Problem Chosen	2024 MCM/ICM Summary Sheet	Team Control Number
C		2406347

Momentum in Tennis: Quantitative Analysis and Strategy Insights Into Momentum Dynamics at Wimbledon

Summary

In the men's singles final at Wimbledon 2023, young player Carlos Alcaraz defeated veteran Novak Djokovic. It was a quintessential display of **"momentum"**. Defined as strength or force gained by motion or by a series of events, momentum plays a key role in tennis, but its quantification analysis is relatively scarce, necessitating further comprehensive research.

For Problem (a), we construct a **Momentum Evaluation System** to quantify momentum indicators, utilizing the **Bayes Optimization Algorithm** to ascertain the respective impact weights of each factor on momentum, achieving an accuracy of **0.7** using the **Mean Square Error**. Visualizing the flow of game, we pinpoint pivotal moments of significant momentum shifts, assessing the performance across various players. The results indicate that four key factors: **score, serving advantage, serving errors** and **consecutive points** exert significant effects on momentum fluctuations. Comparative momentum analysis enables effective evaluation of players' performance and prediction of match trends.

For Problem (b), **Statistical Hypothesis Testing** is employed to quantitatively analyze momentum's influence and its randomness. Subsequently, we construct the distribution of actual and virtual momentum scores through calculations and **Random Race Simulation**, achieving that P-value is 0.0105, which indicates the high prediction of the model. Upon analyzing the significant differences between the two subgraphs, we reject the null hypothesis H_0, concluding that the effects of momentum changes on player performance and game trends are not entirely random.

For Problem (c), based on **Random Forest Regressor,** we analyze the most influential factors of momentum change employing **Feature Importance Analysis** and the **Heteroscedasticity Test,** concluding the model's predictive accuracy and sensitivity to be relatively high. In conclusion, the integration of players' historical match data into this model enhance our understanding of both our own and our opponents' strengths and weaknesses, enabling formulation of targeted strategies for competitive advantage.

For Problem (d), we employ the **R-squared Test** to assess the model's rationality in prediction which demonstrates a well-fitted model. By introducing new features for further optimization and assessing their sensitivity, the model's applicability and generalization is significantly improved. Subsequently, we discover that the model exhibits considerable efficiency in the predicting the outcomes of women's singles matches, indicating that the model possesses high applicability and generalizability.

Against the backdrop of **Data-driven Analysis**, our constructed model can be leveraged to quantitatively assess momentum and gain strategic insights, evaluating player performance and predicting match outcomes. This bears profound significance and impact for the realm of future sports research.

Keywords: Momentum; Momentum Evaluation System; Statistical Hypothesis Testing; Random Forest Regressor; R-squared Test

Contents

1 Introduction

1.1 Problem Background

In the 2023 Wimbledon men's singles final, 20-year-old Spanish phenom Carlos Alcaraz overcame the seasoned Serbian champion Novak Djokovic, ending his near-decade-long reign at the All England Club since 2013[1]. This victory marked the emergence of a new era in tennis, epitomized by the pivotal concept of "momentum"—a term frequently invoked in sports analytics. characterized by the dynamic force accumulated through a sequence of events or actions, is a critical factor in tennis, yet its precise measurement remains a challenge, yet seldom quantified. Defined as the strength or force garnered through movement or a succession of events[2], " Momentum " is a critical factor in tennis, but its precise measurement remains a challenge.

1.2 Restatement of the Problem

Considering the background information and restricted conditions identified in the problem statement, we need to solve the following problems:

- Develop a model that captures and visualizes the flow of the game, pinpoint the specific time as well as the player who gets the upper hand and quantify their relative superiority.
- Employ a model to analyse the existence and influence of "momentum" within the game and discern whether it is a random fluctuation.
- Formulate a model to find the most relative indicators that lead to shifts in the "momentum" — transitions that could sway the game's advantage from one player to the other. Offer advises for players preparing for a new match against opponents.
- Test the model to infer accuracy of the model prediction as well as the validity and generalization of the model application.

1.3 Our Works

For problem (a), we initially constructed a **Momentum Evaluation System** and define the momentum calculation method, considering **score**, **serving advantage**, **serving errors** and **consecutive points**. Then we use **Bayes Optimization Algorithm** to determine the respective **impact weights** of each factor and analyze the influence of each individual factor on momentum. By aggregating and visualizing the influence effects, we eventually quantify player's momentum score, analyze the specific events leading to a significant momentum shifts, and compare player types and their game performance.

For Problem (b), we employ **Statistical Hypothesis Testing** to quantitatively analyze the influence of momentum and its randomness, considering four key factors identified in the preceding Momentum Evaluation System. Subsequently, we construct the distribution of actual and virtual momentum scores through calculation and **Random Race Simulation**. After analyzing the significant differences between two subgraphs, we reject the null hypothesis H_0, concluding that the effects of momentum changes on player performance and game trends are not entirely random.

For problem (c), to predict the flow of game and explore the most relevant influencing factors of the momentum change, we utilize the **Random Forest Regressor.** We initially define the feature variable X and the target variable Y, and subsequently construct the game prediction equation Y_{pred}. **Feature Importance Analysis** reveals that for the player studied in the model, serving advantage and ace score exert the greatest influence on momentum change. Finally, we assess the accuracy of the model predictions and evaluate the sensitivity by conducting a **Heteroscedasticity Test**, and offer a preparation strategy for the players.

For Problem (d), we employ the **R-squared Test** to assess the model's rationality in prediction, and R-squared analysis (0.74) demonstrates a well-fitted model. By introducing new features for further optimization and assessing their sensitivity, the model's applicability and generalization of significantly improved. Subsequently, we discover that the model exhibits considerable efficiency in the predicting the outcomes of women's singles matches, indicating that the model possesses high applicability and generalizability.

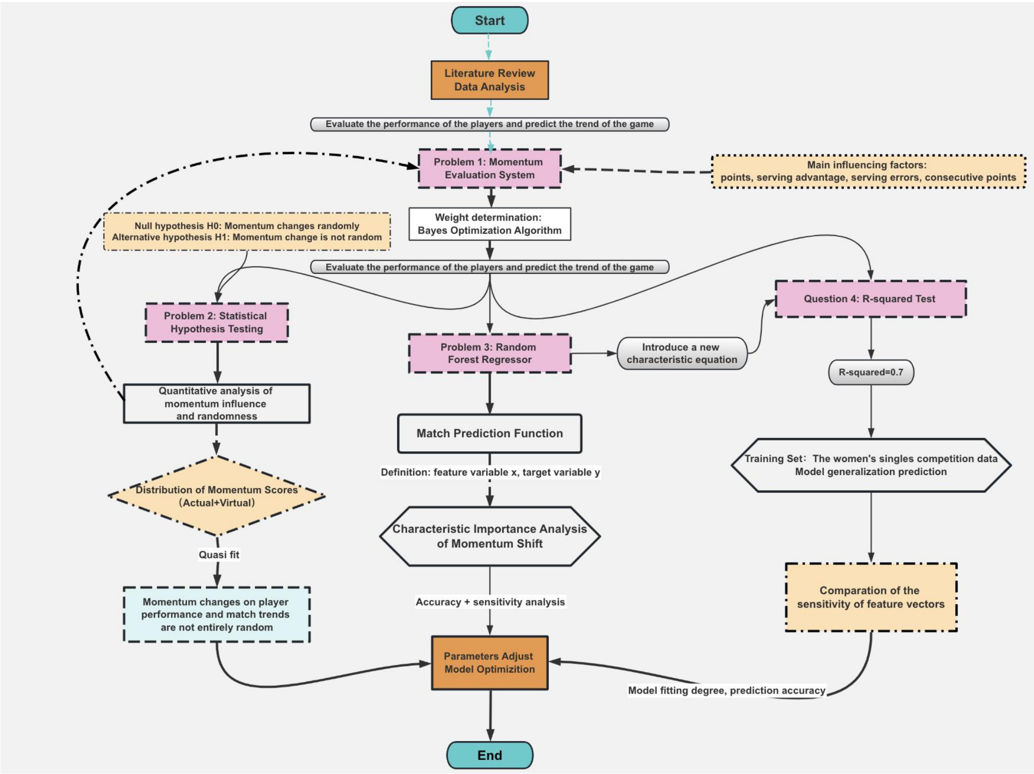


Figure 1: Model Framework

2 Assumptions and Justifications

Assumption 1: We treat scoring point outcomes as random for momentum randomness analysis.

Justification: In the Random Race Simulation, we initially assume randomness across all player scores, and that player performance remains relatively stable, unaffected by extraneous factors.

Assumption 2: Model simplification in the initial analysis in the first question may lead to variable adjustments in subsequent modeling.

