

Tennis Momentum Fluctration Model Based on Time Series Bradley-Terry Model

Summary

The momentum received more and more attention, but it is hard to study due to its intangible nature. However, we not only overcome this problem by connecting momentum with some meaningful features that refer to a generalizable performance score model but also predict the fluctuations and outcome of the match through momentum.

Firstly, after data pre-processing and extracting further features from raw data, we employ the dynamic Bradley-Terry model with time series to score the players' performance which is related to the specific further features. By processing the model, we not only calculate the performance score of players with a set of final dynamic parameters:[0.47150762, 0, 0, 0, 6.06662347, -6.06591313] but also generalize the model to the women's matches. Simultaneously, we also acquire several dominant features which have strong relevance to the match flow. When these features vary, the performance score has a remarkable change.

In the following stage, the method of computing momentum is proposed with better features. Using the data from this method, we successfully prove the correlation between momentum and success, winning rate, and the swings in matches through the Logistic Regression Algorithm to refute the random idea of some coaches. In addition, we apply GBDT to plot the decision tree model which reflects the correlation of features and the winning percentage. Through the above models, we not only calculate the Point-Biserial Correlation Coefficient(0.87838) and P-value(0) which embody high correlation but also gain the correlation heatmap between features.

Further, we need to study the most related feature to fluctuations of the matches with the XGBoost model and predict the fluctuations in the matches with the SARIMAX model. By these models, we not only contrast the relation between the volatility and each feature to obtain the most related feature(p1_win_streak) but also predict fluctuations accurately which is seen in the figure in the model result of the SARIMAX model.

Eventually, we give suggestions for the serve and return strategy to a player through the data statistics. Through analyzing the direction of the serve, the depth of the serve, and the depth of the return, we understand how the momentum gap is influenced by these three features which are not further features in use. The specific suggestions are given in the suggestion of the player strategy.

Overall, this paper has dealt with the momentum definition, analysis, and prediction which could be generalized to other single matches for predicting the result of the matches and giving strategy. In the meantime, this paper also introduces a performance scoring model which could be applied to the athletes ranking.

Keywords: Momentum; Dynamic Bradley-Terry Model; SARIMA Model; XGBoost Model; Prediction; Correlation

Contents

1	Introduction	3
1.1	Background and Restatement of the Problem	3
1.2	Literature Review	4
1.3	Our Work	4
2	Assumptions and Justifications	5
3	Notations and Glossary	5
3.1	Notations	5
3.2	Glossaries	6
4	Data Description	7
4.1	Data Pre-processing	7
4.2	Data Visualization	9
4.3	Description Statistical Analysis of the Data	9
5	Player Performance: Dynamic Bradley-Terry Model with Time Series Prediction	10
5.1	Model Establishment	10
5.2	Model Process	11
5.3	Result of Model	12
5.4	Model Generalization	13
6	Player Momentum: Momentum Estimation with Two Models	13
6.1	Momentum Calculation Model	13
6.1.1	Model Establishment	13
6.1.2	Result of Model	15
6.1.3	Correlation Analysis	15
6.2	Momentum Fitted by SARIMAX Model	15
6.2.1	Model Establishment	15
6.2.2	Model Results	16
6.2.3	Correlation Features Analysis	17
7	Player Strategy: Strategy Choosing Based Statistic Analysis	17
7.1	Data Statistic Analysis	17
7.2	Suggestions	18
8	Sensitivity Analysis	19
8.1	Player Performance Model	19
8.2	Player Momentum Model	19
9	Model Evaluation	20
9.1	Strengths	20
9.2	Weaknesses	20
10	Conclusion	21

1 Introduction

1.1 Background and Restatement of the Problem

During the 2023 Wimbledon Gentlemen's final, Novak Djokovic's winning run at Wimbledon was over.[1] The live situation is shown in Figure 1. But in many sports, some spontaneous things push athletes to gain an easy victory. More specifically, those athletes with a clear advantage are more likely than opponents to win with no difficulty. This is what people call momentum. Adler (1981) has defined psychological momentum as a bidirectional concept, affecting either the probability of winning or the probability of losing as a function of the outcome of the preceding event.[2] Obviously, the momentum that is associated with athletes' psychological pressure and their actual performance is vital to sports matches. However, it is challenging to verify the impacts of momentum in theory due to the intangible nature of mental activity. To overcome this challenge, we usually need to employ an appropriate model to study the relation between momentum, performance, and the result of the match. According to the background described



Figure 1: Live Capture Novak Djokovic [3]

above and specific constraints in the problem statement, the problem could be restated as follows.

- **Establish a performance assessment model for players and visualize results.** To evaluate players' performance, we need to find related features to build the model based on given data. Then, the result of the model will be visualized for understanding directly. In the following step, generalizing this model to other matches is necessary.
- **Construct a correlation model between momentum and features.** In order to emphasize the importance of momentum, we need to establish a correlation model. Utilizing this model we will prove the certainty of the success and swings if momentum has a high correlation with features that decide the winning percentage.
- **Build a model to predict the exchange of advantage because of the fluctuation of momentum.** Predicting the change of advantageous demands to construct a momentum model that could calculate momentum. After that, we need to analyze the relevance between the change and features to identify the most correlated feature.
- **Strategy analysis of athletes' competition based on the momentum model.** Through the model results above, the best strategy will be offered to athletes.

1.2 Literature Review

Various models are utilized by researchers to pursue more accurate and reasonable predictions and assessments.

Through analyzing the data they found, Almarashi, Abdullah M, and Daniyal [4] discovered an extremely strong correlation between winning percentage and game dominance which has a similar meaning in psychology. In addition, they took time into account and emphasized the importance of time. Ben Moss and Peter O'Donoghue also found that momentum will influence players[5]. Both of their studies show that momentum could decide the outcome of the match to a certain extent, which is same with our idea.

Aiming to score players' performance, Gorgi proposed a dynamic model based on the Bradley-Terry model to express players' strengths[6]. Meanwhile, Victor K. R. Matsudo used the six tiered competitive sports model to assess physique and performance[7]. All of these studies provided a stable foundation for our research undoubtedly.

1.3 Our Work

According to the restatement of the problem, the tasks we need to resolve include data cleaning, extraction of further features, four models, and visualization. The general approach framework is plotted in Figure 2.

First of all, we clean the data to ensure that there is no disadvantages to the model computing. After feature crosses, we begin with visualization which includes an instrument panel describing the data characteristics.

Then, the dynamic Bradley-Terry model with a time series is built for scoring players' performance. In this part, we not only formulate the model equation but also analyze the result. Furthermore, this model generalizes to the women's competition data which seems to perform better.

Thirdly, for the sake of researching the importance of momentum, we adopted a multiplicative model to quantify momentum, which is followed by the correlation analysis between momentum and victory through Logistic Regression Algorithm.

After that, the fluctuation of the flow of the game is predicted by the seasonal autoregressive integrated moving average with exogenous factors (SARIMAX) model. Based on the data generated by the model above, the extreme gradient boosting (XGBoost) model is employed to study the most related features to the swings.

In the following stage, we give players some persuasive suggestions by analyze data in last three columns.

Finally, we optimize the model for identifying the influence of new factors on results, as well as to analyze the sensitivity, error, advantages, and disadvantages of the model.

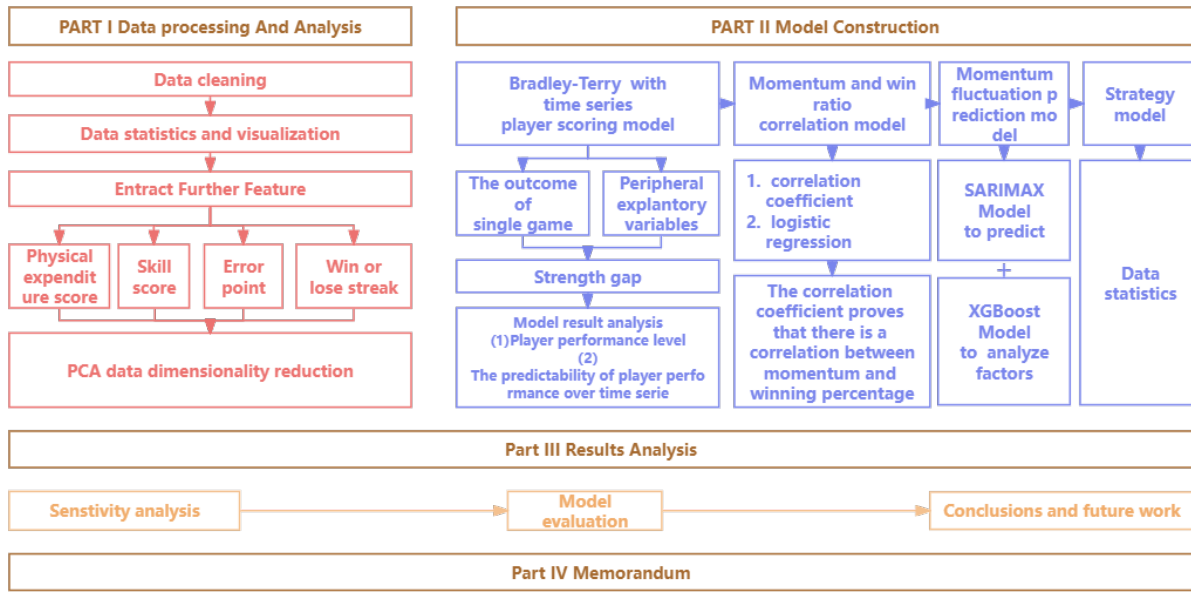


Figure 2: Our Solution

2 Assumptions and Justifications

According to analysis of the problem and model requirement, we proposed assumptions as follows:

- **Momentum and player strength is primarily influenced by the match situation, but not by external factors.** Since the dominant research target is momentum in a specific match, other factors are allowed to be neglected. Meanwhile, this behavior could also simplify our model.
- **Initial Strength in each match is zero.** The initial strength of each player should be given in need of the first model. According to the first assumption, we do not need to consider external circumstances. Hence, the initial strength is set as zero.
- **The Physical Could be Totally Recover in the Break Time.** Generally speaking, a 90-second break is entirely enough to recover the athlete's physical power.
- **The Time Series are not Influenced by Other Random Parameters.** To simplify the Model, we ignore other random parameters' influence on the time series model.
- **Momentum consisted of various factors that influenced players in a short time.** For decrease the weight of the variables long before, we remove these variables influence.

3 Notations and Glossary

3.1 Notations

Vital notations used in this paper are listed in Table 1.

Table 1: Notations Sheet

Notation	Definition
$x^{(i)}$	Row i in the Data Table
k	the Number of Feature Dimensions after the Reduction of Dimension
$x_{RD}^{(i)}$	Row i in the Data Table after the Reduction of Dimension
m	Data Rows Number
V_c	Cumulative Explanation Variance
a, b	The Bib Number of the player
$c_{ab}^{(i)}$	Whether Player a Win the Point of Row i during the Match of Player a and Player b
$\alpha_a^{(i)}, \alpha_b^{(i)}$	The Strength of Player a or b
$\delta_{ab}^{(i)}$	The Strength Gap between Player a and Player b
$p_{ab}^{(i)}$	The Probability of Player a Win the Match of Player a and Player b
$s_a^{(i)}, s_b^{(i)}$	The Impact Factor of Row i Perfomance on The Strength of Player a or b
τ	The Constant Varied only with Different Match
$S_a^{(i)}, S_b^{(i)}$	The Performance Score of Player a or b in the Point of Row i
φ	Dynamic Parameter Vector of Explanatory Variable
$M_a^{(i)}, M_b^{(i)}$	The Momentum score of player a or b in the row i
$W^{(i)}$	The Winner Number in the Row i
P	Percentafe of serve winning set as 1.2
$w_a^{(i)}, w_b^{(i)}$	Winning Score of Player a or b
$l_a^{(i)}, l_b^{(i)}$	Losing Score of Player a or b
$C_a^{(i)}, C_b^{(i)}$	Streak of Player a or b
γ	Parameter of Streak

3.2 Glossaries

Important glossaries used in this paper are explained as follows:

- **P1 Skill Score:** The *P1 Skill Score* is described as skill strength of Player 1.
- **P2 Skill Score:** The *P2 Skill Score* expresses player 2's skill level.
- **P1 Error Score:** The *P1 Error Score* is *p1_unf_err* in raw data table, which means failing

to serve.

- **P2 Error Score:** The *P2 Error Score* is *p2_unf_err* in raw data table, which indicates lose points surprisingly.
- **Point Victor:** The *Point Victor* is described as *point_victor* in raw data table.
- **P1 Win Streak:** The *P1 Win Streak* reflects the winning streak of Player 1 which including positive and negative number.
- **P2 Win Streak:** The *P2 Win Streak* means Player 2's winning streak.
- **Elapsed Time Second:** The *Elapsed Time Second* is the result of the unit conversion of *elapsed_time*, which turns unit into second.
- **P1 Physical Score:** The *P1PhysicalScore* signifies the degree of physical exertion of Player 1.
- **P2 Physical Score:** The *P2PhysicalScore* symbolizes how exhausted Player 2 feels.

4 Data Description

Out of convenience, we use the data that has been supplied by the problem.

4.1 Data Pre-processing

Due to the characteristics of the raw data, we carry out preprocessing data. Out of the convenience of analysis, this process is divided into two steps: data cleaning and feature crosses. The flow of data cleaning is shown in Figure 3. Data preprocessing could be divided into three steps: Remove useless data, Missing value processing, Outlier processing, and Coding. When processing missing values, we mainly use median and mode to fill separately *speed_mph* column and other columns.

After cleaning the raw data, we use Principal Component Analysis(PCA) to find the best feature number. Based on Equation 1, we could plot a line chart of V_c versus k which is shown in Figure 4.

$$V_c = \frac{\sum_{i=1}^m \|x_{RD}^{(i)}\|^2}{\sum_{i=1}^m \|x^{(i)}\|^2} \quad (1)$$

According to Figure 4, five features are basically enough to express all features in the data table. But for the sake of reliability, we still utilize six further features in glossaries generated by primary features to predict and evaluate. These further features include P1 Skill Score, P2 Skill Score, P1 Error Score, P2 Error Score, Point Victor, P1 Win Streak, P2 Win Streak. Based on the result of PCA, we could calculate those further features through the primary features' weight. The formulas of them are displaced as Equation 2,3,4,5, 6,7.

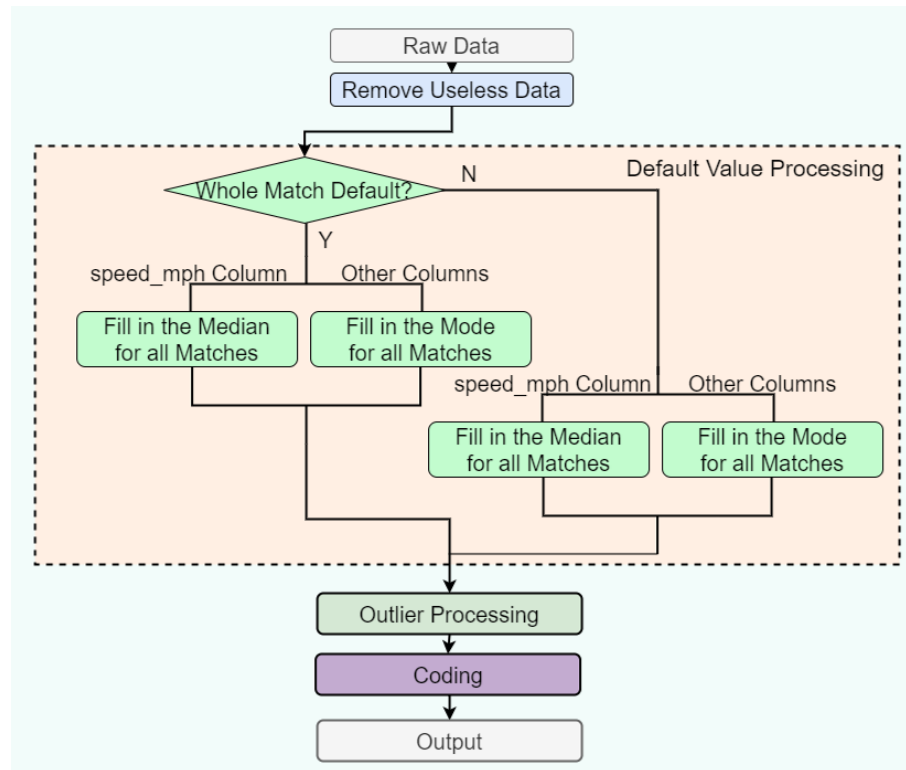


Figure 3: Data Cleaning Flow-chart

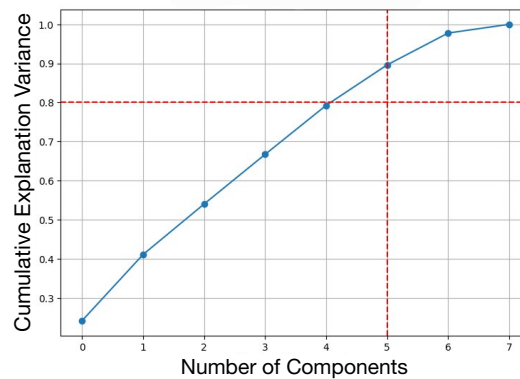


Figure 4: Explained Variance by PCA Components. Based on the Principal Component Analysis result, the explanatory of different number of components is depicted.

