Unleashing the Power of Momentum: Analyzing Tennis Match Flow through Data Analysis

Yuxuan Li¹, a, Niantai Qu², b and Yingzhi Bao², c

¹School of Statistics and Mathematics, CUFE, Beijing 102206, China;
² School of Statistics and Mathematics, CUFE, Beijing 102206, China.

¹linyuanecho@163.com, ²tai12307@163.com, ³1173972378@qq.com

Abstract. This paper aims to analyze the impact of momentum in the Wimbledon Men's Singles Final, uncovering its influence on the game's trajectory and providing strategic advice. To analyze player performance and game pace in tennis, we identified key variables within the dimensions of Consumption, Serving, Accuracy, and Difference. Initially, we developed a momentum prediction model in tennis using logistic regression, with an accuracy of over 73%. In line with the actual match, our visualizations provide valuable insights into the flow of play and momentum shifts in tennis matches. Secondly, we conducted a randomness analysis using the Run test on two metrics: CW (swings in play) and DPW (runs of success by one player), indicated that neither CW nor DPW occurs randomly. Furthermore, we employed Vector Auto-regressive Models (VAR) and found that momentum significantly affects the states of players and their winning streaks. Thirdly, we created the Momentum Swing Prediction Model using the Random Forest approach, introducing a dependent variable, TP, to capture the swings in momentum during a match. By analyzing the variable, we have identified four key factors: win/loss of the serve, rally length, serve speed, and distance. Moreover, we offer three practical recommendations for players entering a new match. Finally, we tested the Momentum Swings Prediction model on matches that were not included in the 2023 Wimbledon Men's model, and proposed incorporating player strength and injury information into future models to enhance the model's performance in different scenarios. Through testing on the 2022 Wimbledon Women's matches, we conclude that our model exhibits generalization across various types of matches. In conclusion, these findings highlight the significance of momentum and how it can impact the flow of play. By understanding these dynamics, coaches can better prepare players to respond to events that impact the match's trajectory.

Keywords: Momentum; Logistic Regression; VAR; Random Forest; Tennis.

1. Introduction

In the 2023 Wimbledon men's singles final, young tennis prodigy Carlos Alcaraz withstood the pressure to defeat defending champion Novak Djokovic, marking the end of an era. Despite Djokovic's initial dominance, Alcaraz staged a remarkable comeback, winning a tense tie-break in the second set and continuing to build momentum. Djokovic briefly regained control in the fourth set, but Alcaraz ultimately emerged victorious with a score of 6-4 in the deciding set.

The concept of "momentum" plays a crucial role in sports events, influencing a player's performance and the score's fluctuations. However, capturing and comprehending this elusive phenomenon can be challenging, as it can stem from factors such as scoreboard advantages, psychological states, physical stamina, or even crowd support. By analyzing the impact of momentum in-depth, we can gain valuable insights to better understand and control the dynamics of a match.

Through our analysis and modeling of the Wimbledon Men's Singles Final, we aim to uncover the precise influence of momentum on the game's trajectory. This knowledge will enable us to offer more strategic advice to players and enhance our ability to predict match outcomes with greater accuracy. Considering the background information, we need to solve the following questions:

Can we measure the impact of momentum using indicators and develop a predictive model based on this analysis to capture the flow of play and identify the player performing better at any given time? Are swings in play and runs of success by one player random, or are they influenced by momentum? Is it possible to develop a model that can accurately predict when the flow of play will change from
favoring one player to the other? How can we advise a player going into a new match against a different player, considering the differential in past match momentum swings?

2. Research Design and Data Description

2.1 Research Design

Our aim is to analyze the dataset of Wimbledon 2023 Gentlemen's singles matches and address the following questions:

1. Measure the impact of "momentum" by analyzing relevant indicators such as serve advantage, serve success rate, percentage of breakpoints won, etc. We need to establish the correlation between these indicators and the match outcome and develop a predictive model based on this analysis. Additionally, we should capture the flow of the race and present the results using visualizations. This model will help us identify trends in player performance by using momentum as a metric.

2. Identify indicators that measure swings in play and runs of success by one player, and assess their randomness. We should construct a model to further investigate the dynamic correlation between these indicators and momentum. If the results show significance, it suggests that momentum plays a substantial role in swings in play and runs of success by one player.

3. Predict the swings in the match and determine the factors that are most closely related to these swings. Given the difference in past match "momentum" swings. Give a player some advice when entering a new match, to help him get into momentum well, and thus achieve better results.

