Problem Chosen	2024 MCM/ICM Summary Sheet	Team Control Number
C		2408451

Build the Momentum——Assessing and Lifting Match Performance

Summary

The purpose of this question is to explore the patterns of changes in the situation that occur within a tennis match and attempt to predict such changes. In the first question, we need to establish a momentum model for both athletes to explore how well they perform and how well they perform. After analyzing the data, we selected some variables and compared their impact on future games with the overall trend to roughly determine the degree of impact on momentum. Based on this, we create a model for calculating momentum. In the second question, we need to first refute the coach's words, and we will achieve this goal through a randomness test. Firstly, we test the randomness of the score data of all players through a run randomness test, and then compare the 10000 randomly set point change values with the point change values of a certain game through the KS randomness test. Both of them refute the coach's statement; Furthermore, we calculate the accuracy by comparing the final momentum values obtained from various factors and the relationship between victory and defeat in each game. The higher accuracy reflects that the overall changes in points are influenced by various internal factors. In the third question, we first determine which factors may lead to a change in the situation, and then define some turning points based on the difference between the situation before and after each point. We use decision trees and random forest models for machine learning, and compare these data to obtain two points that can predict where future turning points may occur based on the previous data in a game; Next, based on the importance of these independent variables, we will provide suggestions and strategies to athletes to help them achieve better results in competitions; In the fourth question, we applied these two models to other data to explore whether this model still performs well in other tennis matches. Based on the characteristics of the selected data, we analyzed other factors such as field factors and found that the final field factor had no significant impact on our model. We also analyzed the feasibility of applying it to other matches, Based on the characteristics of the independent variables we selected, we ultimately concluded that our model is suitable for games where both sides have a serve and each small game requires a score, such as badminton, table tennis, etc.

Keywords: Momentum   Turning Point   Decision Tree   Identification and Prediction

Contents

1 Introduction

1.1 Background

In many sporting events today, people are keen to follow the performance of the players and the results of the matches. In the case of tennis players, we can use websites to look at the number of points and sets won in each match, and to quantify how well a player is doing in a given match. However, even within the same match, a player's performance can change dramatically rather than remain consistent.Within the framework of the theory of complex systems, athletic performance can be viewed as a dynamic system containing numerous factors that are constantly exchanging and exchanging information with the external environment, resulting in moment-to-moment changes within the system. These minor system changes may be amplified at some point in time when certain stimuli are present during a match. In this case, one side of the athletes may show a clear advantage, their technical and tactical applications are more skillful, and scoring is continuous. On the other hand, the other side may fall into a disadvantage, which is manifested in low success rate of technical and tactical skills, low athletic status, and continuous loss of points. However, as the game progresses, the momentum of both sides may change. The original superior team may fall into a bad position, while the inferior team may seize the opportunity to realize a reversal and obtain the superior position. During the game, both sides are always in the alternating changes of equilibrium and imbalance, and such changes are ultimately reflected in the fluctuation of the game phenomenon.[1] On the evening of May 20, 2023, the Chinese badminton team met Japan in the Sudirman Cup semifinals. In the desperate situation of being 1-2 down in the total score and 16-20 down in the set score, Liu Yuchen/Ou Xuanyi team saved the match points continuously and took 6 points in a row to complete the miraculous reversal. As a matter of fact, the Chinese duo made many mistakes in the first half of the match and faced great pressure on the match points, but in the last moment, they completely mastered the situation and realized the comeback with 6 consecutive points, which was a great test for the players' mental quality.

The incredible volatility of the players' performances is a topic we have to look into. "Momentum", in the dictionary, means force or power acquired through movement or a series of events. And this momentum, to a certain extent, affects the player's subsequent performance.[2]

First of all, it is important to recognize that there are countless factors that can affect a player's performance, including, but not limited to, the training and preparation before the match, the level of rest the player has before the match, the player's own technique and tactics, the psychological pressure and anxiety brought about by the match itself, the environment of the match venue and the weather, and many other unknown factors.[3] The interaction of their physical condition, psychological state, or other external factors will ultimately show whether the players perform better or worse than usual. Through the study of momentum, we hope to find out the reasons for the large fluctuations in players' performance.

Second, by analyzing the problem and developing a model, we will investigate how this model can be used to suggest better strategies and recommendations for players in the Grand Slam and other tennis-related tournaments, as well as in other sports.

The model has many applications: it can be used to evaluate how players are likely to perform in a match, and it can be used to anticipate possible match performances for spectators and fans, thus predicting the more likely winning team or player before the match. In addition, the model can also be used to suggest better game strategies for athletes, thus reducing the work pressure of coaches, saving labor costs, and contributing to the power of science and technology in sports.

Figure 1: A Tennis Game

1.2 Restatement of Problems

1. Establish a model that enables the model to make corresponding changes in the momentum of the players after the occurrence of events such as scoring or losing points, set the weights of the independent variables of each event that affect the momentum accordingly, and establish a relationship model between the events and the momentum. Eventually, the model is applied to multiple matches, so that the momentum of the players can be clearly represented at each time point in each match, and the flow of each match is displayed in the form of visualized data.

2. According to the question, we need to verify whether the coaching view is correct. The view is that fluctuations in the game and a player's success are random, i.e., a player scores a point, loses a point, or none of the other events have an effect on the final outcome of the game. Therefore, we need to prove that the outcome of the game is related to various event factors in the game. We need to prove whether event factors such as scoring are related to momentum by using the model, and then prove the view correct or incorrect by using momentum as a bridge to connect the event factors to the outcome of the match.

3. i) Using data from greater than or equal to 1 game, build a model which predicts the emergence of future turning points from data that has already occurred in the game and find out which of the many factors has the greatest weight on the impact of the game's fluctuations.

   ii) For the data of previous matches, analyze its fluctuation and make suggestions for his future matches based on it.

4. Apply the model to more matches and even other fields (not only tennis) to verify its universality. If the predicted results deviate from the actual performance of the match, analyze the reasons for the deviation and identify possible missing factors affecting the fluctuation of the match performance.

1.3 Our Work

For the first question, we first make data observations and identify possible variables that affect momentum as independent variables in the model. At the same time, we explore the weights of influence, set the weights of each event independent variable to influence momentum accordingly, and build a model for the calculation of momentum: the Comparing Planning Momentum Model. Ultimately, we use the visualization data to determine the reasonableness of the model. For the second question, we use the Comparing Planning Momentum Model to get some momentum values, and as we compare the momentum of the two players with the data, and then use a variety of randomness test methods, we prove that the coach's point of view is wrong, and that the score is strongly correlated with the momentum.

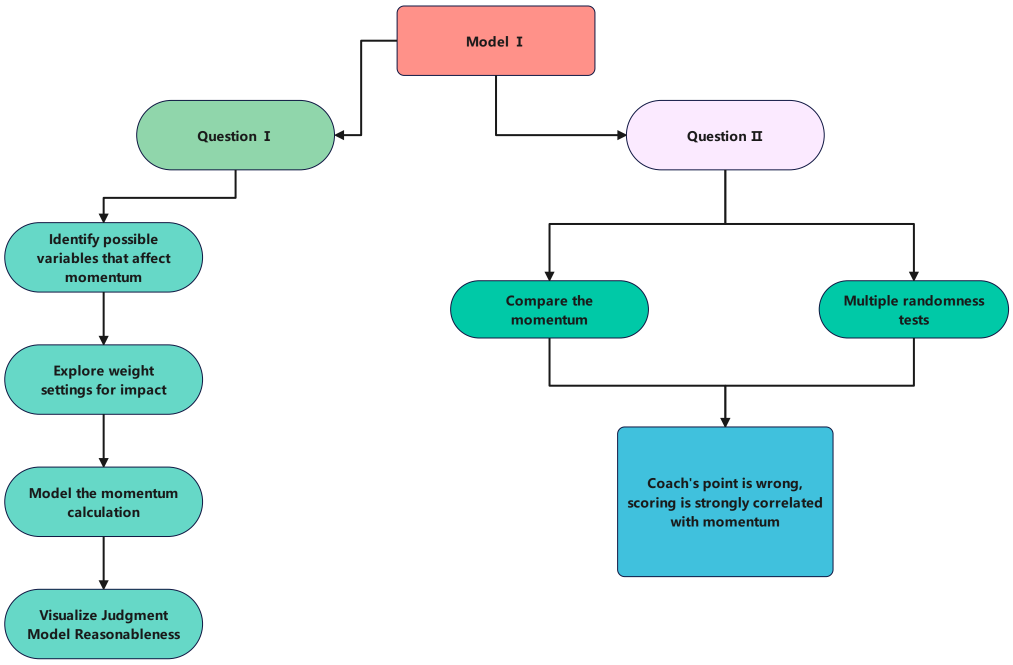

Figure 2: Model I Overview

For the third question, we would like to find an algorithm that can represent the turning point in order to solve the third question. First, we find potentially relevant data, and then we use decision trees and random forests to machine learn the data in the question to obtain a model, and finally we compare it with the given data to obtain a confusion matrix, determine the reasonableness of the model, and based on this model, we give reasonable advice to the athletes. On the basis of the third question, we apply the model obtained from machine learning to other data, use visualization to judge the universality of the model's application, and finally apply the model to other types of competitions to judge whether it is general or not.

