Problem Chosen	2024	Team Control Number
C	MCM/ICM Summary Sheet	2411120

Summary

As for problem 1, we first selected several positive and negative factors base on our research and comprehension towards tennis matches. We then standardized the factors by Min-max normalization. Moving on, we attribute different weights to each factor, and compute a performance score for each player at every point. The difference in the performance score represents the extend that one player outperforms the other player. We visualized two figures representing the difference in performance and the difference in points gained by the two players correspondingly. The result shows a relativity between the performance and points won by the player, proving our effectiveness of our performance calculation.

As for problem 2, we defined a metric called "leverage", which captures the importance of a point/game contributing to a player who is winning the game/set. We constructed a set winning probability matrix and a point winning probability matrix by dynamic programming to help compute the leverage. We utilized the leverage metrics to compute the dynamic momentum during the match. Moving on, we investigated the relationship between momentum and the occurrence of success and/or swing. The result shows that the momentum captured by our model performs well in predicting the occurrence of swing and success in tennis matches, especially for the swing and the success of a game.

As for problem 3, we first extracted the factors that we think can predict swings from the dataset. Then we train the Random Forest, GBDT, XGBoost, and CATBoost Models with the factors data. We told from the result that each of the model has performed a satisfactory prediction. Therefore, we developed our model by aggregating the four models. We assigned the weight to each of the model according to their accuracy. Then, we took the weighted average and ran our model to see the result. The result shows that whether being the winner of an untouchable shot, gaining a net point won, and being the server are three main factors that contribute to a swing.

As for problem 4, we first ran our model using data from other matches in the tournament. Our model has achieved a satisfactory performance in most of the other games. To further enhance our model, we listed the factors that can be taken into consideration in the future. We further analyze the generality of our model by analyzing women's tennis match, tennis match on other court surfaces and other sport matches. Since our model is designed based on specific dimensions of factors in men's tennis match, it might not work well if being applied directly to other sports. Therefore, we analyzed the aspects to consider when applying our model in other circumstances, such as adjusting factor weights or adding new factors, making it easier to generalized our model.

Keywords: Dynamic Programming; Leverage, Momentum; Binary Classification; Ensemble learning; Feature Engineering

Summary

As for problem 1, we first selected several positive and negative factors base on our research and comprehension towards tennis matches. We then standardized the factors by Min-max normalization. Moving on, we attribute different weights to each factor, and compute a performance score for each player at every point. The difference in the performance score represents the extend that one player outperforms the other player. We visualized two figures representing the difference in performance and the difference in points gained by the two players correspondingly. The result shows a relativity between the performance and points won by the player, proving our effectiveness of our performance calculation.

As for problem 2, we defined a metric called "leverage", which captures the importance of a point/game contributing to a player who is winning the game/set. We constructed a set winning probability matrix and a point winning probability matrix by dynamic programming to help compute the leverage. We utilized the leverage metrics to compute the dynamic momentum during the match. Moving on, we investigated the relationship between momentum and the occurrence of success and/or swing. The result shows that the momentum captured by our model performs well in predicting the occurrence of swing and success in tennis matches, especially for the swing and the success of a game.

As for problem 3, we first extracted the factors that we think can predict swings from the dataset. Then we train the Random Forest, GBDT, XGBoost, and CATBoost Models with the factors data. We told from the result that each of the model has performed a satisfactory prediction. Therefore, we developed our model by aggregating the four models. We assigned the weight to each of the model according to their accuracy. Then, we took the weighted average and ran our model to see the result. The result shows that whether being the winner of an untouchable shot, gaining a net point won, and being the server are three main factors that contribute to a swing.

As for problem 4, we first ran our model using data from other matches in the tournament. Our model has achieved a satisfactory performance in most of the other games. To further enhance our model, we listed the factors that can be taken into consideration in the future. We further analyze the generality of our model by analyzing women's tennis match, tennis match on other court surfaces and other sport matches. Since our model is designed based on specific dimensions of factors in men's tennis match, it might not work well if being applied directly to other sports. Therefore, we analyzed the aspects to consider when applying our model in other circumstances, such as adjusting factor weights or adding new factors, making it easier to generalized our model.

Keywords: Dynamic Programming; Leverage, Momentum; Binary Classification; Ensemble learning; Feature Engineering

Contents

1 Introduction

1.1 Problem Background

Momentum, defined by strength or force gained by motion or by a series of events, plays an important part in various sports games. A positive shift in momentum results in increased self-efficacy and motivation, which in turn results in enhanced performance and possibly future success. A negative shift in momentum results in decreased self-efficacy and performance, which may lead to defeat. In tennis matches where the outcome is determined by a complex set of rules, investigating the effect of momentum is crucial. Understanding and harnessing momentum can be the key to mastering the strategic and mental aspects of the game, making it a central element in the sport's dynamics.

In this report we mainly focused on the role of momentum in tennis matches. We utilized the 2023 Wimbledon Gentlemen's singles matches dataset with a wide dimension of data for every point to evaluate the performance and momentum of the athletes. We investigated the relationship between momentum and the swing in a play and the runs of success. In the end, we summarized our findings and wrote a memo to the coach with our suggestions.

1.2 Clarification and Restatement

In this problem, we are given a data set of every point from all Wimbledon 2023 men's matches after the first 2 rounds. The data set includes a wide range of data recorded throughout the game.

We should only use these data to solve the following problems:

- Develop a model that evaluate the performance of each player as points occur and compare how much better they are outperforming their opponent. Apply the model to one or more of the matches and provide a visualization to depict the match flow based on the model.
- Use the model/metric developed in task 2 to assess the claim that swings in play and runs of success by one player are random.
- Using the data provided for other matches to develop a model that predicts the swings in the match. Identify What factors seem most related (if any).
- Investigate the momentum swings data for one player, and give some suggestions to the player based on the findings.
- Test the model on one or more of the other matches. Evaluate the generality of our model and identified factors that might need to be included in future models.

2 Problem 1

2.1 Analysis of Problem 1

For problem 1, our task is to develop a model that can identify which player is performing better at a given time as points occur, as well as how much better they are performing.

To develop such a model, We need a standardized score to measure the performance of each player. To achieve this, we need to select several factors and normalized them. The standardized value of each factor is used to compute a weighted score for each player at every point. Since there's no reference standard, we should determine the weights of factors by common sense. Furthermore, noticing better performance of a player usually results in his winning a point, we visualize the difference in two players' performance and the points gained by two players for the whole match to see if there's any correlation between them.

2.2 Factor Selection and Weights Determination

Among all the data in the given dataset, we selected 12 factors in computing the performance score base on three aspects: technical skills, psychological states, and fatigue level. Those selected factors and their weights (the sign represents positive or negative effect) for calculating the overall performance are shown in Table 1-3.

- For technical skills, we consider whether the player hit untouchable winning shots, whether he made unforced errors and so on. Those factors can reflect his performance in a game and directly lead to his winning or losing a point/game.
- For psychological states, we consider current score advantage (or disadvantage) as positive (or negative) effects since it can boost or undermine one's confidence and influence their mood.
- For fatigue level, we calculated the sum of the distance ran by the player during the last 3 points. Considering there are breaks between each sets, we set the fatigue level into 0 at the beginning of each set.

Factor Name	**Weight**	**Explanation**
Serving status	+0.0667	Player's serving speed (mph) or 0 (if he does not serve)
ACE	+0.0667	1 (if the player achieve an ACE) or 0 (if not)
Untouchable winning shot	+0.1000	1 (if the player hit an untouchable winning shot) or 0 (if not)
Net point	+0.1000	1 (if the player won a point near the net) or 0 (if not)
Break point won	+0.1333	1 (if the player won at the break point) 0 (if not)
Break point missed	- 0.1333	1 (if the player lost at the break point) 0 (if not)
Unforced error	- 0.1000	1 (if the player made an unforced error) 0 (if not)
Double fault	- 0.1000	1 (if the player committed a double fault) 0 (if not)

Table 1: Technical Factors

Factor Name	Type	Explanation
Sets won (dis)advantage	± 0.0333	The difference of the leading player's number of winning sets in the current match to that of the trailing player, with a positive value for the winning player and the negative for the trailing player
Games won (dis)advantage	± 0.0333	The difference of the leading player's number of games in the current set to that of the trailing player, with a positive value for the winning player and the negative for the trailing player
Points won (dis)advantage	± 0.0667	The difference of the leading player's winning points in the current game to that of the trailing player, with a positive value for the winning player, and the negative for the trailing player

Table 2: Psychological Factors

Factor Name	Type	Explanation
Distance ran	- 0.0667	Distance the player ran during last 3 points

Table 3: Fatigue Level Factors

2.3 Data Normalization

After selecting suitable factors and determining their weights, we then get the feature values for each player at each point. Noting the range of the values of each factor vary widely in the dataset, which might lead to those factors varying largely dominating the variation of performance values, we do the min-max normalization, scaling the values of each factor into [0, 1].

