
Momentum Matters!

Summary

The incredible swings, sometimes for many points or even games, that occurred in the player who seemed to have the advantage are often attributed to "momentum". In this paper, we establish several models to quantify the value of momentum and assess the impact of momentum on the game. This model can be used for reference by coaches to guide athletes to achieve better performance in competitions.

For problem 1. Firstly, we go on the **data preprocessing**, dealing with missing value and non-numeric data. Then we regard the value of momentum as the key to quantify the performance in every moment. So we select 11 dominant parameters and employ **PCA** to derive 4 main components. Further we utilize **judgement matrix** to gain the weight for each element. Finally we compute the value of momentum and take the difference of two players' momentum as the indicator to illustrate the match's flow with a chart. Besides, we compare it with another model, which judges solely based on score difference, thus highlighting the practicality of our momentum model.

For problem 2, We first use the main variables of question 1 to construct a **Binary Logistic Regression Model**, and obtain the score probability of player1 at each time. Then we draw the scatter plot with the momentum value of question 1 as the abscissa and the score probability as the ordinate, and intuitively see the relationship between momentum and score probability. We analyzed Momentum and Scoring Trend using the **Spearman Correlation Coefficient**, and found a high correlation of 80.2% between the two. We can answer coach that, the swings in match and runs of success by one player are not random. momentum really counts in that.

For problem 3, first we conduct **data mining** based on known variable and create 19 new traits that better evaluate players' performance in the match. Then we combine the original variables with the new ones obtained through data mining to construct the **Random Forest Model**, and get the winning probability curve. We put the probability and momentum value in a graph, and define the maximum or minimum point of the curve as the turning point of the game. It can be obviously seen that momentum has a great positive impact on the turning point of the match. Then we screen out 10 factors that have the greatest impact on the turning point according to the feature importance value of the Random Forest Model. The dominant factor is service order. According to the conclusions above, we have offered suggestions to players in the paper.

For problem 4, based on the previous Random Forest Model, we have analyzed many other matches. In most cases, the results match expectations. According to the observation, the large deviation often occurs when the competition time is too long or too short. In addition, it is also affected by the individual characteristics. In the future model, we will take into account the physical parameters of the players, the rest time during the game, etc. so as to make the results more accurate. The AUC value is 0.78, which reveals the strong generalization ability.

Keywords: PCA, Binary Logistic Regression, Random Forest, Spearman

Contents

1	Introduction	3
1.1	Problem Background	3
1.2	Restatement of Problem	3
1.3	Our Work	4
2	Assumptions and Justifications	4
3	Notations	5
4	Data Preprocessing	5
4.1	Basic Data Processing	5
4.2	Data Mining	6
5	Task1: Momentum Calculation	6
5.1	PCA	6
5.2	Weight Calculation	8
5.2.1	Building Judgment Matrix	8
5.2.2	Working Out Weight	8
5.3	Momentum Calculation	9
5.4	Visualization	9
5.5	Other Model for Comparison	10
5.6	Conclusion	10
6	Task2: Score Probability Prediction	11
6.1	Binary Logistic Regression	11
7	Task3: Turning Point Analysis	12
7.1	Random Forest Model	12
7.2	Advice to Players	14
8	Task4: Model Extension	15
9	Sensitivity Analysis	15
10	Model Evaluation	17
10.1	Strengths	17
10.2	Weaknesses	17
11	Conclusion	17
	Memo	19
	References	20
	Appendices	21
	<u>Report on Use of AI</u>	23

1 Introduction

1.1 Problem Background

The 2023 Wimbledon Gentlemen's final witnessed the ending of glorious run for Djokovic, who is among the all-time great players in Grand Slams. Undoubtedly, the battle between Carlos Alcaraz and Djokovic, was rather a fabulous match, filled with ups and downs. Initially Djokovic gained dominance over Alcaraz with 6-1, but in the next two sets, Alcaraz clinched victory with 7-6 and 6-1 unexpectedly. Then another reverse occurred. Djokovic won the fourth set and led the match to a tie. The finals ended with Alcaraz's victory. It did arouse our curiosity that why situation of game changes so frequently and what is the possible factor that affects the trend of match.

1.2 Restatement of Problem

There is a conception named "momentum" in tennis, which has a great impact on players' performance. It is a generalization of the influence in manifold aspects like mental stress and residual energy. So the fluctuation of momentum is the most probable factor that reveals the trend of match. However, the momentum is not easy to be quantified for it includes many subjective indicators, and there are few models that can be directly used^[4], so we decide to cut in the following questions from statistical analysis and data procession:

- Develop a model that captures the flow of play as points occur and apply it to one or more of the matches. Your model should identify which player is performing better at a given time in the match, as well as how much better they are performing. Provide a visualization based on your model to depict the match flow. *Note: in tennis, the player serving has a much higher probability of winning the point/game. You may wish to factor this into your model in some way.*
- A tennis coach is skeptical that momentum plays any role in the match. Instead, he postulates that swings in play and runs of success by one player are random. Use your model/metric to assess this claim.
- Coaches would love to know if there are indicators that can help determine when the flow of play is about to change from favoring one player to the other.
 - 1) Using the data provided for at least one match, develop a model that predicts these swings in the match. What factors seem most related (if any)?
 - 2) Given the differential in past match momentum swings how do you advise a player going into a new match against a different player?
- Test the model you developed on one or more of the other matches. How well do you predict the swings in the match? If the model performs poorly at times, can you identify any factors that might need to be included in future models? How generalizable is your model to other matches (such as Womens matches), tournaments, court surfaces, and other sports such as table tennis.
- Produce a report of no more than 25 pages with your findings and include a one-

to two-page memo summarizing your results with advice for coaches on the role of momentum, and how to prepare players to respond to events that impact the flow of play during a tennis match.

1.3 Our Work

We preprocess the data provided before analyzing matches and mainly utilize 3 models. The specific work was as follows:

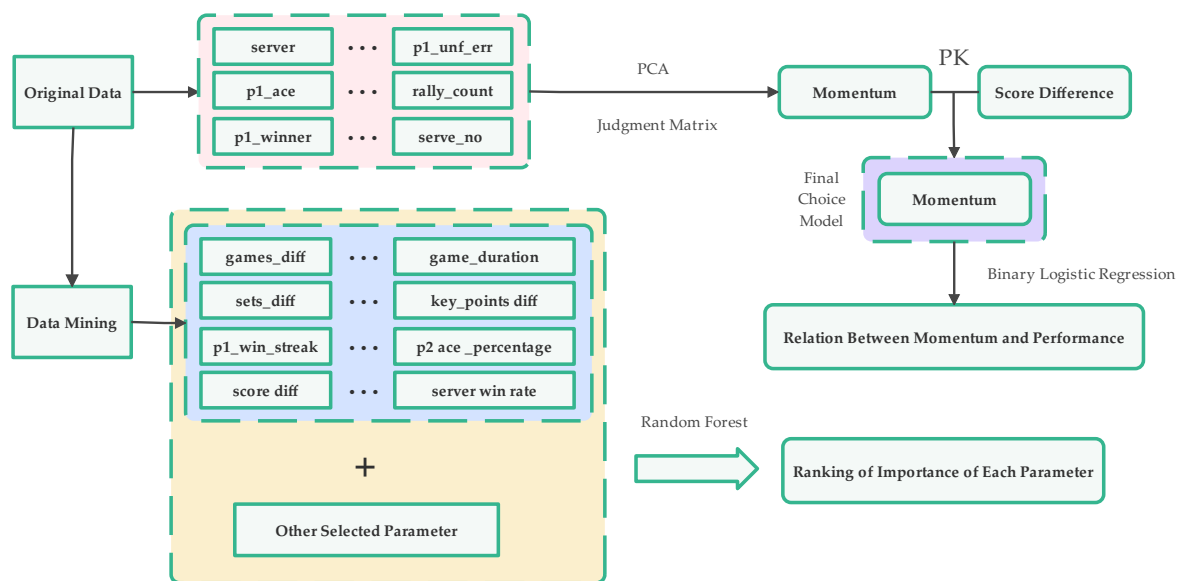


Figure 1: Flow Chart of Our Work

2 Assumptions and Justifications

Through the full analysis of the problem, in order to simplify our model, we make the following reasonable assumptions.

Assumption 1 *Athletes' physical parameters, such as endurance and bodily functions, show no significant differences.*

The data obtained is specific to parameters for each serve in a match, forming the basis for our model. Our model relies solely on data provided, and introducing adjustments based on individual athletes' physical parameters would be overly subjective and challenging to quantify, due to the inherent difficulty in measuring physical parameters.

Assumption 2 *Matches are independent of each other.*

Athletes' performance in previous matches does not affect their performance in the current match, thereby not influencing predictions. This is attributed to the substantial interval between matches, making it unlikely for short-term psychological changes to occur. Additionally, the model's training data is match-centered, and combining

data from different matches for training would introduce bias due to varying match parameters (timing, athlete skill levels, etc).

Assumption 3 *Athletes' match performance is influenced by specific serving behaviors, such as achieving aces, double faults, or unforced errors, resulting in corresponding positive or negative effects.*

Incorporating these factors allows for the consideration of an athlete's psychological state in enhancing model's comprehensiveness.

Assumption 4 *Ignore any effect of coach guidance on athletes' performance.*

The provided dataset lacks information on coaches' actions, and quantifying the influence of coaches on athletes is impractical, making it unsuitable for inclusion in model analysis.

Assumption 5 *Exclude the impact of audience reactions and controversial referee decisions on athletes' performance.*

The dataset does not provide relevant information on these factors, and their inherent randomness, coupled with their opaque psychological impact on athletes, makes them unsuitable for consideration in model analysis.

3 Notations

Important notations used in this paper are listed in Table 1,

Table 1: Notations

Symbol	Definitions	Unit
x	Sample matrix	-
R	Sample correlation matrix	-
ω	The weight of the principal component	-
λ_{\max}	Maximum eigenvalue	-
CI	Consistency index	-
RI	Random consistency index	-
CR	Consistency ratio	-
P	Posibility that player1 scores	-
θ_i	Logistic regression coefficient	-

4 Data Preprocessing

4.1 Basic Data Processing

We hope to unify the dimensions of the selected variables by means of data preprocessing, so as to facilitate the further construction of the model.

- In the column of data for the non-numeric variable return_depth, convert values such as "AD,D,NA" to the numbers 0,1,2, etc.
- In the columns p1_score and p2_score, replace the scores "0,15,30,40,AD" with 0,1,2,3,4.
- For missing data in the table, the forward fill method is used.

4.2 Data Mining

We created 18 new features based on the original variables to be used when modeling with the Random Forest Model:

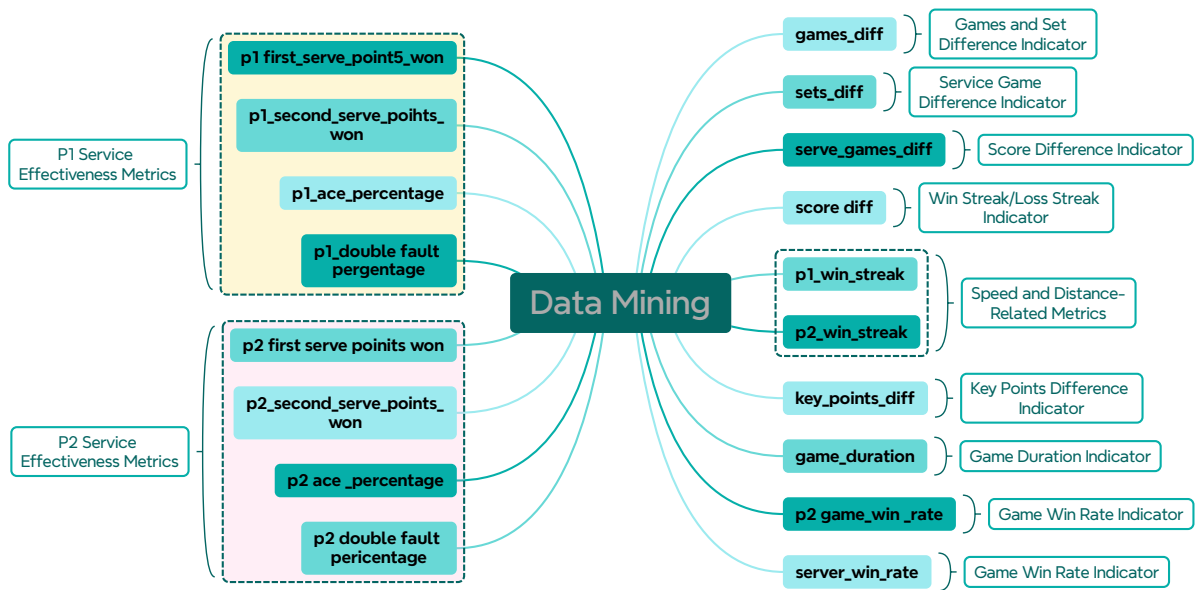


Figure 2: New Variables and Explanations

5 Task1: Momentum Calculation

5.1 PCA

First of all, according to our general consumption and public analysis on indicators^[1], we select 10 main variables that affect the momentum according to the topic, which respectively are: server, serve_no, p1_ace, p1_winner, p1_double_fault, p1_net_pt_won, p1_break_pt_won^[3], p1_break_pt_missed, p1_distance_run, rally_count, p1_unf_err. Then we build a sample matrix x of size 334×10 with these 10 variables as indicators and 334 moments data of the final as samples^[7]:

$$x = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,10} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,10} \\ \vdots & \vdots & \ddots & \vdots \\ x_{334,1} & x_{334,2} & \cdots & x_{334,10} \end{bmatrix} \quad (1)$$

