Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Momentum prediction model of the tennis game based on the random forest algorithm

Zhisang Zhou [1, #, *], Haiwang Gong [2, #], Mingyan Tang [1, #], Wenrui Xiao [1, #], Mengqi Lou [1, #]

[1] Department of medical imaging, HangZhou Medical College, Hangzhou, China, 311300

[2] Department of medical information engineering, HangZhou Medical College, Hangzhou, China, 311300

* Corresponding author: 13819622734@139.com

[#]These authors contributed equally.

Abstract. Momentum in tennis matches can greatly affect the performance of tennis players, and dynamic monitoring of momentum can help players to achieve better results in the competition. This paper constructs a tennis ball momentum prediction model based on the random forest algorithm. The model selects the nine indicators of the process of winning the competition in the current set, whether to score continuously, whether to serve, gain or lose the point by untouchable shot, relative fatigue, error situation, double fault, rally count, speed of serve, uses the random forest algorithm to predict the momentum. To prevent the model overfitting problem, this paper reduces the index in dimension by PCA. Based on the results of model training, this paper puts forward suggestions on how to improve the score: prepare in advance, good attitude, hold the rhythm, reduce error, play to the score, improve technique. The data of 201910260 ATL of quarter 1 from the game data of the 2019-2020 regular season are trained by a momentum prediction model based on the random forest algorithm, which demonstrate the generalization ability of the model. This model can be used in various sports events to help athletes improve their performance.

Keywords: Tennis Match; Momentum Prediction; Random Forest; Momentum Evaluation Index.

1. Introduction

Tennis is a fashionable and elegant sports activity, known as “the world’s second largest ball game”. It has the characteristics of easy learning and sociability [1]. Tennis originated in France, was born in England. Now, in Europe and United States or even other regions, tennis is a popular sport welcomed by people. In the 2023 Wimbledon men’s singles final, 20-year-old Spanish rising star Carlos Alcaraz defeated 36-year-old Novak Djokovic. It was Djokovic’s first Wimbledon loss since 2013 and ended his remarkable run as one of the all-time greats at Grand Slam tournaments. It follows that players control the pace of the entire game, and sometimes a lot of points, and even plays, happen to players who seem to have an advantage, which is often attributed to “momentum” and thus determines the outcome of the game. Since there are many factors that affect the victory of the player in the actual competition, to evaluate the performance of the player.

Tennis has gradually developed into a more complex sport. Some researchers have studied the tennis game itself, the tactics of the tennis game, and the techniques of the tennis game [2]. Some scholars present an in-depth analysis of the serve in men’s singles tennis matches at Wimbledon, with the quantitative tool of the stroke performance relevance (SPR) model providing scientific theoretical support for the tactical application of tennis players and coaches [3].

In addition, there are some studies from the perspective of tennis players, studying the psychological quality, physical quality, the control of the rhythm of the game, the direction of training, the distribution of players' physical fitness [4].

A more scientific and systematic mathematical method can be used to study and analyze the performance and scores of competitors, and develop the running scores and "momentum" of dynamic and flexible players [5]. Therefore, through quantifiable indicators to establish a proper mathematical model to reasonably analyze or predict the performance of players, to control the situation of the game.

Currently, scholars rarely study the momentum through mathematical and statistical methods. Based on this, this paper uses whether to serve, relative fatigued and other nine indicators to predict the momentum, and puts forward some suggestions to improve the performance of players and help players achieve better results in tennis.

2. Establishment of a momentum prediction model based on the random forest algorithm

Momentum in tennis is not random, often correlated with previous fluctuations in momentum, and with the mentality, skill, and fatigue of the game. To construct the correlation analysis model of momentum prediction and indicators, this paper selected nine factors include with the process of winning the competition in the current set, whether to score continuously, whether to serve, gain or lose the point by untouchable shot, relative fatigue, error situation, double fault, rally count, speed of serve, construct Momentum prediction model of the tennis game based on the random forest algorithm.

2.1. Establishment of the momentum model

This paper suggests that the momentum fluctuations are not random but influenced by previous momentum fluctuations. The fluctuation of the scoring rate as Formula (1) can reflect the fluctuation of the momentum. Considering the momentum changes during games, such as consecutive scoring or losing points, the changes in momentum can fluctuate, and the fluctuations vary from person to person. To describe the momentum accurately, the momentum model is constructed as formula (2).

$$P_s = \frac{p_i}{\sum_{i=1}^{n} p_i} \tag{1}$$

P_s represents scoring rate,P_i represents the score obtained for that period.

$$M = \sum_{i=1}^{n} (-1)^{i-1} \times \Delta P_i \tag{2}$$

2.2. Feature vector selection

(1) The process of winning the competition in the current set

In general, when a player is closer to winning, the game mentality is better and the momentum should grow. This paper defines the game progression as formula (3).

$$V_1 = \begin{cases} \dfrac{\text{player's points in this game}}{4}, & \text{opponent's point} \le 2, \ \text{player's point} \le 4 \\ \dfrac{\text{player's points in this game}}{2 + \text{opponent's point}}, & \text{opponent's point} \ge 4, \ \text{player's point} \le 2 \end{cases} \tag{3}$$

(2) Whether to score continuously

This paper can think that winning continuously improves the players mentality and increases the momentum. This paper defines whether to score continuously as formula (4).

$$V_2 = \begin{cases} 1 & \text{player scores} \\ 0 & \text{player doesn't scores} \end{cases} \tag{4}$$

(3) Whether to serve

It can think that untouchable shot wins in the game improves players' mentality and facilitates momentum growth, otherwise, momentum decreases. This paper defines Whether to serve as formula (5).