For feature $\boldsymbol{x_i}$, we calculate the normalized value of x_{ij} as follows:

$$x'_{ij} = \frac{x_{ij} - \min(\boldsymbol{x_i})}{\max(\boldsymbol{x_i}) - \min(\boldsymbol{x_i})} \tag{1}$$

factor, since being the server in each game may have significant impact on one's performance. All the factors and their adjusted weights are listed below in Table 4.

2.4 Performance Calculation and Visualization

We computed the performance score $\boldsymbol{pf}$ by taking the weighted average of the factor values using the formula:

$$\boldsymbol{pf} = \sum_{i=1}^{12} w_i \boldsymbol{x_i} \tag{2}$$

where w_i is the weight of the i-th factor.

The comparison of two players' performance and points won are shown in Figure 1. Four bold black dashed lines separate five sets and light black dashed line separate all the games.

In the upper figure, the two performance curves are drawn and their middle part are filled in two colors to show which player is perform better.

In the lower figure, the point difference of two players in each game are shown, indicating the front runner and the gap of points between two players.

By comparing two figures in Figure 1, we can tell that relativity exists between them. For example, in the end of the first game of the first set, the performance of Djokovic is in the leading position, and correspondingly, he won two more points than Alcaraz. There are many situations like this. The relativity shows that our model effectively depicts the flow of the match.

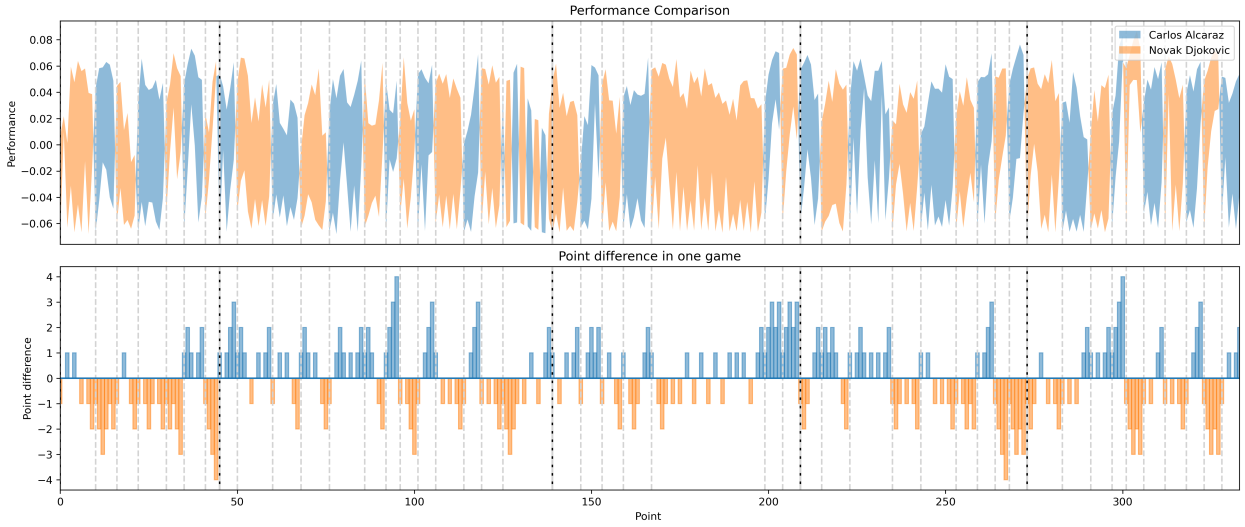


Figure 1: Difference in Performance

3 Problem 2

3.1 Analysis of Problem 2

For problem 2, our task is to investigate if the "momentum" plays any role in predicting the occurrence of swings and success.

To complete the task, we first need to define some metrics based on the understanding of tennis matches. We learnt from some previous research and defined a metric called "leverage", which captures the importance of a game (or point) contributing to a player winning the set (or game). Given the leverage, we further evaluate the momentum for each player at each game (or point) during the match. We compute the player's momentum by an exponentially weighted moving average of the leverage gained by a player. At last, we need to investigate the relativity between momentum and swings as well as success. We implemented a binary classification to verify how well momentum can predict swings and success in a set (or game).

3.2 Assumptions and Notations

To simplify our model and eliminate any confusion, we make the following main assumptions. All assumptions will be re-emphasized once they are used in the construction of our model.

- **Assumption 1:** There is no significant gap between the skill of each player.
- **Assumption 2:** The server of each game has a 0.6 probability of winning the game, and a 0.4 probability of losing the game, correspondingly.

In our following analysis, there are a few symbols to use. We list all important notaions in advance for convenience in the following Table 4.

Symbol	Definition
t_1	The position of the investigated game in a set
t_2	The position of the investigated point in a game
L_{t_1}	Leverage of the t_1-th game in a set
l_{t_2}	Leverage of the t_2-th point in a game
i	The number of games or points won by player A
j	The number of games or points won by player B
p	The probability of player A winning a point in a game
q	The probability of player B winning a point in a game
p_1	The probability of player A winning a game when serving the game
q_1	The probability of player B winning a game when A serves the game
p_2	The probability of player A winning a game when B serves the game
q_2	The probability of player B winning a game when serving the game
P_1	Probability matrix for winning a set
P_2	Probability matrix for winning a game
X_{11}	Probability for player A winning the game at 40:40 (a deuce)
X_{21}	Probability for player A winning the game at AD : 40
X_{12}	Probability for player A winning the game at 40 : AD

Table 4: Notations for Problem 2

3.3 Momentum Model

During our literature review, we found that many of the previous studies on momentum investigated the momentum at some pre-defined key-moments, such as a break point and clutch point. As tennis match is a consecutive process, the limitation of such an approach is that the dynamics of the match is hard to be captured. In a study done by Robert Seidl and Patrick Lucey, the authors created the metrics of "leverage", "clutch point" and "momentum". These metrics are created to capture the importance of a point contributing to a player winning the set and/or match, or the likelihood of a comeback.

3.3.1 Definition of Leverage

Since the winning process for a set and that of a game is different in tennis, we investigated the leverage of a set and that of a game separately. The definition of the two metrics are as follows:

- **Game leverage**: The game leverage (L_{t_1}) of a player during the t_1-th game in a certain set is the change of his winning possibility for the set as a result of winning the game.
- **Point leverage**: The point leverage (l_{t_2}) of a player during the t_2-th point in a certain game is the change of his winning possibility for the game as a result of winning the point.

In short, game leverage reflects the impact of a player's game win on their overall chances of winning the current set, while point leverage measures how winning a single point influences a player's likelihood of winning the current game.

3.3.2 Probability Matrix P_1 and game leverage Computation

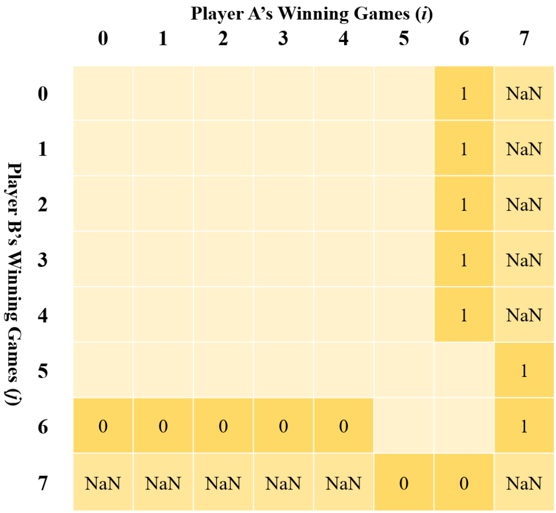
To calculate the game leverage, we first need to figure out the player's probability of winning a set when the number of games won by player A versus that of player B is i:j, for all possible i and j. This can be done by defining a winning probability matrix P_1 and using dynamic programming. Details are shown in (1) - (3). Using the matrix P_1, we can then calculate the game leverage. Details are shown in (4).

