GEORGE BROWN PRESS
JOURNALS, BOOKS & PROCEEDINGS

Journal of Education, Humanities, and Social Research

Article

Quantifying Momentum and Performance Dynamics in Tennis: A Point-Level Study Using Logistic Regression, Momentum Scoring, and Random Forests

Wanhe Huang [1,*]

[1] School of Business, Hunan Normal University, Changsha, Hunan, 410000, China
* Correspondence: Wanhe Huang, School of Business, Hunan Normal University, Changsha, Hunan, 410000, China

Abstract: Does a tennis player's momentum influence match outcomes, and if so, how can it be quantitatively measured? This study investigates the role of momentum in tennis matches using statistical modeling and machine learning methods. A binary logistic regression model was first constructed to predict the probability of a player winning a point, with ten performance indicators as independent variables and point outcome as the dependent variable. The predicted winning probability was used to evaluate point-level performance, and the model achieved an accuracy of 64.7%. A momentum score function was then proposed to quantify players' momentum throughout matches. Pearson correlation analysis between momentum scores and match outcomes revealed a significant positive correlation, indicating that momentum plays an essential role in determining match results. To further explore the factors driving shifts in match situations, a random forest model was applied to predict match outcomes and identify key influencing variables. The results show that first-serve success, distance traveled, serve error rate, and point spread are among the most important factors affecting match outcomes. Finally, four additional matches were used to validate the proposed framework, demonstrating that the binary logistic regression model can effectively predict match outcomes and evaluate player performance. Overall, this study provides a quantitative approach to measuring momentum and analyzing performance in tennis matches.

Keywords: binary logistic regression; Pearson correlation coefficient; sports analytics; performance evaluation; momentum modeling; statistical learning

Received: 02 November 2025
Revised: 26 December 2025
Accepted: 10 January 2026
Published: 15 January 2026

1. Introduction

1.1. Problem Background and Restate of the Problem

The 2023 Wimbledon Championships Men's Singles Final was the tennis championship match of the men's singles competition at the 2023 Wimbledon Championships. In a thrilling five-set match, Carlos Alcaraz demonstrated his extraordinary talent to end Novak Djokovic's four-year reign, and the match attracted fans worldwide, making it one of the most-watched matches at Wimbledon in the last decade. The development of computer technology has given us unprecedented data and computing power, which allows us to analyze further the impact of various factors on the probability of a player's victory based on the match data and to analyze which factors are the key factors that make a difference in the match, which not only helps spectators to understand the process of the match but also better helps athletes to prepare for the match and to adjust their tactics during the match. In order to achieve this goal, we need to realize the following objectives: takes the scoring data and creates models through which to quantify how well a player is performing in real time, while highlighting the impact that serve receive has had [1].

1) Take the scoring data and create models through which to quantify how well a player is performing in real-time while highlighting the impact that serve receive has had
2) Use the proposed model to evaluate the role of momentum in the course of a match to determine whether the fluctuations in a player's form and wins and losses during a match are randomized
3) Utilize match data to build a model that uses selected metrics to predict points in the game when situations shift and fluctuate, and to make recommendations to players
4) Test the accuracy of the proposed model in predicting the outcome of the match, discuss the parameters of influence that may need to be added in the future, and discuss whether they can be migrated to datasets from other sports
5) Summarize the results, advise coaches on the role of Momentum, and instruct players on how to deal with the timing of the game.

2. Assumptions and Notations

2.1. General Assumptions

Assumption1: Does not take into account the influence of wind, light, field and other uncontrollable factors on the performance of the race.

1) Assumption2: The players are determined to win.
2) Assumption3: Attachment data is authentic

2.2. Notations

The key mathematical notations used in this paper are listed in Table 1.

Table 1. Variable description.