$$p1_skill_score = 0.49183874 \cdot p1_ace + 0.3055706 \cdot p1_break_pt_won + 0.12479256 \cdot p1_winner + 0.0777981 \cdot p1_net_pt_won \quad (2)$$

$$p2_skill_score = 0.49183874 \cdot p2_ace + 0.3055706 \cdot p2_break_pt_won + 0.12479256 \cdot p2_winner + 0.0777981 \cdot p2_net_pt_won \quad (3)$$

$$p1_error_score = p1_unf_err \quad (4)$$

$$p2_error_score = p2_unf_err \quad (5)$$

$$p1_win_streak^{(i)} = \begin{cases} 0 & i = 1 \\ p1_win_streak^{(i-1)} + 1 & 1 < i \leq m, point_victor = 1, \\ & p1_win_streak^{(i-1)} \geq 0 \\ -1 & 1 < i \leq m, point_victor = 2, \\ & p1_win_streak^{(i-1)} \geq 0 \\ p1_win_streak^{(i-1)} - 1 & 1 < i \leq m, point_victor = 2, \\ & p1_win_streak^{(i-1)} \leq 0 \\ 1 & 1 < i \leq m, point_victor = 1, \\ & p1_win_streak^{(i-1)} \leq 0 \end{cases} \quad (6)$$

$$p2_win_streak^{(i)} = \begin{cases} 0 & i = 1 \\ p2_win_streak^{(i-1)} + 1 & 1 < i \leq m, point_victor = 2, \\ & p2_win_streak^{(i-1)} \geq 0 \\ -1 & 1 < i \leq m, point_victor = 1, \\ & p2_win_streak^{(i-1)} \geq 0 \\ p2_win_streak^{(i-1)} - 1 & 1 < i \leq m, point_victor = 1, \\ & p2_win_streak^{(i-1)} \leq 0 \\ 1 & 1 < i \leq m, point_victor = 2, \\ & p2_win_streak^{(i-1)} \leq 0 \end{cases} \quad (7)$$

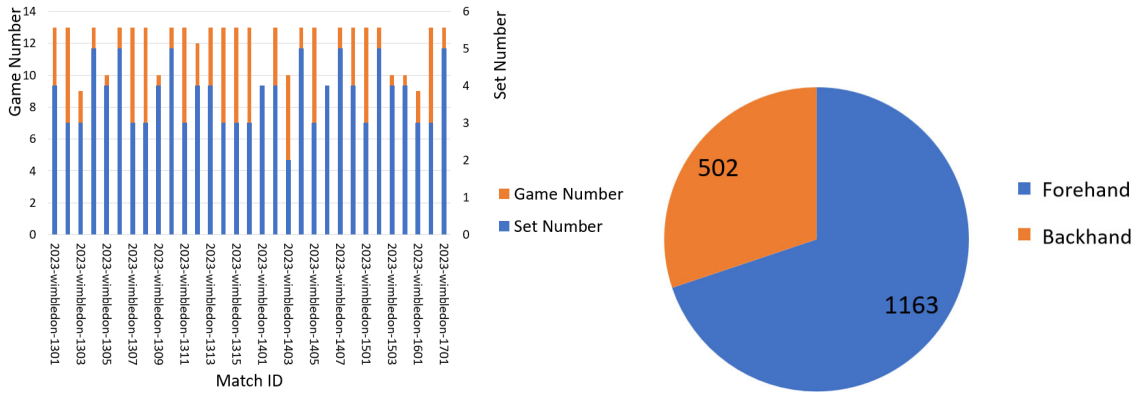
In addition, we still reserve some raw columns to identify different rows, match sessions, and results of the match.

4.2 Data Visualization

Based on the raw data given by the problem, we protract not only the maximum set number and the maximum game number in Figure 5(a) but also the pie chart of forehand hits and backhand hits in Figure 5(b) to analyze the overall features of the data.

4.3 Description Statistical Analysis of the Data

According to the raw data table, the set's score was always three to two and three to one. Meanwhile, nearly one-third of recorded winnings are backhand shots, which means backhand has no disadvantage for winning partly. Specifically, in the matches between Carlos Alcaraz and Novak Djokovic, Djokovic hit more aces than Alcaraz. By contrast, Alcaraz was better at clenching great opportunities. Compared to his ability to grasp chance in the common course of play, Alcaraz struggled with his serve.



(a) The Maximum Set Number and Game Number

(b) Pie Chart of Forehand and Backhand

Figure 5: Raw Data Statistics

5 Player Performance: Dynamic Bradley-Terry Model with Time Series Prediction

5.1 Model Establishment

In the literature review section, we mention the dynamic Bradley-Terry model. Compared to the basic Bradley-Terry model, this innovative model includes a dynamic factor s in the evaluation algorithm, which enhances the precision of the assessment process. By contrast, this dynamic concept, declared as the time varying strength, is not sufficient to explain the relevance between strength and performance. To address this, we strengthen this relevance by incorporating the feature of the *winning streak*. In the following section, the equation of model building will be instructed.

Firstly, we set initial vector variables c_{12} and $\alpha_1^{(1)}, \alpha_2^{(1)}$ as Equation 8, when a new match starts.

$$\begin{aligned} c_{12} &= 2 - point_vector \\ \alpha_1^{(1)} &= 0, \alpha_2^{(1)} = 0 \end{aligned} \quad (8)$$

Then, we could calculate the possibility of the winning of the players with the strength gap $\delta_{12}^{(i)}$ based on the Bradley-Terry model's Equation 9.

$$\begin{aligned} \delta_{12}^{(i)} &= \alpha_1^{(i)} - \alpha_2^{(i)} \\ p_{12}^{(i)} &= \frac{\exp(\delta_{12}^{(i)})}{1 + \exp(\delta_{12}^{(i)})} \end{aligned} \quad (9)$$

In the following place, the dynamic parameter $s_a^{(i)}$ in the dynamic Bradley-Terry model of Gorgi could be reckoned with Equation 10, which originated from the score-driven time varying

parameter model [8, 9].

$$\begin{aligned} s_1^{(i)} &= c_{12}^{(i)} \cdot (1 - p_{12}^{(i)}) - (1 - c_{12}^{(i)}) \cdot p_{12}^{(i)} \\ s_1^{(i)} + s_2^{(i)} &= 0 \end{aligned} \quad (10)$$

Next, a random walk process of the time series model is introduced to figure up the next row strength based on Equation 11, which sets $\tau \cdot s_a^{(i)}$ as a step. Eventually, we not only structure a function($g(x^{(i)})$) declared in Equation 14 to describe the relation between the performance score and some features but also compute the performance score through Equation 12.

$$\alpha_b^{(i+1)} = \alpha_b^{(i)} + \tau \cdot s_b^{(i)} \quad (11)$$

$$S_b^{(i)} = \alpha_b^{(i)} + g(x^{(i)}) \quad (12)$$

$$\varphi = [\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6] \quad (13)$$

$$g(x^{(i)}) = \langle \varphi \mid x^{(i)} \rangle \quad (14)$$

5.2 Model Process

* Initialization Stage

This stage mainly sets initial strength and calculates $c_{12}^{(i)}$ as Equation 8 based on features that are extracted in the data pre-processing section.

* Processing Stage

This stage runs the whole model as the structure block in Figure 6.

