Prediction Of Tennis Players' Court Behaviors Based on The Xgboost Model and The Shap Model

Feiyi Li *, Xinyue Zhou #, Yingwei Liu #

School of Economics, Hebei University, Baoding, China, 071032

* Corresponding Author Email: 19133013097@163.com

#These authors are contributed equally.

Abstract. In tennis competitions, coaches and athletes are in urgent need of quantifiable indicators to comprehensively assess the game situation and make timely tactical adjustments. To meet this crucial demand, this article constructed a "momentum" evaluation and prediction model. Initially, this article conducted an in - depth analysis to screen out variables that impact a player's on - court performance, systematically dividing them into two major categories: game - wide variables and local variables. This article placed particular emphasis on local variables, as they capture the essence of a player's outstanding plays, such as powerful aces or incredible retrievals, which can significantly shift the momentum of the game. This influence is effectively represented through a sophisticated weighting mechanism.Based on the robust XGBoost model, this article's prediction model is not only highly efficient in processing and analyzing data but also highly practical for real - time application during matches. By quantifying the momentum of both players, it provides valuable support for immediate strategy adjustments. Moreover, the model can generate detailed reports and suggestions based on the momentum data, offering targeted guidance for athletes' daily training and helping coaches formulate more scientific training plans, thus enhancing the overall competitiveness of the players.

Keywords: Momentum Quantitative Scoring Model, XGBoost Regression Model, Shap Model.

1. Introduction

Tennis, as one of the important sports events, tennis matches are also highly valued. In some competitions, a dramatic competition process may be demonstrated. For instance, a player who is at a disadvantage at the beginning gradually gains the upper hand as the competition progresses and wins. Therefore, this article speculate that there is something in the game that can reveal the situation on the field, which is called momentum. In tennis, momentum can be defined as a measure of a player's current performance. To analyze the influence of momentum and momentum in a game, this article will analyze a large number of historical data on the wins and losses of matches, including the match time, the winning time, the scoring situation in each match, each set, and each game, the serving player, the consecutive scoring situation, etc., to quantify these indicators and establish a reasonable model to explain the relationship between winning and match time and match state. To provide mathematical data basis and suitable models for reasonably arranging the competition sequence of team members and winning the game.

Tennis, as a highly competitive sport, has a profound impact on the outcome of the game due to both technical and psychological factors. Through a comprehensive analysis of the literature, we can draw the conclusion that momentum and momentum fluctuations [1] in tennis matches are complex and multi-dimensional. The beginning of a tennis match is of vital importance, and the serving technique [2] plays a decisive role at this stage. A good serve [3] can gain an advantage in the game and boost an athlete's momentum. The time pressure [4] also has a significant impact on the performance of athletes, affecting their prediction of the landing point [5] when receiving serves. Research is conducted on the tactical choices during the rallies [6]. Hitting the diagonal ball has a tactical advantage. Compared with the straight ball, it can make the opponent's defense run a longer distance. From a physics perspective, the forehand hit of tennis players [7] emphasizes the influence of the moment of inertia of the body and the contraction intensity of the muscles on the hitting effect.

It provides a theoretical basis for athletes to optimize their hitting by adjusting the swinging of their limbs. In addition, the reasonable regulation of psychological states [8] and spatio-temporal characteristics [9] is conducive to the stability of momentum. Real-time monitoring of heart rate and analysis of facial expressions and other methods provide more comprehensive data support for quantifying the impact of momentum and momentum on the game.

In the research of this article, quantitative analysis [10] of these factors will be conducted to help us gain a deeper understanding of athletes' performance and competition results, so as to improve the scientificity and accuracy of athletes' training and competition decisions.

This article proposed the "momentum" prediction model to illustrate and predict the changes in the momentum of athletes in tennis matches. Provide advice for athletes on the choice of battle strategies. During the model construction process, we use the momentum quantification scoring model and the run-length test method to visualize, digitize and verify the relevant information. Based on the XGBoost model for prediction, we use the shap model to find the variables that have a strong impact on the athletes' momentum, thereby providing strategic suggestions for athletes to seize the momentum advantage. Finally, in the memo, we summarized our results, the role of momentum in the competition, and provided advice for athletes to choose appropriate strategies in the competition.