Symbol	Description
BLR	The Binary Linear Regression
PCCs	Pearson Correlation Coefficient
Y	Independent variable:Whether a player scores a point in a game
MLE	Maximum Likelihood Estimation

3. Task1: Evaluation Model Based on BLR

In this part, we first analyzed and sorted out the existing data and selected appropriate methods to deal with duplicate values and missing values. Then, we selected ten indicators by analyzing the data, taking these ten indicators as independent variables and whether the players scored in the competition as dependent variables to build an evaluation model based on BLR. The probability of an athlete scoring in a game is used as an indicator to evaluate the performance of an athlete at the moment [2].

4. Data Preprocessing

Before conducting data analysis, we must ensure the availability of the data. Regardless of its value, it must be based on reliable data to provide an accurate assessment. Therefore, removing the influence of interference information and extracting practical information from data is a crucial step in model building. The data preprocessing in this paper can be divided into the following steps:

Step1 Check and remove duplicate observations from the original data set

Step2 Determine categorical variables:We use the "factor" function in R language to transform the data into factors, and then determine the categorical variables, in this paper, we determine the categorical variables are mainly server, point-victor, game-victor and so on.

Step3 Handling of abnormal value and missing value:In order to ensure the reliability and rationality of the data, we treated the data coded as categorical variables with missing values.

1) First, convert characters to data in Excel to 1, 2, 3, 4 This type of numbers. For example, for the categorical variable winner_shot_type, we convert the text B and F in the data to be represented by 0,1.
2) Next, the interpolation of missing values was performed using the Random Forest method using the R language. Random forest method is suitable for dealing with data with a large number of variables and there may be complex relationships between the variables, so the interpolation of missing values using this method is more reliable.

4.1. Model Setting of Binary Logistic Regression Model

Binary logistic regression is a generalized linear regression analysis model. The dependent variable is categorical, and a linear combination of independent variables is usually used to predict the logarithmic probability of an event occurring, that is, the natural logarithm of the ratio of the probability of an event occurring to the probability of it not occurring. The model uses Maximum Likelihood Estimation (MLE) to estimate the coefficients and then converts these logarithmic odds into probabilities. Therefore, through the binary logistic regression model, we can predict the probability of an event and then evaluate the player's performance in the game [3].

4.1.1. Indicator Design and Assignment of Values to Indicators

Based on the analysis of the data, we mainly chose ten indicators such as plate point difference, set point difference, single serve error rate, etc. as the independent variables affecting whether a player scores or not.

By analyzing the dataset, we define the following variables to explore the effect of these variables on the dependent variable, as shown in Table 2

Table 2. Specific variable assignment.

Variable	Assignment
Set Margin	Difference in the number of sets won by a player
Game Margin	Difference in the number of games won by a player
First Serve Error Rate	Ratio of single serve errors to total first serves
Total Serve Error Rate	Ratio of the total number of total serve errors to the total number of serve attempts
Number of Aces	The number of player winning serves is 0, 1, 2...
Number of Return Winners	The number of players returning game-winning goals is 0, 1, 2...
Number of Unforced Errors	Number of unforced errors 0, 1, 2, 3...
Net Points Won Percentage	Ratio of the number of points scored by a player at the net to the number of attempts at the net
Break Point Conversion Rate	The ratio of the number of successful breaks of serve by a player to the total number of times the opponent serves for the set
Running Distance	The total length of the run of the players
Serve Indicator	Serves are assigned a value of 1, no serves are assigned a value of 0

In particular, the number of unforced errors refers to the number of points a player loses due to his or her errors without significant pressure from the opponent; the percentage of points scored at the net reflects a player's efficiency at the net, with a high percentage implying that the player is more capable of scoring at the net. The percentage

of broken serves reflects a player's ability to disrupt the opponent's serve when receiving serve. We hypothesize that these variables can influence whether a player scores [4].

4.1.2. Model Construction

Since this paper chooses whether the player scores in the game as the dependent variable, whether the player scores or not is an either/or binary choice, there are only two possible outcomes, i.e., "scoring" and "not scoring," so this paper uses a binary logistic regression model to analyze the situation, and if the player scores, then the definition of "Y=1"; on the contrary, the definition of "Y=0".

