Uncover the Regular Pattern of Momentum Behind Tennis Match

Yaping Yu [1,*], Yuhui Zhou [1], Zhipeng Hou [2]

[1] School of Economics, Nanjing University of Posts and Telecommunications, Nanjing 210046, China

[2] School of Automation, School of Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing 210046, China

* Corresponding Author Email: yyp12081105@163.com

Abstract. In sports, a team or player may feel they have strength or force during a match as a result of which is often attributed to "momentum". We defined a series of indicators and built several models to explore the role of momentum in a tennis match and the factors that influence it. We use a list of data sets of Wimbledon 2023 men's matches after the second round. After preprocessing the given data, we defined six indicators, such as cumulative scoring advantage, success rate of breaking serve and ability to control the ball, which make up the player's performance score in each round. The player's performance score was calculated by the normalized values of the indicators and the weights of the six indicators determined through the Principal Component Analysis (PCA). Next, a tumbling window was used on the player's performance scores, and the size of the tumbling window was set to 3 rounds, which was used to calculate the player's momentum in each period. After that, we chose the match "2023-wimbledon-1701" for the visualisation of the match flow. Then we defined the performance score difference between the two players in each round and used it as the independent variable, and the point victor in each round as the dependent variable. Then we built a Logistic Regression Model to analyze the causal relationship between the two. We found that the swings of the game and the success of the players are not random, it will be affected by the performance scores of the players in the game. Each unit increase in the performance score difference increases the probability of Player 1 winning by approximately 5.087 times. Next, we defined the performance score turning point and used it as the dependent variable. After data processing with normalization and resampling, we built a Softmax Regression Model and combined it with the SHAP method to explain the model. We found that there are indicators that can help determine when the flow of play is about to change from favoring one player to the other such as cumulative rate of unforced error, quality of the serve and cumulative rate of scoring at the net. Of these, the ability to control the ball is the most relevant indicator. After that, we made suggestions for players to play against different players. After that, we tested the model using data from the match "2023-wimbledon-1311", and the predictions were favourable. By constructing the confusion matrix and calculating the four evaluation indicators of accuracy, precision, recall, and f1-score, we found that the model results performed generally well. Then, we proposed six factors that may need to be included in the model, such as direction of serve, depth of return and length of break before this round. Afterwards, we found the dataset of Sebastian Ofner vs. Stefanos Tsitsipas in the men's singles eighth-final of the 2023 French Open Roland Garros to fit the model and found that the model is more generalizable. Finally, we evaluated and refined the model and reported the findings in a memo to tennis coaches.

Keywords: Performance Score, Momentum, Softmax Regression Model, SHAP, PCA.

1. Introduction

1.1. Problem Background

Tennis is a high-intensity sport that requires players to react quickly and move their bodies at high speeds to hit the ball back to the opponent's court in a short time. In this process, the concept of "momentum" is about the athlete's condition and confidence on the field, and its change will directly affect the athlete's performance and mentality. Therefore, "momentum" plays a vital role in giving

athletes an advantage and victory in a round. However, because it is difficult to measure, it is not accurate to grasp it during the match [1].

Wimbledon 2023 men's matches are best of five sets, in which the first player to win six games in each set by two games wins that set, and the first player to win four points in each game by two points wins that game, with points being scored in each round of play.

To summarize, quantifying the momentum from various aspects, studying its changing regularity and its influence on the game results, and predicting it will help to provide athletes with scientific training guidance and tactical advice, which in turn can improve athletes' performance and game results. The study of momentum has important theoretical significance and practical value.

1.2. Restatement of the Problem

In this question, we use a list of data sets of Wimbledon 2023 men's matches after the second round and advise coaches on how to prepare their athletes in terms of momentum. Considering the background information, we need to solve the following tasks:

- **Task 1:** Establish a model to quantify the "momentum" of each player during the match, measuring and describing it qualitatively or quantitatively, and taking into account the serve side. It is also required to provide a visualization of the match flow based on the model.
- **Task 2:** Test whether swings in play and runs of success by one player are random, whether there is a causal relationship between the two and whether momentum plays an important role in the game.
- **Task 3:** Determine when the flow of play is about to change from favouring one player to the other through some indicators.
  - Subtask **1:** Use the data provided to build a model to predict these swings in the match and explore the importance of each influencing factor.
  - Subtask **2:** Consider the differential in past round "momentum" swings and make appropriate recommendations for new matches between one player against a different player.
- **Task 4:** Test the model in other matches, analyze the predictive effects, identify additional factors to include in future models, and explore the generalizability of the model.
- **Task 5:** Write a report summarizing the results of the study with an attached memo advising coaches on the role of momentum and how players can prepare for it.

1.3. Our Work

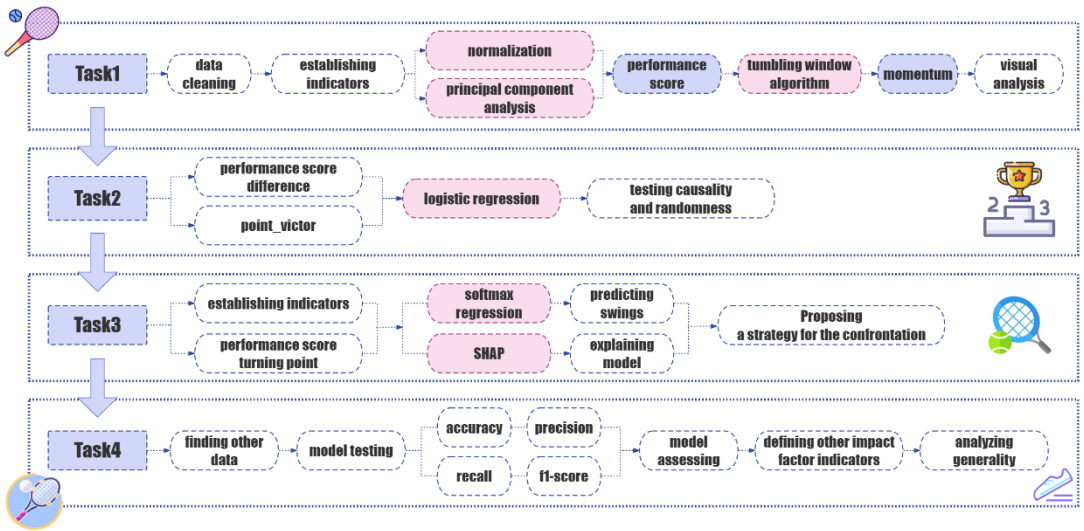

Figure 1. Our Work

2. Assumptions and Justifications

We make several assumptions in our model. Later we may relax some of these assumptions to optimize our model making it more applicable in the complex reality environment.

Assumption 1: Each row of data in the dataset records the current round and the data recorded by players in each game are normal, accurate and reliable.

Justification: To ensure that the indicators we measure are accurate and the model we build based on the dataset can reliably predict patterns of change.

Assumption 2: A player's momentum for each point in a round can be represented by a combination of score situation, mental condition, serve situation, ball control situation etc.

Justification: Therefore, we can use these relevant indicators to build models to quantify, analyze, and predict momentum.

Assumption 3: A player's performance score reflects the swings in the match at the time.

Justification: This allows us to use the performance score turning points of the two players as output variables when building a Softmax regression model to predict match swings.

