2024

Mathematical Contest In Modeling®

Certificate of Achievement

Be It Known That The Team Of

Kunming Wu
Sixian Ding
Yiwei Liu

With Student Advisor

Haotian Luo

Of

Sichuan University

Was Designated As

Meritorious Winner

Paul Kehle, Interim Executive Director

Administered by COMAP

Steven B. Horton, Contest Director

With support from

AMS ASA informs MAA siam

Problem Chosen	2024 MCM/ICM Summary Sheet	Team Control Number
C		2416953

A Comprehensive Framework for Tennis Momentum Analytics

Summary

In sports, a team or player might perceive a sense of momentum, indicating a prevailing "strength" or force during a match or game. Momentum is an essential psychological factor that affects the performance of players and the trend of match in tennis. However, quantifying such a phenomenon proves challenging. Meanwhile, it is not yet clear which indicators influence the generation and variation of momentum. If we can quantify momentum and predict when and how its changes, revealing the mechanisms behind, it will make great contributions to sports research.

First, we build a **Hierarchal Exponential Moving Average model (HEMA)** to quantify momentum and reflect the flow of play. We take the hierarchical nature of tennis matches into consideration and modify the formula of EMA to make it involve both point-level and game-level information. Moreover, to balance the inherent advantages of serving and disadvantages of returning, we utilize the historical data of players in the coefficient to eliminate the influence of serve. By visualization analysis, out method have satisfactory performance.

Then, in order to determine whether momentum changes randomly, we analyze the autocorrelation of momentum. We conduct **ADF test** to evaluate the stationarity of momentum and prove the momentum series is stable. Also, we apply **ACF and PACF test** and discover that the momentum series has autocorrelation.

Next, we construct the **Random Forest**-based analysis framework to predict the swing of momentum. After a broad search of indicators based on feature engineering, we explore sufficient relative indicators. We pay attention to the importance of turning points and define the contextual rule and preceding rule to identify the turning points and the potential points. The accuracy of our prediction model reaches **0.76** and **0.95** on the whole dataset. Furthermore, we propose a **Counter-factual** Analysis Advisor to provide suggestions for players.

Finally, to analyze the generalization of our framework, we apply this framework to experiments with 2011 Australian Open men's/women's match and 2022 NBA data, obtaining favorable results.

In summary, we propose a momentum analysis framework with multiple functionalities. We have proposed a framework for calculating and analyzing momentum, capable of predicting momentum swings, exploring the impact of indicators, providing match advice to players, and exhibiting good performance and versatility. It works well in various men's and women's matches and can also be applied to data mining in team sports like basketball.

Keywords: Hierarchal-EMA; Random Forest; Data Mining; Counter-factual analysis

Contents

1 Introduction

1.1 Problem Background

Momentum is an essential psychological factor that affects the performance of players and the trend of match in tennis. According to Oxford Reference, it is the positive or negative change in affect, physiology, and behavior caused by a series of events, affecting either the perceptions of the competitors or the quality of performance and the outcome of the competition [1].

Players with positive momentum are likely to hit an excellent shot, while those with negative momentum may face a crushing string of points losses. It is very important for players and their teams to recognize changes of momentum and make prompt adjustments to both mindset and strategy. However, measuring momentum is a challenging task, given its abstract nature and rapid changes. Additionally, what events and how they impact momentum remain a mystery.

1.2 Restatement of the Problem

Based on in-depth analysis and research on the background, we can specify the problems to be solved as follows:

- Build a model to track the dynamics and development of play point by point and apply it to one or more of the matches. Access player performance and the real-time advantage over their opponent. Create a visualization to illustrate the match flow.
- Use the model developed before to determine whether momentum affects the swings in play and runs of success.
- Build a prediction model to explain swings in the match based on the given data and explore the most relevant factors. Then, considering past variations in match momentum, give a player going into a new match against a different player some advice.
- Test and evaluate the predictive performance and generalization of the model on other matches.

1.3 Our Work

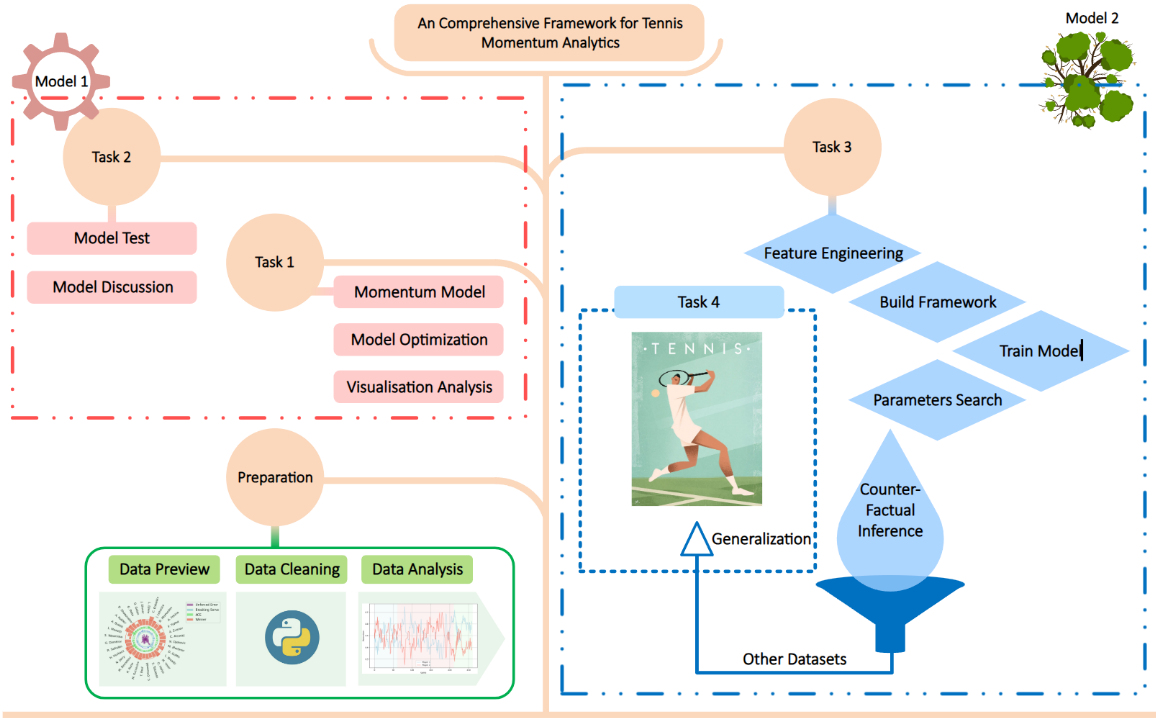


Figure 1: The overall framework of our work

2 Assumptions and Justifications

Assumption: The influence of various indicators on "momentum" is reflected in the outcomes of the matches.

Justification: Momentum represents the trend of match outcomes If an indicator influences momentum, it should affect the match results. The impact of indicators on momentum should manifest in the outcomes of the matches first.

Assumption: The earlier points have a smaller impact on the current momentum.

Justification: Recent events may significantly impact a player's performance, and as the number of matches increases, the impact of past events becomes smaller.

Assumption: Besides the observable turning points in the real data, there are also some potential turning points that do not manifest as actual turning points.

Justification: In the stalemate situations, there might be opportune moments to break the deadlock. However, due to fluctuations in player performance, it's impossible to capture every theoretically potential turning point, leaving some just potential but never actual.

Assumption: Players with similar historical statistical data will exhibit similar performances when facing the same opponent.

Justification: Given sufficient matches, players' performances tend to regress to the mean. If two players have similar historical statistical data, it implies that these two players had similar performances in the past, and we tend to believe that these two players still maintain similar performances at present.

3 Notations

The key mathematical notations used in this paper are listed in Table 1.