The Logistic function is of the form:

$$f(x) = \frac{e^x}{1+e^{-x}} = \frac{1}{1+e^{-x}} \quad (1)$$

The range of values of its independent variable x is(-∞,+∞)

The dependent variable Y takes only two discrete values of 0 and 1 and is unsuitable for the dependent variable in a regression model. We define:

$$\pi_i = f(x) = \frac{1}{1+exp(-(\beta_0+\beta_1 x_i))} \quad (2)$$

$$ln(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 x_i \quad (3)$$

Where is the probability that the dependent variable Y takes 1? Its value varies continuously in the interval [0,1], so it can be used instead of Y as the dependent variable.

We define Y to be a variable of type 0 to 1, and n sets of observations are (), which are random variables taking values 0 or 1. The probability function is

$$P(y_i = 0) = 1 - \pi_i \quad (4)$$

One can define the random probability of

$$P(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}, y_i = 0,1; i = 1,\cdots,n \quad (5)$$

Next, we carry out the constructed likelihood function L and take the likelihood function, which ultimately leads to the binary logistic regression formula for Y:

$$ln(L) = \sum_{i=1}^{n}[y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) - ln(1 + exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}))] \quad (6)$$

Maximum likelihood estimation yields an estimate $\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_p$ of $\beta_0, \beta_1, \ldots \beta_p$

With the constructed model, we determine in real-time how good or bad a player's performance is based on the selected independent variable metrics. We use the metrics to predict situational transitions and fluctuations during the game and advise the player based on these fluctuating differences.

4.2. Result & Analysis

In 3.3.1, we first analyze the results of the binary logistic analysis obtained based on the selected variables to prove the accuracy of the above model and observe and find that the three variables, Set Margin, Number of Unforced Errors, and Serve Indicator, have a significant influence on whether the player scores or not during the match, where the Serve Indicator has the most significant influence on the player and is positive and find the magnitude of influence of this variable on the score of the player.

In 3.3.2, we will analyze the entire game based explicitly on the results obtained, visualize the game's flow, and determine which player performed better at a particular point based on the results obtained [5].

4.2.1. Result

We performed a binary regression analysis of the data based on the variables selected above, and the results are as follows in Table 3

Table 3. Omnibus test of model coefficients.

		Chi-square	df	P-value
Step1	Step	79.851	11	0.000
	Block	79.851	11	0.000

	Model	79.851	11	0.000

a. The estimation was terminated at the 4th iteration because the parameter estimates varied by less than .001.

Table 4 and Table 5 show that the chi-square value is more significant, and the P-value value is less than 0.05. At the same time, Cox Snell R Square, Negolko R. Square are more significant, which indicates that the model results are more accurate.

Table 4. Model Summary.

Step(1)	-2 Log-likelihood	Cox Snell R Square	Negolko R. Square
1	751.925[a]	0.125	0.166

Table 5. Category Table.

Actual measurement			Prediction Y		Correct Percentage
			.00	1.00	
Step1	Y	.00	189	111	63.0
		1.00	101	199	66.3
Overall percentage					64.7

a. The cut-off value is .500

According to Table 6, the model's prediction accuracy of the player's score loss reaches 63%, that of the player's score reaches 66.3%, and the overall prediction accuracy of the player's score reaches 64.7%. The model's prediction accuracy of the player's score is relatively accurate.

Table 6. Variables in the Equation.