According to formula(2), we calculate the sample correlation matrix R of matrix x

$$R = \frac{\sum_{k=1}^{334} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{334} (x_{ki} - \bar{x}_i)^2 \sum_{k=1}^{334} (x_{kj} - \bar{x}_j)^2}} \quad (2)$$

Then the eigenvalues of the matrix R can be calculated :

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{10} \geq 0 \quad (3)$$

Further, the eigenvectors are computed:

$$a_1 = \begin{bmatrix} a_{1,1} \\ a_{2,1} \\ \vdots \\ a_{334,1} \end{bmatrix}, a_2 = \begin{bmatrix} a_{1,2} \\ a_{2,2} \\ \vdots \\ a_{334,2} \end{bmatrix}, \dots, a_{10} = \begin{bmatrix} a_{1,10} \\ a_{2,10} \\ \vdots \\ a_{334,10} \end{bmatrix} \quad (4)$$

Finally, we acquire the principal component contribution rate $\frac{\lambda_i}{\sum_{k=1}^{10} \lambda_k} (i = 1, 2, \dots, 10)$ and the cumulative contribution rate $\frac{\sum_{k=1}^i \lambda_k}{\sum_{k=1}^{10} \lambda_k} (i = 1, 2, \dots, 10)$. We take the first 4 vectors, whose cumulative contribution rate is 57%, as the principal components. Their component matrix is shown in the following table

Table 2: Component Matrix

	a1	a2	a3	a4
server	0.348	0.270	-0.567	0.145
serve_no	0.000	0.251	0.394	0.000
p1_ace	-0.286	-0.523	0.000	0.458
p1_winner	0.000	-0.729	0.255	0.235
p1_double_fault	-0.359	0.403	0.520	0.205
p1_net_pt_won	0.114	-0.471	0.350	-0.469
p1_break_pt_won	0.108	0.000	-0.172	0.662
p1_break_pt_missed	0.112	0.220	-0.363	0.000
p1_distance_run	0.927	0.000	0.210	0.000
rally_count	0.915	0.000	0.308	0.000
p1_unf_err	0.000	0.555	0.505	0.207

It can be seen from the table above that a1 is mainly related to p1_distance_run and rally_count, so a1 is likely to represent the physical strength of players. We define a1 as Physical Strength Variable(PSV). a2 is mainly related to p1_winner and p1_ace, so a2 is likely to represent the skills of players. We define a2 as Technique Variable(TV). a3 is mainly related to server, serve_no, p1_double_fault and p1_unf_err and p1_break_pt_missed, so a3 is likely to represent the mentality of players. We define a3 as Mentality Variable(MV). a4 is mainly related to p1_net_pt_won and p1_break_pt_won, so a4 is likely to represent the motivation of players. We define a4 as Motivation Variable(MTV).

Table 3: Definition of Principal Components

Eigenvector	Definition	Abbreviation
a1	Physical Strength Variable	PSV
a2	Technique Variable	TV
a3	Mentality Variable	MV
a4	Motivation Variable	MTV

5.2 Weight Calculation

5.2.1 Building Judgment Matrix

After analyzing data and consulting experts, we scientifically and properly assign the relative rating and do a series of calculation. Finally we work out the following judgment matrix^[2]:

Table 4: Judgment Matrix

	PSV	TV	MV	MTV
PSV	1	2	2	3
TV	1/2	1	1/2	1/3
MV	1/2	2	1	2
MTV	1/3	3	1/2	1

5.2.2 Working Out Weight

According to $CI = \frac{\lambda_{\max} - n}{n - 1}$ and $CR = \frac{CI}{RI}$, we get $CR = 0.0976 < 0.1$, pass conformance test.

Table 5: Relationship Between n and RI

n	1	2	3	4	5	6	7	8
RI	0	0	0.52	0.89	1.12	1.26	1.36	1.41

Using eigenvalue method, the maximum eigenvalue of the judgment matrix is obtained by calculation, $\lambda_{\max} = 4.2606$. As a result, the corresponding eigenvector can be computed. We normalize the eigenvector and get the weight. The weight matrix is shown in the following table:

Table 6: Weight matrix of Principal Component Variables

Variable	PSV	TV	MV	MTV
Weight	0.4176	0.1242	0.2611	0.197

5.3 Momentum Calculation

We can estimate the value of momentum of player1 with formula below:

$$Momentum_1 = \sum_{i=1}^4 a_i \cdot \omega_{a_i} \quad (5)$$

According to the general consumption, we can also compute the momentum of player2 with formula above. We regard the difference of two momentum (p1 - p2) as the relative momentum of p1, and abbreviate it to momentum. We utilize the conclusion of question 2 in advance, which will be given proof soon that momentum have a great impact on the match. So the value fluctuation of the momentum reveals the swings of match, which is known as the flow of match.

5.4 Visualization

With elapsed time as horizontal coordinate and momentum value as vertical coordinate we fit the curve of the whole game in the chart below:

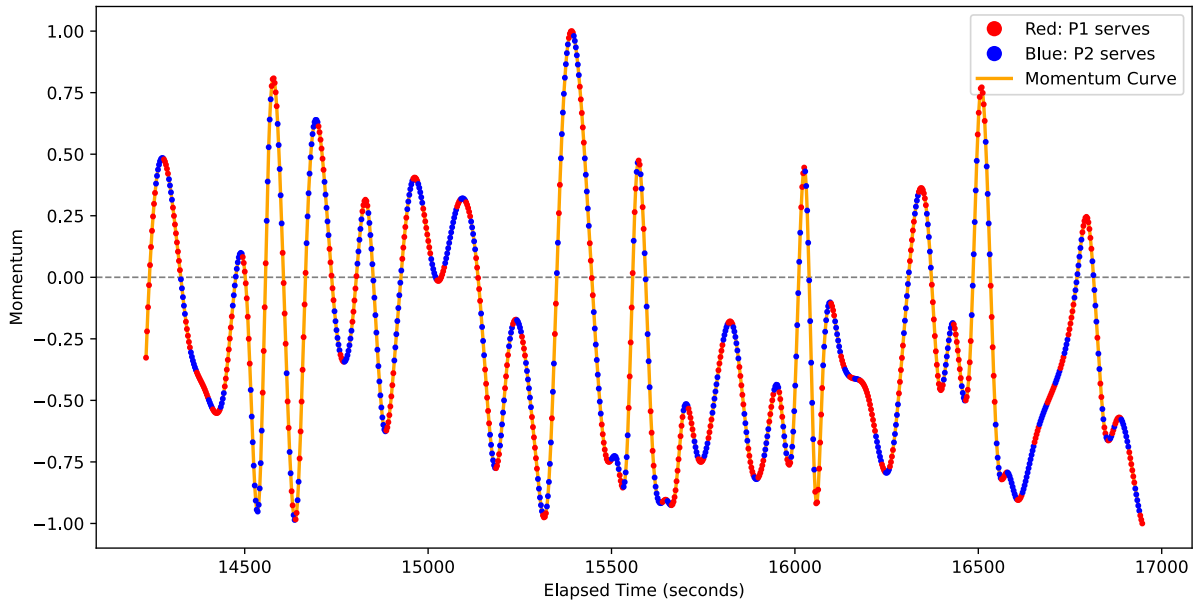


Figure 3: The Fluctuation of Momentum During the Game

The horizontal dotted line is the threshold that differs the two status. The momentum greater than 0 means p1's potential performance is better than p2's, and vice versa. Moreover, the absolute value of momentum can reveal how much better their performance compared to the other. This is how our model decide which one is potentially playing better. As you can see, to better showcase server's role in deciding momentum, we have marked the server with dots in two color (red for p1, blue for p2). It's obvious that red dots deliver more in upper area and blue dots deliver more in lower area, which means server does count, serving as a strong evidence for question 3.

