Problem Chosen	2024 MCM/ICM Summary Sheet	Team Control Number
C		2423370

The Victory Weapon: Quantified Momentum in Tennis Matches

Summary

In competitive tennis matches, understanding game dynamics, particularly **momentum shifts**, is a crucial goal for professional athletes. With the advancement of machine learning algorithms, researchers in related fields are actively seeking versatile and effective **quantitative methods** to capture and forecast changes in match momentum. In response to this challenge, our team has developed a predictive model that combines a **linear model** with the **Random Forest algorithm**.

To start with, our fundamental model inherits a classic featuring method, the linear model, which consists of the 6 factors most relevant to winning, indicated by **PCA analysis**. Based on that, we add **server advantage** co-factors and terms. Co-factors are adjusted to ensure that the final momentum aligns with actual match scores, making this model highly sensitive to changes in match dynamics. Thus, this model is highly sensitive to the situation change of the match.

Leveraging the linear model, we can identify every momentum swing points in each match. To reveal whether these swing points occur randomly, we apply the **Runs Test** to its origin - the momentum change of each player. The **negative** result implies that it is possible to **predict** whether swings will occur. However, predictions based on the linear model require keeping track of a large quantity of data, which is not appealed in real scenarios.

On top of that, since the momentum shifts are usually occurred due to unexpected mistakes on the advantaged side or successful assaults on the backward side, we develop a machine learning prediction model based on related **transient factors**. This model requires only data from one most recent game. After that, we apply **TOPSIS** to analyze the major impacts behind the swings.

Lastly, we rigorously assess the overall sensitivity, robustness, and generalization capabilities of our models. Our linear model incredibly demonstrates impressive stability in capturing match trends and swing points, even when subjected to Gaussian noise. Furthermore, we successfully apply it to the 2023 Wimbledon Women's Tennis Match, capturing both overarching trends and match-specific details, while the classifier maintains highly accuracy in identifying swing points. Consequently, our model exhibits high sensitivity, robustness, and excellent generalization ability.

Keywords: Momentum; Linear Model; Runs Test; Random Forest; TOPSIS

Contents

1 Introduction

1.1 Background

The 2023 Wimbledon Men's Final, featuring the face-off between Spanish rising star Carlos Alcaraz and veteran Novak Djokovic, marked a seminal event in tennis history. The performance of both players demonstrated dramatic fluctuations, with numerous shifts in the lead during each set, a phenomenon often attributed to "momentum" in sports competitions, which is typically defined as "strength or force gained by a series of events or by motion."

In competitive sports, a team or an individual player may feel that they possess momentum or an advantage during a game.Despite the challenges in quantifying momentum, its tangible presence in games significantly influences the direction of the match and the mindset of the players[1]. The thrilling moments of the 2023 Wimbledon Men's Final provide a prime case study for examining the role of momentum in tennis matches and even more competitions.

1.2 Problem Restatement

The 2023 Wimbledon Men's Final offers valuable data and perspectives for a deeper exploration of the patterns and laws governing momentum shifts in sports competitions. Based on the given data, we are going to finish the following tasks:

1. Considering the advantage of serving, develop a model that identifies the better-performing player and how much better at a given time, in which we make some inferences to help modeling. And we give a visualization to depict the match flow.

2. Use the model to check if swings in play and runs of success by one player are random.

3. Identify factors related to the degree of inclination in the balance of victory in the match. Firstly, utilize the provided data for at least one match and build a model predicting swings in a match and check if there are most related factors. Then based on the swings of past matches, give a player advice on how to do in a match against a different player.

4. Test the developed model on one or more matches and its reliability on other matches such as Women's matches and table tennis.

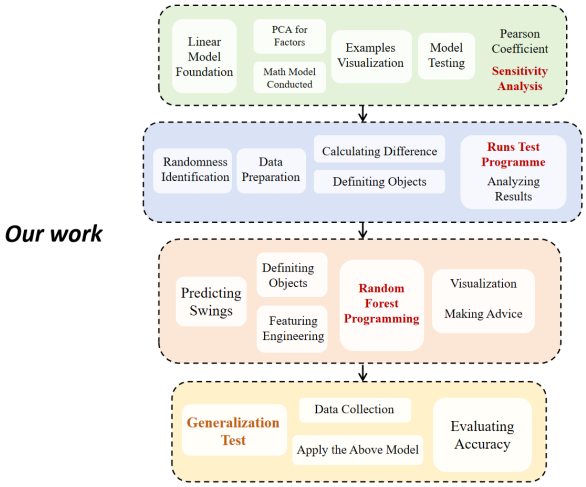

Figure 1: Our work

2 Terminology, Symbols and Assumptions

2.1 Terms

- **Unforced Error[2]:** an unforced error in tennis is a mistake made by a player that was not prompted by the opponent's actions but is solely the result of the player's own shortcomings or errors.
- **Close to line(CTL):** It refers to hitting a shot very near the boundary lines that mark the edges of the tennis court.
- **Deep Return(D)**: A strategy that returns the ball near the end of the court.

2.2 Symbols

See Table 1.

Table 1: Variable description

Symbol	Definition
t	elapsed time
P_t	Momentum of the current time point
P_{t-1}	Momentum of the previous time point
PA	Points advantage
SA	Advantage of sever
UE	Number of unforced error times
BP	Breaking point
WP	Number of winning points
WG	Number of winning games
WS	Number of winning sets
num(A)	Numbers of type A balls
SD	Serve depth
RD	Return depth
CR	Breakpoint converting ratio
CP	Consecutive points
EF	Effective first serve
ACE	Ace rate

2.3 General Assumptions

- **Player Performance Consistency Assumption**:
  Players' skill level and physical condition remain relatively stable during the match, without

significant changes due to non-match factors like sudden injuries.

- **Serving Advantage Assumption**:
  The serving player has an inherent advantage in the game, which can be quantified based on historical data.
- **Environmental and Court Condition Consistency Assumption**:
  It's assumed that external conditions of the match (like weather, humidity) and the court conditions (like grass, hard court) have a consistent impact on all players.

3 Part 1: Modeling and Visualization

3.1 Bulid a Model about Momentum

The Momentum should effectively encompass both the overarching patterns and the intricate fluctuations that occur throughout the course of the match as time progresses.As the first rally plays very important role, as over 1/3 first points are won directly by serving, and over 1/4 first points are won directly by returning.

First we use index 0 and 1 to denote the victory of a certain player. By **PCA**(Principal Component Analysis), some positive-correlated factors to ΔP_t are: SA, BP, WS, WG, WP, the only negative-correlated factor is UE.

Based on the given data and by calculation, **num(CTL) /num(All Serve Balls) = 0.4289, num(D)/num(All Return Balls) = 0.3259**, both can be considered as a powerful strategy for servers or returns. Therefore, we conclude that the first rally has determining impacts on the game trends.