3. Notations

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

Table 1. Notations used in this paper

Symbol	Definition
G_i	The player's scoring advantage in Round i
γ	The success rate of the first serve
ν	The success rate of the second serve
ξ_i	The rate of untouchable winning serve at Round i
D	Performance score difference
B	The probability of player 1 winning
o_j	The linear output of the jth category
w_{ij}	The weight of the ith feature in the jth category
b_j	The total bias of all features in the jth category
N	The number of features
y_j	The predicted probability for each category
C	The number of categories
SH	The SHAP value of the feature

4. Evaluating the Performance of Tennis Players Based on PCA

4.1. Data Preprocessing

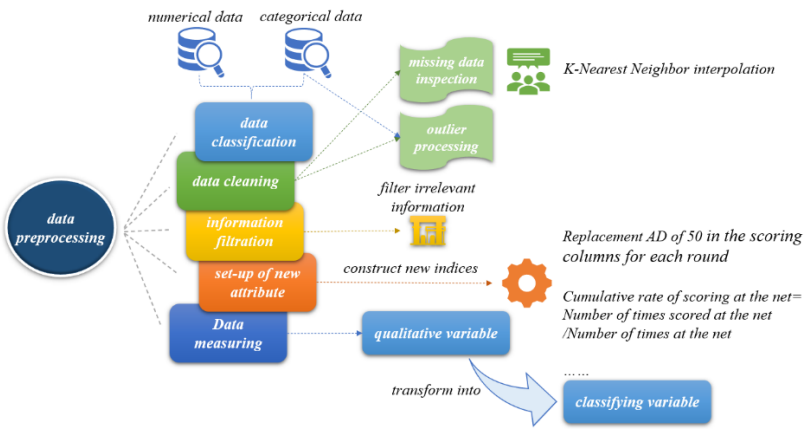

Figure 2. Data preprocessing flow chart

Before analyzing data, it is important to ensure that the data is available. To improve the condition of the data set, it needs to be divided into four steps: data categorization, data cleaning, information filtering, setting of new attributes and data measurement as shown in Figure 2.

By reviewing the given data, we find no outliers, but there are many missing values in the "speed_mph", "serve_width", "serve_depth", and "return_depth" indicators, and their distribution proportion is shown in the following Figure 3.

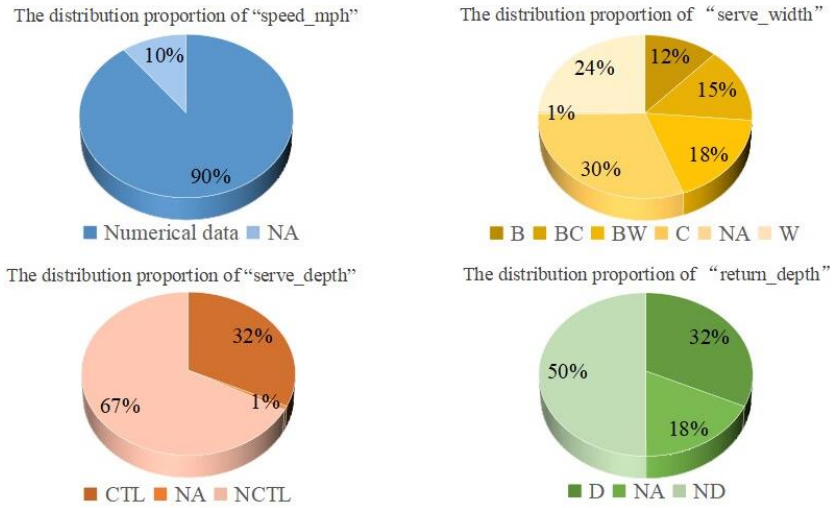

Figure 3. The distribution proportion diagram

As can be seen from the figure, the proportion of missing values in these four indicators is 10%, 1%, 1% and 18% respectively, which is relatively large, and if all of them are deleted, it will lead to incomplete data. Therefore, for numerical data, we have used the Nearest Neighbor interpolation method to fill in the missing values.

At the same time, we recoded categorical variables of columns N and O into dummy variables to facilitate subsequent modelling and analysis. The two new variables generated for column O are named p1_serve_no and p2_serve_no, and column N are named p1_server and p2_server. We also replaced all of the ADs in the scoring columns for each round with 50s to make it easier to compare the size of a player's score in each round.

4.2. Performance Score of Tennis Players

4.2.1 Defining Indicators of the Performance Score

The performance score reflects how well a player performs in a match, and we define six indicators to calculate a player's performance score.

- **Cumulative scoring advantage**

A player's points scored per round is a direct reflection of his performance. The formula for calculating a player's scoring advantage is as follows:

$$G_i = W_i \times p_i \tag{1}$$

Where i denotes the Round i of the match and G_i denotes the player's scoring advantage at Round i. p_i is defined as follows:

$$p_i = \begin{cases} 1, the\ player's\ score\ in\ Round\ i\ is\ higher\ than\ the\ opponent's \\ 0, the\ player's\ score\ in\ Round\ i\ is\ equal\ to\ the\ opponent's \\ -1, the\ player's\ score\ in\ Round\ i\ is\ lower\ than\ the\ opponent's \end{cases}.$$

Considering the recording characteristics of the score within the current game (p1_score, p2_score) in the data (i.e., the result of the score of the current round is recorded in the next row, and the score is recorded in this row as the result of the previous round), we use the score in the next row to represent

the result of the score of the current round. In particular, if Round i is a game point (shown in the data as p1_score=0 and p2_score=0 in the next row), then $p_i = p_{i-1}$.

$$W_i = \begin{cases} \alpha, \ the\ player\ is\ the\ server\ in\ Round\ i \\ \beta, \ the\ player\ is\ the\ receiver\ in\ Round\ i \end{cases}$$

W_i denotes the weight, α is the advantage adjustment factor for the server, and β is the advantage adjustment factor for the receiver, which denotes the relative importance of winning a point for the server and receiver, respectively. Consider that the probability of winning points is much higher for the server compared to the receiver, as the pressure is on the other player to catch up (in particular toward the end of the set) [2]. As a result, the relative importance of the server winning a point is lower than that of the receiver. Referring to Corral and Prieto-Rodriguez (2010) [3], we let $\alpha = 0.8 < 1$ and $\beta = 1.2 > 1$.

The formula for calculating a player's cumulative scoring advantage at Round i is as follows:

$$X_1(i,t) = \sum_{j=t}^{i} G_j \tag{2}$$

Where t denotes that the first round in this set is the tth round of the whole match. Further explanation of t and i can be seen in Figure 4.

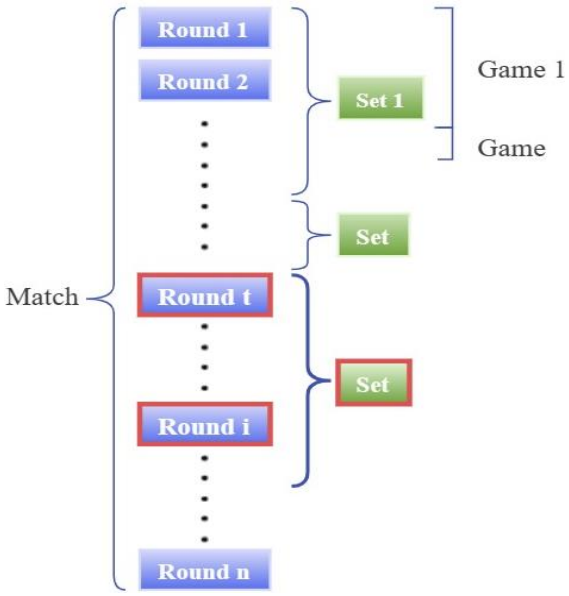

Figure 4. Explanation of t and i

Round i and Round t belong to the same set, but Round t is the first round in this set.