(1) Initialize the boundary values of P_1

To win a set, the player needs to win at most 7 games (either he won 6 games and was leading more than 2 games or he must win one more game after winning 6 games). Noticing this, we can initialize P_1 as in Figure 2. This winning probability matrix focuses on the probablity of player A winning the current game.

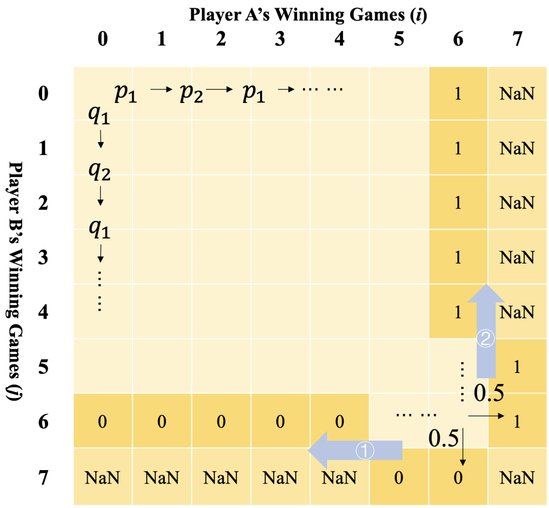
(2) Calculating other values in the middle of P_1

After initializing the boundary, we used four variables: p_1, p_2, q_1, q_2 to fill in the rest of the blocks. The definition of the four variables can be seen on Table 5. Since two players serve alternately in a set, p_1, p_2 and q_1, q_2 appears alternately in a row (which one is first is determined by which player serves first) (see Figure 3) except that when it is 6:6, we assume both players have the same opportunity to win one more game (because they serve in turn for the last game).



Player B's Winning Games (j) \ Player A's Winning Games (i)	0	1	2	3	4	5	6	7
0							1	NaN
1							1	NaN
2							1	NaN
3							1	NaN
4							1	NaN
5								1
6	0	0	0	0	0			1
7	NaN	NaN	NaN	NaN	NaN	0	0	NaN

Figure 2: Initialization of Set Matrix



Player B's Winning Games (j) \ Player A's Winning Games (i)	0	1	2	3	4	5	6	7
0							1	NaN
1							1	NaN
2							1	NaN
3							1	NaN
4							1	NaN
5								1
6	0	0	0	0	0			1
7	NaN	NaN	NaN	NaN	NaN	0	0	NaN

$p_1 \rightarrow p_2 \rightarrow p_1 \rightarrow \cdots\cdots$
q_1
q_2
q_1
⋮
②
0.5
⋯ ⋯
①
0.5

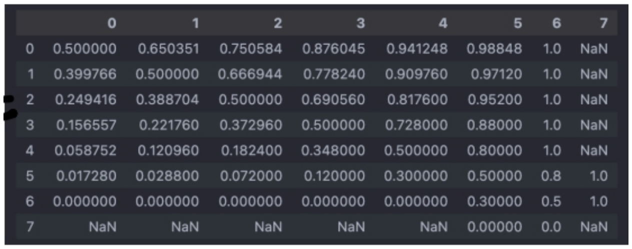
Figure 3: Game leverage Matrix

To calculate other values in the middle, we follow the order of right to left and then bottom to top as indicated by two blue arrows in Figure 3.

$$P_1[j,i] = p \times P_1[j,i+1] + q \times P_1[j+i,i] \tag{3}$$

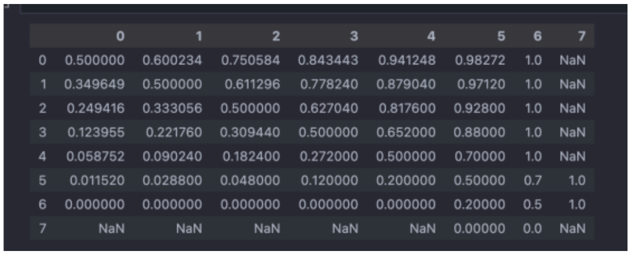
(3) Result of P_1

Since the player who serves has larger probability of winning a game, we set $p_1 = 0.6$ and $p_2 = 0.4$ if player A serves the first game in the current set. The corresponding matrix P_1 is shown in Figure 4. If player B serve first, then we set $p_1 = 0.4$ and $p_2 = 0.6$ with the matrix P_1 showing in Figure 5.



	0	1	2	3	4	5	6	7
0	0.500000	0.650351	0.750584	0.876045	0.941248	0.98848	1.0	NaN
1	0.399766	0.500000	0.666944	0.778240	0.909760	0.97120	1.0	NaN
2	0.249416	0.388704	0.500000	0.690560	0.817600	0.95200	1.0	NaN
3	0.156557	0.221760	0.372960	0.500000	0.728000	0.88000	1.0	NaN
4	0.058752	0.120960	0.182400	0.348000	0.500000	0.80000	1.0	NaN
5	0.017280	0.028800	0.072000	0.120000	0.300000	0.50000	0.8	1.0
6	0.000000	0.000000	0.000000	0.000000	0.000000	0.30000	0.5	1.0
7	NaN	NaN	NaN	NaN	NaN	0.00000	0.0	NaN

Figure 4: Matrix P_1 when player A serves



	0	1	2	3	4	5	6	7
0	0.500000	0.600234	0.750584	0.843443	0.941248	0.98272	1.0	NaN
1	0.349649	0.500000	0.611296	0.778240	0.879040	0.97120	1.0	NaN
2	0.249416	0.333056	0.500000	0.627040	0.817600	0.92800	1.0	NaN
3	0.123955	0.221760	0.309440	0.500000	0.652000	0.88000	1.0	NaN
4	0.058752	0.090240	0.182400	0.272000	0.500000	0.70000	1.0	NaN
5	0.011520	0.028800	0.048000	0.120000	0.200000	0.50000	0.7	1.0
6	0.000000	0.000000	0.000000	0.000000	0.000000	0.20000	0.5	1.0
7	NaN	NaN	NaN	NaN	NaN	0.00000	0.0	NaN

Figure 5: Matrix P_1 when player B serves

(4) Leverage Computation

With the probability matrix in hand, we can calculate the leverage of the t_1-th game in the current set for player A and player B as follows:

$$\begin{cases} {L^A}_{t_1} = |P_1[j,i] - P_1[j,i+1]| \\ {L^B}_{t_1} = |P_1[j+1,i] - P_1[j,i]| \end{cases} \tag{4}$$

Take t_1 = 8 as an example. Assume the number of games won by player A versus that of player B is 4:3. If player A wins the current game, it will become 5:3; Otherwise, it is 4:4. We calculated the game leverage of player A (${L^A}_8$) and that of player B (${L^B}_8$) as follows:

$$\begin{cases} {L^A}_8 = |P_1[3,4] - P_1[3,5]| \\ {L^B}_8 = |P_1[3,4] - P_1[4,4]| \end{cases}$$

(4) Momentum Computation

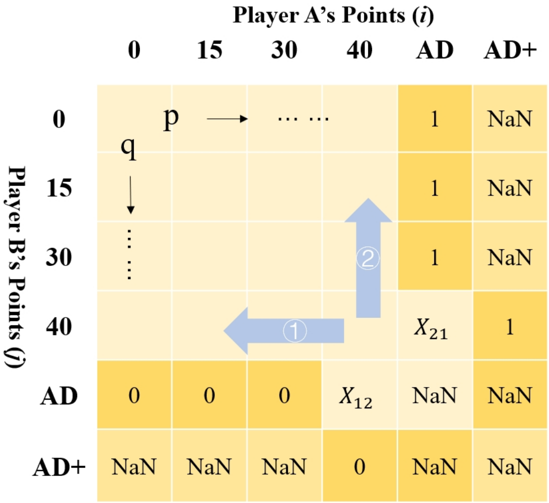
We calculate momentum of a player (M_{t_1}) during the t_1-th game in the current set as an exponentially weighted moving average of his game leverage as follows:

$$M_{t_1} = \frac{S_{t_1-1}L_{t_1-1} + (1-\alpha)S_{t_1-2}L_{t_1-2} + \cdots + (1-\alpha)^{t_1-2}S_1L_1}{1 + (1-\alpha) + \cdots + (1-\alpha)^{t_1-2}} \tag{5}$$

where L_1, L_2, ..., L_{t_1-1} are the game leverages of the last $t_1 - 1$ games with smoothing factor α. S_t is 1 if the player wins the t-th game in the current set and is -1 otherwise. We learnt from the study done by Robert Seidl and Patrick Lucey, and utilize α = 0.33 in our work.