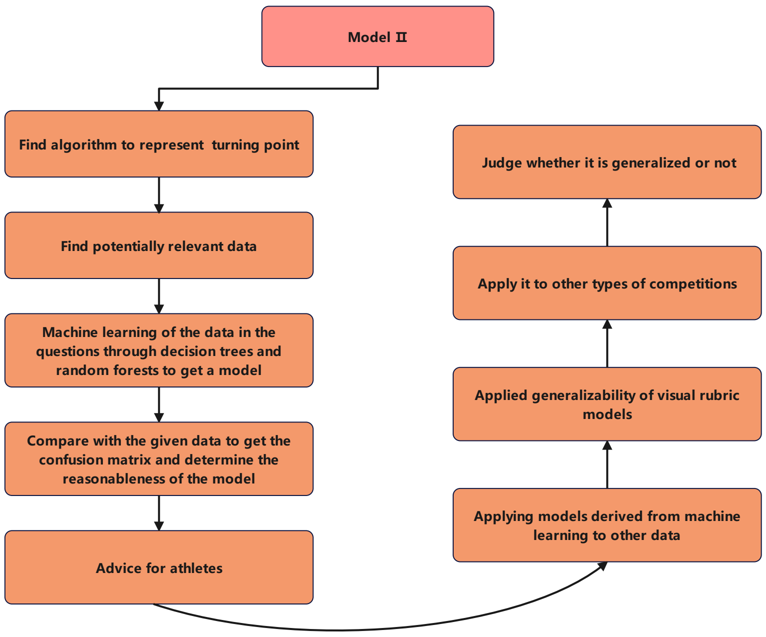

Figure 3: Model II Overview

2 Assumptions and Justifications

2.1 General Assumptions

- **Assumption 1:**Assuming that the data after data cleaning and exception handling is accurate.
- **Assumption 2:**Assuming that these opponents have the same tolerance for various situations in the competition.
  Justification: The players in the dataset are all very experienced, so they have strong psychological tolerance. Few's psychological tolerance will fluctuate significantly[4], so there will be no significant deviation in the results;
- **Assumption 3:** Assuming that the elapsed time,the depth of hitting, angle of hitting, and depth of return have no significant impact on the results.
  Justification: These factors have an indirect impact on the competition and can be roughly covered by the running distance;

3 Question I

3.1 Question Analysis

Since there is no specific definition of this particular performance time period, it is useful to define it as follows: momentum responds to the total performance in an entire match up to a certain small game. From this, we can deduce some related independent variables: firstly, when the opponent is the first to grab a game point, or when the player has lost a series of points, these moments can be regarded as the player's underperformance; and when the player has made various errors, including "double faults" and unforced errors, which indicate that the player's operation is poor. Errors, including "double faults" and unforced errors, indicate that the player's operation has some problems, which can also be recognized as poor performance; when the player loses the game and the set, it can also be judged as poor performance; the degree of fatigue will also affect the player's performance to a certain extent. When a player scores consecutive points or plays a great performance, such as scoring a point at the net, serving a point, or the opponent does not catch the ball, it can be recognized as a better performance; in addition, for the player who wins the game and the set, it can be considered as a better performance. In summary, we can derive some independent variables that lead to the increase or decrease of momentum, and the next step is to set them to a more appropriate value of momentum increase or decrease: by capturing the possible influencing factors of each player's momentum increase or decrease, exploring the scoring rate of the next ten matches and comparing it with the total scoring rate, the proportion of the probability of increasing or decreasing can be taken as the significance of the independent variable.

3.2 Model: Comparing Planning Momentum Model

Based on this, we can derive the following model:

Events	Amount of Change in Momentum
Opponent takes match point	**-1**
Score one point	**+1**
Score two points in a row	**+2**
Score two points in a row	**+2**
Score three or more points in a row	**+3**
Lose two points in a row	**-1**
Lose three points in a row	**-2**
Lose four or more points in a row	**-3**
Lose a game	**-1**
Lose a set	**-1**

Events	Amount of Change in Momentum
Win a game	**+2**
Win a set	**+4**
Faults(Unforced errors or Double faults)	**-1**
Ace gets a wonderful score	**+1**
Win for the opponents don't catch the ball	**+2**
Breaking serve	**+2**
Lose for opponent breaking serve	**-1**
Run every 200m	**-1**

Which is called Comparing Planning Momentum Model. Based on this, for the momentum at the end of each point, the change in momentum of the player over a whole session can be obtained. These four graphs (4) depict the graphs of player 1 and player 2's scores inside all the given sets of matches as well as the graphs of momentum changes based on our model, and it is easy to see that there is a strong correlation between the two.

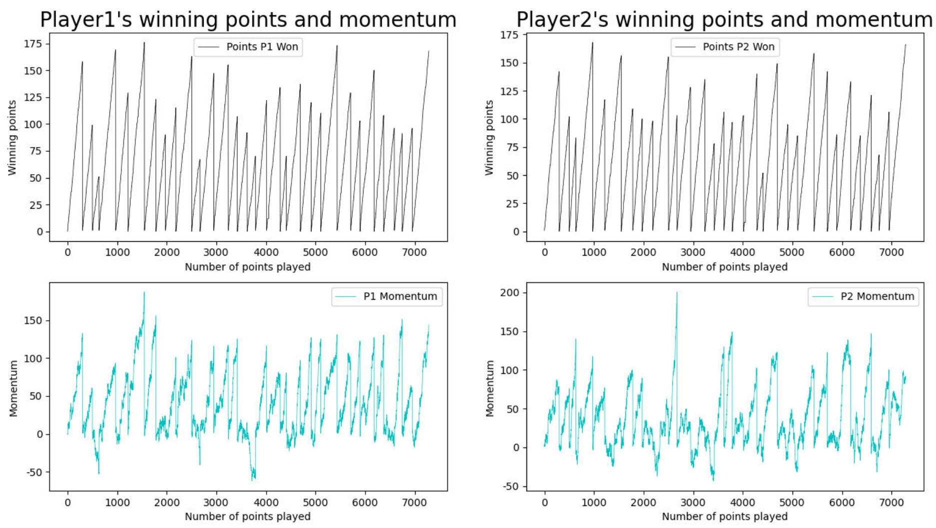

Figure 4: Two Players' Winning Points and Momentum

3.3 Model Validation

However, these graphs alone, by themselves, are not good enough to see the relationship in detail due to their more involved and vague reflection of the data, and more detailed graphs are needed to more concretely reflect the relationship between the scoring situation and the momentum. The graph below (where the different colored dotted lines represent different set) (5) shows the variation of the difference between the scores of the legendary battles of Alcaraz and Djokovic mentioned in the title and the difference in momentum. As can be seen from the graph, there is a very significant positive correlation between the difference in momentum and the difference in points scored between the two, and the trends of the two are basically the same.