- **The success rate of breaking serve**

The ability to seize break-point opportunities and win the game is an important measure of a player's performance. The formula for calculating a player's success rate of breaking serve at Round i is as follows:

$$X_2(i,t) = \frac{\sum_{j=t}^{i} \omega_j}{\sum_{j=t}^{i} \delta_j} \tag{3}$$

Where t denotes that the first round in this set is the tth round of the match. Further explanation of t and i can be seen in Figure 4. ω_i corresponds to "p*_break_pt_won" in the dataset and ω_i for Round i is defined as follows:

$$\omega_i = \begin{cases} 1, \ the\ player\ won\ the\ game\ when\ the\ opponent\ is\ serving \\ 0, \ others \end{cases}.$$

δ_i corresponds to "p*_break_pt" in the dataset and δ_i for Round i is defined as follows:

$$\delta_i = \begin{cases} 1, the\ player\ has\ an\ opportunity\ to\ win\ the\ game\ when\ the\ opponent\ is\ serving \\ 0, others \end{cases}$$

- **Ability to control the ball**

Ball control is the ability to hit the ball to the desired target in a tennis match and it is crucial for a player to win the match. By controlling the ball, a player can move his opponent around and make him run and return the ball passively, thus increasing the player's chances of scoring. We think that in the data of this question, the running distance (p*_distance_run) can reflect the player's ability to control the ball.

The formula for calculating the score of a player's ability to control the ball at Round i is as follows:

$$X_3(i,t) = \varepsilon_i - \phi_i \tag{4}$$

Where ε_i represents the opponent's running distance in Round i, ϕ_i represents the player's running distance in Round i. The higher the score of a player's ability to control the ball, the better the player's performance in controlling the ball in the round.

- **Cumulative rate of unforced error**

An unforced error is a mistake made by a player on his own that results in the loss of a point, independent of the opponent. Unforced errors in tennis are directly related to the score of the match, so fewer unforced errors mean better performance. The formula for calculating the cumulative rate of unforced error of the player at Round i is as follows:

$$X_4(i,t) = \frac{\sum_{j=t}^{i} \varphi_j}{u} \tag{5}$$

$$u = i - t,\ i \neq t$$

Where t denotes that the first round in this set is the tth round of the match. Further explanation of t and i can be seen in Figure 4. φ_i corresponds to " p*_unf_err" in the dataset and φ_i for Round i is defined as follows:

$$\varphi_i = \begin{cases} 1, the\ player\ made\ an\ unforced\ error\ in\ Round\ i \\ 0, others \end{cases}.$$

- **Quality of the player's serve**

The importance of the serve in tennis cannot be overstated, and at some important points, the quality of the serve can even determine the winner of the match. The quality of a player's serve in a match reflects the player's current competitive performance and it consists of the success rate of serve and the rate of untouchable winning serve. The quality of a player's serve is calculated in three steps.

Step1: Calculate the success rate of serve

The success rate of the serve consists of the success rate of the first serve and the success rate of the second serve. The formula for calculating the success rate of first serve at Round i is as follows:

$$\gamma_{i,t} = \frac{\sum_{j=t}^{i} \eta_j}{\sum_{j=t}^{i} \tau_j} \tag{6}$$

Where t denotes that the first round in this set is the tth round of the match. η_i corresponds to "p*_serve_no" in the dataset and η_i for Round i is defined as follows:

$$\eta_i = \begin{cases} 1, the\ player\ is\ the\ server\ in\ Round\ i\ and\ the\ number\ of\ serves\ is\ one \\ 0, others \end{cases},$$

τ_i corresponds to "p*_server" in the dataset and τ_i for Round i is defined as follows:

$$\tau_i = \begin{cases} 1, \ the\ player\ is\ the\ server\ in\ Round\ i \\ 0, \ others \end{cases}.$$

The formula for calculating the success rate of the second serve at Round i is as follows:

$$\nu_{i,t} = 1 - \frac{\sum_{j=t}^{i} \pi_j}{\sum_{j=t}^{i} \theta_j} \tag{7}$$

Where t denotes that the first round in this set is the tth round of the match. π_i corresponds to "p*_double_fault" in the dataset and π_i for Round i is defined as follows:

$$\pi_i = \begin{cases} 1, \ the\ player\ was\ the\ server\ in\ Round\ i\ and\ missed\ both\ serves \\ 0, \ others \end{cases},$$

θ_i corresponds to "p*_serve_no" in data and θ_i for Round i is defined as follows:

$$\theta_i = \begin{cases} 1, \ the\ player\ is\ the\ server\ in\ Round\ i\ and\ the\ number\ of\ serves\ is\ two \\ 0, \ others \end{cases}.$$

Then, the success rate of the first serve and the success rate of the second serve are weighted and added to calculate the success rate of the serve:

$$\sigma_{i,t} = 0.5 \times \gamma_{i,t} + 0.5 \times \nu_{i,t} \tag{8}$$

Step2: Calculate the rate of untouchable winning serve

An untouchable winning serve is one in which the server uses a lot of force, or applies a subtle angle, to force the receiver to be unable to defend, or the tennis racket to be unable to touch the ball. The formula for calculating the rate of untouchable winning serve at Round i is as follows:

$$\xi_{i,t} = \frac{\sum_{j=t}^{i} \psi_j}{\sum_{j=t}^{i} \tau_j} \tag{9}$$

ψ_i corresponds to "p*_ace" in the dataset and ψ_i for Round i is defined as follows:

$$\psi_i = \begin{cases} 1, \ the\ player\ hit\ an\ untouchable\ winning\ serve\ in\ Round\ i \\ 0, \ others \end{cases}.$$

Step3: Calculate the quality of the player's serve

The quality of the serve consists of the success rate of the serve and the rate of untouchable winning serve. The formula for calculating the quality of the player's serve at Round i is as follows:

$$X_5(i,t) = c \times \sigma_{i,t} + d \times \xi_{i,t} \tag{10}$$

The weight of σ and the weight of ξ of each player are calculated by the Entropy Weight Method. Then the weight of σ and the weight of ξ of two players in a match are averaged to get $\bar{c}'$ and $\bar{d}'$. The final weights of σ and ξ of two players in the match are as follows:

$$\begin{aligned} c &= \frac{\bar{c}'}{\bar{c}' + \bar{d}'} \\ d &= \frac{\bar{d}'}{\bar{c}' + \bar{d}'} \end{aligned} \tag{11}$$

- **Cumulative rate of scoring at the net (X6)**

Advance to the net is an effective offence. The player leaves the baseline and goes to the net to return the opponent's ball before it hits the ground. This method can put some pressure on the opponent, so that the opponent may be disorganised and make a mistake in returning the ball. This

method is difficult for players, and a higher cumulative rate of scoring at the net indicates better competitive performance in the current.

