A Data-Driven Approach to Tennis Player Performance Prediction and Evaluation

Jiabao Zhang *

School of Business, Shandong Normal University, Jinan, China, 250358

* Corresponding Author Email: 19861818426@163.com

Abstract. This study presents a comprehensive framework for evaluating and predicting tennis player performance using historical match data. First, to reduce the data dimension, principal component analysis is used to obtain seven principal components that capture more than 85% of the variance of the data. These principal components serve as inputs to a machine-learning model that is used to predict the outcome of tennis scores. Various machine learning algorithms are compared using ROC curves and evaluation metrics. The gradient gradient-enhanced decision tree (GBDT) model performs best. To evaluate athletes' performance comprehensively, an evaluation system based on the derived principal component was constructed by using the entropy weight method and the TOPSIS model. This comprehensive approach achieved 89.04% accuracy in predicting the outcome of tennis matches, demonstrating its effectiveness in capturing match dynamics. In addition, the study introduces a visualization method that combines macro-level score analysis with micro-level performance assessment. This allows a detailed understanding of how players perform at different stages of the game, enabling coaches and players to make informed decisions and optimize strategies.

Keywords: Machine Learning, Gradient Boosted Decision Tree, Entropy Weight Method, Principal Components Analysis.

1. Introduction

With the development of big data and artificial intelligence technology, data analysis has become an important force to promote progress in various fields. Sports data analysis is playing an increasingly important role in athlete training, game strategy formulation, and performance evaluation. As a highly competitive and highly technical sport, tennis performance is affected by a variety of factors, such as speed of serve, depth of return at the net, etc, Therefore, how to accurately predict and evaluate tennis performance has become an important research topic in the field of sports data analysis.

To address the complexity inherent in predicting and evaluating tennis player performance, we chose a gradient-raised decision tree algorithm. Gradient lift trees are known for their superior predictive power and proficiency in handling high-dimensional nonlinear data, proving their superior ability in various fields. The gradient lifting tree algorithm has been widely used in many fields and has achieved remarkable results. In the field of disaster prevention, Song et al. (2022) [1] used a gradient lifting tree algorithm to predict landslide risk, demonstrating its effectiveness in prediction analysis. In the field of physics, Guoqing et al. (2023) [2] applied a gradient lifting tree algorithm to study and predict the tribological properties of MoS2-coated copper. In addition, GBRT is also widely used in the field of materials. For example, Wang et al. (2023) [3] use the gradient lifting tree algorithm to predict the bond strength of concrete superposed surface. In the context of tennis performance prediction and evaluation, GBDT also shows great prospects. For example, Yang (2024) [4] successfully applied a gradient lifting tree to predict individual scoring results in tennis matches and used key indicators such as serve speed and return depth to conduct a granular analysis of match dynamics.

Drawing on these successful precedents, first of all, the data set is preprocessed comprehensively to ensure data quality. Then, the correlation between variables is analyzed, and the dimensionality is reduced by principal component analysis. These steps lay the foundation for machine learning models that are trained and evaluated to determine which algorithms are best suited to predict the outcome

of tennis matches. Finally, this study develops a comprehensive assessment and visualization framework that integrates EWM and TOPSIS [5] to provide a nuanced understanding of tennis match performance.

2. Data Preprocessing and Initial Analysis

2.1. Processing of missing data values

The data of this study comes from: www.comap.com. In the data set provided by the website, null values such as serve_width, serve_depth, return_depth, speed_mph, and rally_count exist marked with "NA". We first deal with the missing values in the table, which are mainly divided into the following two processing methods.

First, for continuous data such as speed_mph, consider filling with the mean value. When the data is normally distributed, the mean value is the central trend of the distribution and describes the overall position of the data set. To ensure that the mean value can reflect this feature, the data need to be tested for normality. After examination, it can be intuitively found from the histogram that this is a left-biased distribution, which is changed into an approximate normal distribution by BOX-COX transformation. The mean is then used to fill in the missing values.

Second, for classified data such as serve_width, we choose a class with a higher proportion of classified data to fill the missing value, that is, fill the missing value with the mode.

2.2. Evaluation of the processing of measurement type data

Normalize two very small metrics, double faults, and unforced errors, and transform them into very large consistency metrics. To minimize the indicator xi(i = 1,2) a translation transformation is applied:x' = Mi−xi, where Mi represents the most possible value, the value that the indicator mi can take.