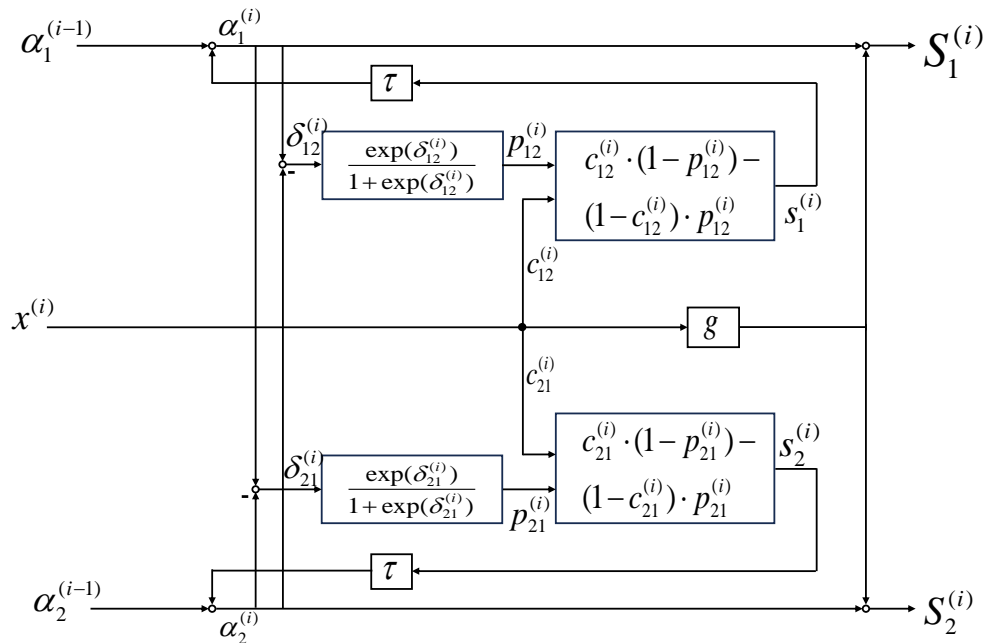


Figure 6: The Structure Block of Score Model

5.3 Result of Model

According to the performance data of players computed by the model, we could plot the line chart of performance score to each point as Figure 7(a). Compared to the won player depicted in Figure 7(b), the results are generally consistent with the variation in performance scores.

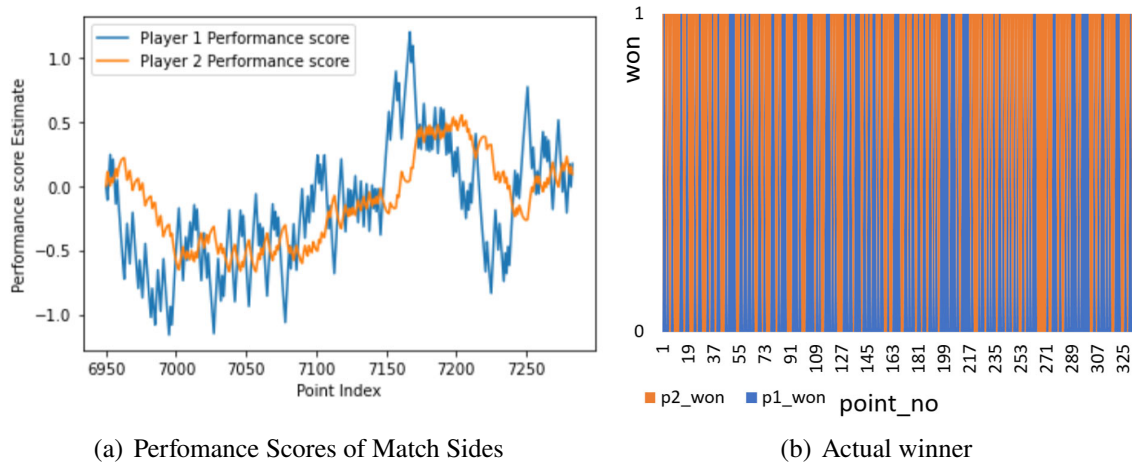


Figure 7: Contrastive Chart between Performance Scores and Winner in Match 1701

The final parameter vector is $[0.47150762, 0, 0, 0, 6.06662347, -6.06591313]$. From the Figure 7, player 1's performance score is less than player 2's performance score when player 2 wins points, vice versa. Thus, the performance of this model could score players' performance reasonably. At the same time, we use this model to generate the data of one set to visualize the statistics and basic situation in Figure 8.

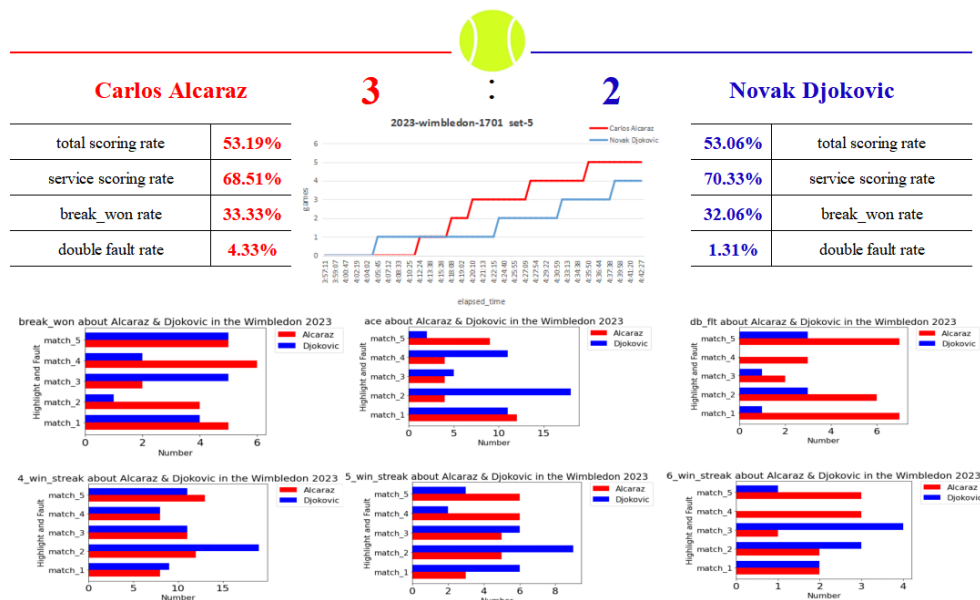


Figure 8: The Instrument Panel of Final Set

5.4 Model Generalization

In order to measure its performance on other data sets, we use the data from women's tennis matches [10]. After pre-processing data, we score the performance of players with the model, which can be drawn in Figure 9. From Figure 9, it can be observed that a strong

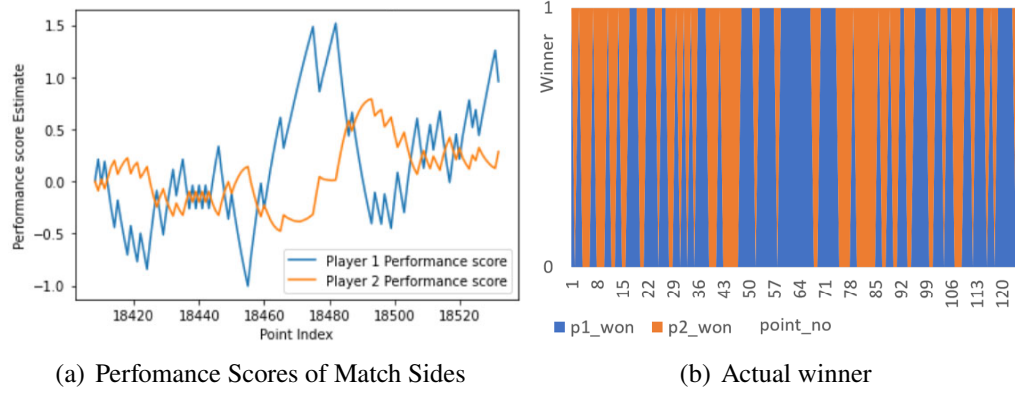


Figure 9: Contrastive Chart between Performance Scores and Winner in 2023 Women WC Match 2701

relation between the scores of the opposite and the match results exists. And the final parameters of the model which generate from the same initial parameter with man's matches is $[1.20848718, 0, 0, 0, -5.96619655, 5.96542469]$. Furthermore, the performance score essentially corresponds to the winning trend with a slight delay which conforms to basic logic.

6 Player Momentum: Momentum Estimation with Two Models

6.1 Momentum Calculation Model

6.1.1 Model Establishment

To respond to some coaches' doubts about the connection between momentum, successes, and swings, we couldn't utilize win status as any parameter, which means to exclude any victor features. However, the momentum still need to quantized.

For this reason, we introduce the Gradient Boosted Decision Tree model and plot a heatmap to choosing better features. In order to find the most fittable features, we plot not only the decision tree model in Figure 10 but also the heatmap in Figure 11. From these figures, we could find that the most relevant features is the $p2_skill_score$ and five features in Figure 11 has the orthogonality which means great features.