$$V_3 = \begin{cases} 1 & \text{player serves} \\ 0 & \text{player doesn't serve} \end{cases} \tag{5}$$

(4) Gain or lose the point by untouchable shot

It can think that untouchable shot wins in the game improves player mentality and facilitates momentum growth, otherwise, momentum decreases. This paper defines Whether to serve as formula (6).

$$V_4 = \begin{cases} 1 & \text{gain the point by untouchable shot} \\ 0 & \text{there's no untouchable shot} \\ -1 & \text{lose the point by untouchable shot} \end{cases} \tag{6}$$

(5) Relative fatigue

In tennis, the player's physical state plays a role on momentum fluctuations. It's generally believed that the more fatigue of the player, the worse the performance, the lower the momentum. At the same time, the fatigue of the opponents in the game is also increasing, and the performance of the opponents also affects the performance of the players. Therefore, to consider the impact of fatigue on momentum, this paper defines relative fatigue as formula (7).

$$V_5 = \frac{\sum_{i=1}^{n} d_i + \sum_{i=1}^{n} d_i'}{\sum_{i=1}^{n} d_i} \tag{7}$$

d_i represents player's distance ran during point.

d_i' represents opponent's distance ran during point.

(6) Error situation

In tennis games, once there is a mistake, the player's mentality is bound to affect it. At the same time, the opponent's mistakes are also conducive to increasing the players' mentality. That is to say, the performance of the opponent also affects the performance of the players. Therefore, to consider the impact of the mistakes of the players on the momentum of the players, this paper defines Error situation as formula (8).

$$V_6 = \begin{cases} 1 & \text{opponnent's error in this point} \\ 0 & \text{no error} \\ -1 & \text{error in this point} \end{cases} \tag{8}$$

(7) Double fault

In tennis, once double fault, the mentality of the players will inevitably be affected. Therefore, to consider the impact of the double fault on the player momentum, this paper defines Double fault as formula (9).

$$V_7 = \begin{cases} 1 & \text{opponnent's double in this point} \\ 0 & \text{no double fault} \\ -1 & \text{double fault in this point} \end{cases} \tag{9}$$

(8) Rally count

Number of shots during the point also effect momentum swing. This paper defines Rally_count as formula (10).

$$V_8 = \text{rally\_count} \tag{10}$$

(9) Speed of serve

It is believed that the speed of the serve reflects the mentality and technique of the server and affects the fluctuation of momentum.

Through the analysis of the data provided by Wimbledon, this paper can find that a considerable part of the data in the game is not continuous array, but classified data, such as whether to score, whether to serve, etc. That is, these data are difficult to build a linear relationship with momentum. To consider the influence of these multi-classification features on momentum fluctuations and with the help of these features, this paper constructed a momentum prediction model based on random forest regression.

2.3. Training model

Random Forest is a supervised machine learning algorithm that creates a “Forest” by combining n decision trees. It generates multiple decision trees and merges them to achieve a preciser prediction [6]. In the random forest, each decision tree is independent and trained on a randomly selected sub_sample, which can effectively reduce the risk of over fitting. Random forest is better at handling categorical variables, compared with random networks and linear discrimination.

The basic principle of random forest is introduced as follows:

If the proportion of class k samples in the current sample set D is p, the information entropy of d is defined as formula (11) [7].

$$Ent(D) = -\sum_{k=1}^{|Y|} p_k \log_2 p_k \tag{11}$$

The random forest decision tree will use the Gini index to partition the properties, and the purity of the data set D can be measured using the Gini values as formula (12)[8].

$$Gini(D) = 1 - \sum_{k=1}^{|y|} {p_k}^2 \tag{12}$$

Assuming that feature a has v possible values, then feature a will divide v nodes on the sample set D.

Gini index of feature a to sample set D is formula (13) [8].

$$Gini\_index(D, a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} Gini(D^v) \tag{13}$$

In decision tree learning, to correctly classify the training samples, the node division process will be repeated, which will lead to the overfitting phenomenon in the training process of random forest decision tree, resulting in the illusion of high result accuracy. To reduce overfitting risks, random forests actively reduce branches, a means called pruning.

Subtree loss function is defined as formula (14).

$$C_\alpha(T) = C(T) + \alpha|T| \tag{14}$$

T represents single subtree.

C(T) represents Error cost of T

|T| represents the number of nodes.

3. Results

3.1. Training model

This paper chose the Carlos Alcaraz competition data, applied to the training and prediction of the random forest model, taking $V_1, V_2, V_3, V_4, V_5, V_6, V_7, V_8, V_9$, as variables. The data is applied to the model and solved in the r language.

To improve the fit of the model, this paper constructed the MES on Cross validation Fold and the MES on iteration, respectively. Model accuracy is higher when the MES approaches to 0.