2.3. Normalization of data processing

So that the relationship before the data is not affected by the dimension size, for n samples xij (i = 1,2,...) Min-Max normalization is performed on the 4 indicators of... 11. j =1, 2,... n). Let Mi = max {xi1, xi2,... , xin}, mi = min {xi1, xi2,... ,xin}. The normalized data value is:

$$x_{ij}^{*} = \frac{x_{ij} - m_i}{M_i - m_i} \in [0, 1] \tag{1}$$

2.4. Person Correlation Coefficient Analysis

To preliminarily explore the strength and direction of the linear relationship between various factors, especially between point_victor, so that key indicators can be more accurately selected in subsequent data processing and analysis, the Pearson correlation coefficient is chosen in this study. The Pearson correlation coefficient provides a quantitative indicator to describe the strength of the linear relationship between two variables. A weak linear relationship does not mean that there is no correlation between two variables. The Pearson's correlation coefficient r of variables X and Y is calculated as follows when the two sets of variables are $X = (x_1, x_1, \cdots, x_n)$ and $Y = (y_1, y_2, \cdots, y_n)$ respectively:

$$r = \frac{\sum_{i-1}^{n}(x_i - \mu_X)(y_i - \mu_Y)}{\sqrt{\sum_{i-1}^{n}(x_i - \mu_X)^2}\sqrt{\sum_{i-1}^{n}(y_i - \mu_Y)^2}} \tag{2}$$

To explore the relationship between indicators, this study conducted a personal correlation analysis and made a correlation heat map, as shown in Figure 1.

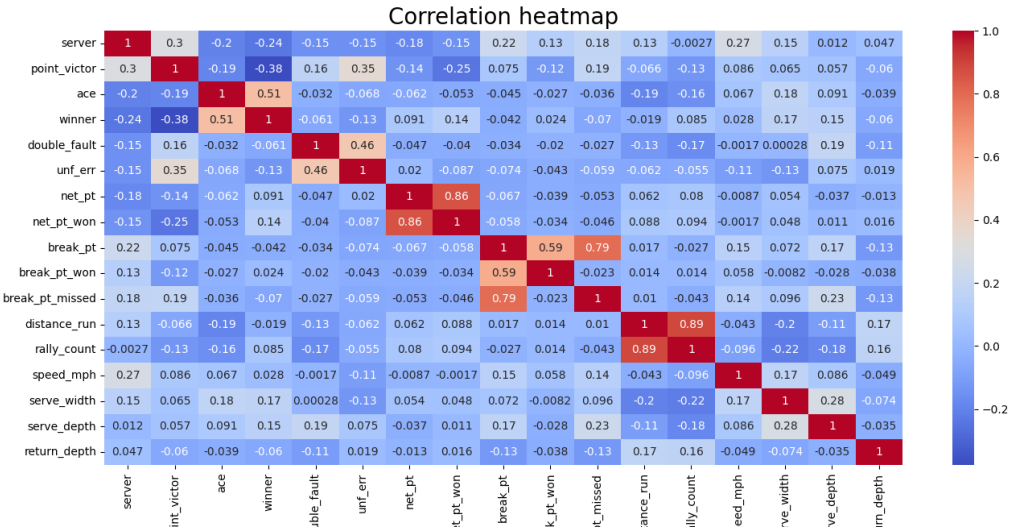


Figure 1. Correlation heat map

According to the thermal map, there is a certain degree of linear correlation between each index, especially the value of 0.3 between server and point_victor, 0.51 between ace and winner, and 0.86 between net_pt and net_pt. The value between distance_run and rally_count is 0.89.

2.5. Data feature extraction and dimensionality reduction

As can be seen from Figure 1, 16 key indicators covering multiple aspects [6] were selected in the initial stage of this study to simplify the problem, reduce the complexity of the analysis, and avoid introducing unnecessary redundant information and potential collinearity problems. Therefore, to focus on the factors with the most explanatory power, 13 indicators were selected through the previous person phase relationship analysis, and principal component analysis (PCA) was adopted for dimensionality reduction. This is a multivariate statistical method, which converts the original indicators with certain correlations into a set of unrelated comprehensive indicators through dimensionality reduction technology. The PCA gravel chart and the cumulative contribution rate chart obtained by replacing the original index are shown in Figure 2.