Table 1: Notations used in this paper

Symbol	Description	Note
EMA_t	Exponential weighted moving average at time t	/
x_t	The raw data at time t	/
α	The smoothing factor of EMA	$\alpha \in (0,1)$
k	The lag factor	/
p, p_t	A binary variable representing whether to win or lose a point (point t)	$p, p_t \in \{0,1\}$
p'	Modified form of p considering service and return	$p' \in (0,1)$
c_t	The current total points won by a specific player as of point t	/
g_t	The current total games won by a specific player as of point t	/
q_t	The relative ratio of the number of games won by a specific player to the number of games won by his or her opponent	/
r_t	Modified form of q_t	/
r'	Modified form of r considering local situation	/
S	The probability of service points won through career	$S \in (0,1)$
R	The probability of return points won through career	$R \in (0,1)$
η	The lag factor of the point-level EMA	/
ξ	The lag factor of the game-level EMA	/
λ	The smoothing factor of the point-level EMA	$\lambda \in (0,1)$
μ	The smoothing factor of the game-level EMA	$\mu \in (0,1)$
m_t	The momentum at point t	/
T	The state variable marking turning points	$T \in \{0,1,2\}$
C	The overlapping rate	/

4 Model I: Real-time Match Flow Analysis Model

4.1 Problem Analysis

To analyze the real-time match trend as points occur, the most important thing is to quantify this abstract phenomenon into a measurable indicator. Scores can provide an intuitive reflection on flow of play and the current performance of both players. Since tennis matches are hierarchies made up of sets containing games, which, in turn, contain points [2], we need to consider the impact of each point and each game simultaneously. Therefore, we choose the gain or loss of each point and the relative ratio of games won by two players as the independent variables. To balance the inherent advantages of serving and disadvantages of returning, we take the historical data of players into consideration, as the assessment of original competence.

Given that the performance of players changes over time, we need to take the results of last several points and games into consideration. It reminds us of the analysis of share prices in the stock market. Therefore, we adopt moving average, a commonly used method that identifies the trend direction in the financial field, to represent flow of play. We modify the moving average to have a hierarchical structure containing point level and game level.

4.2 Hierarchal Exponential Moving Average Model

4.2.1 Exponential Moving Average Method

Simple Moving Average (SMA) is the mean of data points for a specific time period, which is able to filter out noise in the data [4]. On the basis, the Exponentially weighted Moving Average (EMA) is derived. It is a statistic with the characteristic that it gives less and less weight to data as they get older and older [3]. Compared with SMA, EMA is more sensitive to the recent changes and can adapt to the rapid changes quicker. The recursive formula of EMA is as follows:

$$\mathrm{EMA}_t = \alpha x_t + (1-\alpha) \times \mathrm{EMA}_{t-1}, \qquad t > 0 \tag{1}$$

Where t is time, x_t is the raw data, and α is the smoothing factor. Denote the expression for EMA as $F(x_{t-k+1}, \dots, x_{t-1}, x_t)$. We expand formula (1) to get a closed-form formula and then make a slight modification as follows:

$$\mathrm{EMA}_t = F(x_{t-k+1}, \dots, x_t) = \frac{x_t + (1-\alpha)x_{t-1} + (1-\alpha)^2 x_{t-2} + \cdots + (1-\alpha)^k x_{t-k+1}}{1 + (1-\alpha) + (1-\alpha)^2 + \cdots + (1-\alpha)^k} \tag{2}$$

Where $[x_{t-k+1}, x_{t-k+2}, \cdots, x_t]$ is the sequence of the last k raw data at time t.

4.2.2 Point-level EMA

Regarding a particular player, to represent the result of point t, we define a binary variable p_t, where 1 means winning the point and 0 means losing it. As shown below, the value of p_t can be derived from the number of points won by him or her in match:

$$\begin{cases} p_t = c_t, & t = 1 \\ p_t = c_t - c_{t-1}, & t > 1 \end{cases} \tag{3}$$

Where c_t is the current total points won by the player as of point t in match.

To assess the original competence of the player, we collect the probability of service points won and return points won through the player's career, denoted as S and R separately, from https://www.atptour.com/en/players. Based on the two indicators, we further calculate p', which balances the inherent advantage gap between service and return:

$$p'_t = \begin{cases} p_t - (S - 0.5), & service \\ p_t - (R - 0.5), & return \end{cases} \tag{4}$$

The EMA of p' can be calculated according to the following formula:

$$p\_\mathrm{EMA}_t = F(p'_{t-\eta+1}, \dots, p'_t) = \frac{p'_t + (1-\lambda)p'_{t-1} + (1-\lambda)^2 p'_{t-2} + \cdots + (1-\lambda)^\eta p'_{t-\eta+1}}{1 + (1-\lambda) + (1-\lambda)^2 + \cdots + (1-\lambda)^\eta} \tag{5}$$

Where p_EMA_t is the EMA of p' at point t, $[p'_{t-\eta+1}, p'_{t-\eta+2}, \cdots, p'_t]$ is the results of the last η points with smoothing factor λ.

4.2.3 Game-level EMA

Define g_t as the current total games won by a specific player as of point t. Assume the current total games won by two players are g_{1_t} and g_{2_t}, then the relative ratio q is:

$$q_t = \frac{g_{1_t}}{g_{2_t}} \tag{6}$$

To make the data smoother and constrain the data within the vicinity of 1, we define r as follows:

$$r_t = \log_2(q_t + 1) \tag{7}$$

Unlike the point level, the impact of serving in the game level is relatively small because the server changes every game. Take the local situation into consideration, one single game is not able to result in significant impact on the overall picture, instead the combined influence of several games drives the swings. Therefore, in order to reduce the impact of individual game and make the curve smoother, we define r':

$$r'_t = \frac{r_t}{\frac{1}{\xi}\sum_{i=t-\xi}^{t} r_i} \tag{8}$$

The EMA of r' can be calculated as follows:

$$g\_EMA_t = F(r'_{t-\xi+1}, \dots, r'_t) = \frac{r'_t + (1-\mu)r'_{t-1} + (1-\mu)^2 r'_{t-2} + \cdots + (1-\mu)^{\xi} r'_{t-\xi+1}}{1 + (1-\mu) + (1-\mu)^2 + \cdots + (1-\mu)^{\xi}} \tag{9}$$

Where g_EMA_t is the EMA of r' at point t, representing the game level, $[r'_{t-\xi+1}, r'_{t-\xi+2}, \cdots, r'_t]$ is the sequence of r' of the last ξ points with smoothing factor μ.

4.2.4 Momentum Analysis combing two levels

According to the background information, a player will gain “momentum” when winning a point or game, while lose it when losing a point or game. Thus, we define “momentum” as the product of point-level and game-level EMA:

$$m_i = p_{\mathrm{EMA}_t} \times g_{EMA_t} = F(p'_{t-\eta+1}, \cdots, p'_t) \times F(r'_{t-\xi+1}, \cdots, r'_t) \tag{10}$$

Where m_i is the momentum at point t.

4.3 Results and Visualization of Match Flow

Take the 2023 Wimbledon Gentlemen’s final as an example. Carlos Alcaraz defeated Novak Djokovic by 1-6, 7-6, 6-1, 3-6, and 6-4. From the website mentioned above, we collect the S and R of the two players as shown in table 2:

Table 2: S and R of Alcaraz and Djokovic

Player	S	R
Alcaraz	0.66	0.42
Djokovic	0.68	0.42

In the beginning of the match, it is difficult to capture the flow of play which is not that obvious, so we start at point 35, taking the former points as initial states. Both the smoothing factors, λ and μ, are set to 0.1. η is set to15, and ξ is set to 5. We calculate their momentum separately and draw the line chart below:

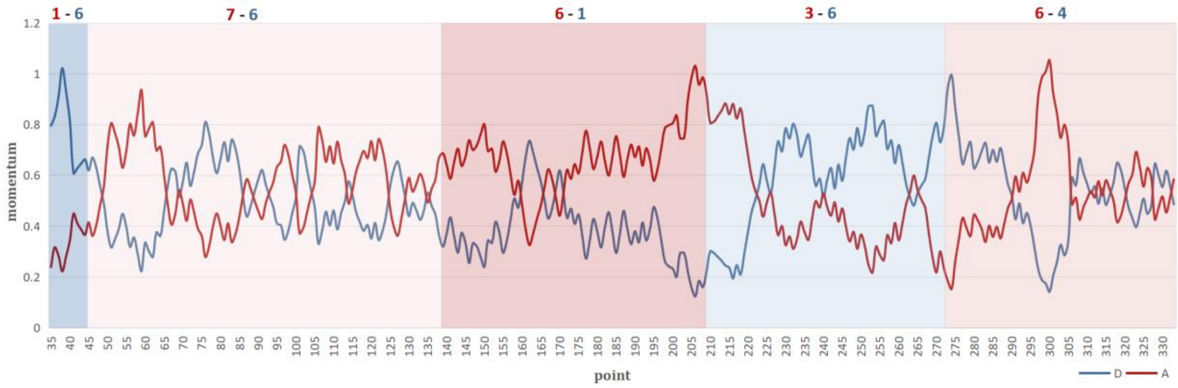
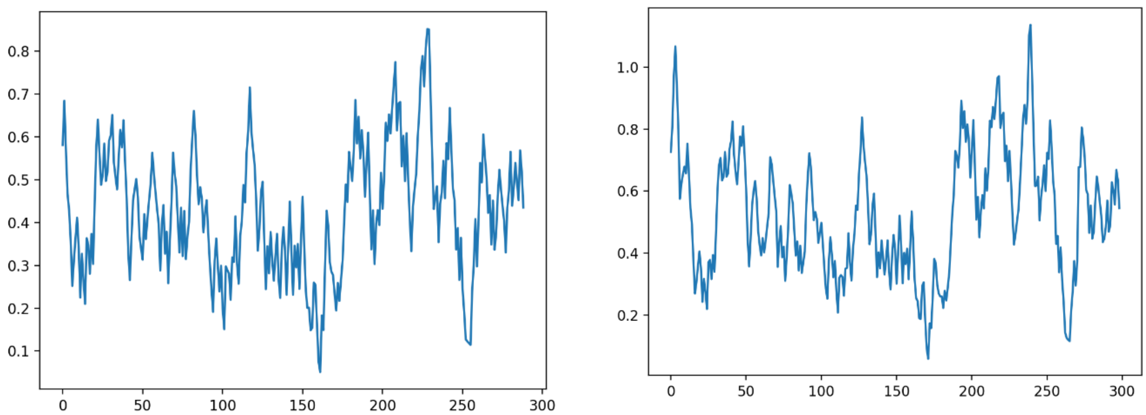


Figure 2: Line chart of the momentum of Alcaraz and Djokovic

As shown in the chart, the red line shows the momentum of Alcaraz as the blue one shows that of Djokovic. The red areas represent Alcaraz won the set and the blue areas represent the winner of set was Djokovic. The shade of color represents the set score difference, which means the greater the lead in set scores, the darker the color.

From the chart, we can clearly observe the dominance of each point and the overall fluctuations. The one who had higher momentum performed better. The larger both the momentum and the magnitude of the difference in his momentum compared to his opponent's, the better he performed. Also, it aligns with the trend of mutual growth and decline.

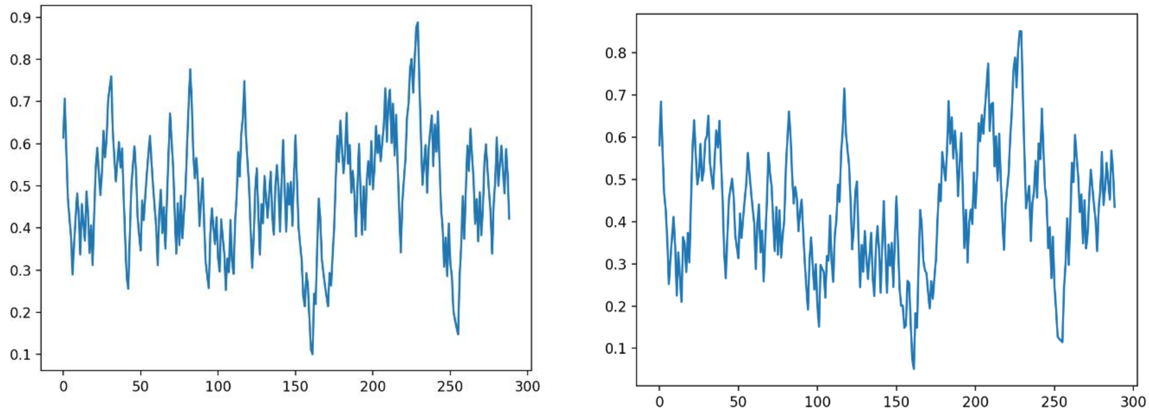
For example, in the first set, Djokovic performed much better, with his momentum consistently greater than Alcaraz's. However, the flow of play kept swinging in the second set, as both players were broken once in the beginning of the set and managed to hold their other serves. Comparing with the original data, the changes of momentum well reflect the shifts in the scores.



(a) calculated from EMA (b) calculated from Hierarchical EMA

Figure 3: The momentum of Djokovic

Moreover, as shown in Figure 3, the hierarchical structure makes the curve smoother, reducing the impact of noisy data.



(a) Considering the situations of serving and receiving together (b) Considering the situations of serving and receiving separately

Figure 4: The momentum of Djokovic

Figure 4 illustrates the differences between considering the inherent advantage gap of service and return or not. Without considering this factor, the curve exhibits numerous abrupt changes, which causes a negative effect on analyzing the short-term momentum. This demonstrates the effectiveness of our hierarchical structure.

4.4 Autocorrelation Analysis of Momentum

In order to determine whether momentum changes randomly, it is necessary for us to analyze the autocorrelation of momentum. To be more specific, we need to assess the time stationarity of momentum and the correlation between the momentum at the current point and itself at different point.

4.4.1 ADF test

Stationarity means that the statistical properties of time series do not change over time. If the series is stationary, then the future value depends on the past information, which makes the series predictable. We use the ADF (Augmented Dickey-Fuller) unit root test to evaluate the stationarity of momentum. If the P value obtained from ADF test is less than 0.05, then momentum series is stable.

Conduct the test on the momentum of Alcaraz and Djokovic in 2023 Wimbledon Gentlemen's final calculated before separately. The results are shown in the table below:

Table 3: Results of the ADF test on momentum

Player	Differential order	P value
Alcaraz	0	0.003***
Djokovic	0	0.002***

Both P values are far less than 0.05, indicating the original momentum series of the two players are stationary.

4.4.2 ACF and PACF

ACF (Autocorrelation Function) reflects the correlation between values of the same sequence at different time lags. To evaluate momentum, the formula of ACF can be written as:

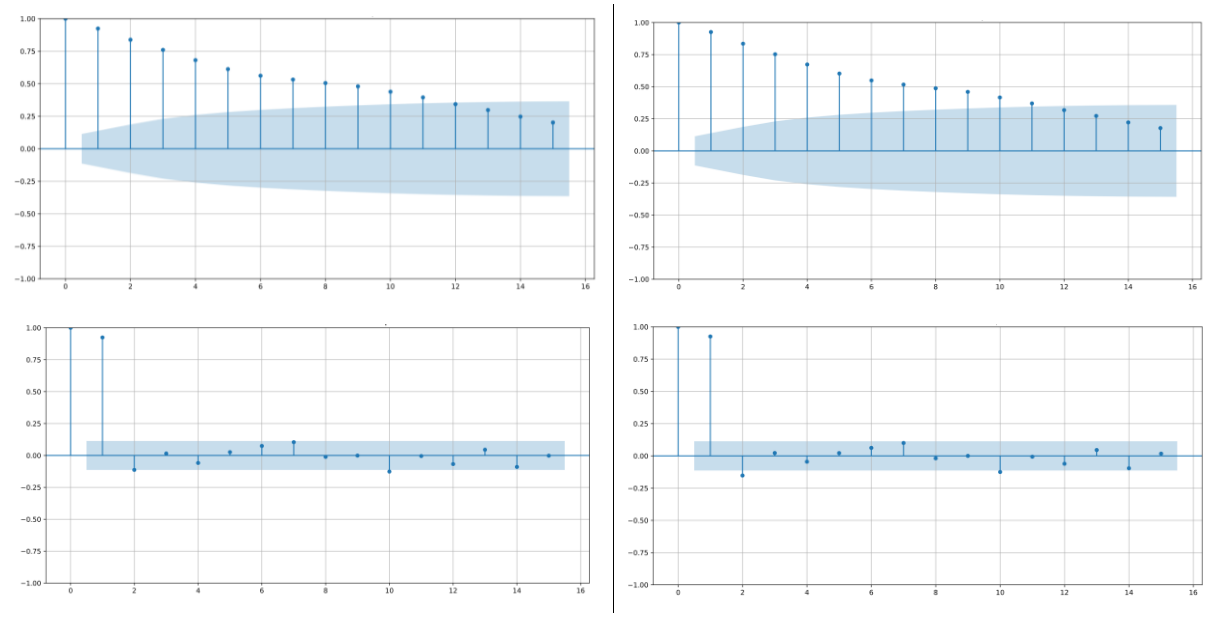
$$ACF(k) = \sum_{t=k+1}^{n} \frac{(m_t - \overline{m})(m_{t-k} - \overline{m})}{\sum_{t=1}^{n}(m_t - \overline{m})^2} \tag{11}$$

Where the k is the lag factor, n is the total number, t means point t, and $\overline{m}$ is the average momentum series.

