Research on the Impact of Momentum on Game Situations Based on Random Forest and XGBoost Models

Haiyang Qiu*

School of Mathematics, Shanghai University of Finance and Economics, Shanghai, China, 200433

* Corresponding Author Email: 1473559172@qq.com

Abstract. The momentum in the realm of sports is like an intangible force, unleashed by a sequence of events. During a game, a team or player may feel they are riding this wave as if victory is within reach. However, the formation of momentum and its impact on the outcome of the game remains a mystery. To study the factors that influence momentum and explore their impact on it, ultimately understanding how they affect the course of a game, This study establishes a Random Forest Model to calculate the weights of these independent variables which we use to define the momentum and XGboost model to detect momentum combined with using SHAP Values to analyze the impact of different feature quantities on match fluctuations to learn about the impact of variable parameter changes on results. Finally, we improve strategic decision-making through genetic algorithm optimization based on the XGboost model to find the optimal allocation of indicators that maximize player momentum. This research has the ability to help us figure out the significance of momentum exploration in competitive scenarios and how to optimize momentum.

Keywords: Random Forest Model, XGBoost, SHAP Analysis.

1. Introduction

The momentum within the sports domain resembles an elusive energy, set in motion by a sequence of occurrences. In the midst of a game, a team or athlete may sense themselves riding this surge, as if victory is on the horizon. However, the precise genesis of momentum and its profound impact on the game's outcome remains shrouded in mystery [1].

Prior research has demonstrated that athletes perceive that momentum exists, but evidence of the effect of momentum on performance within individual athletic contests has proved elusive. This research extends that exploration by looking for momentum over a season. Actual winning and losing streaks for the 28 major league baseball teams and the 29 National Basketball Association teams were compared to streaks that would have occurred under the assumption that game outcome is independent of the outcome of the most recent previous games. The chi-square goodness-of-fit test showed a very close fit of actual streaks to expected streaks under the independence assumption, with none of the ten chi-square probabilities approaching the customary 0.05 significance level. Also, the Wald-Wolfowitz runs test for randomness produced only 5 team season observations out of 86 with significance levels less than 0.05. The results suggest that sports participants and observers place an unjustified importance on momentum as a causal factor in outcomes of sports contests [2]. Prior research has several limitations and shortcomings that should be acknowledged. Its studies rely on the assumption that game outcomes are independent of the previous game outcomes. While this assumption allows for a comparison of actual streaks to expected streaks, it oversimplifies the complex nature of sports performance. Various factors, such as team morale, player injuries, or strategic adjustments, may influence game outcomes and potentially introduce bias into the analysis. the chi-square goodness-of-fit test and the Wald-Wolfowitz runs test used in the study have their own limitations. These statistical tests assume certain conditions and may not fully capture the intricacies of momentum and its impact on game results. Alternative statistical approaches or additional measures could provide a more comprehensive analysis of momentum in sports [3].

However, our study possesses several strengths. Firstly, it demonstrates rigor by thoroughly analyzing the concept of momentum and constructing a random forest model to calculate the weights of various indicators, resulting in a more accurate quantification of momentum [4]. Secondly, the

comprehensiveness of the data is ensured by using one-hot encoding to convert discrete categorical variables into binary vector representations [5]. This allows for a better expression of relationships between different categories in machine learning algorithms, avoiding the introduction of assumptions and ensuring all indicators are considered. Thirdly, the accuracy and persuasiveness of the results are evident as various machine learning model algorithms are compared, with the optimal XGBoost model selected based on their performance in terms of accuracy, stability, and efficiency [6-8]. This ensures the most suitable model is utilized to solve the problem, providing the best prediction or decision-making outcome. Lastly, the study highlights high customizability as genetic algorithms can be tailored to specific problems and needs. By defining the fitness function based on the objective function and constraints and setting appropriate genetic algorithm parameters, personalized optimization of the XGBoost model can be achieved. In summary, these strengths contribute to a thorough analysis of momentum, accurate quantification, and personalized optimization, ultimately providing valuable insights for predicting and maximizing player momentum [9-10].