2. The basic principles of XGBoost regression model and shap model

2.1. Structure of the XGBoost model

XGBoost (eXtreme Gradient Boosting) is an efficient machine learning algorithm based on the gradient boosting framework, which achieves model optimization by integrating multiple decision trees. The core idea is to train the weak learner (decision tree) iteratively, gradually reduce the prediction error, and finally combine it into a strong learner. XGBoost performs exceptionally well in terms of efficiency, accuracy and scalability, and is widely used in classification, regression and ranking tasks. The model structure is shown in Figure 1.

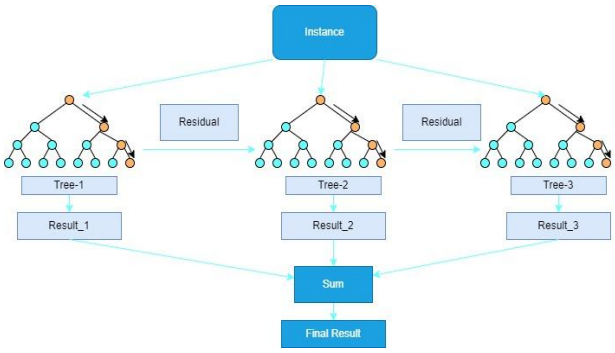

Figure 1. XGBoost Model

The general model of XGBoost is composed of three basic elements, namely:

(1) Input data: As the input layer of the model, it contains various feature data used for training and prediction. These characteristic data will serve as the basis for the model to make decisions and predictions.

(2) Regression tree set: The core processing unit of the model is composed of multiple regression trees. Each regression tree splits the samples through nodes based on different features, divides the samples into different child nodes, and finally gives the predicted values at the leaf nodes. The prediction results of multiple trees jointly determine the final output.

(3) Objective function: It includes the loss function and the regularization term. The loss function measures the difference between the predicted values of the model and the true values; The regularization term is used to prevent overfitting and limit the complexity of the model.

$$L = \sum_{i=1}^{n} loss(y_i, \hat{y_i}) + \sum_{k=1}^{K} R(f_k) \quad (1)$$

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (2)$$

$$w_j^* = -\frac{G_j}{H_j + b} \quad (3)$$

The model mainly consists of three parts: input data, multiple regression trees and the objective function. The input layer data features are processed through multiple regression trees. The tree nodes split the samples based on the features, and the leaf nodes output the predicted values. The objective function consists of a loss function and a regularization term. The former measures the difference between the predicted value and the true value, while the latter avoids overfitting.

During its working process, the predicted values are initialized first, the residuals are calculated and a regression tree is constructed accordingly, and the objective function is optimized through gradient descent. During training, tree pruning and feature subsampling are carried out to prevent overfitting. Finally, the prediction results of multiple trees are weighted and integrated to obtain the final prediction, completing the model training and prediction.

2.2. SHAP Model Structure

SHAP (SHapley Additive exPlanations) is a model interpretation method based on game theory, aiming to quantify the contribution of each feature to a single prediction result. The core idea is to fairly distribute the "benefits" of the prediction results through Shapley values to ensure the fairness and consistency of feature contributions. The SHAP model is applicable to the interpretation of any machine learning model, and its structure is shown in Figure 2.