3.3.3 Probability Matrix P_2 and point leverage Computation

As for the leverage of a certain point, we utilized the similar method in boundary initialization and matrix construction as shown below in Figure 6. We use p to denote the possibility of winning a point for player A, and use q to denote that of player B:



	0	15	30	40	AD	AD+
0	p q		… …		1	NaN
15					1	NaN
30				②	1	NaN
40		①			X_{21}	1
AD	0	0	0	X_{12}	NaN	NaN
AD+	NaN	NaN	NaN	0	NaN	NaN

Figure 6: Point leverage Matrix

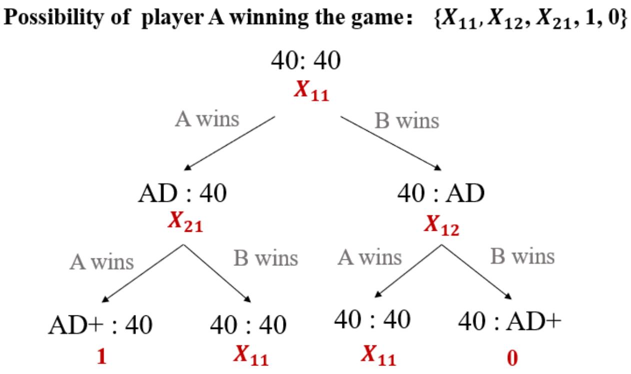


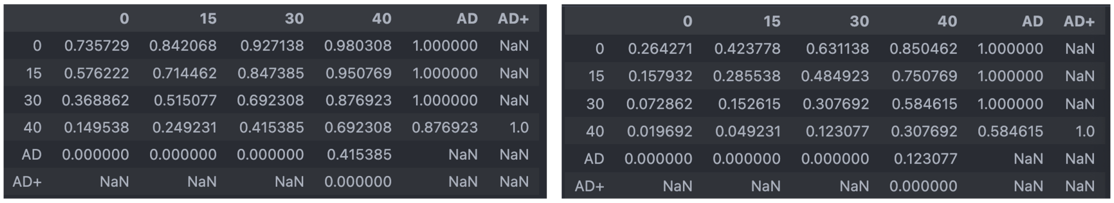
Figure 7: Winning Possibility Tree of player A after a deuce

One thing that is special for the game scenario is the "AD" position. To win a set, the player only need to win over the component for two games, or win the tiebreak. To win a game, the player need to first secure an "AD" position, and can only win the game by scoring another point right after being in the "AD" position (denoted as AD+ here). If there was already a deuce and the leading player failed to maintain the "AD" position, then the score returns to a deuce. Considering this special case, we developed a player winning possibility tree (see Figure 7) to deal with the deuce in a game.

As Figure 7 has shown, the possibility of winning a game after a deuce for a player can be denoted as a set consists of $\{X_{11}, X_{21}, X_{12}, 1, 0\}$. With the relationship shown in the tree and the matrix, we constructed a group of equations to compute X_{11},X_{21}, and X_{12}:

$$\begin{cases} X_{11} = pX_{21} + qX_{12} \\ X_{21} = p + qX_{11} \\ X_{12} = pX_{11} \end{cases} \Rightarrow \begin{cases} X_{11} = p^2/(1-2pq) \\ X_{21} = (p + p^2 q)/(1-2pq) \\ X_{12} = p^3/(1-2pq) \end{cases}$$

When player A serves the game, he has more chances to win a point, so, we set $p = 0.6$ and $q = 0.4$. When player B serve the game, we set $p = 0.4$ and $q = 0.6$. Similarly, we can get the middle values of P_2. The final result of P_2 is shown in Figure 8, with the left one for A serving and the right one for B serving.



	0	15	30	40	AD	AD+
0	0.735729	0.842068	0.927138	0.980308	1.000000	NaN
15	0.576222	0.714462	0.847385	0.950769	1.000000	NaN
30	0.368862	0.515077	0.692308	0.876923	1.000000	NaN
40	0.149538	0.249231	0.415385	0.692308	0.876923	1.0
AD	0.000000	0.000000	0.000000	0.415385	NaN	NaN
AD+	NaN	NaN	NaN	0.000000	NaN	NaN

	0	15	30	40	AD	AD+
0	0.264271	0.423778	0.631138	0.850462	1.000000	NaN
15	0.157932	0.285538	0.484923	0.750769	1.000000	NaN
30	0.072862	0.152615	0.307692	0.584615	1.000000	NaN
40	0.019692	0.049231	0.123077	0.307692	0.584615	1.0
AD	0.000000	0.000000	0.000000	0.123077	NaN	NaN
AD+	NaN	NaN	NaN	0.000000	NaN	NaN

Figure 8: Matrix P_2

Given matrix P_2, we can get the point leverage for player A (${l^A}_{t_2}$) and player B (${l^B}_{t_2}$) at the t_2-th point in the current game as the following three cases.

- When points won by player A versus that of player B is i : j (i : j $\neq$ 40 : AD or AD : 40),

$$\begin{cases} {l^1}_{t_2} = |P_2[j,i] - P_2[j, i+1_{point}]| \\ {l^2}_{t_2} = |P_2[j+1_{point}, i] - P_2[j,i]| \end{cases} \tag{6}$$

*$i + 1_{point}$ refers to the column after i; $j + 1_{point}$ refers to the row after j

- When the score of the game is 40 : AD,

$$\begin{cases} {l^1}_{t_2} = |P_2[AD, 40] - P_2[40, 40]| \\ {l^2}_{t_2} = |P_2[AD, 40] - P_2[AD+, 40]| \end{cases} \tag{7}$$

- When the score of the game is AD : 40, it is similar as the previous case.

$$\begin{cases} {l^1}_{t_2} = |P_2[40, AD] - P_2[40, AD+]| \\ {l^2}_{t_2} = |P_2[40, AD] - P_2[40, 40]| \end{cases} \tag{8}$$

Likewise, we can compute the momentum of the t_2-th point with the formula below:

$$m_{t_2} = \frac{s_{t_2-1}l_{t-1} + (1-\alpha)s_{t_2-2}l_{t_2} + \cdots + (1-\alpha)^{t_2}s_1 l_1}{1 + (1-\alpha) + \cdots + (1-\alpha)^{t_2}} \tag{9}$$

where l_1, l_2, ..., l_{t_2-1} are the point leverages of the last $t_2 - 1$ points with smoothing factor $\alpha = 0.33$. s_t is 1 if the player wins the t-th point in the current game and is -1 otherwise.

3.4 Momentum Model Performance Evaluation

3.4.1 Evaluation Approach

To evaluate how well our momentum model could predict the occurrence of swings and runs of success, we investigate the relationship between the momentum difference of two players and the occurrences of the targeted events. As in previous steps, we broke down the match into game layer and set layer, and investigated the relationship separately. We validate the result by utilizing confusion matrix for binary classification.

3.4.2 Data Processing

Considering the effectiveness and accuracy of the model, we processed some of the data using the following ways:

- We defined a swing as a game (or point) after a tie that won by the player who is also winning the games (or points) before the tie.