As can be seen from Figure 2, it is best to extract 7 principal components when the eigenvalue is around 1, which can occupy more than 85% of the data. Therefore, to reduce the complexity of the modeling process, we can reduce the data to 7 dimensions.

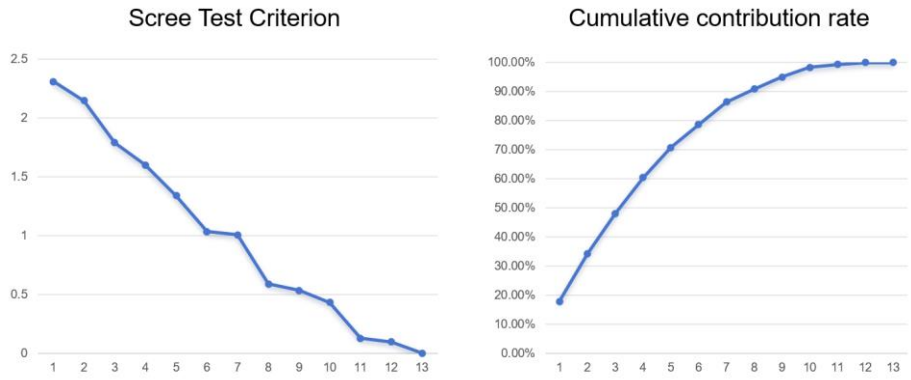


Figure 2. Gravel chart and cumulative contribution rate char

3. Comparative Analysis of Machine Learning Models for Tennis Prediction

3.1. Construction of Machine Learning Model

To better predict the score of athletes, we adopt the machine learning method, using the above characteristic data as input values, whether the player scores as output, training, training set, and test set divided into 7:3. Considering the need to take into account both athletes when using metrics to make predictions, we have eliminated the relative athlete metrics in the process, that is, we can choose one of the indicators of both.

3.2. Evaluation and analysis of results

To evaluate the generalization ability of each prediction model, the ROC curve is introduced. The vertical axis of the ROC curve is TPR, the horizontal axis is FPR, and the maximum value is 1. The size of the area under the ROC curve (AUC) determines the generalization ability of the model. The larger the AUC, the stronger the generalization ability of the model, that is, the better the prediction effect of the model. The better the explanatory ability of binary logistic regression is, the greater the generalization ability of the model is, and the better the prediction effect of the model is. After verification, the total ROC comparison curve is shown in Figure 3.

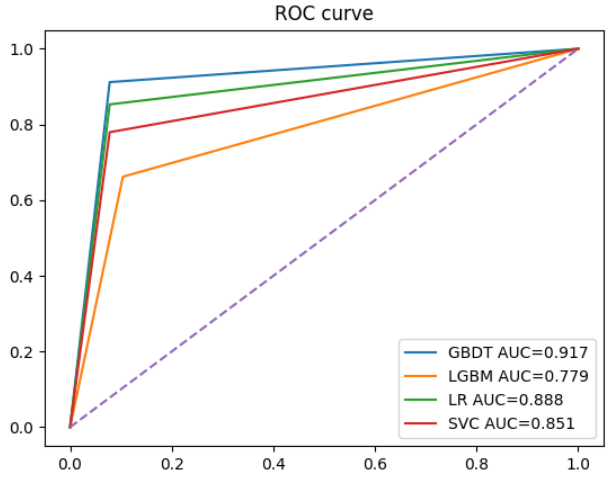


Figure 3. ROC curve

Further, a series of evaluation indicators (Accuracy, Precision, Recall, F-score, RMSE, R_2) are used to compare the performance of each algorithm, and the specific results are shown in Table 1 below:

Table1. Demonstration of the performance of each model

	Accuracy	Precision	Recall	F-score	RMSE	R^2
GBDT	0.897	0.923	0.857	0.896	0.321	1.000
LGBM	0.720	0.912	0.464	0.700	0.529	0.888
LR	0.884	0.883	0.875	0.884	0.341	0.580
SVC	0.797	0.842	0.714	0.796	0.450	0.636

In summary, GBDT performed better at prediction, so we finally decided to use the GBDT algorithm to predict the data. The concept of gradient enhancement is based on the idea of comprehensive learning enhancement, whereby a weak learner is enhanced into a strong learner by a weighted combination of multiple basic learners. On this basis, if a weak learner is generated according to the gradient direction of the loss function (LossFunction), it is called gradient enhancement. The gradient enhancement algorithm first gives a target loss function, and then iteratively selects a gradient direction basis function to approach the local minimum of the function, to achieve the optimization of the loss function.