The momentum effected by the first ball at the server side is defined as

$$P_{0,serve} = C_s * SD \tag{1}$$

where $C_s = x$ is a co-factor to be determined and SD is defined as

$$SD = \begin{cases} 2 & \text{Serve depth is CTL} \\ 1 & \text{Serve depth is NCTL} \end{cases}$$

While the Momentum effected by the first ball at the returner side is

$$P_{0,return} = RD \tag{2}$$

where

$$RD = \begin{cases} 2 & \text{Return depth is D} \\ 1 & \text{Return depth is ND} \end{cases}$$

Hence, we define the contribution to one player's momentum as:

$$\begin{aligned} P_{pi,t} = P_{pi,t-1} + SA_{pi} + \sum_{j \in set} BP_{pi,j} + \sum_{j \in set} WS_{pi,j} + \sum_{j \in set} WG_{pi,j} \\ + \sum_{j \in set} WP_{pi,j} - \sum_{j \in set} UE_{pi,j} + P_{0,pi} \quad i = 1, 2 \end{aligned} \tag{3}$$

or equivalently,

$$\begin{aligned} P_{pi,t} = &\sum_{match} SA_{pi} + \sum_{sets}\sum_{j\in set} BP_{pi,j} + \sum_{sets}\sum_{j\in set} WS_{pi,j} + \sum_{sets}\sum_{j\in set} WG_{pi,j} \\ &+ \sum_{sets}\sum_{j\in set} WP_{pi,j} - \sum_{sets}\sum_{j\in set} UE_{pi,j} + \sum_{sets} P_{0,pi} + P_{pi,0} \quad i = 1,2 \end{aligned} \tag{4}$$

where

$$SA = y * \sum_{set} num(\text{served by pi}) \quad i = 1,2$$

Now we solve for x, y, treating each match as a restriction.
Define $\phi(m)$ as

$$\phi(m) = \begin{cases} 1 & \text{Player 1 won the match m} \\ -1 & \text{Player 2 won the match m} \end{cases}$$

Now the problem is equivalent to

$$\text{s.t.} \begin{cases} \text{determine } (x,y) \in \mathbb{R}^2 \\ \begin{bmatrix} P_{p1,1} - P_{p2,1} \\ \cdots \\ P_{p1,m} - P_{p2,m} \end{bmatrix} \cdot \begin{bmatrix} \phi(1) \\ \cdots \\ \phi(m) \end{bmatrix} > 0 \\ min(|E(x) - \dfrac{1}{0.5711}|) \end{cases} \tag{5}$$

Solving for the multi-objective planning, we found that this restriction is rather loose. Thus, for convenience, we let

$$\begin{cases} x = 1.6 \\ y = 0.1 \end{cases}$$

i.e.

$$P_{0,serve} = 1.6 * SD$$

$$SA = 0.1 * \sum_{set} num(\text{served by pi}) \quad i = 1,2$$

3.2 Visualize the Final Match

We have gained a model about the players' momentum. Here we apply the model to analyze the contest mentioned in the background: The 2023 Wimbledon Men's Final, Carlos Alcaraz v.s. Veteran Novak Djokovic.

Initially, utilizing the precious model, a scatter plot was constructed, denoted as Figure 1, in which blue and red dots represent the momentum of Alcaraz and Djokovic, respectively. This visulization distinctly illustrates the momentum fluctuations experienced by each player throughout the match. Subsequently, the variation in momentum at specific instances is conveyed through the gradation of dot colors within the plot, since the change speed increases as the color gets deeper.

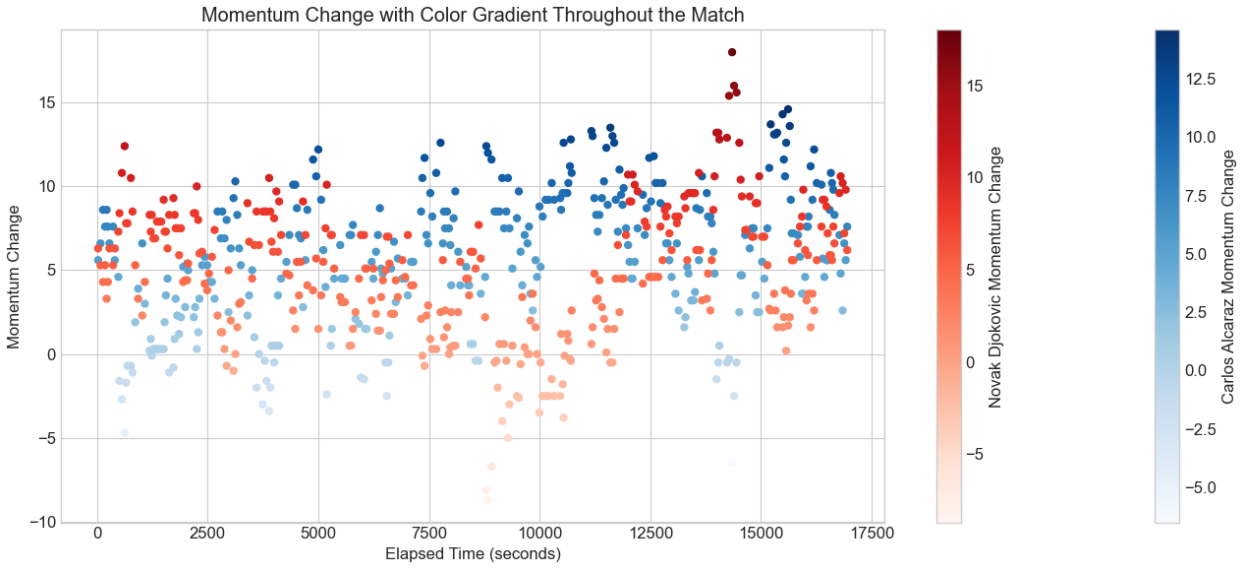

Figure 2: Scatter diagram with gradient colors

To clearly show the changes of momentum, a line chart is made. As depicted in Figure 2, Alcaraz's and Djokovic's momentum are represented by blue and red curves respectively, with match points indicated by blue dots.

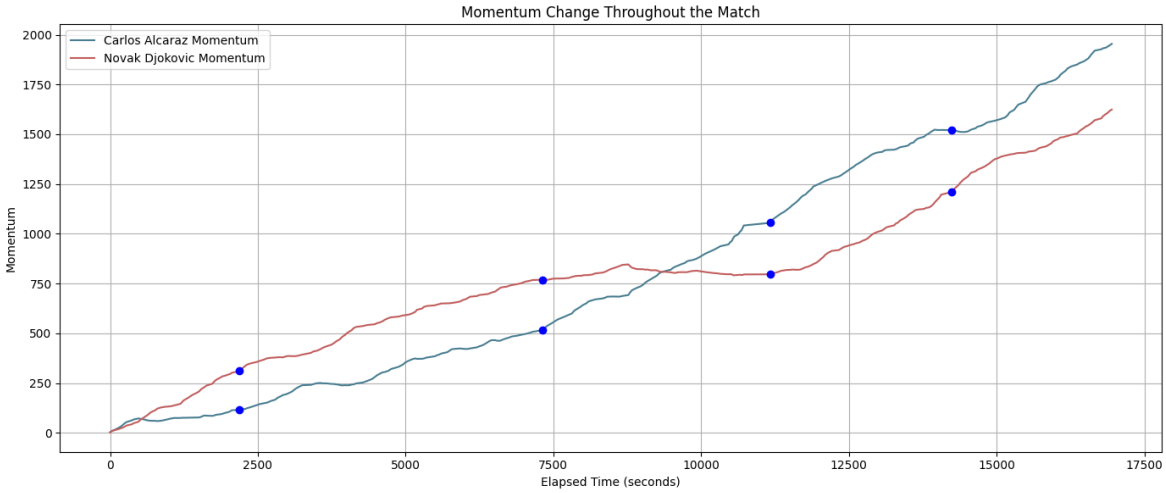

Figure 3: Line chart about momentum change

To test the momentum's effect on match's trend, we define the game's momentum is determined by subtracting Player 2's momentum from Player1's. As shown in Figure 4, a positive value suggests an advantage for Player 1 in upcoming plays, while a negative value indicates an advantage for Player 2.