4. Test the previously developed model in other matches while demonstrating how well we predict the swings in the match. If the model performs poorly at times, we should identify some factors that might need to be included in future models. Moreover, extend the model's application to other matches, tournaments, court surfaces and other sports and analyze the model's effectiveness to determine its generalization.

![Flow chart of our analysis](image)

Fig. 1 Flow chart of our analysis

2.2 Data Processing and Symbols

In order to analyze the fluctuations in player performance and the pace of the game in tennis, we identified several key variables based on the data provided. Table 1 presents 4 crucial dimensions — Consumption, Serving, Accuracy and Difference, which serve as indicators of player performance. Within these dimensions, we established a total of 15 independent variables. To ensure consistency
in our analysis, these variables have been centered and standardized. For different tasks, we choose
different dependent variables to construct the model, which are also shown in Table 1.

The dataset consists of 7,284 entries, reflecting the performance of individual players. Given that
each match involves two players, the original dataset contained 14,568 entries. After removing entries
with missing values, the dataset was reduced to 13,064 entries.

<table>
<thead>
<tr>
<th>Type</th>
<th>Symbols</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consumption</td>
<td>RC</td>
<td>Rallies of each point</td>
</tr>
<tr>
<td></td>
<td>Distance</td>
<td>Player's distance ran during point</td>
</tr>
<tr>
<td></td>
<td>ET</td>
<td>Accumulation time, seconds</td>
</tr>
<tr>
<td>Serving</td>
<td>Server</td>
<td>If a player is serving, it is marked as 1, otherwise it will be marked as 0.</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>Player's scoring rate in first serve each game</td>
</tr>
<tr>
<td></td>
<td>RC</td>
<td>Player's scoring rate in second serve each game</td>
</tr>
<tr>
<td></td>
<td>RB</td>
<td>Player's success rate when he has the opportunity to break his opponent's serve</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Depth</td>
<td>If a player serves or returns deeply, it is marked as 1, otherwise it will be marked as 0.</td>
</tr>
<tr>
<td></td>
<td>SM</td>
<td>Speed of serve</td>
</tr>
<tr>
<td></td>
<td>RN</td>
<td>Player's scoring rate in net each game</td>
</tr>
<tr>
<td></td>
<td>DS</td>
<td>The score difference in current match</td>
</tr>
<tr>
<td></td>
<td>DGC</td>
<td>The score difference in current set</td>
</tr>
<tr>
<td>Difference</td>
<td>DGL</td>
<td>The score difference in last game</td>
</tr>
<tr>
<td></td>
<td>DPW</td>
<td>The score difference in current game</td>
</tr>
<tr>
<td></td>
<td>Server</td>
<td>The count of consecutive wins or losses in the last three rallies, with consecutive wins being positive, consecutive losses being negative, and 0 indicating alternating wins and losses.</td>
</tr>
<tr>
<td>Dependent variables</td>
<td>win</td>
<td>Whether the player wins the point</td>
</tr>
<tr>
<td></td>
<td>MR</td>
<td>Momentum Ratio which is measured by odds.</td>
</tr>
<tr>
<td></td>
<td>PT</td>
<td>A logistic variable which represents for Momentum Swings</td>
</tr>
</tbody>
</table>

3. **Model1 - Momentum Prediction Model based on Logistic Regression**

3.1 **Model Selection and Exploratory Analysis**

To effectively capture the flow of play and analyze the win/lose scenario for each point, we create a dependent variable called win. This variable takes a value of 1 if the player wins a point and 0 if they lose. By tracking the number of rallies won, we can determine the player's performance and their momentum during the match.

Considering that win is a logistic variable, we select factors that may affect it, such as physical exertion, skill level, and point difference. We also take into account the significance of serving in tennis matches and incorporate serving variables when analyzing the match's pace.

In summary, we choose four dimensions and establish 15 variables (as shown in Table 1) to investigate their relationship with win. Given that win is logistic, we opt for a multivariate logistic model. One advantage of this model is its ability to provide probability estimates for each category of the dependent variable, which helps analyze the match flow. The probability of win being 1 represents the likelihood of the player winning a point, indicating better performance at that time.
Before using the multivariate logistic model, we check for multicollinearity among the independent variables. From Figure 2, we observe that only a few correlation coefficients exceed 0.8, while the rest of the variables show weak linear correlation. This indicates the absence of multicollinearity issues. Therefore, we can confidently use the multivariate logistic model to analyze the relationship between the independent variables and the win variable.

