A Study on the Influence Factors of Momentum Based on Random Forest and XGBoost Algorithm

Ruiting Zhao[1,*], Rui Li[2], Jingrui Cai[2]

[1]School of Mathematics Statistics and Mechanics, Beijing University of Technology, Beijing, China
[2]School of Economics and Management, Beijing University of Technology, Beijing, China
[*]Corresponding author: 2905406224@qq.com

Abstract: *This study examines the effect of momentum on the outcome of tennis matches by analyzing data from the 2023 Wimbledon men's singles final. A momentum computational model was established and run test, random forest and XGBoost machine learning methods were applied, and the results showed that momentum has a non-random effect on the match results. It was found that the speed of serve was the most important factor affecting the outcome of the final match. In this match, veteran Novak Djokovic lost to rising star Carlos Alcaraz, demonstrating the latter's potential as the first men's Grand Slam singles champion born in the 2000s. This study fills a gap in research on the factors influencing momentum and provides a new perspective on understanding the mechanisms behind tennis match outcomes.*

Keywords: *Random Forest, XGBoost, CUMSUM Algorithm*

1. Introduction

Momentum, as an important concept in sports competition, has been widely noticed and studied [1]. Past studies have mainly focused on the effect of momentum on the outcome of a match, yet relatively little research has been conducted on the factors affecting momentum [2]. This study aims to fill this research gap and explore the role of momentum in tennis matches by deeply analyzing the data of the 2023 Wimbledon Men's Singles Final. In this paper, we will establish a momentum calculation model and combine running tests, machine learning and other methods in order to explore the influence of momentum non-randomness.

The showdown between veteran Novak Djokovic and rising star Carlos Alcaraz in the 2023 Wimbledon Gentlemen's Final drew a lot of attention. Djokovic, a legendary player with 24 Grand Slam titles, showed his serve dominance in the match, while Alcaraz, a rookie born in the 2000s, emerged as a strong contender for the men's Grand Slam singles title. In this match, the speed of serve was considered as one of the most important factors affecting the outcome of the final. Through this study, we will delve into the effect of momentum on the outcome of tennis matches, providing new insights and implications for the field of sports competition.

2. Momentum calculation model

2.1 Description of momentum

Momentum can be defined as a measure of a player's current performance, relative to the deviation from the average or expected performance in a game. Momentum can be quantified by calculating a player's scoring advantage relative to his opponent over a spe-cific time window. To define a function that calculates a player's momentum in a tennis match, consider the following key metrics: Break Points, Winning Points, Scoring Advantage, Serve situation, Unforced errors, Number of Sets and Sets Won, Double_fault. Thus, the following formula is defined:

$$momentum = points\_advantage + serve\_advantage + break\_points\_won + \\ unforced\_errors + winners + sets\_won + games\_won \tag{1}$$

2.2 Momentum calculation model

After categorizing the data, we used the winner of this score as a reflection of momentum over the course of the game. Through multiple combinations and analyses, we find that all of the following variables in Table. 1 are significantly correlated with the home run winner, also known as momentum.

Table 1: Pairwise correlations

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1) point_victor	1.000						
(2) p1_points_won	-0.041	1.000					
(3) server	0.347	0.013	1.000				
(4) p1_break_pt	0.058	0.041	0.195	1.000			
(5) p1_unf_err	0.388	-0.017	-0.080	0.004	1.000		
(6) p1_winner	-0.446	0.033	-0.234	-0.051	-0.173	1.000	
(7) p1_double_fault	0.133	0.005	-0.133	-0.026	0.343	-0.059	1.000

Correlation analyses were carried out by correlating the different variables that may affect point_victor in a tennis match [3]. First, we found a weak negative correlation between point_victor and p1_points_won (r = -0.041). In addition, we found a moderately strong positive correlation between point_victor and server (r = 0.347). This suggests that server may have some positive effect on the score of winning matches. Secondly, we noticed a slight positive correlation (r = 0.058) between the case of p1_break_pt and the score for winning the match.

In addition, there is a strong positive correlation between p1_unf_err and the score for winning the match (r = 0.388), while there is a strong negative correlation between p1_winner and the score for winning the match (r = -0.446). This suggests that Carlos Alcaraz's errors during the match may have a greater negative impact on his score for winning the match, while the winning score may be negatively correlated with the score. Finally, we also note a degree of positive correlation (r = 0.133) between p1_double_fault and the score for winning the match, although the correlation is not strong.