We found that the model's outcomes align well with the actual match events, particularly noting Alcaraz's **(blue)** substantial lead in the latter half of the games.

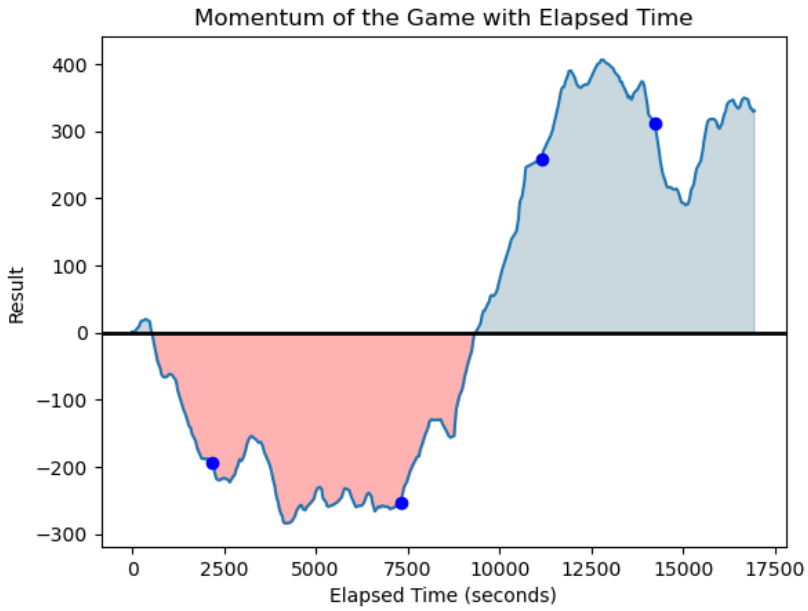

Figure 4: Game momentum change

3.3 Model Evaluation

3.3.1 Correlation and Feature Importance Analysis

By powerful tool computer program of calculation, we debugged this model by the whole data. Now we are going to discover its reliability by calculating each variable's relevance to the defined player's momentum in the match. We use the Pearson correlation coefficient to measure the relevance, whose value is between 1 and -1. A correlation coefficient closer to 1 or -1 indicates a stronger positive or negative linear relationship, respectively. The mathematical calculation formula of the coefficient is :

$$r = \frac{N\sum(xy) - \sum x \sum y}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}}$$

In this formula:

r represents the Pearson correlation coefficient. N is the number of observations. $\sum(xy)$ is the sum of the products of variables x and y. $\sum x$ is the sum of variable x. $\sum y$ is the sum of variable y. $\sum x^2$ is the sum of the squares of variable x. $\sum y^2$ is the sum of the squares of variable y.

Consequently, utilizing the above calculated coefficients, we made a **sankey diagram** to visually represent how various factors we used to quantify the momentum influence two players' momentums.

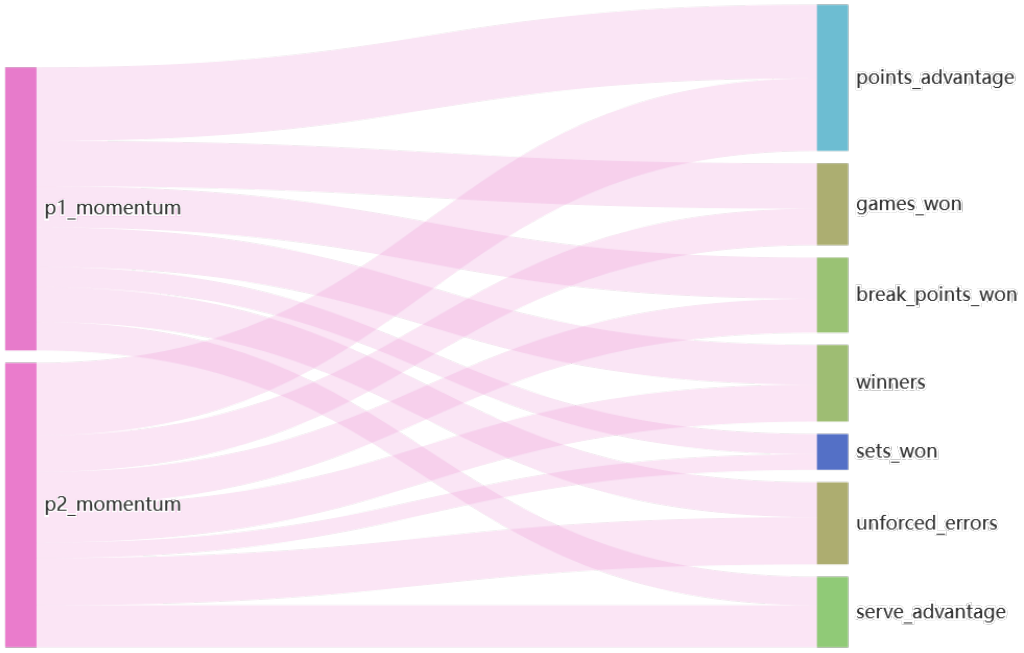

Figure 5: Sankey-simple

In the Sankey diagram, the width of the bands is indicative of the magnitude of flow, which in this context refers to the absolute value of the correlation. Analyzing this sample, we can gain some correlation and feature importance about our quantify model:

- The factor points advantage has the most significant positive impact on the momentum since it is indicated by the widest band.
- The factors games won, break points won, winners, serve advantage and unforced errors also positively influence the momentum.
- The influence of the factor sets won is relatively minor in comparison.

The diagram clearly shows the contribution of different variables to the momentum, suggesting these variables are important in the quantification process and align with our expectations and validates the effectiveness of the quantifying method.

3.3.2 Robustness and Sensitivity Analysis

In this section, we first add **Gaussian Noise** with $(\mu = 0, \sigma = 2)$ to "point_victor". Then we applied our method to noised data to compare with the original data. As Figure6 shows, the trend feature remains significant, the twists and turns, however, become inaccurate. The result proves that our method is both sensitive to the match data and resistant to noise to some extend.

The result of applying our method to noised data (the final match) is provided below.