Justification: The initial model focuses on four key momentum variables, with potential for expansion or modification as the model evolves.

Assumption 3: Invariance Assumption

Justification: We assume that match conditions (e.g., field, player status) are constant throughout the data set, and does not take into account the effects that these external conditions may have on the outcome of the match.

Assumption 4: Overfitting Risk Assumption

Justification: The Random Forest Regressor may exhibit overfitting, potentially because of an excessively large parameter space; too few factors are considered; in an attempt to fit the actual situation, the model may overfit to historical data.

Assumption 5: Model Parameters Assumption

Justification: We assume that the number of trees in a random forest is sufficient to provide good performance and generalization. At the same time, other default parameter settings, such as the minimum number of sample splits for nodes, are considered appropriate for the current data set.

3 Definitions and Notations

3.1 Definitions

Grand Slam: The Grand Slam in tennis is the achievement of winning all four major championships in one discipline in a calendar year. The four Grand Slam tournaments are the Australian Open, the French Open, Wimbledon, and the US Open, with each played over two weeks.

Scoring:[3] Match: best of five sets (for Gentlemen's matches at Wimbledon); Set: collection of games; 6 games win a set, but players must win by two games until the set is tied 6 – 6 when a tie-breaker is played (see below); Game: collection of points; a player wins when reaching 4 points but must win by two. See “scoring a game” below.

Scoring a game:[3] 0 points = Love; 1 point = 15; 2 points = 30; 3 points = 40; Tied score = All (e.g., “30 all”); 40 – 40 = Deuce (players have won the same number of points, at least 3 points each); Server wins a deuce point = Ad-in (or “advantage in”); Receiver wins a deuce point = Ad-out.

Serve: players alternate games as the “server” (the player who hits the initial shot of a point) and “returner”. In professional tennis, the server tends to have a big advantage. A player is given two serves to put the ball in play (into the “serving box”) on each point. Failure to hit a serve in play in two attempts is a “double fault” and the returning player is awarded the point.Breaking serve – when the returning player wins a game; Break point – a point in which if the returner wins, they would win the game; Holding serve – when the serving player wins the game.

Tie-breakers: each set ends when a player has won 6 games, as long as they are ahead by at least two games (i.e., 6 – 4). If not, play continues until a tie at 6 – 6 is reached. At this point a tie-breaker is played. At Wimbledon tie-breakers are first to 7 points (must win by 2 points) except in the 5th set of a match when it is first to 10 points (must win by 2 points).

Rest breaks/sides of court: players switch sides of the court after game 1 and then after every two games. 90 second rest breaks are allowed starting at the 3rd game at every change of sides. During tie-breakers, players change sides every six points. Players also rest for at least 2 minutes after the conclusion of each set. Medical timeouts and one bathroom break are permitted.

3.2 Notations

The key mathematical notations used in this paper are listed in Table 1. The symbols which are not frequently used will be introduced once we use them.

Table 1: Notations used in this paper

Symbol	Description
n	total number of points scored in a game
$w(t_i)$	weight of scores at time point t_i
$P(t_i)$	player's score at time point t_i
α	weight of serving score at time point t_i
$S(t_i)$	serving score at time point t_i
β	weight of serving errors at time point t_i
$D(t_i)$	serving errors at time point t_i
b	weight of consecutive scores at time point t_i
$C(t_i)$	consecutive scores at time point t_i
$H(t_{i-1})$	momentum of the previous time
M_1	momentum score of Carlos Alcaraz
M_2	momentum score of Novak Djokovic
ΔM	momentum difference between two players
$x_1, x_2, ..., x_n$	feature variable of momentum changes
$Imp(x_j)$	feature importance of momentum change

4 Momentum Evaluation System

4.1 Data Prerocessing

- **Data Split:**

Focused on analyzing the final data of Carlos Alcaraz versus Novak Djokovic, we divide it into five sets according to set numbers, removing the outliers of the table data.

- **Processing of missing and outlier data:**

We replace the missing values with the mode and mean of the column data, adjust the weights to make the replaced values more accurate, and initialize the variables.

Moreover, we modify part of the abnormal data, such as the appearance of "NA" and score of players replaced by other data, and delete other outlier data to ensure the validity of the data set.

- **Time processing:**

We observe that the representation of "elapsed_time" lacks uniformity for analytical purposes. Consequently, we transform the "elapsed_time" column into seconds, enabling numerical analysis by converting time-format strings into quantitative values.

4.2 Momentum Evaluation System

- **Model Construction:**

In order to quantify momentum analysis, we first build **Momentum Evaluation System** to define the momentum calculation method. According to the model assumption we've mentioned above, four key factors are considered: **score**, **serving advantage**, **serving errors** and **consecutive points** in the Momentum Evaluation System. Then we determine the respective **impact weights** of the relative factors to explore the extent to which each factor influences momentum and players.

The Momentum Evaluation System here can be written as:

$$\mathrm{M(t)} = \sum_{i=1}^{n} [w(t_i) * P(t_i) + \alpha * S(t_i) - \beta * D(t_i) + b * C(t_i)] + H(t_{i-1}) \tag{1}$$

where n represents the total number of points scored in a game; $w(t_i)$ represents the weight of points at time point t_i, related to the importance of the match and the urgency of the current match score, such as normal points, break points, stocktaking, $P(t_i)$ represents the score at time point t_i, which has a fundamental contribution to momentum ; α represents the weight of the serving score at time point t_i, $S(t_i)$ represents the serving score at time point t_i; β represents the weight of the serving errors at time point t_i, $D(t_i)$ represents a serving errors at time point t_i; b represents the weighting factor for consecutive scores, $C(t_i)$ represents the number of consecutive scores; $H(t_{i-1})$ represents the momentum of the previous time.

- **Variable Weight Assignment:**

Then we import the data set and screen the player score data by using **Bayesian Optimization Algorithm**[5] to assign weights to the relevant influence factors. Assuming that the initial value of all coefficients is 1, we iterate the weight of the impact factor 100 times:

For constant $P(t_i)$, there are three kinds of changes of coefficient $w(t_i)$: for ordinary score, $P(t_i)$ equals to the default value 1, score weight $w(t_i)$ equals to 0.1401; for break point variable "p1_break_pt""p2_break_pt" , $P(t_i)$ equals to 1.5, break_point_weight $w(t_i)$ equals to 0.4089; for safeguard point, $P(t_i)$ equals to 0.5.

$S(t_i)$ is represented by Ace variable"p1_ace""p2_ace", the value of which is 1 if the server scores directly, otherwise 0. Ace_weight α is 0.4553.

$D(t_i)$ is represented by "p1_double_fault""p2_double_fault", the value of which is 1 if the server has a double error, otherwise 0. Double_fault_weight β 0.4202.

$C(t_i)$ is represented by "p1_points_won""p2_points_won", in which, if data of "point_no" changes, it proves that the player scores a goal, and if the change continuous, it can be considered that the player is in a state of continuous scoring, and $C(t_i)$ is going to add up. Consecutive score weight b equals to 0.1964.

4.3 Visual Processing

Through the Momentum Assessment System we constructed, we obtained a single-factor momentum score for five factors: ace effect, break point effect, score effect, double fault effect and consecutive score effect. In Figure 2, the X-axis represents time, and the Y-axis represents the change of momentum score caused by the occurrence of variable effect at this time point, and the change interval is positively correlated with the variable weight.

We scrutinize the oscillations of the five curves and observe that in the final match, ace score, double fault, and break point score exhibit a lower frequency of change in momentum, whereas score and consecutive score show a higher frequency of momentum fluctuations, as illustrated in Figure 2.

This demonstrates the varying degrees and modes of influence that different factors have on momentum, and subsequently, we can utilize the Bayesian Optimization Algorithm to assign weights to these variables in accordance with these influence trends.