3.2 Model Building and Solving

The logistic regression model is represented by the formula:

\[
P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1X_1 + \beta_2X_2 + \ldots + \beta_kX_k)}}.
\]

Here, \(P(Y = 1|X)\) represents the probability of the dependent variable being 1, given the values of the independent variables \(X_1, X_2, \ldots, X_k\). The coefficients \(\beta_0, \beta_1, \beta_2, \ldots, \beta_k\) are estimated to determine the impact of the independent variables on the probability.

To build and solve our model, we first divide the dataset into two groups with a proportion of 70%. This division allows us to have a training set, which will be used to develop the model, and a testing set, which will be used to evaluate the model's performance.

3.2.1 Fit the full model

We implement logistic regression on our training dataset using the glm function in R. To test the null hypothesis, we compare the deviance of the full model to the deviance of the null model. In our analysis, the difference in deviance is statistically significant (\(p < 2.2\times10^{-16}\)), allowing us to reject the null hypothesis. This means that including the independent variables significantly improves the model fit compared to the null model.

After testing, we estimate the coefficients for each independent variable. These coefficients represent the impact of each variable on the log-odds of winning a point. To achieve this, we take the natural logarithm of both sides of the equation 1. The resulting equation, known as the log-odds model, can be written as:

\[
\log\left(\frac{P(Y = 1|X)}{1-P(Y = 1|X)}\right) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \ldots + \beta_kX_k.
\]

To visually present the estimated coefficients and their significance, we draw Figure 3, a Coefficient Plot of the Logistic Regression Model. Significant variables such as Server, DS and DGC.
have coefficients that significantly impact the log-odds of winning. However, variables like DGL and Distance do not show significant effects on the probability of winning.

![Fig. 3 Coefficients of the Full Model](image3)

### 3.2.2 Variable Selection

Next, we employ a stepwise approach, combining both forward and backward elimination techniques to select the most relevant variables. The coefficients of the reduced model are displayed in Figure 4. We proceed by evaluating the accuracy of the reduced model and comparing it to the full model, as shown in Table 2.

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Odds ratio</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Server</td>
<td>9144</td>
<td>2.77 (2.60, 2.96)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>DS</td>
<td>9144</td>
<td>0.64 (0.60, 0.70)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>DGC</td>
<td>9144</td>
<td>0.82 (0.74, 0.91)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>DGL</td>
<td>9144</td>
<td>1.01 (0.92, 1.10)</td>
<td>0.90</td>
</tr>
<tr>
<td>DPW</td>
<td>9144</td>
<td>2.06 (1.89, 2.25)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>RC</td>
<td>9144</td>
<td>0.92 (0.81, 1.03)</td>
<td>0.15</td>
</tr>
<tr>
<td>Distance</td>
<td>9144</td>
<td>1.11 (0.98, 1.25)</td>
<td>0.09</td>
</tr>
<tr>
<td>Depth</td>
<td>9144</td>
<td>1.25 (1.19, 1.31)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ET</td>
<td>9144</td>
<td>1.02 (0.97, 1.07)</td>
<td>0.39</td>
</tr>
<tr>
<td>PP</td>
<td>9144</td>
<td>0.89 (0.85, 0.94)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>CW</td>
<td>9144</td>
<td>0.98 (0.91, 1.06)</td>
<td>0.61</td>
</tr>
<tr>
<td>RF</td>
<td>9144</td>
<td>1.33 (1.26, 1.42)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>RS</td>
<td>9144</td>
<td>1.27 (1.20, 1.35)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>RN</td>
<td>9144</td>
<td>1.18 (1.12, 1.24)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>RB</td>
<td>9144</td>
<td>1.71 (1.61, 1.81)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>SM</td>
<td>9144</td>
<td>1.03 (0.98, 1.09)</td>
<td>0.20</td>
</tr>
</tbody>
</table>

![Fig. 4 Coefficients of the Reduced Model](image4)