In summary, by analyzing the correlations between these variables, we can better understand the impact of different factors on the score of winning matches in tennis.

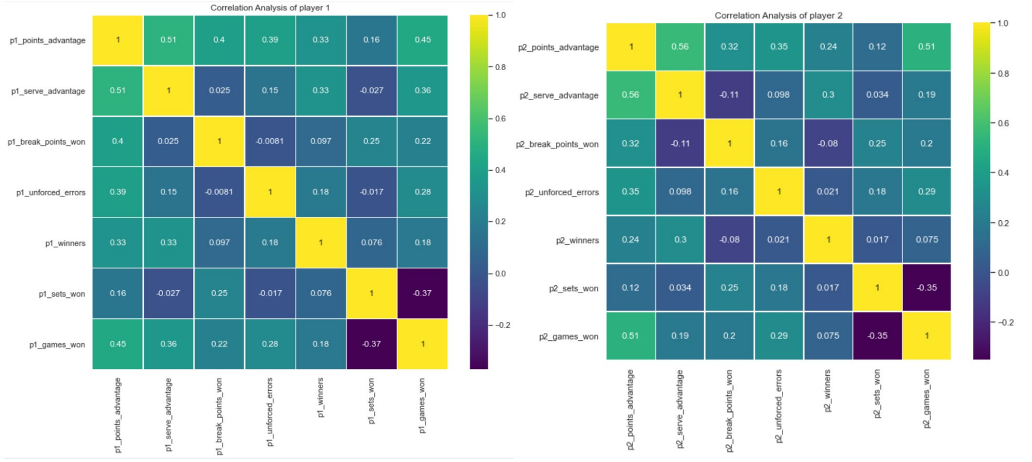


Figure 1: The entropy weight method results of Carlos Alcaraz and Novak Djokovic

Based on the significant variables mentioned above, further analysis was conduct-ed. Through PyCharm analysis, we observed that the impact of "double_fault" on the outcome is minimal, almost negligible, and it's a very small percentage of the 7,285 lines of data we found. Therefore, it was decided to disregard its influence.

Figure 1 shows the results obtained using the entropy weighting method. Each grid represents the value of an indicator on the sample, and the color or shade of the grid indicates the magnitude of that value. The colors and data revealed that there are certain constraints and balances between these characteristics, for example, advantages and errors, victories and defeats during the tennis match.

By employing entropy weight method for the analysis of the final, we can determine the impact of different variables on the two players separately [4]. Subsequently, the weights of each variable can be calculated (as shown in Table. 2.). Generally speaking, the smaller the entropy value, the greater the degree of varication of the indicator, the greater its contribution to the overall evaluation, and therefore the greater its weight be.

Table 2: The entropy weight method calculates the weights of variables in the final

Indicators	Weighting for Carlos Alcaraz	Weighting for Novak Djokovic
points_advantage	0.008978	0.018886
break_points_won	0.331889	0.323592
unforced_errors	0.020202	0.006985
serve_advantage	0.052760	0.033578
winners	0.056377	0.096413
games_won	0.004992	0.005907
sets_won	0.489380	0.477146
composite score	0.035424	0.037493

In order to better validate the definition of momentum, we chose to contrast with another match. Since Novak Djokovic lost the match in the Final and won the match in his semifinal game in 2023 Wimbledon (semifinal in short), comparing his momentum in these matches can provide a better verification of the model's accuracy.

Substitute the data from semifinal into the model. Similarly, we get the weights of the various factors for this match, as shown in Table. 3.

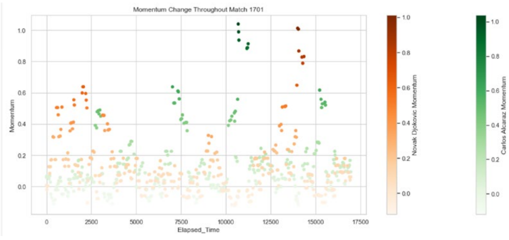
Table 3: The entropy weight method calculates the weights of variables in semifinal

Indicators	Weighting for Carlos Alcaraz	Weighting for Novak Djokovic
points_advantage	0.009817	0.013232
break_points_won	0.466373	0.417538
unforced_errors	0.007181	0.009013
serve_advantage	0.024794	0.017891
winners	0.021475	0.090447
games_won	0.003987	0.010227
sets_won	0.466373	0.441652

2.3 Results

Based on the above metrics, consider selecting the semifinal and final in the data for analysis. We can define a simple momentum calculation model that weights and combines the above factors. The serve side advantage can be represented by increasing the weight of the points won by the serve side in that time window.

