Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 7, No. 2, 2024

Football Momentum Analysis based on Logistic Regression

Zilu Wen [1], Jinyu Liu [1], Chenxi Liu [2]

[1] Department of electrical information, Shandong University of Science and Technology, Jinan, Shandong, 250031, China
[2] Swinburne College, Shandong University of Science and Technology, Jinan, Shandong, 250031, China

Abstract: In tennis, momentum is pivotal and can be quantified using metrics like Consecutive Win Rate (CWR), Unforced Error Rate (UER), Break Point Save Rate (BPSR), and Fatigue Factor (FF). Each metric provides insight into a player's performance and state during a match. CWR is a clear momentum indicator, reflecting a player's game dominance, while UER highlights potential lapses in concentration or physical condition. BPSR evaluates a player's clutch performance in critical situations, and FF gauges physical exertion. Utilizing logistic regression, we can predict a player's probability to win at any scoring point, incorporating these metrics as variables. The coefficients obtained from MATLAB analysis (e.g., p1_cwr at 22.73 and p2_ff at -3.26) reveal the positive or negative correlation of these factors with a player's winning chances. In the case of the "2023-wimbledon-1301" match, the logistic model's predictions showed a symmetrical distribution of win probabilities between players, suggesting a balance in momentum swings throughout the match. Initial volatility in Player 1's success rate indicated a strong start, which diminished over time, possibly due to fatigue or the opponent's improving performance. Despite the fluctuations and a period of deadlock, Player 1's consistent performance and superior win rate for most of the game secured the victory. This outcome underscores the importance of maintaining momentum and physical resilience in tennis.

Keywords: Momentum; Logistic Regression; Model Parameters; Win Probability Prediction; Match Analysis.

1. Introduction

In the men's final of the 2023 Wimbledon Open, 20-year-old Carlos Alcaraz narrowly edged out Novak Djokovic, who had been undefeated at the event for ten years, in a match that was thought to be a non-starter, but the scoreline took an almost bizarre turn for the worse. Djokovic easily won the first set 6-1, and in the following sets, there were many reversals, notably Carlos Alcaraz in the second and third sets, and then the fourth and fifth sets were tightly contested. This strange phenomenon is known as a change in "momentum" and was not only observed in this final, but also in other classic matches, such as the semifinal match between China and Japan in the 2023 Sudirman Cup, and the final match of the women's volleyball team in the 2016 Rio Olympics, among others. volleyball women's final, etc. [1][5]. In competitive sports events, changes in "momentum" often lead to unpredictability in the direction of the game. The purpose of this paper is to explore the factors affecting the generation of "momentum" by analyzing various data in tennis matches, and to propose methods to rationally utilize "momentum" to improve the probability of tennis players winning the matches [6].