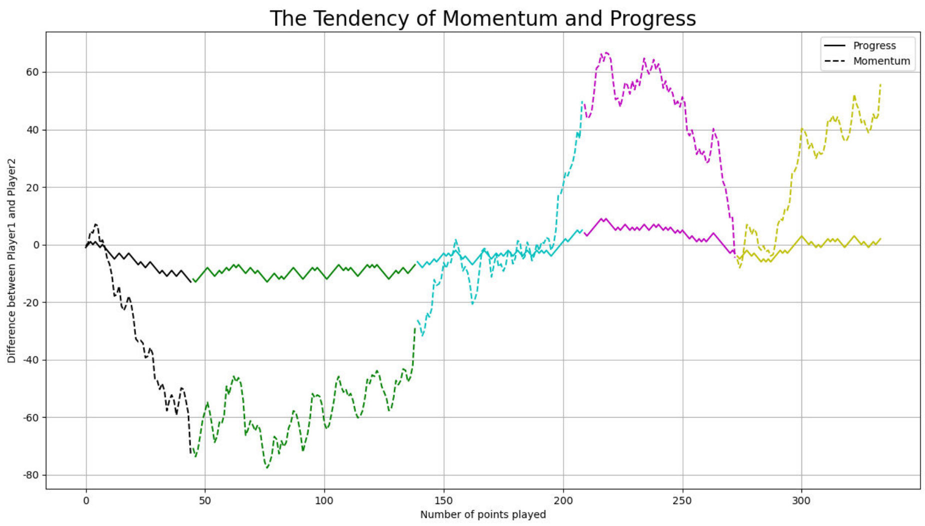

Figure 5: The Tendency of Momentum and Progress (i)

As can be seen from the graph, in this legendary battle, in the first set, Djokovic's momentum significantly exceeded Alcaraz's; in the second set, Alcaraz's momentum gradually recovered and significantly exceeded Djokovic's in the third set; in the fourth set, the gap between Djokovic and Djokovic was further widened, but the latter's brilliant performance managed to get back to a set, allowing the score to level out once again, and finally, in the fifth set, Alcaraz once again made another move to surpass Djokovic's momentum, and ultimately, also achieved a final match victory!

4 Question II

4.1 Randomness Tests

4.1.1 Run Test

According to the run test, which tests whether each score is random or not, finally yields that his probability of randomness is so low that it can be approximated as not random, and it can be inferred that his conclusion is wrong. Fig.6 shows the connectivity between the momentum and score. This number in the result is very close to 0, so we can assume that the relationship is non-randomized as a way of proving the coach's point wrong.

NPar Tests

Runs Test

	point_victor
Test Value[a]	1.50
Total Cases	7284
Number of Runs	3402
Z	-5.613
Asymp. Sig. (2-tailed)	.000

a. User-specified.

1.9857E-8

Figure 6: Results of Runs Test

4.1.2 Kolmogorov-Smirnov Test

Though The Run Test proves that the distribution of point victors is not random, we cannot jump to the conclusion that the appearance of turning points and shift of Momentum are correlated with some in-game factors, instead of being random. Therefore, as a supplement, we performed a Kolmogorov-Smirnov Test to justify that the pattern of normal matches is quite different from the randomly-simulated ones.

We uses Stochastic Process Simulation and Hypothetical Test to establish our model. As usual, we take the legend match between *Carlos Alcaraz* and *Novak Djokovic* as the template of simulation, and the times of serving S_t and Server WinRate

$$W = \frac{\sum server\ \ winning\ \ point}{all\ \ points\ \ played}$$

are considered. After that, we generate the series of cumulative points won by *Alcaraz* at random. Since the increment of cumulative points can signify whether a player is at advantage, we use this index as statistics.

Firstly, we have the null hypothesis

$$N_0 : Transitions\ in\ match\ are\ random.$$

Secondly, we get the random distribution density by running our simulation 10,000 times, and compare it with the actual case using Kolmogorov-Smirnov test, which compares the underlying continuous distributions $F(x)$ and $G(x)$ of two independent samples.

Finally, the result of K-S Test is

$$KS\ statistic = 0.17045479041916167,\ P\ value = 5.923796577159508e - 09$$

Here P-value is lower than the significance index(0.01), suggesting that we can reject the null hypothesis, and that the transitions in matches are not random.

Moreover, we plot the distribution density histogram, which is shown in Fig.7. This shows the fluctuations in real matches are greater than those in random distribution, suggesting that some in-game factors, namely, break pt, momentum, etc. may influence players' performance.

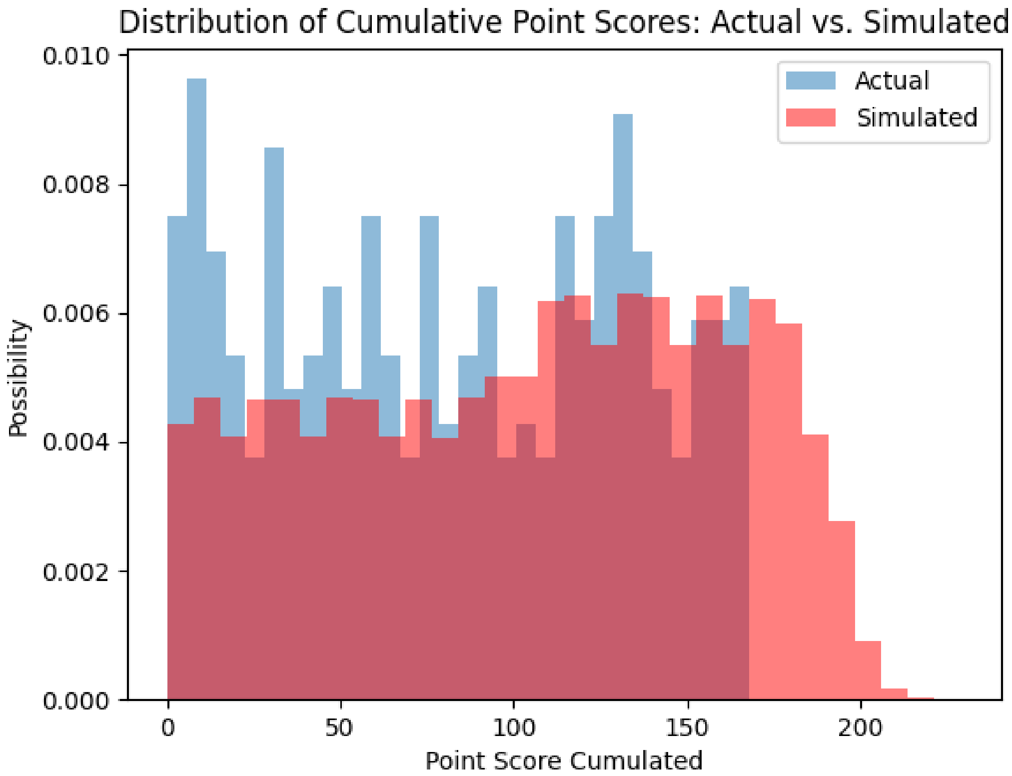

Figure 7: Results of Runs Test

4.2 Effect of Momentum on the Outcome of a Competition

According to the model we proposed in Question I , we calculated the momentum that Player 1 and Player 2 have at the end of each match separately, and subtracted Player 2's momentum from Player 1's momentum to calculate their difference. And whether the model is reliable or not is directly determined by the result of the game. In the general definition, the higher the momentum is, the better the performance of the player, and the higher the possibility of winning the game. Therefore, if the difference is positive, then Player 1 performs better and Player 1 should be the winner. Similarly, if the difference is negative, then Player 1 performs worse and Player 2 should be the winner. We then compared the relationship between the difference and the outcome of the game, and if the player with higher momentum is also the winner, the model prediction is labeled as reasonable; if the player with higher momentum is not the winner, the model prediction is labeled as unreasonable. In the end, we found a total of 30 out of 31 samples to be reasonable model predictions and only 1 sample to be unreasonable model predictions, with an accuracy: $Acc = 0.967741935483871$.

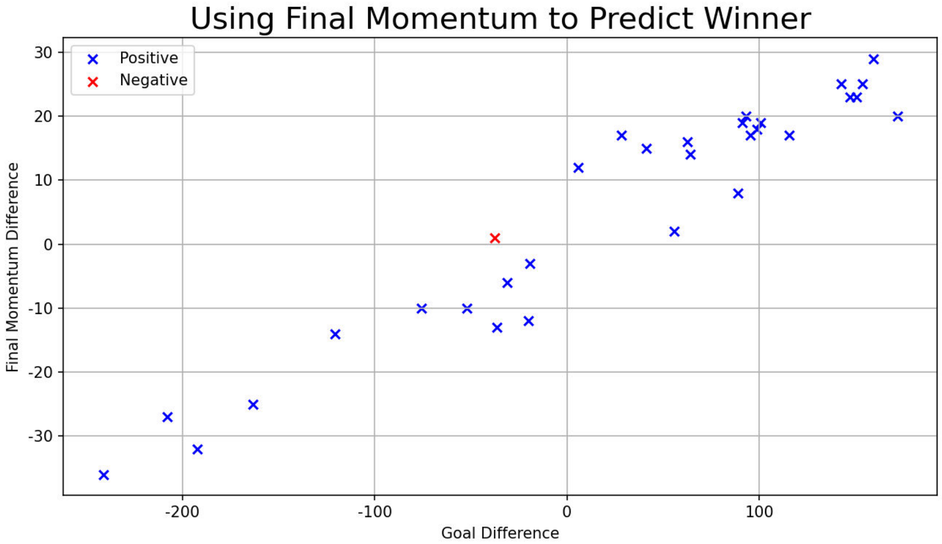

Figure 8: Using Final Momentum to Predict Winner

Based on Fig.9, we can find that although the winner of this game is Player 1, but in the whole point change Player 2 has been ahead of the score, and both of them also fought to the final game, and the data mostly shows that Player 1 won by a small score and Player 2 won by a large score, so this situation shows that the model may have mistakes when facing some rare cases. However, the difference in momentum between the two players is very small, just a slight difference, which means that the model also roughly reflects the situation where Player 1 wins but the total number of points scored in the game is less, and it cannot be said that the model is completely invalid.