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Estimate</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Server</td>
<td>9144</td>
<td>0.22 (0.21, 0.23)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>DS</td>
<td>9144</td>
<td>-0.08 (-0.10, -0.07)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>DGC</td>
<td>9144</td>
<td>-0.04 (-0.05, -0.03)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>DPW</td>
<td>9144</td>
<td>0.14 (0.12, 0.15)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>RC</td>
<td>9144</td>
<td>-0.02 (-0.04, 0.01)</td>
<td>0.1</td>
</tr>
<tr>
<td>Distance</td>
<td>9144</td>
<td>0.02 (-0.00, 0.04)</td>
<td>0.1</td>
</tr>
<tr>
<td>Depth</td>
<td>9144</td>
<td>0.04 (0.03, 0.05)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>PP</td>
<td>9144</td>
<td>-0.02 (-0.03, -0.01)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>RF</td>
<td>9144</td>
<td>0.06 (0.05, 0.07)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>RS</td>
<td>9144</td>
<td>0.05 (0.04, 0.06)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>RN</td>
<td>9144</td>
<td>0.03 (0.02, 0.04)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>RB</td>
<td>9144</td>
<td>0.11 (0.10, 0.13)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

![Table 2 Model Performance](image5)

<table>
<thead>
<tr>
<th>Model</th>
<th>Training Set</th>
<th>Testing Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>0.7334</td>
<td>0.7383</td>
</tr>
<tr>
<td>Reduced</td>
<td>0.7292</td>
<td>0.7337</td>
</tr>
</tbody>
</table>

Upon evaluation, we find that the reduced model performs slightly worse than the full model. Therefore, we have decided to select the full model as our final choice. This decision is based on the fact that the full model includes all the relevant variables, providing a more comprehensive understanding of the factors influencing the outcome variable. Additionally, the full model demonstrates superior accuracy on both the training and testing datasets, indicating its stronger predictive capability.

### 3.2.3 Threshold Selection: Enhancing Predictive Power

Additionally, we examine the selection of the optimal threshold for our model. The figure 5 illustrates the relationship between different decision thresholds and the sum of prediction errors. The graph reveals that a threshold of 0.55 yielded the lowest error count, suggesting its effectiveness in improving the model's performance.

![Fig. 5 Errors in Different Thresholds](image6)
3.3 Analysis and Visualization of Results

To analyze the match flow, we selected a specific match and used logistic regression to calculate the odds for each point. The resulting match flow visualization (Figure 6) displays the dynamics of the match, with the x-axis indicating the points and the y-axis showing the corresponding odds for each player.

![Fig. 6 Match Flow Visualization by Odds](image)

This visualization allows us to identify key moments of momentum shifts and advantages between the players. By examining the plot, we can observe the intersecting lines and fluctuations in odds, indicating changes in momentum. The odds, derived from our logistic regression model, accurately represent the players' chances of winning a point.

The odds, derived from our logistic regression model, accurately depict the probabilities of each player winning a point. The smooth curves, generated using LOESS, provide a cumulative representation of the match flow. In this particular match, we can see Alcaraz's odds smoothing curve (green curve) surpassing that of the opponent in the end, leading to his victory. This demonstrates the effectiveness of our model in capturing the momentum changes in tennis matches.

![Fig. 7 Match Flow Visualization by Odds](image)

To look at the correspondence in detail, we choose 3 sets from the match. The match unfolded fascinatingly. In set 1, Djokovic effortlessly dominated, winning 6-1, as depicted by the blue curve in the graph being significantly higher than the green curve. Moving on to set 4, Alcaraz initially took control, resulting in the green curve being higher than the blue curve at the beginning. However, the match took a turn, and as shown in the graph, the blue curve surpassed the green curve, reflecting Djokovic's resurgence. In set 5, Djokovic started with a lead, with the blue curve being on the top. Nevertheless, Alcaraz managed to gain control and ultimately secured victory, which aligns with the green curve overtaking the blue curve in the graph.
4. Model2 - Dynamic Correlation Model based on VAR

4.1 Randomness Analysis of DPW and CW based on Run test

**Definition of indicators.** This question requires quantifying momentum and analyzing the dynamic correlation between momentum and swings in play, momentum and runs of success by one player. Then determining whether these two metrics are affected by momentum, indicating a not random situation. The metrics are defined below.

1. MR: Momentum Ratio, measured by the odds obtained from Model1. This metric takes values from 0 to positive infinity, with 1 as the cutoff point. A Momentum Ratio greater than 1 indicates that the player is more likely to win; the opposite is true for less than 1.
2. DPW: The score difference between two players in the current game, measuring the swings in play.
3. CW: The number of consecutive games won by a player, representing the runs of success by one player.

