the role of psychological momentum in shaping the course of the match.[3]

These studies collectively highlight the importance of momentum in tennis matches. However, these papers lack strong statistical theoretical support. We have applied more statistical mathematics knowledge and methods, considering events with higher granularity. This provides valuable insights for coaches and players in training and match strategy, prompting them to consider how to effectively leverage or counteract the effects of momentum to optimize performance.

1.4 Our Work

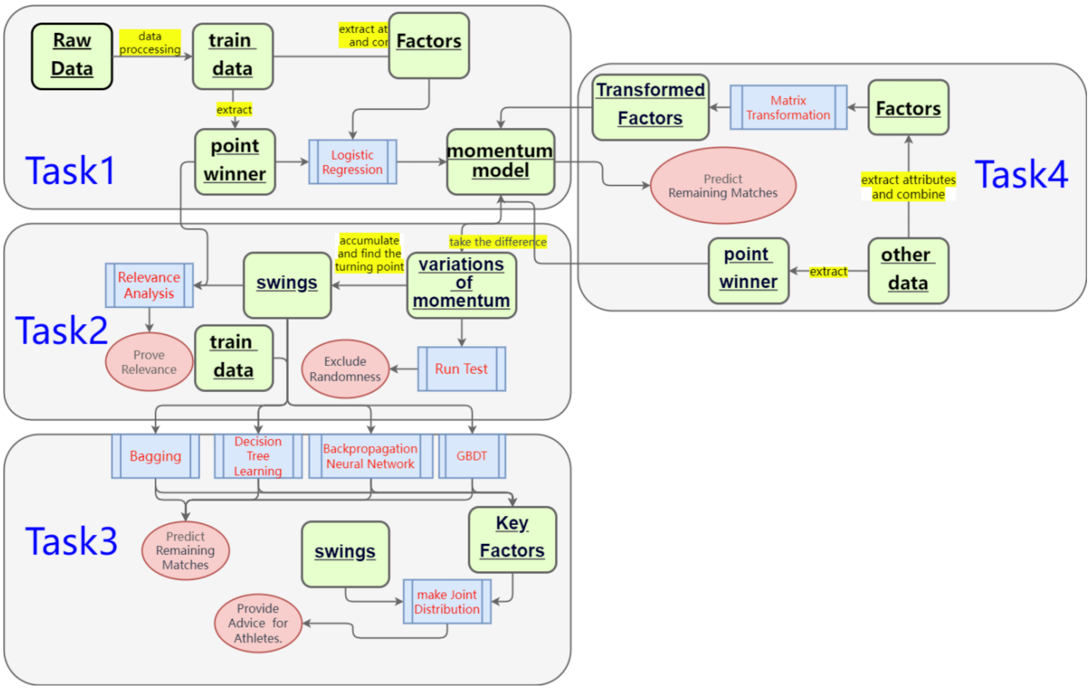

Figure 1: Our Work

2 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 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 points. 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.

3 Notations

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

Table 1: Notations used in this paper

Symbol	Description
Score_diff	The Difference of Score
momen_diff	The Difference of Momentum
Fluc_point	The Point Where the Momentum Reverses

4 Momentum Model

4.1 Data Description

The data is from the 2023 Wimbledon tournament, detailing player names, match statistics like duration, scores, and specific metrics including aces, double faults, unforced errors, net points, player movement, ball exchanges, serve speed, and serve direction. This information aids in analyzing performance, strategizing, and enhancing viewer understanding.

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

Figure 2: Carlos Alcaraz

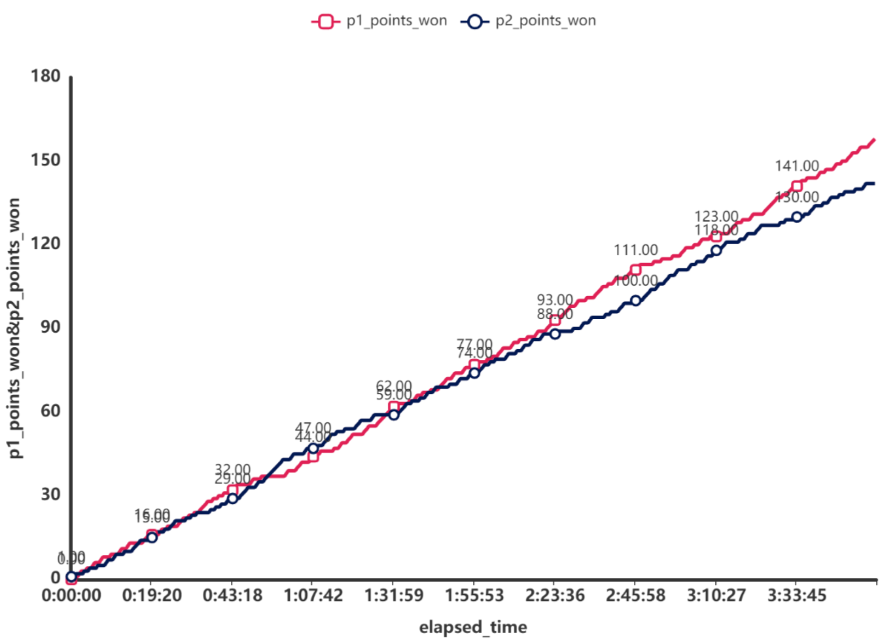

Figure 3: wimbledon-1301

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.