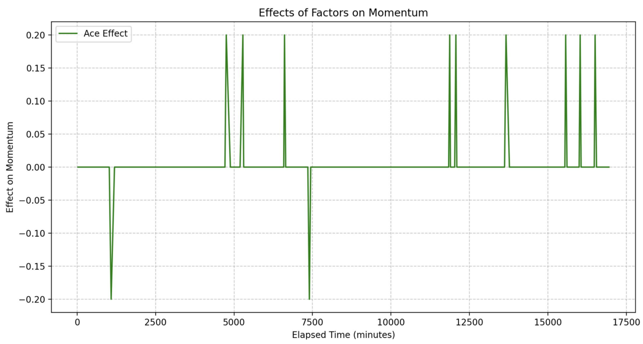


Figure 2(a): Ace Effect on Momentum

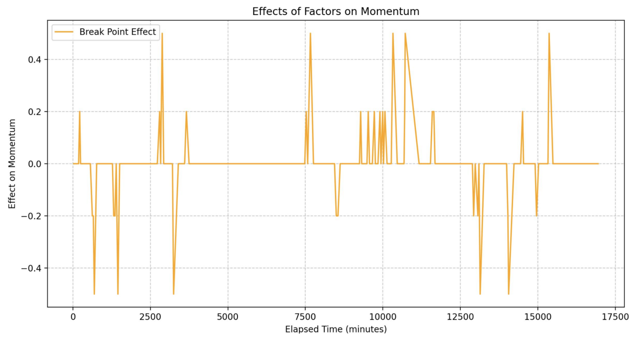


Figure 2(b): Break Point Effect on Momentum

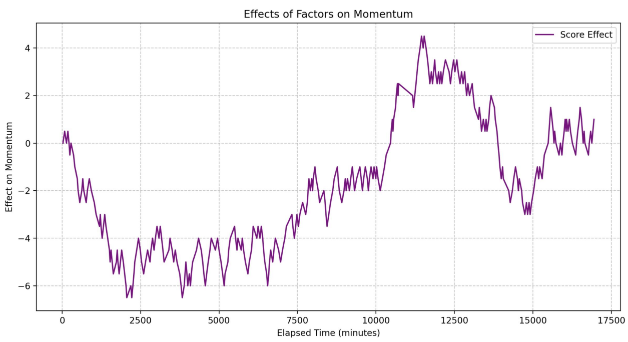


Figure 2(c): Score Effect on Momentum

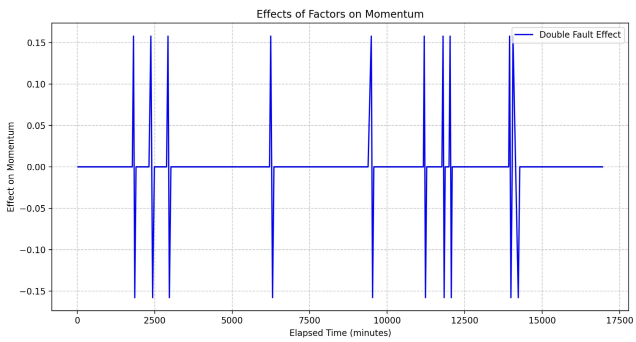


Figure 2(e): Double Fault Effect on Momentum

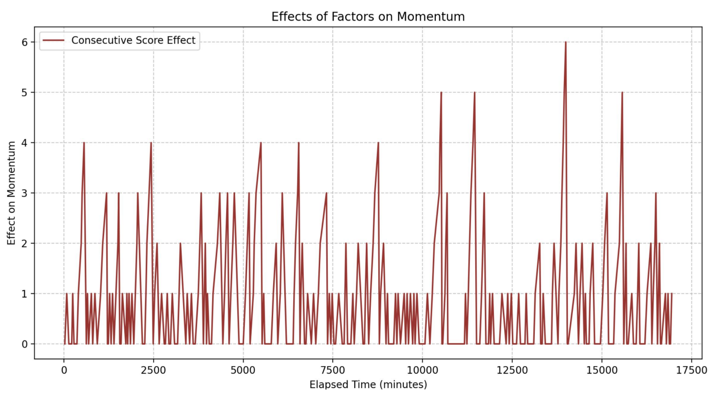


Figure 2(d) : Consecutive Score Effect on Momentum

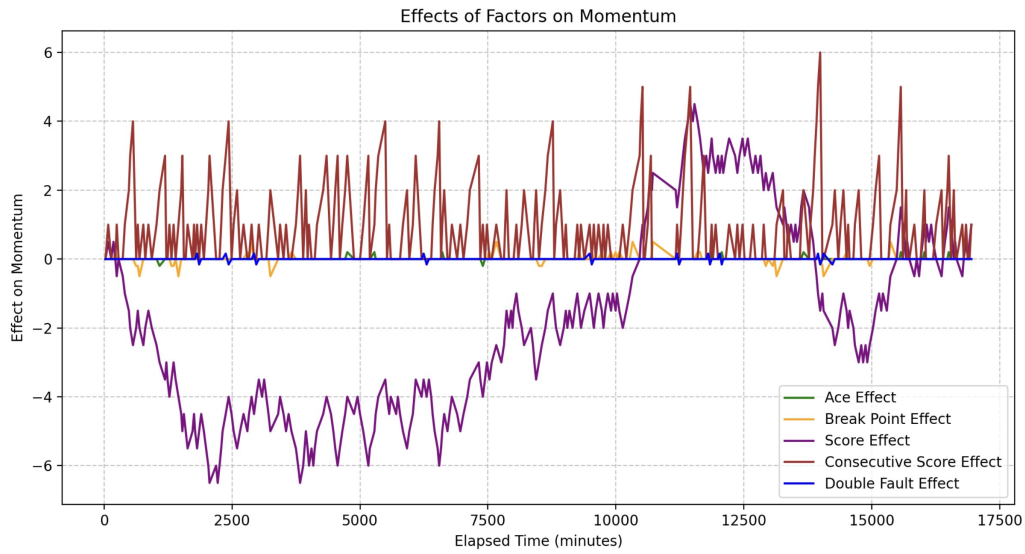


Figure 3: Effects of All Factors on Momentum

4.4 Result of Problem 1

Then we sum the changes in the influence of each single factor on momentum at each time point to get the total momentum score at different time points, as shown in Figure 3. The X-axis shows elapsed time, and the Y-axis of the chart shows the player's momentum score.

According to the graph, we set the point X where there is a significant change in the relationship between two players' momentum scores, and the moment of which is 10340 second, close to the third set of final match. We use M_1 to represent the momentum score of Carlos Alcaraz, M_2 to represent the momentum score of Novak Djokovi, ΔM （$\Delta M=M_1-M_2$） to represent the momentum difference between two players which can compare and evaluate the performance of the players in the match. If $\Delta M>0$, Carlos Alcaraz performed better than Novak Djokovic and vice versa.

The time period before X, during the first half of the first set to the third set, Novak Djokovic's momentum score is substantially higher than Carlos Alcaraz's. There are no significant data reversal phenomenon. The extreme value of ΔM, which is 35.24, appears at the moment of 6250 second, close to the second set of final match. This could be due to the winning of Novak Djokovic's opening match, his serving advantage, high Ace scored, high consecutive points scored, few errors, or other factors such as the veteran's confidence and technical proficiency.

The time period after X, the period from the late part of the third set until the end of the match, Carlos Alcaraz started to gain more momentum scores over Novak Djokovic. During the first half of the third set, Novak Djokovic's momentum scores kept showing intermittent gains, and the momentum points of both teams showed fluctuations. The extreme value appears at the moment of 11340 second, which is 28.23, close to the third set of final match.

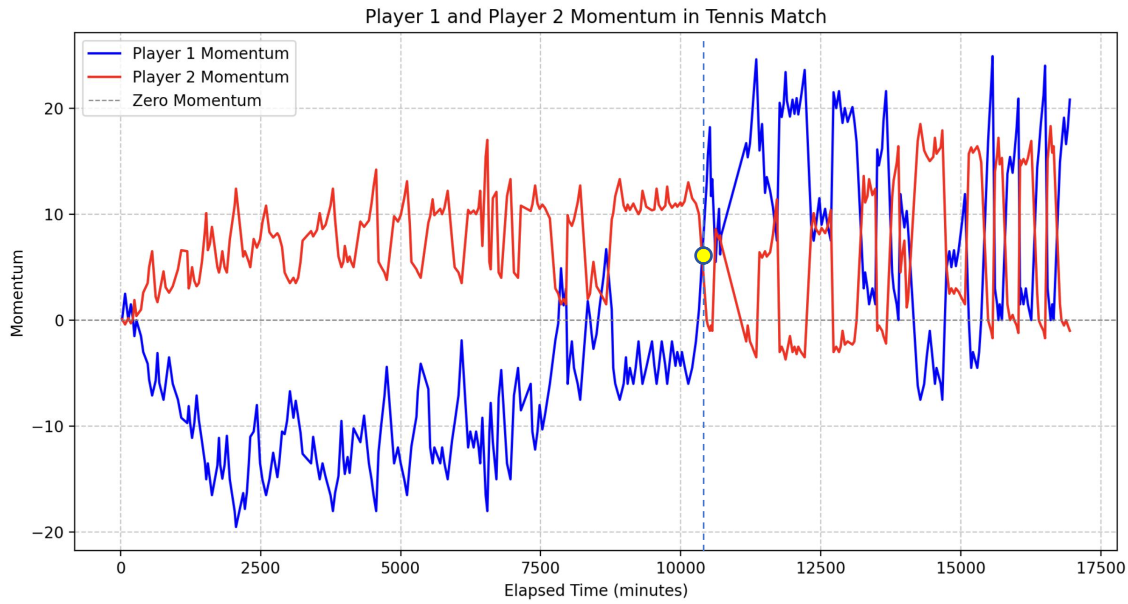


Figure 4: Carlos Alcaraz's and Novak Djokovic's Momentum in Tennis Match

In addition, we describe players' momentum changes and use different colored dots to identify specific match events. When the crease rises, Carlos Alcaraz usually has a streak of scoring and gets more points than Novak Djokovic in that time, which may increase his confidence and game control; when the crease drops, Novak Djokovic scoring more points puts Carlos Alcaraz at a disadvantage in the match. The appearance of nine green dots "p1_ace" and nineteen yellow dots "p1_break_pt" is often accompanied by a small increase in momentum scores, which means that Ace_scores and breaking_scores give Carlos Alcaraz a large psychological and score advantage. The seven red dots "p1_double_fault" are often found in the position of momentum loss, indicating that the double_serve_fault has a negative impact on the momentum of Carlos Alcaraz, including the loss of points and mental state.