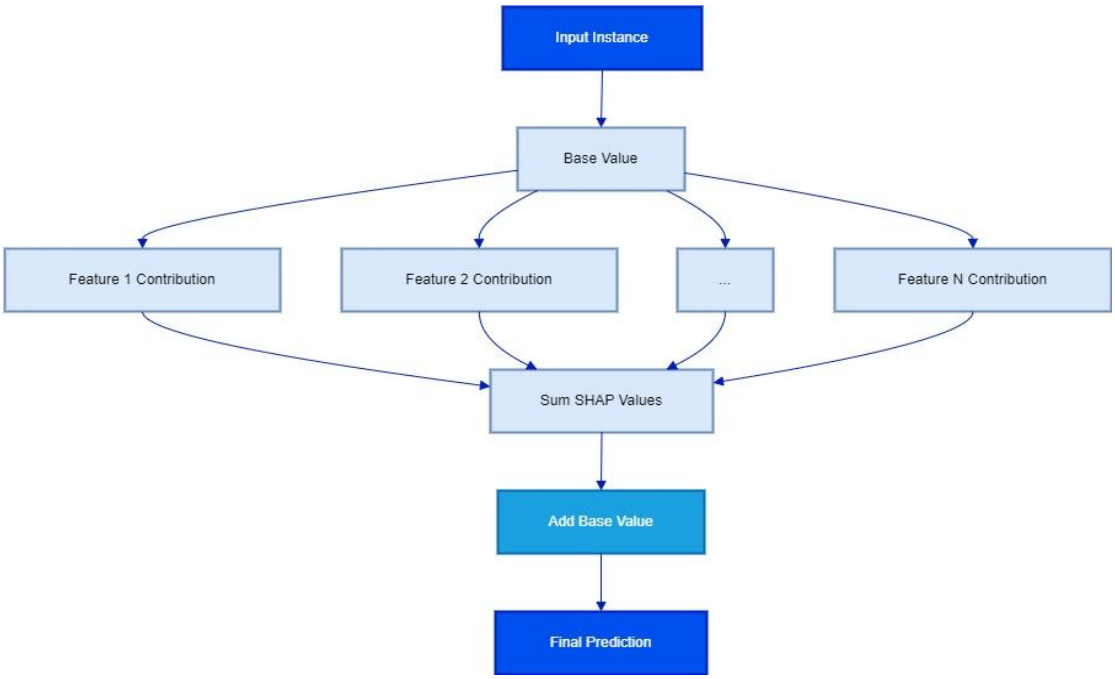

Figure 2. Structure of shap model

The general model of shap is composed of three basic elements, namely:

(1)Shapley value calculation mechanism: Based on the fair distribution principle of cooperative game theory, by considering all possible addition or removal sequences of features, the marginal contribution of features to the model's prediction results is calculated to measure the importance of each feature.

(2)Feature attribution: For each predicted sample of the model, the prediction result is decomposed into the sum of the contributions of each input feature, plus a base value (usually the mean of the predicted values of all samples). The SHAP value of each feature represents the magnitude and direction of the influence of that feature on the final prediction result. A positive value indicates that

the feature increases the predicted value, while a negative value indicates that it decreases the predicted value.

(3) Background dataset: A collection of partial samples extracted from the original training data, used to reflect the overall distribution characteristics of the data, serving as a reference benchmark for calculating feature contributions and simulating various possible scenarios for feature values.

For a model with a feature, the calculation formula of its Shapley value is as follows:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} \tag{4}$$

$$y = w_0 + \sum_{j=1}^{n} w_j x_j \tag{5}$$

The model mainly consists of three parts: the background data set, the calculation mechanism of Shapley values, and feature attribution. The background data set is used to reflect the data distribution and serves as the calculation benchmark; The calculation mechanism of the Shapley value is based on the fair distribution principle of game theory to measure the contribution of features to the prediction results. Feature attribution breaks down the prediction results into the sum of the contributions of each feature and the fundamental values to achieve the interpretation of the model.

During its working process, the background data set is selected first. Then, for the samples to be explained, all feature combinations are considered to simulate the changes in the model's predicted values when features are added or removed. The SHAP values of each feature are obtained by weighted summation using the Shapley value calculation method. Finally, the importance of the features is determined by ranking the absolute values of the SHAP values. Finally, feature contributions are presented through various visualizations to complete the global or local interpretation of the model.

3. Results

3.1. Analysis of the first model results

This article obtained the preliminary test results through XGBoost regression on the data. The average absolute percentage error of the training set was 0.169, and that of the test set was 0.194. The average absolute error was 0.006563508127583203. Clearly, by comparing the average absolute percentage errors of the test sets of the three models, the XGBoost model has better predictive ability, as shown in Figure 3.

Figure 3: True vs.Predict 1

This article will use the prediction model we have built to provide competition advice for athletes. Starting from the adjustable variables of the athletes, we will predict by adjusting different variable combinations to find the combination that can increase the momentum the most and thereby enhance the athletes' momentum. Taking the data at "0:01:31" in the "2023-wimbledon-1301" match as an example, after substituting the data at this moment into the XGBoost model, since the values of the nine variables are only "0" or "1", we can simulate different combinations of the nine variables to seek the combination that maximizes the increase in momentum. After solving with the XGBoost model, this article can get this result:

The optimal combination is p1_ace:1,p1_winner:0,p1_double_fault:0, p1_unf_err:0, p1_net_pt:0, p1_net_pt_won:0, p1_break_pt:1, p1_break_pt:1. p1_break_pt_missed:0

By substituting the other competition data that have been given into the established model, the accuracy of our model is evaluated through indicators such as MSE (Mean Square Error), RMSE (Root Mean Square Error), MAE (Mean Absolute Error), and MAPE (Mean Absolute Percentage Error).

