

Quantifying Tennis Players' Performance and Analyzing Momentum Fluctuations Based on Weighted Logistic Regression Model

Zihan Zhang*

Business and Management College, Jilin University, Changchun, China, 130000

* Corresponding Author Email: z18104326909@outlook.com

Abstract. This paper proposes a dual-model approach to analyze tennis players' performance and predict match outcomes, which combines a weighted logistic regression model and a decision tree model. By examining key factors such as scoring rate and serving status, the models calculate the scoring probabilities at different time points, quantifying the performance of players throughout the match. Furthermore, the paper constructs a decision tree model to predict shifts in the match situation based on momentum fluctuations. The research findings underscore the importance of momentum in determining match outcomes and provide valuable insights for training and match strategies. This dual-model approach offers a fresh perspective on the impact of momentum in tennis matches, allowing players to better understand and adapt to the dynamics of momentum fluctuations throughout a match. Additionally, the study suggests targeted training, simulated match conditions, psychological training, real-time data tracking, and physical preparation as strategies to enhance performance and adapt to momentum changes. This comprehensive approach provides players with valuable insights to optimize their performance and strategies in tennis matches.

Keywords: Weighted Logistic Regression Model, Decision Tree Model, Match Outcome Prediction, Tennis.

1. Introduction

The research paper introduces a dual-model approach to analyze tennis players' performance and predict match outcomes. This approach combines a weighted logistic regression model and a decision tree model to calculate scoring probabilities and predict shifts in match situations. Key factors such as scoring rate and serving status are examined, and the models are trained using match data to quantify player performance and identify factors affecting momentum fluctuations. The study findings highlight the importance of momentum in determining match outcomes and provide insights for training and match strategies. The dual-model approach offers a fresh perspective on momentum in tennis matches, enabling players to better understand and adapt to momentum fluctuations throughout a match.

This innovative approach to quantifying tennis player performance and analyzing momentum fluctuations is built upon prior research in the field. Goyal et al. [1] explored the concept of momentum in tennis matches using statistical models, while Han et al. [2] focused on predicting momentum changes in tennis matches through the application of Markov and random forest algorithms. Depken et al. [3] investigated strategic and psychological momentum at the set-level in professional tennis matches. Additional studies by Meng et al. [4], Wen et al. [5], Cao [6], and others provided new perspectives on the dynamics and evaluation of momentum in sports. Furthermore, research by Gantzler et al. [7] on serve direction, Kuznetsova et al. [8] on match-play characteristics, Stocker et al. [9] on scoring opportunities, and Garrido et al. [10] on shot quality in tennis provided further context for the current study.

The current study addresses limitations in handling nonlinear relationships and leveraging big data by proposing the dual-model approach. The weighted logistic regression model quantifies player performance by calculating scoring probabilities at different time points based on key factors such as scoring rate and serving status. The decision tree model predicts shifts in the match situation based on momentum fluctuations, utilizing feature engineering to identify factors affecting real-time

situation changes. By integrating insights from previous research on momentum in tennis matches and extending the analysis through the use of advanced statistical models, the study offers valuable insights for players to optimize their performance and strategies. The proposed dual-model approach is a significant contribution to the field of tennis analytics, providing players with a comprehensive understanding of momentum dynamics and strategies for adaptation.

2. Model

2.1. The structure of weighted logistic regression model

The weighted logistic regression model is a statistical model utilized for predicting the probabilities of a binary outcome variable. It accomplishes this by mapping continuous independent variables to probabilities between 0 and 1, thereby estimating the likelihood of an observation belonging to a specific category. This model finds extensive application across various fields, particularly in biomedical research, social sciences, and financial analysis.

Given the long duration of tennis matches and the frequent occurrence of key points, players' real-time scoring situations mainly depend on their own technical level and the impact of the game situation on their mentality. Many sports experts and data analysts have studied the factors affecting tennis players' scoring and losing points in matches. Based on historical data and expert analysis, this paper have extracted the following factors affecting players' scoring situations at specific times during matches:

1.Players' real-time scoring rate. The real-time scoring rate reflects both the player's own level and, to some extent, the current performance state and the quality of the situation.