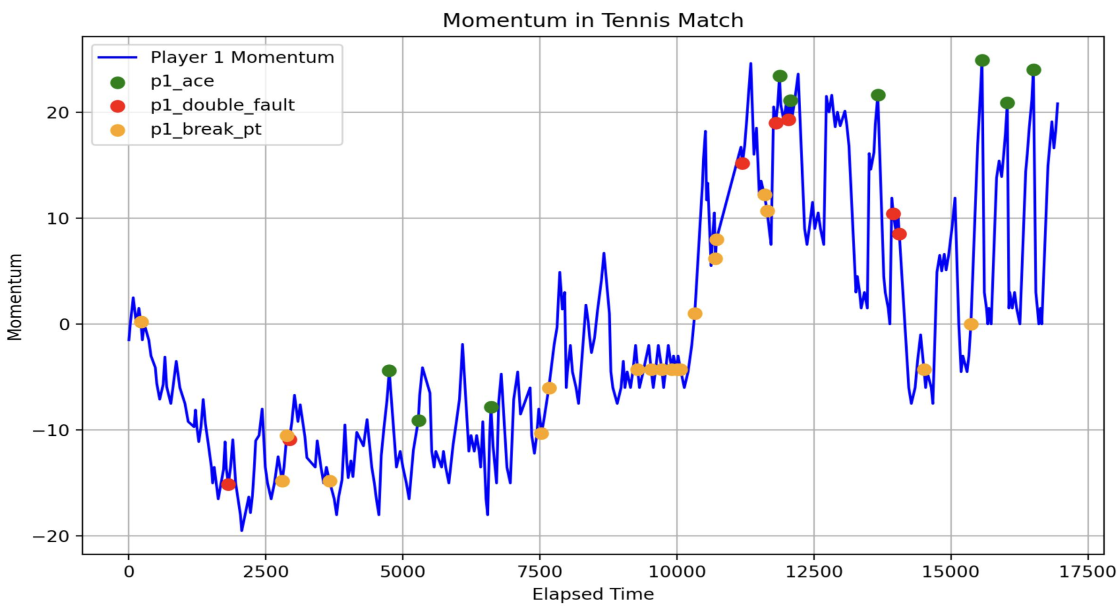


Figure 5: Momentum Changes and Specific Match Events of Carlos Alcaraz

The analysis suggests that Novak Djokovic may lack the psychological readiness to engage in a battle that extends to the deciding set. Psychologically, he tends to become overly cautious when victory seems imminent, which is reflected in the momentum chart by the small player's surge and the veteran's decline over approximately 10,000 seconds. In contrast, Carlos Alcaraz exhibits a high eagerness to score, but this often leads to performance fluctuations after playing an aggressive shot.

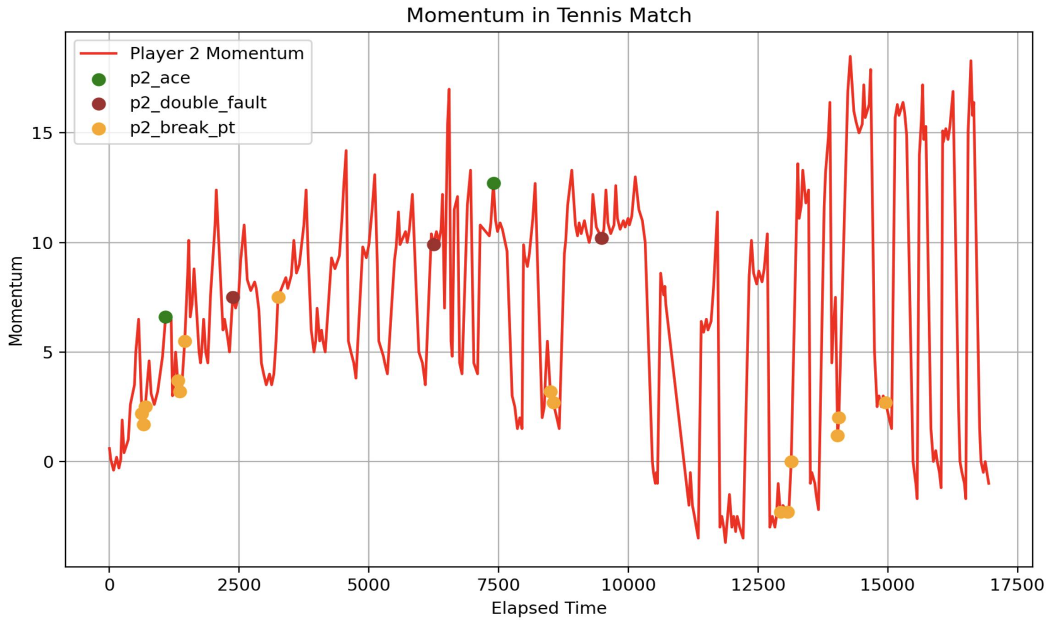


Figure 6: Momentum Changes and Specific Match Events of Novak Djokovic

The results confirm that momentum does play an important role in games, with an increase in momentum associated with improved player performance, while a decrease in momentum may signal a shift in game flow. And we can evaluate player performance through specific match events and momentum score changes, where score, serving advantage, serve miss, and consecutive points have significant effects on player performance and momentum change. The above conclusion is consistent with our previous hypothesis, and we use the Mean Square Error to calculate the accuracy value of the model, which is 0.7.

5 Statistical Hypothesis Testing

5.1 Model Assumption

Assuming that the winner of each scoring point is randomly determined, we only consider the influence of four key indicators: score, serving advantage, serving errors and consecutive points according to the previous Momentum Evaluation System we've constructed.

We are going to examine whether player performance, match trends, and match wins and losses (whether consecutive successes occur) are indeed affected by momentum, and whether the effect of momentum shifts is random by using **Hypothesis Testing**[6]:

Null Hypothesis H0: The influence of momentum change on player performance and game trend is random, and there is no significant difference from the result of random simulation.

Alternative Hypothesis H1: The effect of momentum shifts on player performance and game trends is not random.

5.2 Random Game Simulation

We assign the variable N 1000 to represent the number of matches.

Actual Momentum Calculation： On the one hand, according to the Momentum Evaluation System obtained from the first question, we can calculate the change of the momentum score in the actual game, so that the momentum score can be one-to-one corresponding to the relevant data at the time point. Based on the data, we took momentum score as the horizontal coordinate and corresponding momentum score occurrence frequency as the vertical coordinate, calculating the distribution of different momentum score, and visualized it to obtain the actual game momentum score distribution.

Random Game Simulation: On the other hand, according to the relevant hypotheses of the Momentum Evaluation System in the first place, we obtained the four key factors that have the most significant impact on momentum: score (+), serving advantage (+), serve error (-) and consecutive score (+). Therefore, the probability parameters of winning and losing matches are adjusted. We then run a **Random Race Simulation** to simulate the score distribution of simulated momentum in N random games.

5.3 Result of Problem 2

We adjusted the distance between the subgraphs and conducted a comparative analysis of the two subgraphs. Figure 5 shows the distribution of actual momentum scores and the distribution of virtual momentum scores.

Considering 7,284 games in the raw data, we can see that the actual momentum score distribution is concentrated in the range of -20 to 20, which basically conforms to the **standard normal distribution**. This may be because both players adjust their tactics after scoring or losing points, so that the positive and negative effects on momentum cancel each other out and the actual momentum scores are mostly distributed around 0.