5.5 Other Model for Comparison

To fully depict the degree of the difference between two players from a superficial perspective, we offer another model, deciding only by points. So If the model described above is a professional one with deep insight, this is simply a layman's perspective. A chart of the score difference ($p1 - p2$) in each game during a match is displayed below:

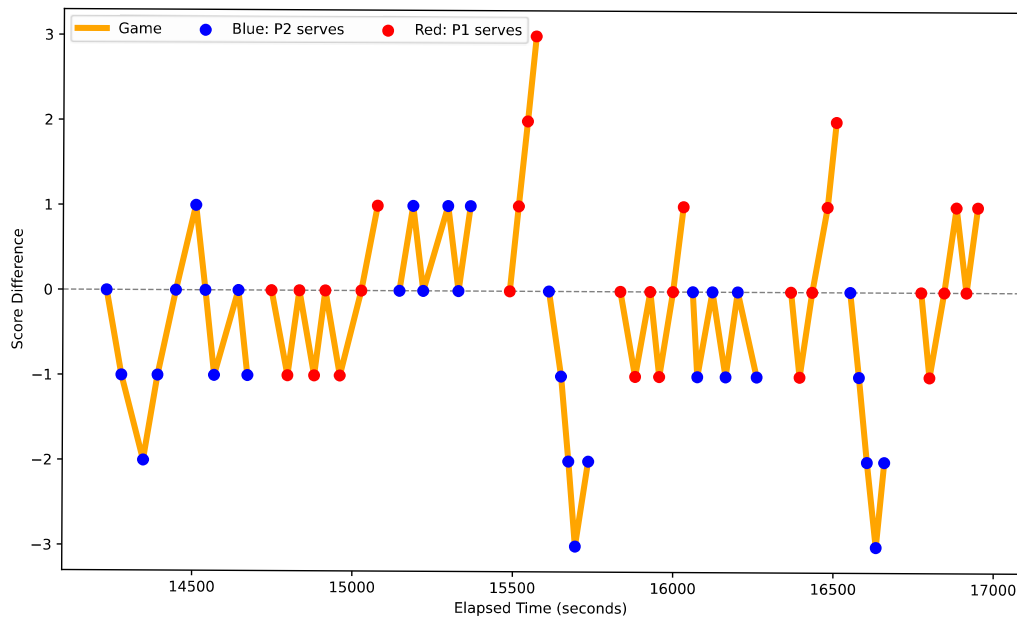


Figure 4: The Score Difference During the Game

As just shown, the score difference will be reset at the beginning of each game, only showing the performance in one game to prevent the influence by the past score. The method to judging is the same as the last model. Having tested many dataset, we learn that this model is unreliable to the question.

5.6 Conclusion

To draw a conclusion, in the preparation period, we adopt the PCA to reduce the number of parameters and then calculate weight as well as momentum. Further, we put forward two schemes, a profound one and a superficial one. After comparing the two model, we confirm our momentum model is better. It can precisely identify which player is performing better at a given time in the match, as well as how much better they are performing.

6 Task2: Score Probability Prediction

6.1 Binary Logistic Regression

We assume that P is the probability that player1 scores and $(1 - P)$ is the probability that player2 scores. $P/(1 - P)$ is the logistic regression ratio of students dropping out of class, and the natural logarithm is taken: $\ln [P/(1 - P)]^{[6]}$.

18 variables such as $Z_1(\text{server})$, $Z_2(\text{server_no})$, $Z_3(\text{p1_ace})$ are taken as independent variables, and P is taken as dependent variable. The logistic linear regression equation is

$$\ln [P/(1 - P)] = \theta_0 + \sum_{i=1}^{18} \theta_i Z_i \quad (6)$$

Where θ_i is the logistic regression coefficient.

According to the above formula, the probability of player1 scoring at each moment is:

$$P = 1 / \left[1 + \exp \left(-\theta_0 - \sum_{i=1}^{18} \theta_i Z_i \right) \right] \quad (7)$$

We take the momentum calculated in the first question as the horizontal coordinate and the P -value as the vertical coordinate to draw the scatter plot as follows

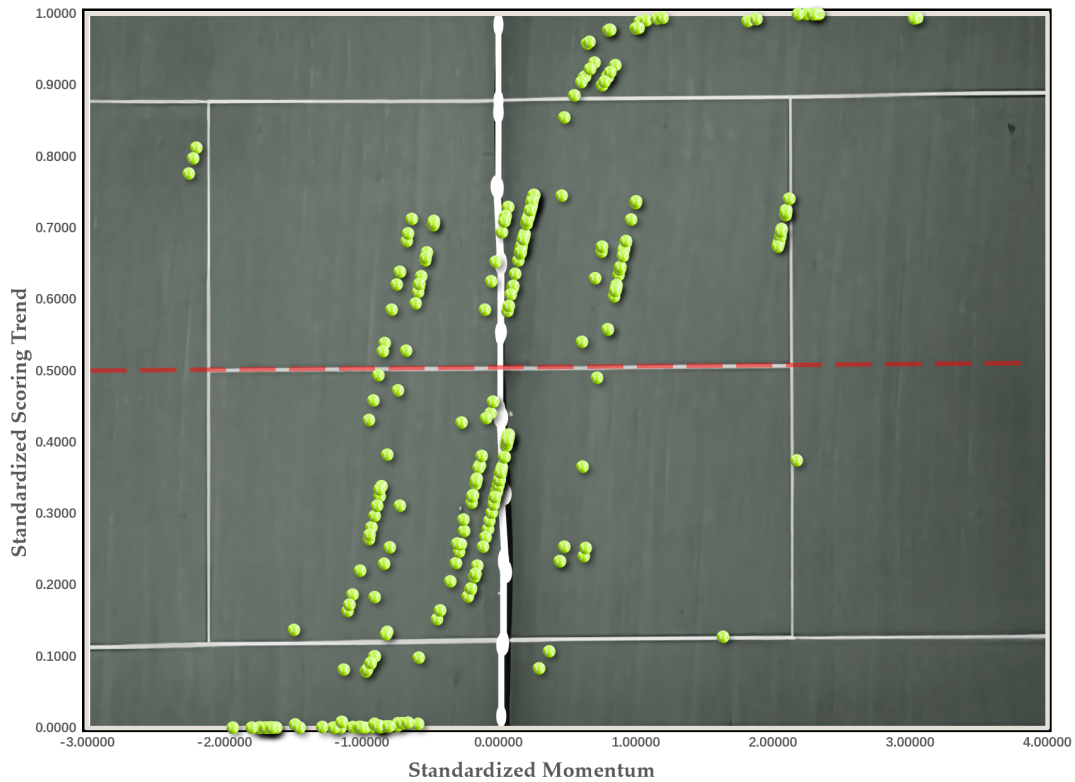


Figure 5: Scatter Plot of Relationship Between Momentum and Score Probability

As can be seen from the graph, if $p = 0.5$ and momentum = 0 are the boundaries, the coordinate graph can be divided into four quadrants. The points in the first quadrant

represent that player1 has a higher probability of scoring and has more momentum, and the points in the second quadrant represent that player1 has a higher probability of scoring but has less momentum. Points in the third quadrant mean that player1 has a lower probability of scoring and has less momentum, and points in the fourth quadrant mean that player1 has a lower probability of scoring but has more momentum.

