| Problem Chosen | 2024 MCM/ICM Summary Sheet | Team Control Number |
|---|---|---|
| C | | 2429507 |

# The Rising Tide: The dance of Tennis Momentum


## Summary

In the 2023 Wimbledon men's singles final, the young Spanish player Carlos Alcaraz successfully defeated Novak Djokovic. Although Djokovic had an overwhelming advantage in the first set, Alcaraz showed strong counterattack ability and came back to win in the subsequent sets. The result ended Djokovic's long reign at Wimbledon and sparked widespread discussion about **momentum**.

Firstly,To quantify and analyze a player's "momentum," we conducted data analysis and mathematical modeling studies. After cleaning and analyzing the given data, we extracted the players' first-serve scoring rate, Ace probability and other indicators in the game, and conducted **Factor Analysis**. The **KMO** value is **0.792**, which is suitable for Factor Analysis. Finally, through analysis, we obtained the player representative factors**F1,F2,F3** , and obtained the player momentum $F_{sum}$**=(0.45664**$F_1$ **+ 0.22117**$F_2$ **+ 0.0763**$F_3$**)/0.75411**, and **the dynamics of momentum during the game are described through visualizations**.

Secondly,To determine the correlation between player momentum and their characteristics, we used **Spearman Correlation Analysis** and drew **heat maps** to illustrate the correlation between indicators. This analysis helps identify factors that influence momentum. In addition, we built a **Random Forest** model to predict players' momentum in other games and derived corresponding evaluation indicators, **(MSE=0.002, RMSE=0.05, MAE=0.041, MAPE=9.167, R2=0.555)**. The results show that momentum is not random but is determined by a combination of player metrics, providing coaches with specific metric analysis.

Thridly,In order to predict fluctuations in momentum, we adopt a sliding window **LSTM** model and predict the overall momentum trend. Through this model, we obtained a root mean square error (RMSE) value of the model of **0.0179**. And we use the **Markov Chain** to predict the winning rate of a game. Since the transition matrix probability of the traditional Markov Chain is fixed, the prediction of the winning rate of the game is poor, so we propose an improved Markov chain. The improved model does not rely on the static state transition matrix, but uses the **Logistic Regression Model** to dynamically calculate the winning probability to update the transition probability of the score state. The model shows higher prediction accuracy than the traditional method. The accuracy reaches **67.86%**, the recall rate is **74.19%**, and the F1 value is **71.88%**, which shows that our model has high accuracy and flexibility.

Next,At the same time, we also considered the impact of weather, venue and other factors on the game to improve the accuracy of the model. Our model also has good transferability to other women's games or volleyball matches and is able to predict the fluctuation trend of the game.

Finally,Based on the above research, we can provide players with suggestions when facing a new game, and apply our model to other games to predict the fluctuation trend and probability of winning or losing in the game.



**Keywords**: Factor Analysis,Spearman Correlation Analysis,RandomForest Model,LSTM,Logistic Regression Model,Markov Chain

# Contents

# 1 Introduction

## 1.1 Background

Young Spanish player Carlos Alcaraz defeated Novak Djokovic in the 2023 Wimbledon men's singles final. Djokovic won the first set overwhelmingly, but Alcaraz showed strong fightback ability in the following sets to win. The result ended Djokovic's long reign at Wimbledon and sparked discussion of "momentum".

Incredible swings in seemingly dominant players, sometimes even multi-point or multi-game swings, are chalked up to "momentum." A team or player may feel they are motivated during a game, but it is difficult to measure this phenomenon. Therefore, analyzing what factors affect the momentum in the game, predicting changes in the situation during the game, and evaluating the performance between players can better provide coaches with suggestions on how to respond to changes in the situation during the game.

## 1.2 Problem Restatement

This question gives us the scores for all men's matches after the first two matches at Wimbledon 2023. This comprehensive tennis match data set includes precise details such as matches, sets, games, game point calculations, serves, breaks, break points, tiebreaks, aces, double faults, and more. Explained to us the definition of "momentum" in sports. Based on the game records, we will now address the following issues:

- Develop a model that captures the process as a score occurs, apply it to one or more games to identify better performing players, and provide visual data. Also take into account the server's high probability of winning the point into the model.
- Using a model or metric to evaluate the claims of a tennis coach who doubts the role of "momentum"
- Explore some indicators of when the flow of a game shifts from favoring one player to favoring another.
- Using the provided match data, build a volatility prediction model and identify relevant factors
- Given the differences in historical match swings, suggest how players might be advised to play against different players in new matches
- Test the developed model in one or more games and evaluate the model's predictive ability and model generality on game trends

  Finally, write a memo to introduce the research results to the coach and make recommendations.

## 1.3 Our Work

In problem 1, After preliminary analysis and processing of the data, we selected eighteen indicators for factor analysis to quantify the role of "momentum" in the game. These eighteen

indicators took into account the impact of serving. Momentum is converted into corresponding player scores. By visualizing the data, you can observe the momentum changes and performance levels of the two players as the game progresses.

We divide problem 2,Based on Spearman correlation analysis and random forest regression, it is determined that "momentum" is not random, but is jointly determined by certain characteristics of the players.

As for problem 3, The traditional Markov chain intermediate probability using a fixed state transition matrix will not change. Therefore, the method we propose uses the probability formula trained in the logistic regression model to dynamically calculate the winning probability when the score changes. This method takes into account Player momentum changes and service rights transitions during the game provide more accurate predictions of the game. At the same time, the LSTM model based on the sliding window is used to predict the fluctuation trend of the game and visualize the fluctuation trend of the game. Provided suggestions for players to participate in future competitions.

With regard to problem 4,We tested the momentum in other games and predicted game fluctuations. At the same time, we took factors such as terrain, weather, and psychology into consideration in future models. We applied the model to other scenarios and concluded that the model is suitable for single-player games and has obvious characteristics. The migration of features is better in the competition.

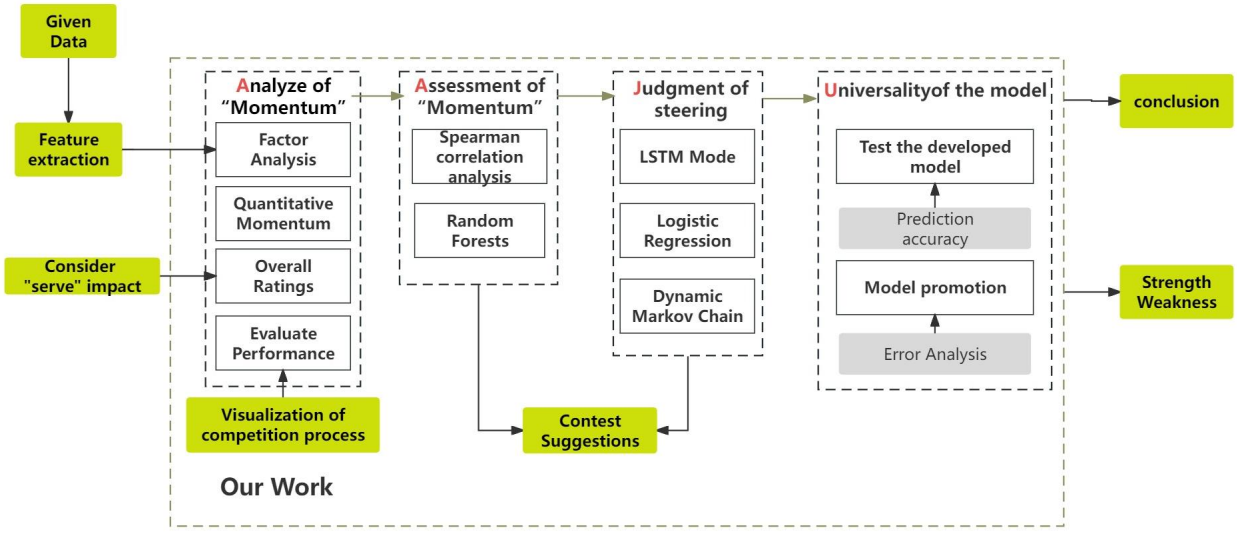



Figure 1: Our Work

# 2 Assumptions and Symbols

## 2.1 Assumptions

- Assume that a player's skill level remains essentially the same over a short period of time. This assumption may not always hold true, as a player's status may be affected by factors such as injuries, mental state, training level, etc. However, for the purposes of analyzing momentum, consider that players' skill levels are relatively stable over the short term.
- It is assumed that momentum changes are possible, i.e. players may experience winning momentum or losing momentum during a match. This hypothesis can be tested by looking at winning or losing streaks in games.
- It is assumed that the real-time information used (such as live game broadcasts, statistics) is accurate. The conclusions of the analysis may be affected by the quality and timeliness of the information.

