Automation and Machine Learning (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2024.050120
ISSN 2516-5003 Vol. 5 Num. 1

Data-Driven Insights into Tennis Match Momentum: A Predictive Study from the 2023 Wimbledon Championships

Xinyu Lu[1,#], Qilong Li[2,#], Shuo Yuan[3,#], Yucheng Mo[4,#]

[1]*College of Control Science and Engineering, China University of Petroleum, Qingdao, 266580, China*
[2]*School of Economics and Trade, Hunan University of Technology, Zhuzhou, 412007, China*
[3]*School of Geosciences, China University of Petroleum, Qingdao, 266580, China*
[4]*School of Economics and Management UPC, China University of Petroleum, Qingdao, 266580, China*
[#]*These authors contributed equally.*

Keywords: Momentum; Logistic Regression; EMA; Run Test

Abstract: Tennis, as a popular sport around the world, has become the focus of sports data science research. The quantification of players' momentum and the prediction of key swings in tennis matches are of great significance for mastering the dynamics of matches and improving the performance of players. Based on the men's singles data of the 2023 Wimbledon Tennis Championship, this study uses comprehensive mathematical methods and models, including confusion matrix, Logistic regression, exponential moving average (EMA), Bessel curve fitting, and run test, and is implemented by Python and MATLAB to obtain the probability of players winning at any time. Momentum is continuous, and the momentum related to the winning streak and turning point of the game is verified by running tests. This paper offers a research direction for studying the fluctuation of player momentum during a match, aiming to quantify player momentum and predict the likelihood of victory in a match.

1. Introduction

The dynamic and unpredictable nature of tennis makes it a challenge to quantify and predict it. Each player's performance in the competition can be affected by a variety of factors, including technique, strategy, physical fitness, and mental state. One particularly critical concept is momentum, which reflects shifts in power and emotion during a match and can greatly influence the direction and outcome of the match. As for the momentum research on sports events, there have been relevant momentum studies based on basketball events before. Therefore, accurately quantifying momentum and incorporating it into match prediction models is of great significance for understanding and predicting match results.

This study employed logistic regression for analysis. Logistic is a generalized linear regression analysis model, which is often used in data mining, economic forecasting, and other fields to estimate

the occurrence probability of time events based on a given data set of independent variables. A multiple regression analysis algorithm that is often used to analyze the relationship between binomial or multinomial classification results and some influential factors. The Exponential Moving Average is a form of weighted moving average commonly referred to as the exponential weighted moving average (EWMA). Similar to WMA, it assigns a fixed series of exponentially decreasing weights to previous values, i.e. the weight coefficient decreases exponentially over time. EMA provides a more visible indicator that reflects the changing trend of the value more quickly.

For this study on momentum in tennis matches, we mainly use a confusion matrix and Logistic regression to assign weights to the indicators that most closely affect the score and consider their impact on the scoring probability. Secondly, considering the volatility of the scoring probability, we mainly use the Bessel curve fitting method to make the probability smoother and obtain the matching model by analyzing the second derivative, inflection point, and continuity of the curve[1-3].

2. Momentum Quantification and Prediction

2.1 Data Pre-Processing

The data used in this study came from the men's singles data of the 2023 Wimbledon Tennis Championship. The collected data set recorded the entire process of the tennis match, including the status of players and various indicators of their performance. The following are a few important indicators that have been summarized:

·**Match_id**: The identifier for each match, aiding in the analysis of different match progressions.

·**Ace and Winner**: Indicate that a player has delivered a decisive serve or shot, boosting the player's confidence and momentum.

· **Break_pt, Break_pt_won, Break_pt_missed**: Correspond to a player gaining, winning, and losing breakpoints, respectively. Winning a breakpoint significantly uplifts momentum, increasing the likelihood of overall victory. Conversely, losing a breakpoint diminishes morale, reducing the chances of success or momentum.