On the basis of ACF, when evaluating the relationship between m_{t-k} and m_t, PACF (Partial Autocorrelation Function) further considers the influence of the variables in between: $m_{t-k+1}, m_{t-k+2}, \dots, m_{t-1}$. The formula of PCFA is as follows:

$$PACF(k) = \frac{Cov(m_t, m_{t-k} | m_{t-1}, m_{t-2}, \dots, m_{t-k+1})}{\sqrt{Var(m_t | m_{t-1}, m_{t-2}, \dots, m_{t-k+1})}\sqrt{Var(m_{t-k} | m_{t-1}, m_{t-2}, \dots, m_{t-k+1})}} \tag{12}$$

We conduct ACF and PACF on the momentum of Alcaraz and Djokovic and set the lag k to 15. The results are shown in Figure 5.



(a) ACF and PACF chart of the momentum of Alcaraz

(b) ACF and PACF chart of momentum of Djokovic

Figure 5: ACF and PACF chart

The blue shaded areas in the ACF and PACF charts represent the 95% confidence interval. From the ACF charts, it can be observed that the current momentum m_t is significantly correlated with the previous eleven momentums $[m_{t-1}, m_{t-2}, \dots, m_{t-11}]$. The PACF charts further reveals that, after eliminating the interference of intermediate variables, the current momentum m_t is still significantly correlated with the previous momentum m_{t-1}.

In conclusion, momentum exhibits autocorrelation. It is influenced by several previous values, especially the previous one, indicating a non-random pattern. Therefore, contrary to the coach's notion of randomness, momentum does play a role in the match and affects the performance of players.

5 Model II: Random Forest Based Analytics Framework

In order to provide a clearer illustration of our approach, here is the overall framework of our model II:

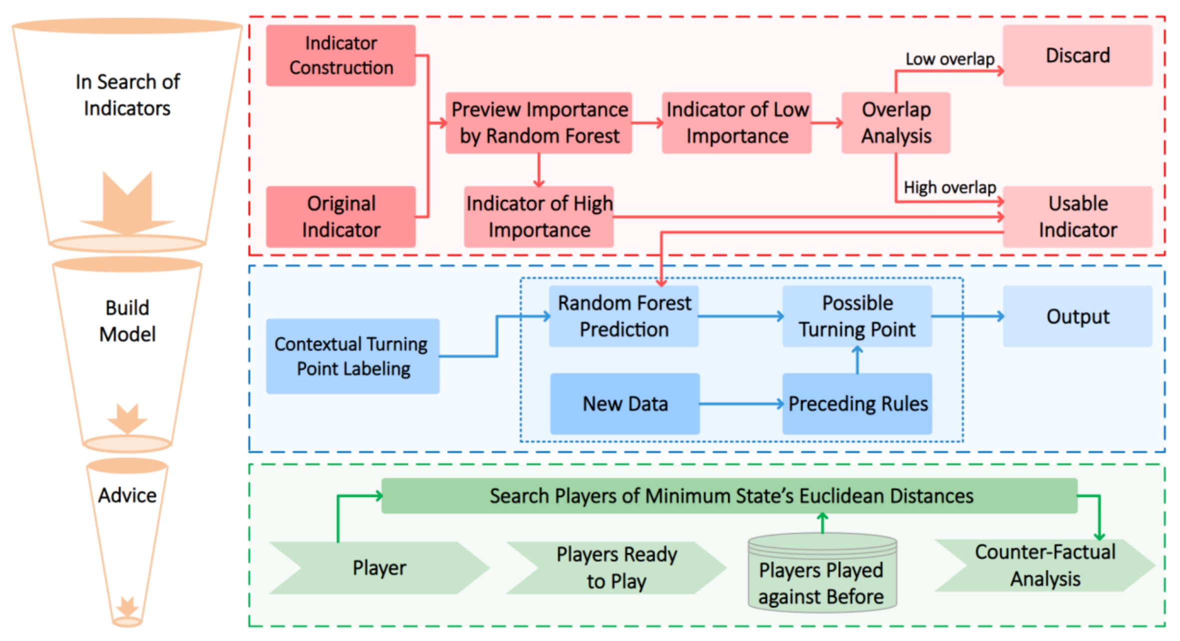


Figure 6: Overall framework of model II

5.1 Data Preprocessing and Feature Engineering

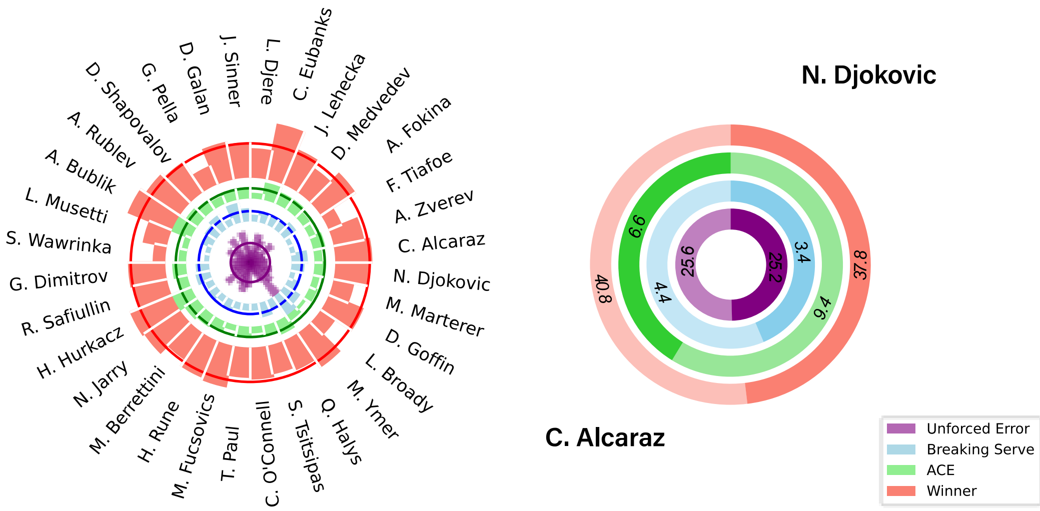


Figure 7: Dataset overview charts

Figure 7 is a statistical representation of the given Wimbledon 2023 men's matches after the first 2 rounds dataset. The left chart is composed of four circle bar charts, showing four core indicators. From inner to outer, they are Unforced Error, Breaking Serve, ACE, and Winner. The dotted lines represent their respective means in the corresponding sample space. The concentric circles on the right highlight the comparisons of the core indicators of finalists Alcaraz and Djokovic (Alcaraz on the left, Djokovic on the right).