2.Players' current score difference with their opponent. Different players have different performances based on personal characteristics when they are leading, trailing, or tied.

3.Whether serving. Based on the higher likelihood of scoring for the server, players in their serving games have more control over the initiative in the match. At the same time, mistakes in serving games may have a greater psychological impact.

4.Ball speed. As a physical factor, it affects players' serving or returning performance.

For the dependent variable, the player's scoring situation at each point in time when scoring occurs, 0 is marked for not scoring, and 1 is marked for scoring. the same time, calculate the values of p_1 and p_2 during the match, and set the weight for scoring on the receiving end to $\frac{p_1}{p_2}$

For each data point x , the calculation formula for the standardized new data point x is:

$$Z = \frac{x - \mu}{\lambda} \quad (1)$$

The data was divided into a training set (70%) and a test set (30%).

To fit the player's real-time winning probability through the scoring situation, this paper introduced a linear model $z = \omega x + b$ to predict y from x . Since the model data is not linearly distributed, this paper introduced the Sigmoid logistic function to confine the values to either 0 or 1. Then let the probability of the sample being a positive case be $y = \frac{1}{1 + e^{-\omega^T x + b}}$, and $1 - y$ be the probability of the sample being a negative case. From this, this paper can deduce:

$$p(Y = 1 | X) = \frac{1}{1 + e^{-\omega^T X + b}} \quad (2)$$

$$p(Y = 0 | X) = 1 - p(Y = 1 | X) = \frac{1}{1 + e^{\omega^T X + b}} \quad (3)$$

For the logistic regression model established, the probability for sample i when the dependent

variable corresponds to 1 is $h(x_i)$, $p(Y=1|X)$, and when the dependent variable corresponds to 0, the probability is $1-h(x_i)$. Construct the maximum likelihood function and take the logarithm to get:

$$L(w) = \frac{1}{n} \sum_{i=1}^n -y_i \ln(h(x_i)) - (1-y_i) \ln(1-h(x_i)) \quad (4)$$

To find its minimum value, this paper can use the gradient descent method, which finds the direction of descent through the first-order derivative of $L(w)$ with respect to w , and iteratively updates the parameters.

Taking match_id “2023-wimbledon-1701”, which is the final match, as an example, this paper fitted individual logistic regression models for the two players using their standardized real-time scoring rates, score differences, serving status, and ball speed as independent variables.

2.2. The structure of decision tree model

This paper capture the real-time changes in the game situation at different time points during one or more matches, i.e., when there is a shift in the high probability winner of the game. Then, use feature engineering to extract factors that may affect the real-time conversion of the game situation on the field. This paper use the extracted features as independent variables and whether there is a shift in the game situation as the dependent variable to build and train a decision tree model, in order to determine which factors affect the change of the game situation and to predict when the flow of play is about to change from favoring one player to the other. The specific process is as follows:

Firstly, based on the logistic regression model, this paper use the time points when the sign of the quantified momentum difference values of the two players changes in the match with match_id “2023-wimbledon-1701” as the dependent variable indicating whether there is a shift in the game situation, with 1 representing a shift in the game situation and 0 representing no shift. Next, apply feature engineering to extract, transform, select, and construct features that may affect the real-time conversion of the game situation on the field, obtaining new feature variables. Finally, train the decision tree with the extracted dependent variable, i.e., whether there is a shift in the game situation, and the feature variables to obtain the importance proportion of each feature variable’s impact on the change of the game situation. According to the real-time momentum quantitative values fitted by the logistic regression model for the two players, this paper performed subtraction operations on them, and regarded the points where the momentum quantitative values changed from greater than 0 to less than zero or from less than zero to greater than zero as the points where the situation fluctuated. For the selection of feature variables, this paper comprehensively considered the differences in the technical, physical, and psychological qualities of the players during the game. Table 1 reflects the final selection and significance of the feature variables.