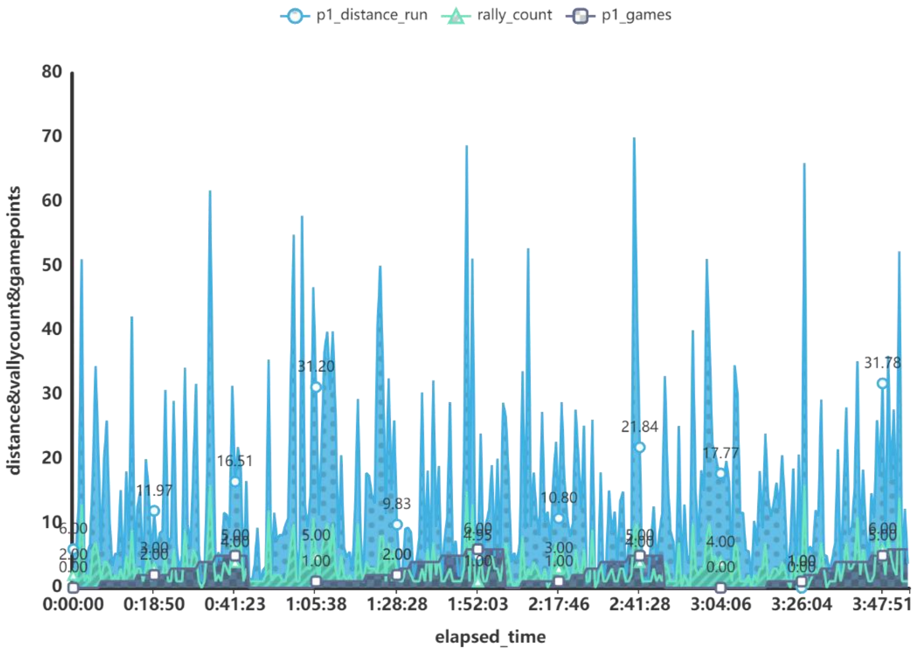

Figure 4: Run distance and rally count

Figure 4 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.

4.2 The Establishment of Momentum Model

4.2.1 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.

4.2.2 Momentum Model

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

$$\text{Momentum} = c \cdot \text{F} \tag{1}$$

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

$$F = \begin{bmatrix} Score\ Factor \\ Service\ Factor \\ Break\ Factor \\ Rally\ Factor \\ Points\ Advantage \\ Serve\ Advantage \\ Unforced\ Errors \\ Winners \end{bmatrix} \tag{2}$$

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

$$c = [c_0 \quad c_1 \quad c_2 \quad c_3 \quad c_4 \quad c_5 \quad c_6 \quad c_7] \tag{3}$$

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.

$$\text{Score Factor} = p1\_games - p2\_games + (p1\_sets - p2\_sets) \cdot 6 \tag{4}$$

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.

$$\text{Serve Advantage} = \text{server where server is 1 if player 1 is serving, else } -1 \tag{5}$$

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.

$$\text{Break Factor} = \frac{\sum(p1\_break\_pt\_won - p1\_break\_pt\_missed) - \sum(p2\_break\_pt\_won - p2\_break\_pt\_missed)}{\sum(p2\_break\_pt + p1\_break\_pt) + \epsilon} \tag{6}$$

Rally Factor represents the long-round ball factor.

$$\text{Rally Factor} = \frac{\sum p1\_long\_rallies\_won}{\sum(\text{if } rally\_count > \text{average\_rally\_count then 1 else 0}) + \epsilon}) \tag{7}$$

Points Advantages represent the goal difference within the time window

$$\begin{aligned}\text{Points Advantage} \\ &= \sum(\text{if } point\_victor == 1 \text{ then 1 else 0}) \\ &- \sum(\text{if } point\_victor == 2 \text{ then 1 else 0})\end{aligned} \tag{8}$$

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

$$\text{Service Factor} = \sum(p1\_ace - p1\_double\_fault) - \sum(p2\_ace - p2\_double\_fault) \tag{9}$$

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

$$\text{Unforced Errors} = -\sum p1\_unf\_err \tag{10}$$

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

$$\text{Winners} = \sum p1\_winner \tag{11}$$

4.3 The Solution of Momentum Model

4.3.1 Coefficient Calculation with Logistic Regression

According to Hypothesis Three, an athlete's momentum will promptly reflect in the current game's scoring situation. Therefore, we target each point's scorer and employ logistic regression to model the relationship between momentum and scoring. This process helps us derive the coefficients corresponding to the eight sub-factors in the momentum expression, each associated with its impact on scoring.

Table 2: Logistic Regression Results

feature	coefficients	SD	Wald	P	OR	OR−95%CI	
						Upper limit	Lower limit
Const	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

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.

Table 3: 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.