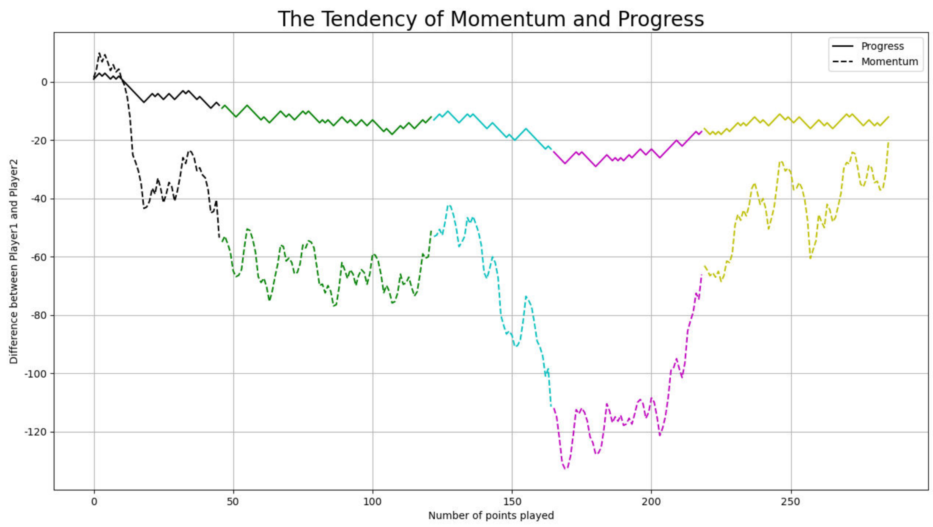

Figure 9: The Tendency of Momentum and Progress (ii)

5 Question III

5.1 Searching for Turning Points

Firstly, we need to define which time points in a model can be used as turning points, and this definition should be related to the scores before and after this point. Therefore, we can use the score of player 1 in the first 20 matches of each point minus the score of player 2 in the last 20 matches to represent the change in score before and after this point: when the absolute value of this value is large, it can be considered a turning point at this point. We will remove the first and last 20 points from each game.Since these points are located at the beginning and end of each game, they cannot be identified as turning points for this game. Next, we will process the remaining data using the method described above, and finally obtain the characterization turning point values for each remaining point. We will sort and observe their absolute values.Due to the fact that the points near the turning point all have higher absolute values for this data, none of the 20 points near this turning point should be considered turning points anymore.

Ultimately,we decide to use points with absolute values$\geq$5 as turning points.We use categorical variables to describe the type of turning point. A value$\geq$5 is set as 1, indicating a favorable turning point for player 2; A value $\leq$-5 is set to -1, indicating a favorable turning point for player1; The rest are set to 0, indicating no significant turning point in the war situation.

The following figure shows the identification of turning points during the first given competition. Since the momentum change is basically consistent with the point change, it is more intuitive to label the identification on the momentum change curve.

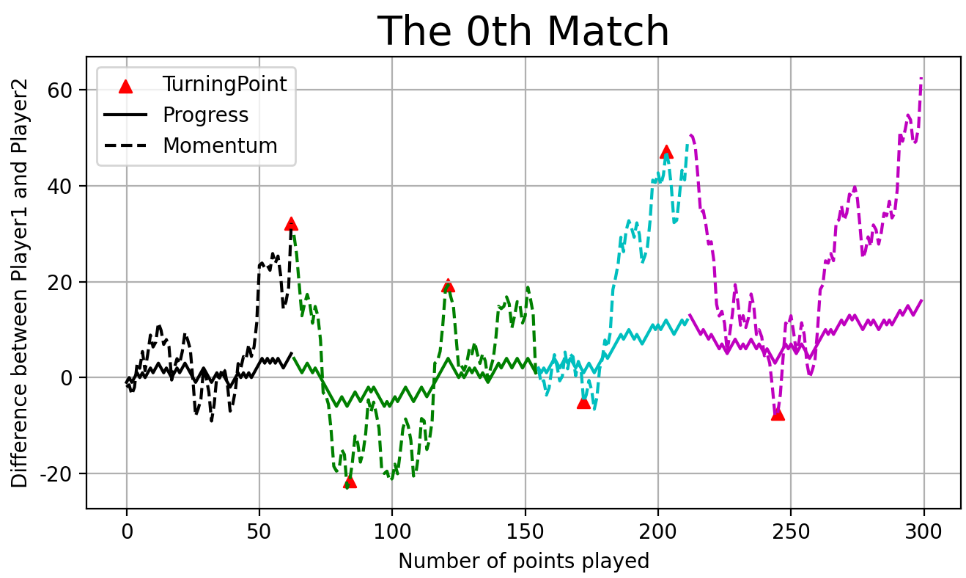

Figure 10: Turning Point

It is not difficult to see from the pictures that this turning point search method is quite reasonable.

5.2 Model Training

Next, we need to predict the occurrence of future turning points based on the given data of the events that have already occurred.

How to find the independent variable? Firstly, it can be determined that the independent variables are certain data before the start of each point, and they should represent the difference between the two states before this point.[5] After careful consideration, we chose two contestants to compare their total number of mistakes (including double mistakes and unforced errors) in the first 20 games of each point, their total number of outstanding performances (including ace and untouchable shots), their number of breaking serve, and their score differences. It should be noted that the 20 points before and after the turning point should not be included in the training range, otherwise due to their independent variables being similar to the turning point, it will lead to the inaccuracy of the learned model.
We choose two models here, namely decision tree and random forest[6], to train the given data.

5.2.1 Decision Tree Model

Through the decision tree model, we can obtain estimated values for each point, which are -1, 0, and 1 respectively. After comparing them with the actual values, we can obtain the confusion matrix of the decision tree model.The confusion matrix is shown in the following figure.