**Randomness analysis on Run test.** To examine the randomness of swings in play and runs of success by a single player, we initially conduct a Run test for both CW and DPW. The Run test is a non-parametric statistical method used to test for the presence of randomness in data. By comparing the difference between runs and expected runs, we can determine the randomness of the data. A bar graph plotting runs and expected runs is shown below.

![Bar graph for Alcaraz](image1)

![Bar graph for Djokovic](image2)

The graphs visualize that the actual runs are all smaller than the expected runs, suggesting a strong indication that both CW and DPW do not occur randomly. And the results of the Run test are shown in the table.

<table>
<thead>
<tr>
<th></th>
<th>Alcaraz</th>
<th>Djokovic</th>
</tr>
</thead>
<tbody>
<tr>
<td>CW</td>
<td>131</td>
<td>130</td>
</tr>
<tr>
<td>DPW</td>
<td>125</td>
<td>115</td>
</tr>
<tr>
<td>Expected runs</td>
<td>160.77</td>
<td>160.77</td>
</tr>
<tr>
<td>162.90</td>
<td>160.53</td>
<td></td>
</tr>
<tr>
<td>Variance runs</td>
<td>78.54</td>
<td>78.54</td>
</tr>
<tr>
<td>80.65</td>
<td>78.29</td>
<td></td>
</tr>
<tr>
<td>Z score</td>
<td>-3.36</td>
<td>-3.47</td>
</tr>
<tr>
<td>-4.22</td>
<td>-5.15</td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>0.001***</td>
<td>0.001***</td>
</tr>
<tr>
<td>0.000***</td>
<td>0.000***</td>
<td></td>
</tr>
</tbody>
</table>

The table clearly indicates that all four Run tests are significant, leading us to reject the null hypothesis. Consequently, we conclude that neither CW nor DPW occurs randomly and that the tennis coach’s claim may be open to examination. To further illustrate the MR performance of runs of success by one player at different levels, we also plotted a Mountain Range plot of the MR metrics on CW, which is displayed below.
Fig. 10 Mountain Range plot

The Mountain Range plot reveals an uneven distribution of the data, thus tentatively suggesting that the effect of momentum may be different for different levels of CW. It also gives a good indication of the differences between the two players’ respective situations, with Djokovic’s own situation fluctuating more relative to Alcaraz.

4.2 Modeling and Solving for Model2

In order to explore the dynamic correlations between MR and CW, MR and DPW, we utilize vector auto-regressive models (VAR), using data from Alcaraz and Djokovic’s match.

- **ADF test.** The ADF test was performed on each time series variable before the VAR model was built. We found that the time series of the respective indicators from both players are stable, enabling us to proceed with building the VAR model.

- **Determine the lag order.** To determine the optimal lag order, we employ two information criteria: AIC (Akaike Information Criterion) and SC (Schwarz Criterion). These criteria are utilized to minimize the values of AIC and SC as the lag order, denoted as k, is increased. The two information criterion formulas are calculated as follows:

$$\text{AIC} = \log \left( \frac{\sum_{t=1}^{T} \hat{u}_t^2}{T} \right) + \frac{2k}{T} \quad \text{SC} = \log \left( \frac{\sum_{t=1}^{T} \hat{u}_t^2}{T} \right) + \frac{k \log T}{T}.$$  \hspace{1cm} (3)

The case of Alcaraz is analyzed and it’s observed that the AIC and SC values of the model of MR and CW are minimized at k=1. Similarly, the AIC value of the model of MR and DPW is minimized at k=4 and the SC value is minimized at k=1. Since a large k leads to a reduction in the degrees of freedom and affects the validity of the parameter estimation, we choose k=1 for our analysis.

Continuing with this analysis, the case of Djokovic is analyzed and it’s observed that the SC values of both the model of MR with CW and the model of MR with DPW are minimized at k = 1. Therefore, we choose k=1 for our analysis in the case of Djokovic as well.

- **Modeling and Parameter Estimation.** A VAR model of MR with CW lagged by 1 period is first developed.

$$\begin{align*}
\{CW_t &= \mu_1 + \pi_{11}CW_{t-1} + \pi_{12}MR_{t-1} + u_{1t} \\
MR_t &= \mu_2 + \pi_{21}CW_{t-1} + \pi_{22}MR_{t-1} + u_{2t}.
\end{align*}$$  \hspace{1cm} (4)

Here $u_{1t}, u_{2t} \sim IID(0, \sigma^2), \text{Cov}(u_{1t}, u_{2t}) = 0$. Similarly, we could also develop VAR model of MR with DPW lagged by 1 period, which is as the same form as with CW.