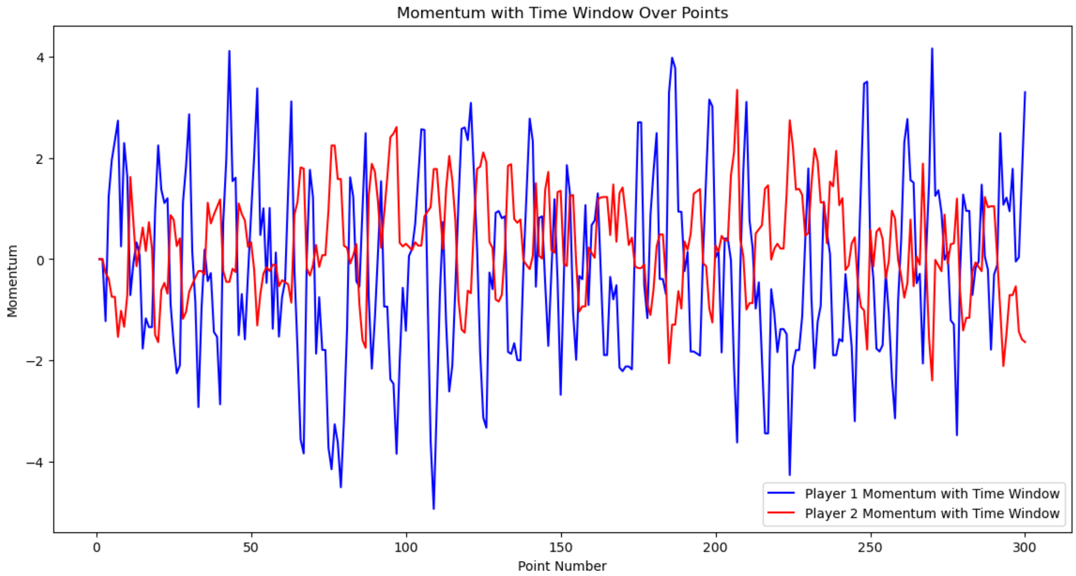

Figure 5: Momentum in 2023-wimbledon-1301

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.

5 Proof of Correlation and Exclusion of Randomness

5.1 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 $v_i \times v_{(i+1)} < 0$, 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:

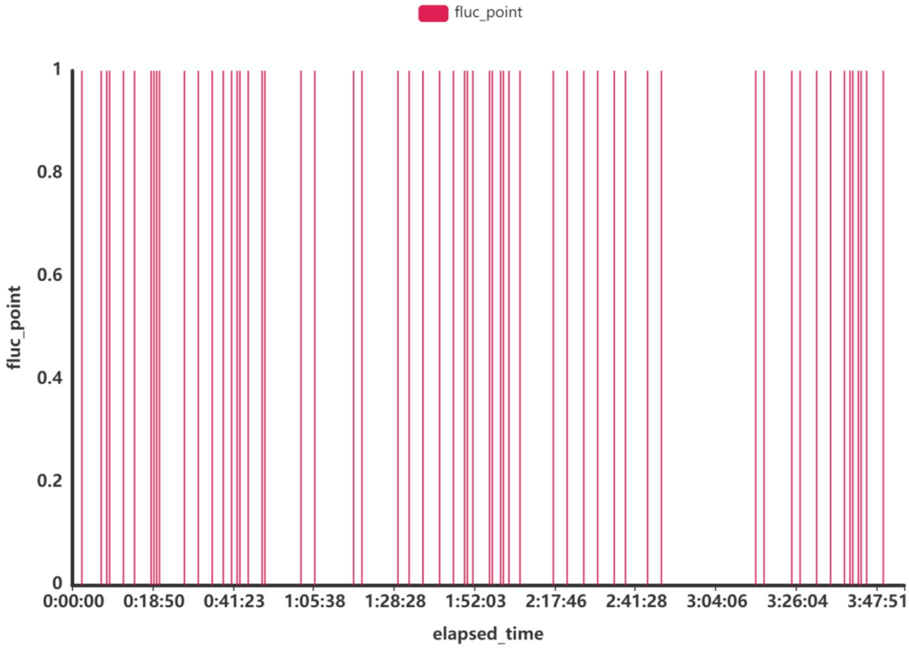

Figure 6: Fluctuation Plot

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:

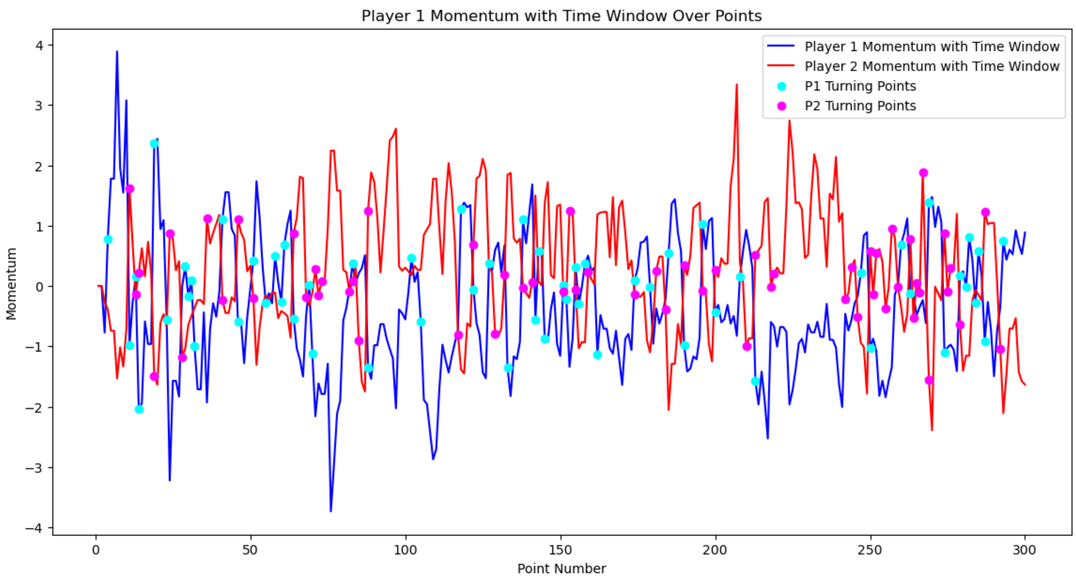

Figure 7: Momentum Plot With Turning Points

5.2 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:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_x \sigma_y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_x \sigma_y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}} \tag{12}$$