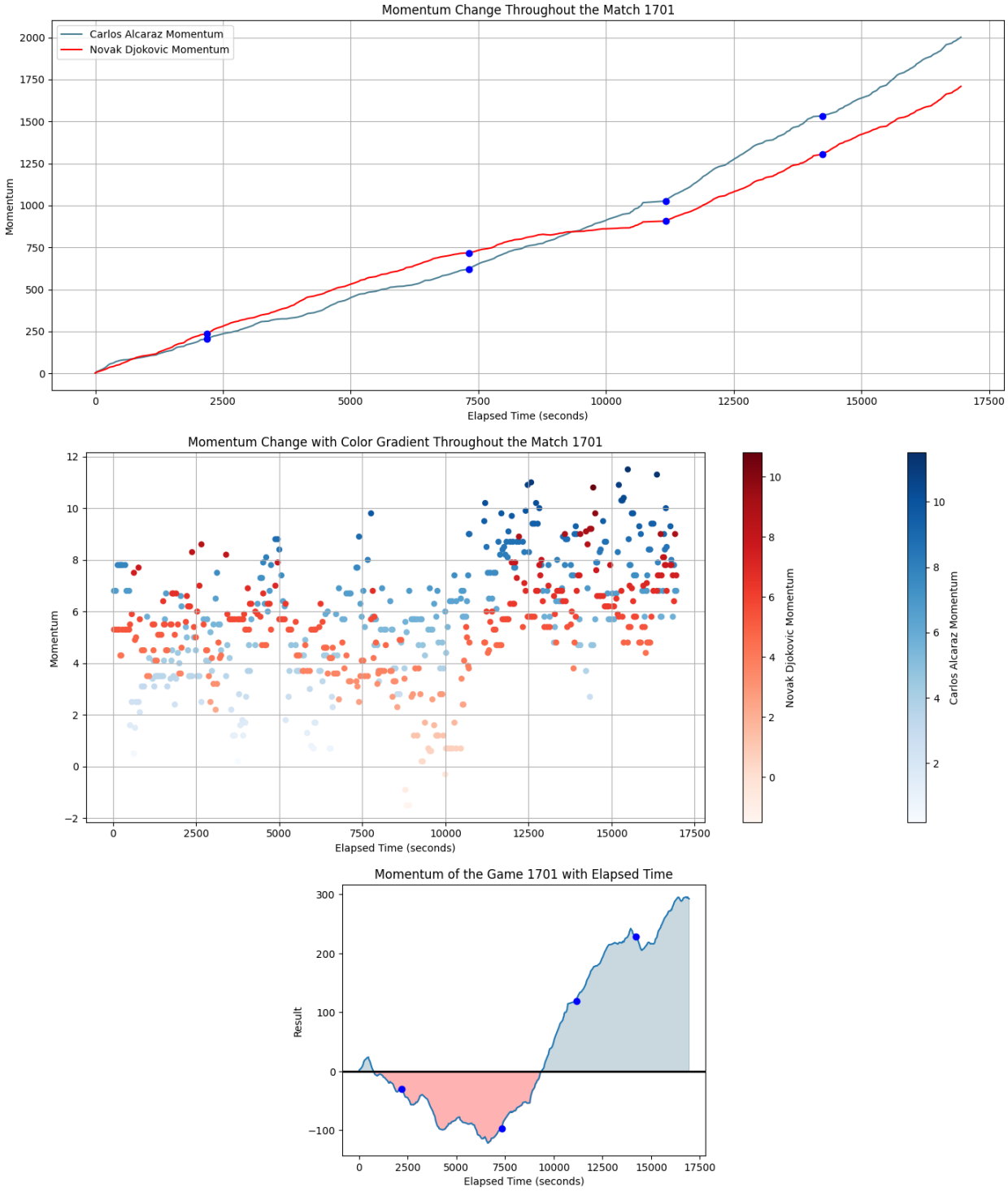

Figure 6: Noised Data

4 Part 2: Test of Randomness

In this part, we evaluate whether the fluctuations in play and sequences of success by one player are random. For this purpose, we utilized the Wald-Wolfowitz Runs Test, a non-parametric statistical method designed to assess the randomness of a sequence.

4.1 Data Preparation

Since we are considering the randomness of performance, we can use the momentum difference to denote the level of the player's current performance. Here we employ the model of Part 1 to visualize the momentum change in the final match.

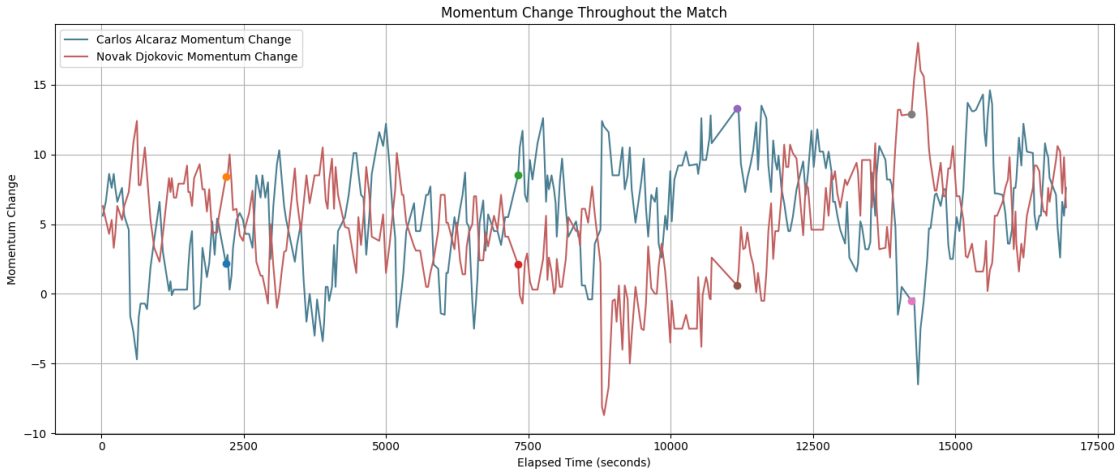

Figure 7: Momentum Difference Chart

The time series was first aligned with the match timeline by assigning an elapsed time in seconds to each recorded change in momentum. To preserve the integrity of the data, we addressed missing or incomplete entries by removing them and then employed interpolation to fill the gaps. This process ensured the continuity and completeness of the series. The cleaned and processed time series was preserved, securing the refined data for future analyses.

4.2 Application of the Runs Test

After the data preparation phase, the refined momentum change data for Player 1 was subjected to the Wald-Wolfowitz Runs Test. This non-parametric test is instrumental in evaluating the randomness within a sequence of observations. It does so by analyzing the arrangement and succession of data points relative to a central value, which, for this analysis, was set at zero. The central concept of the test revolves around the notion of 'runs'. A 'run' is defined as a sequence of consecutive data points all above or all below the central value, with no crossover in between.

With the data duly prepared, we directed our focus towards the application of the Wald-Wolfowitz Runs Test to the refined momentum change data for Player 1. This non-parametric test stands out for its ability to critically evaluate the randomness within a sequence of observations. It achieves this by methodically analyzing the arrangement and succession of data points in relation to a predetermined central value. For the purposes of this analysis, the central value was set at zero.

The fundamental concept driving this test is the notion of 'runs.'The analysis unfolds by first categorizing each data point in the series as either positive or negative. Following this categorization, the test proceeds to count the number of runs present in the data. This observed number of runs is then meticulously compared to the expected number of runs under the null hypothesis, which posits that the sequence of observations is random.

The test then compares the observed number of runs against the expected number under the null hypothesis of randomness. The expectations and variance are calculated based on the following standard formulas:

$$E(R) = 1 + \frac{2n_1n_2}{n_1 + n_2}$$

$$Var(R) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}$$

In these expressions, n_1 and n_2 respectively represent the counts of positive and negative points. The culmination of this analytical process is the calculation of the test statistic Z, which is achieved by standardizing the observed number of runs:

$$Z = \frac{R - E(R)}{\sqrt{Var(R)}}$$

Where R is the observed number of runs.

Under the null hypothesis of randomness, Z follows a normal distribution. Therefore, the significance of the observed sequence can be assessed by comparing the test statistic to a standard normal distribution, resulting in a **p-value**.If the p-value is less than or equal to the significance level(α), we reject the null hypothesis. Conversely we fail to reject the null hypothesis.

4.3 Results

Through our calculation program, the final p-value is 2.43×10^{-10}, which is less the significance level($\alpha = 0.05$) by comparison. So we reject the null hypothesis. That is to say, **the hypothesis of randomness does not hold and swings in play and runs of success by one player are not random.**

In our analysis, the computation of the test statistic and the corresponding p-value based on the above formulas revealed a statistically significant deviation from the expected behavior under the hypothesis of randomness. The notably low p-value indicated a strong likelihood of the observed sequence not occurring by random chance, thereby suggesting a pattern or dependency within the momentum change data.