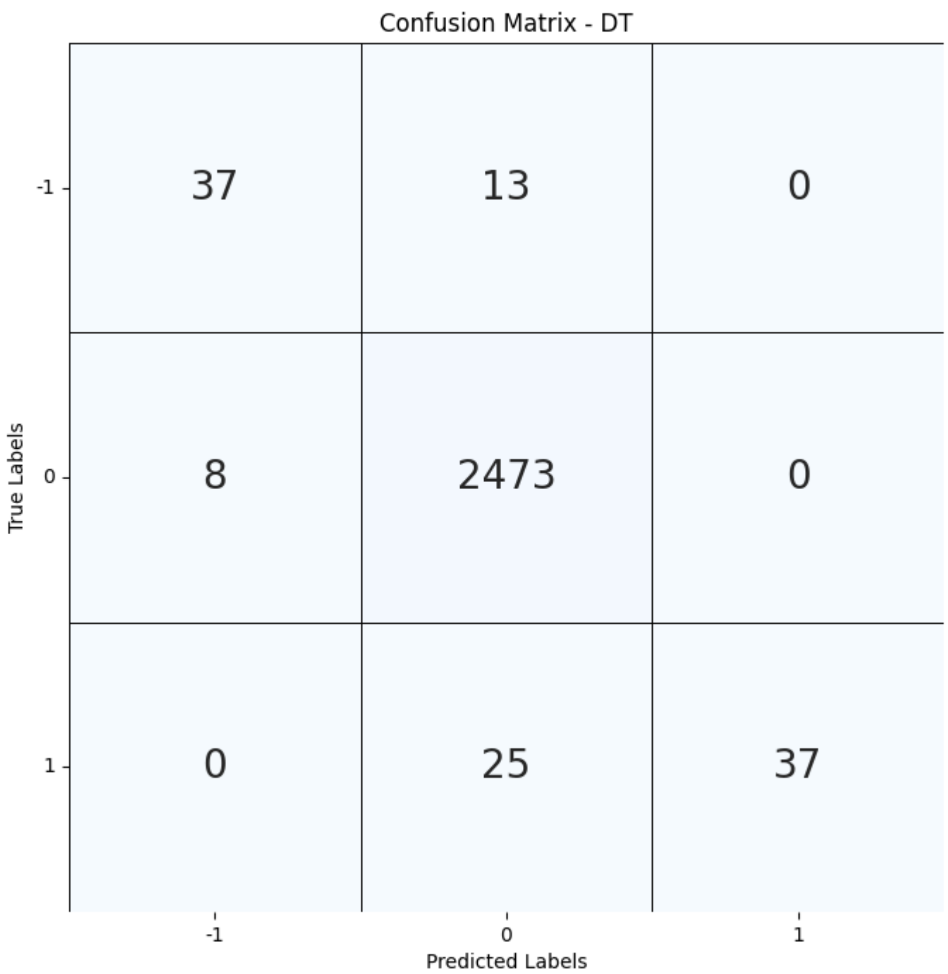

Figure 11: Decision Tree's Confusion Matrix

From this confusion matrix, it is not difficult to see that the model can roughly identify which points are turning points and identify them, while almost no non turning points are identified as turning points, only a small portion of turning points are identified as non turning points, indicating that our model can recognize turning points and improve the recognition standards for turning points.The comparison chart between the predicted turning point obtained from random forest and the actual turning point is shown below.

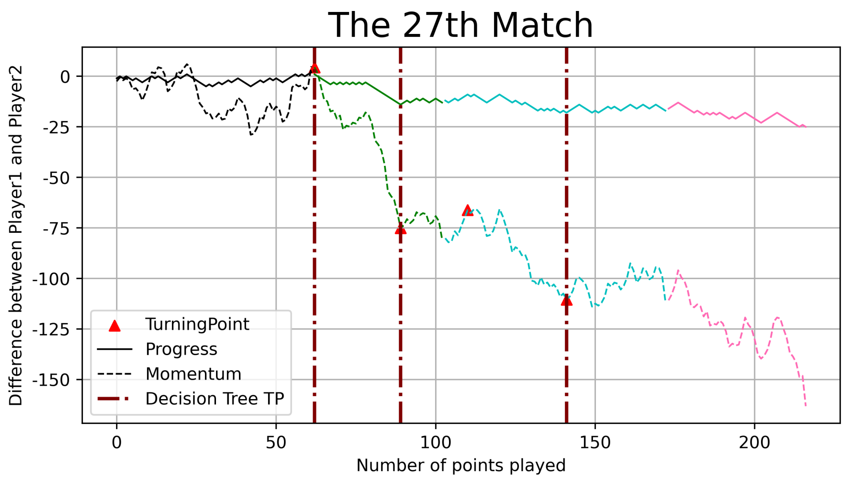

Figure 12: Decision Tree's Turning Point

From this model,We can obtain the importance of the difference in errors, outstanding performance, breaking serve, and score for the first 20 games, which are 0.23, 0.33, 0.03, and 0.41, respectively.

5.2.2 Random Forest Model

Through the random forest model, we can also obtain estimated values for each point, which are -1, 0, and 1 respectively. After comparing them with the actual values, we can also obtain the confusion matrix of the random forest model.The confusion matrix is shown in the following figure.

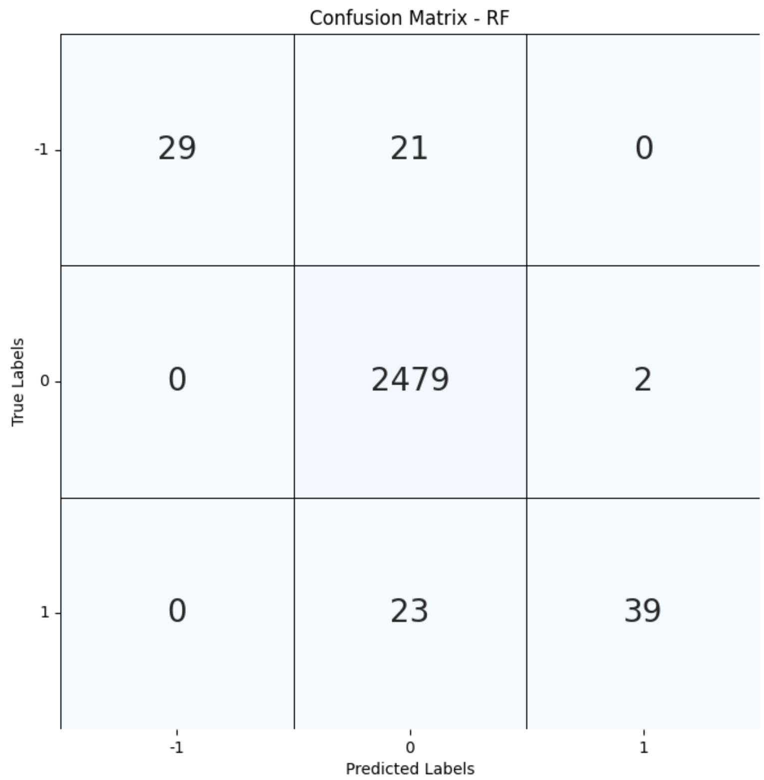

Figure 13: Random Forest's Confusion Matrix

The situation is similar to the one in Decision Tree.From this model,we can improve the recognition standard of turning points without any harm to getting some unreasonable turning points,which does good to our future application.The comparison chart between the predicted turning point obtained from random forest and the actual turning point is shown below.

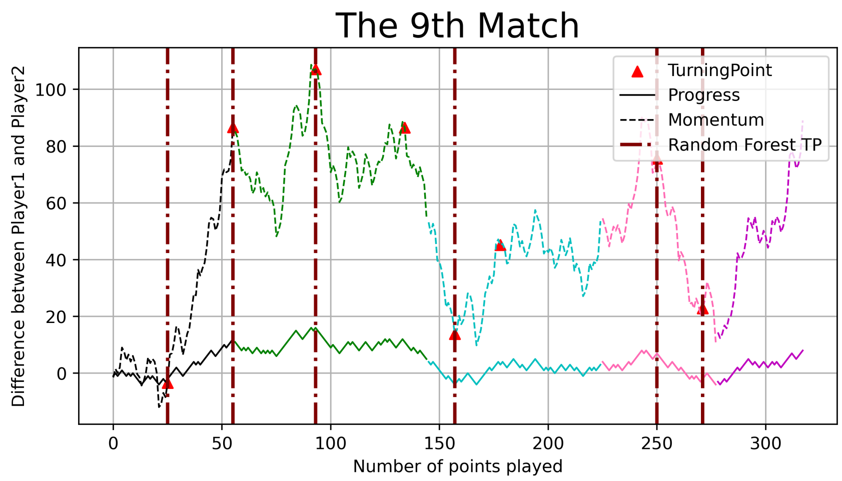

Figure 14: Random Forest's Turning Point

From this model,We can obtain the importance of the difference in errors, outstanding performance, breaking serve, and score for the first 20 games, which are 0.24, 0.28, 0.11, and 0.37, respectively.

5.3 Suggestions for players

Through the analysis and data of the above two models, we can know from the importance of the independent variable that in the first few points of each point, we score more than the opponent, perform more brilliantly than the opponent, and make fewer mistakes than the opponent. Breaking has a positive impact on the momentum of the player, and the importance gradually decreases.Due to the low probability of breaking serve occurrence, it does not appear to have high importance in the data. However, after calculation, the importance/occurrence rate is still very high, indicating that it is still very important.On the contrary,due to the high occurrence rate of the other three and the less significant difference, the importance in fitting the model can be approximated as its impact on momentum.
This can provide suggestions for athletes to face different opponents in future competitions. After being scored consecutively by the opponent, one must find ways to recover the score, otherwise it is likely to become a turning point for one's momentum decline and failure at this stage; Maintain momentum when scoring continuously, otherwise it may lead to being scored continuously by others and losing the advantage; In the competition, you can try to give some exciting performances, which will greatly improve momentum, but also try to reduce the error rate as much as possible, otherwise momentum will also decline;We can practice more on some exciting scoring methods to reduce the error rate while scoring well, which can keep the momentum rising relative to the opponent; And, when there is a possibility of breaking serve, try to strive for it as much as possible, but maintain a calm mind and consider it as an opportunity for oneself. Success will increase momentum largely, and failure will not have a negative impact on one's momentum.