However, there are numerous types of variables in the given dataset and the data is rough, creating difficulties for further analysis. Meanwhile, many of them are qualitative data. Thus, we perform feature engineering with the following flowchart:

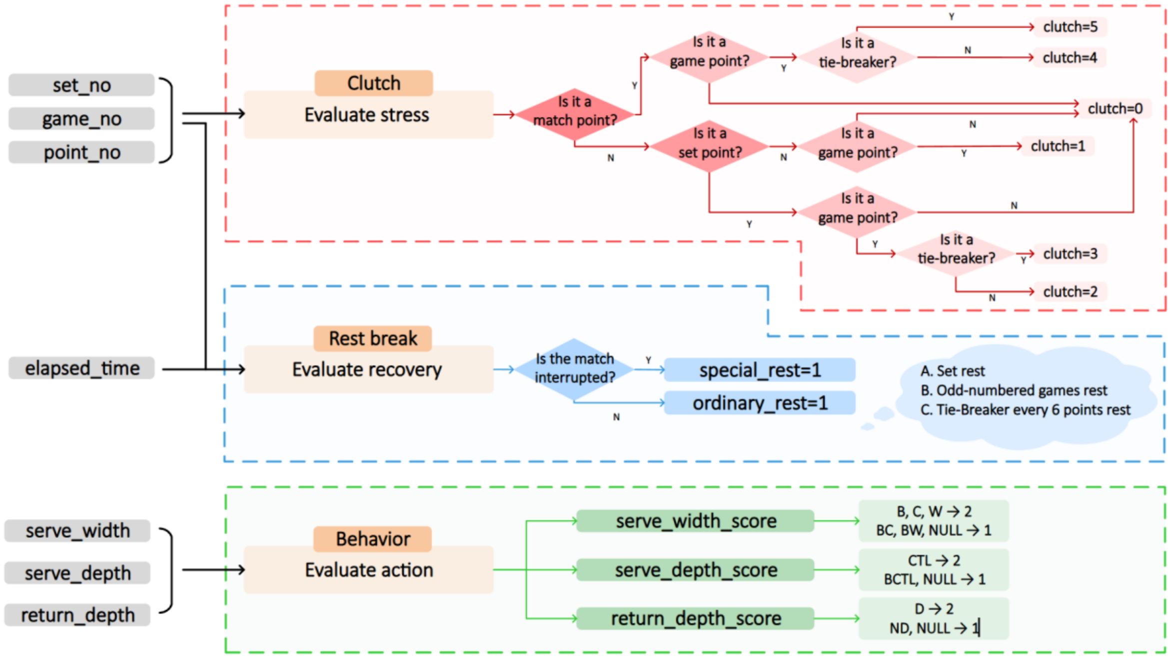


Figure 8: Feature Engineering Flowchart

Specifically, our work can be divided into two parts:

1. Feature extract

a. Clutch

In sports events, people often use “Clutch player” or “Mr. Clutch” to refer to players with big hearts [5]. Therefore, in the key game full of adrenaline, we use set_no, game_no, point_no and the player's score to abstract an indicator named clutch to measure the impact of psychological pressure on the player in the key game. Depending on the importance of the key set, the score of clutch is 5, 4, 3, 2, 1, 0, as shown in the assignment process in the red dotted box in Figure 6.

b. Rest break

Tennis demands significant physical exertion, so the short rest time resulting from side changes or special factors may affect the performance of players. Under the premise of adhering to tennis rules, we construct two indicators, special_rest and ordinary_rest, using set_no, game_no, point_no, and the length of match interruption. Both the new indicators are binary variables, which are scored as 1 if it occurs and 0 if it does not occur.

2. Feature transform

For the three qualitative indicators—serve_width, serve_depth, and return_depth—they capture the player's real-time actions. However, combining ordinal and numerical indicators for categorization can be challenging. Therefore, we perform an order-preserving mapping on them and transform them into discrete numerical values.

5.2 Turning-Point Labeling

After a broad search, we obtain the following potentially relevant indicators:

Table 4: All the potential indicators

Difference between … of two players	**Player1/2's …**	**(Others)**
break point saved	unforced error	Serve width score
total service point won	ace	Serve depth score
break points converted	winner	Rally count
return point won	distance run	Return depth score
1st serve return points won	break point won	Clutch
1st serve points won	break point missed	Whether to have rest.1
	momentum	Whether to have special rest
		Server

To explore when the flow of play changes, we should pay attention to the turning points, which reflect significant changes or reversals in the direction. In other words, our final aim is to predict if a point with several specific states is a turning point and if it is, continue to predict the turning direction. Therefore, we define a turning-point-state variable T, where 0 means non-turning point, 1 means turning upwards, and 2 means turning downwards.

Furthermore, we define two rules to identify turning points:

- **Contextual Rule (true turning points):** According to the contextual information, if the current momentum is bigger or smaller than all the 5 momentums before and all the 5 momentums after, and the difference between the current momentum and that of the next turning point reaches 0.2, then it can be regarded as a true turning point. Designate the last point of match as a turning point.
- **Preceding Rule (potential turning points):** According to the preceding information, if the current momentum is bigger than all the 5 momentums before, then this point can be regarded as a potential turning point. Otherwise, it is unlikely to be a turning point.

The reason why we take potential turning points into consideration is that when predicting, we only have the historical information and not about future, so we also need to pay attention to those which are probable to be a turning. Following preceding rule, we can filter out the points which is probably a turning point. Then, following contextual rule, we can analyze more precisely which ones are true turning points. Label the T of points shifting upwards 1, and those shifting downwards 2. At last, label the others 0, representing points which potentially, but not actually, turning points.

5.3 Indicators Evaluation

We apply the Random Forest classification algorithm and input all these indicators. The random forest can evaluate the importance of each feature by calculating Mean Decrease in Accuracy (MDA). The formula is as follows:

$$MDA = \frac{1}{N}\sum_{i=1}^{N}(Accuracy_{original} - Accuracy_{shuffled_i}) \tag{13}$$

Where N is the number of indicators, $Accuracy_{original}$ is the original model accuracy and $Accuracy_{shuffled_i}$ is the model accuracy after shuffling the i^{th} indicator. Part of the results are shown below:

Table 5: Part of the results of the feature importance evaluation

Indicator	Relative importance	Indicator	Relative importance
Momentum	34.85%	p1 break point won	0.54%
Distance run	10.91%	p2 break point won	0.49%
Rally count	5.36%	p1 break point missed	0.43%
clutch	4.12%	p2 break point missed	0.18%

Indicators like momentum and distance run have high importance. However, to our surprise, the model shows little interest in break point. Although it weighs little, according to common sense, break point should be a crucial potential turning point. We believe this is because the correlation between indicators causes information overlap, and sparse data like break point can't compete with dense data like momentum, resulting in less and less interest from random forest. Therefore, we should apply other methods to reassess the importance of these indicators: unforced error, winner, rest, ace, break point won, break point missed, special rest.

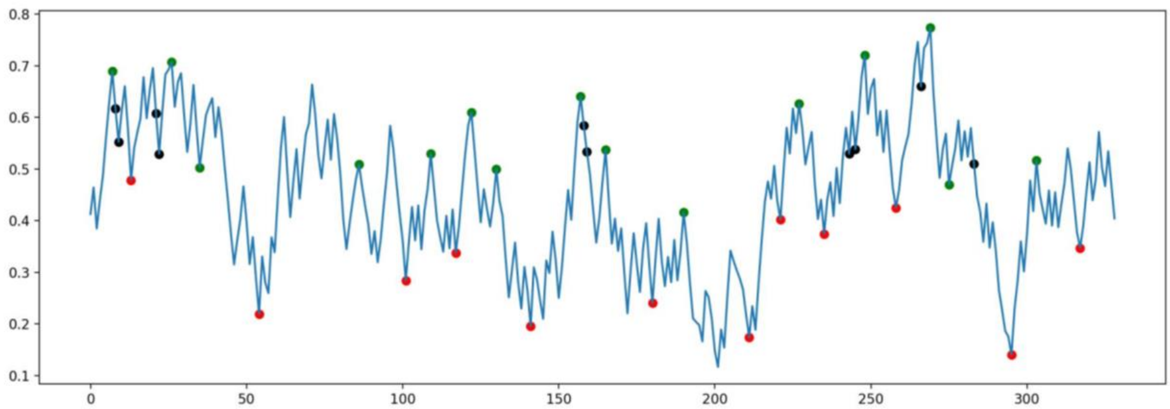


Figure 9: The marked curve of Djokovic's momentum in the final

Take the 2023 Wimbledon Gentlemen's final as an example. Figure 9 shows the momentum of Djokovic. According to the two rules mentioned before, we annotate the true turning points by green (downwards) and red (upwards) dots. Also, we annotate the break point missed by black dots. Hypothesize that point t is a turning point, if there exist black dots in {point t-2, point t-1, point t, point t+1}, then we call it a hit, also overlap. The formula of the overlapping rate is as follows:

$$C = \frac{n}{M} \tag{14}$$

Where C is the overlapping rate, n is the number of hits and M is the number of break points missed. Discard indicators with low overlapping rate and keep the others. We consider an indicator to be upward if it hits more red points and downward if it hits more green points. Part of the results are shown below:

Table 6: Part of the results of the overlapping rate

Indicator	Red dots (upwards)	Green dots (downwards)	Total overlapping rate	Type
Break point won	48%	21%	69%	Up
Break point missed	21%	27%	48%	Down
Ace	35%	25%	60%	Up

After two rounds of screening, we discard indicators like special rest, unforced error and winner, retaining the important ones.