Table 1. The names and meanings of the feature variables

Variable Name	meanings of the feature variables
X_sd	The absolute value of the score difference of the players at each time point
X_speed	
X_cs	Whether to score continuously (with a scoring threshold of three consecutive scores)
X_ace	Whether the player hit an ace ball
X_df	Whether the serving player missed both serves and lost the point
X_win	Whether there was an untouchable winning shot
X_ue	Whether there was an unforced error

The production process of a decision tree is essentially a continuous process of grouping training samples. The branches of the decision tree grow gradually as the data is continuously divided. The core technology of decision tree growth is the test attribute selection problem. For the classification

solution of the decision tree, this paper use the ID3 algorithm. The decision attribute (i.e., the non-leaf nodes of the decision tree) is selected based on the size of the information gain value of the sample subset attribute values, and the branches of the decision tree are generated according to the different values of this attribute. The method is then recursively called on the subsets, and it ends when all subset data belongs to only one category. Finally, based on the generated decision tree model, new, unknown category data objects are classified.

3. Results

3.1. Quantitative Analysis of Player Scoring Performance

1. Feature Extraction and Data Integration: By analyzing a large amount of match data, this paper identified four factors that significantly impact player scoring performance: real-time scoring rate, score difference, whether serving, and serve speed. The real-time scoring rate reflects a player's scoring efficiency in the current match, while the score difference indicates the winning or losing state at the current stage of the match. Serving or not and serve speed directly affect the number of scoring opportunities. This paper standardized these features to eliminate the influence of different dimensions and assigned corresponding weights to different scoring situations to reflect the actual scoring situations in matches.

2. Logistic Regression Model Building and Training: After identifying the influencing factors, this paper established a logistic regression model with the prediction of player scoring situations as the dependent variable and the four influencing factors as independent variables. Through the training of the model with a large amount of match data, this paper obtained the regression coefficients for each influencing factor, with the real-time scoring rate, score difference, and whether serving having a significant impact on scoring situations. This indicates that the higher the real-time scoring rate, the larger the score difference, and the more likely it is for players to score when serving.

3. Quantifying Player Performance: Using the well-trained logistic regression model, this paper calculated the scoring probability at each scoring point in time, obtaining the quantitative results of the player's performance throughout the entire match. By plotting performance line graphs, this paper can visually compare the performance gaps between different players. Additionally, this paper calculated the scoring probability difference between two players at each scoring point in time to represent their performance gap. By plotting the performance gap line graphs, this paper can visually observe the changes in the performance gap between two players during the match. Figure 1 shows the performance curves of players Carlos Alcaraz and Novak Djokovic as they vary with time in the match with match_id "2023-wimbledon-1701".

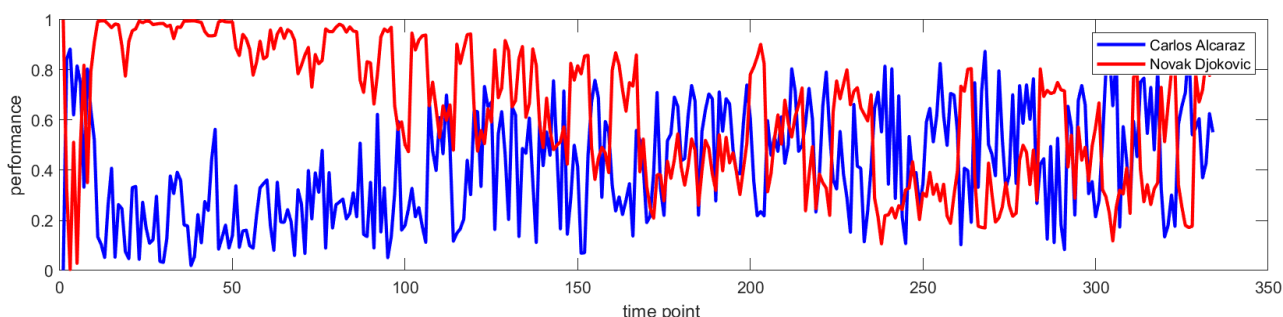


Figure 1. The performance trend over time

Figure 2. represents the difference in performance between the two players. When the y-axis is greater than 0, it indicates that Carlos Alcaraz performed better at that time point. When the y-axis is equal to 0, it means that both players had the same performance at that time point. When the y-axis is less than 0, it indicates that Novak Djokovic performed better.