6 Question IV

In the previous question, we have obtained two models for calculating turning points and demonstrated their feasibility. Next, we will apply them to other scenarios to predict turning points in competitions.

6.1 Analysis for Other Competitions

6.1.1 Men's Group at the 2022 Wimbledon Tennis Championships

We first found the data for the men's group at the 2022 Wimbledon Tennis Championships.(The source of data: https://github.com/JeffSackmann/tennis_slam_pointbypoint) We processed the data we found to obtain the difference in the number of mistakes, outstanding performance, breaking serve, and score for the first 20 games.
After sorting them out, we applied them to the two types of models we trained for prediction. Since these variables are all continuous, we stipulate that 20 points before and after each point that becomes a turning point will no longer become a turning point. After drawing, we obtain the following figures.

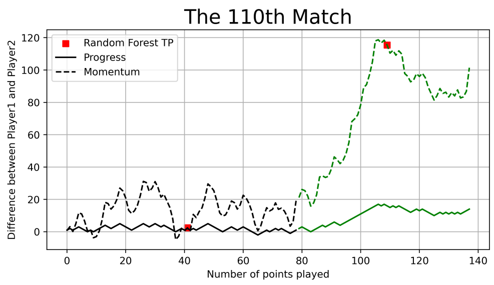

Figure 15: Random Forest's Turning Point 2

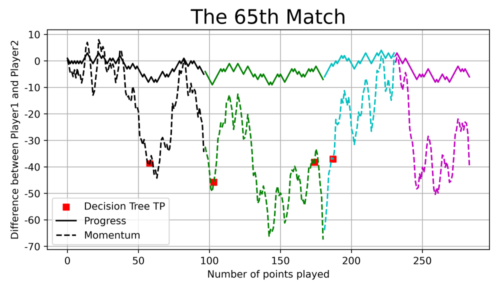

Figure 16: Decision Tree's Turning Point 2

From these graphs, it is not difficult to see that our model can still roughly predict the location of the turning points for the entire game when applied to these competitions. However, a small number of turning points have not been accurately predicted, indicating that there are still some secondary factors that may need to be taken into consideration.

6.1.2 Women's Group at the 2022 Wimbledon Tennis Championships

Next, we will analyze the turning points using the data from the 2022 Women's Wimbledon Tennis Championships. We noticed that the women's competition is a three game, two win system, so the length of the entire competition will be compressed relatively short. Therefore, the previous method of using 20 sets of data to determine the turning points may not be timely enough, and some seemingly obvious turning points may not be reflected. Therefore, we will compress the data from the first 20 games to 15 games for analysis.After drawing, we obtain the following figures.

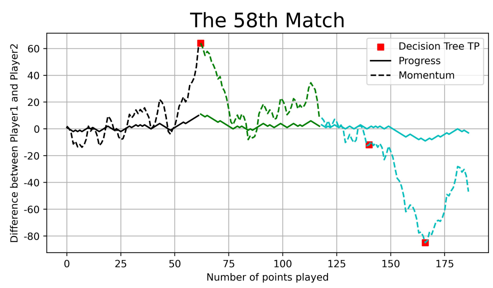

Figure 17: Decision Tree's Turning Point 2

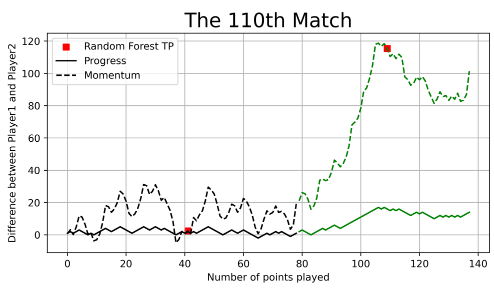

Figure 18: Random Forest's Turning Point 3

It can also be seen from this that our model has the ability to predict turning points in women's competitions, with relatively high stability.

6.2 Other Possible Factors

Through some comparisons of the above models, it is not difficult to find that there are many turning points that are within our expectations, but there are still some turning points that our model did not accurately predict. Therefore, it is necessary to consider other factors related to the possibility of changes in the situation.[7]

First,due to the different performances of different players on the field, some may have poor performance initially and gradually improve later, while others may perform better initially and then perform poorly later. Therefore, it is possible to summarize the characteristics of each player's own competition

based on data from previous player competitions, and weight the evaluation criteria for their turning points at appropriate times to obtain a more accurate model.
Second,during the competition, there may be changes in momentum due to various mistakes or outstanding performances by players, but the magnitude of these changes may not be the same for different players. Historical data can reveal the momentum changes of players when facing these situations. Through the historical data of players, we can assign different weights to the changes of different players, making the prediction of changes more accurate.[8]
Third,the level difference between different players can also be reflected through historical data. If the level of a player can be evaluated through factors such as historical win rates, and the possibility of situation changes can be analyzed based on their level, this is also an optimization method.

6.3 Possibility to Use This Model in Different Competitions

First,we previously applied the model to the 2022 Wimbledon Tennis Championships for both men's and women's groups, and found that it can basically predict the trend of the competition; Moreover, the ground types in these two matches were also different, but our model still performed well in these different ground types, indicating that our model has excellent universality in tennis matches.
Second,we should also consider the rationality of our model in other sports competitions. We evaluate turning points based on each point, so our model can be applied to games where scoring is frequent and there are people scoring in each small game; And we also consider breaking serve as an important evaluation factor, which indicates that our model can be applied to games with serving parties, and our model should also be applicable to games with two participating parties. Otherwise, our data will be the difference between the two parties, and such a difference cannot be obtained.
Due to the stable performance of our model in other data, it indicates that it can be applied to games that meet the above conditions, such as badminton, table tennis, etc[9].

7 Sensitivity Analysis

In the first question, we obtained a momentum model that is related to many factors. We can generally classify these factors into 5 categories: Combo, Brilliance, Mistake, Game point, and milestone, representing consecutive scores or losses, number of outstanding performances, number of mistakes, reaching a game point, and periodic success, respectively.
We control for the same other factors separately and fluctuate the impact of a certain factor on momentum by 10% to calculate the change in the final momentum difference result. This is reflected in the graph and the percentage of impact on momentum difference is calculated. The result is shown in the following figure.

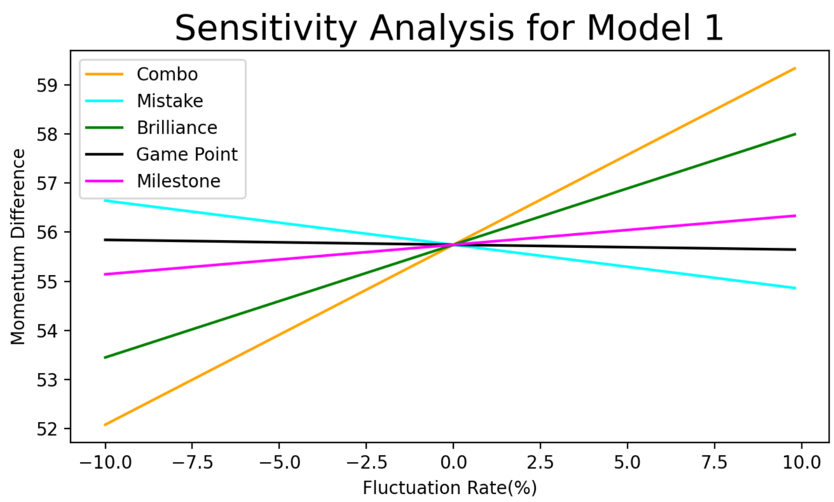

Figure 19: Sensitivity Analysis

It is not difficult to see from this graph that our model has relatively low sensitivity, and the fluctuation of a certain data does not greatly affect the accuracy of the model.
After calculation, the impact of Combo on momentum difference is about 1.3% per 1% change, the impact of miss is about 0.3%, the impact of brilliance is about 0.8%, the impact of game point is about 0.03%, and the impact of milestone is about 0.2%. It can be seen that our momentum model is relatively stable.
In the fourth question, we simulated the women's group competition by modifying the range of data reference, and still obtained a rough turning point, indicating that our turning point model still has good stability and low sensitivity in different scenarios.

8 Model Evaluation

8.1 Momentum Model Evaluation

8.1.1 Advantages

a. Our momentum model can accurately evaluate everyone's performance on the field.

b. The model takes into account many factors and has strong practicality.

c. The model has relatively high stability.

d. The model approach is concise and clear, with high efficiency.

8.1.2 Disadvantages

a. When the performance gap between the two parties is close, the evaluation of the model may have some deviation from the final actual performance.