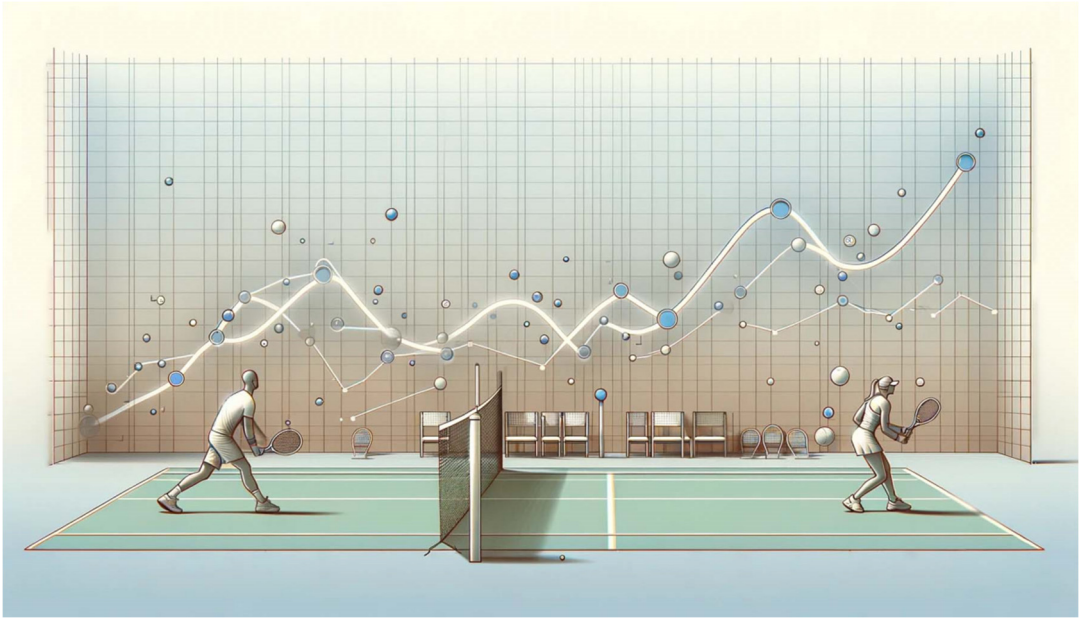

Figure 1. Diagram of Momentum Changes in Tennis Competitions

2. Data Description

The data we use includes the data files given as *Wimbledon_featured_matches.csv*. This file contains almost all the information we need to address the problem. However, before applying it, the data needs to be reprocessed.

Firstly, in a tennis match, when both sides score 40-40 (i.e. Reduce), the winner of the next point will gain an 'advantage' (marked as 'AD'). If the player with an advantage wins another point, he/she will win the game; If the opponent wins the next point, both sides return to Deuce [7][10]. In data analysis, we need to convert 'AD' scores into numerical data, and categorical variables like this need to be converted into numerical variables. We will directly represent the normal scores (0, 15, 30, 40) with their numerical values and assign a value to the 'AD' score. Due to the fact that 'AD' represents a score higher than 40 but not enough to directly win the game, we can choose to represent 'AD' as 45, which not only preserves the order of scores but also avoids confusion with normal scores, and can also assess the anxiety of the game [11].

We not only filled in the Missing values in the *speed_mph* numerical column, and filled with the categorical variable *serve_ Width*, *serve_ depth*, *return_ depth*. Columns like Y and those related to hitting speed, as they did not hit the ball, were filled in as 0. Furthermore, no abnormal values were found and no relevant measures will be taken.

3. Dynamic Scoring System

In predicting a player's winning percentage, a large number of variables can extremely increase the computational time-consuming as well as the model complexity, so the selection of important variables is indispensable (46 variables including match_id, player1, set_no, etc.). Therefore, we choose the more objective and accurate random forest for importance selection, and according to the different weights from large too small to draw the ranking chart. The selection of the importance of variables is closely related to the Gini index [12].

Random forest models primarily use the Gini index to measure the impurity of the set and thus assess the importance of each feature in the model. The importance of each variable is based on the extent to which it reduces the Gini impurity in constructing the model, which is calculated as follows:

$$I(A) = \sum_{t \in T_A} \frac{|D_t|}{|D|} \left(Gini(D_t) - Gini_{split}(D_t, A) \right) \quad (1)$$

Where T_A is the set of all decision tree nodes segmented using feature A, and D_t is the data set at node t.

The weights of the importance of the features can be obtained and plotted in Python, and the weights of the variables whose importance is in the top five variables and the ranking plot will be shown below in Figure 2:

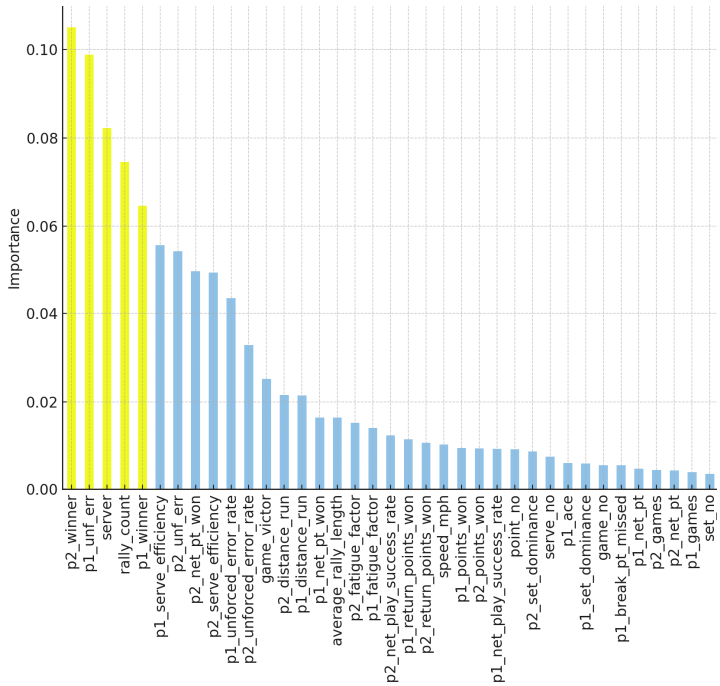

Figure 2. Ranking of the importance of the variables

From the figure, it can be observed that the five variables *p2_winner, p1_unf_err, server, rally_count, and p1_winner* have the greatest influence on the win rate, which is regarded as the main influence, and all the other variables are regarded as secondary influences.

4. Establishment of Momentum Indicators

In tennis, momentum is a multi-dimensional concept that can be understood as a player's or team's ability to achieve continuous success over a period, which may be manifested as winning multiple points in a row, breaking serve, or consecutive service holds [13]. Therefore, based on the top 5 important metrics in terms of importance, we can set further momentum indicators.

CWR (Consecutive Win Rate)

This indicator reflects a player's ability to win consecutive points and is a direct reflection of momentum. A high consecutive scoring rate indicates that a player has dominated the game for a certain period, demonstrating positive momentum.

For player p, the consecutive winning percentage at each

point_no can be expressed as:

$$CWR_p(i) = \frac{\sum_{k=1}^{i} (victor_{k-1} = victor_k = p)}{i} \quad (2)$$

where i represents the index of point_no, p denotes player 1 or player 2, $victor_k$ denotes the winner of the kth point, and 1 is an indicator function that takes the value 1 when the internal condition is true and 0 otherwise [14].

UER (Unforced Error Rate)

Unforced errors are often seen as a sign that a player is losing focus or fitness and can lead to a loss of momentum.

$$UER_p(i) = \frac{\Sigma_{k=1}^{n} unf_{p,k}}{i} \quad (3)$$

where $unf_{p,k}$ denotes the number of unforced errors made by player p in the kth point.

BPSR (Break Point Save Rate)

In tennis, Break Points are critical scoring opportunities that often determine the outcome of a service game.

$$BPSR_p(i) = \frac{\Sigma_{k=1}^{i} bp_{p,k}}{\Sigma_{k=1}^{i} 1(bp_{p,k})} \quad (4)$$

where the denominator is the total number of break points faced by player p and $bp_{p,k}$ is the number of break points saved by player p in the kth point [15].

FF (Fatigue Factor)

The fatigue factor reflects a player's physical exertion and durability during a match and can be a sideways reflection of a player's physical condition.

$$FF_p(i) = \frac{\sum_{k=1}^{n} dis_{p,k}}{\Sigma_{k=1}^{n} \text{rallies}_k} \quad (5)$$

where $dis_{p,k}$ denotes the distance run by player p in the kth score, and rallies_k denotes the number of overs in the kth score.

5. Logistic Regression Model

After establishing the momentum indicator, we can further simulate the process of one or more matches based on the momentum indicator and the winners of the matches at each scoring point, and predict the winning probability of a player at a particular scoring point [16].