## 2.2 Symbols

Table 1: Add caption

| Notations | description |
|---|---|
| $tanh$ | An logit functions,used to purify information |
| $x_t$ | current status informations |
| $h_t$ | the hidden state of the output |
| $\sigma$ | an activation function |
| $P_ij$ | change of status |

# 3 Quantification of Momentum Based on Factor Analysis

## 3.1 Indicator selection

In tennis, many factors can affect the outcome of the game in different ways and in varying depths. This tennis match data set provides a total of 46 items of data for analyzing the influencing factors of the match results.

Preliminary analysis of the indicators included in the data (listed indicators) found that the match markers are only used to distinguish matches and have no direct influencing factors. The number of serving games will not directly affect the outcome of the match, so these data are not included in this analysis of influencing factors.

The remaining data affects the game in the serving link, receiving link, and other links. There are certain internal connections, influencing each other, and jointly determining the momentum of the game. Therefore, we decided to construct twenty indicators for analysis.

For the convenience of subsequent data analysis, these twenty indicators are converted into corresponding ratio indicators. The ratio indicators are divided into the following three links according to different links:

**Serving session** The probability of an ACE being served in the service game $X_1$, the probability of a double fault in the service game $X_2$, the probability of facing a break point in the service game $X_3$, the success rate of the first serve $X_4$, the scoring rate of the first serve $X_5$, and the scoring of the second serve rate $X_6$, success rate in saving break points $X_7$, winning rate in serving games $X_8$

**Catching session** The scoring rate of returning the first serve $X_9$, the scoring rate of returning the second serve $X_{10}$, the probability of break chance in the receiving game $X_{11}$, the break success rate $X_{12}$, and the winning rate of winning the returning game $X_{13}$

**Scoring session** Serve scoring rate $X_{14}$, total scoring rate $X_{15}$, player physical fitness indicators $X_{16}$, unforced errors $X_{17}$, consecutive points $X_{18}$, number of consecutive shots $X_{19}$, and current score advantage $X_{20}$.

Table 2: Indicator system and conversion method

| link | index | Conversion method |
|---|---|---|
| Serving session | Probability of delivering an ACE ball | ACE balls /service games |
| | probability of double error | double faults/service games |
| | Probability of facing a break point | Break points/service games |
| | First serve success rate | first serve/Total first serves |
| | Second serve scoring rate | second serve/Total second serves |
| | Success rate in saving break points | saving break points /break points |
| | Win the serve game | service games won/service games won |
| Catching session | Points scored after return of serve | return serve/returns after a serve |
| | Second serve return scoring rate | returning second serve/total returned second se |
| | Probability of break chance in return game | Break points/number of service games received |
| | Break success rate | Number of break successes/break points |
| | Win percentage of return games | service games won /service games won |
| Scoring session | serve scoring rate | Points scored by serve/Total points scored |
| | Total scoring rate | Total points scored/Total games played |
| | unforced error | Number of unforced errors/Total points scored |
| | Current score advantage | p1_point / (p1_point + p2_point) |

## 3.2 Factor Analysis

This analysis selected 2023-wimbledon-1701 games for analysis, and conducted factor analysis on the two players respectively. The following is the analysis of Carlos Alcaraz (the same is true for Novak Djokovic):

In order to ensure that all indicators contribute positively to the score, before performing factor analysis, the two indicators of the occurrence of the service game (here are the specific negative indicators) should be positively transformed.

First, conduct the first factor analysis. According to the analysis results, only the common factors of the current score advantage have a common degree less than 0.5, and the $X_{20}$ indicator is deleted directly. Conduct factor analysis again until all indicators meet the conditions and retain them all.

The final results include the probability of an ACE being served in the service game $X_1$, the probability of a double fault in the service game $X_2$, the probability of facing a break point in the service game $X_3$, the success rate of the first serve $X_4$, and the scoring rate of the first serve $X_5$. , second serve scoring rate $X_6$, break point saving success rate $X_7$, winning serve rate$X_8$, first serve return scoring rate $X_9$, second serve returning scoring rate$X_{10}$, receiving serve Probability of game break chance $X_{11}$, break success rate $X_{12}$, winning rate of receiving and receiving games $X_{13}$, serve scoring rate $X_{14}$, total scoring rate $X_{15}$, player physical fitness index $X_{16}$, unforced Turnovers $X_{17}$, consecutive points $X_{18}$. Eighteen indicators are used as a technical indicator system for analyzing factors affecting tennis matches.

In order to verify the correlation of the eighteen selected indicators and determine whether they are suitable for factor analysis, the principal component analysis method was used. In this process, the suitability of the selected indicators and the significance of the correlation were first assessed using the KMO test and the sphericity test. The rotation method was based on the varimax method and the regression method was chosen to calculate the factor scores.

### 3.2.1 KMO and Bartiett's test of sphericity

The KMO test is based on calculating the correlation matrix of the data and converting the correlation of indicators into commonality. It measures the degree to which each indicator is related to other indicators to determine whether there is enough commonality for use in factor analysis.

Table 3: KMO and Bartlett's Test

| | | |
|---|---|---|
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy | | 0.792 |
| Bartlett's Test | Approximate chi-square | 7444.857 |
| | degrees of freedom | 190 |
| | Significance | 0.000 |

According to the results of our KMO test and Bartlett's sphericity test, we got the following conclusion: the KMO test value is 0.729, which is close to 1. At the same time, the result of Bartlett's sphericity test shows that the significant P value is 0.000, which is less than the significance level of 0.01. This means that there is a significant correlation between the eighteen selected indicators and enough commonality to be suitable for factor analysis.

### 3.2.2 Extract Common Factors

For the common factor of common degree: serve scoring rate, winning rate of serving games, first serve scoring rate, second serving scoring rate, unforced errors, ACE ball scoring, first serve success rate, winning rate of returning serve games, these indicators The extracted information is greater than 0.8, which indicates that they can be well expressed by common factors. The scoring rate after serving, the probability of facing a break point in the service game, the total scoring rate, the break success rate, physical fitness indicators, and the current score advantage are between 0.6-0.7, and can be better expressed by the common factor. The probability of consecutive points, second-serve return scoring rate, and break-serve opportunity is also greater than 0.5. Although there may be noise, it can also be expressed.

### 3.2.3 Variance Explained

As shown in the table below, we sorted the characteristic roots of the eighteen selected indicators. The eigenvalues of the first three factors are greater than 1, and the cumulative variance contribution rate reaches 75.441%. Therefore, we decided to extract these three factors as main factors, recorded as $F_1$, $F_2$, and$F_3$ respectively.