In the final Novak Djokovic lost to Carlos Alcaraz, while in the semifinal Novak Djokovic won. Using these two matches to compare Novak Djokovic's change in momentum, the following Figure 2 and Figure 3 can be drawn through PyCharm programming.



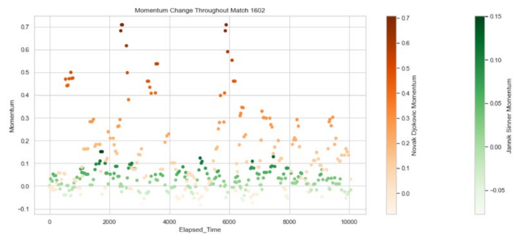


Figure 2: Momentum changes throughout the final (Left)

Figure 3: Momentum changes throughout semifinal (Right)

The horizontal coordinate represents the time of the race in seconds. The vertical coordinate represents momentum. By looking at the two graphs above, we can compare the momentum trends of the two players during the match. If the scatterplot is darker for one of the players in the graph, it means that the player has a greater advantage at that moment. By comparing the momentum changes and scatter plots of the two players, we can tell which player has more advantage or more momentum in the game.

By way of comparison, we can see that Novak Djokovic's momentum in semifinal is significantly higher than that in the Final, and as it happens, Novak Djokovic won in the former and lost in the latter.

Thereby, the modeling is more reliable.

Players usually perform better when their momentum is strong. This is because a strong momentum can make a player more confident, focused and motivated, as well as improving their technical and tactical skills. Conversely, if a player's momentum is weak, then their performance may suffer. They may feel nervous, anxious and lack confidence, and their technical and tactical levels may drop.

3. The run test

3.1 Description of turning points

Momentum can be understood as a player's ability to score consecutive points at a given point in a game. Turning points are key points in the game where the outcome of the game may change significantly, such as a shift from consecutive points allowed to consecutive points scored. Momentum has been calculated more explicitly in the first question. At this point, we identify turning points by a significant change in the difference test. When momentum shifts from a positive to a negative value, it is seen as a turning point, and the reverse holds true.

3.2 The run test

Are fluctuations in a player's performance and successive occurrences of success random? To this end we subject player1 to a brief test. The point victor is used as an explanatory variable, while the momentum defined in relation to section 2 is quantified by applying an existing dataset. Ordinary OLS regression is applied to test whether there is a significant relationship between the variables, which in turn reveals whether there is some regularity and non-randomness.

The output of the ordinary OLS regression method using STATA is shown in Table. 4.

Table 4: OLS regression results

	(1)
	point_victor
P1_momentum	-0.000***
	(-3.13)
_cons	1.518***
	(139.14)
N	7284
r2_a	0.001

With the above output it can be concluded that the expression for the relationship between point_victor and p1_momentum is:

$$point\_victor = -0.00044p1\_momentum + 1.518 \tag{2}$$

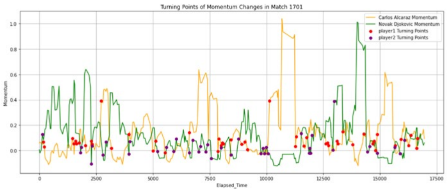
The coefficient of p1_momentum and the constant term is significant at the 1% level of significance. It can therefore be argued that the statistical significance of a player's points gains or loss, i.e. performance status, in relation to what we have defined as momentum is not random but rather follows a pattern. Therefore, we do not think that a tennis coach is right to be sceptical about the role of momentum in a match. Therefore, we then used Python for a more rigorous proof.

The Run Test is a non-parametric statistical hypothesis test that can be used to determine whether a data set exhibits randomness and is essentially a test of independence. Its requirement is that the data is dichotomous data (data 0 or 1). The principle of the test is to divide the data into two categories to see whether the former case will affect the latter case, and then conclude whether the data is random or not.