This study uses a dataset to analyze match progression, player performance, and changes in game dynamics. The system effectively handles missing values to ensure data accuracy and improve model training through visualization techniques like heat maps. As shown in figure 1:

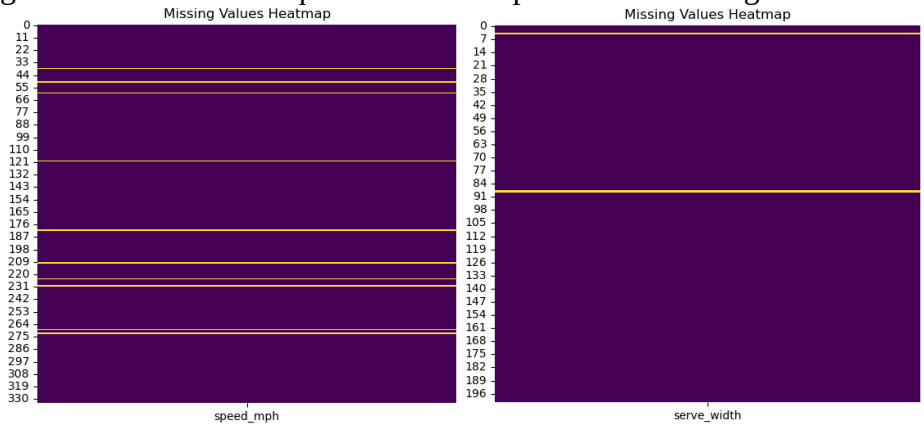

Figure 1: Missing value identification result

2.2 Index Selection

When selecting indicators, we take into account factors such as error rate, breakpoint points, net field goal points, and passing ACES to capture a player's form and performance. To assess a player's

momentum and likelihood of winning, we will utilize the confusion matrix method for index selection. Due to space constraints, only partial results are given as figure 2:

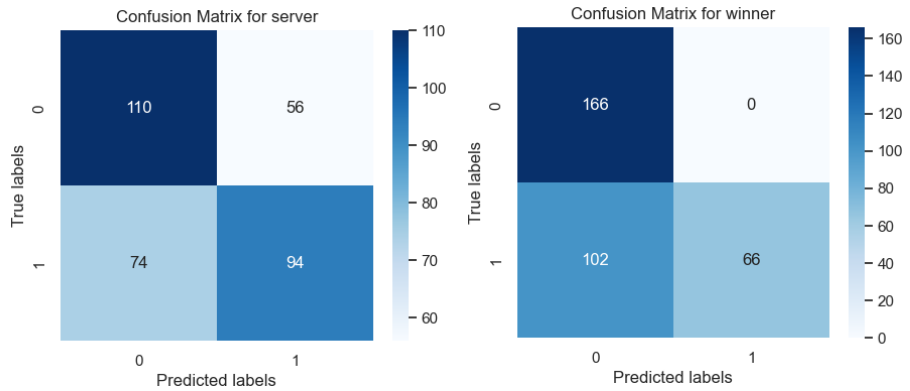

Figure 2: Confusion matrix

Certain performance indicators have a significant impact on a player's scoring. This study will utilize these metrics in a logistic regression model to evaluate a player's momentum, with the dependent variable being whether they win or not. As shown in Table 1:

Table 1: The independent and dependent variable indicators of the logistic regression model

Number	Index	Symbol
1	Player One Win or Not	Y_1
2	Player Two Win or Not	Y_2
3	The difference in the set	X_1
4	The difference between a game	X_2
5	The difference in point	X_3
6	Break_pt_won	X_4
7	Break_pt_missed	X_5
8	ACE	X_6
9	Winner	X_7
10	Net_pt	X_8
11	Net_pt_won	X_9
12	Double_fault	X_{10}
13	Serve	X_{11}

The performance indicators encompass the player's current status, encompassing unreturned serves, successful breakpoints, and effective net shots. In our logistic regression models, we will assign greater significance to the indicator of the player serving first in light of their inherent advantage.

2.3 Logistic Regression Model

Logistic regression is akin to multiple linear regression, with the primary difference being the inclusion of logical variables in logistic regression. Utilizing the relevant indicators selected earlier, this paper proceeds to construct a logistic regression model for both player one and player two:

$$\begin{cases} Y_{1t} = \beta_{10} + \sum_{i=1}^{11} X_{it} + \varepsilon_{1t} \\ Y_{2t} = \beta_{20} + \sum_{i=1}^{11} X_{it} + \varepsilon_{2t} \end{cases} \quad (1)$$

Where Y_{1t} and Y_{2t} indicate whether Player 1 and Player 2 win at the current moment.