Table 4: Total variance explained

| Total variance explained | | | | | | |
|---|---|---|---|---|---|---|
| Element | Variance explained rate before rotation | | | Variance explanation rate after rotation | | |
| | characteristic root | Variance explanation rate (%) | Cumulative variance explanation rate (%) | characteristic root | Variance explanation rate (%) | Cumulative variance explanation rate (%) |
| 1 | 8.22 | 45.664 | 45.664 | 768.656 | 42.703 | 42.703 |
| 2 | 3.981 | 22.117 | 67.782 | 444.049 | 24.669 | 67.372 |
| 3 | 1.373 | 7.63 | 75.411 | 144.702 | 8.039 | 75.411 |
| 4 | 0.994 | 5.525 | 80.936 | | | |
| 5 | 0.79 | 4.389 | 85.326 | | | |

The rotated $F_1$ has a higher load in terms of the second serve scoring rate, the probability of serving an ACE ball, the winning rate of winning the serving game, the winning rate of winning the returning serve, and the probability of facing a break point in the serving game. These indicators are used for An indicator of the stability of the serve, so it is named the serve stability factor. $F_2$ has higher loadings in terms of probability indicators of first serve scoring rate, first serve success rate, break point saving success rate, and serve scoring rate. These are all indicators of serving efficiency, so it is named the serving efficiency factor. $F_3$ has a high loading in terms of probability indicators of continuous scoring and current score advantage. These are all indicators of measuring the game advantage score, so it is named the game advantage factor.

### 3.2.4 Factor Score

According to the factor score coefficients obtained in the above table, calculate the total score of factors affecting tennis matches. The calculation formula is as follows:

$F_1$= 0.091$Y_1$ + 0.046$Y_2$ + 0.023$Y_3$ - · · · · · +0.124$Y_{18}$

$F_2$= 0.08$Y_1$ + 0.16$Y_2$ + 0.183$Y_3$ - · · · · · -0.103$Y_{18}$

$F_3$= 0.016$Y_1$ + 0.031$Y_2$ + 0.023$Y_3$ - · · · · · +0.001$Y_{18}$

$F_{sum}$=(0.45664$F_1$ + 0.22117$F_2$ + 0.0763$F_3$)/0.75411

Table 5: Ingredient matrix table

| Ingredient matrix table | | | |
|---|---|---|---|
| name | Element | | |
| | Ingredient 1 | Ingredient 2 | Ingredient 3 |
| Second Serve Points Won | 0.091 | 0.08 | 0.016 |
| Total Service Points Won | 0.046 | 0.16 | 0.031 |
| First Serve Points Won | 0.023 | 0.183 | 0.023 |
| First Serve | -0.066 | 0.222 | -0.06 |
| Aces/Service Games Played | 0.116 | 0.036 | -0.117 |
| hot streak | -0.024 | -0.027 | 0.513 |
| physical fitness indicators | -0.121 | 0.053 | 0.164 |
| score advantage | -0.033 | -0.042 | 0.571 |
| Service Games Won | 0.093 | 0.078 | 0.025 |
| Return Games Won | 0.101 | 0.046 | -0.003 |
| Break Points Saved | -0.024 | 0.204 | -0.049 |
| Break Points Converted | 0.081 | 0.066 | 0.063 |
| unforced error | 0.103 | -0.203 | 0.034 |
| Break Points Opportunities/ReturnGames Played | -0.085 | 0.063 | -0.166 |
| Total Points Won | 0.102 | -0.026 | 0.141 |
| Second Serve Return Points Won | -0.105 | 0.021 | 0.171 |
| First Serve Return Points Won | 0.117 | -0.095 | -0.029 |
| Break Points Faced/Service GamesPlayed | 0.124 | -0.103 | 0.001 |

F1 represents the service factor, F2 represents the 2 factor, and F3 represents the scoring factor. The contribution rate of F1 in the game reaches 42.703%, the contribution rate of the F2 factor reaches 24.669%, and the contribution rate of the F3 factor reaches 8.039%.

### 3.2.5 Visualize the Competition Process

Calculate the score of each ball in the game for the two players, and draw the following chart by visualizing the data:

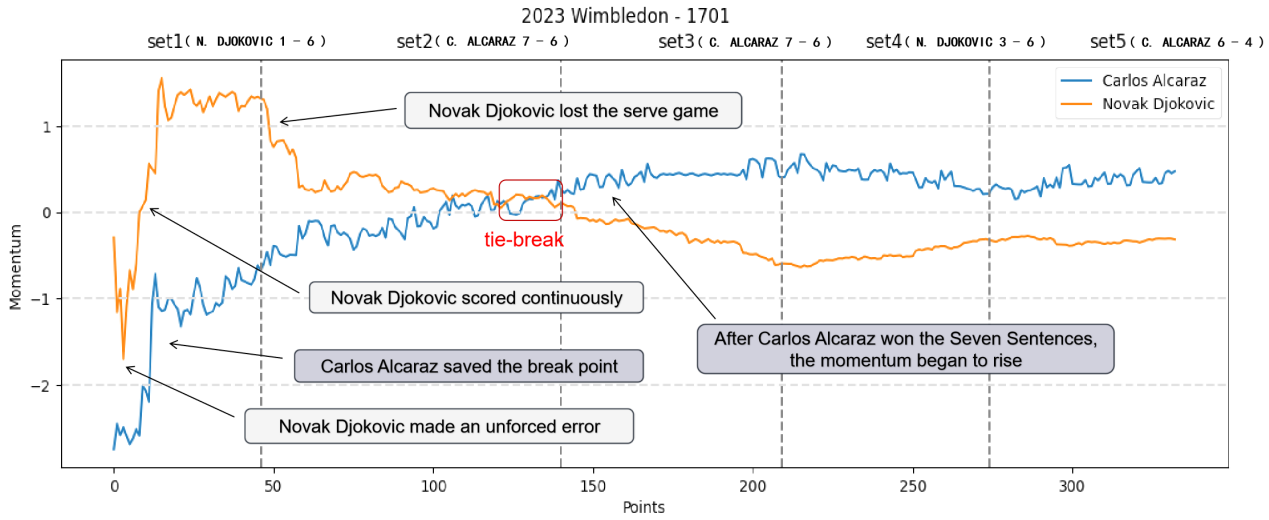



Figure 2: Momentum change curves of players on both sides in 2023-wimbledon-1301

The picture marks several important moments of the game:

Novak Djokovic made an unforced error at the start of the first set and the momentum dropped. Carlos Alcaraz saved a break point in the match and the momentum increased significantly. In the process of scoring consecutively, Novak Djokovic gradually gained momentum. Novak Djokovic lost his serve at the beginning of the second set and his momentum dropped significantly. He gained momentum and overtook Novak Djokovic after Carlos Alcaraz won the tie-break.

Player performance:

Novak Djokovic (orange line) made an error at the start of the first set and the momentum dropped. However, his momentum picked up as the game progressed, especially after scoring consecutive points.

Carlos Alcaraz (blue line) had a relatively smooth second set, and a significant surge in momentum after winning the set in the tie-break showed he was gaining both a psychological and scoreline advantage.

It can be seen from the figure that although the momentum kept changing during the game, Alcaraz won more sets when the momentum was higher, especially after winning the second set tiebreaker, his overall momentum increased and Finally won the game. Djokovic's momentum picked up a few times early in the game, but fell off in a few big moments later.