Logistic regression is a widely used binary classification model, which can necessarily predict the probability of winning or losing a tennis match, mainly through the logistic function to determine the probability of winning or losing, it is assumed that the probability of player 2 to win is P(Y=1), then:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_4 X_4)}}$$

At this point the characteristic variables $X1, X2, \ldots, X4$ represent the four main influencing factors, while $\beta 0, \beta 1, \ldots, \beta 4$ represent the respective corresponding model parameters, which are solved by MATLAB and stored in the Table 1:

Table 1. Model parameters for the four main influencing factors

Variables	Model parameter
p1_cwr	22.73
p2_cwr	-11.74
p1_uer	11.91
p2_uer	-4.66
p1_bpsr	-0.24
p2_bpsr	-0.07
p1_ff	1.88
p2_ff	-3.26

A positive coefficient means that the characteristic variable is positively correlated with the score, while a negative coefficient is negatively correlated with it. For example, a high positive coefficient for *p1_swr* means that the winning streak is positively correlated with winning match points when Player 1 has a winning streak. the fatigue factor coefficient for Player 2 is -3.255, which may indicate that Player 2's fatigue factor has some negative effect on his winning match points.

The probability of player 1 and player 2 winning the game "2023-wimbledon-1301" with consecutive *point_no*'s is shown below in Figure 3:

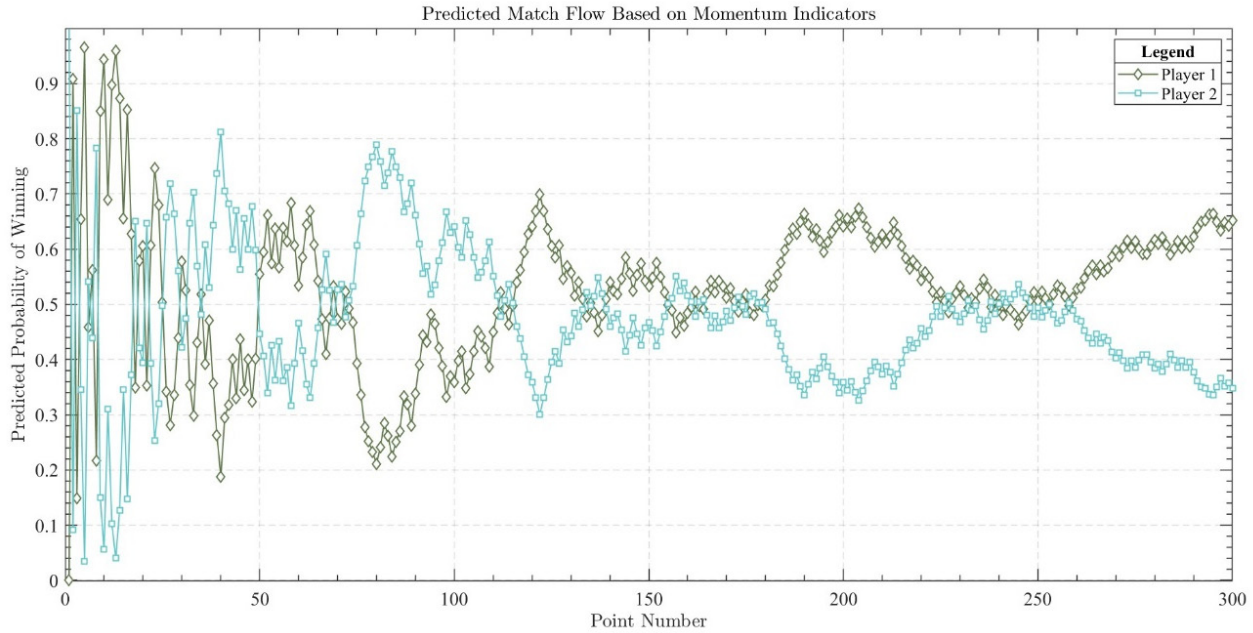

Figure 3. Probability of player 1 and player 2 winning the game

Upon scrutinizing Figure 4, a symmetrical distribution in the win rates of the contenders is discernible, indicating an

inverse relationship between the two metrics. Initially, Player 1 exhibited substantial volatility in his probability of success, embarking on the match with a commendable performance which waned progressively, suggesting either a decline in stamina or an ascendency in his adversary's play. The match subsequently evolved into a deadlock, with neither player able to secure a definitive advantage. Nonetheless, Player 1 ultimately clinched victory, having maintained a superior win rate for the majority of the contest. This observation may be indicative of a consistently better overall performance throughout the duration of the match.