5 Part 3: Search for Indicators and Make advice

5.1 Problem Analysis

Now we are curious about indicators related to swings in a match. Since swings are not directly shown in the data or previous model, firstly we should find out what the swings in a match are. We have seen fluctuations in a match from the previous part while the players' momentum changes match the flow of the match. Hence, **we define swings as changes of momentum** while the **swing point is referred to the moment when the momentum crosses x-axis, aka the match flows to another player**. As depicted in **Figure 4**, there are three swing points where the momentum of game has a shift in positiveness or negativity .

Similar to Figure 3, the game momentum change situation of several matches are visualized in the following figure:

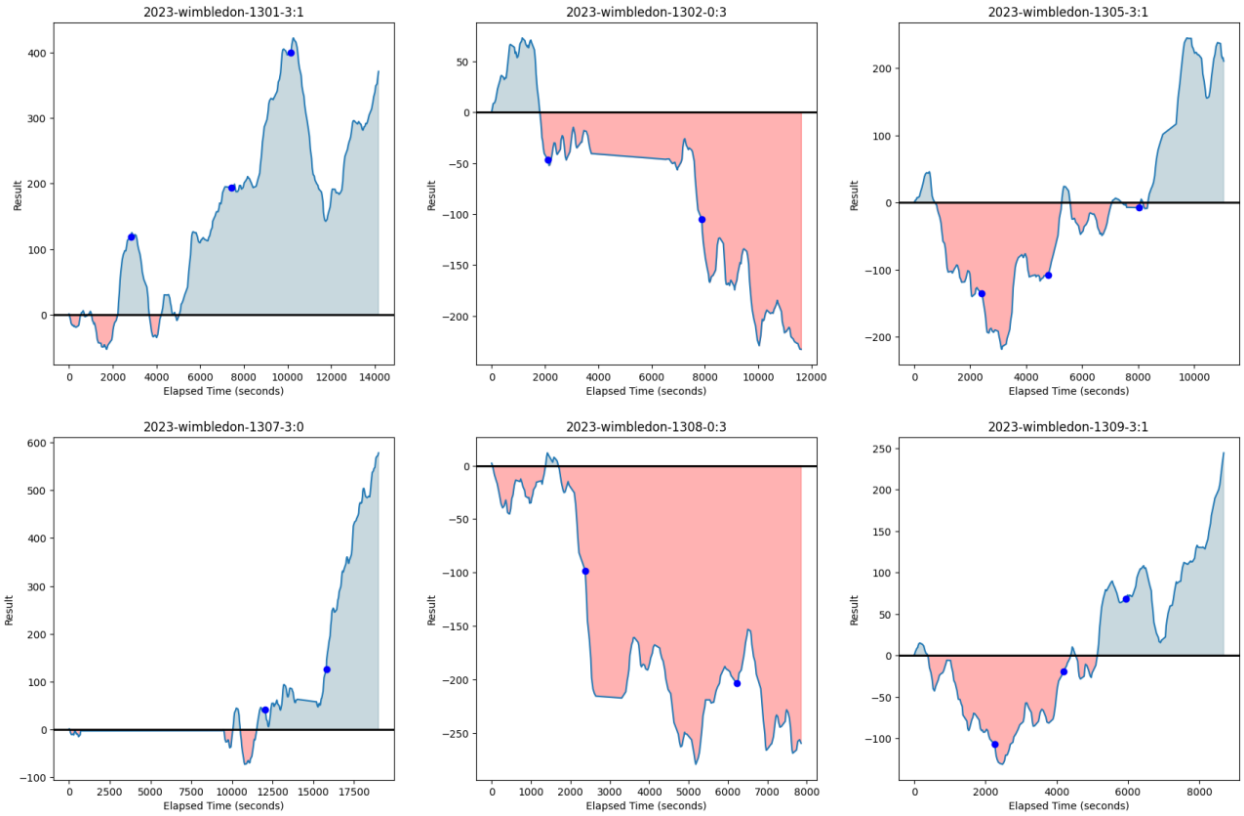

Figure 8: Game Momentum Examples

In these figures, with blue points denoting to the point where a set ends, the curve can be used to tell the flow of the match and at every swing point the advantage of matches changes.

From Part 2's test we attained an idea that there is a pattern or dependency within the momentum change and Part 1 provides us with a method to quantify momentum. It is time to search for some

important indicators with relation to the changes of momentum to get a deeper understanding of the pattern.

5.2 Feature Engineering

The original data provides extremely detailed information about each rally and it is not so convenient to directly cope with. Hence, we are going to extract more complicated features by featuring engineering to help searching for indicators.

Start from original variables from Part 1, the more complicated factors we are going to use are as follows:

- **Cumulative unforced errors**: The sum of a player's unforced error numbers from the start of the match to the on going point game, calculated by $\sum UE$.
- **Break point convert ratio**: The ratio of a player's total won break points over the total break points, calculated by $\sum BP_x won / \sum BP_x$.
- **Consecutive points**: The number of the consecutive points won by a player until the on-going game.
- **Effective first serve**: The ratio of the effective first serves over a player's total first serves, calculated by $\sum Serve_x / \sum Serves$.
- **Ace rate**: The ratio of the Ace serves over a player's total effective first serve, calculated by $\sum Win_x / \sum Serve_x$.

The selection of these advanced factors was conducted through our rigorous discussion and research, whose primary rationale lies in the fact that, although the raw data is comprehensive, directly utilizing this data for model training and analysis may not be sufficiently effective or intuitive. This approach is able to refine the predictive accuracy and interpretability of the model. Specifically:

- **Cumulative Unforced Errors**

  1.Attained by summing the total number of unforced error from the beginning of the match to the current point, this factor reflects a player's stability and resilience under pressure.

  2.Since the cumulative count of unforced error provides a crucial indicator of a player's psychological state and subsequent performance, it serve as a key factor in understanding the rhythm of the match and the player's condition[3].

- **Break Point Conversion Ratio**:

  1.This metric represents a player's performance during critical points in the match, particularly during break opportunities, and is one of the most telling indicators of a player's ability to perform under high-pressure situations.

  2.Additionally,he ability to capitalize on break points often dictates the overall direction of the match and the psychological dominance between the competitors[4], making it a pivotal factor in the match's dynamics.

- **Consecutive Points**:

  The count of consecutive points a player wins directly represents "dynamic trend" which is a key for identifying short-term trends and the game momentum shifts, essential for understanding and forecasting the game's progression.

- **Effective First Serve Rate**:

  A player's first serve success rate not only denotes their serving skill but also their overall command over the match. Effective first serves grant players an upper hand and marks a substantial competitive edge.

- **Ace Rate**

  Aces, which are immediate point-scorers in tennis, indicate a player's powerful serve and offensive capability. By computing the percentage of aces relative to successful first serves, we can get an essential factor in analyzing the momentum.

And the symbols we added are in Table 2:

Table 2: Variable description

Symbol	Definition
UE	Number of unforced error times
CR	Break point converting ratio
CP	Consecutive points
EF	Effective first serve
ACE	Ace rate

Through strict computer program calculation, we attained a new CSV file with the statistics of the quantified changes of players' momentum, UE, CR, CP, EF, ACE based on each point number. And the target indicators are going to be selected among the five factors.