# 4 Spearman Correlation Coefficient and RF to correlation analysis

## 4.1 Spearman Correlation Coefficient

In order to prove that "momentum" is not random but determined by relevant characteristics, we apply the correlation analysis method to the 18 indicators and momentum evaluation constructed above (measured by Spearman ranking correlation coefficient). The idea is as follows:

The Spearman correlation coefficient for the variables x and y is actually calculated using the sorted order of the two columns of numbers. Get a sorted sequence by sorting variables in ascending or descending order$r = \{r_1, r_2...r_n\}$,The rank sequence corresponding to variable y is$s = \{s_1, s_2...s_n\}$.Subtract the corresponding elements in the series r and the series s and substitute them into the Spearman rank correlation coefficient formula to obtain the rank difference series$d = \{d_1, d_2..d_n\}$Finally,we get the level coefficient:

$$\rho = 1 - \frac{6\sum {d_i}^2}{n(n^2-1)}$$

### 4.1.1 Correlation Coefficient Heat Map

The picture below describes the relationship index percentage between the above indicators and momentum. Among them, the first serve scoring rate, the winning rate of serving games, the total points rate, and the winning rate of winning receiving and serving games are closely related to "momentum". There is no obvious relationship between indicators such as the number of stalemate shots, score advantage, and success rate in saving break points and other indicators, and the remaining indicators are highly correlated. We found that physical strength is negatively correlated with momentum, indicating that the side with more athletes running during the game will have less momentum.

## 4.2 Random Forest to Analyze

We use random forest to further explain the relationship between indicators and "momentum". First, we use bootstrap sampling to extract 70% from the original training set as training samples, and the remaining 30% as test samples. Without disturbing the data set, cross-validation was used to build 100 decision tree models for each sample. Finally, each record is voted on based on 100 classification results to finally determine the momentum. The following figure shows the simplified process implementation diagram of random forest.

After training, the model evaluation index MSE is 0.002 RMSE is 0.05 MAE is 0.041 MAPE is 9.167 $R^2$ is 0.665, indicating that the evaluation accuracy of the model is high. At the same time, we used the trained random forest model to predict the momentum of the same athlete in other games. The model evaluation index MSE was 0.188, RMSE was 0.434, MAE was 0.236, MAPE was 274.375, $R^2$ was 0.555, and it was found that the predicted value was consistent with the true value. The evaluation momentum is basically consistent. The final feature importance is also closely related to the "momentum" of the first-serve scoring rate, serve winning rate, total scoring rate, and return-of-serve winning rate analyzed by Spearman's correlation mentioned above.