Given that Y is a logical variable, we introduce the Bernoulli distribution, using conditional probability to represent the probability of a logical variable occurring, thereby addressing endogeneity issues. As follows:

$$\begin{cases} P(Y=1|X)=F(X,\beta) \\ P(Y=0|X)=1-F(X,\beta) \end{cases} \tag{2}$$

Where F is a function defined on [0,1], and we choose the Sigmoid function, as follows:

$$F(X,\beta)=\frac{e^{X_i^T\beta}}{1+e^{X_i^T\beta}} \tag{3}$$

Finally, work out $P(Y_{1t}=1|X)$ $P(Y_{2t}=1|X)$ and. Representing the probability of player 1 and player 2 winning at time t, respectively. We then weight the probabilities of both winning at time t as follows:

$$\begin{cases} PW_{1t}=\dfrac{P(Y_{1t}=1|X)}{P(Y_{1t}=1|X)+P(Y_{2t}=1|X)} \\ PW_{2t}=\dfrac{P(Y_{2t}=1|X)}{P(Y_{1t}=1|X)+P(Y_{2t}=1|X)} \end{cases} \tag{4}$$

The logistic regression model requires estimating overall regression coefficients using sample observations. After introducing control variables, we will use Ordinary Least Squares (OLS) to calculate the regression coefficients and the F-statistic to test their joint significance.

Utilizing the logistic model, we ultimately derive the probabilities of player 1 and player 2 winning at any given moment. These probabilities serve as indicators of the excellence of the player's performance at the current moment, reflecting their momentum[4-6].

2.4 Exponentially Weighted Moving Average

The momentum of players in a tennis match is always changing, so we need to consider not just their performance at one moment but also how it evolves throughout the match.

This study introduces the EMA model, which uses a weighted average to capture dynamic variation in winning probabilities. The core advantage of the EMA model lies in its ability to assign higher weights to recent data while giving decreasing weights to earlier observations. The model construction is as follows:

$$PW_{1t}=\alpha PW_{1,t-1}+(1-\alpha)PW_{1,t} \tag{5}$$

Where α is the smoothing coefficient, the value range is $0<\alpha<1$. We will employ hyperparameter optimization to calculate the optimal smoothing coefficient.

By using the EMA model, we can calculate a player's winning probabilities based on their current match situation and performance trend, leading to a more accurate measurement of their momentum.

2.5 Model Solving

The study uses a Python model to predict player scoring based on data, using logistic regression and EMA for momentum quantification. Binary classification tasks typically use the cross-entropy loss function for computation, the formula is as follows:

$$\mathcal{L}(\theta)=-\frac{1}{n}((\log(f(X\theta))^T y+(\log(1-f(X\theta))^T(1-y)) \tag{6}$$

Where, n is the total amount of data, $f(\cdot)$ is the function of the logistic regression model, $and\ \theta$ is a learnable parameter. Take the first derivative of $\mathcal{L}(\theta)$, the formula is as follows:

$$\frac{\delta\mathcal{L}(\theta)}{\delta\theta_j} = \frac{1}{n}X^T(f(X\theta) - y) \tag{7}$$

Implementing Stochastic Gradient Descent (SGD) to solve the logistic model. The calculation formula for SGD is shown as follows:

$$\theta_j = \theta_j - \gamma\frac{\delta\mathcal{L}(\theta)}{\delta\theta_j} \tag{8}$$

Where, γ is the learning rate, which is set to 0.01 in the experiment. The number of model epochs is set to 10000.

We divide the data into training and testing sets for each match, then use half of the data for training and the rest for testing. For example, in match 1302 with 201 data points, we split the data in half for training and testing. The results are shown in figure 3:

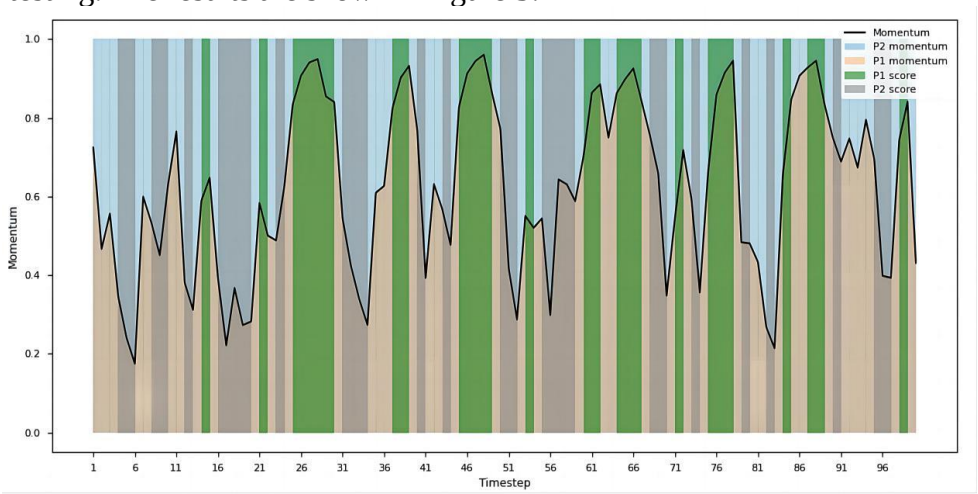

Figure 3: Momentum changes on test set

3. Momentum Verification through Win Probability Changes

3.1 The Definition of Swings

In tennis, swings on the court show important shifts in a player's momentum. We can identify these transitions by calculating the first-order difference of the Bezier curve, the formula is as follows:

$$\nabla B = B(t+1) - B(t) \tag{9}$$

Where $B(t)$ represents the value of the Bezier curve at time t.This study examines momentum changes near transition points to confirm the presence of swings. Stronger changes indicate a higher likelihood of a potential swing in the match. The study quantifies this trend by calculating the second-order difference of the Bezier curve, the formula is as follows:

$$\nabla^2 B = \nabla B(t+1) - \nabla B(t) \tag{10}$$

If the second-order difference value at a specific transition point is notably large, that point is considered a swing in the match.

Based on the comprehensive analysis above, we define a swing as follows:

If the absolute value of the second-order difference of the Bezier curve at that point is greater than the mean absolute value of the historical data's second-order difference, then that point is identified

as a swing in the match.

3.2 The Definition of Runs of Success

Runs of success in tennis matches are crucial for analyzing the game and players' performance, especially when the players are evenly matched and the game is tightly contested. To illustrate this point, this paper provides an example of match 1302, as shown in Figure 4:

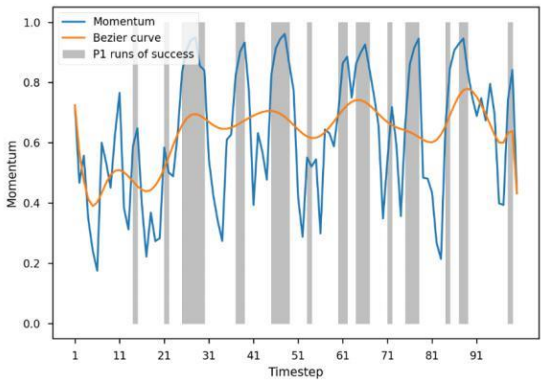

Figure 4: Match 1302 Player 1's momentum, swings, and runs of success

Figure 4 illustrates Player 1's success in Game 1302, showing a consistent winning probability greater than 0.5 and a Bessel curve momentum trend exceeding 0.5. We define the success point of a run by simultaneously satisfying the following conditions:

1) At that specific moment, the player's winning probability is continuously greater than 0.5, signifying a sustained high winning rate.

2) The trend of the player's momentum changes (the value of the Bezier curve) at that moment exceeds 0.5, indicating that the player's performance during this period surpasses that of the opponent.

3) To ensure the continuity and stability of runs of success, this paper stipulates that the variation in scoring probability between any two consecutive points should not exceed 0.1.

3.3 Identification of Swings and Runs of Success

Taking game 1701 as an example, this paper uses the above definition to predict the swing and successful run of player 1 in the game. The test results are as shown in figure 5:

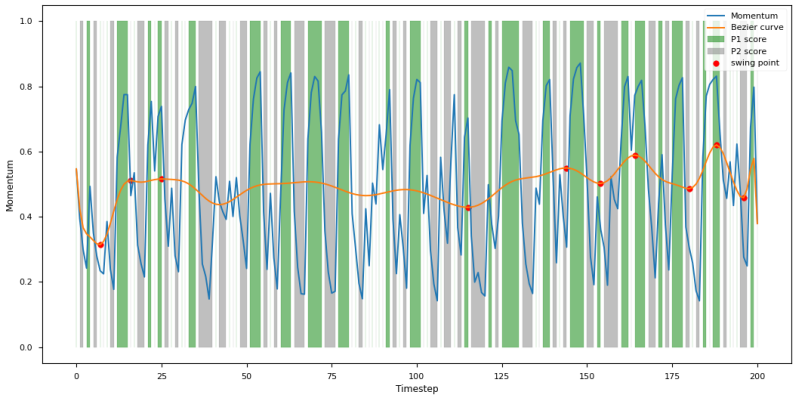

Figure 5: Identification of Swings and Runs of Success

In Figure 5, the green and grey sections show the success of the two players, while the red dots represent changes in the game. Player 1's momentum has a clear downward trend from the start of the game to their first swing, indicating they did not have consecutive victories before that.