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 $\rho_{X,Y}$=0. But on the contrary, if $\rho_{X,Y}$=0, we cannot judge whether X and Y are independent. For example, consider $Y = X^2$, 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:

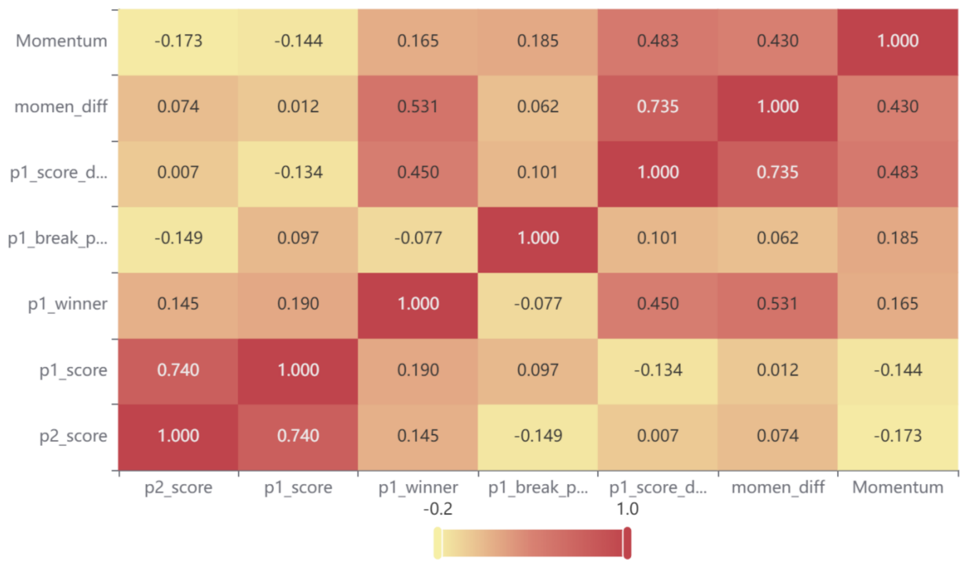

Figure 8: Pearson Correlation Coefficient Heat Map

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 make an impact on the 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 R_x and R_y respectively, the calculation formula of their Spearman coefficients is as follows:

$$\rho = \frac{1-6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{13}$$

Where d_i is the level difference between X and Y, that is, $d_i = R_x - R_y$. The rank R_{v_i} of a number v_i is the position of v_i 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:

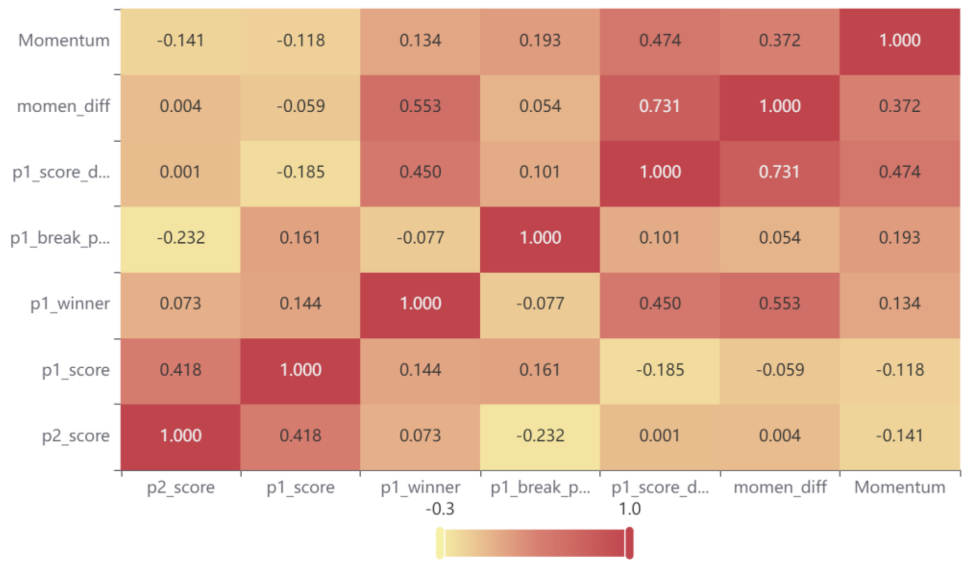

Figure 9: Spearman's Rank Correlation Coefficient Heatmap

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.

5.3 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 n_1 and the number of elements in the sequence Y is n_2.

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

$$P(R = 2k + 1) = \frac{C_{n_1-1}^{k-1} \times C_{n_2-1}^{k} + C_{n_1-1}^{k} \times C_{n_2-1}^{k-1}}{C_n^{n1}} \tag{14}$$

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

$$P(R = 2k) = \frac{2 \times C_{n_1-1}^{k-1} \times C_{n_2-1}^{k-1}}{C_n^{n1}} \tag{15}$$

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

$$\begin{aligned} \mu_R &= E(R) = \frac{2n_1n_2}{n_1 + n_2} + 1 \\ Var(R) &= \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)} \end{aligned} \tag{16}$$

And Z~N(0,1), so

$$Z = \frac{R - \mu_R}{\sqrt{Var(R)}} \tag{17}$$

At this time, the rejection criterion is $Z < -Z_{\frac{\alpha}{2}}$ or $Z > Z_{1-\frac{\alpha}{2}}$

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

Table 4: 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.