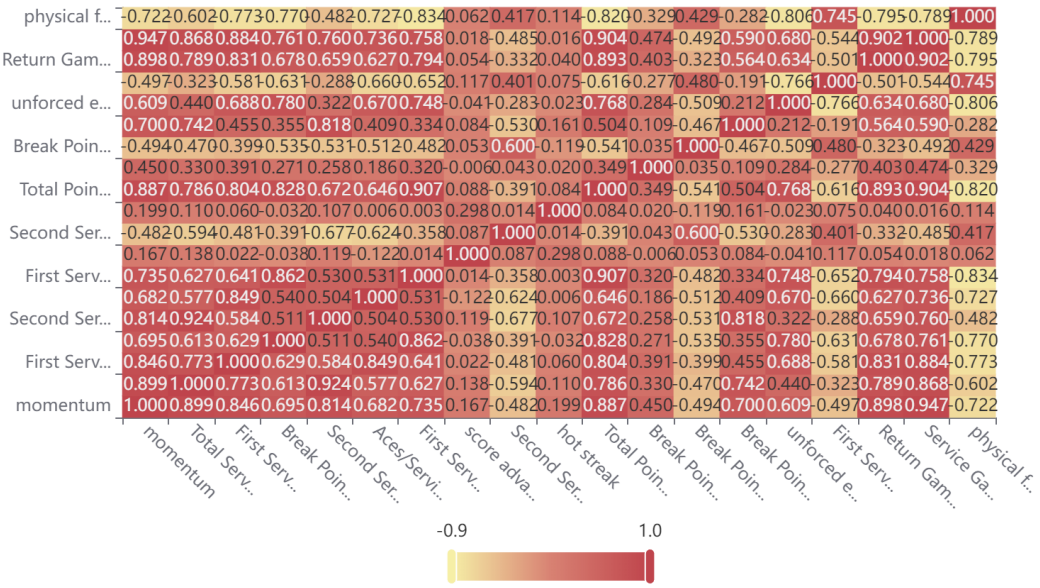



Figure 3: Heat Map

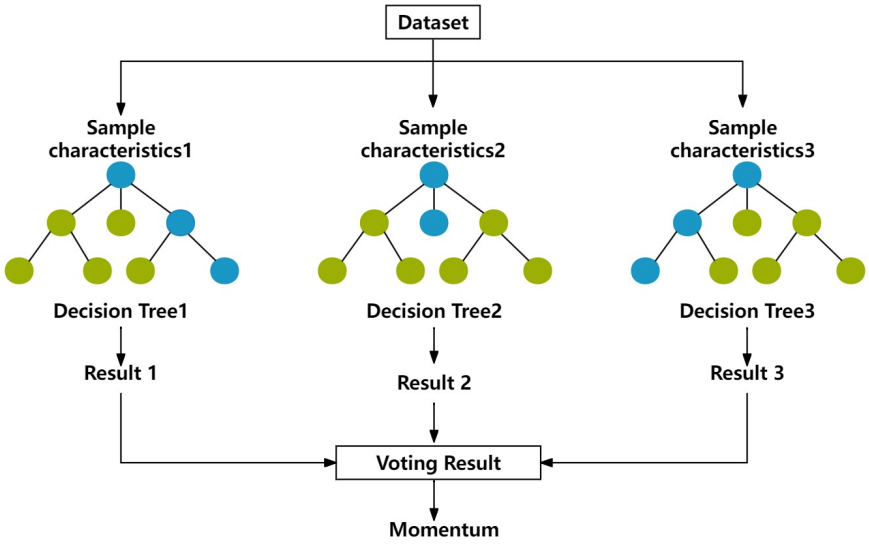



Figure 4: simplified process implementation diagram

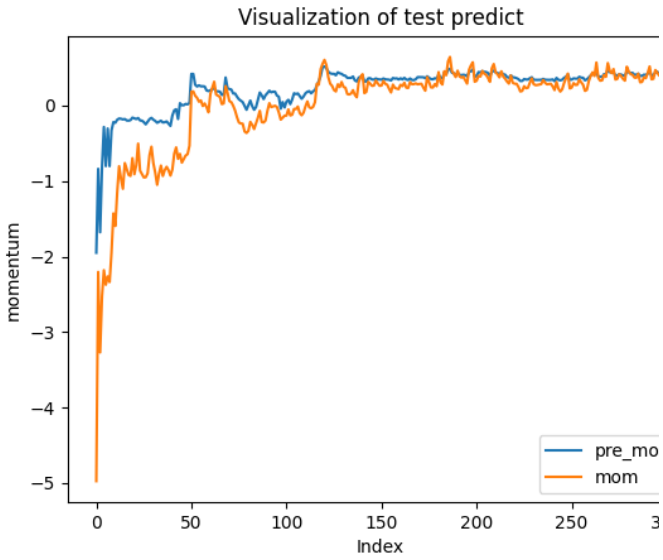



Figure 5: Evaluation of momentum in different games

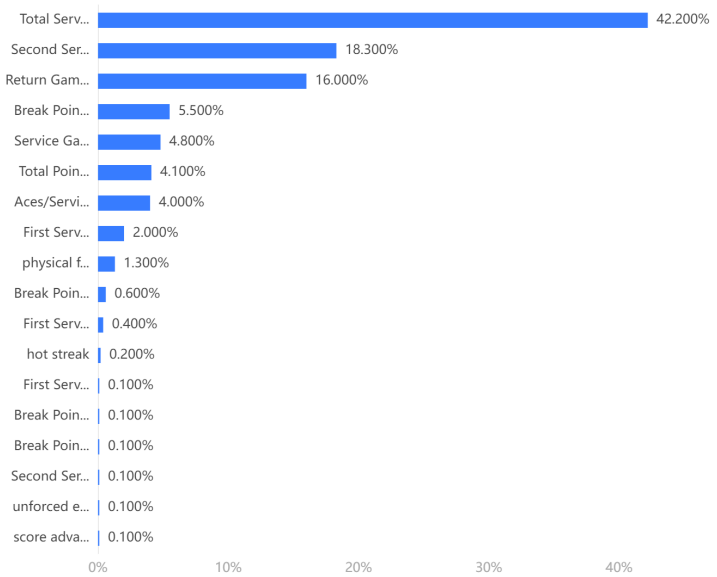



Figure 6: Indicator weight sorting

# 5 LSTM model predicts match fluctuations

## 5.1 Model implementation

The LSTM network can store remote time-dependent information and can perform appropriate mapping between input data and output data, making it powerful in time series data prediction. The LSTM network structure is different from the traditional perceptron architecture. It contains 1 internal state storage unit and 3 gates (input gate, forget gate and output gate) that control the flow of information. Different from other types of artificial neural networks, LSTM combines Long-term and short-term memory ensure the stability and accuracy of the model along the time series.This unique feature can be shown as a repeating unit in Figure.6. Our team uses LSTM neural network to propose a prediction model based on sliding window LSTM network. Here we will briefly describe the principles used in our prediction model.

## 5.2 Long-term and Short-term Memory

- **oblivion door**

  In order to understand the output of long-term memory, we should understand the function of the forget gate. The following is the formula of the forget gate:

  The object of the forgetting gate is the cell state$C_t$,Its function is to control the selective forgetting of information in the cell state,and decide which parts need to be discarded and which parts need to be retained.

- **input gate**

  The target of the input gate is also the cell state. Its function is to decide what new information is to be stored in the cell state, that is, what new memories are selectively added to the cell

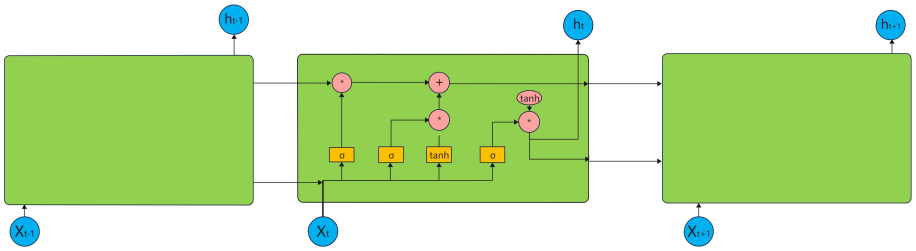
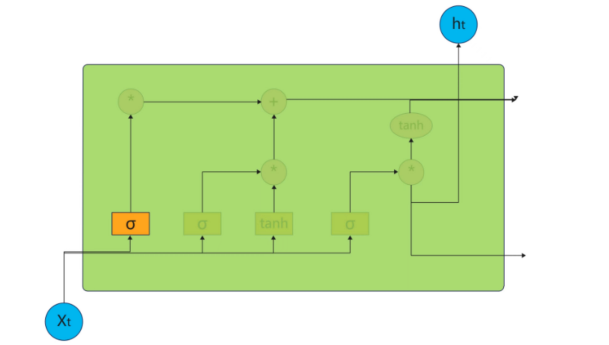


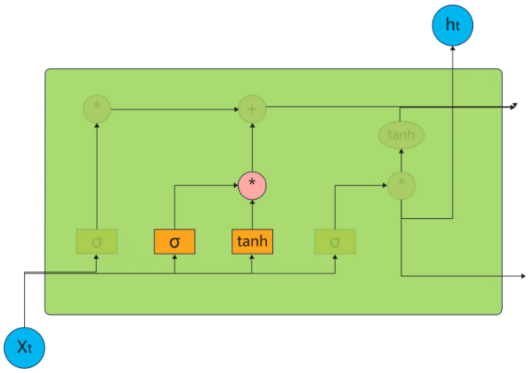

Figure 7: LSTM implementation principle




$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

Figure 8: oblivion door

state. To decide what new information to add, the input gate is performed in two parts. The first part builds an input gate that determines what information is added to the cell state as new memory. Enter the gate formula:




$$i_t = \sigma W_i \cdot [h_{t-1}, x_t] + b_i$$

Figure 9: input gate

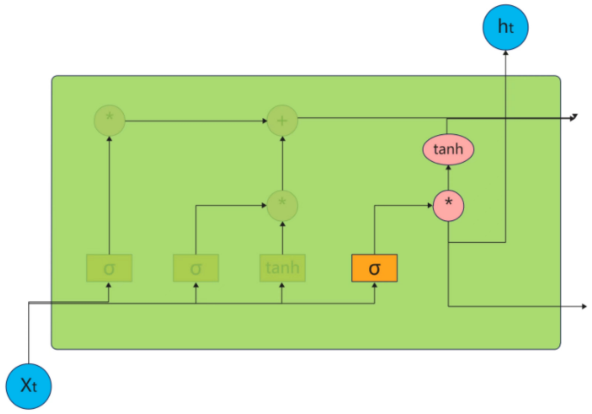
The second part constructs a candidate cell state$\widetilde{C}_t$,it saves $x_t$and$h_{t-1}$mmessage,Then multiply it with the value of $i_t$ to determine which memories are useful. Still,0 in $i_t$ represents complete discarding, and 1 represents complete retention. In this way, the retained informa-

tion is added to the new memory as a new memory. Cell state,so $\widetilde{C}_t$ here is called candidate cell state. It is just one waiting for selection, and only useful ones will be added as new memories. The formula of candidate state cell $\widetilde{C}_t$:

$$\widetilde{C}_t = \tanh(W_c \cdot [h_{t-1} x_t] + b_C)$$

- **output gate**

  The function of the first part of the output gate is to determine the final output, which is $h_t$. This output is based on the new cell state $c_t$ and is also divided into two parts. The first part still uses the sigmoid layer to get $O_t$, such as $O_t$=[0,0,0,1], to determine which part of the cell state needs to be output, and only the required memory is output, not all are output.




$$o_{t=}\sigma(W_o \left[h_{t-1}, x_t\right] + b_o)$$

Figure 10: output gate

The second part processes the new cell state through the tanh function, changes the output value to [-1. 1], and then multiplies $O_t$ to control which part needs to be output.

$$h_t = O_t * \tanh(C_t)$$

## 5.3 LSTM implementation

From this, we can assume that the number of games available is Q, and the length of the sliding window is set to N. The length of the sliding window is the length of the player's momentum input during each training session. The LSTM network outputs a predicted player momentum data for each training, and uses the "remove the old and add the new" mode to iteratively update the player momentum data in the sliding window, that is, use the training output value as the latest data to replace the oldest data in the sliding window. The sliding window LSTM network is trained iteratively for Q - N times, and the parameters are continuously optimized to obtain the player momentum prediction model. When predicting the game fluctuation trend, use the player momentum data from Q-N to Q games to predict the momentum data of Q+1 games in the future, and then use the player momentum data from Q-N+1 to Q+1 to predict Q+2 games. Player momentum data for the game. At the same time, the input data of the LSTM model should be

normalized. Normalized by the Min-Max method, the input data of the LSTM model is finally obtained. Due to the small sample size, we only used two LSTM layers. Each LSTM layer has 100 neurons (units) and the activation function is 'ReLU'. and two Dropout layers. Then, we adjusted the "look back" parameter, and finally found the best look back performance time series to be 3. If it is too small, it will be difficult to evaluate.