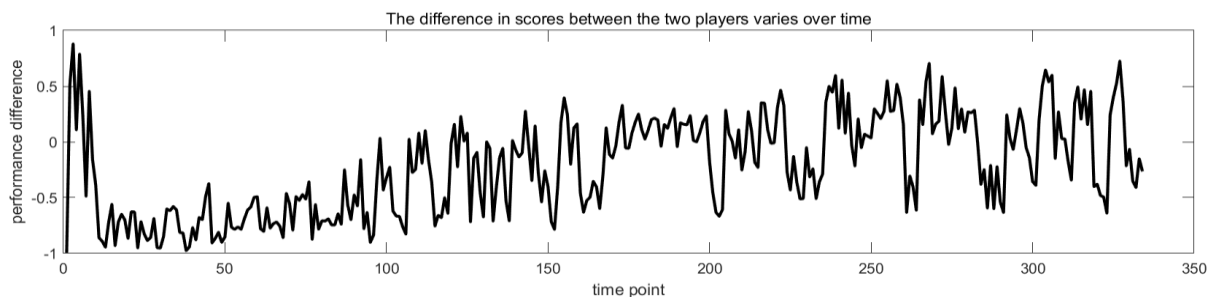


Figure 2. The difference in scores between the two players varies over time

3.2. Applying the Decision Tree Model to Predict Shifts in the Match Situation

After determining the solution algorithm for the decision tree model, this paper imported the prepared training dataset and sought the initial classification, quantifying the goodness of each attribute domain split, calculating the best split, and repeating this process until each leaf node's records belong to the same class, growing into a full tree. The calculation resulted in a maximum number of leaf nodes of 50 and a maximum depth of 10 for the tree. The complete decision tree model is shown in Figure 3. The importance ratio of each feature is shown in Figure 3.