(1) Solving the VAR model for Alcaraz yields:

$$\begin{align*}
\{CW_t &= 0.327 + 0.296CW_{t-1} + 0.069MR_{t-1} \\
MR_t &= 1.134 - 0.131CW_{t-1} + 0.049MR_{t-1}.
\end{align*}$$  \hspace{1cm} (5)
\[
\begin{align*}
\{ & \text{DPW}_t = -2.971 + 0.265\text{DPW}_{t-1} + 0.377\text{MR}_{t-1} \\
& \text{MR}_t = 1.015 - 0.003\text{DPW}_{t-1} + 0.081\text{MR}_{t-1} \}.
\end{align*}
\]

The eigenroots of the two models were calculated to be 0.09 and 0.25, 0.09 and 0.26 respectively. The eigenroots are all located within the unit circle, indicating that the VAR model is stable.

(2) Solving the VAR model for Djokovic yields:

\[
\begin{align*}
\{ & \text{CW}_t = 0.468 + 0.271\text{CW}_{t-1} + 0.02\text{MR}_{t-1} \\
& \text{MR}_t = 1.782 + 0.172\text{CW}_{t-1} + 0.095\text{MR}_{t-1} \}.
\end{align*}
\]

\[
\begin{align*}
\{ & \text{DPW}_t = 2.06 + 0.322\text{DPW}_{t-1} + 0.134\text{MR}_{t-1} \\
& \text{MR}_t = 1.916 + 0.024\text{DPW}_{t-1} + 0.049\text{MR}_{t-1} \}.
\end{align*}
\]

The eigenroots of the two models are 0.08 and 0.29, 0.04 and 0.33 respectively, which are equally stable. The models can be further subjected to impulse response analysis and variance decomposition.

4.3 Analysis and Visualization of Results

• Impulse Response Analysis. To further explain the practical significance of the parameters in each VAR model, we observe the impulse response function and variance decomposition of the system. The impulse response function illustrates the response of an endogenous variable to an error shock, and it represents the effect of a shock of one standard deviation in size acting on the random error on the current and future values of the endogenous variable. Through the lag operator, we can write the above VAR(1) model in the form of VMA(∞):

\[
y_t = \mu + \sum_{i=0}^{\infty} \psi_i u_{t-i}.
\]

At this point the impulse response function (IRF) is a function of the time interval, denoted as and describes the response process of to a single shock to \( y_{i,t} \), while holding other variables constant in period , as well as in previous periods. The expression is:

\[
\frac{\partial y_{i,t+s}}{\partial u_{it}}, s = 1,2,3,....
\]

• Analysis of Results. The impulse response functions of Alcaraz and Djokovic are plotted separately in figure 11 and figure 12. The impact of the MR shock on CW and DPW within the 10th order of lag is shown. A side-by-side comparison reveals that the states of both Alcaraz and Djokovic are affected by momentum in roughly the same trend. Both players’ winning streaks are most affected by their momentum in the previous one period, and as the lag order increases, momentum in the last 5 games gradually converges to 0 on a player’s winning streak.
A comparison between two players shows that Alcaraz is significantly influenced by momentum, as evident from the wider range of vertical coordinates in the left graph. A comparison between different indicators for the same player reveals that the indicator DPW is more affected by momentum than CW. Thus, we suggest that momentum is more likely to influence shifts in game situations.

5. Model 3 - Momentum swings Prediction model based on Random Forest

5.1 Variable Creation and Model Selection

5.1.1 Variable Creation

To effectively predict the swings in the match, we first establish a variable called TP (Turning Point) to accurately measure the swings in the match. TP signifies the point at which the flow of play is about to transition from favoring one player to the other. The creation of TP is illustrated in Figure 13.

Firstly, we choose 3 points as the standard of judgement because if one players wins 3 points in a row, it indicates a significant shift in momentum and often lead to at least a draw in one game, placing them/he in a favorable position. Conversely losing three points in a row has the negative result.

Next, we filter matches that end with a score of 3-2 or 2-3, aiming to control the difference in player’s strength between the players in the match. In these matches, the impact of player entering Momentum phase has a greater influence on the match outcome.