## 5.4 Model Predict

After 150 times of training, we get the loss value to converge to 0.14 and the prediction graph is as follows

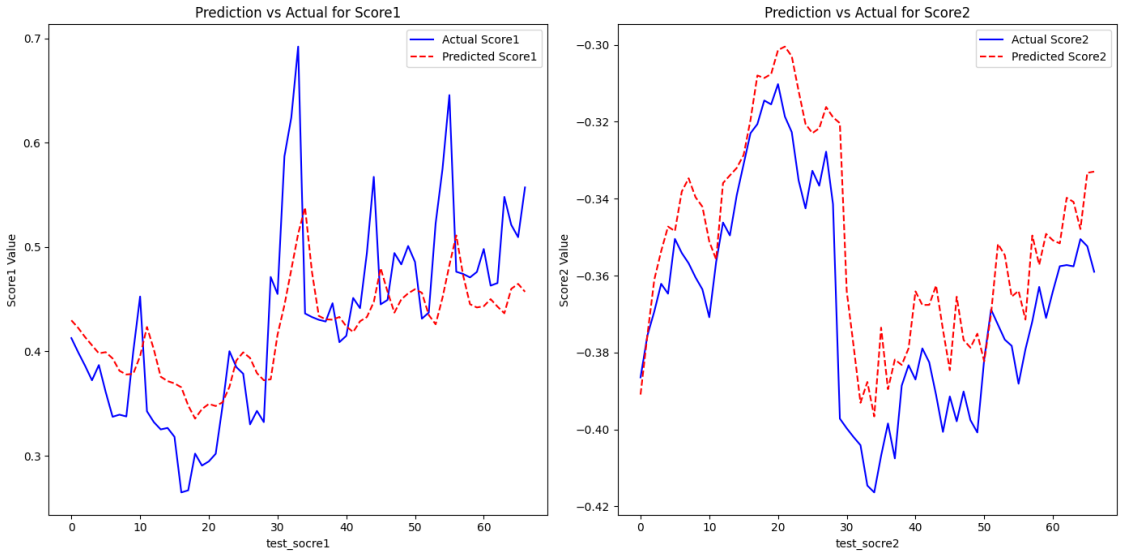



Figure 11: LSTM Predict

## 5.5 Model evaluation

The root mean square error (RMSE) indicator based on momentum information is used to evaluate the stability of the model's trend prediction. The root mean square error represents the standard deviation of the model's predictions. The smaller the evaluation index value, the higher the model prediction accuracy. The root mean square error in this test is shown in the following table:

Table 6: RMSE

| RMSE | Player 1 evaluation results | Player 2 evaluation results |
|---|---|---|
| 3 sliding windows | 0.077148359 | 0.017963428 |
| 5 sliding windows | 0.079469785 | 0.021394204 |

We found that if the athlete's momentum fluctuates greatly, the prediction will be less accurate. Finally, through the momentum prediction of the previous few games, we can bring the following dynamic Markov chain based on logistic regression to predict the win or loss. Finally, we can analyze the difference in momentum between the two sides. Comparison can determine winning or losing very accurately, so we substituted this difference into model training and found that the accuracy of predicting winning or losing can reach 62.3%.

# 6 Dynamic Markov model prediction of results based on logistic regression

According to the tennis scoring rules, a tennis match is divided into points, games and sets. A game is made up of points, a set is made up of different sets, and a match is made up of different sets. This structure is well suited to be modeled using Markov methods. Once we have the probability of two players winning a point on their respective service games, we can recursively calculate the probability of each player winning the game, set, and match. Next, a Markov chain is used to calculate the winning probability of a game based on the probability of each player serving and scoring.

First, let's analyze one game in a set. In a game, if a player scores four or more points first and scores two more points than his opponent, this player can win the game.

## 6.1 Calculate the probability of winning a game

p represents the probability that player A wins B by one point when player A serves,$1-p$ Indicates the probability that player B wins A by one point when player A serves. Remember that when player A serves, the probability that he wins player B with a score of (a, b) is

$$P(a,b) = p \cdot P(a-1,b) + (1-p) \cdot P(a,b-1)$$

if $a = 4$, $b \leq 2$ ,The boundary value is $P(a,b) = 1$;if $b = 4$, $a \leq 2$ ,Boundary value$P(a,b) = 0$.When the score is $40:40$ The problem arises when, in this case, we use the formula proposed for the above bounds to infinite recursion. Nonetheless, the value of P(3,3) can be calculated unambiguously. We can deduce the value of $P(3,3)$. The given conditions are:

if $a = 4, b \leq 2$,Then the boundary value $P(a,b) = 1$. if $b = 4, a \leq 2$,Then the boundary value $P(a,b) = 0$.

According to these boundary conditions, we have $P(5,3) = 1$ and $P(3,5) = 0$.When the score is $40:40$, we use the given formula to calculate $P(3,3)$:

$$P(3,3) = p^2P(5,3) + 2p(1-p)P(4,4) + (1-p)^2P(3,5)$$

Because $P(5,3) = 1$ and $P(3,5) = 0$,and$P(3,3) = P(4,4)$,we can $P(4,4)$ Replace with $P(3,3)$ to solve the equation, we get:

$$P(3,3) = p^2 \cdot 1 + 2p(1-p)P(3,3) + (1-p)^2 \cdot 0$$

$$P(3,3) = p^2 + 2p(1-p)P(3,3)$$

Since we need to solve for $P(3,3)$, let's tidy up the equations:

$$P(3,3) - 2p(1-p)P(3,3) = p^2$$
$$P(3,3)(1-2p(1-p)) = p^2$$
$$P(3,3) = \frac{p^2}{1-2p(1-p)}$$

Similar to p and P , q is the probability that player B wins a point when he serves, and Q is the probability that player B wins a game when he serves. How to estimate the probability (p and q) of a player winning a point when he serves is the biggest difficulty in predicting the result of the Markov model. In the improved model of Markov chain below we use the trend of player momentum to estimate the values of p and q.

## 6.2 Winning rate logistic regression

A fixed state transition matrix is used in traditional Markov chains, resulting in the intermediate probability not changing. Therefore, we propose a method to dynamically calculate the probability of winning when the score changes based on a logistic regression model. The method takes into account player momentum changes and service transitions during a match, providing a more accurate prediction. We use logistic regression models to predict key factors in the game, including players' momentum trends, serving situations, and win-loss results, to predict the impact of subsequent wins. The player's momentum and serving situation are used as dependent variables, and the outcome of the game is used as the independent variable. By taking a weighted sum of player momentum and serve, and applying a logistic function to map the output of the linear regression model between 0 and 1, we can predict the probability of player 1 winning. The formula for predicting the probability of victory by the logistic regression model can be expressed as:

$$P(S)_{i,j} = \frac{1}{1 + \mathrm{e}^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_{server})}}$$

Among them$P(S)_{i,j}$is the probability of a player winning the game given the input feature X, e is the base of the natural logarithm,$\beta_0$ is the intercept, $\beta_1$, $\beta_2$, $\beta_3$ is the parameters of the model corresponding to the characteristic coefficients respectively. $x_1$ (momentum of player 1), $x_2$ (momentum of player 2), $x_3$ (who serves), i is the current score status, j is the next score status, we train on the 2023-wimbledon-1601 data, 80% is used as training set and 20% is used as test set. Through training and testing the model, we obtained the model parameters. The intercept term is: -0.577, and the three characteristic coefficients are: 0.366, -0.177, 1.545, Putting these parameter values into the logistic regression formula, it is the probability of player 1 winning under certain conditions.