		B	SE	Wald	df	P-value	Exp(B)
Step[a 1]	Set Margin	-0.818	0.282	8.389	1	0.004	0.442
	Game Margin	-0.244	0.092	6.992	1	0.008	0.784
	First Serve Error Rate	-0.274	1.558	0.031	1	0.860	0.760
	Total Serve Error Rate	-11.111	9.268	1.437	1	0.231	0.000
	Number of Aces	0.299	0.125	5.687	1	0.017	1.348
	Number of Return Winners	0.093	0.098	0.903	1	0.342	1.097
	Number of Unforced Errors	-0.193	0.055	12.364	1	0.000	0.824
	Net Points Won Percentage	0.481	0.618	0.606	1	0.436	1.618
	Break Point Conversion Rate	-0.225	0.429	0.275	1	0.600	0.798

	Running Distance	0.000	0.001	0.158	1	0.691	1.000
	Serve Indicator	1.275	0.182	48.937	1	0.000	3.577
	Constant	-1.209	0.515	5.515	1	0.019	0.298

a. Variables entered in step 1: Set Margin, Game Margin, First Serve Error Rate, Total Serve Error Rate, Number of Aces, Number of Return Winners, Number of Unforced Errors, Net Points Won Percentage, Break Point Conversion Rate, Running Distance, Serve Indicator, Constant

According to Table 6, from the significance results, it can be seen that Set Margin, Game Margin, Number of Aces, Number of Unforced Errors, and Serve Indicator have a significant effect on whether the player scores or not. Further, by looking at the values of these five variables, we can see that the values of Number of Aces, Serve Indicator, are positive, while the values of Set Margin, Game Margin, Number of Unforced Errors are all negative, and we can basically determine that the Number of Aces. Serve Indicator on player's score is positive, i.e., the more the number of player's winning serves, the higher the probability that the player will score; while Set Margin, Game Margin, Number of Unforced Errors on player's score is negative, indicating that the difference of set points, the greater the difference of set points, the greater the difference of unforced errors, and the greater the difference of set points. The larger the Set Margin, the more the number of Unforced Errors, the lower the probability of the player scoring a point

Focusing on Number of Aces, the Exp(B) value of Serve Indicator, we can get the strength of the influence of these two factors on the player's scoring, which is shown as follows: for each rank increase in the player's Ace score, the probability of scoring increases to 1.348 times of the original one; if the player serves, the probability of that player's scoring will increase to 3.577 times of the original one.

4.2.2. Analysis

From the above results, it can be seen that the difference in the sets won by the player, the difference in the number of sets won by the player, the number of winning serves of the player, the number of unforced errors of the player, and whether or not the player serves have a significant effect on the performance of the player in the match, in which the number of Aces, Serve Indicator and other factors have a positive effect on the score of the player, and the player's performance can be well predicted by observing and analyzing the number of winning serves and whether the player serves or not. By observing and analyzing the number of winning serves of the players and whether the players serve or not can well predict the performance of the players in the match, and the probability of winning the score of the players in the process of the match can be predicted according to these factors, thus evaluating the performance of the athletes in the process of the match is good or bad.

4.3. Visual Analysis & Sensitivity Analysis

4.3.1. Competition Flowchart

Below is a flowchart of this tennis match, which describes the flow and rules of the tennis match so that we can better understand the performance of the players at various points of the match (As shown in Figure 1).

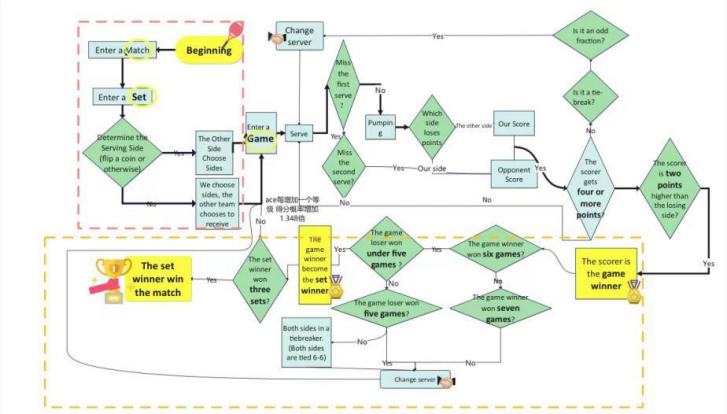

Figure 1. Flowchart of Tennis Match Scoring and Game Progression.