As a result, we build a formula shown in Equation 15 to calculate momentum. If the momentum computed by this method is still highly related to the successes and swings, the possible explanation will be that the swings and success are not random but correlate to momentum.

$$M^{(i)} = W^{(i-1)} \cdot P + w^{(i)} + l^{(i)} + \gamma \cdot C^{(i)} \quad (15)$$

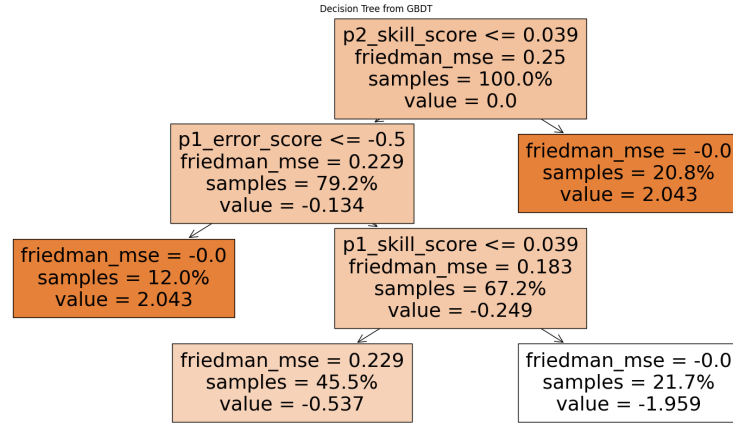


Figure 10: The Decision Tree from GBDT

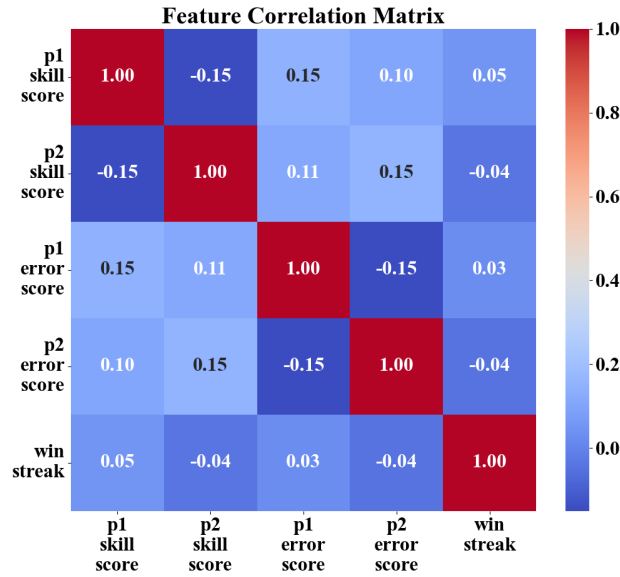


Figure 11: The Correlation Heatmap

Where

$$\begin{aligned}
 C^{(i)} &= p1_win_streak \\
 W^{(i-1)} &= serve \\
 P &= 1.2 \\
 w_1^{(i)} &= 2 - point_victor \\
 w_2^{(i)} &= point_victor - 1 \\
 l^{(i)} &= w^{(i)} - 1 \\
 \gamma &= \begin{cases} 0.5 & win_streak^{(i)} \cdot win_streak^{(i-1)} < 0 \\ 0.1 & other \end{cases} \\
 C^{(i)} &= \begin{cases} win_streak & win_streak \leq 4 \\ sign(win_streak) \cdot (4 + \sqrt{|win_streak| - 4}) & win_streak > 4 \end{cases}
 \end{aligned} \tag{16}$$

6.1.2 Result of Model

Followed by calculating momentum using Equation 15, the value of $p1 - p2$ is given as a momentum gap, which reflects the flow of the match. In Figure 12, one side will be the winner if he has a positive momentum gap, which verifies the relevance of momentum and success.

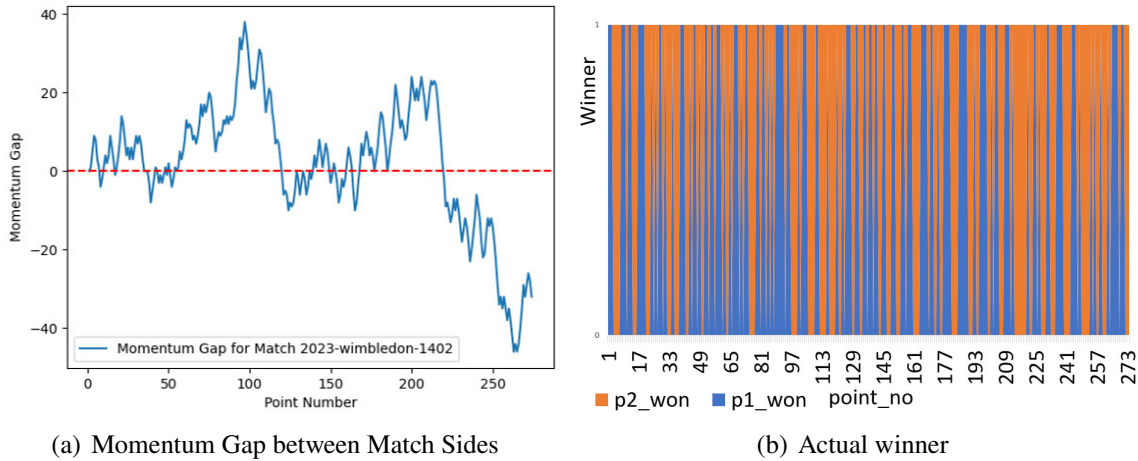


Figure 12: Contrastive Chart between Momentum Gap and Winner in Match 1402

6.1.3 Correlation Analysis

Based on the momentum computed in the calculation model, we apply the Logistic Regression Algorithm to predict the winning rate to verify the correlation between momentum and success with Point-Biserial Correlation Coefficient and P-value, which refutes the coach's idea of the problem. The prediction results are shown in Table 2. The result of the point-biserial correlation coefficient

Table 2: Prediction Results of XGBoost Model

Prediction Target	Precision	Recall	F1-score	Support
Player 1 winning	0.94	0.94	0.94	1063
Player 2 winning	0.94	0.94	0.94	1123
accuracy		0.9387		2186
Point-Biserial Correlation Coefficient			0.87838	
P-value			0	

is 0.87838 and the P-value is 0, which indicates the strong correlation between momentum and winning rate.

6.2 Momentum Fitted by SARIMAX Model

6.2.1 Model Establishment

SARIMAX, seasonal autoregressive integrated moving average with exogenous factors model, is a kind of time series model that combines seasonal autoregressive integrated moving average with

exogenous factors. The SARIMAX model has promising performance in terms of categorization and consideration, as it has significantly improved forecasting accuracy compared with the simpler autoregressive integrated moving average-based techniques.[11]

Based on the SARIMA model, the SARIMAX model which structured as Equation 17 introduces exogenous factors.

$$RA \cdot SRA \cdot (1 - G)^d \cdot (1 - G^s)^D \cdot x_t = \alpha_k y_{k,t} + MA \cdot SMA \cdot e_t \quad (17)$$

The $(1 - G)$ and $(1 - G^s)$ are non-seasonal and seasonal time series. e_t refers to regression error, $y_{k,t}$ refers to the number of external exogenous factors at time t and α_k is the correlation coefficient value of the external exogenous input factors. The RA , SRA mean autoregressive component and seasonal autoregressive component, and the MA , SMA are declared as moving average component and seasonal moving average component.

As long as set the number of (p, d, q) and $(P, D, Q)^S$, we could use this to fit momentum as the winning rate. On the basis of the predicted momentum data, we could judge whether the fluctuation lies in the game.

6.2.2 Model Results

After processing the SARIMAX model, we gained a series of momentum gap. For analyzing the accuracy, we plot the scatter diagram as Figure 13. The figure has plotted the error band which

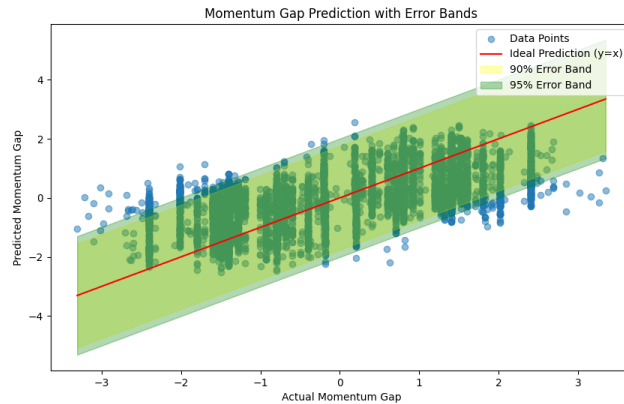


Figure 13: The Result of the Prediction of the Momentum Gap

including 10% error band in green and 5% error band in yellow. The width of bands are as shown as Table 3, which refers to a good accuracy.