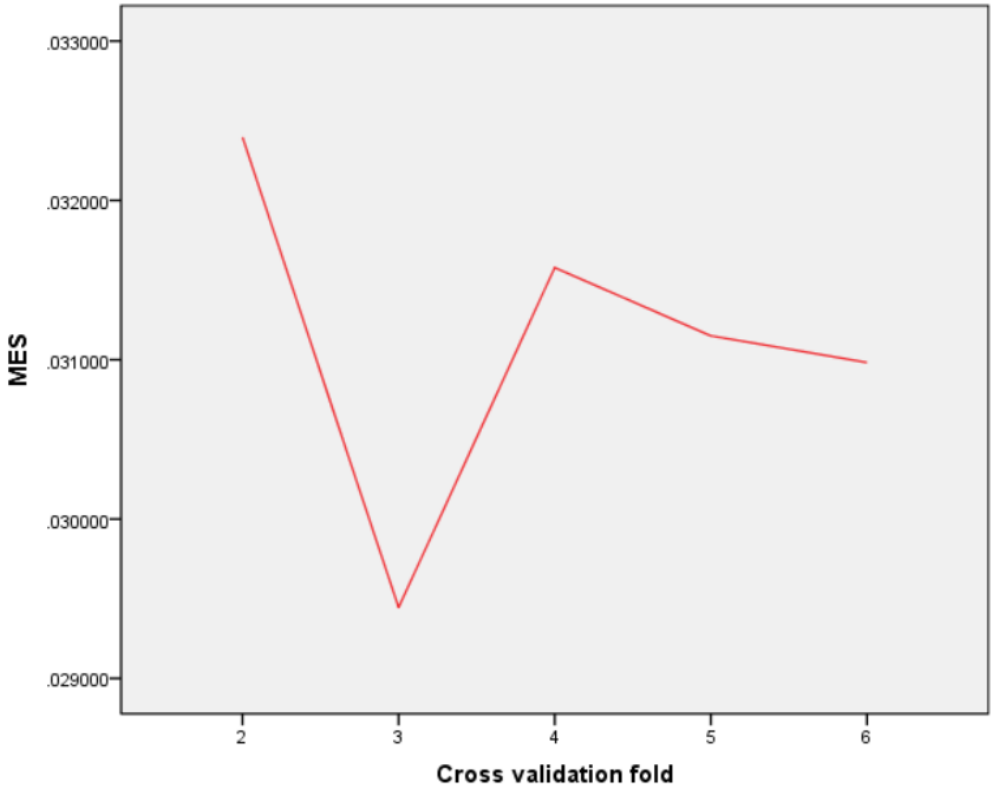

Figure 1. Cross Validation Fold

Analysis of the data in the Figure 1 reveals that the smallest MES and the best model fit when Cross validation Fold equals 3.

Model accuracy is higher when the MES approaches to 0. Analysis of the data in the Figure 2 reveals that the smallest MES and the best model fit when Iteration equals 1.

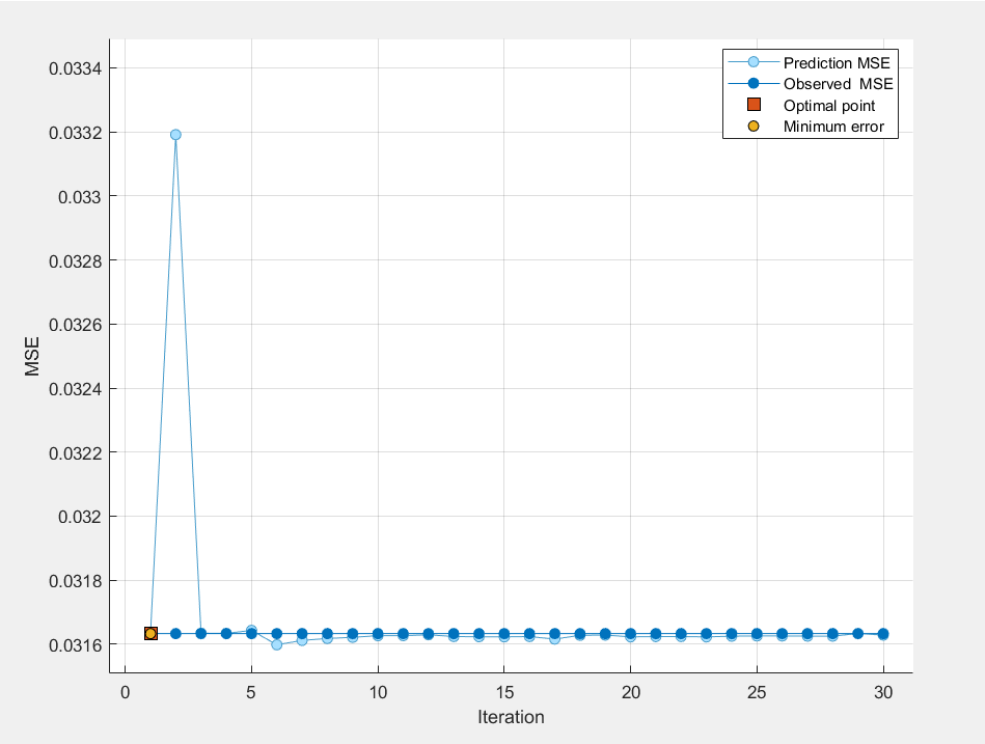

Figure 2. Minimum MSE

Based on the above results, this paper performed the following parameter settings for the random forest Model as Table 1.

Table 1. Random forest parameter setting

Number of cross-validated folds	iteration	min samples leaf
3-fold cross-validation	30	143

After adjusting the parameters, the output model results are shown in the Figure 3.