4.3.2. Sensitivity Testing

The Roc curve can be used to assess the ability of a classification model to perform under different thresholds by comparing the True Positive Rate (TPR) and the False Positive Rate (FPR), which visualizes the change of the model sensitivity with the False Positive Rate. Whereas sensitivity refers to the ability of the model to correctly recognize instances, the curve can be used to roughly test the sensitivity of the model.

From the Figure 2 we can observe that Set Margin, Game Margin, Number of Aces, Number of Unforced Errors, Serve Indicator and other indicators are farther away from the diagonal line, among which Serve Indicator is the farthest away from the diagonal line, which indicates that whether a player scores or not is sensitive to these factors, which proves that the prediction effect of our model is good.

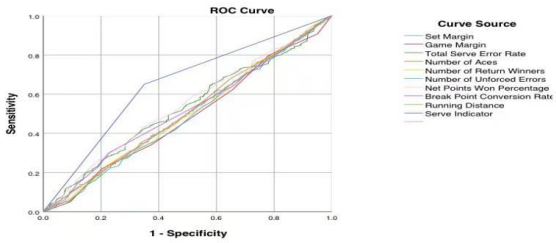

Figure 2. ROC Curves for Key Indicators (Sensitivity Analysis).

5. Task2: Momentum Score Model

Since the coach's point of view is that the players' momentum changes during the game are random, and he questions the role that momentum plays in the game, we can utilize statistical knowledge to verify the veracity of the coach's point of view. In this part, we first quantify the momentum to get the momentum score and then use the Person correlation test to conduct the correlation analysis between the momentum score and the dependent variable to observe whether the momentum is correlated with the player's game performance and what the correlation is like, so as to test the authenticity of the coach's viewpoint.

5.1. Defining Momentum

In the previous model, we analyzed the impact of various factors on the player's game score; these factors work together throughout the whole game, making the game volatile. However, these fluctuations are not random; they are closely related to the game

at various points of the serve, points, and errors; therefore, we will formally define and quantify momentum here.

In order to validate the effect of momentum on a match, we first assume that momentum exists in a match and that the initial momentum score is 0. Momentum is related to scoring and is a continuous process whereby for each point won by a player, the total momentum score increases, and the total momentum score is the original momentum score plus the relevant specific value. On the contrary, if the player loses a point, the total momentum score decreases, and the total momentum score is the original momentum score minus the relevant specific value.

We use W_t to denote the momentum at moment t, i.e., the momentum at point t

5.2. Quantifying Momentum

Whether a player scores or not is related to a number of factors, so we can quantify momentum by considering the impact of these factors on momentum and constructing a scoring system for momentum.

The main influencing factors of momentum are service points, service errors, unforced errors, net points, breaking points, and so on. Serve, a high level of scoring ball will cause a positive momentum increase for the player, while losing, mistakes, etc., will cause the player to lose momentum. Therefore, we can construct serve indicators, miss indicators, and control indicators based on these factors to quantify momentum. Now, we will look at what each of these terms represent.

5.2.1. Serving Target (St)

As summarized in Table 7, the serving target S_t is assigned according to the player's serving status and point outcome. Different values are used to reflect the asymmetric win–loss probabilities associated with serving and non-serving situations.

Table 7. The values of S_t and the conditions under which they are taken.

Prerequisites	St
The player is on the serve side and loses point	0.327
The player is not on the serving side and wins the point	0.673
The player is on the serve side and loses point	-0.673
The player is not on the serving side and wins the point	-0.327

Note:0.327, 0.673, -0.673, and -0.327 respectively, based on the win/loss probability of whether or not it is a serve.

5.2.2. Failure Indicators (ET)

The value of Et is the sum of the value of the service error and the value of the unforced error.