Table 3: The Width of Error Bands

The Width of 90% band	The Width of 95% band
3.52	4.01

6.2.3 Correlation Features Analysis

After prediction, we evaluate the F score of each feature with SARIMAX model, which counts in Figure 14. From this figure, we could know the most related feature is $p1_{win_streak}$.

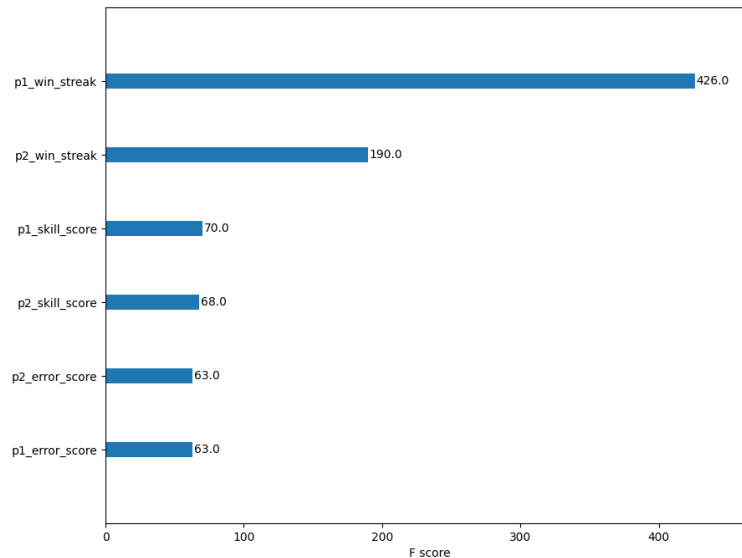


Figure 14: The F Score of Each Feature to Momentum Gap

7 Player Strategy: Strategy Choosing Based Statistic Analysis

7.1 Data Statistic Analysis

Based on the data in the raw table, we could use the last three columns to make a directed serve-and-return strategy. While analyzing the data, we also utilize the momentum gap to evaluate the different trends to strengthen the advantages and reverse the disadvantages. As a result, we divided the data of Alejandro Davidovich Fokina into three parts to compare their difference based on the momentum gap.

First of all, we analyze the direction of the serve when the player occupies advantage, disadvantage, and the turning part when standing on the disadvantage. According to the pie chart plotted in

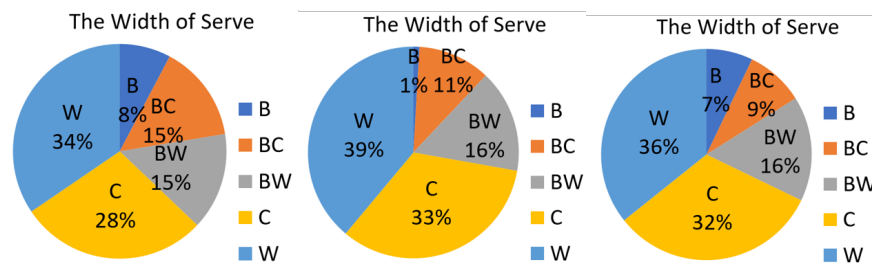


Figure 15: The Direction of the Serve

From left to right are the dominant position, the inferior position and the turning point

Figure 15, we find that the body and body/center direction may be related to the superiority while the center and wide direction mean the disadvantage.

Next, we analyze the depth of the serve. As the same of the analysis of the direction, we also contrast the different pie charts in Figure 2 to obtain results. From the pie chart, we find out the

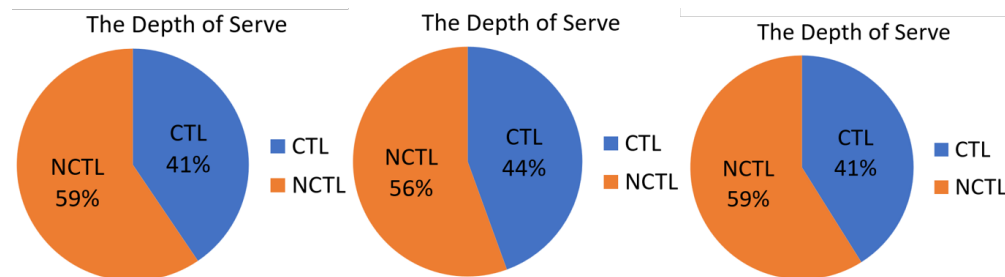


Figure 16: The Depth of the Serve

From left to right are the dominant position, the inferior position and the turning point

serve not close to the line seems to have more momentum to the winning. At the same time, this kind of serve could save the player from the chasm of the disadvantage.

Finally, we analyze the depth of return by picturing the pie chart in Figure 17 as well. This pie chart shows an obvious preference for the not deep return, though it has no benefit for the game. The

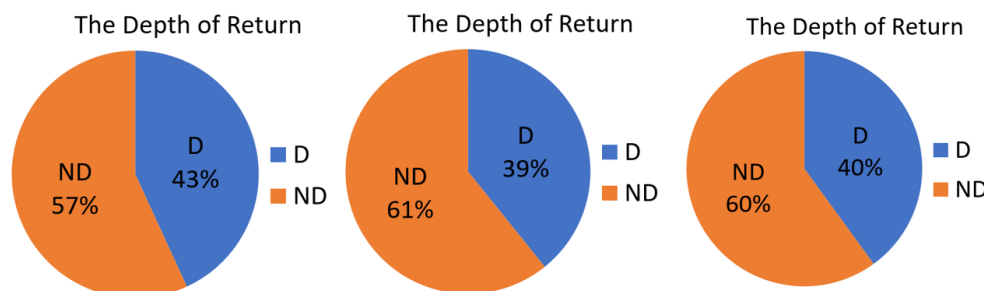


Figure 17: The Depth of the Return

From left to right are the dominant position, the inferior position and the turning point

feature of the figure is clear to observe, the deep return has preponderance in terms of momentum, which helps the player a lot to gain the victory.

7.2 Suggestions

In accordance with the analysis above, the suggestions on serve and return are clear to not only Alejandro Davidovich Fokina but also us.

Initially, though he is good at serving the center and the side opposite him, the most useful strategy is still body and body/center direction. When persuing the convenience on serve, the rival easily returns this serve which gives the initiative to the rival. This behavior will lead to inferiority at the beginning.

Secondly, try to serve close to the line, although this may be very difficult. Technically speaking, the line serve will mislead the rival's determination which leads to victory.

Eventually, avoid hitting a net return when others serve, as a result of the long distance which needs time to move. When the opponent finishes serving, he must choose to go forward or backward. Whatever he chooses, the long-distance shot is worth hitting.

8 Sensitivity Analysis

8.1 Player Performance Model

After changing the features $p1_win_streak$, we calculate the score percentage to the original score, then normalize it. After that, we plot the change of the score to the feature in Figure 18. From the figure, we found that the score changed rapidly around the original feature data, which

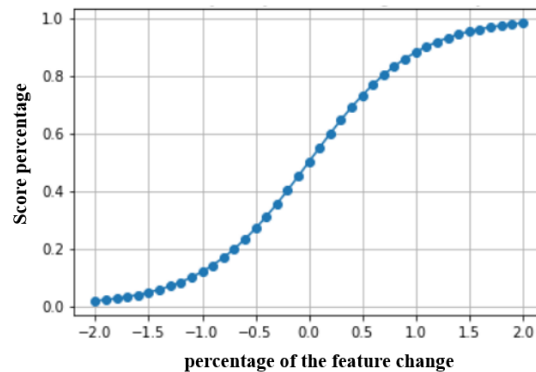


Figure 18: Player Performance Model Sensitivity Analysis

means that our performance model has great sensibility to $p1_win_streak$.

8.2 Plyer Momentum Model

After changing the features $p1_win_streak$, we calculate and normalize the momentum gap to the original gap. After that, we plot the change of the momentum gap to the feature in Figure 19.