This article evaluated the model by substituting the data of the three matches, 2023-wimbledon-1302, 2023-wimbledon-1304, and 2023-wimbledon-1305, from the match data into the model we created. The evaluation results are as show in Table.1, Table.2 and Table.3:

Table.1. the result of 2023-wimbledon-1302

	MSE	RMSE	MAE	MAPE	R^2
Training set	0	0.001	0.001	17.94	1
Test set	0	0.001	0.001	18.108	1

Table.2. the result of 2023-wimbledon-1304

	MSE	RMSE	MAE	MAPE	R^2
Training set	0	0.001	0.001	12.26	1
Test set	0	0.001	0.001	12.194	1

Table.3. the result of 2023-wimbledon-1305

	MSE	RMSE	MAE	MAPE	R^2
Training set	0	0.013	0.01	16.059	1
Test set	0	0.013	0.01	16.075	1

From the above three examples of data, it can be concluded that the evaluation result of the model we established is good, and it can predict the change value of the momentum on the competition field relatively accurately. If more data are used further, the evaluation of our model will be even more accurate.

3.2. Analysis of the second model results

In the future, due to the limited number of influencing variables provided by the data, there is a probability that the model will perform poorly in the future. At this point, we can consider several factors that have a significant impact on the prediction model as analyzed by the previous shap model.The result is shown in Figure 4.

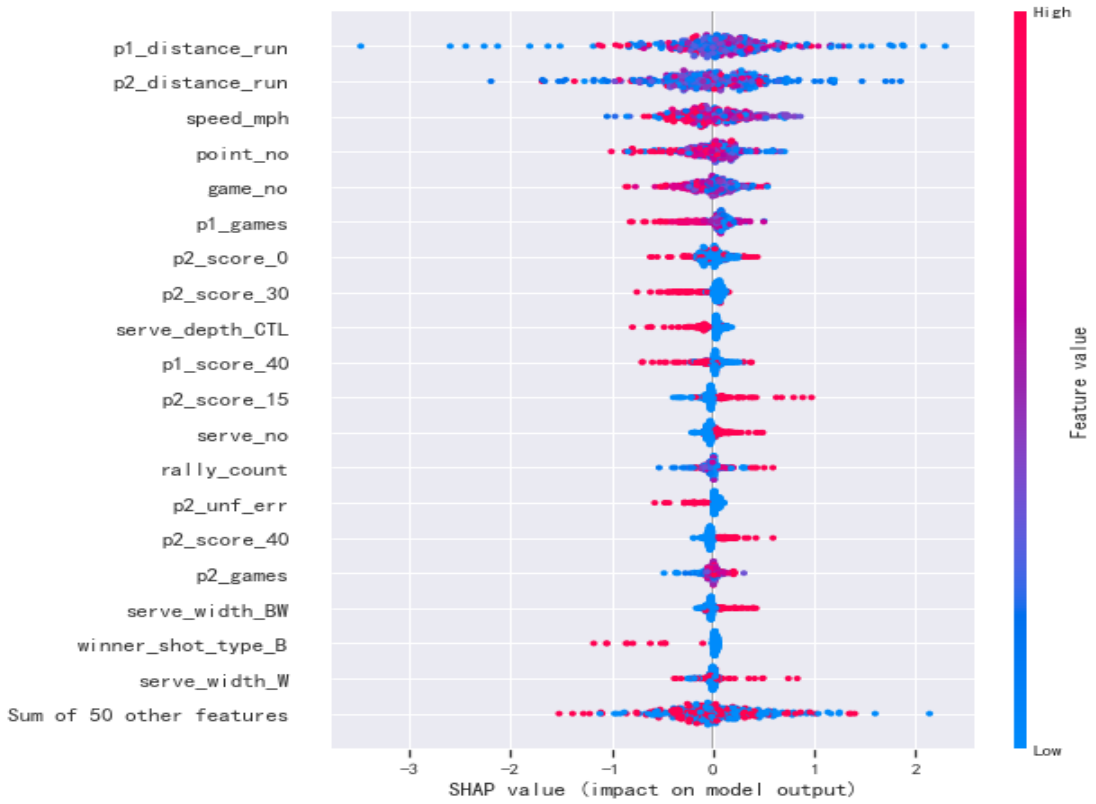

Figure 4. the factors in the eigenvalue analysis graph

As shown in Figure 4, the eigenvalue analysis graph indicates that p1_distance_run and p2_distance_run are the most critical factors. Consequently, when predicting model failure, we should prioritize examining variables associated with these two metrics, such as the proportion of front-field running versus back-field running. This approach helps clarify whether a player tends to stay near the baseline or actively advance toward the net. For instance, the frequency of short-distance sprints can effectively reflect a player's explosive power and ability to handle critical moments. Furthermore, these factors may all serve as potential indicators of momentum changes in the game.