The formula for calculating the cumulative rate of scoring at the net in Round i is as follows:

$$X_6(i,t)=\frac{\sum_{j=t}^{i}\varpi_j}{\sum_{j=t}^{i}\varrho_j} \tag{12}$$

Where t denotes that the first round in this set is the tth round of the match. ϖ_i corresponds to " p*_net_pt_won " in the dataset and ϖ_i for Round i is defined as follows:

$$\varpi_i=\begin{cases}1, \textit{ the player won the point while at the net}\\ 0, \textit{ others}\end{cases}.$$

ϱ_i corresponds to " p*_net_pt " in the dataset and ϱ_i for Round i is defined as follows:

$$\varrho_i=\begin{cases}1, \textit{ the player made it to the net}\\ 0, \textit{ others}\end{cases}.$$

4.2.2 Principal Component Analysis

We determined the weights of the six indicators in the performance score through the PCA method. The main steps are as follows [4].

Assuming there are n samples and 6 indicators, a sample matrix X of size $n\times 6$ can be formed:

$$X=\begin{bmatrix}x_{11} & x_{12} & \cdots & x_{16}\\ x_{21} & x_{22} & \cdots & x_{26}\\ \vdots & \vdots & \ddots & \vdots\\ x_{n1} & x_{n2} & \cdots & x_{n6}\end{bmatrix}.$$

$$x_{ij}^{*}=\frac{x_{ij}-\bar{x}_j}{S_j},\ i=1,2,\cdots,n;\ j=1,2,\cdots 6,$$

Where

$$\bar{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij},$$

$$S_j^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_j\right)^2.$$

Next, the matrix of correlation coefficients of the variables is established:

$$R=(r_{ij})_{6\times 6},$$

$$R=X^TX,$$

Where

$$r_{ij}=\frac{1}{n-1}\sum_{k=1}^{n}x_{ki}^{*}x_{kj}^{*}.$$

Then, we calculated the characteristic equation $|R-\lambda I|=0$ and solved for the characteristic root $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_6>0$ of R and its corresponding unit eigenvector. Where the characteristic root is the variance of the principal component Z_j. The higher the variance, the greater the contribution to the total variance.

We need to determine the weights of the indicators through PCA so we don't reduce the number of factors.

The principal components are as follows:

$$Z_j=a_{1j}X_1+a_{2j}X_2+\cdots+a_{6j}X_6,\ j=1,\cdots,6.$$

Z_j represents the j th new feature and $a_j=(a_{1j},\ a_{2j},\cdots,\ a_{6j})\ (\ j=1,\cdots,6)$ represents the contribution of all original features in the jth new feature. Further, $a_{ij}\ (i=1,2,\cdots,6;\ j=1,\cdots,6)$ represents the angle of rotation of the coordinate system, $a_{ij}>0$ represents a counterclockwise

rotation of the coordinate axes and $a_{ij} < 0$ represents a clockwise rotation of the coordinate axes. Therefore, we used $|a_{ij}|$ to represent the magnitude of the contribution of each original characteristic.

Next, we calculated the mean of the magnitude of the original characteristic contribution:

$$\bar{a}_j = \frac{\sum_{i=1}^{6} |a_{ij}|}{6},\ j = 1, \cdots, 6 \tag{13}$$

Finally, the percentage of $\bar{a}_j$ was calculated as the weight of each indicator for the player's performance score.

$$\mu_j = \frac{\bar{a}_j}{\sum_{j=1}^{6} \bar{a}_j} \tag{14}$$

4.2.3 Calculating the Player's Performance Score

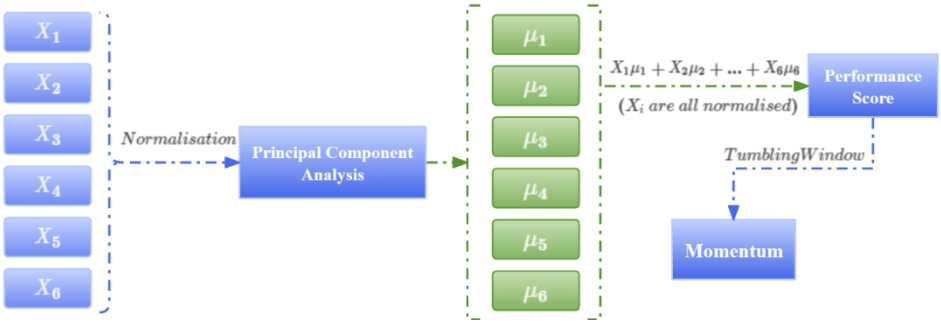

Figure 5. Performance score and momentum design

After principal component analysis, we calculated the weights of each indicator of the performance score:

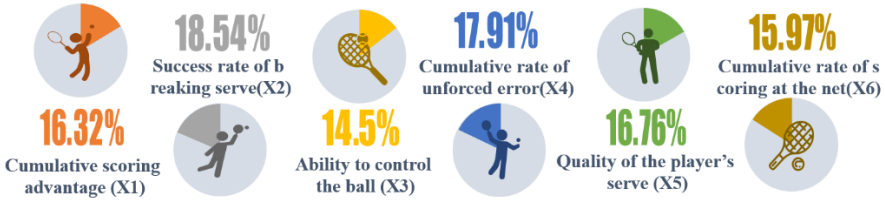

Figure 6. Weights of indicators

Based on the weights and the normalised value of indicators, we can calculate a player's performance score in Round i:

$$P_i = X_1(i,t) \times \mu_1 + X_2(i,t) \times \mu_2 + \cdots + X_6(i,t) \times \mu_6 \tag{15}$$

In the match 2023-wimbledon-1701, Novak Djokovic and Carlos Alcaraz's Performance Scores per round are as follows.

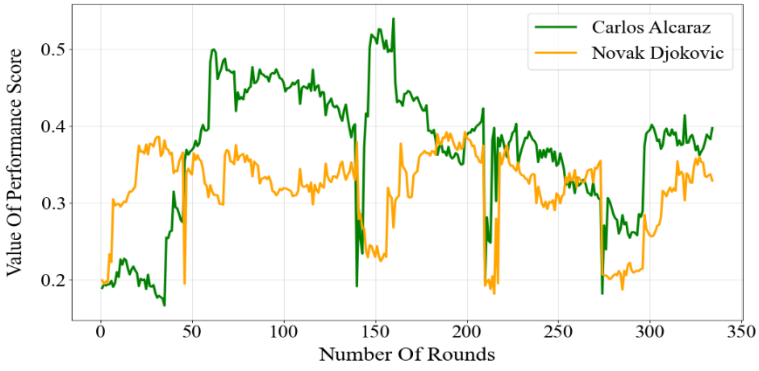

Figure 7. Comparison of performance score

In the first 45 rounds (the first set of the match), Djokovic's performance points were significantly higher than Alcaraz's. Starting in the second half of the first set, Alcaraz's performance scores increased rapidly and were well ahead of Djokovic in the second set. In the last three sets of the match, Alcaraz's performance score was higher than Djokovic's, except in a few games.

4.3. Calculating the Momentum of Tennis Players

A tumbling window is used on the player's performance scores to calculate the momentum of the player, setting the size of the tumbling window to 3 rounds. The calculation process is shown as follows:

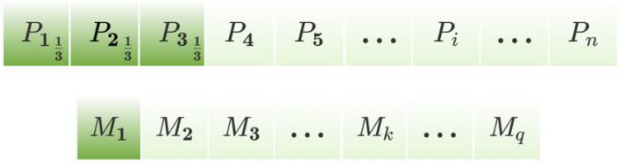

Figure 8. Calculation process for a tumbling window

P_i in Figure 8 represents the player's performance score in Round i and M_k represents the player's momentum in the kth period. The scrolling window is scrolled by mask on the performance score bar in lengths of 3 rounds at a time. The coverage of the mask in the figure includes P_1, P_2 and P_3, then the momentum of the first period is calculated as follows:

$$M_1 = P^T \times m \tag{16}$$

Where

$m = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)^T$,

$P = (P_i, P_{i+1}, P_{i+2}),\ i = 1.$

By the above method, we can calculate a player's momentum over all periods in a match. In the match 2023-wimbledon-1701, Novak Djokovic and Carlos Alcaraz's momentum is as follows.