It is not difficult to find that scattered points are mainly concentrated in the first and third quadrants, which indicates that the more momentum, the greater the score probability, and vice versa, that is, momentum has a significant positive relationship with the score probability. So we can answer coach swings in play and runs of success by one player are not random, momentum plays a big part in that.

We analyzed Momentum and Scoring Trend using the Spearman correlation coefficient, and found a high correlation of 80.2% between the two.

7 Task3: Turning Point Analysis

7.1 Random Forest Model

Random Forest is an ensemble learning model based on decision tree as the basic classifier. It includes several decision trees trained by Bagging Ensemble Learning technology. When the samples to be classified are input, the final classification result is decided by the output result of the individual decision tree.^[5]

In a Random Forest Model, feature importance measures the impact of each feature on the performance of the model. This metric is usually a number between 0 and 1, where 0 means that the feature has no effect on the model, and 1 means that the feature has a large effect on the model.

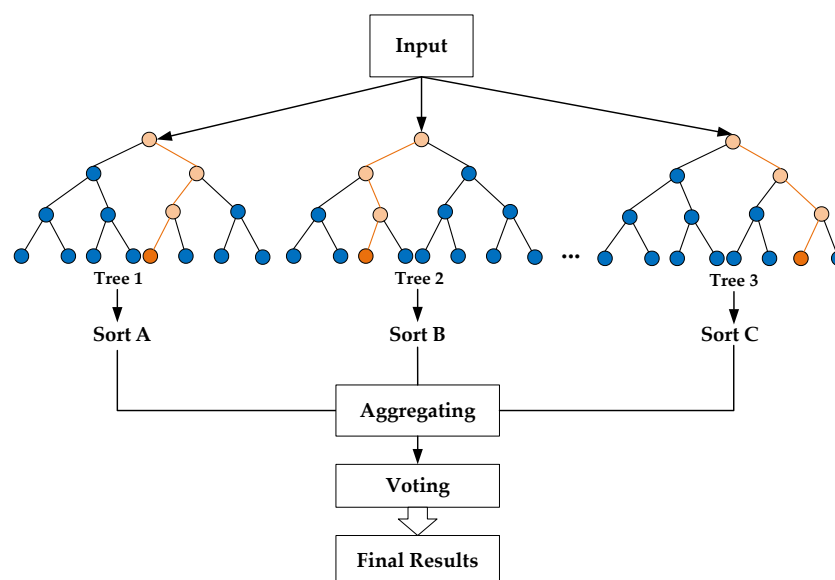


Figure 6: Schematic Diagram of Random Forest Model

We built the model by using 'games_diff', 'sets_diff', 'serve_games_diff' and other

indicators after data mining and the original indicators (49 in total), we get the score probability curve, we choose the maximum and minimum of the score probability curve as the turning point of the field. We plot the momentum curve obtained from the first question in a graph as follows:

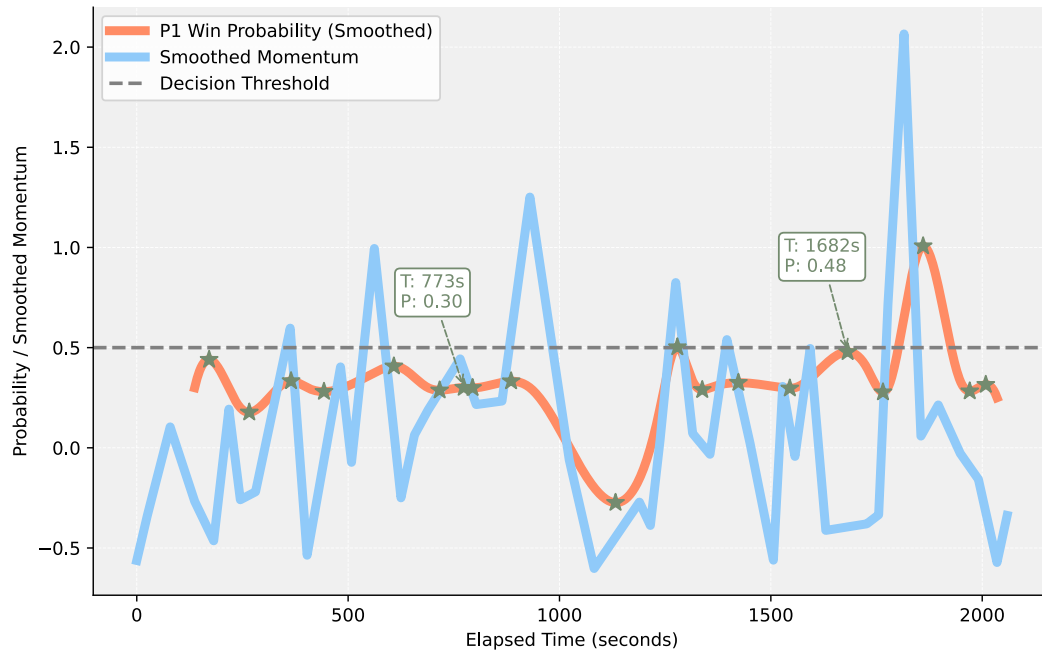


Figure 7: The Chart of Match Turning Point

As can be seen from the graph, the momentum curve is also at the extreme value at most turning points, and it can be concluded that momentum is closely related to the turning point of the game.