Secondly, when the model performance is not very ideal, this article can also introduce more other variables, such as the different types of competitions causing the contestants to be unaccustomed to the competition system, or the relationship between the contestants and their opponents. If the contestants lose to their opponents in previous competitions, it may cause psychological pressure on them when they compete again, thereby affecting the contestants' performance in the competition.

4. Conclusions

In this paper, this article has improved the traditional XGBoost model to adapt to the rapidly changing court situation, identified the key factors for analysis, and provided a direction for subsequent model optimization.This article’s model mainly has two parts of functions, namely prediction and decision-making. The prediction results of the XGBoost model are input into the decision-making model. Output the decision-making results this article needs for the players to adjust their tactics. Finally, this article evaluated the prediction model and decision-making model this article established, providing guidance for the optimization direction and main ideas of the models.

When conducting self-evaluation, the "momentum" prediction model of this paper has many advantages but also some shortcomings. For example: Large requirement for the quantity of available data: Since the main component of this model is the XGBoost model, it also has the defect of excessive demand for computing resources at the same time; Sensitive to outliers: When this model is in use, data preprocessing is required to handle outliers; otherwise, the probability of prediction errors will increase. For these deficiencies, this article will also continue to improve and perfect our

model. To address the above shortcomings, future improvements will focus on three aspects: At the data level, multi-source fusion will be used to expand samples, adaptive cleaning algorithms will be introduced to handle outliers, and small-sample technologies such as transfer learning will be integrated to reduce reliance on massive data. At the model level, XGBoost will be lightweighted through feature selection, and federated learning will be combined to enable distributed computing, reducing resource consumption. For robustness, noise-resistant loss functions will be adopted, and a multi-model collaborative mechanism will be built to enhance tolerance to outliers and prediction stability.

References

[1] Wang L. Construction of Evaluation Model of Tennis Skills and Tactic Level and Application of Grey Relational Algorithm[J]. Journal of Sensors, 2022, 2022(1): 9446175.

[2] Niu W. Kinematic Analysis of Excellent Female Tennis Players' Serving Technique[C]. //2024 2nd International Conference on Algorithm, Image Processing and Machine Vision (AIPMV). IEEE, 2024: 310-313.

[3] Qingxian S. Research on the Training of Serving and Receiving Techniques for Table Tennis Players[J]. Frontiers in Sport Research, 2025, 7(2).

[4] Mlakar M, Kovalchik S A. Analysing time pressure in professional tennis[J]. Journal of Sports Analytics, 2020, 6(2): 147-154.

[5] Beckmann J, Fimpel L, Wergin V V. Preventing a loss of accuracy of the tennis serve under pressure[J]. Plos one, 2021, 16(7): e0255060.

[6] Tang K, Zheng Q, Jiang Z, et al. Tennis Enhancing Tennis Match Strategies through Momentum Change Analysis and Prediction Models[C]//2024 International Conference on Interactive Intelligent Systems and Techniques (IIST). IEEE, 2024: 187-195.

[7] Crespo M, Martínez-Gallego R, Filipcic A. Determining the tactical and technical level of competitive tennis players using a competency model: a systematic review[J]. Frontiers in Sports and Active Living, 2024, 6: 1406846.

[8] Duan C, Shu Z, Zhang J, et al. Real-Time Prediction for Athletes' Psychological States Using BERT-XGBoost: Enhancing Human-Computer Interaction[J]. arxiv preprint arxiv:2412.05816, 2024.

[9] Vikram Y, Senthilvel P G. Efficient framework for sports result prediction using random forest and compare the accuracy with XGBoost[C]//AIP Conference Proceedings. AIP Publishing LLC, 2025, 3252(1): 020031.

[10] Zhong M, Liu Z, Liu P, et al. Searching for the Effects of Momentum in Tennis and its Applications[J]. Procedia Computer Science, 2024, 242: 192-199.