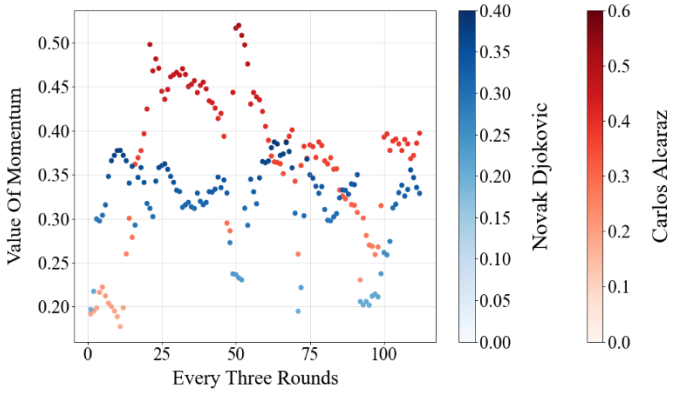

Figure 9. Comparison of momentum

Djokovic's momentum was significantly higher than Alcaraz's in the first 15 time slots, from which he was judged to be better in this period, with momentum peaking at 0.3776 in the 12th time slot. Alcaraz's momentum was at its lowest in the 11th time slot, and increased rapidly from the next time slot, countering Djokovic's momentum from the 16th time slot and far outstripping Djokovic's momentum for the next two sets in the match. He was well ahead of Djokovic in this period, with his momentum peaking at 0.5198 in the 51st time slot. Alcaraz's momentum was also higher than Djokovic's in the next two sets, except for a handful of time slots in the match.

5. Relationship between Swings in Performance Score and Success

Currently, we know the performance scores and results of the players in each round. To test whether swings in the game and player success are random, we need to explore whether fluctuations in a player's performance score affect the player's result for the round. From this, we decided to use Logistic regression to test the causal relationship between the two [5].

5.1. Logistic Regression Model

We defined the outcome of each round as the 0-1 variable *O* which is a categorical variable:

$$O=\begin{cases}1,\ player\ 1\ won\\0,\ player\ 2\ won\end{cases}. \tag{17}$$

We made a difference between the performance scores of the two players in each round to obtain the performance score difference, defined as the variable *D*, which is a numerical variable.

At the same time, we normalize the performance score difference to allow the parameters of the logistic regression to converge faster and make the model more accurate and stable.

Because O itself only takes on discrete values of 0 and 1, it is not suitable to be used directly as the dependent variable in a regression model. Instead of O, we use the probability B of O equal to 1 as the dependent variable. Then, we use logistic transformations to get the logistic regression model:

$$B=\frac{\exp(\beta_0+\beta_1 D)}{1+\exp(\beta_0+\beta_1 D)}, \tag{18}$$

Where, $\beta_0,\ and\ \beta_1$ are the parameters of the model.

5.2. Results Analysis

By using maximum likelihood estimation, the final fitted regression model is:

$$\widehat{B}=\frac{\exp(-0.9186+1.8061D)}{1+\exp(-0.9186+1.8061D)}. \tag{19}$$

The results of the significance tests for the model as a whole and for the regression parameters are collated in Table 2 and Table 3.

Table 2. Parameter results

	B	**S.E.**	**Wald**	**Sig.**	**Exp (B)**
D	1.8061	0.628	2.878	0.004	6.087
Constant	-0.9186	0.342	-2.683	0.007	0.399

Table 3. Model results

	P-value	**Log-Likelihood value**	**P-value for the HL test**
Model	0.003	-227.15	0.982

First, the log-likelihood value of the model is -227.15, which is smaller, indicating a better model fit. While, the p-value of the model as a whole is 0.003, which is significant at the 5% level of significance, indicating that the model is statistically significant in general. In addition to this, we conducted the Hosmer-Lemeshow test and the p-value = 0.982 > 0.05, which accepts the original hypothesis and indicates that the model fit is good.

Second, the Wald test statistic for the coefficient of the performance score difference = 2.878, p-value = 0.004, which indicates that the coefficient is significant at the 5% level of significance. The coefficient of the performance score difference = 1.8061 > 0, which is positive, indicating that the difference in performance scores between player 1 and player 2 is positively related to the probability that the round will result in a win for player 1. That is the higher the performance score of player 1 compared to the performance score of player 2, the higher the probability that player 1 will win.

Meanwhile, Exp(B) = 6.087, suggesting that each unit increase in the performance score difference increases the probability of Player 1 winning by approximately 5.087 times.

To summarize, the swings of the game and the success of the players are not random, they will be affected by the performance scores of the players in the game, and this effect is significant, the degree of influence is large, which shows that the "momentum" plays an important role in the game.

6. Match Swings Prediction Model Based on Softmax-SHAP

Exploring the impact of influences on match swings is a modelling problem with multiple input variables and a single output variable. While, SHapley Additive exPlanations (SHAP) provides a way to quantify the contribution of features to model predictions, helping to identify the most important features. Therefore, we decided to use Softmax Regression to derive which factors influence match swings, to predict match swings, and to explore the importance of these factors using SHAP [6].

6.1. Softmax Regression Model

We make differences in the 6 indicators set up in the first model as the features, which are set as cumulative scoring advantage difference (ΔX_1), the success rate of breaking serve difference (ΔX_2), ability to control the ball difference (ΔX_3), the cumulative rate of unforced error difference (ΔX_4), quality of the player's serve difference (ΔX_5), the cumulative rate of scoring at the net difference (ΔX_6). At the same time, we normalized each feature to make the model converge faster and be more accurate and reliable.

The output variable is the performance score turning point (*tp*) and is a categorical variable. We need to assign values to it before substituting it into the model and we assign values to each category as follows:

$$tp = \begin{cases} -1, \ performance\ score\ difference\ changes\ from\ positive\ to\ negative \\ 0, \ the\ sign\ of\ the\ performance\ score\ difference\ has\ not\ changed \\ 1, \ performance\ score\ difference\ changes\ from\ negative\ to\ positive \end{cases} . \quad (20)$$

We reviewed the dataset and listed the number of each of the three categories of the dependent variable and found that the dataset is not balanced. The number of samples for *tp*=0 is 6789, the number of samples for *tp*=1 is 252, and the number of samples for *tp*=-1 is 243. That is, the number of samples for *tp*=0 is much larger than the number of samples for the other classes, which affects the classification results and biases the basic classifier towards the majority class. Therefore, we performed a resampling of *tp*=0. Since *tp*=0 represents the non-turning point of the performance score, in an actual game, the non-turning point should occur more often than the turning point. So we defined the optimal number of samples by specifying that we use the maximum of the number of samples for *tp*=1 and *tp*=-1 plus 6, which is 252 plus 6 equals 258. This balances the number of samples for the three categories. This treatment resulted in better model performance.