6 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：

6.1 Factor Importance Screening

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：

6.1.1 Random Forest model

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:

$$\mathrm{Gini}(D) = 1 - \sum_{c=1}^{|C|} P_c^2 \tag{18}$$

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

$$\mathrm{Gini\ Index}(D, A) = \sum_{v \in \mathrm{Value}(A)} \frac{D_v}{D} \cdot \mathrm{Gini}(D_v) \tag{19}$$

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:：

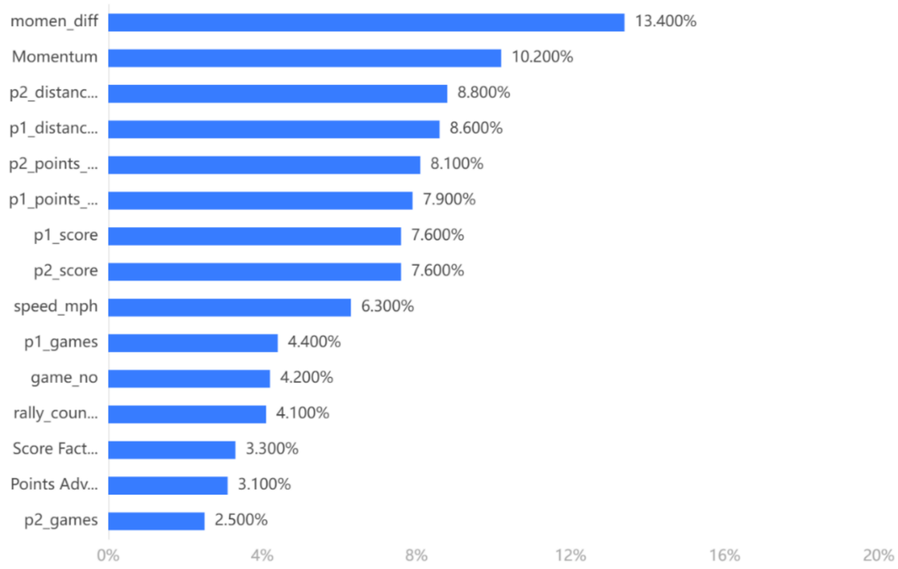

Figure 10: Feature Importance Ranking

Table 5: 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.

6.1.2 GBDT

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：

$$\mathrm{F}_{\mathrm{t}}(x) = F_{\mathrm{t}-1}(x) + r_t \tag{20}$$

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：

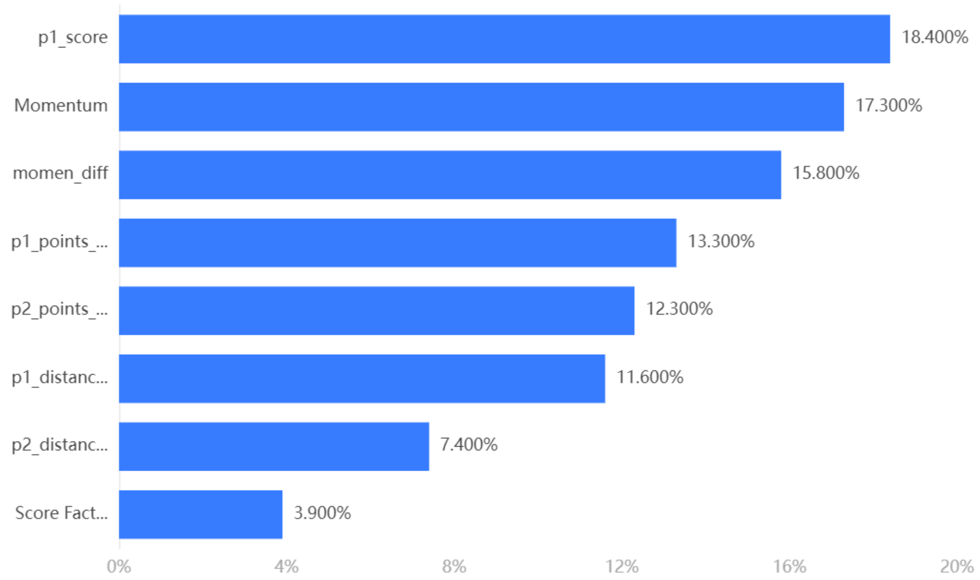

Figure 11: Feature Importance Ranking

Table 6: 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.

6.1.3 BP Neural Network

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：

$$\frac{\partial \mathrm{Loss}}{\partial \mathrm{w_i}} = \frac{\partial \mathrm{Loss}}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial \mathrm{a}} \cdot \frac{\partial \mathrm{a}}{\partial \mathrm{w_i}} \tag{21}$$

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：

Table 7: 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.

6.1.4 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.。

6.2 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：

Table 8: Friedman Test Result

Variable Name	Sample Size	Median	Standard Deviation	Statistic	P	Cohen's f Value
momen_diff_1	334	0	1.46			
momen_diff _2	334	0	1.473			
momen_diff _3	334	0	1.628	0.65	0.957	0.002
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.