The value of serve error (in tennis):If double_fault=1, the value of a serve error is -0.25; if double_fault=0, it is unchanged

Unforced error value: If unf_err = 1, the unforced error value is -0.3; if unf_err = 0, it is unchanged

5.2.3. Control Force Indicators (CT)

Since net_pt_won, break_won and break_lost_won are more reflective of a player's skill and more determinative of the course of the match, demonstrating the player's control of the match, the value of the control indicator is the sum of the values of net_pt_won, break_lost_won and break_lost_won.

Points at net: if net_pt_won=1, the value of service errors is 0.25; if net_pt_won=0, it is unchanged

Points scored on broken serves: if break_pt_won=1, the value of serve errors is 0.65; if break_pt_won=0, no change

Breaking serve points lost: if break_pt_won=1, the value of serve error is -0.1

5.3. Momentum Score Modeling Formula

Based on the above indicators, we built a momentum score model and summarized it in:

$$W_t = W_{t-1} + S_t + E_t + C_t \quad (7)$$

By using the formula we can observe that the momentum of the player at the tth point depends on the momentum at the (t-1) th point with the values of the serve indicator, error indicator and control indicator at the tth point. This formula allows us to calculate the player's momentum at each point and, based on this, to understand how the player's momentum changes over the course of the match. Correlation Analysis

5.3.1. Pearson Correlation Coefficient (PCCs)

In statistics, PCCs are used to measure the correlation between two variables with values between -1 and 1.

We express the correlation between two variables through the Pearson correlation coefficient. The closer the correlation is to 1 or -1, the greater the absolute value of the correlation coefficient and the stronger the correlation; the closer the correlation coefficient is to 0, the weaker the correlation.

We begin with a null hypothesis and alternative hypotheses about the overall parameters:

Null hypothesis (H0): the quantitative data of momentum is not related to whether the player scores or not

Alternative hypothesis (H1) Quantitative data of momentum is correlated with whether a player scores or not.

Selection of Significance Level: this is the probability of the error we are willing to make before rejecting H0

Next we calculate the correlation coefficient according to the formula to get the results of the correlation analysis

5.3.2. Result of Correlation Analysis

We used the data on the momentum score of the player at each point to correlate with whether the player scored or not using PCCs, defined as the momentum score of player 1, the momentum score of player 2, Y_1 as whether player 1 scored or not, and Y_2 as whether player 2 scored or not.

We get the following results:

The resulting Pearson correlation coefficients are reported in Table 8, illustrating the relationships between each player's momentum score and point-level scoring outcomes.

Table 8. Correlation.

		W_t_1	W_t_2	Y_1	Y_2
W_t_1	Pearson Correlation	--			
	Number of cases	300			
W_t_2	Pearson Correlation	-0.981**	--		
	Significance (two-tailed)	0.000			
	Number of cases	300	300		
Y_1	Pearson Correlation	0.121*	-0.114*	--	
	Significance (two-tailed)	0.036	0.049		
	Number of cases	300	300	300	
Y_2	Pearson Correlation	-0.121*	0.114*	-1.000**	--

Significance (two-tailed)	0.036	0.049	0.000	
Number of cases	300	300	300	300

**.At the 0.01 level (two-tailed), the correlation was significant. *. At the 0.05 level (two-tailed), the correlation was significant.

Observing the significance values of the variables in the table, we find that there is a more significant correlation between all four variables, and we focus on analyzing the correlation between with Y_1, and with Y_2. We can see that the correlation coefficient between the momentum score of player 1 and whether player 1 scores is 0.121, and the correlation coefficient between the momentum score of player 2 and whether player 2 scores is 0.114, which indicates that there is a significant positive correlation between the momentum score of player 1 and player 2 and the player's game performance.

Based on the above results, we can refute the professor's argument. The fluctuations in the match and the result of the match are not random, there is a correlation between the momentum of the players in the match and the players' match scores and the relationship between the two is positive.

6. Task3: Random Forest Prediction Model

According to the model established in the previous section, we can know that several variables, such as set point difference, set point difference, and total service error rate, have an impact on whether a player scores or not. In this section, we will make a Random Forest prediction based on the data of whether a player scores or not based on these factors, in order to test the accuracy of the model and analyze the degree of influence of each variable on the model, so as to provide a reference suggestion for the players.