The mechanism of the Softmax regression model is shown below:

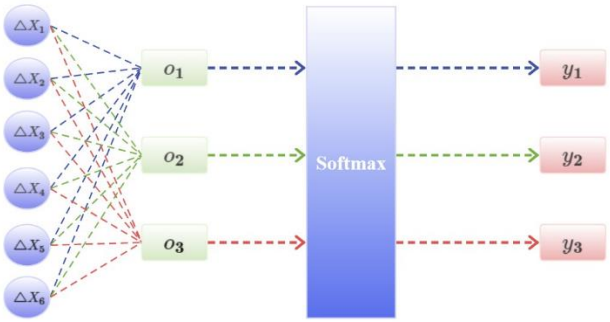

Figure 10. Softmax regression model mechanism diagram

To solve the categorization problem, we output the affine function of the features corresponding to each of its own categories:

$$o_j = \sum_{i=1}^{N} \Delta X_i w_{ji} + b_j, \tag{21}$$

Where o_j represents the linear output of the *j*th category, w_{ij} represents the weight of the *i*th feature in the *j*th category, and b_j represents the total bias of all features in the *j*th category. *i* denotes the *i*th feature and *j* denotes the *j*th category of the output variable. *N* represents the number of features.

The Softmax function $\hat{y} = softmax(o)$ is used for numerical processing to output the probability of each category of the feature to obtain expressions for the model:

$$y_j = \frac{\exp(o_j)}{\sum_{j=1}^{C} \exp(o_j)}, \tag{22}$$

Where, y_j is the predicted probability for each category, and *C* stands for the number of categories.

6.2. Rsults Analysis

The parameters are calculated by using the gradient descent method, and the final fitted regression model is:

$$\begin{aligned} o_1 &= -0.802\Delta X_1 - 0.500\Delta X_2 - 1.504\Delta X_3 - 0.547\Delta X_4 - 0.111\Delta X_5 - 1.264\Delta X_6 + 2.221 \\ o_2 &= 0.723\Delta X_1 + 0.547\Delta X_2 + 0.004\Delta X_3 + 0.132\Delta X_4 - 0.130\Delta X_5 + 0.296\Delta X_6 - 0.622 \\ o_3 &= 0.079\Delta X_1 - 0.047\Delta X_2 + 1.500\Delta X_3 + 0.415\Delta X_4 + 0.241\Delta X_5 + 0.968\Delta X_6 - 1.599 \end{aligned} \quad . \tag{23}$$

The magnitude of w_{ij} indicates the rate of change of the category *j* probability relative to the feature ΔX_i. The larger the value of w_{ij}, the more likely it is that a change in the feature ΔX_i in that direction will result in a large change in the category *j* probability.

Category 3 indicates that Player 1's performance score has increased and surpassed that of Player 2. In *tp*=1, w_{33}=1.5, the maximum. This indicates that the ability to control the ball has the most important effect on player performance. A player's performance score grows with ball control ability. As the ball control of Player 1 becomes progressively more accurate during the game, the flow of the play gradually changes from favouring Player 2 to favouring Player 1. This is followed by the cumulative rate of scoring at the net, quality of the player's serve, cumulative rate of unforced error, and cumulative scoring advantage, all of which have the same positive path of influence on match swings.

6.3. Model Explanation Using SHAP

SHAP constructs an additive explanatory model where all features are considered "contributors". For each predicted sample, the model produces a predicted value, and the SHAP value is the value assigned to each feature in that sample. Assuming that the i th feature is $\Delta X(i)$, the i th feature of the n th sample is $\Delta X(i, n)$, the i th feature of the n th sample in the j th category is $\Delta X(i, n, j)$, the model's predicted value for that sample is SH_n, and the baseline for the entire model (usually the mean of the target variable across all samples) is SH_{base}, the SHAP value obeys the following equation:

$$SH_n = SH_{base} + f[\Delta X(i,n,1)] + f[\Delta X(i,n,2)] + \ldots\ldots + f[\Delta X(i,n,j)], \tag{24}$$

Where $f[\Delta X(i,n,j)]$ is the SHAP value of $\Delta X(i,n,j)$. Intuitively, $f[\Delta X(1,n,j)]$ is the contribution of the 1st feature in the *n*th sample to the final predicted value SH_n.

When $f[\Delta X(i,n,j)] > 0$, it means that the feature enhances the predicted value, and also has a positive effect; and vice versa, it means that the feature makes the predicted value lower, and has a negative effect. The biggest advantage of the SHAP value is that it can reflect each of the feature's

influence in the sample and also shows the positive and negative effects. Our calculated SHAP values and plotted SHAP diagram are shown in Figure 11:

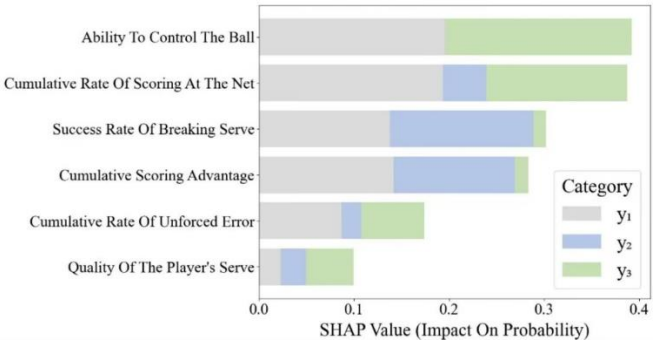

Figure 11. Comparison of SHAP values for the six indicators

As can be seen from Figure 11, in order of relevance to match swings, these factors are the ability to control the ball (X_3), cumulative rate of scoring at the net (X_6), success rate of breaking serve (X_2), cumulative scoring advantage (X_1), cumulative rate of unforced error (X_4), quality of the player's serve (X_5). That is, an increase in player performance score is most related to the ability to control the ball (X_3).

6.4. Proposal of Strategy for the Confrontation

For the player's strategies against different players in a match, we decide to start with the important factors that lead to positive swings in the momentum of the game.

- Cumulative scoring advantage

For defensive players, try to gradually overwhelm the opponent through baseline hit-ting and patient tactics; for offensive players, try to take the initiative and be more focused on fighting for every point to gain a scoring advantage quickly.

- Ability to control the ball

When playing tennis against a full-skilled opponent, players need to have a breakthrough in ball control. This can be achieved by adjusting the players' playing strategy, increasing the depth and accuracy of return and increasing the running distance of the opponent. As to skilled opponents who are good at controlling the tempo of the game, players need to focus more on ball control and minimise errors. Against weaker opponents, players can be more proactive in attacking and putting pressure on their opponents.

- Cumulative rate of unforced error

During a match, with physical exertion, the error rate of a player will significantly increase. Players can reduce their running distances during a match by improving their ball control, which in turn reduces physical exertion and reduces the rate of unforced errors.

- Quality of the player’s serve

Facing opponents with strong receiving ability, players need to improve the quality of their serves, which can be achieved by increasing the speed of the serve, changing the direction of the serve, and adjusting the depth of the serve to minimise the opponent's chances of returning the ball. Facing opponents with weak receiving ability, attacking strategies can be used, such as serve and volley, and high-speed flat serve to increase the rate of untouchable winning serve.

- Cumulative rate of scoring at the net

Facing opponents who favour baseline tactics, players can try to change the game situation by advancing to the net. Players can go to the net at the right time to suppress opponents and increase scoring chances. Against opponents with strong baseline ability, they need to pay more attention to the rate of scoring at the net and break the opponent's defence through volley. Against a weaker opponent, players can be more flexible in utilizing net play to increase scoring opportunities.