Assuming that a regression tree has n features, each feature$S_i(i \in (1,n))$ values, the space is partitioned by exhaustively enumerating each take of each feature until the take of feature j, s, is taken such that the loss function is minimized ie:

$$\min_{js}\left[\min_{c_1} Loss(y_i, c_i) + \min_{c_2} Loss(y_2, c_2)\right] \tag{3}$$

If the input space is divided into N cells: $R_1, R_2, \cdots, R_n$ then the output value of each region is the mean of all y-values in that region:

$$C_n = ave(y_i \mid x_i \in R_n) \tag{4}$$

The final regression tree generated is:

$$f(x) = \sum_{n=1}^{N} C_n I(x \in R_n) \tag{5}$$

Based on the above analysis, this paper chooses to use the GBDT model for prediction.

4. Comprehensive Evaluation and Visualization of Athlete Performance

When constructing the comprehensive evaluation model of athletes' performance, we consulted the opinions of tennis experts and related enthusiasts and found that although there was no significant relationship between some indicators and athletes' performance, these indicators would still have an impact on athletes' status and on-court conditions. Therefore, we considered using the TOPSIS model based on entropy weight. Although the TOPSIS method is good at ranking a limited number of evaluation objects according to how close they are to the desired goal, it faces the challenge of determining the weight of the indicator. In this regard, the EWM method has effectively alleviated this difficulty to some extent. TOPSIS excels at considering a variety of factors at the same time, while the EWM method provides a more comprehensive weight determination for each influencing factor.

4.1. Evaluation system construction

As mentioned above, 13 indicators [7] were selected from the data in this study, and 7 principal components were obtained by using principal component analysis (PCA) for dimensionality reduction. Therefore, the seven principal components mentioned above are used as the criterion layer of the evaluation system, and the 13 indicators are used as the index layer of the evaluation system. The 13 indicators are shown in Table 2 below:

Table 2. Description of indicators

variables	explanation
server	the server of the point
ace	player hit an untouchable winning serve
winner	The player hit an untouchable winning shot
double_fault	The player missed both serve and lost the point
unf_err	The player made an unforced error
net_pt	The player made it to the net
net_pt_won	player won the point while at the net
break_pt	one player has an opportunity to win a game another player is serving
break_pt_won	onr player won the game other player is serving
break_pt_missed	one player missed an opportunity to win a game another player is serving
distance_run	player's distance ran during point (meters)
rally_count	number of shots during the point
speed_mph	speed of serve (miles per hour; mph)

This section is based on the extraction rate of seven principal components, seven criteria layers, and seven criteria. The advantage mainly includes server and break_pt, which is a huge advantage to simulate in these two situations. ace and winner are also used to simulate psychological incentives in these two situations [8]. error mainly includes double_fault, unf_err, in these two cases there are mainly mistakes, net victory mainly includes net_pt, net_pt_won, in these two cases, is mainly the advantage of the net [9]. opportunity mainly includes break_pt and break_pt_won. In these two cases, the main concern is whether the opportunity is grasped. The activity intensity is distance_run, rally_count, In these two cases, it mainly indicates the athlete's sports activity, and the power is mainly speed_mph, which mainly shows the athlete's hitting speed and power [10].

4.2. Entropy weight method to determine the weight

To determine the relative importance of each index to the total score of athletes, we use the entropy weight method to calculate the weight of each index. The entropy weight method is an objective valuation method, which is explained away from the basic principles of information theory. Information is a measure of the degree of order in the system, and entropy is a measure of the degree of disorder in the system. The specific steps for calculating the weight of each index by using the entropy weight method are as follows:

(1) Suppose all the given indicators are$X_1, X_2, \ldots, X_j$, where $X_i = \{X_1, X_2, \ldots, X_n\}$. Assuming that the standardized value of each indicator data is $Y_1, Y_2, \ldots, Y_j$, it can be obtained by standardizing the original matrix $X = (x_{ij})_{(m\times n)}$:

$$\boldsymbol{Y_{ij}} = \begin{cases} \frac{\boldsymbol{x_{ij} - min(x_{ij})}}{\boldsymbol{max(x_{ij}) - min(x_{ij})}}, \boldsymbol{x_{ij}\, is\, +} \\ \frac{\boldsymbol{max(x_{ij}) - x_{ij}}}{\boldsymbol{max(x_{ij}) - min(x_{ij})}}, \boldsymbol{x_{ij}\, is\, -} \end{cases} \tag{6}$$

(2) Calculation of the weight of the value of the ith point of the jth indicator about the indicator p_{ij} :

$$\boldsymbol{P_{ij}} = \frac{\boldsymbol{Y_{ij}}}{\sum_{i=1}^{m} \boldsymbol{Y_{ij}}} \tag{7}$$

(3) Calculate the entropy value of the jth indicator E_j when $P_{ij} = 0$, $P_{ij} \ln P_{ij} = 0$:

$$\boldsymbol{E_j} = -\frac{\boldsymbol{1}}{\boldsymbol{ln\, m}} \sum_{i=1}^{m} \boldsymbol{P_{ij}\, ln\, P_{ij}}, (\boldsymbol{j = 1, 2, \cdots, n}) \tag{8}$$

(4) Calculate the coefficient of variation G_j for the jth indicator:

$$\boldsymbol{G_j = 1 - E_j} \tag{9}$$

(5) According to the formula of information entropy, the information entropy of each indicator is calculated as $E_1, E_2, \ldots, E_j$, and the weight of the jth indicator, W_j, is calculated by information entropy:

$$\boldsymbol{W_j} = \frac{\boldsymbol{G_j}}{\sum_{j=1}^{n} \boldsymbol{G_j}}, (\boldsymbol{j = 1, 2, \cdots, n}) \tag{10}$$

4.3. Entropy weight method combined with topics scoring system

Generate weighted data according to the above, determine the evaluation indicators of players' scores, and ensure that all evaluation indicators are in a positive trend, and carry out TOPSIS analysis of weighted data. TOPSIS method is used to process the same trend of the original data, and normalization is carried out to eliminate the influence of inter-scale indicators. The ranking results make full use of the original data information and quantitatively reflect the pros and cons of the comprehensive evaluation value of each index, which is more intuitive and reliable.

(1) A weighted standardization matrix can be obtained from the standardization matrix and the weights of the indicators:

$$\boldsymbol{Z_{ij} = W_j \times Y_{ij}} \tag{11}$$

(2) Calculate positive and negative ideal solutions:

$$\boldsymbol{I_j^+ = max\left(z_{1j}, z_{2j}, \cdots, z_{mj}\right), I_j^- = minz_{1j}, z_{2j}, \cdots, z_{mj}} \tag{12}$$

(3) Euclidean distance of each indicator from positive and negative ideal solutions:

$$\boldsymbol{D_i^+ = \sqrt{\sum_{j=1}^{n} \left(i_j^+ - z_{ij}\right)^2}, D_i^- = \sqrt{\sum_{j=1}^{n} \left(i_j^- - z_{ij}\right)^2}} \tag{13}$$

(4) Calculate the composite assessment value:

$$\boldsymbol{C_i = \frac{D_i^-}{D_i^+ + D_i^-}} \tag{14}$$

After the score is obtained, the score is assigned to the range [1,10]. From the score, we can infer two athletes who perform better and are more likely to get points. The evaluation visualization is shown in Figure 4. Therefore, if we think that the side with the higher performance rating will win the set, by comparing the data of the scorer in one of the 301 rounds, we find that the accuracy of the judgment is 89.04%, and some of The Times that the judgment is wrong is because the server fails to score when the two sides have no obvious advantages and mistakes.