6.3 Suggestions and method given

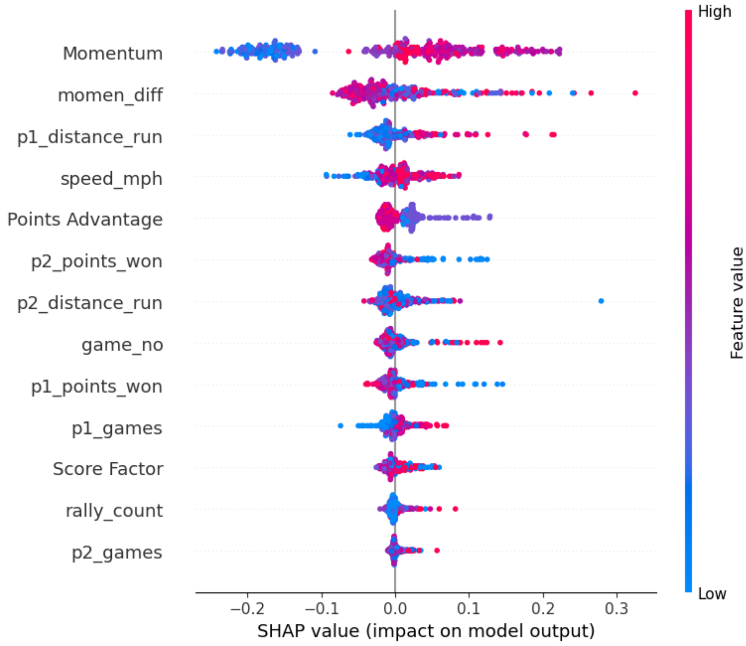

Figure 12: Shap value for features

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 [11][4], 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[5]. 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 [6]. 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[7]. 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.

7 Prediction of Other Matches and Generalization Evaluation

7.1 Predictions for multiple games on different schedules

7.1.1 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.。

7.1.2 Model prediction

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.

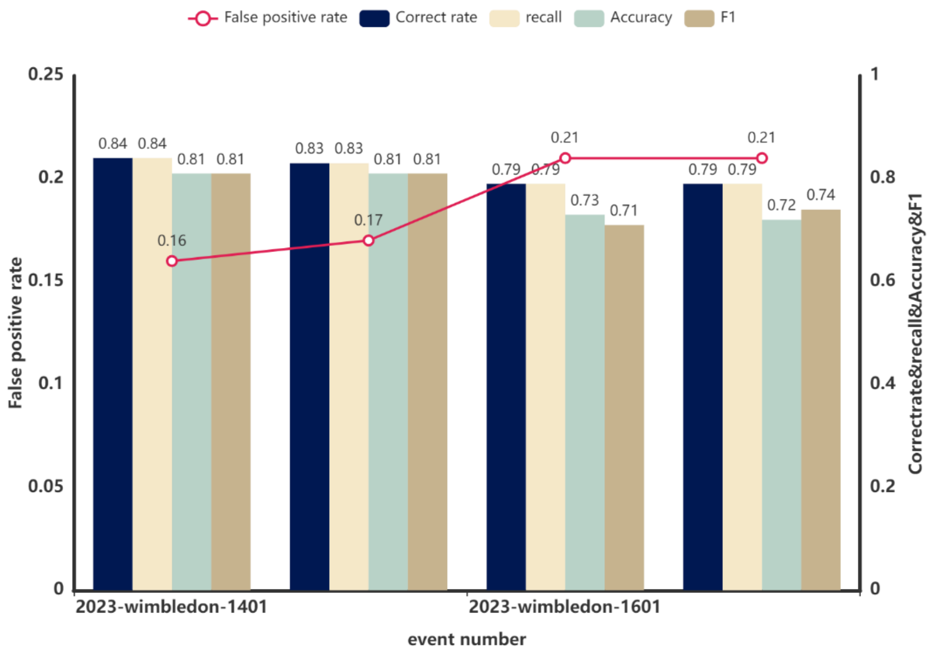

Figure 13: Subsequent Match Prediction

Table 9: Result of following match's prediction

event	correct	recall	accuracy	F1	False positive rate
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.

7.2 Assessment of generalization to badminton

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.

7.2.1 Data processing

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.

7.2.2 Parameter Mapping Matrix Transformation.

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：

$$\mathrm{A} = \mathrm{B}\mathrm{C}^{\mathrm{T}} + E \tag{22}$$

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

$$min_C||A - BC^T||_2^2 \tag{23}$$

$||\cdot||_2$ represents the L2 norm, which is the Euclidean distance. The solution to the above problem can be expressed as：

$$C^T = (BB^T)^{-1}B^TA \tag{24}$$

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.。

7.2.3 Model input and generalization evaluation

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.：

Table 10: 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.

8 Sensitivity Analysis

With this, we have completed the establishment, evaluation, and optimization of the model. Considering the perturbation of parameter changes, we introduced and constructed a sensitivity analysis model: based on the most important feature parameter set calculated in question three, we selected parameters for overall transformation to observe whether the trend's computational curve undergoes significant fluctuations and errors. Here, we chose the rally_count parameter, multiplying it by 1.1 to achieve a 10% increase, and applied random perturbations within an absolute range of 0.01. The trend curves before and after are as follows:

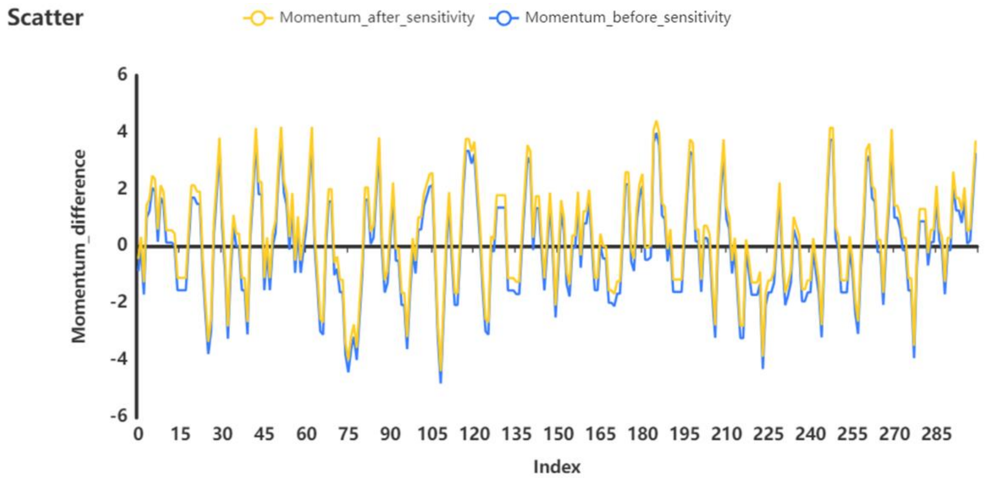

Figure 14: Momentum_sensitivity_analysis

From the above figure, it can be seen that even with the application of growth-oriented factor perturbations, the overall trend of the model has not changed significantly. It is equivalent to an overall enhancement, indicating that the perturbation of important factors has not disrupted the waveform structure of the model output, demonstrating good robustness and stability.

9 Model Evaluation and Further Discussion

9.1 Strengths

- Our model introduces the logistic regression method in machine learning in the process of training parameters, and adjusts the optimal weight through iterative training instead of manually specifying it randomly.。
- The negative weight in the weight coefficient trained by the model successfully represents the psychological pressure of the players when they are leading, taking into account the rigor of data theory and humanistic care.
- When fitting the model parameters, we did not use the simple addition method to linearly weight all factors. Instead, we comprehensively considered short-term factors and long-term factors, and intercepted a period of time through the time window to evaluate the player's momentum performance.
- We use GBDT, random forest, and BP neural network methods to screen important features

of model parameters, and select the top N most important feature sets based on the prediction accuracy of the three methods to optimize the model structure and remove redundant features.

- In the process of model generalization, we matched parameter lists with different dimensions through matrix transformation, completed the interface for data processing of different ball sports competitions, and achieved good generalization and practicality.

9.2 Weaknesses

- Without taking too much into account the influence of the players' own strength, coaching team, weather, venue, etc., the actual results may not be ideal
- The model does not consider comprehensively the random factors inside and outside the field, and simply discards the deviation processing of the data. It may not be able to predict small probability events in the game through momentum.

9.3 Further Discussion

- The model struggles with generalization across sports, showing a 15% drop in prediction accuracy compared to tennis. To address this, data correlation matching before matrix transformation could specify correlations between different sports through manual labeling, mitigating accuracy loss during data transformation. However, this approach requires significant resources, beyond our current capacity. We aim to explore this in the future to improve the model's conversion interface.

10 Conclusion

We developed a momentum model using logistic regression, analyzing factors like score, serve, and unforced errors, which significantly impact winning. The model's reliability is evidenced by a 77% accuracy rate and an 84% AUC, offering insights into real-time momentum changes through data from the 2023 Wimbledon matches.

Further analysis involved using models such as random forest, GBDT, and BP neural network to refine the model's sensitivity and robustness, with random forest showing superior parameter importance evaluation. This led to actionable strategies for athletes to enhance their performance, focusing on improvements in short rounds, physical recovery, and activity in service games, based on Carlos Alcaraz's data and momentum variance analysis.

We extended our model to badminton, maintaining accuracy above 70%, showcasing its predictive strength and generalization capability across different sports. This research provides strategic insights and practical advice for enhancing competitive performance, with future plans to improve the model's accuracy and practicality by expanding the dataset. Our goal is to advance sports science research and training, aiding athletes in achieving superior competition outcomes.

11 References

[1] Moss B, O'Donoghue P. Momentum in US Open men's singles tennis[J]. International Journal of Performance Analysis in Sport, 2015, 15(3): 884-896.

[2] Dietl H, Nesseler C. Momentum in tennis: Controlling the match[J]. UZH Business Working Paper Series, 2017 (365).

[3] Iso-Ahola S E, Blanchard W J. Psychological momentum and competitive sport performance: A field study[J]. Perceptual and Motor Skills, 1986, 62(3): 763-768.

[4] 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.

[5] 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.

[6] 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.

[7] 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

12 Report on Use of OpenAl (Feb 5, 2024 version, GPT 4.0)

We mainly use the Chatgpt large model to polish the abstract and literature review of the first draft of the paper. We ensure that the first draft of the article is handwritten, and no information from Chatgpt was referenced during the modeling process.