Detection of turning points in input sequences using differential and cumulative sum methods.

3.3 Results

We use the CUMSUM algorithm to identify turning points in PyCharm and mark them on the player's momentum value graph as shown in Figure 4 and Figure 5, noting turning points as 1 and non-turning points as 0.



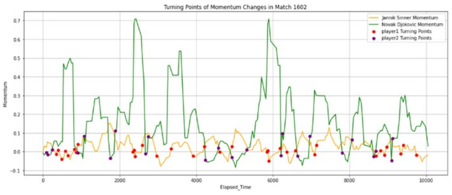


Figure 4: Turning points of momentum changes in the final (Left)

Figure 5: Turning points of momentum changes in semifinal (Right)

Thus, we use the run test method to test whether momentum and turning points are random. If random, the coach's view that the turnaround and winning streak are random events holds. Conversely, the coach's view is wrong.

Table. 5 shows the results of the model test with sample size, statistics and significance p-value (SPSSPRO 2021). Each analytical term is analyzed to see if the p-value is significant, i.e., p less than 0.05. If it presents significance then the original hypothesis is rejected, indicating non-random data. The opposite indicates that the data presents randomness (see Zeng Guang 2008).

Table 5: Run test results

Run test results			
Designation	Sample size	z	P
p1_momentum	334	-12.165	0.000***
p2_momentum	334	-11.287	0.000***
p1_turning_points	334	-0.515	0.607
p2_turning_points	334	-0.615	0.538

The results of the run test obtained according to SPSSPRO show that.

Based on the variable p1 momentum, the significance p-value is 0.000**, which presents significance at the level and rejects the original hypothesis, therefore the data is non-random.

Based on the variable p2 momentum, the significance p-value is 0.000**, presenting significance at the level and rejecting the original hypothesis, therefore the data is non-random.

Based on the variable p1 turning points, the significance p-value is 0.607, which does not present significance at the level and the original hypothesis cannot be rejected, therefore the data is random data.

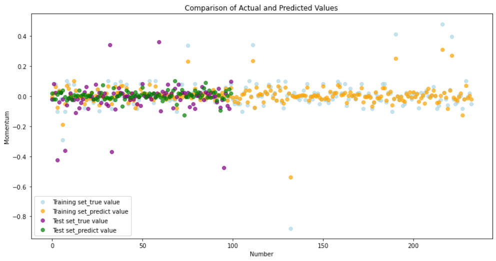
Based on the variable p2_turning_points, the significance p-value is 0.538, which does not present significance at the level and the original hypothesis cannot be rejected, therefore the data is random.

Momentum is a non-random probability.

4. Machine learning models

4.1 Modelling and analysis

Since the main study is the Final, the match between Carlos Alcaraz and Novak Djokovic was selected. As shown in Figure 6, by using random forest regression, it can be obtained that the average absolute percentage error for the training set is 1.728 [5]. The average absolute percentage error of the test set is 8.467. The average absolute error is about 0.0616, the accuracy of all three is high.



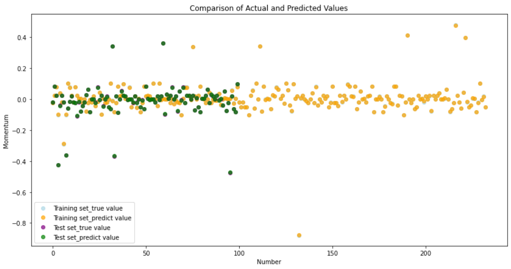


Figure 6: Comparison of actual and predicted values in the final by random forest (Left)

Figure 7: Comparison of actual and predicted values in the final by XGboost (Right)

It was further analyzed with XGboost regression for subsequent SHAP Model calculations [6]. As shown in Figure 7, an average absolute percentage error of 0.222 was obtained for the training set and 0.165 for the test set, giving an average absolute error of about 0.0009.

Analyzed using the SHAP model, Figure 8 is derived.