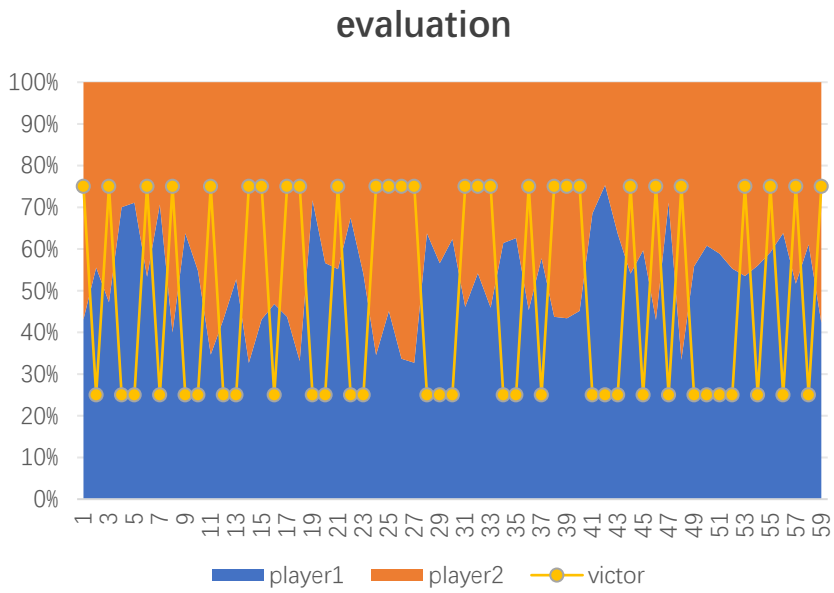


Figure 4. Evaluation visualization

4.4. Visualization of Athlete Performance

In providing model-based visualization to depict the game flow, we first understand the macro situation by analyzing the specific score curve between the two athletes. As can be seen from the total score in Figure 5, the results of the two athletes are very close, and the relationship between them is not intuitive and obvious.

Using the win field change chart, it can be seen that the players in the game still have high performance when they are at the point, but the win field change chart still cannot account for the performance of the players at every point on the court, that is, the player's score situation, the difference between the score situation and the gain field change can only be used to judge the macro field situation, we need to proceed to one Step microscopic analysis.

Therefore, This paper changes the score graph to the score difference graph to strengthen the relevant features. After changing the performance difference chart, we can see the change in the situation, but still cannot identify the athlete's performance. Since the athlete's performance is affected by many factors, even if they are temporarily behind in the score, they may have a relatively bright result in the later competition, which shows that the difference in scores alone cannot describe the athlete's performance. On the other hand, micro-analysis requires the use of a comprehensive evaluation model to score the performance of athletes, that is, through the combination of the TOPSIS model based on entropy weight, it can capture which athlete performs better and how much better at different scoring points at specific times in the competition [11].