- For both analysis, we excluded the $0:0$ tie, since there is no winner before such tie.
- For game layer analysis, we excluded the first 3 games in each set due to the "cold starting problem" (the meaning of this term is stated in **3.4.3** with an example).
- For point layer analysis, we excluded the first 3 points in each game due to the "cold starting problem".
- We defined momentum difference as the momentum of the winner in the previous game minus that of the opponent's (see the equations below).

$$\Delta M(or\Delta m) = \begin{cases} Momentum_{playerA} - Momentum_{playerB}, & \text{If player A won previously} \\ Momentum_{playerB} - Momentum_{playerA}, & \text{If player B won previously} \end{cases} \tag{10}$$

- We validate the result utilizing confusion matrix consisting of the following formula:

$$\begin{cases} Accyracy = \frac{TP+TN}{TP+TN+FP+FN} \\ Precision = \frac{TP}{TP+FP} \\ Recall = \frac{TP}{TP+FN} \end{cases} \tag{11}$$

3.4.3 An Example to Illustrate Using Momentum Model

We here show how to use our momentum model in the game layer. More precisely, we give an example of using the momentum of player A ($M_{t_2}^{A}$) and that of player B ($M_{t_2}^{B}$) during the t_2-th game in the current set to predict swings of the number of games won by two players as well as the game victors.

In Figure 9, we visualized the number of games won by two players and their momentum in the match of Alcarz and Djokovic. In set 4 of this match, there are two ties, namely 1:1 and 2:2.

- When it is 1:1, momentum of player A versus that of player B is -0.1183 and -0.1019. The latter one is larger than the former one ($\Delta M < 0$), so we predict that it is Djokovic to win one more game and expect a swing. However, the truth is the opposite. Here our model gets and error.

  The reason for making such an error is possibly that only 4 games have been played, which makes it hard for people to tell who is performing better in this set and also makes it hard for the momentum to capture the "strength" or "force" gained by the previous few games in the current set. We call this a "cold starting problem" as mentioned in the data processing part. And this is also the reason that we excluded the first several games or points for our model.
- When it is 2:2, momentum of player A versus that of player B is -0.1183: -0.1019. The latter one is larger than the former one ($\Delta M < 0$), so we also predict that it is Djokovic to win one more game and expect a swing, which is actually the truth. Our model predicts correctly in this case.

Other than looking at the momentum value at a tie, we can also use momentum to predict the game victor not at a tie. For example, during the 8-th game in set 4, the momentum of Djokovic is larger, so we predict he will win one more game. The truth is that he is the victor and it turns from 3:5 to 3:6.

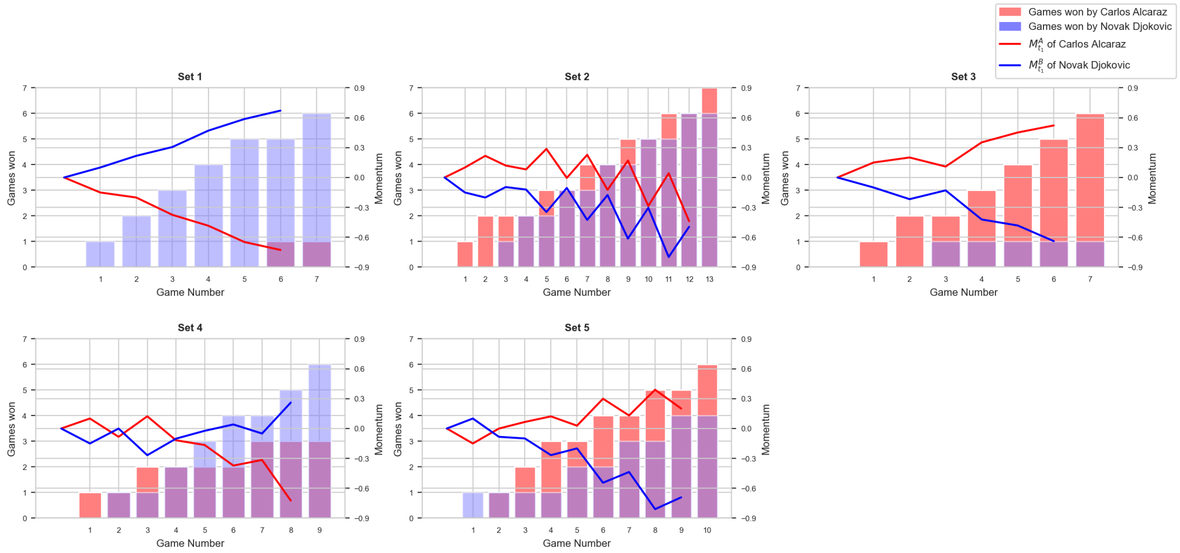


Figure 9: Comparison of Games Won and Momentums in each Set

3.4.4 Result Analysis

Momentum and swing:

According to the definition of swing, and momentum difference, the expected value for the momentum difference at a tie before a swing may be above 0. Likewise, the expected value for the momentum difference at a tie without any swing may be bellow 0. As we can see in Figure 10 and 11, the area cover with green indicates that the momentum made appropriate predictions towards the occurrence of swings. Binary classification evaluation table is shown in Table 5.

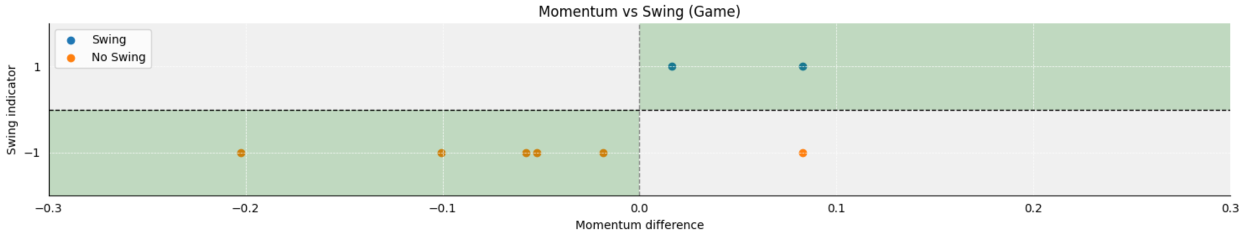


Figure 10: Momentum Difference and Swing Occurrence (Game)

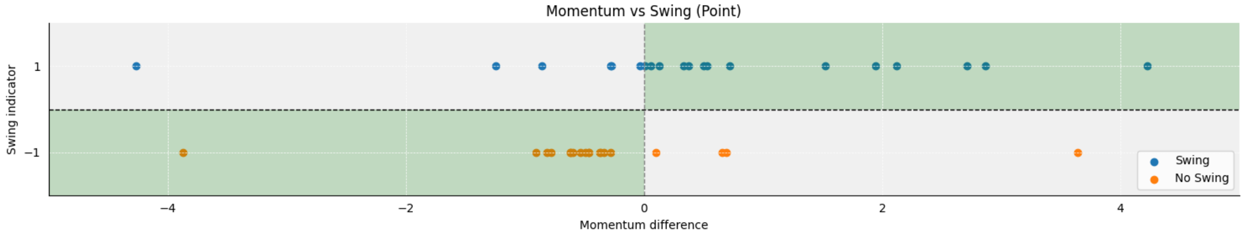


Figure 11: Momentum Difference and Swing Occurrence (Point)

Layer	Accuracy	Precision	Recall
Game	0.875	0.667	1.000
Point	0.744	0.778	0.700

Table 5: Binary Classification Evaluation for Using Momentum to Predict Swings

Momentum and victory:

The expected value for the momentum difference when the player wins the point may be above 0. Likewise, the expected value for the momentum difference when the player lose may be bellow 0. As we can see in Figure 12 and 13, the area cover with green indicates that the momentum made appropriate predictions towards the winner's success. Binary classification evaluation table is shown in Table 6.