## 6.3 Markov for dynamic probabilistic prediction

A Markov chain is a set of discrete random variables that randomly transfers from one state to another in the state space and is memoryless, that is, given the current knowledge or information, the

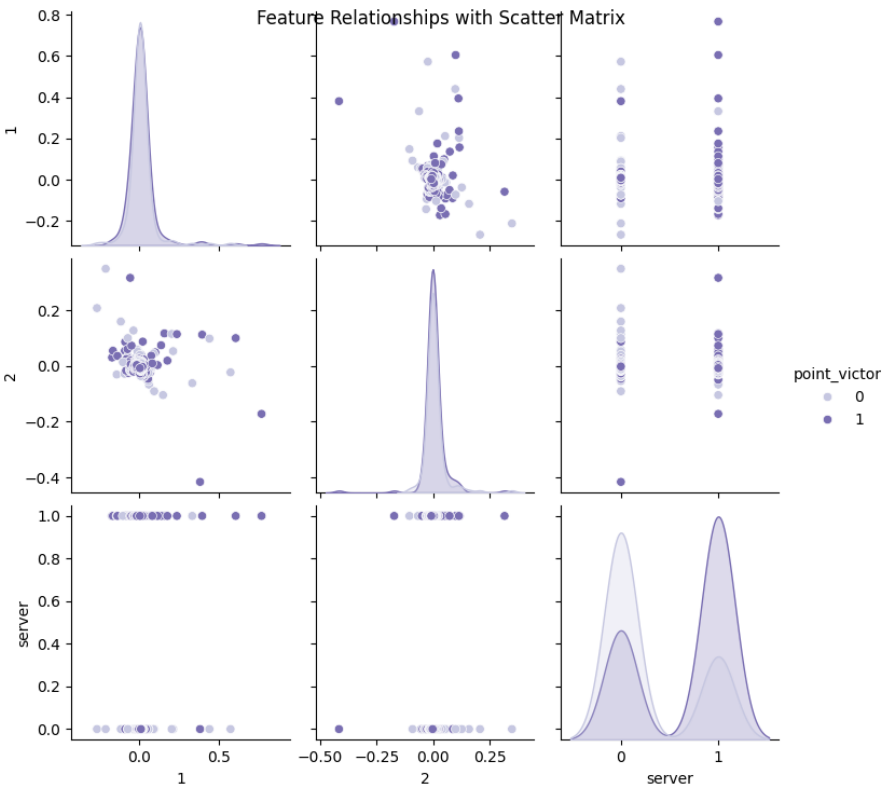



Figure 12: Feature Relationships with scatterplot matrix

next state Only related to the current state, not the past state. 3.2.1 Define status and transformation rules In tennis score prediction based on Markov chain, state definition and transition rules are the core parts of model construction. Each state represents a specific score in the game. For example, 30-40 means that the server has 40 points and the receiver has 30 points.

Transition rules describe the probability of transitioning from one state to another. In the framework of Markov chains, these probabilities only depend on the current state and have nothing to do with past states, which is called the Markov property. For example, the probability of transitioning from state i to state j can be expressed as Pij. These transition probabilities are real-time, and the winning probability when the score changes is dynamically calculated through the probability formula trained in the logistic regression model above.

After defining the state and transition rules, a state transition matrix can be constructed, where each element Pi→j represents the probability of transitioning from state i to state j. This matrix contains all possible score states. We can calculate the probability path from the beginning of the score to any specific score state. By analyzing these paths, we can predict the probability of which player will ultimately win. The specific state transition matrix is shown in the figure below:

The accuracy of the model reached 67.86%, the recall rate was 74.19%, the F1 value was 71.88%, and the AUC value was 0.67. It can be concluded that it is reliable to use the player's momentum and serving situation to predict the results of tennis matches.

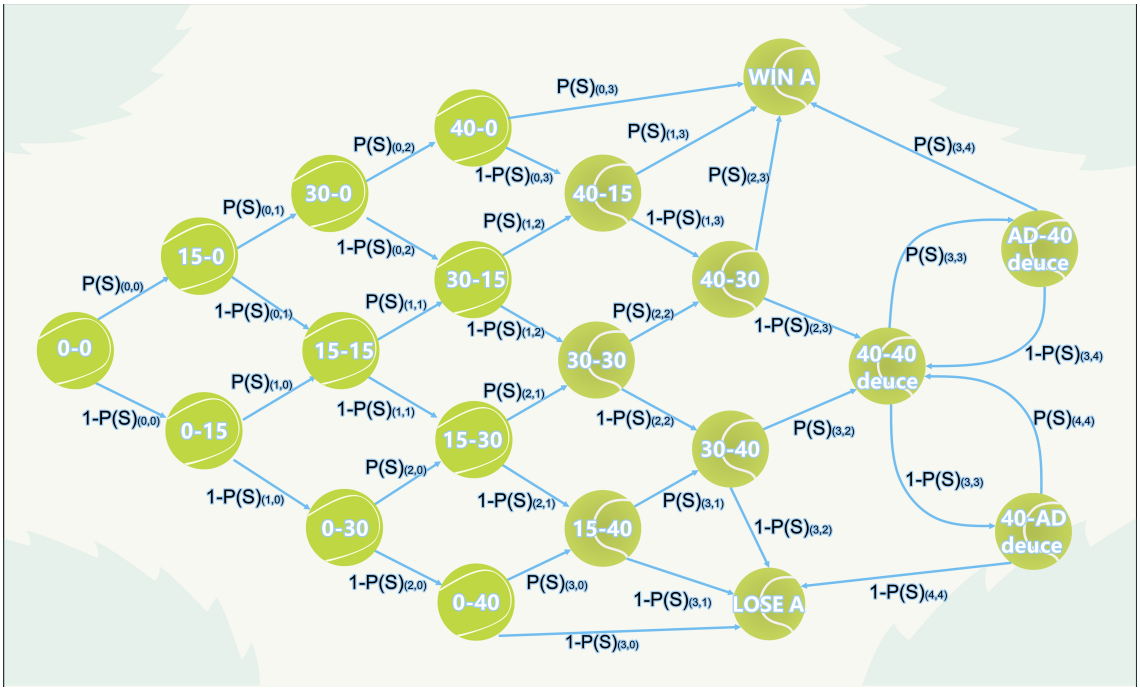



Figure 13: state transition matrix

Applying the improved Markov chain model, we are able to dynamically predict score changes and final results during the game. The model shows higher prediction accuracy than traditional methods, can effectively reflect dynamic changes in the game, and provides valuable information for coaches and players to formulate strategies. interest. The following is the prediction of the second game score of the first set of the 2023-wimbledon-1301 game based on the improved Markov chain model:

| predict_pt | point_victor |
|---|---|
| 2 ✓ | 2 |
| 2 ✓ | 2 |
| 1 ✗ | 2 |
| 1 ✓ | 1 |
| 1 ✓ | 1 |
| 1 ✗ | 2 |

Figure 14: predict score

In a tennis match, the “ebb and flow” of momentum reflects the dynamics of the relative performance of the players on both sides. This change will be affected by competitive indicators and cause fluctuations. Therefore, coping with such changes, especially strategic choices when momentum shifts, becomes the key to winning in the game. Therefore, we give strategic suggestions for players to face another player in a new game based on changes in momentum during the game:

When the momentum drops, that is, when one's side is in a disadvantageous position, the key is to quickly identify the advantages and disadvantages of the competitive potential of both parties. Players need to adjust their technical and tactical structure, such as changing their serving strategies, increasing offense at the net, or adjusting their baseline play, in order to find and exploit their opponent's weaknesses. At the same time, maintain mental stability to avoid increasing unforced errors. When the momentum of both sides stabilizes during the game, the game situation should be carefully analyzed to identify and change the unfavorable factors in the game. At the same time, players should remain on high alert to prevent their opponents from attacking when their momentum is down. When momentum is rising, stay calm to avoid losing momentum. This requires players to predict the opponent's adjustment strategy in time and make corresponding preparations, such as adjusting their position to respond to possible changes in the opponent's serve. In general, the basic strategy in tennis is to strive for momentum, that is, to constantly seek and expand the advantage of the game while avoiding a loss of momentum. When the two sides are in a stalemate, the key is to look for opportunities, break the balance, and achieve a change in momentum. At the same time, players need to observe the changes in the opponent's momentum during the game, find the main factors that affect the opponent's momentum, and tilt the situation in their favor.