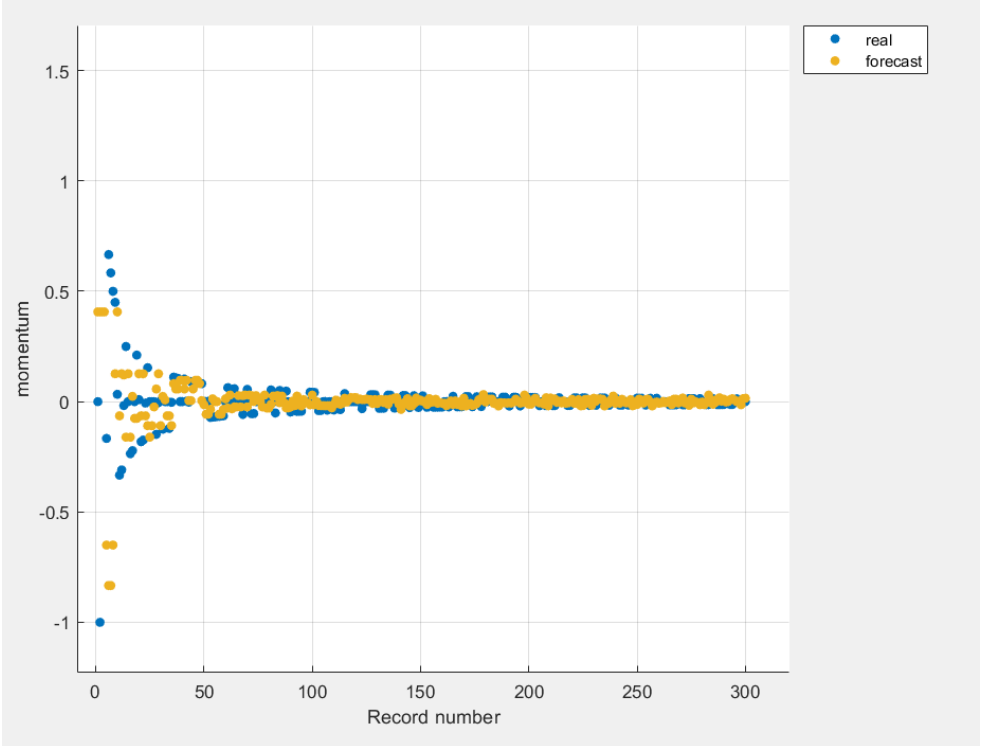

Figure 3. Momentum Prediction Result

The analysis of the prediction results in the figure shows that most of the predicted values and the true value are small, indicating that the prediction effect of the model is ideal. As for the weight of the features, as shown in the Figure 4.

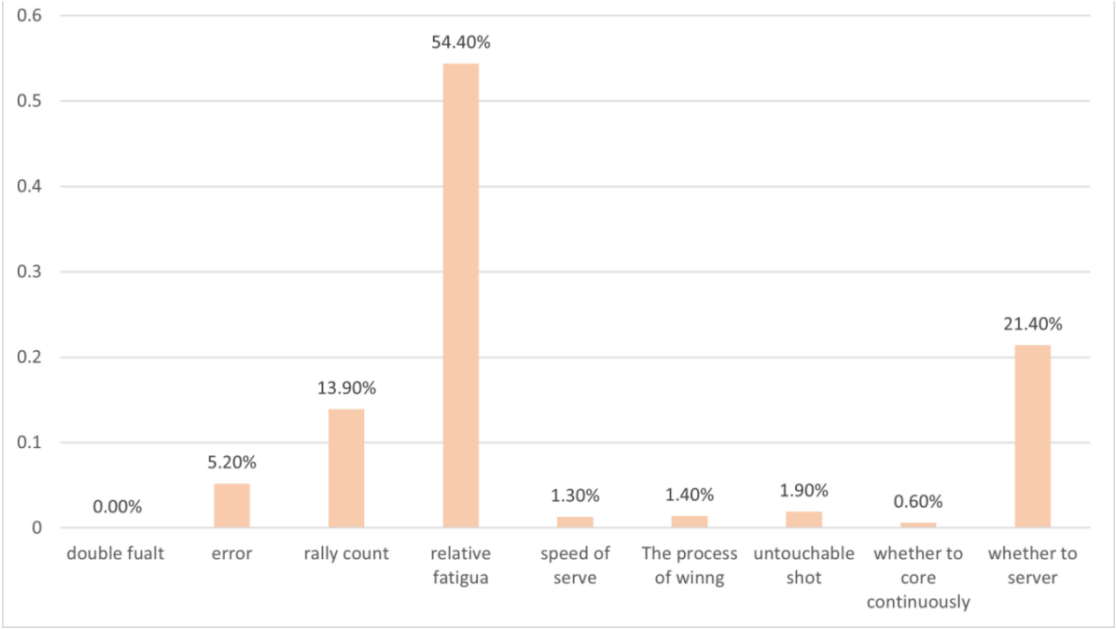

Figure 4. Weight Of Nine Indicators In Momentum Prediction Model

Analyzing the results, it's founded that the four factors of relative fatigue, whether to serve, serve speed and error had a significant influence on momentum fluctuations, while other factors had less influence. At this time, this paper needs to consider whether too many factors with less correlation will cause the overfitting phenomenon of the model. Therefore, the dimension of the model ought to be reduced.

3.2. Testing model

To test the accuracy of the model, this paper calculated regression index.

(1) MSE: Mean Squared Error [8].

$$\mathrm{MSE}=\frac{1}{m}\sum_{i=1}^{m}(y_i-y_i')^2 \tag{15}$$

y_i' represents prediction data, y_i represents real data.

(2) RMSE: Root Mean Squard

$$\mathrm{RMSE}=\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i-\hat{y}_i\right)^2} \tag{16}$$

It can be considered that the regression effect deviates from the true value on average

(3) MAE: Mean Absolute Error

$$\mathrm{MAE}=\frac{1}{m}\sum_{i=1}^{m}|(y_i-\hat{y}_i)| \tag{17}$$

When the predicted value perfectly matches the true value, it is the perfect model; the smaller the error, the smaller the value.