After **Hypothesis Testing**, we figure out that average P-value from all simulations is 0.0105, which is much more smaller than 0.5. So that the distribution of actual momentum scores is significantly different from which of the virtual momentum scores. Therefore, we reject the null hypothesis H0 and accept H1 that **momentum shifts in matches are not random**. It’s reasonable to draw the conclusion that player performance, momentum shifts, and winning or losing a game are highly affected by momentum which is not random.

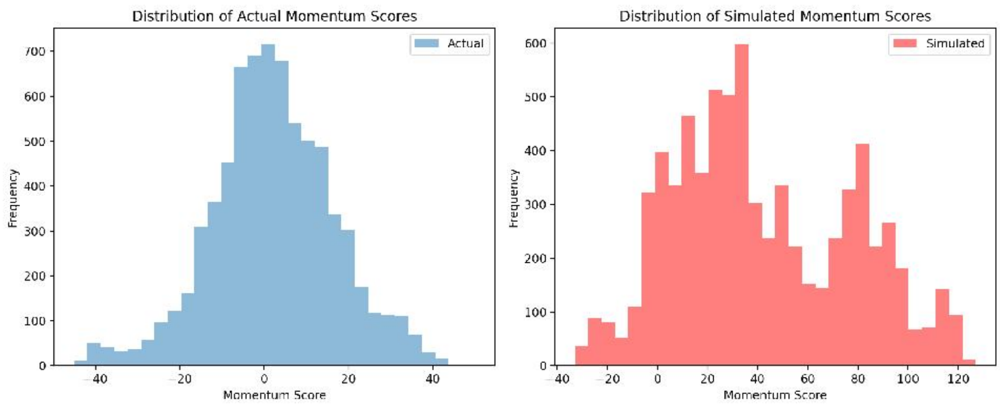


Figure 7(a): Distribution of Actual Momentum Scores Figure 7(b): Distribution of Virtual Momentum Scores

6 Random Forest Regressor

6.1 Data Prerocessing and Feature Engineering

We first utilizing the data set "Wimbledon_featured_matches.csv"provided by IBM and the United States Tennis Association to make time processing, transforming the "elapsed_time" column into seconds which is consistent with our previous steps. We then partition the data set into training and testing subsets through random sampling to ensure equitable data distribution for model evaluation, usually using a ratio of 70%-30% or 80%-20%.

Next, we define the feature $x_j = [x_1, x_2, ..., x_n]$, among all the feature variables, x_{ace} represents the ace_score difference ($x_{ace1} - x_{ace2}$), x_{df} represents the double_fault_score difference ($x_{df1} - x_{df2}$), x_{bp} represents the break_points_score difference ($x_{bp1} - x_{bp2}$), x_{cs} represents the consecutive_ points_score difference ($x_{cs1} - x_{cs2}$). Then we set "momentum_shift" Y as the target variable to represent the momentum change value.

6.2 Model Construction

To predict the fluctuations in the match and explore the most relevant influencing factors, we use the **Random Forest Regressor**[7] to analyze momentum shifts and incorporate match performance metrics.

We aim to build a **Match Prediction Function** $mode(x_{ace}, x_{df}, x_{bp}, x_{cs})$ that simulates the possibility of wins and losses Y_{pred}. In the framework of the Random Forest Regressor, the function here can be written as:

$$Y_{pred} \approx mode(x_1, x_2, ..., x_n) = \sum_{i=1}^{N} w_i \cdot h_i(x_1, x_2, ..., x_n) \tag{1}$$

Among them, $mode(\cdot)$ is the majority voting function that returns the most frequent element in the set, h_i is the prediction function of the first decision tree, w_i is the corresponding weight, and N is the number of decision trees. Each decision tree h_i is built based on the features in the training data set, and the weight w_i is usually assigned based on the predicted performance of each tree.

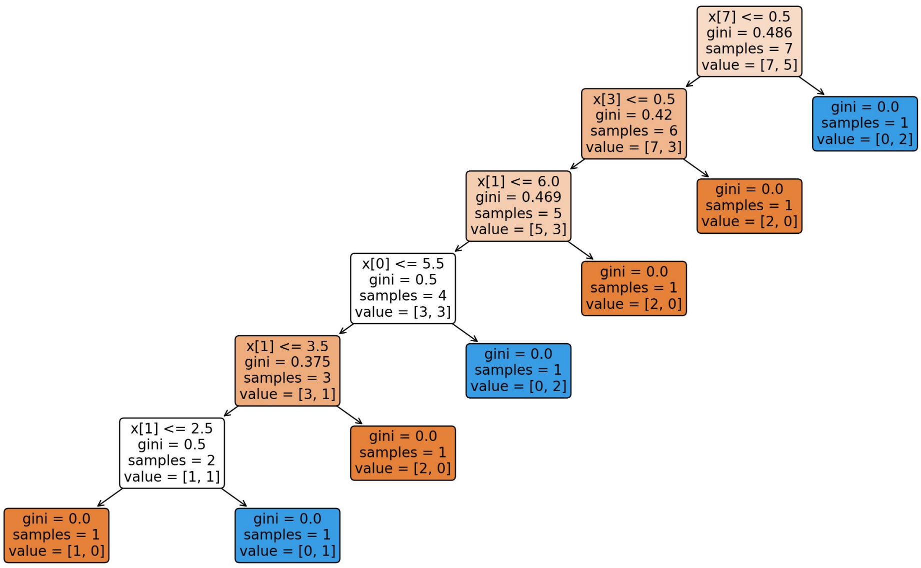


Figure 8: Random Forest Regressor Diagrammatic Drawing

6.3 Sensitivity Analysis and Heteroscedasticity Test Method

The **Random Forest Regressor** can evaluate the feature importance to perform the sensitivity analysis of the model. The importance of feature $\mathrm{Imp}(x_j)$ can be quantified in the following ways:

$$\mathrm{Imp}(x_j) = \sum_{i=1}^{N} \Delta\mathrm{perf}(T_i, x_j) \tag{1}$$

where $\Delta\mathrm{perf}$ represents the increase in performance (such as reduced impurity) when the feature x_j is used for segmentation in the decision tree T_i. Feature importance $\mathrm{Imp}(x_j)$ represents the sum of the average performance improvement of the feature x_j across all trees, reflecting the overall importance of the feature for predicting momentum shifts.

Then we plotted the Partial Dependence Plot (PDP) of the relationship between eigenvector and momentum to analyze the sensitivity coefficient of match features:

The slope size represents the sensitivity of the variable, and the higher the slope, the higher the sensitivity; the blue shaded areas are confidence intervals; the horizontal axis is the zero degree line, the closer the curve is to the zero degree line, the lower the sensitivity of the variable, and the smaller the influence on the momentum. If the curve is higher than the zero degree line, it has a positive influence, and if it is negative.

As shown in Figure 9, for player 1, "p1_net_pt_won", "p1_break_pt_won", "server_advantage" are positively correlated, "p1_ace", "p1_unf_err" are negatively correlated. "p1_net_pt_won" and "p1_ace" have the most negative influence, while the correlation of other factors is small, close to the zero point line.

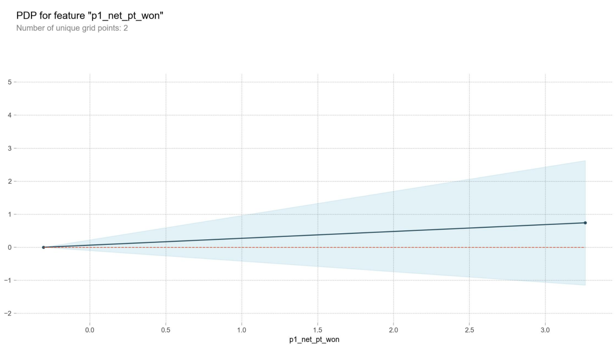


Figure 9(a): PDP for feature “p1_net_pt_won”

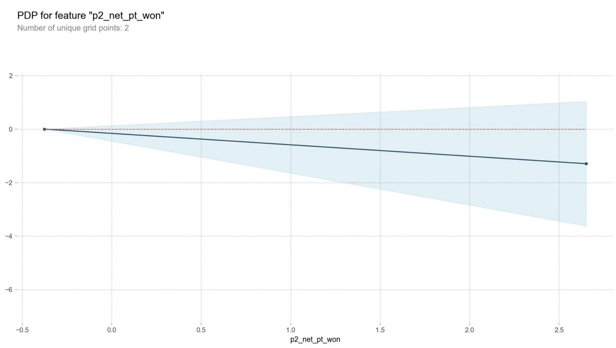


Figure 9(b): PDP for feature “p2_net_pt_won”