2. The establishment of the models

2.1. The basic fundamental of the Random Forest Model

Random forest regression is a process of generating numerous decision trees, where each tree is constructed by randomly sampling observations and features from the modeling dataset. Each sampled result represents a separate tree, and each tree generates its own set of rules and decision values based on its attributes. The random forest algorithm combines the rules and decision values from all the decision trees to perform regression. The Random Forest structure is shown in Figure 1.

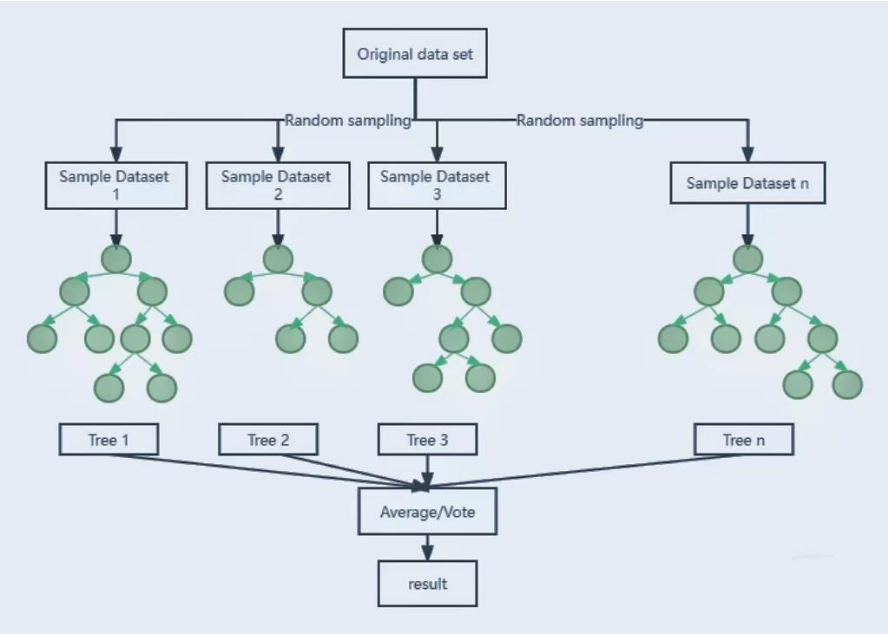


Figure 1. Random Forest structure

Random forest is an ensemble learning method consisting of multiple decision trees. Here is an explanation of its principles:

1. Random Sampling: Randomly sample the training data set with replacement to generate multiple distinct training subsets, each containing a portion of the original data.

2. Decision Tree Construction: For each training subset, independently construct a decision tree model. During the construction process, features are recursively split based on selection criteria to maximize predictive accuracy.

3. Ensemble Prediction: For regression problems, random forest combines the prediction results of all decision trees by averaging or weighted averaging to obtain the final regression prediction. The

weights of each decision tree's prediction result can be adjusted based on the tree's performance or sample weights.

For regression problems, the random forest's prediction result can be obtained using the following formula:

$$\hat{Y} = \frac{1}{N}\sum_{i=1}^{N} f_i(X) \tag{1}$$

where $\hat{Y}$ is the random forest's prediction result, N represents the number of decision trees, and $f_i(X)$ indicates the prediction result of the i decision tree for input sample X.

2.2. The establishment of the XGboost Model

XGBoost is an efficient implementation of Gradient Boosting Decision Trees (GBDT). Unlike traditional GBDT, XGBoost introduces regularization terms to the loss function. Additionally, XGBoost handles loss functions that are difficult to compute derivatives for by using the second-order Taylor expansion to approximate the loss function.