(4) R-Square

$$\mathbf{R2} = -1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\overline{y_i} - y_i)^2} \tag{18}$$

The calculation results of regression index of momentum prediction model based on random forest algorithm are shown as Table 2:

Table 2. Calculation of the model evaluation index

RMSE	R-Square	MSE	MAE
0	0.98	0.061633	0.07075

The results, when RMSE = 0, MSE = 0.61633 shows that the model fits ideal, almost no error; R^2=0.98, close to 1, ideal fitting result, small error, MAE = 0.07075, ideal fitting result and small error.

3.3. Dimensin Reduction-PCA

Considering that the random forest regression model selects too many indicators, it tends to cause the overfitting phenomenon. To avoid this, this paper reduced the feature dimension using PCA. PCA as a statistical method for reducing the dimensionality of data and retaining information at the same time. It is widely used in many fields, including machine learning, data mining, and image processing. Here is a summary of the principles of PCA [9]:

Suppose there have m row n-dimensional data:

step1. The raw data is formed into columns into n rows m column matrix X

step2. Each row of X is averaged at zero, by subtracting the mean of this line

step3. To find out the covariance matrix $C = \frac{1}{m} XX^T$[10]

step4. The eigenvalues of the covariance matrix and the corresponding eigenvectors are obtained

step5. The feature vectors are arranged into a matrix from top to bottom according to the corresponding feature value size, and the first k rows constitute the matrix P.

step6. $Y = PX$ That is, the data after the dimension reduction to the k dimension

After the post-dimension reduction index, this paper applied it to the random forest model for retraining.

Comparing the prediction effect before and after dimension reduction, this paper can get the results as Figure 5 and Figure 6 shown below:

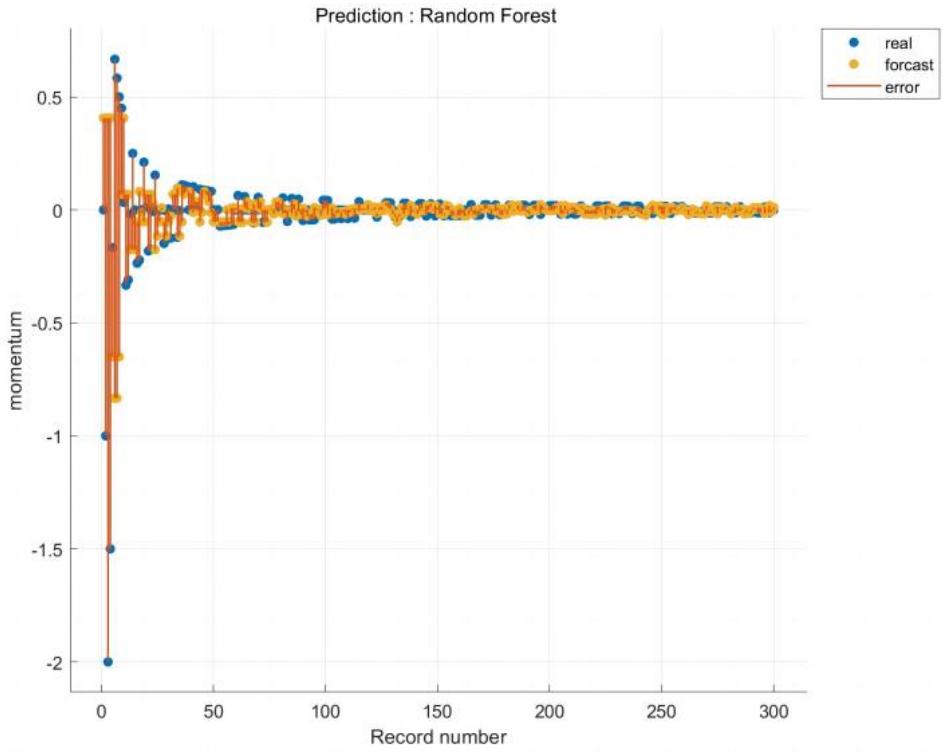

Figure 5. Error Before Dimensional Reduction

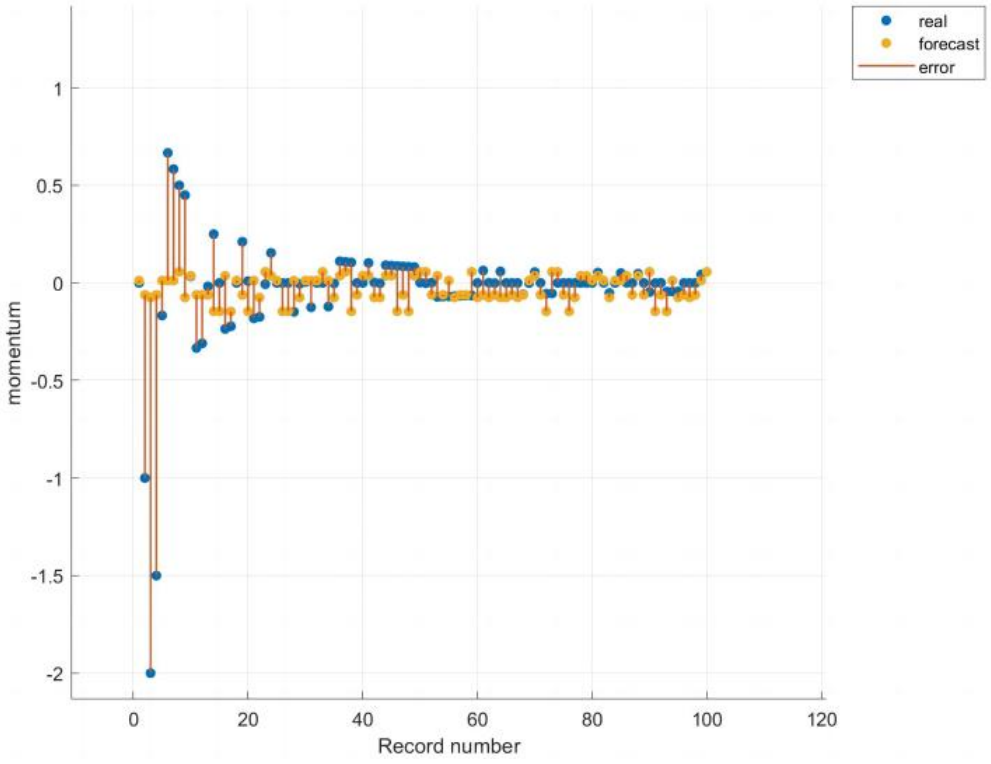

Figure 6. Error After Dimensional Reduction

Analysis the results in the Figure 5 and Figure 6, this paper concludes that the model error increases significantly after dimension reduction, indicating excessive model indicators, resulting in the model too fit to some extent, which should be paid attention to in the practical application.

3.4. Validation of the generalization ability of the momentum prediction model

The momentum model based on the random forest algorithm this paper constructed in the previous question achieved excellent prediction effect in the competition data of 2023 Wimbledon Gentlemen's final. However, in order that the model can be widely used in practice, it must have excellent generalization ability and application. To test the generalization ability of the model, this paper selected 201910260 ATL of quarter 1 from the game data of the 2019-2020 regular season to train the model. Considering the practical application of the model, to improve its generalization ability, this paper needs to consider other factors that may be included in the future model.

3.4.1. Feature vector selection.

(1) Attempt

This paper guesses that the type of shot scored would have an impact on the team's momentum fluctuations, and this paper assign the shot type as formula (19)

$$w_1 = \begin{cases} 2 & \text{attempt is 2} \\ 3 & \text{attempt is 3} \end{cases} \tag{19}$$

(2) The orientation of the pitch

This paper suggests that the orientation of the pitch influences the momentum fluctuations. Take the basket as the center, and the coordinate axis is established, and the projection of the ball on the axis when throwing the ball is taken as the indicator of affecting the momentum.

(3) Distance

This paper expects that the pitch distance influences the momentum fluctuations. Use the distance between the ball and the basket as an indicator of momentum.

(4) Score point

This paper argues that the scoring momentum fluctuations have an impact. The score is taken as an indicator of the momentum.

3.4.2. Training model.

This paper cleans the data, applied to the training and prediction of the momentum model. The data is applied to the model and solved in the r language.

To improve the fit of the model, this paper constructed the MES on Cross validation Fold and the MES on iteration as Figure 7.

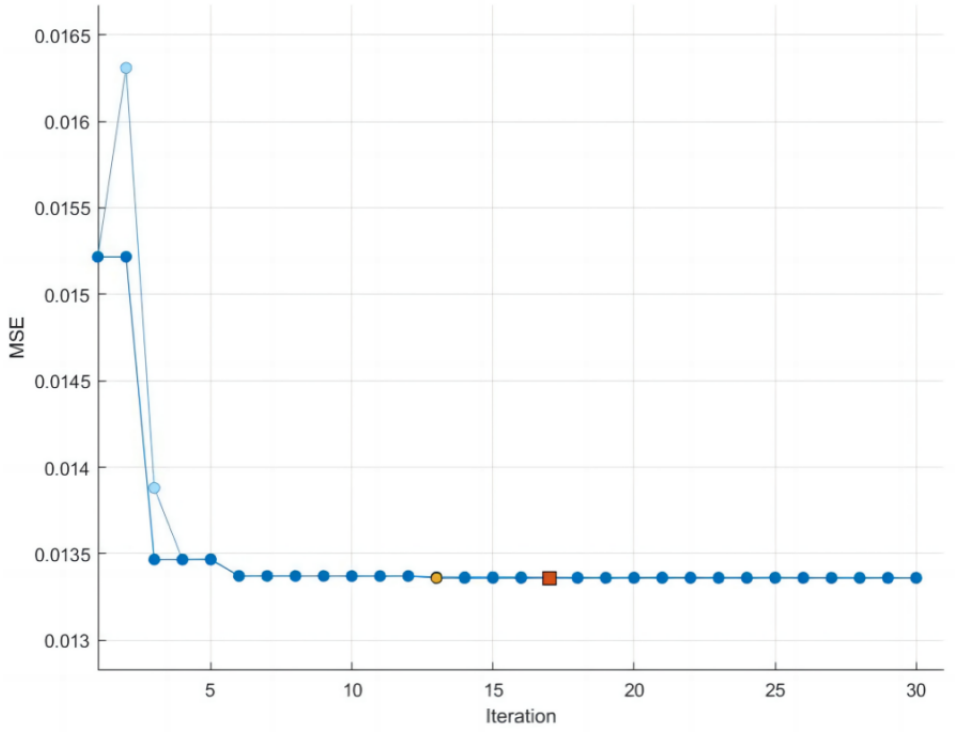

Figure 7. Minimum MSE

Based on the above results, this paper performed the following parameter settings for the random forest Model as Table 3.

Table 3. Random forest parameter setting

Number of cross-validated folds	iteration	min samples leaf
3-fold cross-validation	17	542

3.4.3. result analysis.

After adjusting the parameters, the output model results are shown in the following Figure 8.