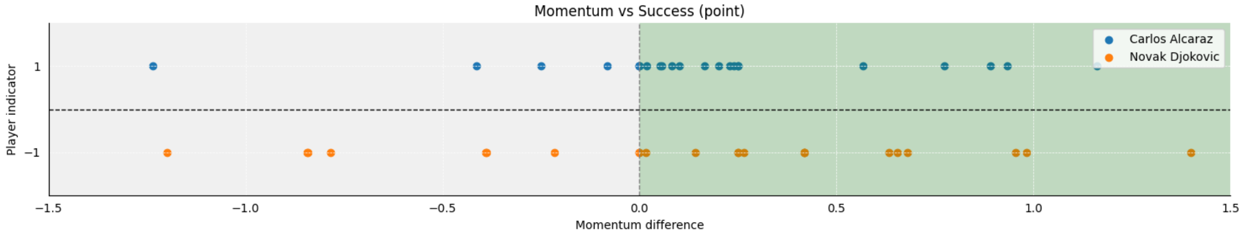


Figure 12: Momentum Difference and Success (Game)

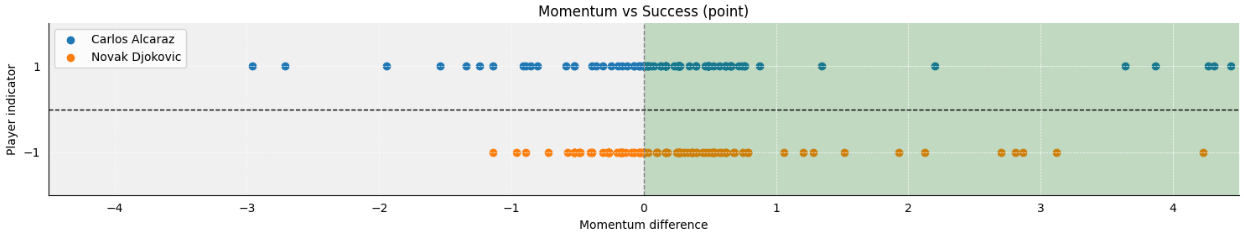


Figure 13: Momentum Difference and Success (Point)

Layer	Accuracy	Precision	Recall for Player A	Recall for Player B
Game	0.761	0.731	0.826	0.695
Point	0.706	0.689	0.778	0.631

Table 6: Binary Classification Evaluation for Using Momentum to Predict success

As the four figures shows, most the indicator points fell into the right prediction area. The value in the confusion matrix further confirms our conclusion that momentum plays an important role in tennis matches, helping people predicting the swings in play and runs of success by one player. Hence, we claim that the tennis coach is wrong.

4 Problem 3

4.1 Analysis of the Problem

To see what factors (other than momentum we demonstrated in the previous question) are most related with swings in the match, we need to develop models to predict the swings in the match and figure out the factors important in our prediction models. To achieve this, we developed 4 common machine learning models which are usually used by people for classification prediction tasks. Furthermore, we compared the performance of those 4 models, and combined those models by taking the average of factor importance values generated by them and select the factors with large importance values.

In the second part of this problem, we need to look into the detailed data of one player's "momentum" swings during past matches, and give suggestions to the player. We first analyzed the problem by investigating the momentum gap, clutch points and break points to identify the critical moments of the match for the player and thus make some suggestions. Then, we also looked into the performance data of a player such as return depth, serve width, number of ACE and so on, to deliver more personalized advise.

4.2 Prediction Models

4.2.1 Data Processing

Considering that the swing data exhibits a significant skewness due to the imbalance between (0,1), which is not conducive to training, we utilized the "point_vector" as the dependent variable. If find that certain features have a substantial impact on the prediction of "point_vector", it naturally show their significance to the occurrence of swing points, as swings fundamentally represent the outcomes at a given point in a game.

The factors we chose to construct the independent variable is listed in the table below:

Chosen Data
server
ace
winner
double_fault
unf_err
net_pt
net_pt_won
break_pt
break_pt_won
break_pt_missed
distance_run
rally_count

Table 7: Data Chosen in Training the Prediction Model

Server denote for the player who is serving in the game. Other chosen data are taken both from player A and player B. We took the average of those data, since the data can represent the average performance of each player in this way. We divided the data set into two parts for machine learning process. 0.8 of the dataset becomes the training set while 0.2 of it becomes the testing set.

4.2.2 Model Selection

To develop the prediction model, we chose four of the machine learning models: Random Forest, GBDT, XGBoost, and CatBoost to train with the dataset. Then, we combined the four models together by taking a weighted average.

The rationale behind selecting each model is that, the importance of features are based on the gain associated with a feature when it is used as a split node. The more frequently a feature is chosen as a split node and the greater the gain it brings, the higher its importance. These algorithms calculate the importance of features based on the structure and training process of decision trees. By tallying the number of times a feature is used as a split node, along with its gain and contribution to the loss function, we can evaluate the extent of each feature's contribution to the model's predictions and thus determine the importance of the features.

Combining the insights from these four models, we find that ensemble models can take the prediction results from multiple base models and aggregate them through weighting or voting, thereby achieving a more robust ranking of feature importance.

4.2.3 Result Analysis

The result of the four models are shown in the figure and table below:

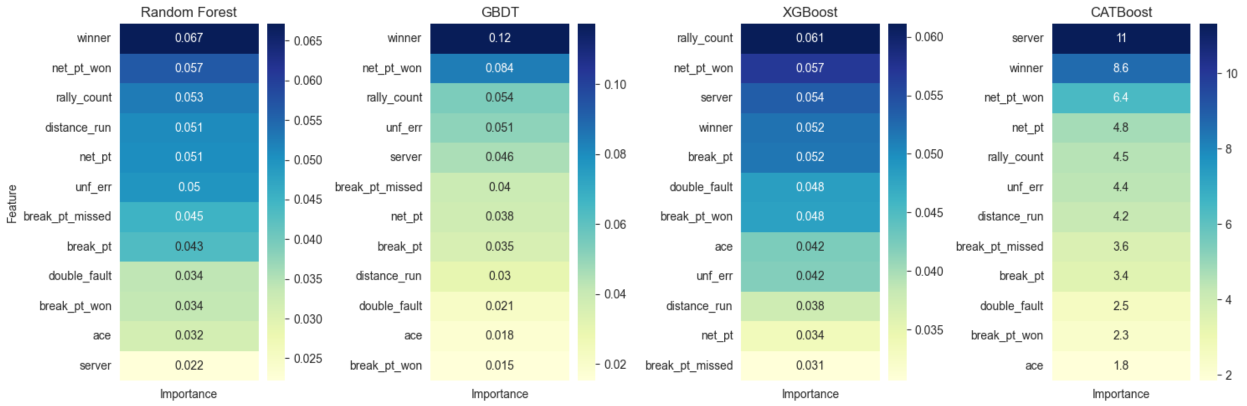


Figure 14: Results of the Machine Learning Models

Model Name	Testing Set Accuracy	Training Set Accuracy
Random Forest	0.687	1.000
GBDT	0.597	0.996
XGBoost	0.641	0.993
CATBoost	0.671	1.000

Table 8: Data Chosen in Training the Prediction Model

As the highest test set accuracy of the four models has reached 0.67, it means that the models are appropriate to be used in building a prediction model. We normalize the feature importance of each model into [0, 1]. We build our model by taking the weighted average of the four models, using the accuracy of each model as the weight distributed to it. The final result is shown in the table below:

Feature	Importance
winner	0.378158
net_pt_won	0.305174
server	0.270220
rally_count	0.247715
unf_err	0.218660
net_pt	0.200876
break_pt	0.191984
distance_run	0.189264
break_pt_missed	0.178621
double_fault	0.149495
break_pt_won	0.138881
ace	0.127967

Table 9: Data Chosen in Training the Prediction Model

4.3 Advise for the Player

4.3.1 Suggestion 1

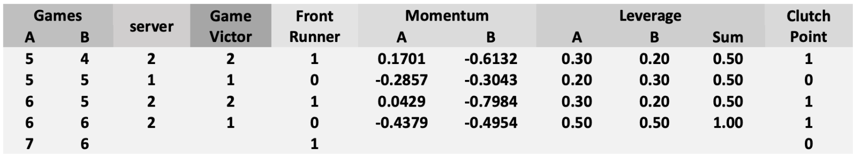
We first define the **clutch point**s as the moments when the sum of leverages of two players is larger or equal to 0.5. The clutch point is an important moment for the match because two players winning or losing a game (or point) can change the players' winning probability, and thus greatly affects the winner of the match.

If we are given the data for the match of Alcaraz (player A) and Djokovic (player B), and we need to make suggestions for two players for their next match, we can analyze the past match for them in this way.

- There are four clutch points in the match, three of which is in set 2 as shown in Figure 15. Thus we think performing well in set 2 is very important for winning the whole match.
- Furthermore, observing the data of set 2 (Figure 15), the first clutch point occurs in the 10-th game with 5:4. So we think this as the important moment in set 2.
- Next, observing the data of the 10-th game (Figure 16), all of the leverage sums are smaller than 0.5, indicating that there's no clutch point. Besides, there's no break point. The largest leverage sum is 0.4615 when it is 30:30 or 40:40. At those two time, winning one more point will influence the whole game greatly.
- At 30:30, the momentum of player A is larger than the momentum of player B, so it is very likely that player A will win the point. While at 40:40, the situation is exactly the opposite.