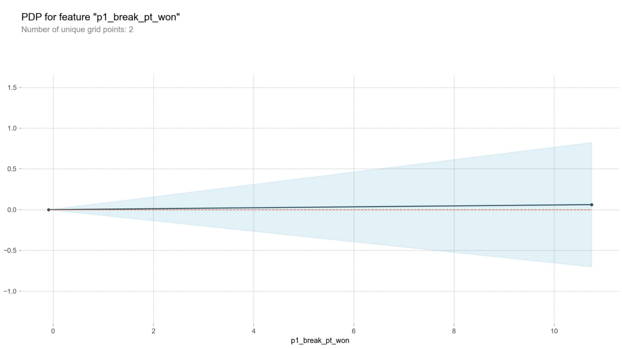


Figure 9(c): PDP for feature “p1_break_pt_won”

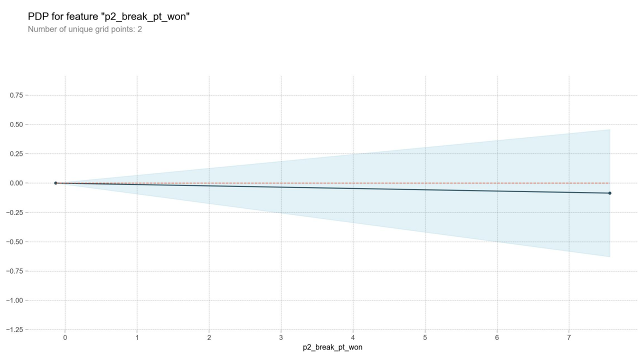


Figure 9(d): PDP for feature “p2_break_pt_won”

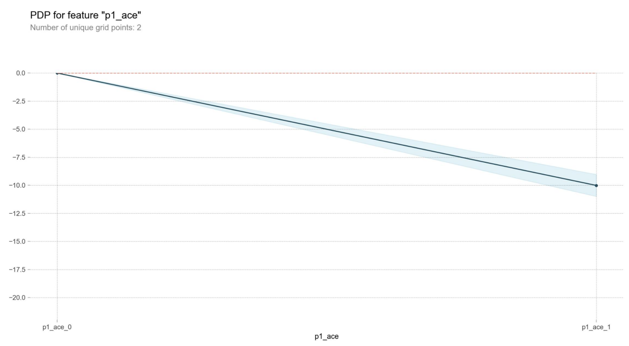


Figure 9(e): PDP for feature “p1_ace”

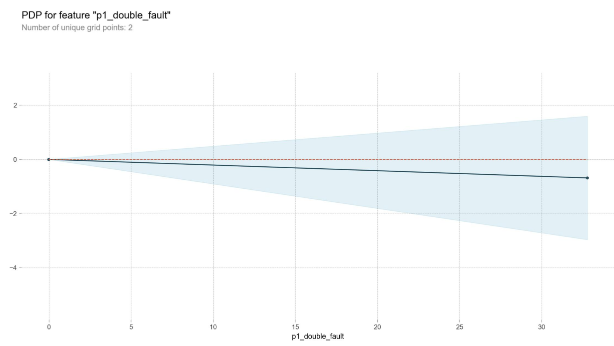


Figure 9(f): PDP for feature “p1_double_fault”

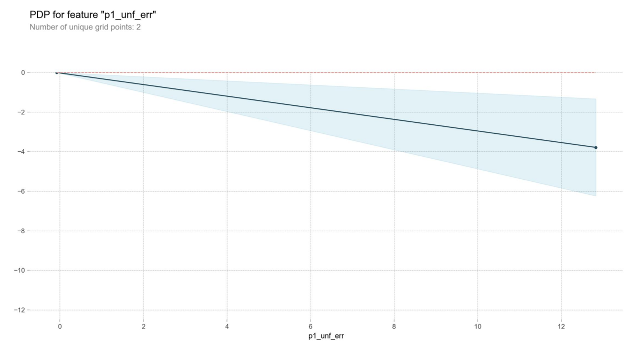


Figure 9(g): PDP for feature “p1_unf_err”

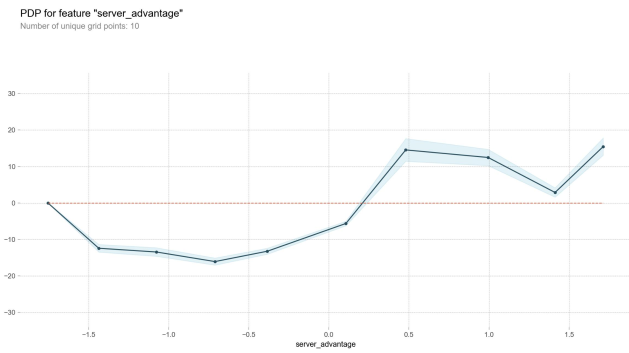


Figure 9(h): PDP for feature “server_advantage”

Figure 10 shows the relationship between the actual value and the predicted value of the momentum shift. Ideally, if the model prediction is completely accurate, all points will fall on a 45-degree line through the origin.

In Figure 10, we can observe the following: the data points are clustered near the diagonal line, highlighting the high predictive accuracy of the model. Additionally, the symmetric distribution of the data points relative to the center line indicates that the systematic prediction error is negligible. Despite the presence of some outliers, the fact that a large number of observations are close to the ideal prediction line demonstrates the model's excellent performance.

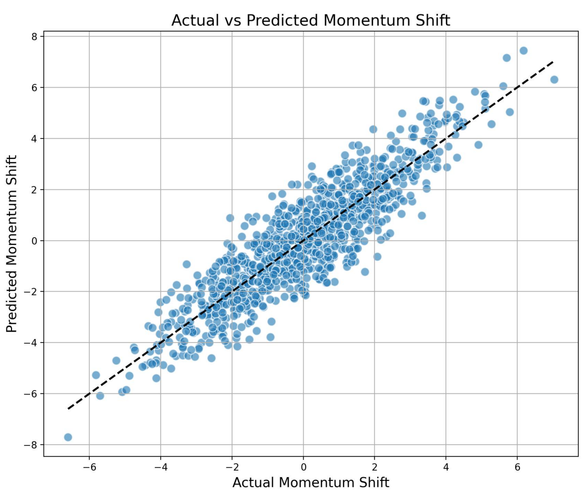


Figure 10: Scatter Plot of Actual/Predicted Value of Momentum

The residual plot is used to evaluate the accuracy of the prediction of the regression model, and each point in the residual plot represents the difference between the actual momentum value and the predicted momentum value. If all the residuals are tightly distributed around the horizontal axis, the model prediction is valid.

In Figure 11, we can observe the following:

- **Residual Concentration:**

Most residuals appear to be concentrated between the x-axis (predicted values) near -2 and 0. This indicates that within this range of predicted values, the model's predictions are relatively close to the actual values.

- **Outlier Residuals:**

There are some significant outlier residuals, especially when the predicted values are close to 0. This may suggest that the model's predictions are inaccurate at these points. These large residuals could be caused by outliers, an inappropriate model specification, or important variables missing from the model.

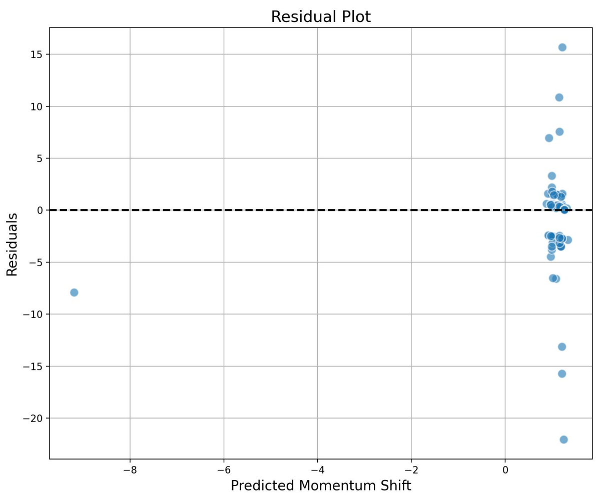


Figure 11: Residual Diagram of Actual/Predicted Value of Momentum

6.4 Tactical Strategy Advice

In order to help coaches and analysts better understand the dynamics of the game and the impact of different events on the momentum of the game, we first analyze the actual data from the final game. For each single game, we first analyzed the scores of the two players, including the main scoring types and key scoring time points, so as to reflect the types of players and the types of players who are good at/not good at. Then, the score types of the two players were analyzed longitudinally.

In the one hand, we carry out statistical analysis and visual processing of the data.

As shown in Figure 12, Carlos Alcaraz's scores most in "net_point_won" type, following with his "ace_score", representing that he is good at gaining pressure attack points. While his "double_fault" more frequently as well, especially in his losing game, which is why he performed worse than Novak Djokovic in the set 1 and 4. In conclusion, Carlos Alcaraz is a **strong aggressive player** and has a strong ability to finish the match, but he needs to **pay more attention on defense**.