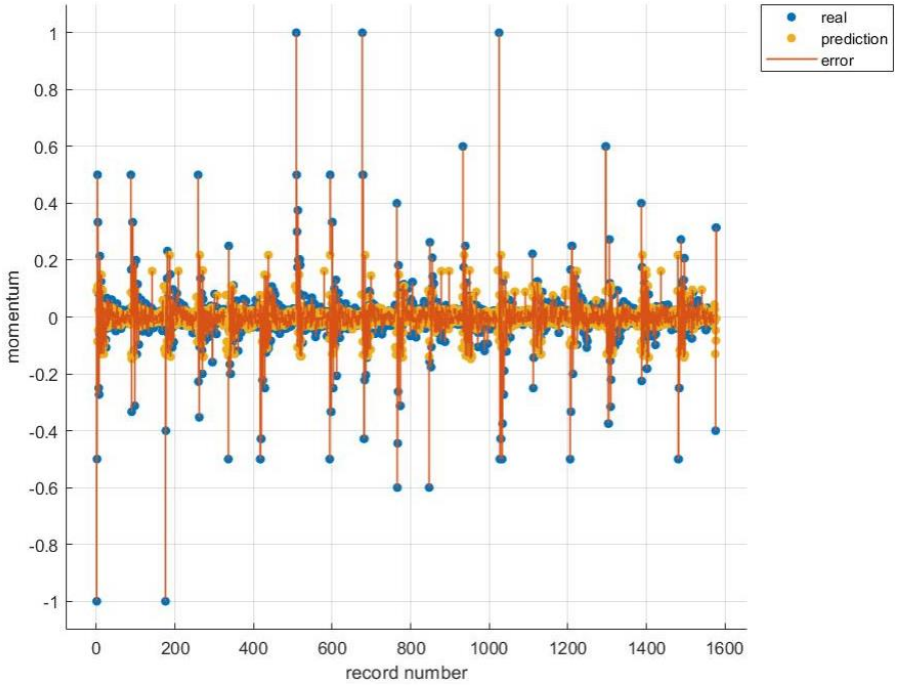

Figure 8. Result of NBA Players' Momentum Prediction

Analyse the data in the figure, this paper can conclude that most of the predicted values have small errors, but there are still some errors. However, this paper can consider that the model predicts better momentum for basketball players. The model should have an ideal generalization ability.

At the same time, this paper gets the weight value of each index of the model as Figure 9.

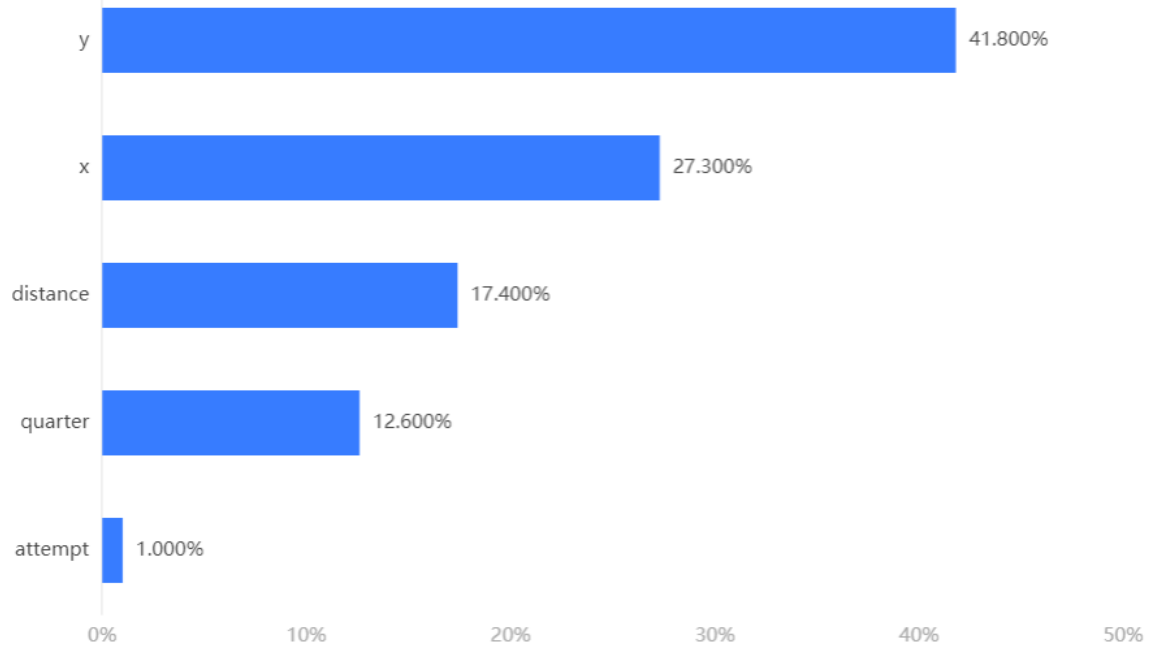

Figure 9. Weight of Four Indicators in Momentum Prediction Model

After analyzing the index weight, this paper can find that x, y and distance indicators have a large weight, so this paper can suggest that an ideal serve position and an appropriate serve distance will be conducive to the growth of momentum.

3.4.4 model test.

This paper calculated the regression metrics to further verify the generalization ability of the model. The calculation results of regression index of momentum prediction model based on random forest algorithm are shown as Table 4:

Table 4. Calculation of the model evaluation index

RMSE	R^2	MSE	MAE
0.12247	0.89	0.015	0.01075

The results, when RMSE = 0.12247, MSE = 0.01075 shows that the model fits ideal; R=0.89, ideal fitting result, small error, MAE = 0.01075, ideal fitting result and small error.