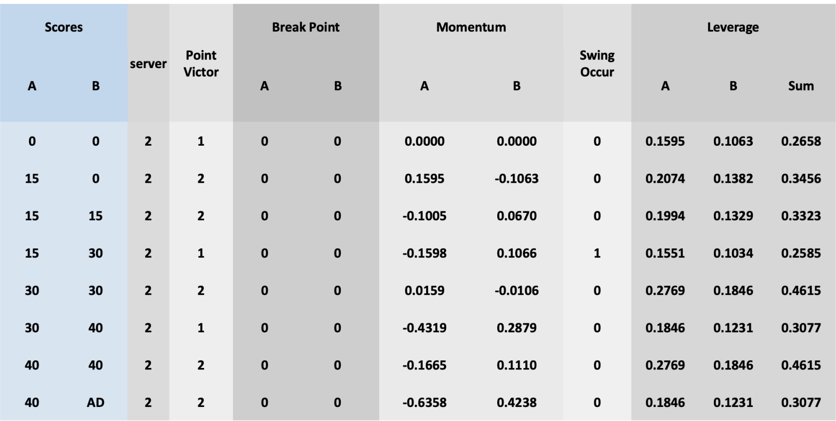
After identifying the key moments in the match, we strongly suggest that at those key moments, players consider taking a rest utilizing the single chance of bathroom. Besides, players should be clear that they are ought to try their best to win the next point. They can also cool down and analyze the situation seriously to know how to win the next point. However, they can never be nervous at those moments; instead, they should encourage themselves more.

Though the above is the suggestion for the last match, it can be an experience to learn for the next match. They should feel what time is the key moments in the new match and take those suggestions.



Games		server	Game Victor	Front Runner	Momentum		Leverage			Clutch Point
A	B				A	B	A	B	Sum	
5	4	2	2	1	0.1701	-0.6132	0.30	0.20	0.50	1
5	5	1	1	0	-0.2857	-0.3043	0.20	0.30	0.50	0
6	5	2	2	1	0.0429	-0.7984	0.30	0.20	0.50	1
6	6	2	1	0	-0.4379	-0.4954	0.50	0.50	1.00	1
7	6			1						0

Figure 15:



Scores		server	Point Victor	Break Point		Momentum		Swing Occur	Leverage		
A	B			A	B	A	B		A	B	Sum
0	0	2	1	0	0	0.0000	0.0000	0	0.1595	0.1063	0.2658
15	0	2	2	0	0	0.1595	-0.1063	0	0.2074	0.1382	0.3456
15	15	2	2	0	0	-0.1005	0.0670	0	0.1994	0.1329	0.3323
15	30	2	1	0	0	-0.1598	0.1066	1	0.1551	0.1034	0.2585
30	30	2	2	0	0	0.0159	-0.0106	0	0.2769	0.1846	0.4615
30	40	2	1	0	0	-0.4319	0.2879	0	0.1846	0.1231	0.3077
40	40	2	2	0	0	-0.1665	0.1110	0	0.2769	0.1846	0.4615
40	AD	2	2	0	0	-0.6358	0.4238	0	0.1846	0.1231	0.3077

Figure 16:

4.3.2 Suggestion 2

After analyzing the momentum, we looked into specific factors in order to conduct more personalized analysis.

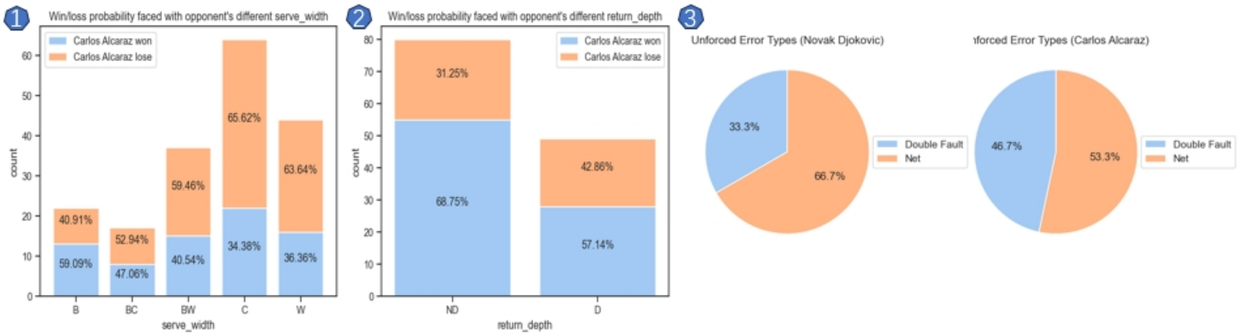


Figure 17: Sample Performance Factors for Alcaraz and Novak Djokovic

Figure 17 shows three examples of the factor we investigated in. From the figure on the left, we observe that Alcaraz has a noticeably lower success rate when returning the opponent's first serves directed towards the center of the court (C) and towards the sidelines (W), indicating a weakness in returning serves in these two position.

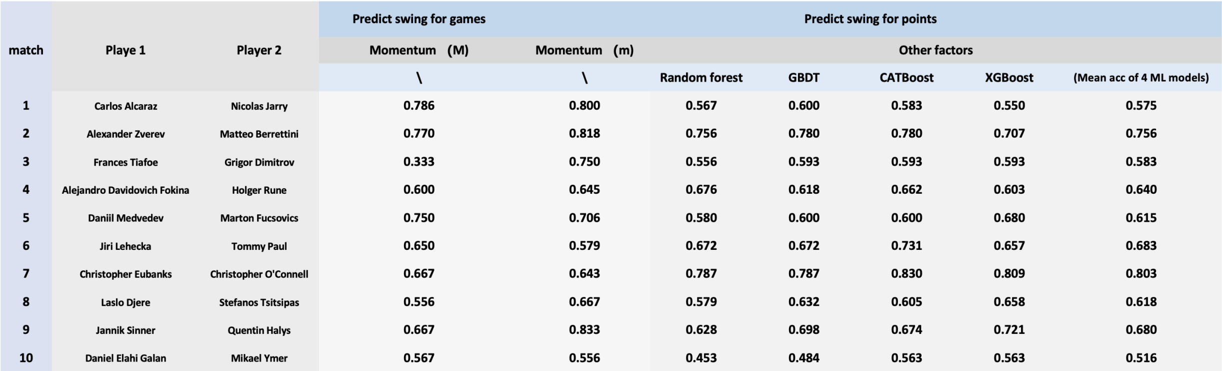
From the figure in the middle, we observe that Alcaraz has a significantly lower success rate when receiving deep returns (D) from the opponent, indicating his weakness in hitting back deep return.

From the figure on the right, we observer that the success rates of the two players are similar with close averages for different rally counts. This indicates the two players may have comparable endurance.

Considering the indications shown in Figure 17, we would advise Alcaraz to pay more attention on the center of the court and the sidelines and be well-prepared for deep returns. This approach can be utilized to analyze a player from other various aspects as well.

5 Problem 4

5.1 Performance Analysis on other Matches



match	Playe 1	Player 2	Predict swing for games	Predict swing for points					
			Momentum (M)	Momentum (m)	Other factors				
			\	\	Random forest	GBDT	CATBoost	XGBoost	(Mean acc of 4 ML models)
1	Carlos Alcaraz	Nicolas Jarry	0.786	0.800	0.567	0.600	0.583	0.550	0.575
2	Alexander Zverev	Matteo Berrettini	0.770	0.818	0.756	0.780	0.780	0.707	0.756
3	Frances Tiafoe	Grigor Dimitrov	0.333	0.750	0.556	0.593	0.593	0.593	0.583
4	Alejandro Davidovich Fokina	Holger Rune	0.600	0.645	0.676	0.618	0.662	0.603	0.640
5	Daniil Medvedev	Marton Fucsovics	0.750	0.706	0.580	0.600	0.600	0.680	0.615
6	Jiri Lehecka	Tommy Paul	0.650	0.579	0.672	0.672	0.731	0.657	0.683
7	Christopher Eubanks	Christopher O'Connell	0.667	0.643	0.787	0.787	0.830	0.809	0.803
8	Laslo Djere	Stefanos Tsitsipas	0.556	0.667	0.579	0.632	0.605	0.658	0.618
9	Jannik Sinner	Quentin Halys	0.667	0.833	0.628	0.698	0.674	0.721	0.680
10	Daniel Elahi Galan	Mikael Ymer	0.567	0.556	0.453	0.484	0.563	0.563	0.516

Figure 18: Model Accuracy on other matches

When testing the model on other matches, the overall performance is good, but there have been instances where predicting swings has shown varying levels of success.