6. Conclusion

The analysis of football momentum using logistic regression, as interpreted from Figure 4, yields a profound insight into the dynamic interplay between opposing teams' win probabilities over the course of a match. The symmetrical distribution of win rates observed suggests a zero-sum nature of the competition where the increase in one player's likelihood of winning inversely affects the other's chances. Player 1's initial erratic performance, with a high degree of fluctuation in win probability, points to a strong opening in the game that diminished over time, possibly due to a decline in physical stamina or the opponent's improved strategy and form. The logistic regression model captures this turning point and the ensuing equilibrium phase, where a statistical stalemate reflects the evenly matched contest. This deadlock situation, a common occurrence in football matches, is effectively represented by the stability in the predicted win rates, showing the model's ability to capture the changing dynamics as the players' performances plateaued.

Despite the stalemate, the eventual triumph of Player 1, as revealed by the model, underscores the importance of sustaining a higher win probability over the match's duration. This consistent advantage, even in the face of decreasing margins, highlights Player 1's resilience and perhaps a superior tactical approach, which logistic regression can quantify and use to predict outcomes. The conclusion drawn from the model's analysis is that while fluctuations in win probability can indicate temporary shifts in a match's momentum, the overall higher win rate maintained by Player 1 suggests a stronger performance that ultimately led to victory. In essence, logistic regression offers valuable quantification of the momentum swings within a football match, serving as a potent tool for analyzing performance, predicting outcomes, and understanding the complexities of competitive sports dynamics.

7. Discussion

Logistic regression is widely used in sports data analysis, especially when analysing momentum in sports such as tennis, where the model can reveal key trends and turning points in the game. Momentum, as a multidimensional concept, not only reflects a player's ability to achieve consecutive successes over a period of time, such as consecutively winning multiple points, breaking serve, or consecutively retaining serve, but it is also reflected in specific statistical indicators. Important momentum indicators include consecutive win rate (CWR), unforced error rate (UER), break point protection rate (BPSR) and fatigue factor (FF). These metrics quantify a player's performance and are further modelled by logistic regression to predict win rates on specific points.

A logistic regression model uses a logistic function to estimate a player's win rate at a particular point, converting a range of influencing factors such as CWR, UER, etc. into probability values for win rates. For example, in tennis, a player's ability to score consecutive points (CWR) is a direct reflection of his or her dominance of the game, while a high winning streak indicates that the player has positive momentum. Unforced errors (UER), on the other hand, are signs that a player may be losing focus or becoming less fit and can lead to a loss of momentum. Break Point Protection Rate (BPSR) reflects a player's ability to resist pressure on key scoring points, while Fatigue Factor (FF) is a side-by-side reflection of a player's endurance and physical exertion during a match.

The estimation of model parameters was performed via MATLAB, with each variable corresponding to a coefficient, where a positive coefficient indicates that the characteristic variable is positively correlated with scoring, and a negative coefficient indicates a negative correlation. For example, Player 1's winning streak (p1_cwr) has a high positive coefficient, which indicates that when Player 1 scores consecutive points, this is positively associated with winning match points. On the other hand, Player 2 has a negative coefficient of fatigue (p2_ff), which may indicate that Player 2's fatigue has some negative impact on his winning match points.

Through the logistic regression model, we are able to observe the changes in the winning percentage during the match. In the specific case of "2023-wimbledon-1301", by analyzing Figure 4, we can see that the distribution of the winning percentage of the two players is symmetric, which indicates that there is an inverse relationship between the two metrics. player 1's winning percentage fluctuates greatly at the beginning of the match, but gradually decreases as the match progresses, which may indicate a decrease in his stamina or an increase in his opponent's performance. The match eventually became a stalemate, with neither side able to secure a clear advantage. However, Player 1 maintained a high winning percentage for most of the match and eventually won, which may indicate that his overall performance was consistently better throughout the match.

Overall, logistic regression provides us with a powerful tool to quantify and predict momentum changes in tennis matches. With accurately calculated model parameters, we can better understand each turning point in a match and how different momentum metrics affect a player's performance and winning percentage during a match. Such analyses are very important for coaches and players to develop tactics and improve training and match strategies.