6.1. Random Forest Prediction Model Construction

Random Forest (Random Forest) is a classic Bagging model with a weak learner as a decision tree model. The Random Forest model will randomly sample the original data set to constitute n different sample data sets, and then build n different decision number models based on these data sets, and finally obtain the final results based on the average of these decision tree models.

In this paper, 500 data as a training set, and the first two data as a test set, we can see from the training set there are a total of 10 input parameters such as plate point difference, set point difference, total serving error rate, and one output parameter, i.e., whether or not the player scores points. We built a random forest model based on these training sets.

6.1.1. OOB Error & Accuracy

Random Forest models can be evaluated using the out-of-bag (OOB) estimate, which assesses generalization performance by predicting the samples not selected in each bootstrap draw. This provides an internal validation mechanism and is particularly useful when an explicit hold-out validation set is limited or unavailable. As shown in Table 9, the OOB estimate of the error rate is 35.63%, indicating an overall predictive accuracy of 64.37% (1 − 0.3563). This accuracy is broadly consistent with the performance obtained from the binary logistic regression model in the previous section, suggesting that the predictive signal captured by the selected indicators is stable across modeling approaches.

Table 9. OOB estimate of error rate.

OOB estimate of error rate:35.63%			
	0	1	class.error
0	321	180	0.3592814
1	177	324	0.3532934

6.1.2. Confusion Matrix & Metrics

The confusion matrix summarizes how well the model's predicted classes match the reference (actual) classes. In Fig. X, with Y = 1 (player scores the point) defined as the positive class, the entries are: TP (actual 1, predicted 1), FP (actual 0, predicted 1), TN (actual 0, predicted 0), and FN (actual 1, predicted 0). Based on the figure, the values are TP = 437, FP = 64, TN = 484, and FN = 17. Using these values, the model achieves an Accuracy of (TP+TN)/(TP+FP+TN+FN) = 91.92%, a Recall (Sensitivity) of TP/(TP+FN)=96.26%,and a Specificity of TN/(TN+FP)=88.32%.These results indicate that the model identifies scoring points with high sensitivity while maintaining a relatively strong ability to correctly recognize non-scoring points (As shown in Figure 3).

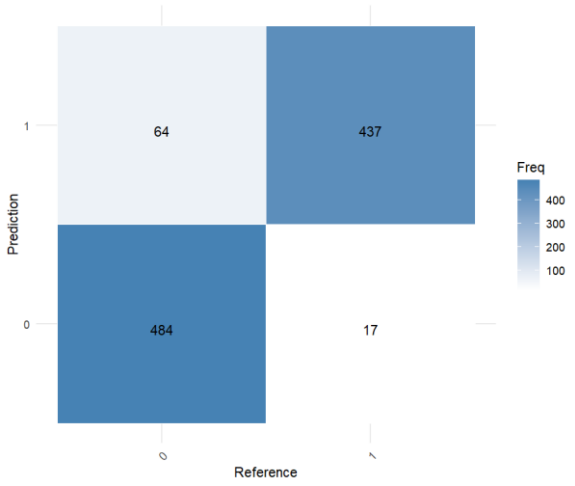

Figure 3. Confusion Matrix Comparing Actual vs. Predicted Point Outcomes.

6.1.3. Degree of Influence of Each Variable on the Outcome

To interpret the Random Forest results, we examined the model's feature-importance output, which ranks predictors by their contribution to reducing classification error (i.e., higher importance indicates a larger contribution to the split quality and overall predictive performance). The results show that Serve Indicator and Number of Unforced Errors are among the most influential variables, with importance contributions of 39.7% and 8.9%, respectively, and variables such as Set Margin, Game Margin, and Number of Aces also exhibit relatively high influence. Notably, these key predictors are consistent with the significant variables identified in the binary logistic regression model, providing cross-method support for their importance.