5.3 Model Development

Since we have determined the influential factors, we are going to build a model to evaluate whether a swing will happen, which only require the data from nearest one single point of match.

5.3.1 Preparation of modelling

We defined a function f to examine whether current match is at the swing point:

$$f(t) = \begin{cases} 1 & \text{t is a swing point} \\ 0 & \text{t is not a swing point} \end{cases}$$

With momentum data prepared, we are able to identify every swing point. To derive the same result from the match data, we define another binary function g as

$$g(UE_{pi,t}, CR_{pi,t}, CP_{pi,t}, EF_{pi,t}, ACE_{pi,t}) = \begin{cases} 1 & \text{t is a swing point} \\ 0 & \text{t is not a swing point} \end{cases} \quad (i = 1, 2)$$

5.3.2 Random Forest Algorithm

Since the relationship between the parameters and the result is implicit, **Random Forest Classifier** is applied to determine g so that agree with f.

Random Forest Classifier is a widely utilized ensemble machine learning algorithm[5], and its adaptability to diverse domains and suitability for handling high-dimensional datasets make it a prominent choice for addressing complex classification challenges in the academic and practical domains. Therefore, we trained a Random Forest Classifier on 6557 data samples. Moreover, we applied **TOPSIS** (entropy weight method) to analyze the influence of each factor.

5.3.3 Results

Utilizing the present model, we are going to decide the most related factors to the swings in a match.

In our case, we first calculate $f(t)$ for each game. Then we randomly divided the data into a training set(90% of total dataset) and a test set. With careful hyper-parameter tuning, the average accuracy on the test set is **0.97803**. The error rate $\epsilon < 0.05$ and **AUC=0.82** (Figure10) which demonstrated that our classifier is effective.

On top of that, we calculate each factor's importance through **the entropy weight method**. The results are shown in the following table:

Player\Factor	Unforced Error	Ace Rate	Consecutive Pts	Effective First Serve	Break_pt CR
Player 1	0.18037	0.10822	0.07920	0.05166	0.03602
Player 2	0.26495	0.04540	0.01460	0.15437	0.06574

From the table and comparison, the most related indicator to the swings of a match is the unforced error(UE). Notice that player1 is always the first one to serve throughout the match.

5.3.4 Visualization

To more clearly gain the related importance of these factors, some visualization steps are conducted.

A demonstration of one decision tree (the estimator of Random Forest Classifier) and the ROC curve is provided below.

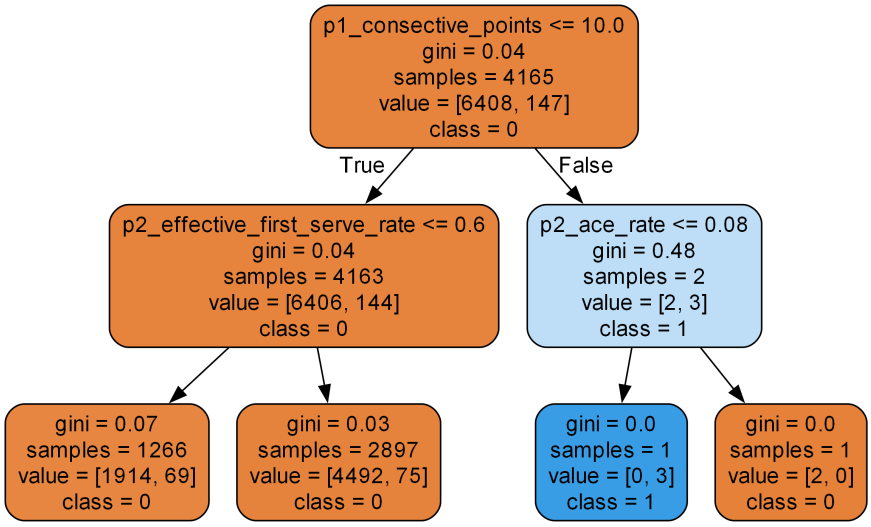

Figure 9: A decision tree

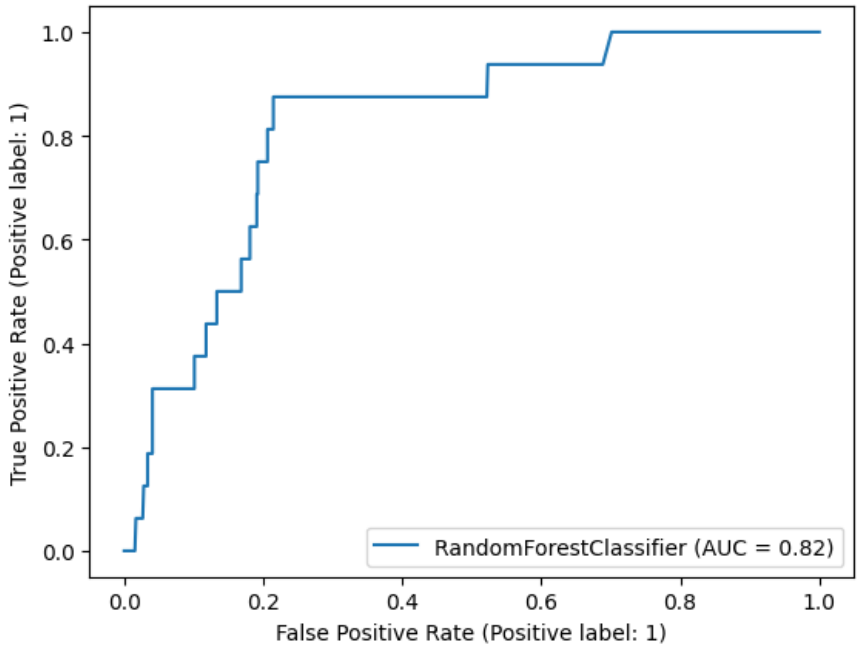

Figure 10: ROC Curve

Here we visualize the importance table with a pie chart where each distinct feature is matched with a circular sector:

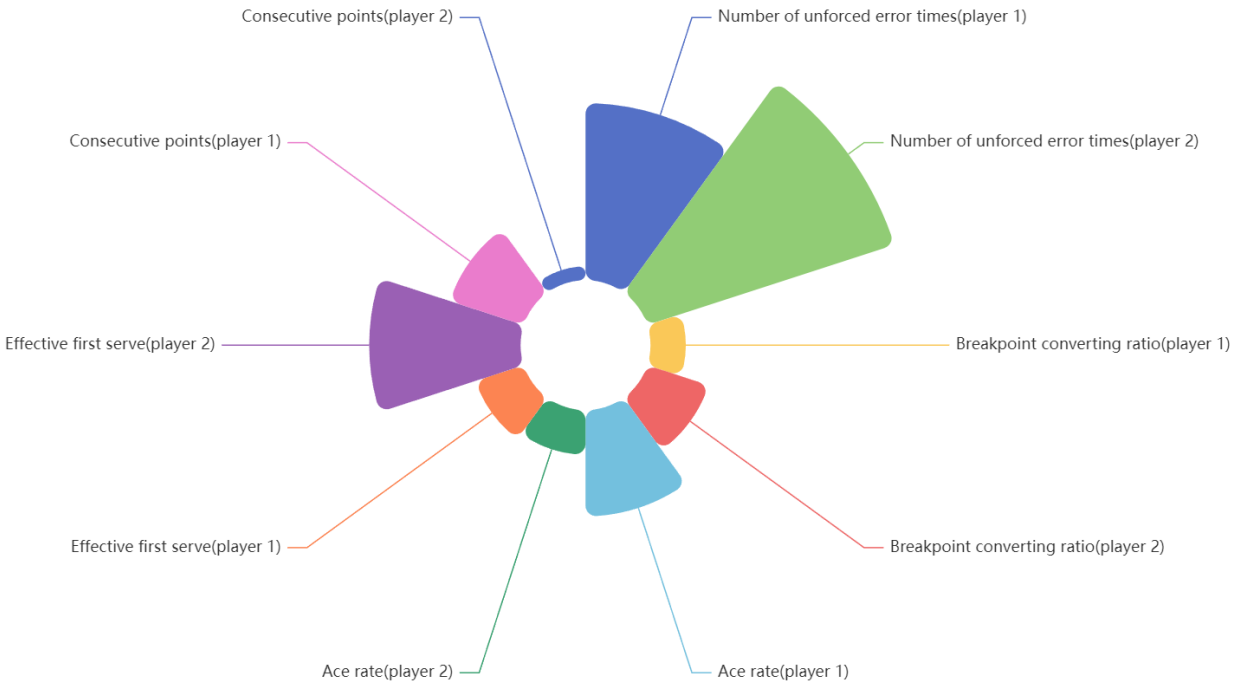

Figure 11: Importance Pie Chart

From Figure 7, the related importance of each is depicted by the circular's area.

5.4 Advice to Players

Based on our previous results, here we conclude the advice from identified key parameters that are sorted by the importance. We analyze this step by step. First, please focus on the unforced error, which plays the most important role in swings in the match and that's why the player is supposed to concentrate on avoiding making enforced error. In this way, the player may adopt a more conservative game strategy to avoid unnecessary risks. And the player had better go through simulations under high pressure before the formal match, aiming at improving the stability of playing techniques.

Next, let us turn to server advantage. Due to the existence of the obvious advantage of serving, improving the quality of the serve is crucial for enhancing the momentum and finally becoming the winner. The success rate of the first serve can significantly increase the probability of scoring. The player is supposed to focus on training the serving techniques under different match conditions, including speed and accuracy, especially the effectiveness of the first serve. If the player can make more ACE balls than the opponent, there is no doubt he takes great advantage in the tennis match.

Then, let us focus on the parameter break point converting rate which reflects the player's ability to capitalize on opportunities of crucial moments. It means a lot at the crucial point's momentum change. And this can be enhanced by large amounts of simulations.

Combining these factors, players can do the following preparations for new matches :

Psychological Preparation:

The shift in momentum is not just about technique and strategy, but also involves psychological factors. The players need to focus on strengthening mental training to improve psychological consistency for calmness during matches, particularly when the momentum is likely to shift and the pressure is the highest the player goes through.

Opponent's and Personal Data Analysis :

Players can firstly analyze the data from past matches of the opponent they are about to face, especially focusing on their weaknesses, such as high rates of unforced error or a low effective first serve rate.

Also, players should analyze the situations of momentum shifts in their own past matches to understand when they are easier to make mistakes or when they can better turn the situation around for further training.

Targeted Training:

Now the player has done opponent's analysis, he can do targeted training by simulating matches with the opponent's weaknesses. For example, if the opponent has a low break point conversion rate, the player can increase his own serve practice during crucial points to improve service hold rate. The player is supposed to improve the consistency of baseline play in the training, identify and exploit opportunities for consecutive scoring in the match.

Strategy Adjustment:

During the match, players can adjust strategies in real-time based on the opponent's performance. Following this principle, players can disrupt their opponents' rhythm by many methods like varying the ball's trajectory or spin, which can increase the probability of winning.

Physical Preparation:

Since the shift in momentum may require a long time, a player's physical training is also very important to ensure maintaining a high level of performance throughout the match. Players are supposed to do more physical training.

6 Part 4: Evaluate the Model's Generalization

In this section, in order to evaluate the generalization performance of the model, it was applied to the 2023 Wimbledon Women Match. The result turns out that both the Momentum model and the classifier worked as expected. The Momentum model once again successfully captured trends and twists of the match, demonstrating high consistency with the scores. Meanwhile, without modifying the classifier, it performed at the accuracy of **0.9828** on women's marches dataset. Both the scores indicate the great capability of our model in generalization.

6.1 Women's Matches

We first downloaded data from the Wimbledon Championship official site [https://www.wtatennis.com/], then we applied the same Momentum algorithm to the match data, the result is provided below. As the figure shows, the Momentum agrees with the final result. However, even though the classifier provided with high accuracy, the **AUC** rate drops to 0.69, which implies further training is required.

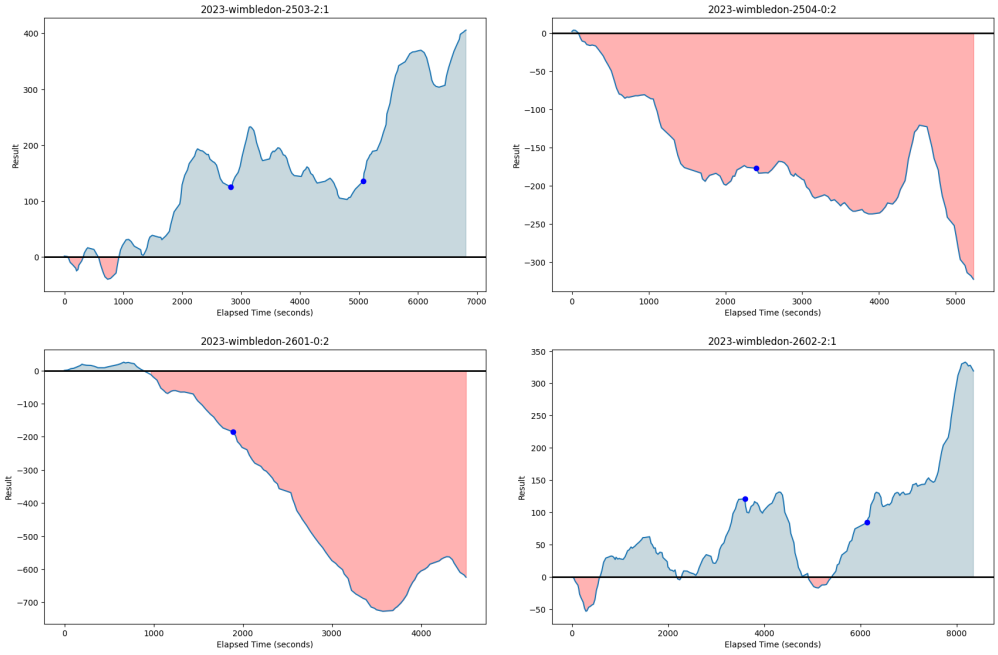

Figure 12: Overview

7 Strengths and weaknesses

7.1 Strengths

- Considering and testing physical and psychological aspects of tennis matches, we build the model to quantify dynamic momentum, which adequately utilized the whole real-match data so that we ensured empirical grounding, enhancing the model's relevance and accuracy.
- Our model gives us interpretable indicators for people to understand tennis even other matches, thus giving some inspiration to the audience and athletic teams.Additionally, it represents the important order of player's technique factors which leads to more advanced aspects of tennis sports evaluations.
- Our second model is incredibly accurate in predicting swings in the tennis match, which can be used to judge the flow of matches.
- Beyond theoretical modeling, we provide actionable insights for players and coaches, directly applicable to training and match strategies, highlighting its utility.
- Our model has an acceptable ability for generalization and can be optimized so that it can be used to quantify momentum in other types of matches.