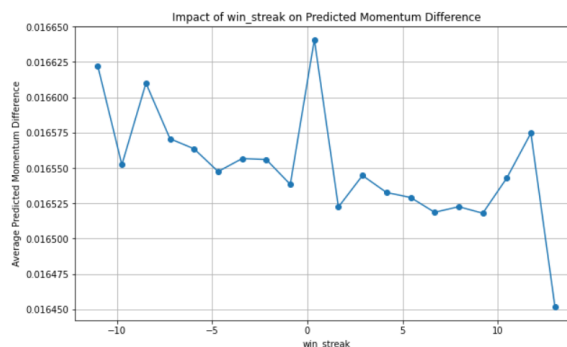


Figure 19: Player momentum Model Sensitivity Analysis

From the figure, the slope of the curve is rise and fall dramatically, which means sensibility to *p1_win_streak*.

9 Model Evaluation

9.1 Strengths

After analysis, our models' advantages are as follows.

- **Identification of Momentum's Impact:**

Our model has effectively highlighted the crucial influence of momentum within matches, providing coaches and players with vital indicators for match preparation. By recognizing the ebb and flow of a game's dynamics, the model equips the team with insights that are essential for strategic planning and execution. Generalizability: Exhibiting a commendable level of generalization, the model learns and predicts based on advanced features extracted from scoring points and various technical aspects of play. This approach renders the model applicable across a broad spectrum of matches, offering valuable lessons and strategies that can be adapted to diverse competitive scenarios.

- **Robustness:**

Demonstrating strong robustness, the model tolerates errors and new variables with remarkable resilience. This attribute ensures that the model remains reliable and accurate even when faced with unforeseen factors or data inconsistencies, making it a dependable tool for analysis.

- **Real-time Predictive Capability:**

The model serves as a foundation for real-time predictions of momentum shifts and match trajectories, facilitating instantaneous observation of win probabilities and enabling players to adjust their strategies on the fly. This real-time applicability enhances decision-making processes during crucial moments of the game.

9.2 Weaknesses

By contrast, our models also have some disadvantages.

- **Dependence on High-Quality Data:** Like many predictive models, the effectiveness of our analysis hinges on the availability of high-quality, comprehensive data. Incomplete or inaccurately recorded match details can limit the model's accuracy, underscoring the importance of robust data collection practices.

- **Potential for Overfitting:** While the model boasts strong predictive power, there's always a risk of overfitting the training data, particularly when dealing with intricate feature sets. Overfitting can reduce the model's ability to generalize to new, unseen matches, potentially affecting its practical application.

10 Conclusion

Eventually, we will summarize our results so that the reader can grasp the point rapidly. In the data description, we not only find that backhand shots victory account for one-third of all but also find that five features are almost enough to express this problem. In the player performance, the dynamic Bradley-Terry model with time series prediction could indirectly predict the winning situation. The prediction data and some basic situations are also given in the instrument panel. At the same time, this model performs excellently on women's data, which proves this model's generalization. For the momentum part, we have successfully extracted the best feature including *p1_skill_score*, *p2_skill_score*, *p1_error_score*, *p2_error_score*, *win_streak* in virtue of GBDT and heatmap. Then, a calculated model was given with these features. The most important is that we utilize the momentum gap to accurately predict the winning rate with the high Point-Biserial Correlation Coefficient(0.87838), which refutes that coach's idea. Based on this model's data, we predict the fluctuation with the SARIMAX model and obtain the most correlating feature through the XGBoost model: *p1_win_streak*. In the player strategy model, we statistic the data of serving and return of someone, and give him some suggestions about serving and returning. Finally, we analyze the strengths and weaknesses which give others a reference direction to study.

References

- [1] Jamie Braidwood. TITLE novak djokovic has created a unique rival – is wimbledon defeat the beginning of the end? <https://www.independent.co.uk/sport/tennis/novak-djokovic-wimbledon-final-carlos-alcaraz-b2376600.html>, 2023.
- [2] John M. Silva, Charles J. Hardy, and R. Kelly Crace. Analysis of psychological momentum in intercollegiate tennis. *Journal of Sport and Exercise Psychology*, 10(3):346 – 354, 1988.
- [3] Ronald Chettiar. TITLE live streaming, novak djokovic vs carlos alcaraz: Wimbledon 2023 men’s singles final - where to watch. <https://olympics.com/en/news/wimbledon-2023-tennis-men-singles-final-live-streaming-telecast-india>, 2023.
- [4] Abdullah M Almarashi, Muhammad Daniyal, and Farrukh Jamal. A novel comparative study of nnar approach with linear stochastic time series models in predicting tennis player’s performance. *BMC Sports Science, Medicine and Rehabilitation*, 16(1):1–11, 2024.
- [5] Ben Moss and Peter O’Donoghue. Momentum in us open men’s singles tennis. *International Journal of Performance Analysis in Sport*, 15(3):884–896, 2015.
- [6] Paolo Gorgi, Siem Jan Koopman, and Rutger Lit. The analysis and forecasting of tennis matches by using a high dimensional dynamic model. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 182(4):1393–1409, 2019.
- [7] Ricardo E. Rivet Victor K. R. Matsudo and Mônica H. N. Pereira. Standard score assessment on physique and performance of brazilian athletes in a six tiered competitive sports model. *Journal of Sports Sciences*, 5(1):49–53, 1987.
- [8] Drew Creal, Siem Jan Koopman, and André Lucas. Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, 28(5):777–795, 2013.
- [9] Andrew C Harvey. *Dynamic models for volatility and heavy tails: with applications to financial and economic time series*, volume 52. Cambridge University Press, 2013.
- [10] JeffSackmann. TITLE tennis_slam_pointbypoint. https://github.com/JeffSackmann/tennis_slam_pointbypoint, 2023.
- [11] Fahad Radhi Alharbi and Denes Csala. A seasonal autoregressive integrated moving average with exogenous factors (sarimax) forecasting model-based time series approach. *Inventions*, 7(4), 2022.

Memorandum

MEMORANDUM

To: Coach

From: Team # 2423444 of 2024 MCM

Subject: Comprehensive Analysis and Strategic Recommendations Based on 2023 Wimbledon Gentlemen's Tournament Data

Date: February 6, 2024

I trust this note finds you well. Following our thorough examination of the 2023 Wimbledon Gentlemen's Tournament data, with a particular focus on Alejandro Davidovich Fokina's performances as well as other leading players, our objective has been to discern the impact of momentum on match outcomes and to formulate corresponding strategies. This nuanced, multidimensional research approach is designed to offer a holistic and authoritative perspective for players and coaching teams, aiming to refine match strategies and mental preparations.

Identification and Utilization of Momentum

1. Recognition of Momentum Shifts: We encourage players to develop an acute sense of identifying momentum shifts within matches, marking critical transitions from a series of wins to losses, or vice versa. Such moments demand timely psychological and strategic adjustments to either consolidate or reverse the flow of the game.

2. Momentum Management in Training: Enhancing the ability to recognize and leverage momentum shifts and pivotal points during matches through training is crucial. Our analyses reveal that sustaining a winning streak, along with superior technical scores such as break points, aces, and unreturnable serves, significantly bolsters momentum, thereby improving win rates.

3. Steady Mindset and Technical Scoring: Faced with momentum pivots, maintaining a calm demeanor and capitalizing on technical scoring and minimizing errors are essential for swiftly reversing momentum. Moreover, disrupting the opponent's flow and pursuing an aggressive follow-up are key to securing victory.

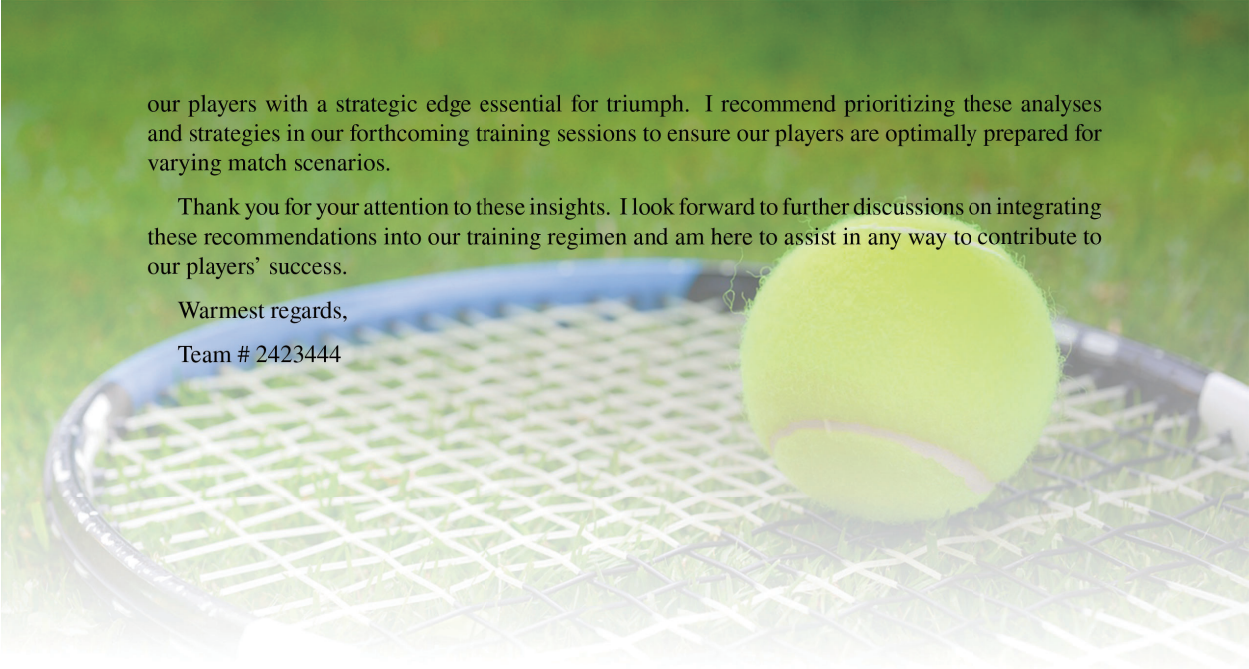
4. Customized Strategies: Drawing from an extensive analysis of past match data, we can tailor strategies to enhance individual performance, adapting our approach to the unique dynamics of each game.