For a dataset containing n instances with m features, the XGBoost model can be represented as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i),\ f_k \in \mathcal{F}, (i = 1, 2, \ldots, n) \tag{2}$$

$$F = \{f(x) = w_q(x)\}\ \left(q: \mathbb{R}^n \to 1,2,\ldots,T, w \in \mathbb{R}^T\right)] \tag{3}$$

Here, F is a collection of CART decision tree structures, q represents the mapping of samples to leaf nodes in the tree structure, T is the number of leaf nodes, and w is the real-valued score of the leaf nodes. When constructing an XGBoost model, the optimal parameters need to be found based on the principle of minimizing the objective function to establish the best model. The objective function of the XGBoost model consists of two terms: the loss function term L and the model complexity term Ω. The objective function can be written as:

$$\text{Obj} = L + \Omega \tag{4}$$

$$\begin{aligned} L &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ \Omega &= \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \end{aligned} \tag{5}$$

where γT is the L1 regularization term and $\frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2$ is the L2 regularization term

During the optimization training of the model using the training data, the original model is preserved while introducing a new function f into the model to minimize the objective function. The detailed process is as follows:

$$\begin{aligned} \hat{y}_i^{(0)} &= 0 \\ \hat{y}_i^{(1)} &= \hat{y}_i^{(0)} + f_1(x_i) \\ \hat{y}_i^{(2)} &= \hat{y}_i^{(1)} + f_2(x_i) \\ &\cdots\cdots \\ \hat{y}_i^{(t)} &= \hat{y}_i^{(t-1)} + f_t(x_i) \end{aligned} \tag{6}$$

where: $\hat{y}_i^{(t)}$ is the prediction of the t-th model and $f_t(x_i)$ is the newly added function at the t-th iteration. The objective function at this point is given by:

$$Obj^{(t)} = \sum_{i=1}^{n}\left(y_i - \left(\hat{y}_i^{(t-1)} + f_t(x_i)\right)\right)^2 + \Omega \tag{7}$$

In the XGBoost algorithm, to quickly find the parameters that minimize the objective function, the objective function is approximated by a second-order Taylor expansion, resulting in an approximate objective function:

$$Obj^{(t)} \approx \sum_{i=1}^{n}\left[\left(y_i - \hat{y}_i^{(t-1)}\right)^2 + 2\left(y_i - \hat{y}i^{(t-1)}\right) f_t(x_i) - hf_t(x_i)\right] + \Omega \tag{8}$$

after discarding the constant term, it can be observed that the objective function only depends on the first and second derivatives of the loss function. At this point, the objective function can be represented as:

$$\begin{aligned} Obj^{(t)} &\approx \sum_{i=1}^{n}\left[g_i w_q(x_i) + \frac{1}{2}h_i w_q^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)w_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right)w_j^2\right] + \gamma T \end{aligned} \tag{9}$$

If the structure part q of the tree is known, the objective function can be used to find the optimal Wj and obtain the optimal objective function value. This essentially becomes a problem of minimizing a quadratic function. The solution is given by:

$$w_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda} \tag{10}$$

$$Obj = -\frac{1}{2}\sum_{j=1}^{T}\left[\frac{(\sum_{i\in I_j} g_i)^2}{\sum_{i\in I_j} h_i + \lambda}\right] + \gamma T \tag{11}$$

Obj can serve as a scoring function to evaluate the model, where a smaller Obj value indicates better model performance. By recursively calling the tree construction method described above, a large number of regression tree structures are obtained, and Obj is used to search for the optimal tree structure to be incorporated into the existing model, thus establishing an optimal XGBoost model.

3. Results

3.1. Analysis of Random Forest Results

For setting weights, we first process the data in the Excel spreadsheet and organize the considered indicators into separate columns according to the calculation formula (organized in the table below). Then, we select the data for the match with the ID 2023-wimbledon-1701, import it into SPSS Pro, perform a random forest regression, and calculate feature importance, which is shown in Figure 2.