In conclusion, it is important to maintain a stable mindset when facing different opponents. Players should focus on their game performance maintain a positive mindset and control their emotions.

Avoid getting caught up in guessing and speculating about the opponent's strength and the outcome of the match, learn to stay calm under pressure.

7. Analysis of the Predictive Effectiveness and Generality

7.1. Model Test

Now, we have built a Softmax regression model that can predict match swings with multiple influencing factor indicators. We then tested the model using the dataset of 2023-wimbledon-1311, and the results of the prediction are shown in Figure 12.

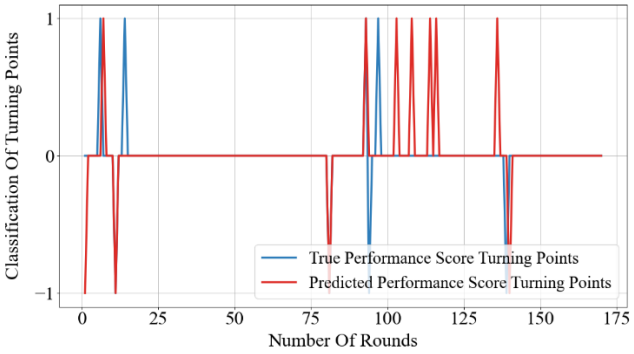

Figure 12. Prediction and reality comparison diagram of 2023-wimbledon-1311

As can be seen from the figure, the prediction result is more similar to the real value, and it can be considered that the prediction result is better.

To evaluate the performance of the model on the problem of predicting other matches, we measure it by constructing a confusion matrix [8] and obtain the corresponding evaluation indicators such as accuracy, precision, recall, and f1-score. The confusion matrix and evaluation indicators are as follows:

Table 4. Confusion matrix

Predicted value	**Real value**		
	Real Positive	Real Negative	Total
Predicted Positive	True Positive (TP)	False Positive (FP)	Total marked positive (TP+FP)
Predicted Negative	False Negative (FN)	True Negative (TN)	Total marked negative (TN+FN)
Total	Total actually positive (TP+FN)	Total actually negative (FP+TN)	Total number of samples in the test set (TP+FP+FN+TN)

Accuracy is the rate at which the model predicts all samples correctly.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}. \tag{25}$$

Precision is the rate at which the model predicts positive samples correctly.

$$Precision = \frac{TP}{TP + FP}. \tag{26}$$

Recall is the rate at which the model predicts actual positive samples correctly.

$$\mathrm{Re}call = \frac{TP}{TP + FN}. \tag{27}$$

F1-score is the reconciled mean of precision and recall.

$$F1-score=\frac{3}{\frac{1}{Precision}+\frac{1}{\mathrm{Re}call}}=\frac{2*Precision*\mathrm{Re}call}{Precision+\mathrm{Re}call}. \tag{28}$$

All four evaluation indicators are the closer to 1, the better the model effect and performance. The confusion matrix and evaluation indicators of turning points are shown in Figure 13 and Table 5.

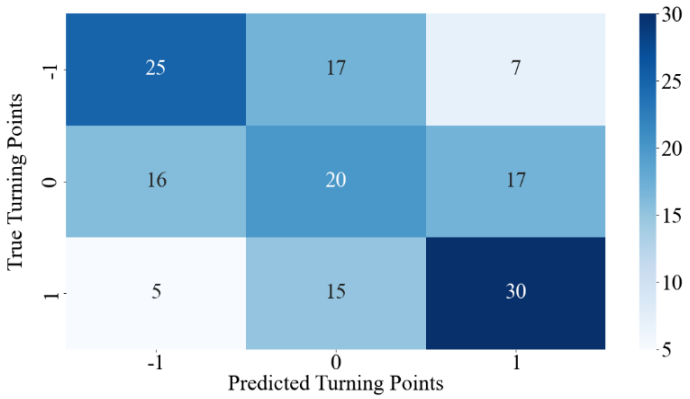

Figure 13. Confusion matrix of turning points

Table 5. Evaluation indicators for Softmax models

Accuracy	Precision	Recall	F1-score
0.493	0.492	0.493	0.492

This suggests that based on the indicators we have provided so far, this model is not performing well enough in the classification task. Therefore, it is suggested that we will perform further adjustments and evaluation of the model.

7.2. Definition of Other Impact Factor Indicators

As can be seen from the table, by combining multiple evaluation indicators mainly in terms of accuracy, the performance of the model in predicting swings in other matches has a large room for improvement, so we propose the influencing factors that may need to be included in the model in the future:

- Direction of serve
- Depth of serve
- Depth of return
- Speed of serve

Each player's technical strengths and weaknesses are different. The server can increase the likelihood of scoring points by changing the direction and depth of the serve and increasing the speed of the serve to find the receiver's technical weaknesses. Similarly, the receiver can change the depth of return to find the weakness of the opponent and increase the possibility of scoring.

- Number of shots during the point
- Length of break before this round

Tennis is a high-intensity sport, the number of shots during the point directly determines the amount of physical exertion of a player. With physical exertion, the error rate of a player will significantly increase, thus affecting the score of the match. Resting before the rounds can help players recover their physical energy and improve their athletic status, which will affect the trend of the score. In summary, the above factors will affect the swings in the match and should be included in the model.

7.3. Generality Analysis

To analyze the generality of the model, we used it to fit the dataset of Sebastian Ofner vs. Stefanos Tsitsipas in the men's singles eighth-final of the 2023 French Open Roland Garros and the results are shown below.

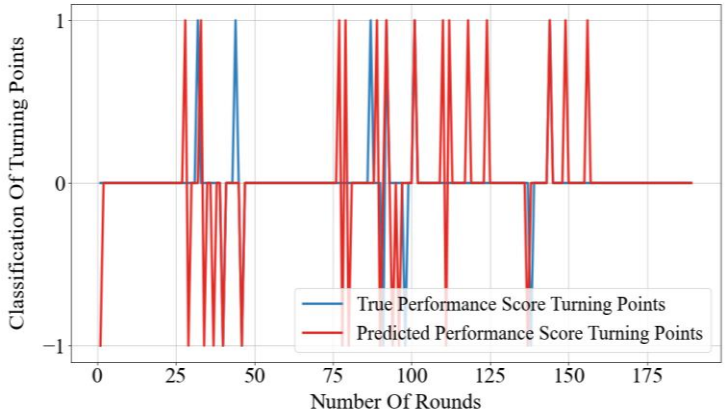

Figure 14. Prediction and reality comparison diagram of Ofner vs. Tsitsipas court

As can be seen from Figure 14, the predicted results are closer to the true values. This is due to the fact that the essence of ball games is similar. Athletes will have "momentum", and these "momentum" will be affected by factors such as scoring advantage, ability to control the ball, quality of serve, rate of error and so on. The model we constructed covers the influence of many factors, so it is widely used and has a high degree of generalizability.

Our model can easily have the same intrinsic mechanism across different matches. Softmax regression, as a deep learning method, is a classification algorithm for multi-categorization problems [9]. Compared with other classification algorithms, Softmax regression can better adapt to the distribution of complex data and can effectively solve multiclassification problems. In addition, it can be used in combination with other algorithms, such as convolutional neural networks, recurrent neural networks, etc., to adapt to different application scenarios [10].