3.5. Proposal

According to the prediction results and feature weight values of the fit of the random forest regression model, this paper makes the following suggestions for the players to better face a new game:

(1) Prepare in advance

Collect historical data of new players, adjust mentality and “momentum”. To understand the situation of the opponent players, formulate the optimal training schedule, and conduct multiple simulation matches before the game is to better simulate the change of “momentum”, to find the best rhythm of the game and win the game.

(2) Good attitude

If player wins in the last set, the winning player should maintain the current mentality and quickly adjust the distribution of physical strength, re-formulate strategies and physical strength distribution. To let the players maintain the mentality of the last set, this paper hopes the players can continue the state of the last set in the next set, grasp the situation of the court as the previous set, and quickly establish their own advantageous situation. The players who lost the last set should adjust their mentality in time to avoid mistakes on the field, re-analyze the situation on the field, formulate a new game strategy, let themselves start again, and slowly save the situation on the field [10].

(3) Hold the rhythm

The two players before the new match, whether it is their own service, maintain a good mood, and quickly prepare coping strategies. It is to make the players better face some mistakes and lost points in the game. So as not to interfere with their own game rhythm. Can still stabilize their own mentality to reduce mistakes and lost points, grasp the players in the game of their own rhythm, adjust and grasp the "momentum" changes, to control the rhythm of the field, the formation of their own favorable game situation [10].

(4) Reduce Error

When the two players play a new game, it is "momentum" changes, to regain the rhythm of the game, take back the situation on the field and reduce mistakes.

(5) Play to the score

In the face of continuous scoring on the court, the scoring player should maintain the current rhythm, and develop strategies for a new round of games, adjust the rhythm of the game. Avoid strength due to improper physical distribution, and grasp their own "momentum" changes, resulting in changes in the situation on the field of play.

(6) Improve technique

Players should enhance their own skills before the game, such as: (1) Regular training is required. During the training, the accuracy of the service is paid more attention, including the control of the drop point, the training of different service modes such as flat stroke. (2) Put the weight of the whole body on the back leg and try to hold the racket more comfortably and easily, so that the power is better transmitted. (3) and master training or participate in more competitions and different people, and at the end of the humbly consult, and then to make up for the shortcomings of the real body. (4) Pay attention to their own physical training, such as improving their own muscle endurance, improving the ability to bounce, but also to strengthen the joint, improve core strength and flexibility, and for the back, legs and shoulders and other key parts of the planned strength training [11].

4. Conclusion

With the popularity of tennis, scholars have studied the concept of how to improve the performance of tennis players. In view of the momentum of tennis competition, the momentum model prediction based on random forest algorithm is established. Using the data of 2023 Wimbledon competition and nine main indicators are selected for the prediction and analysis of momentum. To eliminate the information overlap between independent variables, PCA is used to reduce the dimension of the indicators. This paper trains and verifies the NBA regular season data, further proves the generalization ability of the model, and provides a new idea for athletes in various sports events to improve their performance.

As the number of indicators increases, the model is better, but the model overfitting will be more obvious. Therefore, when making the analysis of the main indicators and establishing the model prediction, the main indicators should be effectively selected according to the specific situation of the predicted demand. The momentum prediction of tennis sport is affected by many factors. In this paper, only 9 indicators are introduced to establish the model prediction, and the variable indicators need to be further improved. How to introduce more correlated variables and make the variable index reflect

the momentum of tennis sport more comprehensively and scientifically needs further exploration and research. However, when the random forest algorithm data is used for training and analysis, there is often a long running time, and how to use a more efficient model needs further research.

References

[1] Shiguang H, Yuping J. Current Situation and Countermeasures of College Tennis Culture Development [J]. The Frontiers of Society, Science and Technology, 2020, 2 (5): 45 - 48.

[2] Rafael M, F J G, Miguel C et al. Technical, tactical and movement analysis of men's professional tennis on hard courts. [J]. The Journal of sports medicine and physical fitness, 2017, 59 (1):

[3] Rouli Y, Delong Z, Meng Z, et al. Nash equilibrium and tennis serve performance: a game theory analysis [J]. International Journal of Performance Analysis in Sport, 2023, 23 (6): 515 - 526.

[4] Sun Haifang, Hu Tianheng, Ma Yaxin, et al. Research on the flow soft measurement technology of gas pipeline compressor based on random forest [J]. Petrochemical Automation, 2024, 60 (01): 21 - 24.

[5] Liu Yunxiang, Wu Hao. A flood early warning model based on the random forest algorithm [J]. People's Yellow River, 2018, 40 (8): 4.

[6] Cheng Miaohai, Lou Qiao, Wang Qiong, et al. The prediction method of distribution network based on random forest algorithm [J]. Computer system applications, 2016 (9): 7.

[7] Xiong Biao, CAI Ting. Almost unbiased split k-d estimate of the semi-parametric regression model [J]. Journal of Hubei Normal University (Natural Science Edition), 2017, 37 (02): 62 - 67.

[8] Sun Ping'an, Wang is prepared for the war. The PCA Dimensionality Reduction Method in Machine Learning and its Application [J]. Journal of Hunan University of Technology, 2019, 33 (1): 6.

[9] Baiyun, Yan Hua, Wei Yuankun. Temperature field reconstruction based on PCA dimension reduction and iterative regularization [J]. Automation and instrumentation, 2023, 38 (9): 21 - 26.

[10] Zhang Shaokun, Lu Feifei. Investigation and research on the status quo of college tennis sports programs in Nanchang city [J]. Sports & Sports and Technology, 2020, (01): 21 - 22.

[11] Jiang Yang. Research on the psychological influencing factors of tennis players and their training methods [J]. Exercise and Health, 2023 (3): 012 3- 0125.