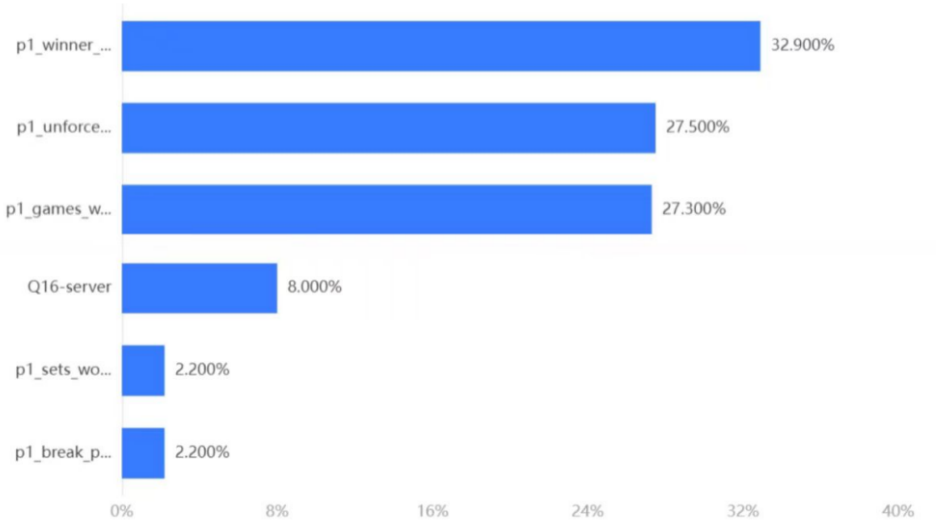


Figure 2. Feature importance proportion

From Figure 2, we conducted multiple groups of random forest regressions using data from different competitions and players to highlight more significant features. Therefore, we set the serve advantage weight to 0.1 to define and calculate the momentum.

Next we visualize the momentum as shown in the following Figure 3.

According to Figure 3, we observe that the momentum value of Player 1 peaks at 10.4 around 4:18:35 and dips to its lowest at -10.9 by 3:53:55. Conversely, Player 2 experiences their lowest momentum at -10.8 by 2:02:01, with their momentum reaching a high of 10.4 at 3:51:26.

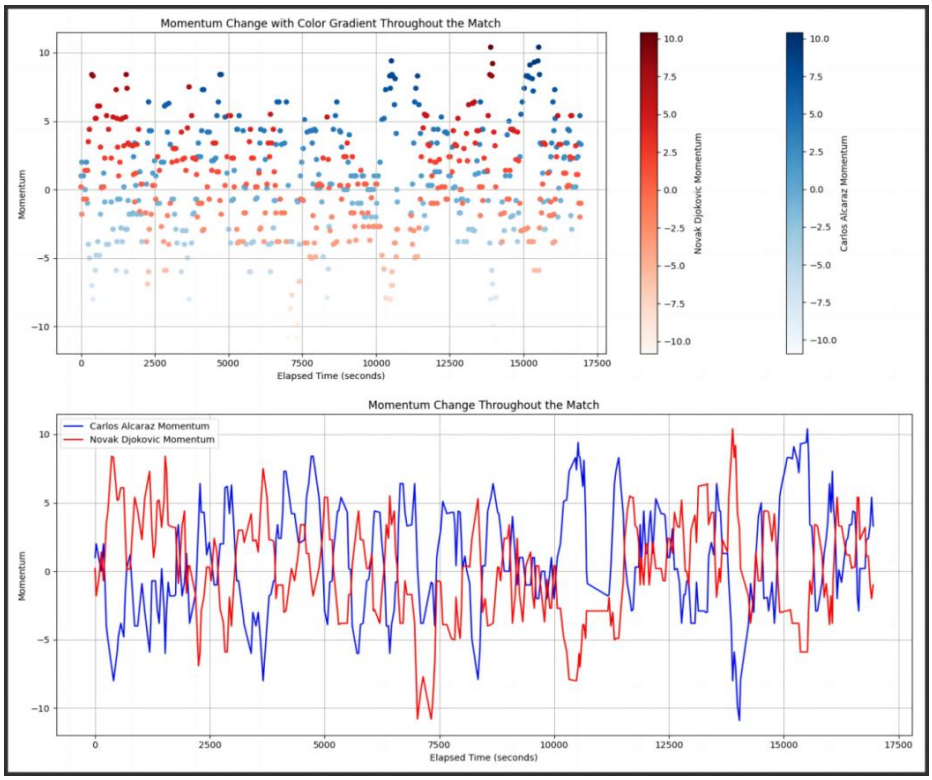


Figure 3. Line chart and Scatter plot of the momentum change of the match

3.2. Analysis of XGboost results

We convert discrete categorical variables into binary vector representations and utilize the XGBoost model, using Python programs to implement the models and generate data, which is presented in Figure 4.