5.4 Random Forest Turning-Point Prediction Model

Random Forest (RF) can handle large-scale data and high-dimensional data. It integrates predictions from multiple decision trees and help identify the most important indicators. What's more, RF is resistant to overfitting and provides better generalization, which is beneficial to model promotion. All records of 31 matches are shuffled as the original dataset and randomly divided to form a training set and a test set, with a ratio of 80-20. The settings are shown below:

Table 7: Random forest model settings

Number of estimators	Random state	Number of different labels		
		0	1	2
100	42	580	286	287

First, we use the original dataset to train the model to determine whether a point is a potential turning point without determining its tendency, the accuracy on test set reaches **0.92**. Then, we pick out all the true turning points as dataset and train the model to determine the turning direction of true turning point, the accuracy on test set reaches **0.83**. At last, combining these two functions, we train the RF on the original dataset, and require it to learn to screen the right potential turning points and predict the turning-point state T. The accuracy reaches **0.76**.

Table 8: Results of the predictions

Function	Accuracy
Identify turning points	0.92
Make predictions on turning point state	0.83
Identify turning points and then make predictions on turning point state	0.76

From the results and manual evaluation, the preceding rule actually can get 90% true turning point involved. So, in practical applications, for any given point, if it is identified as a turning point, the probability of correctly determining its direction is approximately 70%. If it is a non-turning point, assuming 5% points of all points are turning points, then the probability of correctly identifying it would be above 98%, which is fairly high and satisfactory. Generally speaking, our model performs well with high accuracy.

According to the importance analysis of indicators, the momentum, distance run, and 1st serve points won all play very important roles in the turning point, among which the one with biggest effect is momentum.

5.5 Counter-factual Analysis Advisor

If a player is going to a new match against a different player, he can follow these steps to help himself better prepare for the coming match:

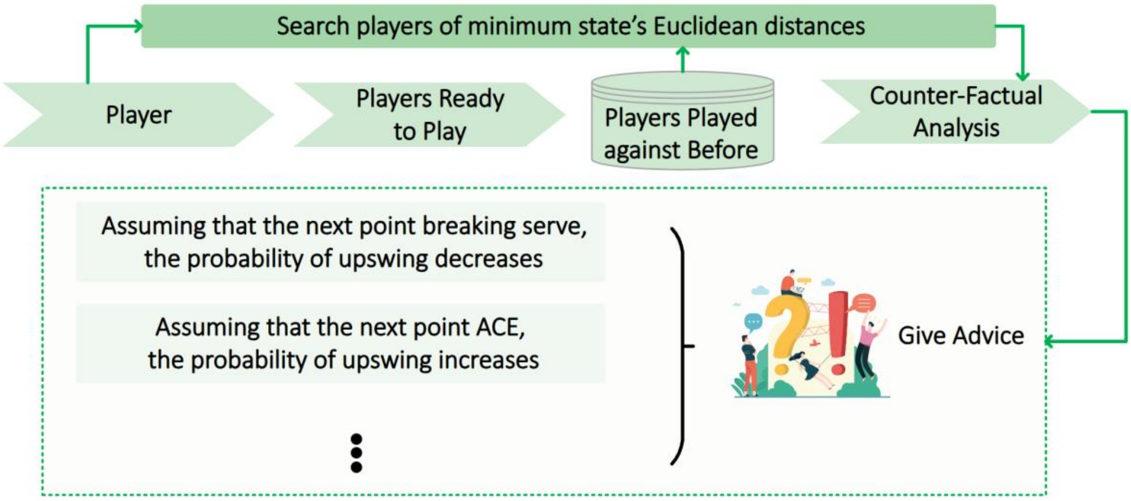


Figure 10: The flowchart of counter-factual analysis

Suppose you are player A, and your next opponent is player B.

Step 1: Find a player X who is most similar to you Based on the historical performance, and has played against your next opponent in previous rounds (or historical matches if there exists the data). Utilize the Euclidean distance (as shown in the formula below) to measure individual competence by the following indicators: break points saved, total service points won, break points converted, return points won, 1st serve points won, 1st serve return points won.

$$d(A,X) = \sqrt{(A_1 - X_1)^2 + (A_1 - X_1)^2 + \cdots (A_n - X_n)^2} \tag{15}$$

Where A_i represents the indicators of yourself, X_i represents those of other players, $d(A,X)$ is the Euclidean distance from you to another player.

Step 2: Find all the true turning points in the previous match between X and B based on contextual rule.

Step 3: Apply the counter-factual method at the turning points. Once change the value of one indicator and predict the results again. If the performance of opponent B is worse or better in this season, then run a grid search to adjust the historical data to be more consistent with the current situation. Observe the magnitude of the changes in results and determine which indicators have significant impacts. These indicators are the key to making and adjusting your strategy.

For example, assume the quarterfinals had just concluded and Jannik Sinner was going against Djokovic. For Sinner, the most similar player is Andrey Rublev, with the Euclidean distance 0.053. Run the grid search on Djokovic's state. As a result, the total serve points won and the return points won all decrease by 0.01. With counter-factual analysis, the crucial indicators are aces, break points won, distance run and rally count. In conclusion, if Sinner has the opportunity to break serve but requires running a long distance, there is no need to go for the break, especially when his momentum is relatively higher. Try to hit aces as much as possible while ensuring that the number of steps taken is not large.

6 Generalization Analysis

6.1 Overview

In order to test the generalization of our model, we collect three sets of data:

Table 9: Dataset descriptions

No.	Match name	Sports	Data Source	Generalization
1	2011 Australian Open Men’s Match	Tennis	https://github.com/JeffSackmann/tennis_slam_pointbypoint	to other men's tennis matches
2	2011 Australian Open Women's Match	Tennis	https://github.com/JeffSackmann/tennis_slam_pointbypoint	to women's tennis matches
3	2022 NBA Playoffs	Basketball	https://www.basketball-reference.com/playoffs/2022-nba-finals-celtics-vs-warriors.html	to other types of sports matches

These three sets of data evaluate the model's generalization ability gradually.

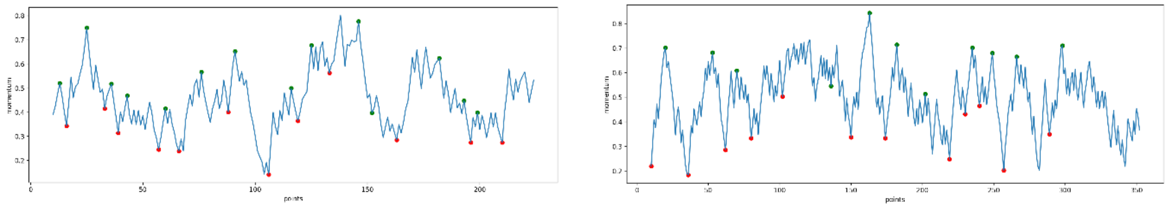
Table 10: Different predictions made on the three datasets

No.	Match name	Test Contents
1	2011 Australian Open Men’s Match	a) Evaluation based on momentum, contextual rule, and preceding rule b) Assessment of overlapping rate c) Effect of directly testing the model trained on 2023 Wimbledon Gentlemen's data d) Test results of retraining our model on new data
2	2011 Australian Open Women's Match	
3	2022 NBA Playoffs	e) Evaluation based on momentum, contextual rule, and preceding rule a) Assessment of overlapping rate

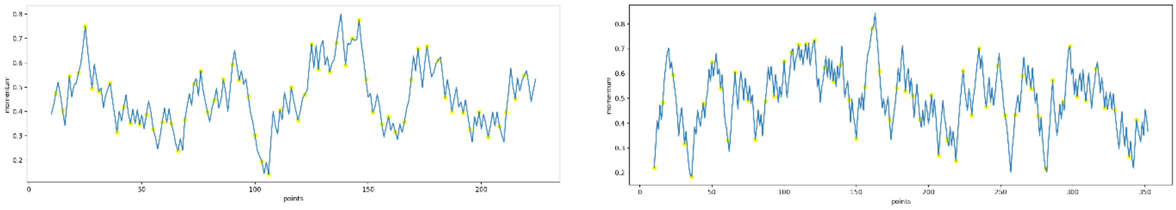
Table 10 shows the test contents to be done on the three datasets. The datasets of the 2011 Australian Open men's and women’s match contain all the indicators required for prediction, so we conduct complete predictions on them. While there is no similar data to tennis matches for the 2022 NBA Playoffs, we cannot apply random forest algorithm to make accurate predictions. Therefore, for this dataset, we only conduct first two tests.