Once the conditions are met, we assign a value to TP. If a player wins at least 3 points in a row after losing one point, the TP value for the first win in the sequence is recorded as 1. On the other hand, if a player wins at least 3 points in a row after losing one point, the TP for the first win in the sequence is recorded as -1. In short TP=1 means the player is in Momentum whereas TP=-1 means the player is out of Momentum.

5.1.2 Model Selection

It is evident that TP is a logistic variable. In order to predict the change of TP and analyze what factors are most related, we use Random Forest model to classify and make predictions, with TP as the dependent variable and 15 independent variables from four dimensions in Table 1. The Random Forest model combines the prediction results of multiple decision trees to improve the overall prediction ability. This model is well-suited for capturing nonlinear relationships and doesn’t necessitate explicit assumptions. These characteristics make it an ideal choice for addressing the situation presented in Problem 3.

It should be noted that we chose the multivariate logistic model for similar problems in Model 1. This decision was motivated by our primary focus on the odds and the intention to capture the player
's momentum through the odds. However, it requires greater emphasis on predicting the swings in the match.

It has a better effect on the classification and shows better robustness when confronted with noisy data and outliers, enabling us to better predict the player's momentum swings in the tennis match.

5.2 Model Principle and Momentum Swings Prediction

5.2.1 Model Principle

Random forest is an integrated learning algorithm and its principle in this problem can be expressed as follows.

1. From original training set using the self-sampling method, generate N new training datasets $D_1, D_2, \ldots, D_N$.

2. Train a decision tree for each decision tree $D_i$. In a Random Forest Model, each tree typically has different nodes. When constructing each node of the decision tree, the algorithm is to randomly select a subset of the independent variables with a specified try and chooses the optimal independent variables form this subset.

3. Each tree $T_i$ independently predicts the samples $x$ to vote for a category $C_i(x)$. In this problem, is 1 or -1. The final classification result is obtained by majority voting, where the highest number of votes is selected as the final prediction category. The formula for the voting process can be expressed as follows:

$$C(x) = \text{mode}\{C_1(x), C_2(x), \ldots, C_N(x)\}. \quad (11)$$

Here, mode indicates choosing the category that occurs the most times in all the prediction results.

5.2.2 Momentum Swings Prediction

A total of 628 data are collected in this model, with 614 remaining after removing missing values. Based on existing literature studies, this model takes a random sampling approach to divide the dataset into a training set (70% of data) and a test set (30% of data).

<table>
<thead>
<tr>
<th>Actual/Predicted</th>
<th>-1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>203</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>206</td>
</tr>
</tbody>
</table>

Table 4 Results of Random Forest model

We use R language to solve the random forest model and the results is shown in Table 4. The results indicate that, with a configuration of 300 trees and trying 6 variables at each split, the Random Forest model has an Out-of-Bag error estimate of 4.66%. The confusion matrix reveals that for instances belonging to class -1, the model accurately predicted 203 instances and misclassified 10 as class 1. Similarly, for instances of class 1, the model correctly predicted 206 instances and misclassified 10 as class -1. Overall, the model exhibits a high accuracy rate of 95.34%.

Furthermore, we apply test set on the model, and the model demonstrated a high accuracy of 93.51%. This outcome indicates that the Random Forest model possesses excellent predictive capabilities, consistently performing well on both the training and the test sets.

In order to better show the prediction process of Random Forest model, we visualize the categorization process of one decision tree. As shown in Figure 14 firstly the decision tree splits starting from the root node based on Server. Subsequently, each child node continues with further splits based on other variables such as RC, RF, etc. Each split is based on the optimal split point to maximize the increase in the purity of the child nodes as much as possible.

In the Random Forest model, similar decision trees are constructed several times, and the prediction process for each tree is based on a random subset drawn from the original dataset and a subset of the random variables considered in each split.
5.3 Related Factors with Momentum Swings

In a Random Forest model, the importance of an independent variable indicates its contribution to the accuracy of the model's predictions. Variables with high importance are considered to be strongly related to Momentum Swings. There are four variables whose importance is above 20. Specifically, server has the highest importance value of 52.34. Following that, RC ranks second, SM ranks third and Distance ranks fourth.

To further analyze the magnitude and direction of the impact of the above 4 variables on Momentum Swings, we can utilize the Partial Dependence Plot of the respective variables which can reflect the variable's average marginal impact on TP. As shown in Figure 16, Server has a high positive effect on TP for a value of 1, and a high negative effect for a value of 0.