# 7 Strengths and weaknesses

**Strengths**

- Factor analysis can reduce the dimensionality of data and extract common structures in features by combining original variables into fewer potential factors. By extracting the main latent factors, factor analysis can be used to simplify the data set, and the factor analysis results are more interpretable, making it easier for researchers to understand and explain "momentum".
- Random forests can output the importance of features and help understand which features "momentum" is more sensitive to. Random forest has good adaptability to multi-category classification problems and can effectively handle category imbalance.
- LSTM is particularly good at processing time series and time data. It can capture momentum and predict, effectively solving the problems of gradient disappearance and gradient explosion in RNN.

**Weaknesses**

- Factor analysis is more sensitive to outliers, which may have a greater impact on the extracted potential factors.
- Although random forest can provide feature importance, random forest as a whole is still a relatively black box model, and it is not easy to explain each specific decision-making process.
- LSTM requires a large amount of data to train, and may cause overfitting. The network structure is relatively complex, and it is difficult to explain the specific role of each neuron under the influence of different features, making it a relatively black box model.

# 8 Conclusion

We use data analysis and modeling methods to study momentum changes and fluctuation trends in games. First, factor analysis is used to quantify the momentum, and a dynamic Markov model and an LSTM model are proposed to predict the fluctuation and probability of winning or losing the game. Our team also found correlations between match momentum and multiple metrics and used a random forest model to identify the most relevant factors. Finally, the model is applied to different games and suggestions are made to help players understand the role of momentum in the game and deal with the factors that affect the game.

# References

[1] Qi Junfeng. Research on power generation power classification prediction method based on Markov chain [J]. Popular Electricity, 2023, 38(10):54-55.

[2] Wang Yukuan, Xie Xinlian, Ma Hao, et al. Ship track prediction based on sliding window LSTM network [J]. Journal of Shanghai Maritime University, 2022, 43(01): 14-22. DOI: 10.13340/j.jsmu.2022.01.003.

[3] Fang Kuangnan, Wu Jianbin, Zhu Jianping. A review of research on random forest methods [J]. Statistics and Information Forum, 2011, 26(03): 32-38.

[4] Jiang Qifei, Zheng He. Construction and analysis of regression prediction model for men's tennis singles skills and comprehensive strength [J]. Jilin Institute of Physical Education Journal, 2015, 31(02):39-43.DOI:10.13720/j.cnki.22-1286.2015.02.009.

[5] Newton P K, Aslam K. Monte Carlo Tennis: A Stochastic Markov Chain Model[J]. Journal of Quantitative Analysis in Sports, 2009, 5(3): 1-44.

# Appendices

## Appendix A Memo

Dear, coach

In the 2023 Wimbledon Gentlemen's final, Carlos Alcaraz, a 20-year-old Spanish rising star, defeated Novak Djokovic, ending Djokovic's remarkable winning streak at Wimbledon since 2013. The match was a remarkable battle with significant momentum shifts. Djokovic dominated the first set 6-1 but lost the second set in a tense tie-breaker, won by Alcaraz. The third set saw Alcaraz winning convincingly 6-1, but Djokovic took control in the fourth set, winning 6-3. However, Alcaraz regained control in the final set and secured a 6-4 victory. The incredible swings in momentum throughout the match are often attributed to this phenomenon, which is difficult to measure and understand fully. Therefore, our team combined data from multiple tennis matches to extract characteristic indicators for factor analysis, analyzing the momentum of players and the fluctuation trends of tennis matches, and further investigated, proposing the following models:

- Factor analysis was used to reduce the dimensions of multiple data indicators, transforming them into comprehensive indicators, while calculating factor composite scores to assess "momentum" and player performance.
- A dynamic Markov model based on logistic regression was used to calculate the winning and losing probabilities of players.
- An LSTM model based on a sliding window was used for predicting the fluctuation of the match.

Firstly, based on the definition of momentum, we believe that it is most important to quantify momentum to facilitate subsequent analysis progress. "Momentum" is influenced by a variety of factors. Initially, eighteen indicators were extracted for factor analysis after reviewing the literature and consulting materials, showing significant correlations among the eighteen factors. They interact with each other, collectively affecting the trend of the match. Through factor analysis, these eighteen measured variables were transformed into three comprehensive indicators: the serving impact factor, receiving impact factor, and scoring impact factor, with the serving impact factor contributing the most at 42.702%.

Furthermore, we discovered that "momentum" in matches is not random. We conducted a correlation analysis of the aforementioned indicators, measuring their relationships with a rank correlation coefficient. By drawing a heatmap, we found that there are positive or negative relationships among the indicators.

To predict the fluctuation trend of the match and the role of momentum, we used an LSTM model based on a sliding window. Setting the window size to 3 with the momentum of both players in each

match as the input, we predicted the fourth game using the previous three games, incorporating the fourth game into the old values and continuing to predict the fifth game from games 2-4, and so on, thus predicting the fluctuation trend of the match. The root mean square error of prediction was 0.01796, indicating high accuracy.

Traditional dynamic Markov chains use a fixed state transition matrix, where the intermediate probability does not change. To consider the impact of "momentum" changes and service rights on the probability of winning or losing, we used a logistic regression model to train the probability formula, dynamically calculating the probability when the score changes. The model was trained using a split training and testing set, and overall, the predicted values were close to the actual values with a small error rate, giving us confidence in our model.

To determine which factors are most related to the fluctuation of the match, we used a random forest for prediction, finding that match fluctuation is most related.

Finally, we applied our model to different matches to predict fluctuation trends, considering the impact of the venue and weather on the match when the model performed poorly, concluding that our model has high generalizability and transferability.

Based on this research, we hope to offer the following suggestions for players entering a new match, implemented through the on-the-spot command of the coach and the self-adjustment of the athlete:

1. The on-the-spot command of the coach involves using timeouts, substitutions, and off-court guidance to strategically and tactically direct the team, aiming for victory. During tennis matches, coaches can use gestures or expressions to signal players to stabilize their "momentum."

2. During the beginning or critical moments of a match, non-forced errors can lead to over-control and excessive effort, hindering automatic action execution and ultimately lowering "momentum." Coaches and players are advised to maintain composure during matches.

3. When the match reaches a stage where both sides are in a good competitive state and have adapted to each other's techniques and adjustments, disrupting the opponent's rhythm to lower their "momentum" can reduce the effectiveness of their techniques and tactics.

We hope our model and these suggestions will help you understand the role of "momentum" in matches and assist players in dealing with factors that affect tennis matches. Wishing you success in the future.

Sincerely yours,

Team 2429507