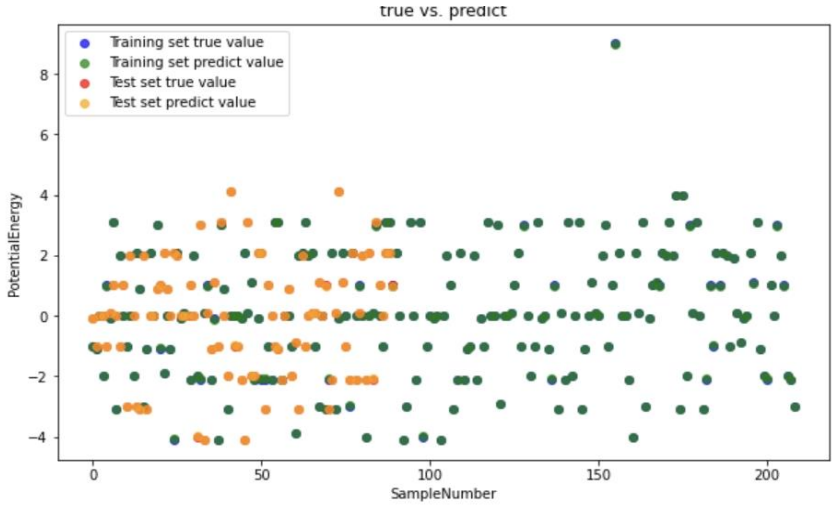


Figure 4. XGBoost Regression

In the XGBoost regression model, the mean absolute percentage error (MAPE) on the training set is 0.177 and on the testing set is 0.193. The mean absolute error (MAE) is 0.004, indicating the average absolute difference between the predicted and actual values. Furthermore, the R2 score of 0.99 suggests that the model explains 99% of the variance in the target variable.

Then, we make use of this model to calculate the SHAP value, The final SHAP values for each feature can be visualized as the Figure 5. From Figure 5, it can be seen that p1_distance_run, p2_distance_run, and p2_score have the greatest impact on the fluctuations.

Through the analysis of SHAP values and the relative contribution of each feature they reveal to the prediction outcome, we can assess the predictive capacity of the XGBoost model. In this analysis, we observed that certain features, such as ”p1_distance_run” with a SHAP value of 0.355936, have a significant impact on the model’s predictive results, indicating that this feature plays an important role within the model. Other features like game_no, point_no, serve_width, and winner_shot_type also exhibit higher SHAP values, further confirming their importance in the model’s decision-making process.

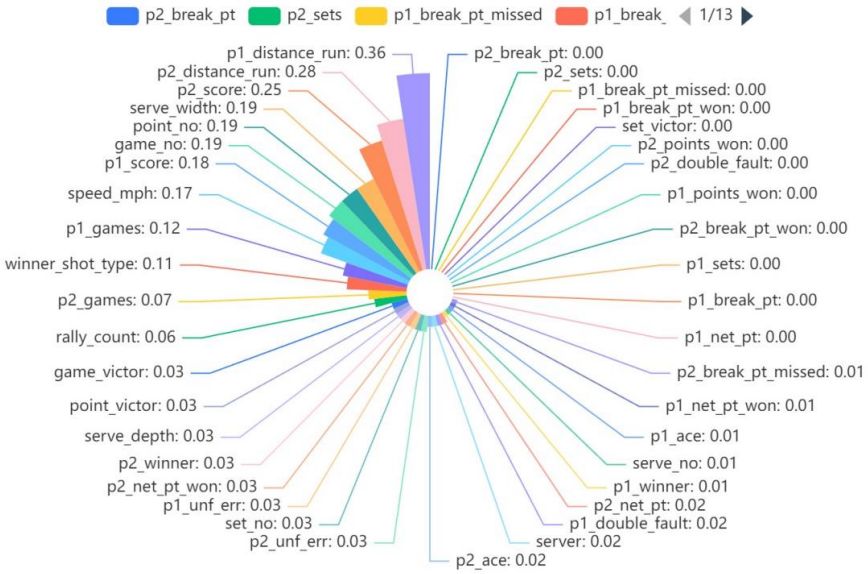


Figure 5. The Nightingale Rose Chart of SHAP values

3.3. Analysis of Genetic Algorithm Results

In our experiments, we set the population size to 100, the number of generations to 50, and the mutation rate to 0.01. By calling the ‘genetic_algorithm‘ function, we obtained the best parameter combination and the best score found.