8.2 Turning Point Model Evaluation

8.2.1 Advantages

a. We can accurately find the important turning point in training data.

b. The model can be applied to other competitions with high stability to predict the possible turning point.

c. Our model can use the data from the previous scene to predict the turning points in the future, without relying on historical data.

8.2.2 Disadvantages

a. Due to the lack of reference to historical data, some turning points cannot be well identified.

b. When this model is applied to other competitions, it is necessary to adjust each variable in the model according to the length of the competition to establish the number of previous games used, otherwise it will lead to significant deviation in the results.

9 Conclusion

9.1 Question I Result

- Momentum is a quantity that reflects a player's performance over a specific period of time, and by analyzing the problem, we can define it as follows: Momentum reflects the total performance of a player over the course of an entire game up to a certain inning.

- A player's performance is clearly related to his or her scoring, and scoring or losing points affects his or her momentum. We set the amount of change in momentum by different events in the game by comparing the back ten and the whole game scoring rate, as a way to get the change in momentum of a player over a whole game.

- According to the visualized data graph, it can be learned that the player's score is strongly and positively correlated with momentum, i.e., the higher the player's momentum, the better the player's performance and the stronger the scoring ability, which verifies the reasonableness of the model.

9.2 Question II Result

According to the question, the coach assumes that the momentum has no direct impact on the outcomes of the competitions. First, we use randomness tests to find whether each score is random or not.

- The result of Run Test shows the connectivity between the momentum and score. The number in the result is very close to zero so we can assume the relationship is not randomized.

- The result of KS test suggests that the transitions in matches do not occur randomly in most cases. Furthermore, the test indicates that there may be some fixed patterns and correlation between the occurrence of turning points and in-game features.

Meanwhile, based on Comparing Planning Momentum Model, we calculated the momentum of Player 1 and Player 2. Then we compare both players' momentum and test if the one with higher momentum can perform better and win the game(score more). If the player with higher momentum is the winner, the model prediction is labeled as reasonable. Totally, 30 out of 31 samples are proved reasonable and accuracy is 0.967741935483871, so we can come to the conclusion that scoring is cosely related to the momentum.

9.3 Question III Result

- We obtained a model for predicting turning points based on the difference in the number of outstanding performances, errors, scores, and breaks performed by two players in the first 20 games. We also found that breaks and the number of points scored over a period of time are the two most important factors affecting the situation.

- Through this model, we provide suggestions to the contestants.Maintain continuous scoring, try to block the opponent's scoring trend, strive for excellent performance, and avoid mistakes. Practice some excellent scoring methods to generate excellent performance while avoiding mistakes; Strive for the opportunity to break serve, and don't be discouraged even if you fail.These suggestions enable contestants to win more and perform better.

9.4 Question IV Result

- After testing in other competitions, we have determined the applicability of the model, which can be applied to other tennis matches; We also tested the women's group and found that by modifying the range of independent variable values, the model is still applicable. Through analyzing the characteristics of the model, we determined that the model can be applied to games such as table tennis and badminton, where both sides compete, score frequently, and the server is present,and other factors like types of grounds won't bring much difference to the prediction.

- We can predict which other factors can be used to predict the overall trend of the game. Through the analysis of the competition situation, we ultimately obtain the reactions of players to excellent performance and mistakes in their historical data, the scoring trends of players in their historical matches, and their level, which will play a certain role in further optimizing our model.

References

[1] Ji Chengkun, Quantitative analysis of the ”up and down” phenomenon of the score of world high-level women’s volleyball matches[D]. Beijing Sport University, 2022.DOI:10.26961/d.cnki.gbjtu.2020.000342.

[2] Walid Briki, Rethinking the relationship between momentum and sport performance: Toward an integrative perspective, Psychology of Sport and Exercise, Volume 30, 2017, Pages 38-44

[3] Jim Taylor & Andrew Demick (1994) A multidimensional model of momentum in sports, Journal of Applied Sport Psychology, 6:1, 51-70, DOI: 10.1080/10413209408406465

[4] Klaassen, Franc, and Jan R. Magnus, ’The method of moments’, Analyzing Wimbledon: The Power of Statistics (New York, 2014; online edn, Oxford Academic, 16 Apr. 2014)

[5] P. Gorgi, S. J. Koopman, R. Lit, The Analysis and Forecasting of Tennis Matches by using a High Dimensional Dynamic Model, Journal of the Royal Statistical Society Series A: Statistics in Society, Volume 182, Issue 4, October 2019, Pages 1393–1409

[6] Boehmke B , Greenwell B . Interpretable Machine Learning[M]. 2019.

[7] Scheibehenne, Benjamin, and Arndt Bröder, ’Predicting Wimbledon 2005 Tennis Results by Mere Player Name Recognition’, in Gerd Gigerenzer, Ralph Hertwig, and Thorsten Pachur (eds), Heuristics: The Foundations of Adaptive Behavior (2011; online edn, Oxford Academic, 1 May 2011)

[8] Wilkens, Sascha. ‘Sports Prediction and Betting Models in the Machine Learning Age: The Case of Tennis’. 1 Jan. 2021 : 99 – 117.

[9] Romain Gauriot, Lionel Page, Does Success Breed Success? a Quasi-Experiment on Strategic Momentum in Dynamic Contests, The Economic Journal, Volume 129, Issue 624, November 2019, Pages 3107–3136

To:coaches
From:Team 2408451
Date:2024.02.05
Subject:Giving some suggestions to coaches
Message:

Our team has recently analyzed some match data in tennis matches and would like to share our research findings with you to assist you in guiding players in the future.

From our research, we found that momentum is a key factor affecting the outcome of a competition, and it is closely related to many aspects of the field, such as consecutive scores, impressive performances, and mistakes, the success of breaking serve, and the distance a player runs.

These factors also greatly affect the direction of the competition, leading to the emergence of turning points, I would like to give you a few suggestions for players and tell the coach how you should train regularly:

In our research, it was found that continuous scoring in a short period of time can greatly improve one's own momentum and reduce the opponent's momentum. Therefore, in daily training, players should train themselves to score continuously when serving, first reduce the opponent's momentum , and improve their own momentum. This way, when receiving the ball, it is more likely to achieve breaking serve, and breaking serve has a significant positive effect on momentum in our research. Therefore, training a player's serving ability is very important.In regular training, coaches can motivate players to score continuously, such as organizing them to compete in consecutive scoring matches, where the winner is rewarded.

We have also noticed that impressive performances, such as obtaining an ace or failing to catch the ball when returning, have a significant positive effect on momentum. Therefore, players should think about how they can achieve such outstanding performance by serving or returning the ball. Coaches can also organize training specifically to practice such outstanding performance during regular training.

However, our research also found that widening the score gap, making mistakes, being snatched from the match point or being broken, these short-term disadvantages are not objectively significant , but in research, they often become turning points in a game, leading to the failure of the entire game. Therefore, players should maintain a calm mind towards these temporary failures and think about ways to break the deadlock instead of being defeated by frustration. Coaches should also pay attention to the subsequent performance of different players in the face of defeats during regular training, and provide them with timely guidance.

In addition, different players may have different main scoring periods, some players may enter the state slowly, while others may start scoring high but lack momentum. Coaches should grasp the characteristics of each player, provide targeted encouragement to them in the event of defeat, and give warnings in the event of victory.

These are some methods that I have developed through my research to improve the performance of athletes. I hope your athletes can become more and more outstanding!

<u>Report on Use of AI</u>

1. OpenAI *ChatGPT* (Nov 5, 2023 version, ChatGPT-3.5)

Query1: **Where can i get similar dataset for other matches?**

Output: To find similar datasets for other tennis matches, you can explore various sources, including sports data repositories, tennis organizations, and open data platforms. Here are some places where you might find tennis match datasets:

1. **ATP/WTA Official Websites:** The official websites of the ATP (Association of Tennis Professionals) and WTA (Women's Tennis Association) may provide official match data. Look for dedicated sections related to statistics and data.

2. **Grand Slam Websites:** Explore the websites of Grand Slam tournaments such as Wimbledon, US Open, French Open, and Australian Open. They often provide detailed match statistics and historical data.

3. **Kaggle:** Kaggle is a platform for predictive modeling and analytics competitions. You can find various datasets related to sports, including tennis matches. Search for tennis-related datasets on Kaggle.

4. **Sports Data APIs:** Some sports data providers offer APIs that allow you to access tennis match data programmatically. Examples include Sportradar, Stats Perform, and others.