Queryl: < I'll give you some English paragraph and ask you to simplify it. 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' 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 dur-ing 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>

Output: < The data is from the 2023 Wimbledon tournament, detailing player names, match statistics like duration, scores, and specific metrics including aces, double faults, unforced errors, net points, player movement, ball exchanges, serve speed, and serve direction. This information aids in analyzing performance, strategizing, and enhancing viewer understanding.>

Query2: < In the 2023 Wimbledon men's singles final, the young Spanish player Carlos Alcaraz thrillingly defeated the legendary Novak Djokovic, ending the latter's dominance at Wim-bledon. This match highlights the critical role of momentum changes in tennis. The research aims to analyze the Wimbledon 2023 men's competition data and build a model to reveal the patterns and influencing factors of momentum changes, provide coaches with deep insights, and provide athletes with more effective coping strategies to improve competition competi-tiveness>

Output: < In the 2023 Wimbledon men's singles final, Carlos Alcaraz beat Novak Djokovic, marking a shift in Wimbledon's dynamics. The study examines the 2023 Wimbledon men's data to identify momentum change patterns and factors, offering coaches and athletes strategies to enhance competitiveness.>

Query3: < In Task 1, we used the momentum modeling formula, taking into account eight factors including score, serve, break, round, goal difference, service quality, unforced errors and winning shots. The coefficients are solved through logistic regression to obtain the influence of each attribute on momentum. Experimental results show that Points Advantage, Serve Advantage and Unforced Errors have a significant impact on player winning. The evaluation indicators of the logistic regression model show that it has good reliability in judging ath-letes' momentum status in real time, with an accuracy rate of 77% and an AUC of approxi-mately 84%. In addition, we also took the 2023 Wimbledon match as an example to analyze the momentum state change diagrams of the athletes on both sides, demonstrating the appli-cation and explanatory capabilities of the model in actual matches. >

Output: < In Task 1, we developed a momentum model based on eight factors—score, serve, break, round, goal difference, service quality, unforced errors, and winning shots—using logistic regression to determine each factor's impact. Points Advantage, Serve Advantage, and Unforced Errors significantly affect winning. The model proved reliable for real-time momentum assessment, with a 77% accuracy rate and an 84% AUC. We also analyzed the 2023 Wimbledon matches, showcasing the model's application and explanatory power through athletes' momentum state change diagrams.>

Query4: < In Task 3, based on the model and data of Task 2, we analyzed the game data of player Carlos Alcaraz and proposed a method of screening important parameters through models such as random forest, GBDT and BP neural network to optimize the sensitivity of the model. and robustness. By comparing the evaluation results of different models, it was found that random forest performed best in evaluating the importance of model parameters, so the im-portance parameter list provided by it was selected. In addition, we also analyzed the mo-mentum difference of player Carlos Alcaraz in the past five games, and found through the Friedman test that there is no significant difference in momentum difference. This shows that the volatility changes presented by players in past games are relatively consistent, which can be used as a joint reference and provide support for model recommendations. Fi-nally, based on the importance parameters derived from the model, we gave guidance sug-gestions for players, including improving abilities in short rounds, effectively using game intervals for physical recovery, and running more actively in service games, etc. These rec-ommendations are based on the model's analysis results and provide targeted guidance for athletes' performance in competition.>

Output: < In Task 3, we refined our analysis with Carlos Alcaraz's game data, employing models like random forest, GBDT, and BP neural network for parameter optimization to enhance model sensitivity and robustness. Random forest excelled in parameter importance evaluation, guiding our selection. Analyzing Alcaraz's momentum variance over five games with the Friedman test revealed consistent performance fluctuations, offering joint reference for model advice. Based on key parameters, we suggested improvements in short rounds, physical recovery during game breaks, and increased activity in service games. These strategies are tailored to boost competitive performance, derived from model insights.>

Query5: < In task 4, we used the game data of athlete Carlos Alcaraz and verified the predictions of other games by the model. We found that as the game time increases, the accuracy of the model gradually decreases, but the overall accuracy can still be maintained at around 80%. The model was further extended to badminton game data. After matrix transformation, the model can still maintain an accuracy of more than 70% in badminton games. Taken together, the model demonstrates excellent prediction and generalization capabilities in different competition types and sports provides strong support for accurate prediction of athlete per-formance trends, and has certain potential for practical application >

Output: < In Task 4, using Carlos Alcaraz's data, we verified our model's predictions across different matches. We observed a slight decrease in accuracy with longer game durations, yet it remained around 80%. Extending the model to badminton, with matrix transformation, it maintained over 70% accuracy. This underscores the model's strong predictive and generalization abilities across sports, offering robust support for forecasting athlete performance trends, highlighting its practical application potential.>

Query6: < In conclusion, our research can provide a comprehensive momentum analysis and pre-diction method for the field of sports competition, and also provide scientific and practical guidance and suggestions for coaches and athletes. In the future, we will continue to im-prove the accuracy and practicality of predictions by further optimizing the model and ex-panding the data set. We sincerely hope that our research can provide useful reference for sports scientific research and practical training, and promote athletes to achieve better re-sults in competitions>

Output: < In conclusion, our research offers a thorough method for analyzing and predicting momentum in sports competitions, providing valuable insights and practical advice for coaches and athletes. We plan to enhance prediction accuracy and utility by optimizing the model and broadening the dataset. Our goal is to support sports science research and practical training, ultimately helping athletes excel in their competitions>

Query7: < The model does not generalize well enough, and its prediction success rate is nearly 15% lower than in tennis. Therefore, the correlation matching of the data can be performed in advance on the basis of the original matrix transformation, and the correlation of the game data of two different ball types can be specified through manual labeling. This can effectively reduce the problems caused by the expansion and contraction of the data during matrix transformation. Loss of ac-curacy. However, this will consume a lot of manpower and material resources. It is difficult for us to do it with our current time and manpower. We strive to have the opportunity to realize it in the future and provide a better conversion inter-face >

Output: < The model struggles with generalization across sports, showing a 15% drop in prediction accuracy compared to tennis. To address this, data correlation matching before matrix transformation could specify correlations between different sports through manual labeling, mitigating accuracy loss during data transformation. However, this approach requires significant resources, beyond our current capacity. We aim to explore this in the future to improve the model's conversion interface.>