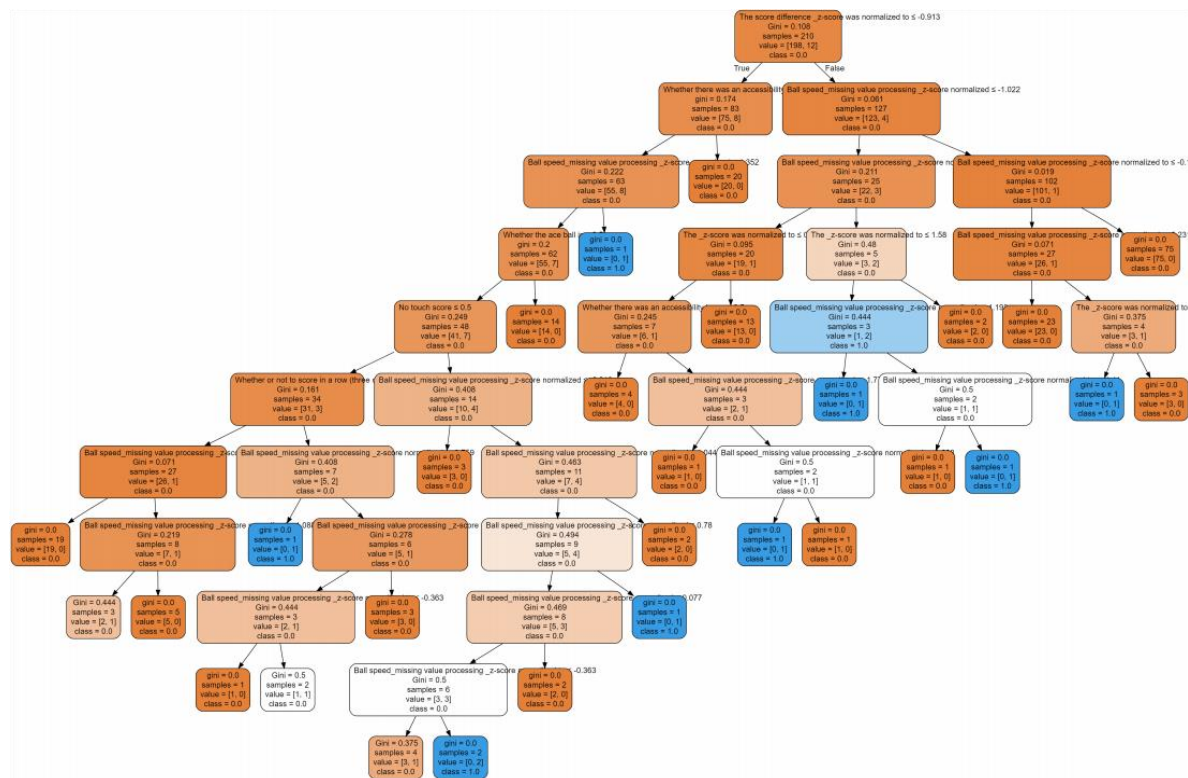


Figure 3. Decision tree structure

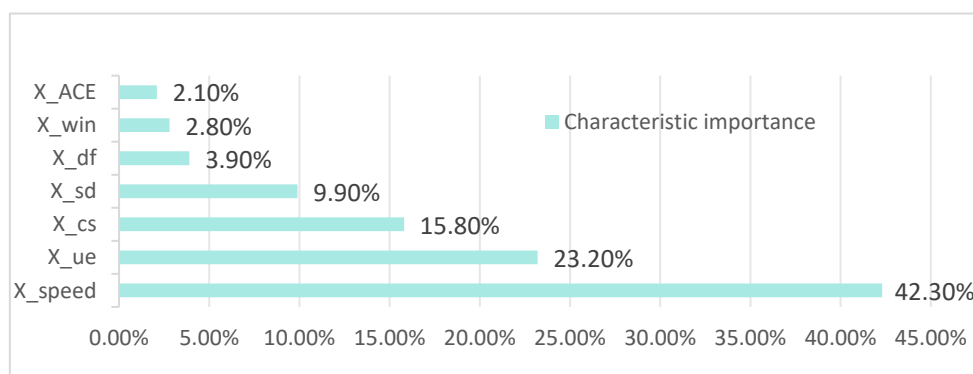


Figure 4. Characteristic importance

From Figure 4, it can be seen that the seven feature variables have an impact on the real-time fluctuation of the game situation, of which the serve speed has the greatest impact on the situation. This paper analyze that there may be the following reasons:

Direct impact of serve speed: The serve speed directly affects the scoring situation in the game. Generally speaking, the higher the serve player's ball speed, the more difficult it is for the opponent to receive the ball, so the player with a higher serve speed is more likely to score. Therefore, players with higher serve speeds tend to have an advantage in the game, thereby affecting the fluctuation of the situation.

Opponent's response strategy: When a player's serve speed is higher, the opponent may adopt different response strategies to deal with the threat of serving. They may choose more conservative tactics, focusing on stable receiving rather than trying to actively counterattack. This adjustment in strategy can also lead to fluctuations in the game situation.

Psychological pressure and self-confidence: A higher serve speed may put psychological pressure on the opponent, making them feel nervous or uneasy during the game. Conversely, players with higher serve speeds may gain an advantage in self-confidence, which may affect their game performance and to some extent affect the fluctuation of the situation.

The classification evaluation metrics in Table 2 show the performance of the decision tree model on both the training and test datasets. The model achieves high accuracy (97% on training, 87.1% on test), high recall (97% on training, 87.1% on test), and high precision (97.1% on training, 88.8% on test). The F1 score, which balances precision and recall, is also good (96.4% on training, 82.7% on test). However, the decrease in performance from training to test indicates some overfitting. Further model tuning, such as reducing tree depth or adding more features, may be needed to improve the model's generalization ability on unseen data.