5. **University Research Repositories:** University research repositories or data archives may have datasets related to sports analytics, including tennis match data.

6. **GitHub:** Check repositories on GitHub for sports-related datasets. Researchers and data enthusiasts often share datasets related to various sports, including tennis.

7. **Open Data Platforms:** Explore open data platforms like Data.gov or other regional open data portals. They may have datasets related to sports and tennis matches.

Remember to check the terms of use for each dataset to ensure compliance with any licensing or usage restrictions. Additionally, consider reaching out to sports analytics communities or forums where enthusiasts share and discuss datasets related to tennis matches.

Query2: **If there are two kinds of points, namely, positive and negitive, how can i show that in the legend?**

Output: To distinguish between positive and negative points in a scatter plot and display it in the legend, you can use different markers or colors for positive and negative points. Here's an example:

```python
import matplotlib.pyplot as plt

# Sample data
x_positive = [1, 2, 3, 4, 5]
y_positive = [10, 12, 5, 8, 20]

x_negative = [1.5, 2.5, 3.5, 4.5, 5.5]
y_negative = [-5, -8, -3, -10, -15]

# Create scatter plot for positive points
plt.scatter(x_positive, y_positive, label='Positive', color='blue', marker='o')

# Create scatter plot for negative points
plt.scatter(x_negative, y_negative, label='Negative', color='red', marker='x')

# Set labels for the axes
plt.xlabel('X-axis')
plt.ylabel('Y-axis')

# Show legend
plt.legend()

# Display the scatter plot
plt.show()
```

In this example, positive points are marked with blue circles (`'o'`), and negative points are marked with red crosses (`'x'`). The `label` parameter is used to assign labels to each set of points. The `plt.legend()` function is then used to display the legend.

You can customize the colors, markers, and other aspects of the scatter plot and legend based on your preferences.

Query3: how to sort by the absolute value of a column without altering it?
Output: To sort a DataFrame by the absolute value of a specific column without altering the original DataFrame, you can create a temporary column containing the absolute values and then sort based on that. Here's an example:

```python
import pandas as pd

# Sample DataFrame
```

```
data = {'Column_Name': [10, -5, 8, -15, 3]}
df = pd.DataFrame(data)

# Create a temporary column with absolute values
df['Absolute_Column'] = df['Column_Name'].abs()

# Sort the DataFrame based on the absolute values
sorted_df = df.sort_values(by='Absolute_Column', ascending=True)

# Drop the temporary column if needed
sorted_df = sorted_df.drop(columns=['Absolute_Column'])

print(sorted_df)
```

In this example, a new column 'Absolute_Column' is created using the `abs()` method on the 'Column_Name'. The DataFrame is then sorted based on this temporary column, and if you don't need the temporary column in the result, you can drop it using the `drop` method.

This approach ensures that the original DataFrame remains unchanged, and you get a sorted DataFrame based on the absolute values. Adjust the column names as needed for your specific DataFrame.

Query4: How to use SVM, LGBM and Decision Tree
Output: Certainly! Let's adapt the example to use Support Vector Machines (SVM), LightGBM, and Decision Tree as alternative models.

Using Support Vector Machines (SVM):

```python
from sklearn.svm import SVR

# Step 5: Choose a Model
model_svm = SVR()

# Step 6: Train the Model
model_svm.fit(X_train, y_train)

# Step 7: Make Predictions
y_pred_svm = model_svm.predict(X_test)

# Step 8: Evaluate the Model
mse_svm = mean_squared_error(y_test, y_pred_svm)
```

```python
print(f'SVM Mean Squared Error: {mse_svm}')
```

Using LightGBM:

```python
import lightgbm as lgb

# Step 5: Choose a Model
model_lgbm = lgb.LGBMRegressor()

# Step 6: Train the Model
model_lgbm.fit(X_train, y_train)

# Step 7: Make Predictions
y_pred_lgbm = model_lgbm.predict(X_test)

# Step 8: Evaluate the Model
mse_lgbm = mean_squared_error(y_test, y_pred_lgbm)
print(f'LightGBM Mean Squared Error: {mse_lgbm}')
```

Using Decision Tree:

```python
from sklearn.tree import DecisionTreeRegressor

# Step 5: Choose a Model
model_dt = DecisionTreeRegressor()

# Step 6: Train the Model
model_dt.fit(X_train, y_train)

# Step 7: Make Predictions
y_pred_dt = model_dt.predict(X_test)

# Step 8: Evaluate the Model
mse_dt = mean_squared_error(y_test, y_pred_dt)
print(f'Decision Tree Mean Squared Error: {mse_dt}')
```

Feel free to experiment with these models, tune hyperparameters, and compare their performance to choose the one that best suits your specific problem. Additionally, you can explore ensemble methods or more advanced techniques for further improvements.

Query5 : The NN model?
Output: For a neural network model, you can use the confusion matrix similarly as for other classification models. However, instead of using the `y_pred` directly, you might need to convert the model's probabilities into class predictions.

Here's an example using a neural network model with TensorFlow/Keras:

```python
import numpy as np
from sklearn.metrics import confusion_matrix
import seaborn as sns
import matplotlib.pyplot as plt
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense
from sklearn.model_selection import train_test_split

# Assuming you have already created and trained your neural network model (model_nn)
# Replace X_train, X_test, y_train, y_test with your actual variables

# Example data (replace this with your actual data)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Example neural network model
model_nn = Sequential([
    Dense(64, activation='relu', input_dim=X_train.shape[1]),
    Dense(1, activation='sigmoid')  # Assuming binary classification; adjust for
multiclass
])

model_nn.compile(optimizer='adam', loss='binary_crossentropy')  # Adjust for multiclass

# Train the model
model_nn.fit(X_train, y_train, epochs=10, batch_size=32, validation_split=0.2)

# Make predictions
y_pred_nn = model_nn.predict(X_test)

# Convert probabilities to class predictions (binary classification)
y_pred_nn_classes = np.round(y_pred_nn)

# Create and display the confusion matrix
confusion_mat_nn = confusion_matrix(y_test, y_pred_nn_classes)

# Display the confusion matrix as a heatmap
```

sns.heatmap(confusion_mat_nn, annot=True, fmt='d', cmap='Blues', cbar=False)
plt.xlabel('Predicted Labels')
plt.ylabel('True Labels')
plt.title('Confusion Matrix - Neural Network')
plt.show()
```

This example assumes binary classification, so it uses a sigmoid activation function in the output layer and rounds the predicted probabilities to obtain class predictions. Adjust the code accordingly for multiclass classification by changing the model architecture, output layer activation function, and how you convert probabilities to class predictions.

**Query6: Can i use permutation test to examine whether a list of num is generated at random?**
**Output:** Yes, you can use a permutation test to examine whether a list of numbers is generated at random. A permutation test is a non-parametric statistical method that assesses the null hypothesis by comparing the observed statistic with the distribution of the statistic under all possible permutations of the data.

Here's a general outline of how you could perform a permutation test to check if a list of numbers is generated at random:

1. **Define the Null Hypothesis:**
   - Null Hypothesis (H0): The list of numbers is generated randomly.

2. **Define the Test Statistic:**
   - Choose a test statistic that represents the aspect of randomness you want to test. For example, you might choose the sum of the numbers.

3. **Calculate the Observed Test Statistic:**
   - Compute the observed test statistic from the original list of numbers.

   ```python
   import numpy as np

   observed_statistic = np.sum(your_list_of_numbers)
   ```

4. **Permute the Data:**
   - Create a large number of permutations of your list of numbers and calculate the test statistic for each permutation.

   ```python

permutations = [np.random.permutation(your_list_of_numbers) for _ in range(num_permutations)]
```

5. **Compare Observed vs. Permutated Test Statistic:**
- Compare the observed test statistic to the distribution of permuted test statistics. Determine the p-value, which represents the probability of observing a test statistic as extreme as the one calculated from the original data, assuming the null hypothesis is true.

```python
permuted_statistics = [np.sum(permuted_list) for permuted_list in permutations]

p_value = (np.sum(permuted_statistics >= observed_statistic) + 1) / (num_permutations + 1)
```

Here, `num_permutations` is the number of permutations you generate.

6. **Interpret Results:**
- If the p-value is below a chosen significance level (e.g., 0.05), you may reject the null hypothesis and conclude that the list of numbers is not generated at random.

Keep in mind that the choice of the test statistic and the significance level is crucial and depends on the specific aspects of randomness you are interested in. Also, the permutation test is computationally intensive, especially with a large number of permutations, so be mindful of performance considerations.
```