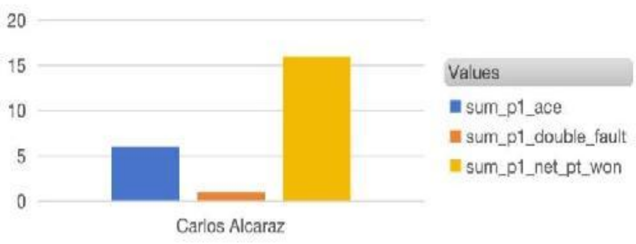
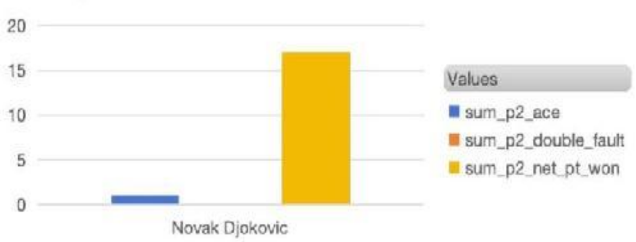


Figure 12(a): Carlos Alcaraz's Score Type For Wins **Figure 12(b): Carlos Alcaraz's Score Type For Losing**

As shown in Figure 13, Novak Djokovic is also good at oppression of the net interception score, but his "ace_score" is less, indicating that he is an aggressive player, but not as aggressive as Carlos Alcaraz. However, the frequency of his "double_fault" is very low, which proves that he is good in defense. Therefore, Novak Djokovic is an **offensive and defensive type of player,** but if he wants to play more prominently in the game, he needs to **strengthen his match finishing ability**, especially to strengthen the "ace_score" training.



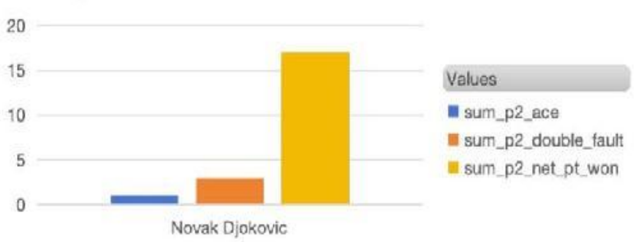


Figure 13(a): Novak Djokovic's Score Type For Wins Figure 13(b): Novak Djokovic's Score Type For Losing

In the other hand, utilizing the Random Forest Regressor, we have assessed the feature importance in the dynamics of momentum shifts for Carlos Alcaraz. As depicted in Figure 14, the characteristics ranked by their impact on momentum are ace score, double fault, and break point, with ace score being the most influential factor. The correlation between ace score and momentum change, as illustrated in Figure 9(c), is negative, suggesting that a significant downward momentum trend follows an ace. This could be attributed to the physical exertion required for an ace or a psychological effect, such as a temporary lapse in concentration post-ace. Additionally, Carlos Alcaraz's propensity for double fault further contributes to the negative momentum, indicating a need for him to improve consistency in the match. Overall, Carlos Alcaraz's performance exhibits instability, particularly around the occurrence of ace, where attention to physical management, error minimization, and mental focus is paramount for maintaining momentum.

While for Novak Djokovic, the characteristics ranked by their impact on momentum are break point, double fault, and ace score, with break point being the most influential factor. Therefore, his aggression may be relatively low, but the error rate is also low, because of the experience of the game, his performance is more stable, but he needs to be extra wary of the opponent's strong attack.

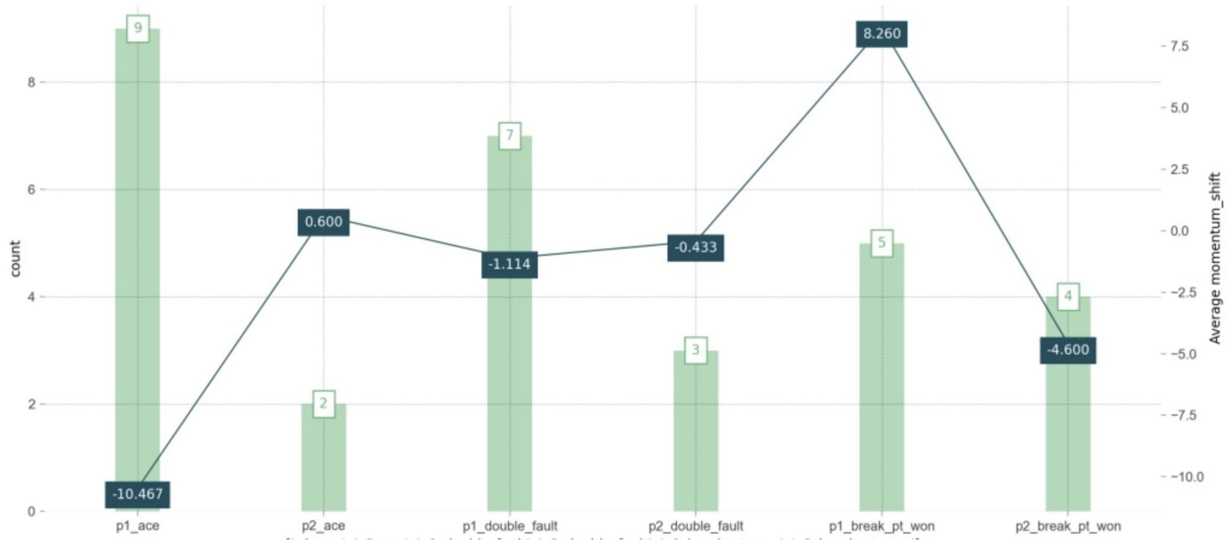


Figure 14 All Feature Importance For Momentum Changes

Finally, we compare the actual results with the predicted analysis results, and find that the model prediction fits well, so that we can provide detailed and effective tactical strategies for coaches and players based on the above analysis. By understanding how these key events affect momentum, coaches can better guide athletes to adjust strategies to take advantage of or combat momentum changes during a game.

7 R-squared Test

7.1 Model Prediction Analysis

In statistical modeling, we use **R-squared Test** to assess the reasonableness of the model in predicting the outcomes of men's tennis matches. In which, the metric reflects the proportion of variance explained by the model relative to the total variance. If R-squared is closer to 1, it indicates that the more variation the model explains and the better it fits to reality.

The coefficient of determination is calculated by the ratio of Sum of Squares for Regression (SSR) to the Total Sum of Squares (SST):

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{1}$$

where SSR (Sum of Squares for Regression) is the sum of squares explained by the model, that is, the sum of squares of the difference between the model's predicted value and the mean of the actual observed value; SSE (Sum of Squares for Error) is the sum of squares of residuals, which is the sum of squares of the difference between the predicted value of the model and the actual observed value; SST (Total Sum of Squares) is the sum of squares of the difference between the actual observed value and its mean.

Then we use **Gradient Boosting** to iterate R-squared:

$$F(x) = f_0(x) + \sum_{m=1}^{M} \beta_m \cdot h_m(x) \tag{2}$$

where $F(x)$ is the final prediction function; $f_0(x)$ is the initial predicted value is usually a simple model; β_m is the weight of the m-th weak learner; $h_m(x)$ is the prediction function of the m-th weak learner; M is the total number of weak learners.

In each iteration, a new weak learner $h_m(x)$ is trained to minimize the residual. The formula for calculating residual error is:

$$r_i = y_i - F(x_i) \tag{3}$$

where y_i is the true value of the i-th sample; $F(x_i)$ is the value predicted by the current model for sample i. The value of R-squared from all simulations is 0.46.

7.2 Model Improvement and Model Extension

To enhance the model's performance, we introduced additional feature factors based on the random forest model as new feature equations. By analyzing the sensitivity of these new factors to the model output, if they have a significant impact, this result will guide us to prioritize and incorporate such factors in future model construction to improve the model's explanatory power and predictive accuracy.

Through model improvement, we found that the following factors may have a high impact on enhancing the precision of our model's predictions:

1. Player's condition and physical fitness:

In addition to the already considered running distance and ball reception count, factors such as match duration, consecutive match days, and rest periods can be included; the player's injury history and current health condition.

2.Psychological factors:

The player's psychological state, such as stress, tension, and confidence, can be analyzed through non-verbal cues like facial expressions and body language; the historical match records between opponents and the real-time match situation can influence the player's psychology and strategy.

3.Tactical strategy and technical level:

The player's tactical choices, such as serving strategy (angle, speed, spin, depth of serve, and return position), and return strategy; the player's tactical adjustments at crucial moments, such as changes in strategy during break points.

4.Environmental factors:

Weather conditions, such as temperature, humidity, and wind speed, which can affect the trajectory of the ball and the player's performance; the type of court surface, as different surfaces (hard, grass, clay) have a significant impact on the match.