Table 2. Classification evaluation metrics

	Accuracy	Recall	Precision	F1
Training set	0.97	0.97	0.971	0.964
Test set	0.871	0.871	.888	0.827

4. Conclusion

This paper introduces a dual-model approach, combining a weighted logistic regression model and a decision tree model, to quantify the performance of tennis players and predict fluctuations in match outcomes. The research findings indicate that factors such as real-time scoring rate, score difference, and serving status have a significant impact on the scoring probability of players. By utilizing the well-trained logistic regression model, the paper calculates the scoring probability of players at each point in time during the match, thereby quantitatively evaluating the overall performance of the players. Furthermore, the paper establishes a decision tree model that successfully predicts shifts in the match situation by analyzing momentum changes at critical moments in the game. The results suggest that serve speed and consecutive scoring significantly affect the real-time fluctuations in the match situation.

This study provides valuable insights for tennis players' training and match strategies and offers suggestions for further improving their performance. Specifically, players should focus on targeted training to strengthen weak areas such as consecutive scoring and serving speed. Training under simulated match conditions can help athletes better adapt to the pressures of actual competition. Additionally, psychological training is essential to manage emotions and stress during matches, as momentum fluctuations are influenced by both technical and psychological factors. Real-time data tracking and adjustment allow players to respond quickly to changes in the match. Comprehensive physical training is crucial to reduce the risk of injury and enhance overall performance. Analyzing an opponent's momentum fluctuations can help in formulating more effective countermeasures.

The research methods and results presented in this paper make a significant contribution to the field of tennis analytics and provide a foundation for future in-depth studies. The research content and

findings hold instructive significance for future tennis players, with the potential to assist them in better adapting to momentum changes during matches and devising more rational training and match strategies.

References

- [1] Goyal A, Simonoff J S. Hot Racquet or Not? An Exploration of Momentum in Grand Slam Tennis Matches [J]. arxiv preprint arxiv:2009.05830, 2020.
- [2] Han S, Shayahaxi A, Yu C. Research on Predicting Momentum Changes in Tennis Matches Based on Markov and Random Forest Algorithms [J]. Highlights in Science, Engineering and Technology, 2024, 94: 586-595.
- [3] Depken C A, Gandar J M, Shapiro D A. Set-level strategic and psychological momentum in best-of-three-set professional tennis matches [J]. Journal of Sports Economics, 2022, 23(5): 598-623.
- [4] Meng X, Tian L, Zhang L. Unveiling Momentum Dynamics in Tennis [J]. Advances in Engineering Technology Research, 2024, 10(1): 586-586.
- [5] Wen Z, Liu J, Liu C. Football Momentum Analysis based on Logistic Regression [J]. Frontiers in Computing and Intelligent Systems, 2024, 7(2): 60-64.
- [6] Ling Y, Yanyan W. Analysis of tennis techniques and tactics based on multiple linear regression model [J]. Applied Mathematics and Nonlinear Sciences, 2023, 8(2): 2061-2068.
- [7] Almarashi M A, Daniyal M, Jamal F. A novel comparative study of NNAR approach with linear stochastic time series models in predicting tennis player's performance. [J]. BMC sports science, medicine rehabilitation, 2024, 16(1): 28-28.
- [8] Aisha F, Rizwana M, Ali A, et al. Forecasting Tennis Match Results Using the Bradley-Terry Model [J]. International Journal of Photoenergy, 2022, 2022.
- [9] Tea, Peter, Swartz, et al. The analysis of serve decisions in tennis using Bayesian hierarchical models [J]. Annals of Operations Research, 2022, (prepublish): 1-16.
- [10] Jacob G. Forecasting serve performance in professional tennis matches [J]. Journal of Sports Analytics, 2021, 7(4): 223-233.