Player 1 won consecutive games and their scoring probability increased. After the third swing, Player 1 consistently scored high and secured a victory by a large margin, which makes sense in the game situation[7-8].

3.4 Model Construction

In this research, a randomly selected dataset of player match data, processed for momentum quantification and including swing points and runs of success, was organized into a binary sequence. We created a structure for hypothesis testing based on the principles of statistical hypothesis testing.

H_0 (null hypothesis): Momentum exerts no significant influence within matches.

H_1 (alternative hypothesis): Momentum significantly influences match outcomes.

Rejecting the null hypothesis indicates that momentum has a significant impact on match outcomes. A Run Test showed non-randomness in the total run count, which may lead to rejecting H_0 based on the statistics.

We also checked the actual consecutive wins in the dataset to confirm their validity. Running tests on these sequences in Python yields, the result as shown in table 2 and table 3:

Table 2: Results of the Run Test for swing and runs of success during match 1701

Variable	Sample size	Z-value	P-value
wins_points_pred	167	-7.877	0.000***
wins_points_true	167	-4.754	0.000***
swing_points	167	0.338	0.035*

Note: ***, ** and * represent significance levels of 1%, 5% and 10% respectively

Table 3: Results of the Run Test for swing and runs of success during match 1302

Variable	Sample size	Z-value	P-value
wins_points_pred	100	-5.829	0.000***
wins_points_true	100	-4.609	0.000***
swing_points	100	0.554	0.028*

Note: ***, ** and * represent significance levels of 1%, 5% and 10% respectively

The run test results in Tables 2 and 3 show that swing points and success points in H1701 and 1302 matches are not random, but significantly correlated with momentum. This supports the validity of the momentum quantization model proposed in this study.

4. Conclusions

This study aims to develop a model utilizing men's singles data from the 2023 Wimbledon Tennis Championships for the purpose of predicting momentum and match advantage. Logistic regression, the EMA method for momentum continuity, and the Bezier curve for changing game probabilities were employed in our analysis. Subsequently, we validated the momentum using the run test method. By this, we find that our data is contrary to the null hypothesis, so we find that the winning streak and turning point in the game are closely related to the momentum.

This paper provides a research idea and framework to prove the feasibility of Logistic regression, exponential moving average (EMA), and Bessel curve fitting for momentum prediction of tennis players in the fields of sports science and sports psychology.

References

[1] Silva J M, Hardy C J, Crace R K. Analysis of psychological momentum in intercollegiate tennis[J]. Journal of sport

and exercise psychology, 1988, 10(3): 346-354.
[2] Meier P, Flepp R, Ruedisser M, et al. Separating psychological momentum from strategic momentum: Evidence from men's professional tennis[J]. Journal of Economic Psychology, 2020, 78: 102269.
[3] Richardson P A, Adler W, Hankes D. Game, set, match: Psychological momentum in tennis[J]. The Sport Psychologist, 1988, 2(1): 69-76.
[4] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).
[5] Han X A, Ma Y C, Huang X L. A novel generalization of Bézier curve and surface[J]. Journal of Computational and Applied Mathematics, 2008, 217(1): 180-193.
[6] Song X, Liu X, Liu F, et al. Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis[J]. International journal of medical informatics, 2021, 151: 104484.
[7] Zhang Minjie. Study on the effect of physical exercise on Mental Health [C]// China University Sports Association. The 29th National college track and field research paper report paper album. School of Physical Education, Inner Mongolia Normal University; 2019:2. DOI: 10.26914 / Arthur c. nkihy. 2019.099968.
[8] Quan Zheng, Fan Shuhai, Xu Bin. Universal exponential weighted moving average control chart design [J]. Journal of statistics and decision, 2023, 33 (08) 6:30-34. DOI: 10.13546 / j.carol carroll nki tjyjc. 2023.08.005.