5.Audience and home advantage:

The level of audience support, as playing at home may provide additional motivation for the player; the noise level of the audience, which can affect the player's concentration.

We select the **WTA Tennis 2023 women's singles competition data**[4] to generalize our model. The study found that the model demonstrates considerable efficacy in predicting the outcomes of women's singles matches, indicating a high degree of applicability and generalization of the model. This cross-gender predictive capability suggests the reliability of the model, providing a solid foundation for further model validation and iteration.

8 Model Evaluation and Further Discussion

8.1 Strengths

1. Parameter Space Exploration:

We use ensemble learning for efficient global optimization algorithms, such as random forests, which can handle high-dimensional data and provide measures of important features, so that the model can find the best combination of parameters in the parameter space.

2. Data-driven:

Using data-driven and ensemble learning methods, the model can be constructed based on actual data and the momentum score can be calculated to provide support for empirical analysis.

3. Alternate Winning Mechanism:

The Statistical Hypothesis Testing model simulates the dynamic change of scores in actual matches by introducing the probability of alternating wins, which simplifies the workload and difficulty of data processing.

4. Statistical Test Application:

We assess the difference between simulated scores and actual scores using the Wilcoxon rank sum test, a non-parametric test method that does not require assumptions about the normality of the data distribution.

8.2 Weaknesses

1. Model Simplification:

Compared with the actual situation, the factors considered by the model are partly missing. And we tend to consider static factors while ignoring some dynamic changes of the game, such as the influence of off-field factors or psychological state etc., and the accuracy of the model predictions needs to be improved.

2. Adaptability in Specific Situations:

Due to the limitation of training set data selection, the construction of the model may be applicable to some specific types of games or players.

3. Fixed Parameter Setting:

In the random match simulation function, we assume that the probability of winning points is fixed, and it is not adjusted according to different players or match conditions, which may deviate from the actual situation to a small extent.

8.3 Possible Improvements

1. Improve the Predictive Power of the Model:

We can add new features, adjust the complexity of the model by considering more influencing factors, or employ other regression algorithms.

2. Further Diagnosis:

Due to the residuals existence, we can further diagnose the model to determine if there are specific patterns or trends that may reveal assumed irregularities or if there are potential data problems.

3. Verify Deficiencies in Data Preprocessing:

Verify whether there are data entry errors, outliers, or influencing factors that we fail to capture.

References

[1] Braidwood, J. (2023). Novak Djokovic has created a unique rival – is Wimbledon defeat the beginning of the end? The Independent. https://www.independent.co.uk/sport/tennis/novak-djokovic-wimbledon-final-carlos-alcarazb2376600.html

[2] Merriam-Webster. (n.d.). Momentum. In Merriam-Webster.com dictionary. Retrieved September 28, 2023, from https://www.merriam-webster.com/dictionary/momentum

[3] Rivera, J. (2023). Tennis scoring, explained: A guide to understanding the rules, terms & point system at Wimbledon. The Sporting News. https://www.sportingnews.com/us/tennis/news/tennis-scoring-explained-rules-system-pointsterms/7uzp2evdhbd11obdd59p3p1cx

[4] Sackmann, J. (2023). WTA matches 2023 [Data set]. GitHub. https://github.com/JeffSackmann/tennis_wta/blob/master/wta_matches_2023.csv

[5] Cui, J., & Yang, B. (2018). Bayesian optimization algorithm and application review. Journal of Software, 29(10), 3068-3090. https://doi.org/10.13328/j.cnki.jos.005607

[6] Yang, Y., Wang, L., & Song, L. (2012). Discrete distribution parameter hypothesis test of the optimal rejection region. Journal of Jilin Normal University (Natural Science Edition), 33(3), 82-83. https://doi.org/10.16862/j.cnki.issn1674-3873.2012.03.024

[7] Liu, R., Mao, J., & Li, Y. (2016). Facial feature points localization based on random forest regression. Journal of Electronic Measurement and Instrument, 30(5), 684-693. https://doi.org/10.13382/j.jemi.2016.05.004

Memo

To: Tennis Coaches and Players
From: Team # 2406347
Date: February 5, 2024
Subject: Strategic Insights on Momentum Dynamics and Game Flow in Tennis

Dear tennis coaches and players,

I hope this message finds you well.

Like many, you may have been impressed by the 2023 Wimbledon men's singles final, where Spanish rising star Carlos Alcaraz defeated Novak Djokovic, a result that likely surprised many. Currently, the predominant question may be: Why would a seasoned veteran lose to a young player? Indeed, this can be attributed to "momentum" - a critical yet infrequently quantified concept in tennis. Unfortunately, the strategic complexity of tennis renders the definition of momentum ambiguous for many, leading to a lack of deep and comprehensive understanding of its impact on player performance and match outcomes. Consequently, our team has utilized extensive match data and a data-driven algorithm to construct a series of momentum quantification models.

Initially, the Momentum Evaluation System examines key factors significantly impacting the match's trend, including score, serving advantage, serving errors, and consecutive points. Furthermore, we ascertain that momentum fluctuations in such matches are strongly correlated with the influence on player performance and match trends, demonstrating non-randomness. Additionally, our model employs real-time data for iterative optimization and autonomously adjusts weights and parameters, which boasts not only high accuracy in game predictions but also applicability to other games, demonstrating considerable robustness and adaptability.

From an intuitive standpoint, taking Carlos Alcaraz for example, Figure 15 illustrates the simplified diagram highlighting the importance of different feature vectors on momentum.

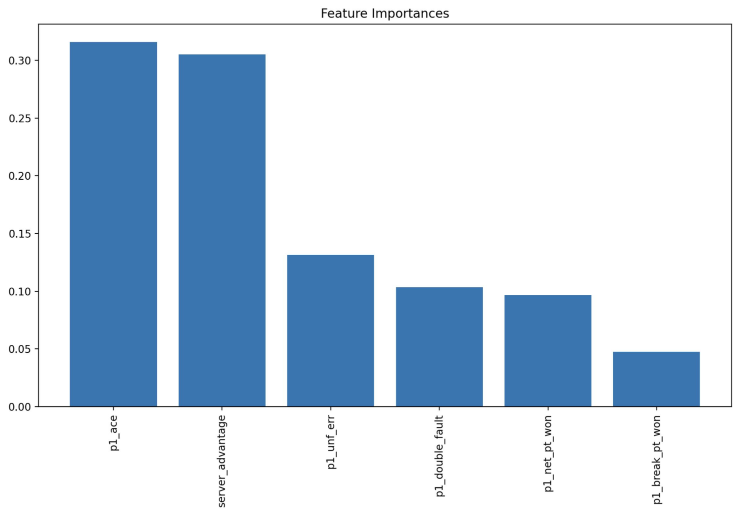


Figure 15 Simplified diagram of feature importance

Among these factors, serving advantage and ace score tend to exhibit a higher correlation and have a more profound impact on momentum changes. Drawing on our research findings, Carlos Alcaraz's example will be used to elucidate how tennis coaches can leverage momentum's role in matches for crafting targeted player preparations and strategies, as well as to optimize players' daily training schedules. Furthermore, this analysis aids players in better understanding their strengths and weaknesses and in preparing accordingly. The momentum change Comparison of Carlos Alcaraz and Novak Djokovic during the Wimbledon men's singles final is analyzed, as depicted in Table 2:

Table 2: Comparison of players' score types

	Sum of ace	Sum of double_fault	sum_net_pt_won
Carlos Alcaraz	9	7	25
Novak Djokovic	2	3	34

- **Technical and Tactical Aspects:**

Alcaraz's proficient net play and robust baseline game enable him to score opportunistically, often signaling an ascent in his momentum.

- **Physical Fitness:**

Alcaraz's superior fitness and recovery capabilities ensure sustained peak performance, fostering mental and emotional resilience amidst challenges—crucial for navigating high-pressure situations in pivotal matches.

- **Opponent Dynamics:**

Facing a seasoned adversary such as Novak Djokovic necessitates Alcaraz's preparedness to counter strategic shifts and psychological tactics. Djokovic's mastery in identifying opponents' vulnerabilities and dictating match tempo significantly influences the competition.

As is mentioned above, Alcaraz's performance in the 2023 Wimbledon final hinges on integrating his technical, physical, and psychological strengths with grass court dynamics and opponent strategies. He must continually adapt his tactics, leverage his speed and power, and exercise patience, seizing opportunities to strike.

We posit that through the application of these strategic insights, players akin to Alcaraz can adeptly harness motivational dynamics, navigate the game's fluctuations, and enhance their overall performance. This analysis is also applicable to other competitors.

We are looking forward to your good news.

Yours Sincerely,
Team # 2406347