6.2 Analysis on 2011 Australian Open Men’s or Women’s Match

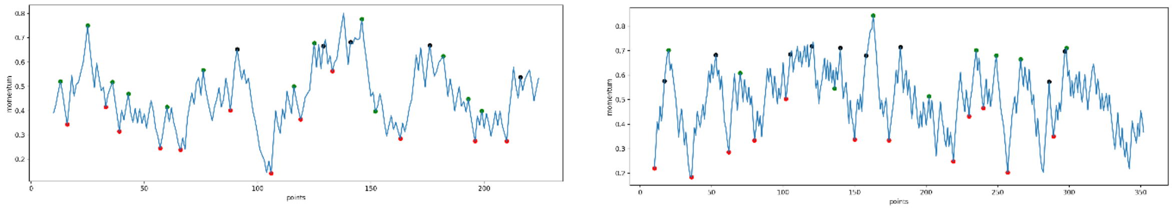
On test a and b, for dataset 1 we choose the match between Novak Djokovic and Roger Federer as example, and for dataset 2 we choose the match between Svetlana Kuznetsova and Francesca Schiavone as example. On test c and d, for both dataset 1 and 2, we use all the records of dataset to make predictions. The results are shown below:



(a) Visualization of test a, following contextual rule to annotate the turning points



(b) Visualization of test a, following preceding rule to annotate the turning points



(c) Visualization of test a, with the indicators annotated

Figure 11: Results of all the tests (men left, women right)

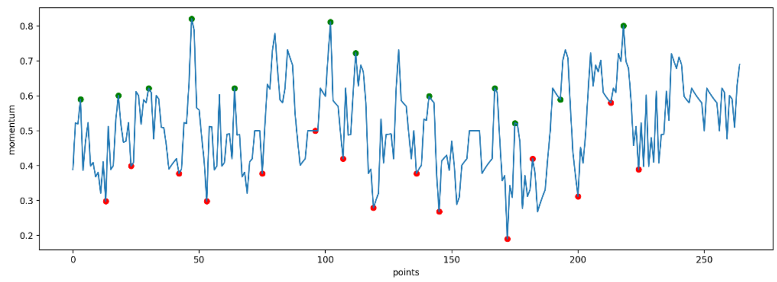
By indicators exploration, on both datasets, we get the same conclusion as the 2023 Wimbledon Gentlemen's match, and keep the same indicators.

For dataset 1: We then conduct test c and reaches the accuracy of 0.708. Conduct test d and the model reaches the accuracy of 0.733 on the test set. According to our analysis, this gap probably raises from the changes of players ability during the decade, which causes the decrease of representation ability of the difference between players states and leads to the reduction in model predictive power.

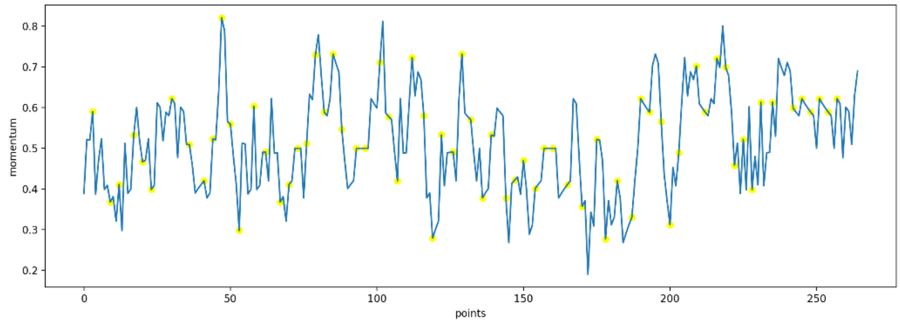
For dataset 2: Conduct test c and the accuracy reaches 0.663. We then conduct test d and reaches the accuracy of 0.712. According to our analysis, in addition to the changes in players ability discussed before, this gap is also affected by gender. It can be seen from the example momentum that women's tennis competitions are more volatile and more difficult to predict.

6.3 Analysis on 2022 NBA Playoffs

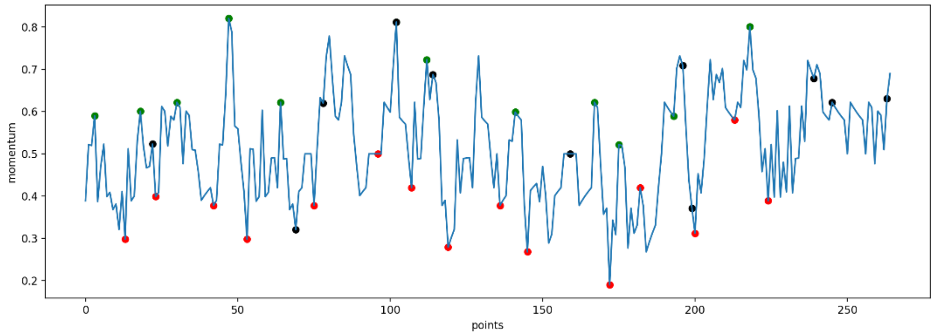
On test a and b, for dataset 3, we choose the Playoffs Finals Game 1, Boston Celtics vs Golden State Warriors, as example.



(a) Visualization of test a, following contextual rule to annotate the turning points



(b) Visualization of test a, following preceding rule to annotate the turning points



(c) Visualization of test a, with the indicators annotated

Figure 12: Results of tests

Take the indicator “player” as an example, we discover that player substitutions and the turning points highly overlap, which is consistent with the fact that in actual games, team coaches generally choose to use substitutions to break the deadlock when the offense and defense are in a stalemate.

7 Sensitivity Analysis

The parameters of our model derive from the parameters of momentum model, contextual rule and preceding rule.

For the parameters of the momentum model, the adjustable options include the lag factor and smoothing factor of point level and game level.

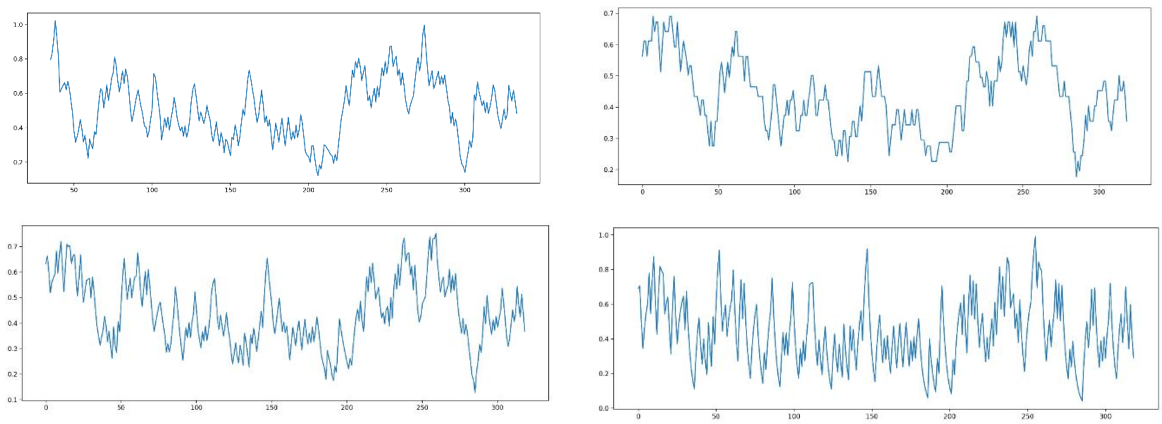


Figure 13: the same momentum with different game-level and point-level lag factors and smoothing factors

From visualization analysis, we can find out that the smaller lag factor and the bigger smoothing factor is, the worse the ability to smooth small movements, but the more sensitive it is to respond to changes.