Therefore, mastery of the serve has a key role in getting the player into Momentum. The effects of the remaining 3 variables on TP show nonlinear relationship. High values of RC and DPW have negative impact on TP, suggesting that they take the player out of Momentum at high values. Conversely, in most cases Distance has a positive impact.
All in all, the most influential factor with swings in the match are Server, RC, DPW and Distance. Practically, the win or loss of the serve has the biggest impact on the swings. Proper point difference and rallies help a player get into Momentum, whereas they have negative impact when they are too high. Moderate runs can also help a player’s get into momentum.

5.4 Momentum Test on Other Matches

Due to previous considerations of controlling player’s strength differentials, the model is built on only matches whose set score ended with 3-2. Matches whose score ended with 3-0 or 3-1 are not considered. Therefore, we will test the effectiveness of the model on these data in a number of ways, analyzing how well this model predicts the swings in other matches.

In order to initially test the effectiveness of the model, firstly we calculated some performance evaluation metrics. The accuracy of the model for TP prediction is 81.50%, indicating that the model has a high overall prediction correctness on other datasets. Recall are calculated according to Formula 12. A recall of 0.86 indicates that the model has a high ability to recognize positive class samples.

\[
\text{Recall} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalseNegatives}}. \tag{12}
\]

Here Positive represents for TP=1 whereas negative represents for TP=-1. True and False represent whether the model prediction is true or not.

In order to fully assess the effectiveness of the model, we also plotted the ROC curve and calculated the AUC value. As shown in the Figure 17, the ROC curve rises rapidly and approaches the upper left corner, which implies that the model achieves a high rate of true classes while maintaining a low rate of false positive classes. The value of AUC is 0.94, which is close to 1, indicating that the model has a good performance in Momentum Swings prediction on other matches.

Overall, the validation results of the model on other matches have a certain decline compared with the original dataset, but still maintain a good prediction on Momentum Swings.

Fig. 17 ROC curve for other competitions’ results
6. Sensitivity Analysis

6.1 Cross-validation: Assessing Logistic Model Performance

To assess the robustness and reliability of Model 1 the logistic model, we perform a sensitivity analysis through cross-validation. The whole dataset is divided into 10 folds, and each fold is used as a testing set while the remaining folds are used for training.

![Cross Validation Results](image)

Based on the above Figure 18, the accuracy and F1 scores remain stable. Moreover, the accuracy obtained from the process are averaged, resulting in a mean accuracy of 0.7324 and an F1 score of 0.7370. These results indicate that our model performs well in predicting the momentum of a tennis match based on the selected variables. This sensitivity analysis provides valuable insights into the stability and consistency of the logistic model, reaffirming its reliability in capturing and analyzing the dynamics of a tennis match.

6.2 Parameter Values Modification: Assessing Random Forest Model

In order to test the stability of the Random Forest Model, we change values of Ntree and mtry respectively, the variation range of Ntree is 250:350, and the variation range of mtry is 3:9. According to the changed parameter values, we re-establish the random forest model, compute its error, and plot the curve of the variation with the parameter values. From Figure 19 and Figure 20, it can be seen that the model ERROR rate with the parameter values is always stable at less than 5%, and the range of changes are small, indicating a high level of stability.
7. **Suggestions**

Given the differential in past match "momentum" swings, we offer three following advice based on our findings for players when going into a new match:

1. **Focus on winning the serve and enhancing serving skills.** Winning the serve or delivering a high-speed ball is conducive to get into momentum. Players should strive to improve the quality of their serves by analyzing their own performances in previous matches. By winning the opponent’s serve and disrupting their game rhythm, players can effectively counter their opponent’s momentum.

2. **Do not dwell too much on the point difference.** Our model indicates that a significant point difference can hinder players from gaining momentum. However, it is important for players to remember that they cannot change the point difference of the previous game. Instead of being influenced by past failures, players should focus on the next points and maintain their confidence. Even when the score is down, players need to play at their normal level, avoiding any interference with their rhythm of play.

3. **Manage consumption and run strategically.** According to our model too much consumption has disadvantages for the player to enter the momentum state, whereas proper running can help the player to enter the momentum. In the second half of the game the player will inevitably decline, and it is hard to maintain his technical accuracy. What the player has to do is to make a reasonable distribution of physical strength, he can change his strategy in the new game according to his physical consumption in the previous game. Additionally, players should pay attention to their running style and avoid unnecessary energy consumption by minimizing excessive runs.

**References**