Initially, the parameter combination for the features (p1_ace, p1_winner, p1_double_fault, p1_unf_err, p1_net_pt, p1_net_pt_won, p1_break_pt, p1_break_pt_won, p1_break_pt_missed) was [0, 1, 0, 0, 0, 0, 0, 0, 0], and it achieved a score of 2.19. However, after improvement using the Genetic Algorithm, the best parameter combination was found to be [1, 1, 1, 1, 1, 0, 0, 1, 1], resulting in a higher score of 2.42.

Therefore, through the genetic algorithm, we can make timely adjustments to Player 1's behavior based on relevant data to achieve the maximum momentum value, thereby reaching optimal performance.

4. Conclusions

This report summarizes a comprehensive analysis of momentum in tennis matches using detailed data from the 2023 Wimbledon tournament. By employing advanced analytical modeling and machine learning techniques, we quantified momentum and visualized its impact on match dynamics. Through the utilization of a Random Forest model and one-hot encoding, we transformed categorical variables into binary vectors. The XGBoost model emerged as the most suitable for predicting momentum changes, with 'p1_distance_run', 'p2_distance_run', and 'p2_score' identified as significant factors affecting match dynamics through SHAP value analysis. To optimize the model's performance, we employed a genetic algorithm to determine the optimal combination of features. The resultant model demonstrated high predictive accuracy, which was further validated by successfully predicting outcomes in a dataset of women's singles tennis matches. Overall, this study showcases the importance of momentum in tennis matches and provides insights into the relationship between momentum shifts and match results. The findings highlight the potential of utilizing match data patterns and indicators to predict and understand momentum shifts in tennis.

References

[1] Araújo D, Davids K, Renshaw I. Cognition, emotion and action in sport: an ecological dynamics perspective [J]. Handbook of Sport Psychology, 2020: 535-555.

[2] Virgin R C. Winning Streaks in Sports and the Misperception of Momentum[J]. Journal of Sport Behavior, 2000, 23(2)

[3] .Jones M I, Harwood C. Psychological momentum within competitive soccer: Players' perspectives[J]. Journal of Applied Sport Psychology, 2008, 20(1): 57-72.

[4] Shaik A B, Srinivasan S. A brief survey on random forest ensembles in classification model[C]// International Conference on Innovative Computing and Communications: Proceedings of ICICC 2018, Volume 2. Springer Singapore, 2019: 253-260.

[5] Okada S, Ohzeki M, Taguchi S. Efficient partition of integer optimization problems with one-hot encoding [J]. Scientific Reports, 2019, 9(1): 13036.

[6] Tsai CC, YiuHT. Investigation of Laser Ablation Quality Based on Data Science and Machine Learning XGBoost Classifier [J]. Applied Sciences, 2023, 14(1).

[7] Nadh S S, Sarit C, Mohammad A, et al. Explainable XGBoost–SHAP Machine-Learning Model for Prediction of Ground Motion Duration in New Zealand [J]. Natural Hazards Review, 2024, 25(2).

[8] NasrinE, M. SB, Changhyun J, et al. Estimation and interpretation of equilibrium scour depth around circular bridge piers by using optimized XGBoost and SHAP [J]. Engineering Applications of Computational Fluid Mechanics,2023,17(1).

[9] Xiangcheng L, Jialong W, Zhirui G, et al. Short-term Wind Power Prediction Method Based on Genetic Algorithm Optimized XGBoost Regression Model [J]. Journal of Physics: Conference Series, 2023, 2527(1).

[10] Katoch S, Chauhan S S, Kumar V. A review on genetic algorithm: past, present, and future [J]. Multimedia tools and applications, 2021, 80: 8091-8126.