To summarize, our model can be used not only in Wimbledon 2023 Gentlemen's singles matches but also in many similar matches, such as Women's matches, tournaments, table tennis and so on. We can also apply our model to solve more practical problems. A large amount of data can enrich our sample collection, and a good model can better help us to solve the prediction problem of multi-categorization, which strengthens the rationality of the whole model.

8. Sensitivity Analysis

We explored model and forecast uncertainty using sensitivity analyses that adjusted model parameter methods. For parameter adjustment, we fixed the other variables unchanged and adjusted the variable of ability to control the ball, which had the most significant effect in the Softmax regression model, in the range of [0.7, 0.9] and at intervals of 0.05. We re-predicted the data using the model after adjusting the parameters, and the results are shown in Table 6.

As shown in Table 6, changes in the model parameters affect the predicted results, but not to a high degree, indicating a low level of uncertainty in the model and predictions. It can be seen that all of these parameters have a reasonable impact on the results, and the model is less sensitive. That is, small changes in the input data will not have a significant impact on the model's prediction results, which makes the model more robust and reliable.

Table 6. Sensitivity and uncertainty analysis

Value of change in the ability to control the ball	Probability of reduced performance score	Probability of constant performance score	Probability of increase in performance score
0.7	0.204	0.375	0.421
0.75	0.186	0.369	0.445
0.8	0.169	0.361	0.470
0.85	0.153	0.353	0.495
0.9	0.138	0.343	0.519

9. Model Evaluation and Further Improvement

9.1. Strengths and Weakness

- In quantifying the performance score, we constructed many new indicators based on existing indicators, synthesized various factors, and used Principal Component Analysis to determine the weights, which is more comprehensive in consideration and more objective in results. When measuring the momentum, we use the Tumbling Window Algorithm, which allows the data to be processed in real time and the model is novel.
- In building the Logistic regression model, we make a difference in the player's performance score as the independent variable and normalize it to reduce the influence of the feature scale on the model, which is conducive to faster convergence of the model and makes the model more accurate and reliable.
- We combine Softmax regression and SHAP to provide a more detailed and intuitive explanation of feature importance for multi-category classification models, which helps to deeply understand the decision-making process of the model and makes the model prediction results more convincing. It can also guide the optimization of the model, making the model more consistent with practical application scenarios.
- In predicting match swings, we used the constructed six indicators but did not consider the categorical variables as independent variables, such as direction of serve and depth of return, and did not explore their impact on match swings in depth.

9.2. Further Improvement

- In subsequent studies, we can add more categorical variables into the model and test the significance of the coefficients, eliminating the non-significant variables and retaining the significant ones, thus making the model more accurate and effective.
- In real life, the influences on match swings are more complex than we think, and we should consider a more comprehensive model to describe them.

10. Conclusion

In this article, we employed different models to analyze the information of match results. In general, we have achieved the following conclusions:

- Momentum can reflect the performance of a player at a given time in the match, and momentum can be quantified by six indicators: cumulative scoring advantage, the success rate of breaking serve, ability to control the ball, cumulative rate of unforced error, quality of the player’s serve and cumulative rate of scoring at the net.
- Swings in play and runs of success by one player are not random. Each unit increase in the performance score difference increases the probability of the player winning by approximately 5.087 times, which shows that "momentum" plays an important role in the game.
- Six indicators can help determine when the flow of play is about to change from favouring one player to the other. An increase in a player’s performance score and momentum is most related to the ability to control the ball, followed by the cumulative rate of scoring at the net and cumulative rate of unforced error.
- The model we developed had favourable prediction results in other competitions and was more generalizable, but the performance results could be improved. Therefore, we proposed six factors that may need to be included in the model, such as direction of serve, depth of serve, depth of return, length of break before this round, number of shots during the point, and speed of serve.

11. The Memo

Date: Feburary 5th, 2023
To: All tennis coaches
From: Team # 2402382
Subject: Recommendations for tennis matches

In sports, a team or player may feel they have strength or force during a match which is often attributed to “momentum.” We define a series of indicators and build several models to explore the role of momentum in a tennis match and the factors that influence it.

Momentum can reflect the performance of a player at a given time in the match, and momentum can be quantified by six indicators: cumulative scoring advantage, the success rate of breaking serve, ability to control the ball, cumulative rate of unforced error, quality of the player’s serve and cumulative rate of scoring at the net.

The swings of the game and the success of the players are not random, they will be affected by the performance scores of the players in the game, and this effect is significant. Each unit increase in the performance score difference increases the probability of the player winning by approximately 5.087 times, which shows that "momentum" plays an important role in the game. In terms of factors, we find that an increase in a player’s performance score and momentum is most related to the ability to control the ball, followed by the cumulative rate of scoring at the net and cumulative rate of unforced error. The model we developed had favourable prediction results in other competitions and was more generalizable.

Given the important role of momentum in winning matches and the important factors in improving a player's performance score and momentum, we have the following recommendations for you:

- Help players recognise the important role of momentum in the game and help them improve their abilities and factors related to momentum swings through training.
- Focus on improving the players' ability to control the ball during pre-match training. By controlling the ball, a player can move his opponent around and make him run and return the ball passively, thus increasing the player’s chances of scoring and the player’s momentum.
- Focus on advancing to the net, which is an effective offence. The player leaves the baseline and goes to the net to return the opponent's ball before it hits the ground. This method can put some pressure on the opponent, thus increasing the rate of scoring at the net during the match, which in turn increases the momentum of the player.
- Take care to improve the players' stamina. With physical exertion, the rate of unforced error will significantly increase, thus affecting the score of the match. Therefore, good stamina can reduce the rate of unforced error in the match to some extent and does not negatively affect a player's momentum.
- Before the match, players should be instructed to reduce unforced errors, take advantage of break chances, attack aggressively to the net when they have a chance, and make short adjustments through the bathroom break if they are out of form.

References

[1] Dietl H, Nesseler C. Momentum in tennis: Controlling the match [J]. UZH Business Working Paper Series, 2017 (365).

[2] Page, Lionel. "The momentum effect in competitions: field evidence from tennis matches." Econoindicator Society Australasian Meeting. 2009.

[3] Corral, J., & Prieto-Rodriguez, J. (2010). Are differences in ranks good predictors for Grand Slam tennis matches? International Journal of Forecasting, 26, 551–563.

[4] Davies A, Fearn T. Back to basics: the principles of principal component analysis [J]. Spectroscopy Asia, 2005, 1 (1): 35-38.

[5] Montagna S, Orani V, Argiento R. Bayesian isotonic logistic regression via constrained splines: an application to estimating the serve advantage in professional tennis [J]. Statistical Methods Applications, 2020, 30 (2): 1-32.

[6] Henrikki T, M. HRP, Elsa A, et al. Explaining a century of Swiss regional development by deep learning and SHAP values [J]. Environment and Planning B: Urban Analytics and City Science, 2023, 50 (8): 2238-2253.

[7] Wan Lei, Tong Xin, Sheng Mingwei. Review of Image Classification Based on Softmax Classifier in Deep Learning [J]. Navigation and Control, 2019, 18 (6): 1-9.

[8] Townsend J T. Theoretical analysis of an alphabetic confusion matrix [J]. Perception & Psychophysics, 1971, 9: 40.

[9] Michael Nielsen. (2015). Neural networks and deep learning. Determination Press.

[10] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.