For this model: Accuracy = 0.9286, Precision = 1.0000, Recall = 0.8000, F1 Score = 0.8889, AUC = 0.9444, its ROC curve is shown in the figure below:

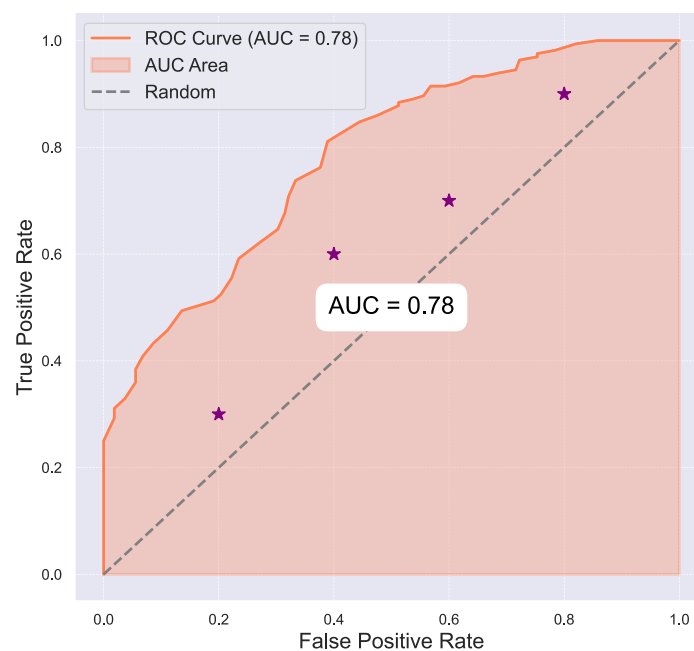


Figure 8: The Curve of ROC

Therefore, we can see that the model is very accurate and has strong generalization ability.

Then, by comparing the feature importance, we select 10 factors that have the greatest influence on the turn of the game.

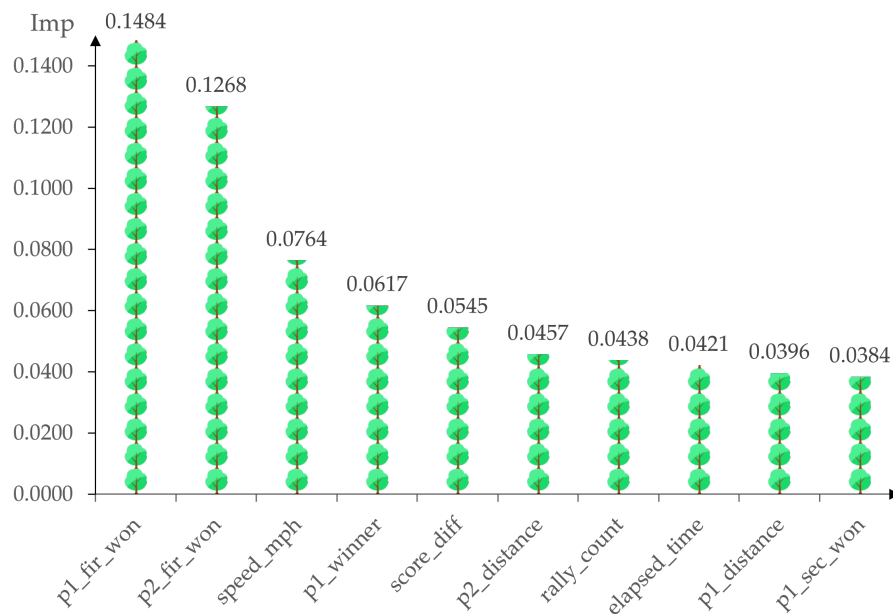


Figure 9: Bar Chart of Variables Ranked by Importance

According to the analysis of the data, we know that the biggest factor affecting the winning rate of the game is the service order, which means that the person who serves first has a greater advantage to win. This is consistent with our previous conclusions. In addition to this factor, the biggest impact is the running speed. Through our investigation and model prediction, the correlation between the momentum of the game and the change of the game situation is very large. Usually, the turning point of the game occurs in the process of a certain player's momentum change, and the win rate of the game will also change.

It is important to note, however, momentum does not correlate positively with changes in game winning percentage. Sometimes their momentum is better, but will increase the opponent's win rate; Sometimes the momentum of the opponent is low, and you do not necessarily have a chance to improve the win rate. However, according to the previous scatter plot analysis, when the potential energy is extremely high or low, the impact on the game situation is significant and positively correlated, while if the value is not extreme, the random component is larger.

7.2 Advice to Players

If you are at disadvantage in a match, it is necessary to take advantage of the opportunity of momentum change on both sides to improve your chances of victory. If the momentum of other party becomes higher, remember that this is also an opportunity for you to turn the tide and improve your chances of winning.

If you are superior players with a higher overall level, you also need to pay attention to potential energy and improve your potential energy to make yourselves compete in the best posture, which can significantly improve your winning probability.

8 Task4: Model Extension

Based on the Random Forest-based winning probability prediction model has just been built, we predict a number of other matches, which in most cases were more in line with the expectation, that is, the winning rate fluctuation and the potential energy fluctuation are consistent, but there would be some situations with large deviations from the expectation. According to the observation, the large deviation often occurs in the case of too long or too short competition time, in addition to the influence of different players, caused by personal characteristics.

According to the previous conclusion, the most influential factor in the winning rate of the game is service order, followed by service speed. It is speculated that this may be due to the different training results and personal qualities of each person, and the sensitivity to the service is different. In addition, due to the physical differences of players, the service speed will be greatly different, and the physical loss caused by the service speed is also different, so the accuracy of the prediction data of different games will be greatly different.

In the future model, we should take into account the physical parameters of the players, such as endurance, explosive power, etc. (data differences between different genders can also be analyzed from these perspectives) to balance the differences in model prediction performance caused by personal factors. At the same time, the rest time during the race is also a very important parameter that should be taken into account. The rest time is related to the time given in the question. An algorithm can be designed to quantify the rest time in the future.

Since the AUC value of our model is 0.94, the model has strong generalization ability. We have reason to believe that the model works for other matches such as Women's matches, tournaments, court surfaces, and other sports such as table tennis.

9 Sensitivity Analysis

Each path of a decision tree represents a decision process, but in practical applications, it is difficult to directly find how each branch of the tree structure affects the final prediction. Sensitivity analysis can help us understand the relative importance of each feature to the model output, providing interpretability for the model's decisions.

Sensitivity analysis helps to detect instability or bias in the model under specific input conditions, helping to identify possible problems and thereby improving the robustness of the model. We understand the contribution of each feature to each prediction, leading to a more complete understanding of model behavior.