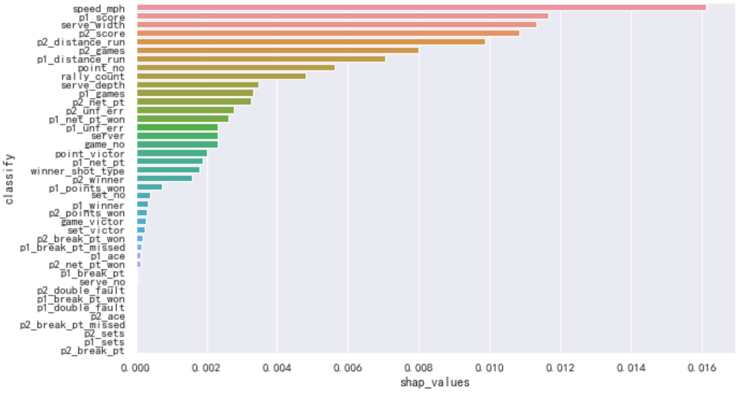
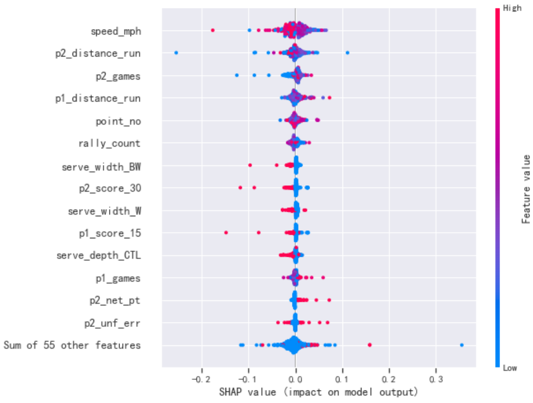


Figure 8: Sorting variables by SHAP value in the final

Combined with the analysis of the above figure can be found that different factors have different degrees of influence on the change of the situation. The length represents the size of the influence. The greater the degree of influence, the longer the graph corresponding to the factor. The graph in the 2023 Wimbledon Gentlemen's Final, the biggest influence is speed_mph. Speed_mph is affected by the match environment, weather factors, tactics, opponent level, physical condition, technical level, match stage and other factors, which affects the situation change.



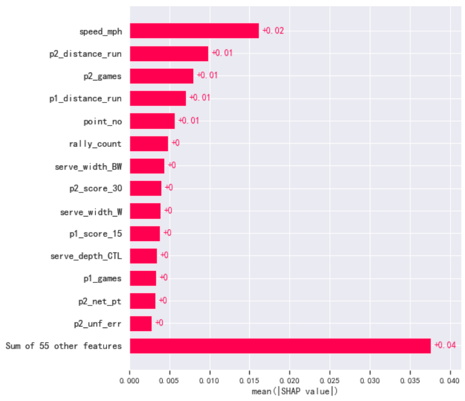


Figure 9: Variable interaction diagram *Figure 10: 15 variables sorted by mean of SHAP value*

The data we found had 63 variables, and we take the first 15. In Figure 9, red color represents positive influence and blue color represents negative influence. The two are interleaved with each other, indicating a stronger interaction. It can be seen that the distance_run interaction is stronger and the p1_score and p2_score positive effects are higher.

Figure 10 represents the average degree of influence of different variables on momentum. In this case, the horizontal coordinate is the SHAP value, and the degree of influence of the variables is ranked from largest to smallest to obtain the above graph.

4.2 Suggestions to the players

The model is retrained, and the optimal momentum prediction value is obtained by the objective programming function under reasonable parameter configuration. Genetic algorithm is used for calculation, and the output results are shown in Table. 6.

According to the value of each variable mean, we can get the ranking of the importance of momentum factors from the largest to the smallest. According to the size of mean, we can conclude that an excellent tennis player should pay special attention to the influence of p1_winner, p1_double_fault, p1_unf_err, p1_net_pt_won four factors on the momentum of his own game. To this end, we offer the following suggestions for players to play against different opponents in the new tournament:

Table 6: The output results of genetic algorithm

	Real value	Projected value	p1_ace	p1_winner	p1_double_fault	p1_unf_err	p1_net_pt	p1_net_pt_won	p1_break_pt	p1_break_pt_won	p1_break_pt_missed
Count	299.000	299.000	299.000	299.000	299.000	299.000	299.000	299.000	299.0	299.0	299.0
Mean	0.010	0.781	0.027	0.656	0.542	0.522	0.050	0.151	0.0	0.0	0.0
Std	2.028	1.741	0.162	0.476	0.499	0.500	0.219	0.358	0.0	0.0	0.0
Min	-4.100	-3.871	0.000	0.000	0.000	0.000	0.000	0.000	0.0	0.0	0.0
25%	-1.100	-0.349	0.000	0.000	0.000	0.000	0.000	0.000	0.0	0.0	0.0
50%	0.000	0.773	0.000	1.000	1.000	1.000	0.000	0.000	0.0	0.0	0.0
75%	1.950	2.029	0.000	1.000	1.000	1.000	0.000	0.000	0.0	0.0	0.0
Max	9.000	8.631	1.000	1.000	1.000	1.000	1.000	1.000	0.0	0.0	0.0

Focus on p1_winner: In the face of different opponents, timely use p1_winner to enhance the kinetic energy of the game. Watch the opponent's position and movement, look for gaps in the shot, and try to hit the ball so that it cannot be touched by the opponent. Such offensive ability helps to break down the opponent's defense, while boosting their own confidence and momentum.

Reduce p1_double_fault: Reduce the occurrence of P1_double_faults, especially in critical situations and tight scores. Strengthen service training, improve the accuracy and stability of service. In the game, pay attention to the technical details of the serve, stay focused and control the strength and di-rection of the serve.

Reduce p1_unf_pt: Reducing the number of p1_unf_pt can be achieved by reducing unforced errors. Stay calm and focused during the game and avoid taking excessive risks or aggressive shots. Pay attention to steady play, try to choose a high probability of hitting the way, reduce un-necessary mistakes.

Use p1_net_pt_won: Enhancing p1_net_pt_won's ability can increase advantage in matches. Learn to effectively attack the ball in front of the net and improve your performance and skills in front of the net. Observe the opponent's position and movement, choose the right time to go online, and flexibly respond to the opponent's counterattack, and increase the chance of winning the score in front of the net.

5. Conclusions

Through the in-depth analysis and study of the 2023 Wimbledon Men's Singles Final data in this paper, we have drawn some important conclusions. Firstly, momentum does have an impact on the outcome in tennis matches and this impact is non-random in nature. The speed of serve was identified as one of the most important factors affecting the outcome of the final, which highlights the importance of technical details in competitive sports.

Using running tests and machine learning methods, we found that players' winning streaks in matches have a non-random nature, while turning points may have a random probability. The most influential factors were identified by genetic algorithms, which provided relevant suggestions for players, which provided strong support for improving match performance.

In summary, this study fills the research gap on the factors influencing momentum and provides a new perspective for understanding the mechanisms behind the results of tennis matches. These findings are not only important for the field of sports competition, but also provide useful references and insights for future research.

References

[1] Morgulev E, Avugos S. Beyond heuristics, biases and misperceptions: the biological foundations of momentum (hot hand) [J]. International Review of Sport and Exercise Psychology, 2023, 16(1): 155-175.

[2] Stickle A M, DeCoster M E, Burger C, et al. Effects of impact and target parameters on the results of a kinetic impactor: predictions for the Double Asteroid Redirection Test (DART) mission [J]. The planetary science journal, 2022, 3(11): 248.

[3] Schlickeiser S, Schwarz T, Steiner S, et al. Disease severity, fever, age, and sex correlate with SARS-CoV-2 neutralizing antibody responses [J]. Frontiers in immunology, 2021, 11: 628971.

[4] Mukhametzyanov I. Specific character of objective methods for determining weights of criteria in MCDM problems: Entropy, CRITIC and SD [J]. Decision Making: Applications in Management and

Engineering, 2021, 4(2): 76-105.

[5] Ghunimat D, Alzoubi A E, Alzboon A, et al. Prediction of concrete compressive strength with GGBFS and fly ash using multilayer perceptron algorithm, random forest regression and k-nearest neighbor regression [J]. Asian Journal of Civil Engineering, 2023, 24(1): 169-177.

[6] Alabdullah A A, Iqbal M, Zahid M, et al. Prediction of rapid chloride penetration resistance of metakaolin based high strength concrete using light GBM and XGBoost models by incorporating SHAP analysis[J]. Construction and Building Materials, 2022, 345: 128296.