Optimization of Technical Details

Your insights on serving and receiving have laid a solid foundation for tactical discussions. Emphasizing the refinement of these techniques will further augment our players' ability to command momentum, especially in navigating critical junctures with skill and mental fortitude.

Conclusion and Recommendations

Momentum plays a decisive role in tennis matches. By accurately identifying and harnessing momentum shifts, combined with continuous enhancements in serving and receiving, we can equip

A photograph of a tennis racket and a tennis ball on a grassy field. The racket is blue and white, and the ball is yellow-green. The background is a blurred green field.

our players with a strategic edge essential for triumph. I recommend prioritizing these analyses and strategies in our forthcoming training sessions to ensure our players are optimally prepared for varying match scenarios.

Thank you for your attention to these insights. I look forward to further discussions on integrating these recommendations into our training regimen and am here to assist in any way to contribute to our players' success.

Warmest regards,

Team # 2423444

Report on Use of AI

1. OpenAI ChatGPT (Nov 5, 2023 version, ChatGPT-4)

Query1: Can the final output of XGBoost, a decision tree model with a sliding window, include fluctuations in time series trends? I want to use XGBoost only to output the probability of winning. Or are models like SARIMAX better in this regard?

Output: Using XGBoost with a sliding window technique can indeed capture fluctuations in time series trends by incorporating historical data as features. This approach allows XGBoost to utilize past information to predict future outcomes, such as the probability of winning. However, SARIMAX and other time series models are specifically designed to model and forecast based on time series data, capturing trends, seasonality, and autocorrelation directly. While XGBoost can handle time series data with appropriate feature engineering, SARIMAX might offer more nuanced insights into time-dependent patterns for purely time series forecasting tasks.

2. OpenAI ChatGPT (Nov 5, 2023 version, ChatGPT-4)

Query1: What is the Kruskal Wallis H-test and what should be noted when using it in a tennis match model?

Output: The Kruskal-Wallis H-test is a non-parametric statistical test used to determine if there are statistically significant differences between the medians of three or more independent groups. It is often used as an alternative to the one-way ANOVA when the data do not meet the assumptions necessary for ANOVA, particularly normal distribution and homogeneity of variances. The test works by ranking all data points across all groups together and then comparing the sum of ranks between groups. It is especially useful for data that are ordinal, not normally distributed, or when the sample sizes are small or not equal across groups.

Key Points of the Kruskal-Wallis H-test:

- **Non-parametric**: It does not assume a normal distribution of the data. - **Independent Samples**: It compares three or more groups that are independent of each other. - **Rank-based**: The test uses ranks of data points rather than their actual values, making it less sensitive to outliers. - **Medians Comparison**: It tests for differences in the medians of the groups, not the means.

Considerations for Using in a Tennis Match Model

When applying the Kruskal-Wallis H-test in a tennis match model, several specific considerations should be taken into account:

1. **Data Type and Distribution**: Ensure the data being analyzed (e.g., scores, duration of matches, number of aces) does not follow a normal distribution, making the Kruskal-Wallis H-test a suitable choice. The test is ideal for ordinal data or continuous data that is not normally distributed.

2. **Group Independence**: The test requires that the groups (e.g., matches played on different types of courts, or performances in different weather conditions) are independent. Each group should represent a different condition with no overlap in participants (i.e., the same player's performances under different conditions can be compared, assuming each condition is treated as an independent sample).

3. **Sample Size**: While the Kruskal-Wallis H-test can handle small sample sizes better than parametric alternatives, having a reasonably sized sample in each group can help in achieving more reliable results. In the context of tennis matches, ensure that there are enough matches in each category being compared.

4. **Outcome Interpretation**: The test will indicate whether there's a statistically significant difference in the medians across groups, but it won't specify where the differences lie. If the test indicates significant differences, post-hoc pairwise comparisons will be necessary to identify which specific groups differ from each other.

5. **Match Specific Factors**: In a tennis match model, consider factors that could influence the outcome, such as player fatigue, psychological factors, or environmental conditions. These factors should be evenly distributed across the groups being compared or controlled for if possible.

6. **Use of Ranks**: Since the test is based on ranks, the magnitude of differences between scores is not considered, just the order. This means that very large or very small differences are treated the same as long as their rank order is consistent.

Applying the Kruskal-Wallis H-test to a tennis match model can provide insights into whether different conditions (like surface type, player rank, or tournament level) significantly affect match outcomes, such as match duration, points won, or other performance metrics. However, careful consideration of the test's assumptions and limitations is crucial for accurate interpretation and application of the results.

3. OpenAI ChatGPT (Nov 5, 2023 version, ChatGPT-4)

Query1: When using the Bradley Terry model to estimate the winning rate of tennis players, the results have too much variability, often very close to 1 or 0. Can you help me analyze the possible reasons and solutions?

Output: The Bradley-Terry model is a probabilistic model that estimates the relative strengths of players in paired comparisons—such as matches between tennis players—with the goal of predicting the outcomes of future matches. When the estimated winning probabilities are extremely close to 1 or 0, it indicates that the model perceives a very large difference in skill levels between the players. Here are some possible reasons for such extreme variability and potential solutions:

Possible Reasons

1. **Small Sample Size**: If there are not enough matches involving the players, the estimates may be unreliable and skewed towards the extremes. 2. **Unbalanced Data**: If top players mostly play against much weaker opponents (or vice versa), the model may overestimate the strength differences. 3. **Model Parameters**: The choice of model parameters, including the prior distributions (if using a Bayesian approach), can significantly affect the outcomes. Overly strong priors might bias the results. 4. **Data Quality**: Errors in the data, such as incorrect match outcomes or mixing matches from significantly different levels of play (e.g., mixing professional and amateur matches), can distort the model's estimates. 5. **Lack of Contextual Factors**: The Bradley-Terry model in its basic form does not account for contextual factors like player fatigue, surface preference, and recent form, which can influence match outcomes.

Solutions

1. ****Increase Sample Size****: Ensure that the model is trained on a sufficiently large and representative dataset to improve the reliability of the estimates.
2. ****Balance the Data****: Include a mix of matches involving opponents of varying strengths to provide a more balanced view of each player's performance.
3. ****Adjust Model Parameters****: Experiment with different model parameters, including priors, to see how they affect the outcomes. In Bayesian versions, consider weaker priors to reduce bias.
4. ****Improve Data Quality****: Clean the dataset to remove any errors and ensure that only relevant matches are included in the analysis.
5. ****Extend the Model****: Incorporate additional factors into the model. For instance, modify the Bradley-Terry model to account for player-specific characteristics or match conditions. This could involve creating separate models for different surfaces or adjusting player strengths based on recent performance.
6. ****Regularization****: Introduce regularization to prevent overfitting, especially in cases where the data might not fully represent the true strength levels of players.
7. ****Hierarchical Models****: Consider hierarchical or mixed-effects models that can share strength information among similar players, potentially stabilizing estimates.

By addressing these potential issues, you can improve the accuracy and reliability of the winning rate estimates produced by the Bradley-Terry model for tennis players. Remember, the key is to ensure that the model has access to high-quality, comprehensive data and that it accounts for the complex realities of competitive sports where many factors can influence the outcome of a match.