6.2. Analysis

The establishment of the above Random Forest prediction model and the analysis of the results show that the results of the game are related to Set Margin, Game Margin, Number of Aces, Number of Unforced Errors, Serve Indicator and other factors, and the fluctuation of these variables determines the fluctuation of the game and the transition of the situation. Based on this, we will make the following recommendations to the players based on the results of the analysis and combined with the previous analysis of momentum:

1) Since the opponent's serve is an important factor that leads to the loss of momentum for their side, it is more important to focus on finding ways to respond when the opponent is serving
2) Players should concentrate on scenarios where they may lose points
3) Players should make sufficient preparation before the match to improve their skills and control the situation.

4) When the score difference is large, you need to adjust your mindset, because this is the time when the opponent's momentum is high, and your side is more likely to lose points.

7. Task4

We considered utilizing data from 4 more games and validated it with logistic regression, and finally found that our model had an accuracy of 67.7%, and found that the number of aces, the rate of errors, the number of game-winning kicks returned, and whether or not the first kick was made had a significant effect on the outcome of the game.

8. Conclusion

This study develops a point-level performance analytics framework for tennis by combining binary logistic regression, a momentum quantification scheme, and a Random Forest model. Using ten engineered indicators, the binary logistic regression model provides a probabilistic evaluation of whether a player scores at each point and achieves an overall classification accuracy of 64.7% on the main dataset.

Across modeling approaches, the results consistently emphasize the importance of serve-related and error-related factors. In the logistic regression analysis, Serve Indicator and Number of Aces are positively associated with scoring probability, whereas variables such as Set Margin/Game Margin and Number of Unforced Errors exhibit negative effects; the estimated effect sizes further suggest a substantial advantage when the player is serving (Exp(B) = 3.577) and when ace performance improves (Exp(B) = 1.348).

To evaluate the role of momentum, we constructed a momentum score and tested its association with point outcomes. The Pearson correlation analysis indicates a statistically significant positive relationship between momentum and scoring for both players ($r = 0.121$ and 0.114, $p < 0.05$), supporting the conclusion that match fluctuations are not purely random and that momentum is meaningfully linked to performance dynamics.

Finally, the Random Forest model offers a complementary, nonparametric perspective on predictive performance and variable importance. The out-of-bag (OOB) estimate yields an error rate of 35.63% (accuracy 64.37%), and the feature-importance output ranks Serve Indicator and Number of Unforced Errors among the most influential predictors (importance contributions of 39.7% and 8.9%, respectively), consistent with the regression-based findings. A further robustness check on four additional matches attains an accuracy of 67.7%, providing additional evidence that the identified drivers of scoring are stable across matches.

References

1. E. C. Zabor, C. A. Reddy, R. D. Tendulkar, and S. Patil, "Logistic regression in clinical studies," *International Journal of Radiation Oncology* Biology* Physics*, vol. 112, no. 2, pp. 271-277, 2022. doi: 10.1016/j.ijrobp.2021.08.007
2. A. Zaidi, and A. S. M. Al Luhayb, "Two statistical approaches to justify the use of the logistic function in binary logistic regression," *Mathematical Problems in Engineering*, vol. 2023, no. 1, p. 5525675, 2023. doi: 10.1155/2023/5525675
3. N. A. Saran, and F. Nar, "Fast binary logistic regression," *PeerJ Computer Science*, vol. 11, p. e2579, 2025. doi: 10.7717/peerj-cs.2579
4. J. R. Wilson, K. A. Lorenz, and L. P. Selby, "Introduction to binary logistic regression," *Modeling binary correlated responses: Using SAS, SPSS, R and STATA*, pp. 3-18, 2024. doi: 10.1007/978-3-031-62427-8_1
5. E. Beacom, "Considerations for running and interpreting a binary logistic regression analysis-a research note," *DBS Business Review*, vol. 5, 2023.