References

[1] Ötting, M., Langrock, R., & Maruotti, A. (2021). A copula-based multivariate hidden Markov model for modelling momentum in football. AStA Advances in Statistical Analysis, 1-19.

[2] Jamil, M., Liu, H., Phatak, A., & Memmert, D. (2021). An investigation identifying which key performance indicators influence the chances of promotion to the elite leagues in professional European football. International Journal of Performance Analysis in Sport, 21(4), 641-650.

[3] Gifford, M., & Bayrak, T. (2023). A predictive analytics model for forecasting outcomes in the National Football League games using decision tree and logistic regression. Decision Analytics Journal, 8, 100296.

[4] Cox, J., Schwartz, A. L., Van Ness, B. F., & Van Ness, R. A. (2021). The predictive power of college football spreads: Regular season versus bowl games. Journal of Sports Economics, 22(3), 251-273.

[5] Augustus, S., Hudson, P. E., & Smith, N. (2021). The effect of approach velocity on pelvis and kick leg angular momentum conversion strategies during football instep kicking. Journal of Sports Sciences, 39(20), 2279-2288.

[6] Krolo, A., Gilic, B., Foretic, N., Pojskic, H., Hammami, R., Spasic, M., ... & Sekulic, D. (2020). Agility testing in youth football (soccer) players; evaluating reliability, validity, and correlates of newly developed testing protocols. International journal of environmental research and public health, 17(1), 294.

[7] Liu, T., García-de-Alcaraz, A., Wang, H., Hu, P., & Chen, Q. (2021). Impact of scoring first on match outcome in the Chinese Football Super League. Frontiers in Psychology, 12, 662708.

[8] Giannakoulas, N., Papageorgiou, G., & Tjortjis, C. (2023, June). Forecasting Goal Performance for Top League Football Players: A Comparative Study. In IFIP International Conference on Artificial Intelligence Applications and Innovations (pp. 304-315). Cham: Springer Nature Switzerland.

[9] Guan, S., & Wang, X. (2022). Optimization analysis of football match prediction model based on neural network. Neural Computing and Applications, 1-17.

[10] Maerlender, A., Masterson, C. J., Norris, R., & Hinthorne, A. (2020). Validating tackle mechanics in American football: improving safety and performance. Annals of biomedical engineering, 48(11), 2691-2700.

[11] Mann, J. B., Bird, M., Signorile, J. F., Brechue, W. F., & Mayhew, J. L. (2021). Prediction of anaerobic power from standing long jump in NCAA division IA football players. The Journal of Strength & Conditioning Research, 35(6), 1542-1546.

[12] Tian, C., Zhou, Q., & Yang, B. (2022). Reform and Intelligent Innovation Path of College Football Teaching and Training Based on Mixed Teaching Mode. Mobile Information Systems, 2022.

[13] Matsuo, H., Funasaki, K., & Yamada, S. (2022). The Applicability of the Risk Score Approach to Competitive Sport: Development of a Physical Success Score for the Canadian Football League Combine. Journal of Strength and Conditioning Research, 36(3), 695-701.

[14] da Costa, I. B., Marinho, L. B., & Pires, C. E. S. (2022). Forecasting football results and exploiting betting markets: The case of "both teams to score". International Journal of Forecasting, 38(3), 895-909.

[15] Ab Rasid, A. M., Muazu Musa, R., Abdul Majeed, A. P., Musawi Maliki, A. B. H., Abdullah, M. R., Mohd Razmaan, M. A., & Abu Osman, N. A. (2024). Physical fitness and motor ability parameters as predictors for skateboarding performance: A logistic regression modelling analysis. PLoS one, 19(2), e0296467.

[16] Arboix-Alió, J., Aguilera-Castells, J., Buscà, B., Miro, A., Hileno, R., Trabal, G., & Pena, J. (2021). Situational variables in elite rink hockey: effect of match location, team level, scoring first and match status at halftime on the competitive outcome. International Journal of Performance Analysis in Sport, 21(6), 1101-1116.