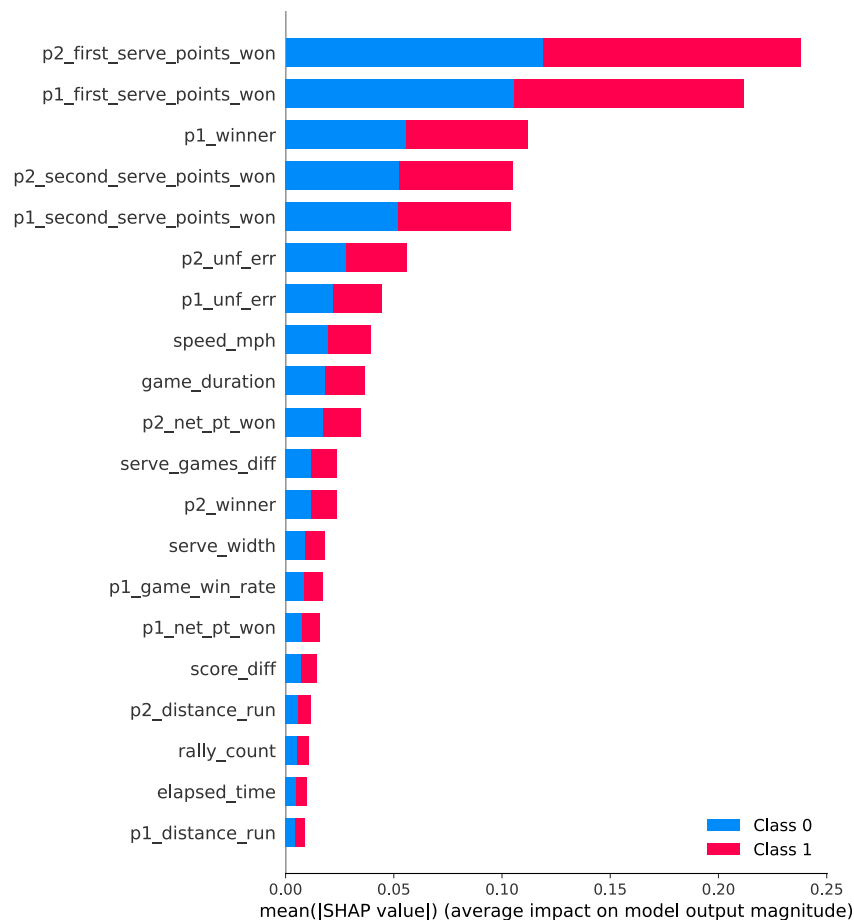


Figure 10: Sensitivity Ordering of Variables

This graph ranks the variables involved in Random Forest training in descending order according to their sensitivity. We can see that the variable that has the most significant impact on the model directly is p1/2_first_serve_points_won, which corresponds to our previous conclusion.

Class 0 and Class 1 in the legend represent the influence of variables on each result of the model in the binary classification process. Since our predicted odds of winning are mutually exclusive, the degree of influence on both sides is similar.

In addition, we also find that some factors affecting the psychological state of players, such as p1_winner, p2_unf_err and other variables, will also greatly affect the sensitivity of the model prediction, so it can be inferred that the change of athletes' mentality will also affect the winning rate of the game to a certain extent. This is also consistent with the influence factors of the Momentum model that we have previously obtained.

10 Model Evaluation

10.1 Strengths

1. PCA effectively reduces data dimensionality, eliminating redundancy and enhancing computational efficiency.
2. We used a large number of variables in building the Random Forest Model, including new variables that we built ourselves, making the results more robust.

10.2 Weaknesses

1. We only use the data from file "Wimbledon_featured_matches.csv", which may result in incomplete comprehensiveness.

11 Conclusion

- Task 1

We use the weighted assignment method to calculate the momentum value and use it to represent the player's performance. In order to compare the performance of two players more intuitively, we use the difference method to get the final relative momentum value.

From the final visualization, it can be intuitively seen that the momentum changes over time. The two players have the advantage almost alternately, and the height difference between the curve and the dividing line at each moment indicates the difference in the performance of the two players.

We use dots of different colors to represent the serving players at different times. It can be seen from the figure that the serving order has a great impact on the situation of the match.

- Task 2

We use binary logistic regression to calculate the score probability, and draw a scatter plot with momentum as the horizontal coordinate and score probability as the vertical coordinate. The whole plot is divided into four quadrants by the two dividing lines of score probability = 0.5 and momentum = 0. It can be intuitively seen from the graph that most points are concentrated in the first and third quadrants. This suggests that the higher the momentum, the greater the probability of winning, so we can tell skeptical tennis coaches that swings in play and runs of success by one player are not random.

- Task 3

We use the data variables obtained from data mining and the original data variables to build a Random Forest Model, fit the winning probability curve, and draw the momentum curve and the winning probability curve in a graph. We set the maximum or minimum point of the momentum curve as the turning point

of the game, and find that in most of these points, momentum curve is also in the extreme position, so we can judge that momentum value plays a crucial role in the turn of the match situation. We use the Random Forest Model to calculate the characteristic importance of variables, so as to judge the impact of variables on the turn of the match. The result shows that the service order has the greatest impact on the turn of the match. Second is the speed of the service.

Based on these findings, we recommend that players, regardless of their level, should pay attention to the influence of momentum in the game, reduce mistakes and seize opportunities when it is their turn to serve, because serving is more likely to affect the trend of the game.

- Task 4

Based on the Random Forest-based winning probability prediction model has just been built, we predict a number of other matches, which in most cases were more in line with the expectation, that is, the winning rate fluctuation and the potential energy fluctuation are consistent, but there would be some situations with large deviations from the expectation.

According to the observation, the large deviation often occurs in the case of too long or too short competition time, in addition to the influence of different players, caused by personal characteristics.

In the future model, we will take into account the physical parameters of the players, such as endurance, explosive power, etc. (data differences between different genders can also be analyzed from these perspectives) to balance the differences in model prediction performance caused by personal factors. At the same time, the rest time during the race is also a very important parameter that will be taken into account. The rest time is related to the time given in the question. An algorithm can be designed to quantify the rest time in the future.

Since the AUC value of our model is 0.94, the model has strong generalization ability. We have reason to believe that the model works for other matches such as Women's matches, tournaments, court surfaces, and other sports such as table tennis.

Memo

To: Tennis Coaches

From: Team # 2405186

Date: February 5th, 2024

Subject: Effects of Momentum on Matches

Our recent study delved into the intricate dynamics of momentum in tennis matches, exploring how it influences the outcome and flow of play. Key findings indicate that momentum, characterized by a series of consecutive points won, significantly impacts a player's performance and psychological state during a match.

Our study revealed a direct correlation between momentum shifts and players' performance. Positive momentum boosts confidence, while negative momentum can lead to frustration and a decline in performance.

Our research found that service order and service speed have the greatest impact on game performance. Coaches should pay special attention to these two on-court factors and remind players of grasping the opportunity and maintain a good attitude when serving.

Before matches, coaches should emphasize the importance of mental preparation alongside physical training and teach players techniques for maintaining focus and re-setting their mindset after unexpected events. During matches, coaches should use coaching interventions strategically in rest breaks to guide players to adapt to the changing dynamics of the game. After matches, coaches should help players actively learn from the previous game, ease their mental stress, and embrace the next game with a peace mind.

Understanding the role of momentum and preparing players to respond effectively to disruptions in the flow of play are crucial for success in tennis matches. By incorporating mental training, utilizing rest breaks, and post-game rituals into coaching strategies, coaches can allow players to harness dynamic shifts and improve their overall performance.

For a more in-depth discussion of the study results and personalized coaching strategies, please feel free to reach out for a one-on-one consultation.