For contextual rule and preceding rule, the adjustable options include the length of contextual information and the threshold determining whether the change at the point is sharp enough to make it a turning point.

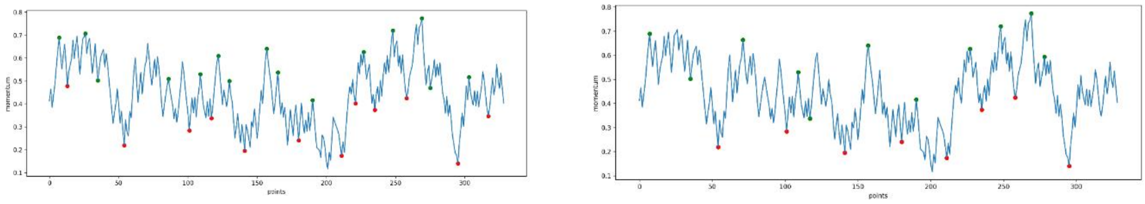


Figure 14: the same momentum with different length of contextual information and threshold

From visualization analysis, the larger the length, the fewer points will be found but has higher accuracy. A smaller length and a smaller threshold for turning points will result in more points, but lower accuracy.

8 Model Evaluation and Further Discussion

8.1 Strengths

For Model I, the HEMA model not only captures trend information at the point level but also incorporates trend information at the game level, minimizing small fluctuations and amplifying significant ones. Additionally, we differentiate between the server and returner, adjusting point-level values based on the historical win rate data, ensuring the model less affected by server-returner transitions.

For Model II, following preceding and contextual rules, we pre-screen possible turning points using a learnable random forest model, effectively addressing issues of data imbalance and long-tailed distributions in real-world scenarios. The counter-factual approach provides flexible and effective suggestions and can be employed for dynamic recommendations.

8.2 Weaknesses

For Model I, although the model attempts to represent momentum information based on historical data, the utilization of information may be suboptimal. The model employs exponential weighting to assess the importance of each point, but in reality, the importance of each point may vary.

For Model II, the model's training relies on contextual and preceding rules. If these rules perform poorly on specific datasets, the model's effectiveness can significantly decrease. The model requires numerous parameters, making it challenging to obtain comprehensive data for accurate predictions, impacting its generalization to other sports such as basketball. Due to the limitations of the random forest model, recommendations derived from counterfactual reasoning might not always be accurate.

8.3 Further Discussion

For Model I, it could be beneficial to explore alternative methods, such as measuring the importance of each point using corresponding indicators instead of exponential weighting. This approach would provide more flexibility to momentum and allow for optimization in utilizing various types of information.

For Model II, it needs to further integrate input information into the model, enhancing its generalizability. Similar to momentum, other crucial information could be represented in a similar way, utilizing preceding information. Additionally, considering classifiers other than random forests and refining the counterfactual reasoning model could contribute to improvement.

9 Conclusion

Our article begins with an exploration of momentum in tennis matches. At the first stage, we build a Hierarchal Exponential Moving Average model considering both point level and game level to quantify "momentum", and then conduct autocorrelation analysis to demonstrate its impact in match. Then, on the basis of the momentum model, we explore the potential indicators of turning points identified by preceding and contextual rules and use effective indicators to train a random forest to predict turning points. Next, counter-factual analysis is conducted on the trained prediction model to provide reasonable suggestions for players. Finally, we test our model on three other datasets. The results show that our model still have good performance on other datasets, indicating that our method has generalization ability.

References

[1] Oxford Reference. Psychological momentum. https://www.oxfordreference.com/display/10.1093/oi/authority.20110803100352743

[2] Simmonds, Emma, and Peter O'Donoghue. "Probabilistic models comparing Fast4 and traditional tennis." International Journal of Computer Science in Sport 17.2 (2018): 141-162.

[3] Hunter, J. Stuart. "The exponentially weighted moving average." Journal of quality technology 18.4 (1986): 203-210.

[4] Stock Charts. Moving Averages -- Simple and Exponential. https://school.stockcharts.com/doku.php

[5] Wikipedia. Clutch (sports). https://en.wikipedia.org/wiki/Clutch_(sports)

10 Memorandum

TO: MCM

FROM: Team # 2416953

DATE: February 6, 2024

SUBJECT: Momentum Swings Analytics in Tennis

Tennis is a popular sport, in which the trend of the game can be quantitatively characterized as "momentum". In order to evaluate the performance of players better, we designed a framework based on statistical perspective and machine learning approach, which allows players or their coaches to make timely and favorable behavioral changes during the match, thus potentially changing the flow of play. Specifically, they can use the framework to achieve the following functions:

1. Identifying and responding to turning points in match flow

A player's coach can quantitatively characterize "momentum" based on our proposed hierarchical exponentially weighted moving average model and utilize this indicator to help the play to have more precise control over the pace of match. Also, they can specify whether it is a crucial break point and whether it is a key game based on the real-time situation of the game. We believe that "momentum" already contains enough information about the flow of play, which attributes most to the occurrence and direction of turning points in the match. Therefore, to make predictions on the current situation, it is sufficient to feed the real-time point-by-point data into the Random Forest classifier pre-trained by our methods. Following the preceding and contextual rules, the prediction model outputs whether the next point is a turning point, and if it is, what is the turning tendency. The generalization of this classifier is quite strong that it can even be applied to Grand Slam from a decade ago.

Meanwhile, the coach can use the counter-factual method to count over a large number of past match records to statistically identify the most favorable strategy for the player at present and then provide suggestions. For instance, if the classifier predicts that the probability of upward turning point on the next point is 0.2 and the probability of downward turning point is 0.8, the coach can use the counter-factual method to count past matches and find that at this very point, if the player just hits an ace on the next point, the probability of facing downward turning point can be reduced to 0.6.

It is also worth noting that the indicators of the enumeration must be controllable. For example, the order of serves is uncontrollable, but the mindset can be adjusted. The psychological pressure of both sides will multiply in the key game. If the player can adjust his or her mindset in time and make calm response to the "clutch" points, it is possible to change the direction of momentum. It also means that the coach needs to have a counter-factual searcher ready in advance to be able to identify those potential turning points when they come.

2. Predicting games against new opponents and getting prepared in advance

An important task for coaches before a match is to investigate and advise on how to deal with their opponents, but those players who have never played against each other before make this task difficult. Meanwhile, the sample space searched by counter-factual methods is not specific to a particular opponent's match record but a generalized space, which is moderate for generalization, but insufficient for analyzing specific opponent performance. To address this problem, we propose that the coach search all the players and calculate the closest Euclidean distance of the indicator "state" for matching, i.e., this player or players can be used as a "proxy" to help us replace the sample space searched by the counter-factual method with the corresponding opponent's record, which makes the model intentionally "overfitted" and increases the prediction ability for specific opponents.

3. Some interesting conclusions

People have always been curious about what indicators change the flow of play. Based on the variable importance analysis conducted by Random Forest, we discover that the indicators contributing to the prediction of volatility are our manually constructed quantitative indicators "momentum", which is about 40%. The second most important factor is who serves, which is intuitively consistent with the fact that the person who serves has a head start. Tied for third place are distance run per point, rally count per point, and the stress indicator "clutch" in key games. It means that physical exertion in the previous point and mindset in key games have a significant effect on the next point. When these three indicators appear high, the probability of a turning point becomes higher. The accuracy of our model is 76%, and the above five indicators contribute about 70% of the accuracy.

Through data mining, we also come up with a conclusion about break points that is different from the traditional view: after reviewing numerous completed matches, pick any two break points in a match. If the time interval between the two points is long (except for matches interrupted by special factors), then the first break point is likely to lead to a drop in “momentum”. Conversely, if the interval is relatively close, then it is possible that it will not lead to a drop. This also means that the event of a break point in tennis, is not the node that establishes one's advantage. Instead, it could be a potential sign that one's advantage is about to end. If at this point that player cannot continue to maintain the advantage and hit a break point in a short period of time, then the balance of victory will be tilted in favor of the other side.

Yours sincerely,
Team # 2416953