The model 3 performs poorly in predicting swings at the game level (only 0.333) due to the small number of games and high volatility. For instance, the presence of only one swing at the game level undermines the effectiveness of the metric used for evaluation. Additionally, the small sample size of 134 points in this particular match may not provide sufficient information to accurately capture the underlying patterns and dynamics that contribute to swing predictions, thus resulting in subpar performance.

5.2 Factors for Future Model Improvement

Through weighted fusion machine learning, we found that "server" ranks first in the importance of learned features. We can consider increasing the weight of the "server" feature in predicting swings. For matches with low occurrences, it is advisable to incorporate additional factors such as:

- Injury status: Whether a player is currently injured or has a history of injuries, as such conditions can significantly impact their performance and physical condition during the match.
- Technical characteristics: A player's individual technical traits and style, including offensive and defensive capabilities, as well as their overall ball sense.
- Mental state: The player's psychological state in historical matches, including confidence, ability to handle pressure, and match status.

- Opponent strength: the level of skill and tactical approach adopted by the opponent.
- Game environment: Factors such as the location of the match, weather conditions, and audience support may also affect the game.

By incorporating these factors into future models, we aim to enhance the evaluation process and improve the accuracy of swing predictions.

5.3 Model Generalizability

5.3.1 Women's matches

Our model has high generalizability due to the similarity in game rules.. However, we need to make the adjustment of model parameters to accommodate different genders in tennis matches:

- Match Duration: Considering that women's tennis matches typically last shorter, we can decrease the weight of endurance.
- Serve Speed and Power: Considering Male players generally have higher serving speed and power, we can lower the weight of speed parameters.

. It can enhance model's adaptation to the characteristics of men's and women's tennis matches, and improving the accuracy and reliability of predictions.

5.3.2 Court Surfaces

Each surface type has distinct characteristics that can affect factors such as ball speed, bounce, and elasticity. Different players may perform better on different surface types, making it an important factor to consider when predicting match outcomes. Due to the lack of consideration of new key factors, the generalization ability of our model has decreased.

5.3.3 Other Sports

Since our specific model is designed based on the complex game_set_match progression structure of tennis, it may not be directly applicable to other sports at the moment. However, considering different levels of performance and incorporating the concept of momentum into the design are ideas that are definitely worth considering and can be applied in various contexts.

6 Strengths and Weaknesses

6.1 Strengths

The model's strengths lie in

- Incorporation of Momentum: The model's ability to incorporate momentum is a significant strength, as it allows for accurate estimation of swing and success runs. This can provide valuable insights into the dynamics of a tennis match and inform strategic decisions.

- Interpretability: High interpretability is another strength of the model, allowing coaches and spectators to understand how the model arrives at its predictions. This enhances transparency and trust in the model's outputs.
- Reasonable Level of Generalization Capability: The model's reasonable level of generalization capability is another strength, as it allows for accurate prediction across different tennis match scenarios and players.
- Ensemble Models for Feature Importance Ranking: The utilization of ensemble models for feature importance ranking enhances the robustness of the results, reducing the impact of noise or outliers in the data. This can improve the accuracy and reliability of the model's predictions.
- Real-time Application Potential: The model's potential for real-time predictions during live matches is a significant advantage, offering value for coaches and spectators alike. This can enhance decision-making during the match and deepen spectators' understanding of the game.

6.2 Weaknesses

However, the model has some weaknesses that should be considered:

- Susceptibility to Low Match Numbers: The model may be influenced by factors related to a low number of matches, such as a limited number of swings, which could potentially make the metric less effective.
- Cold Start Issues: Initially, the model relies solely on the server's win rate without considering the abilities of both players, leading to cold start issues. In the early stages where historical data is scarce, momentum may not accurately reflect the current state of the game.
- Data Availability: It may heavily rely on the availability and quality of historical match data, which could limit its effectiveness in certain contexts.

These weaknesses suggest areas for improvement in future research and model development. Additional strengths and weaknesses could be identified through further analysis and testing of the model in different scenarios and datasets.

7 Conclusion

In this study, we successfully constructed a definition of momentum and conducted in-depth analysis of various features through carefully designed feature engineering. By using ensemble models to sort feature importance, we successfully utilized momentum and features to predict the occurrence of swing in tennis matches, achieving an accuracy rate of approximately 75%. These results provide coaches with reference guidance for real-time decision-making and provide valuable information for the gambling industry to calculate player winning rates.

Our definition of momentum is based on a novel method of weighting historical matches (especially key wins in each game). However, we also recognize that due to the model being built on the consideration of complex rules in tennis, some generalization ability has been lost. This needs to be further explored and improved in future research.

Overall, this study provides new ideas and methods for the use of machine learning technology to predict results in tennis matches. Our research results not only provide valuable decision support for sports coaches but also provide more accurate tools for the gambling industry to calculate winning rates. Future research can further expand the scope of the model's application, improve generalization ability, and explore more potential applications of momentum in sports matches.

8 Memorandum to the Coach

Memorandum

To: Mr. Coach

From: 2024 MCM Team

Subject: Role of Momentum in Tennis Matches

Date: March 29, 2024

Dear Mr.Coach,

In the realm of tennis,"momentum" has been a topic of ongoing discussion for a considerable period of time. It is understandable that some professionals might be skeptical about the role of "momentum" in tennis matches, since it does seem a bit mystical at first glance. However, after thorough research and data modelling process, we observe that relativity exists, to a certain degree, between "momentum" and the flow of the match. we are here pretty glad to have the opportunity to introduce our research finding and suggestions to you, with the hope that it may give you some insight into future training strategies.

(1) Pay attention to players' performance at the break point and try the best to avoid error: We aim to investigate factors influencing tennis match performance in problem 1. The analysis revealed that assigning higher weights to break point won(or missed), unforced error, and double fault effectively make the model to better correlates with the actual change of the player's score. Therefore, we would like to recommend coaches to prioritize these indicators' impact during matches.

(2) Momentum can predict swing occurrences during the matches: In problem 2 and 3, we built a machine learning model to predict swing and success using momentum, the result turned out to be in high accuracy. Therefore, we suggest the coach to consider this information and better prepared for the change of the score.

(3) Momentum data in past match can be utilized to get a better preparation for future game: In problem 3, we tested our model using the data of the player's past match, and find it still showing

high accuracy in predicting the result. Therefore, we suggest the coach to adjust tactics or strategies to better respond to changes during or before the game.

We really appreciated this opportunity to give you some suggestions base on our findings. We are convinced that our analysis of the data and suggestions can be utilized to the improvement of future training process. If you have any questions, please feel free to contact us!

Best regards,

MCM 2024 Team

References

[1] Seidl, Robert & Lucey, Patrick. (2022), Live Counter-Factual Analysis in Women's Tennis using Automatic Key-Moment Detection.

[2] Barnett, Tristan & Clarke, Stephen. (2005). Combining player statistics to predict outcomes of tennis matches. IMA Journal of Management Mathematics. 16. 113-120. 10.1093/imaman/dpi001.

[3] Braidwood, J. (2023), Novak Djokovic has created a unique rival – is Wimbledon defeat the beginning of the end, The Independent, https://www.independent.co.uk/sport/tennis/novak-djokovic-wimbledon-final-carlos-alcaraz^b2376600.html.

[4] https://www.merriam-webster.com/dictionary/momentum

[5] Rivera, J. (2023), Tennis scoring, explained: A guide to understanding the rules terms & point system at Wimbledon, The Sporting https://www.sportingnews.com/us/tennis/news/tennis-scoring-explained-rules-system-pointsterms/7uzp2evdhbd11obdd59p3p1cx.