7.2 Weaknesses

- Though the models are comprehensive, the complicated explanations might block understanding and practical application.

- Since the models are only based on several tennis matches and some factor values are not properly gathered to statistics or calculated thus leading to limitations of our models. If the model pattern is going to be used more widely, it needs more refinement and training.
- Despite addressing psychological aspects like unforced errors, the model might not fully encapsulate the nuanced psychological and emotional factors in tennis, possibly affecting its accuracy and comprehensiveness.
- Due to limited data, we made some ideal assumptions like the environments' equal impact on both players. There are some more factors need to be considered.

8 Conclusion

The 2023 Wimbledon's Gentlemen's final provides us a visual feast with unpredictable fluctuations during the match. With the utilization of Quantifying, Runs test, Random Forest Algorithm and kinds of visualization methods, our work checked the notion of tennis players' momentum and used a large number of factors to reasonably quantify the momentum, in order to attain the model of better understanding the trends and current situations of matches and predicting players' future performance. Additionally, through linear correlation and sensitivity analysis, we tested our model's reasonableness and applicability. The robustness of our methodologies and the detailed provided data have enabled us to optimize the momentum's effectiveness in tennis matches.

Our models not only offer a method to understanding and predicting dynamic gentleman's tennis matches, with more different sports types of data and training, but also it can be expanded to other sports matches like badminton or table-tennis matches. These variables' relations remove the coverage of pattern and insights inside the tennis match, offering guidance on decision-making and strategic planning.

As we project the future trajectory of quantifying the momentum and even more implicit factors, it becomes important to consider the disadvantages of our methods in sports matches. There exists the necessity for collecting more information and optimizing our models.

In conclusion, our work clarifies the notion and quantifying method of athletes' momentum while making some contribution to the stage of future explorations of data science in sports area to further refine people's understanding and methods of sports improvement. And the path forward involves a continued commitment on data-driven analysis and simulations.

9 Memo to the Tennis Coaches and Analysts

To: Tennis Coaches and Analysts
From: MCM Team #2423370
Subject: Summary of Momentum Research and Proposal
date: February 5, 2024

To whom it may concern,

The 2023 Wimbledon's Gentlemen's final not only offers us a visual feast with unpredictable fluctuations during the match but also stimulates people's curiosity about the abstract indicator "Momentum," which has been appreciable for a long time but there has not been a specific method to describe it. In accordance with your requests, we would like to share the findings and insights derived from our analysis of players' momentum and the flow of play in tennis matches, as requested in the research project.

Results

We utilized two different models to gain a more comprehensive and accurate method of capturing and forecasting the flow of "Momentum." We found that the momentum can be quantified by six detailed tennis-match factors: points advantage, winning games, winning break points, winners of each point, winning sets, unforced error, and serving advantage, and the swings in play and runs of success by one player are not random. Moreover, we selected five more advanced indicators among which we found the most related to the swings of a match indicator is cumulative unforced errors. Finally, we tested the generalization of the model by using it in a women's match, leading to the discovery that though the model has high accuracy but the ability of expansion is not so good.

Proposal

Based on our above work, we propose the following for coaches and analysts:

- You can enhance technical precision, focus on advancing serving and point-winning strategies, while strengthening players' mental resilience for high-pressure situations.
- You can prioritize data-driven analysis for player-customized training and match strategies, ensuring players are adaptable, physically fit and mentally prepared before the matches. That is to say, you are suggested to do research about the players' momentum and technical detailed data at the stage of preparation.
- You can get involved in a cooperation with professional data institution for the most advanced momentum-evaluation method that could be applied to your players so that you can plan the most scientific training program for them.

This comprehensive strategy aims to optimize player performance by leveraging insights from our momentum analysis. It's about boosting players' understanding and response to the game's flow, aiming for a competitive advantage in crucial tennis matches.

Best,

MCM Team #2423370

References

[1] Jordan Truman Paul Noel, Vinicius Prado da Fonseca, and Amilcar Soares. A comprehensive data pipeline for comparing the effects of momentum on sports leagues. *Data*, 9(2), 2024.

[2] The Editorial Team. What does unforced error mean? – meaning, uses and more, What Does Unforced Error Mean?, 2023. https://fluentslang.com/unforced-error-meaning/, Last accessed on 2023-09-19.

[3] Brad Gilbert and Steve Jamison. *Winning Ugly: Mental Warfare in Tennis–Lessons from a Master*. Simon Schuster, 1993.

[4] Franc Klaassen and Jan R. Magnus. *Analyzing Wimbledon: The Power of Statistics*. Oxford University Press, 2014.

[5] Tavish Srivastava. Introduction to random forest – simplified. https://www.kdnuggets.com/2020/01/random-forest-powerful-ensemble-learning-algorithm.html, 2020. Last accessed on 2020-01-22.

10 Appendix

Here are simulation programmes we used in our model as follow.

Input Python source:

```
elapsed_seconds = match_data['elapsed_time']
.apply(lambda x: sum(int(a) * 60**index for index,
a in enumerate(reversed(x.split(":")))))
plt.figure(figsize=(14, 6))
plt.plot(elapsed_seconds, match_data['p1_momentum'],
label=match_data['player1'].iloc[0] + " Momentum", color='blue')
plt.plot(elapsed_seconds, match_data['p2_momentum'],
label=match_data['player2'].iloc[0] + " Momentum", color='red')
for i in range(1,len(match_data)):
    if match_data['set_no'][i] != match_data['set_no'][i-1]:
        plt.plot(elapsed_seconds[i], match_data['p1_momentum'].iloc[i], 'bo')
        plt.plot(elapsed_seconds[i], match_data['p2_momentum'].iloc[i], 'bo')
plt.legend()
plt.title('Momentum Change Throughout the Match')
plt.xlabel('Elapsed Time (seconds)')
plt.ylabel('Momentum')
plt.grid(True)
plt.tight_layout()
plt.show()
```

```
from statsmodels.sandbox.stats.runs import runstest_1samp
# remove the NaN values and interpolate missing values
# set elapsed_seconds as the index
p1_momentum_change.index = elapsed_seconds
p1_momentum_change = p1_momentum_change.dropna().interpolate()
p1_momentum_change.to_csv('p1_momentum_change.csv')
# Applying the Runs Test to the differenced data
test_statistic_diff,
p_value_diff = runstest_1samp(p1_momentum_change, cutoff=0, correction=True)
test_statistic_diff, p_value_diff
```

Random Forest.py

```
df = pd.read_csv('Wimbledon_featured_matches_processed.csv')
# train a model to predict swing
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn import preprocessing
from sklearn.model_selection import cross_val_score
X = df[df.columns[9:]].copy()
X.drop(['swing'], axis=1, inplace=True)
y = df['swing'].copy()
X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.1, random_state=42)
clf = RandomForestClassifier(n_estimators=100, max_depth=2, random_state=0)
clf.fit(X_train, y_train)
```

```
y_pred = clf.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
accuracy
from sklearn.model_selection import GridSearchCV
param_grid = {
    'n_estimators': [100, 200, 300],
    'max_depth': [2, 5, 10, 20],
}
grid_search = GridSearchCV(clf, param_grid, cv=5, n_jobs=-1)
grid_search.fit(X_train, y_train)
print(grid_search.best_params_)
grid_search.best_score_
feature_importance = clf.feature_importances_
feature_importance
```