Yours Sincerely,

Team #2405186

References

- [1] Richardson Peggy A, Adler William, and Hanks Douglas. Game, set, match: Psychological momentum in tennis. *The Sport Psychologist*, 2(1):6976, 1988.
- [2] ZHU Qianxiang and Luo Pengping and Wang Longpeng et al. Intelligent drilling process identification based on analytic hierarchy process [j]. *Coal Geology and Exploration*, 52(03):184–190, 2024.
- [3] Philippe Meier, Raphael Flepp, Maximilian Ruedisser, and Egon Franck. Separating psychological momentum from strategic momentum: Evidence from mens professional tennis. *Journal of Economic Psychology*, 78:102269, 2020.
- [4] Elia Morgulev. Streakiness is not a theory: On momentums (hot hands) and their underlying mechanisms. *Journal of Economic Psychology*, 96:102627, 2023.
- [5] Dong Shishi, Huang Shihishi, and Guo Xindong. Analysis on random forest theory [j]. *Integrated Technology*, 2(01):1–7, 2013.
- [6] Guo Wenfeng, Fan Chao, and Guo Xindong. Mooc retreat lesson based on binary logistic regression model to predict[j]. *Journal of computer age*, 12:50–53, 2017.
- [7] Cheng X. Application of fuzzy-principal component analysis method in groundwater quality evaluation of petrochemical enterprises [j]. *Science and Technology Innovation and Application*, 14(06):128–131, 2019.

Appendices

Code of Random Forest

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import roc_curve
from scipy.interpolate import interp1d
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score,
roc_auc_score, roc_curve
import seaborn as sns
import shap
import random

def evaluate_classification_model(model, X_train, y_train, X_test, y_test):
    y_pred = model.predict(X_test)
    y_proba = model.predict_proba(X_test)[:, 1]

    accuracy = accuracy_score(y_test, y_pred)
    precision = precision_score(y_test, y_pred)
    recall = recall_score(y_test, y_pred)
    f1 = f1_score(y_test, y_pred)
    roc_auc = roc_auc_score(y_test, y_proba)

    print(f'Accuracy: {accuracy:.4f}')
    print(f'Precision: {precision:.4f}')
    print(f'Recall: {recall:.4f}')
    print(f'F1 Score: {f1:.4f}')
    print(f'ROC AUC: {roc_auc:.4f}')

    fpr, tpr, thresholds = roc_curve(y_test, y_proba)

    # Rest of the code remains unchanged

def plot_feature_importance(model, X, top_n=None):
    feature_importance = model.feature_importances_
    feature_importance_df = pd.DataFrame({'Feature': X.columns, 'Importance':
feature_importance})
    feature_importance_df = feature_importance_df.sort_values(by='Importance',
ascending=False)

    if top_n:
        top_n_features = feature_importance_df.head(top_n)
    else:
        top_n_features = feature_importance_df

    top_n_features.to_csv('top_n_features.csv')
```

```

top_n_features['Feature'] = top_n_features['Feature'].apply(lambda x: x[:15]+'...' if
len(x) > 15 else x)
top_n_features['Cumulative_Importance'] = top_n_features['Importance'].cumsum()

plt.figure(figsize=(12, 6))
plt.bar(top_n_features['Feature'], top_n_features['Importance'], edgecolor='black',
linewidth=0.5)
plt.xlabel('Feature')
plt.ylabel('Importance')
plt.xticks(rotation=0, fontsize=10)
plt.grid(True, linestyle='--', alpha=0.5)
plt.twinx()
plt.plot(top_n_features['Feature'], top_n_features['Cumulative_Importance'],
color='lightcoral', marker='o', label='Cumulative Importance')
plt.ylabel('Cumulative Importance', color='lightcoral')
plt.legend(loc='upper left')

plt.show()

def analyze_shap_values(model, X, feature_names):
    explainer = shap.TreeExplainer(model)
    shap_values = explainer.shap_values(X)

    shap.summary_plot(shap_values, X, feature_names=feature_names)
    plt.show()

def analyze_match_or_set(data, match_id=None, set_number=None, detail=False,
comment=False, analyze=False):
    features = [
        # ... (remaining features)
    ]
    target = 'point_victor'

    if match_id:
        match_data = data[data['match_id'] == match_id]
        if set_number and not analyze:
            match_data = match_data[match_data['set_no'] == set_number]
    else:
        match_data = data[data['match_id'] == data['match_id'].iloc[0]]

    match_data['elapsed_time'] =
pd.to_timedelta(match_data['elapsed_time']).dt.total_seconds()

    weights = [0.1970, 0.2611, 0.4176, 0.1242]
    match_data['Momentum'] = np.dot(match_data[['momentum_1', 'momentum_2', 'momentum_3',
'momentum_4']], weights)
    scaler = MinMaxScaler(feature_range=(-1, 1))
    match_data['Normalized_Momentum'] =
scaler.fit_transform(match_data['Momentum'].values.reshape(-1, 1)).flatten()

```

Report on Use of AI

1. OpenAI ChatGPT (Dec 1, 2022 version, ChatGPT-4)

Query1: What aspects should I modify the drawing logic without changing the code logic, so that the icons can be very flashy and eye-catching

Purpose: Some drawing logic is based on AI suggestions, which allows us to make our drawings more beautiful.

Output: To make the drawing more flashy, you can enhance the visual effect of the chart by adjusting the styles of colors, lines, markers, and other elements. Here are some modification suggestions: Color and Style: Use more vivid and contrasting colors and lines. For example, you can try using more gradients and transparency in the colors of curves and points. Marking points: You can try using larger and more eye-catching marking points, or use different marking shapes. Line type and line weight: Try using different line types and line weights to make the chart more exaggerated. You have already used dynamic line weights in the code, and you can further adjust or add other elements. Background and Grid: Modify the color and style of the background and grid lines to better match the overall style. Annotations and Text: Add more annotations, text, or arrows to highlight key information in the chart.

Query2: Is there a recommended tennis dataset source that includes detailed information about each serve of a player, such as the first and second hand, score, running distance, and serving posture? CSV format required

Purpose: We attempted to inquire with AI about machine learning datasets for tennis matches, and AI provided several related websites. Unfortunately, we ultimately did not find any high-quality data available.

Output: As of my knowledge deadline (January 2022), I am unable to provide specific tennis dataset links. However, you can try searching for tennis datasets in the following places: ATP official website: ATP (Association of Tennis Professionals) may provide some official tennis datasets. You can search for relevant statistics and data on its website. The official website of ITF: The International Tennis Federation (ITF) may also provide some tennis datasets. Check their official website to see if there is any information you need. Kaggle: Kaggle is a data science competition platform where users sometimes share various datasets, including sports data. You can search for relevant tennis datasets on Kaggle. UCI Machine Learning Library: UCI machine learning libraries typically provide various types of datasets, including sports related data. You can check their website to see if there is a tennis dataset that suits your needs. GitHub: Some developers and researchers may share their collected tennis datasets on GitHub. You can search on GitHub and see if there are any projects that meet your requirements. Please ensure that you comply with relevant laws and regulations regarding data usage, and review the license agreement for the dataset to ensure that you have the right to use it. Also, remember to check if the dataset contains the specific information you need, such as first and last, score, running distance, and serving posture.