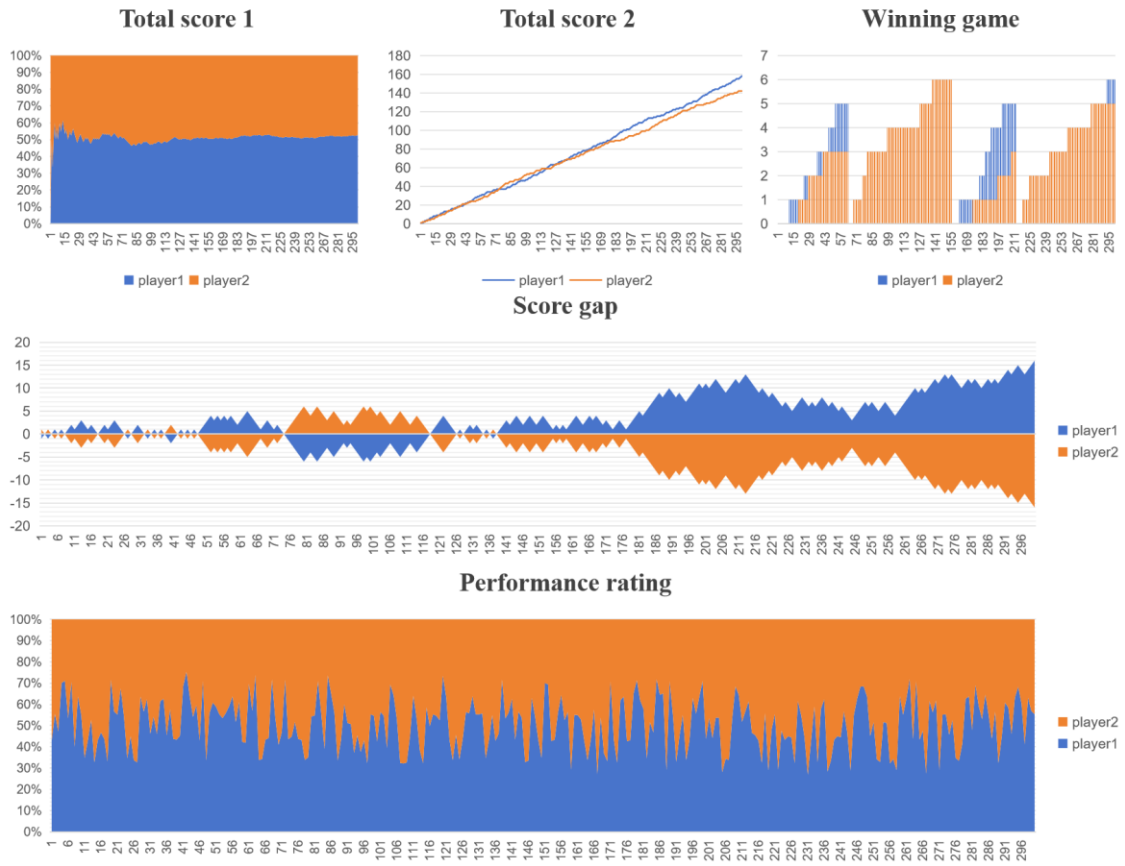


Figure 5. Visualization of athlete performance

As shown in Figure 5, the combination of on-site macro analysis and micro judgment of athlete performance in this paper allows us to capture the flow of the game and provide a visual process.

5. Conclusions

Based on data from previous competitions, this paper constructs mathematical models to evaluate and predict athletes' fragrances during competition and visualizes how various activities in competition create or change performance evaluations. In this study, Pearson correlation analysis was used to explore which indicators were correlated with point_victor and analyze their correlation, to further study the relationship between indicators and point_victor. PCA dimensional degradation process was used for principal factor selection and various machine algorithms were used for verification. The results are compared with other machine algorithms to find that the GBDT model performs best. Build a GBDT model to evaluate and predict the performance of athletes in real-time. The entropy weight method is used to quantify the momentum, calculate the weighted average, and determine the importance of each index in the evaluation. By comparing the theoretical calculation results of the topsis model with the actual win/loss situation, the accuracy rate reaches 89.04%. Next,

we consider combining on-the-spot macro analysis with micro judgments of player performance, which can capture the flow of the game and provide a visualization of the process.

This paper provides a research idea and framework for the field of motion pattern analysis, which combines the two advanced algorithms of gradient lifting decision Tree (GBDT) and entropy weight method technology priority Ranking and Ideal solution Similarity Ranking (EWM-TOPSIS), and proves the feasibility of the method. Through a series of well-designed simulation tests and empirical analysis, this study not only successfully demonstrated the excellent ability of the proposed model to capture subtle differences in sports patterns, accurately predicting and evaluating athletes' performance, but also further revealed its excellent generalization ability and robustness, which not only enriched the theory and method system in the field of sports pattern analysis. It provides strong technical support for practical application and has far-reaching significance for promoting the intelligent and scientific development of sports.

References

[1] Song J, JinYuan L, Sai Z, et al. Landslide risk prediction by using GBRT algorithm: Application of artificial intelligence in disaster prevention of energy mining [J]. Process Safety and Environmental Protection, 2022, 166 384-392.

[2] Guoqing W, Yuling R, Hongxing W, et al. Tribological performance study and prediction of copper coated by MoS2 based on GBRT method [J]. Tribology International, 2023, 179

[3] Wang Jianmin, Ye Yurong, Rao Chaomin, et al. Prediction and analysis of bond strength of concrete superposition surface based on GBDT algorithm [J]. Journal of Building Materials, 2023, 26 (02): 150-155+171

[4] Yang J. Research on Tennis Match Strategies Based on LightGBM and BP Neural Network Models [J]. Academic Journal of Computing & Information Science, 2024, 7 (4): 103-109

[5] Li Z, Yin Y. A Quantitative Assessment Model for Momentum in Tennis Based on Exponential Decay and EWM-TOPSIS Method [J]. Academic Journal of Computing & Information Science, 2024, 7 (5): 136-142

[6] Zhou Xin. Construction and Analysis of Winning Factor Model of Women's Tennis by Logistic Regression Model -- Taking Australian Open as an example [J]. Sports science and technology literature bulletin, 2022, 30 (03): 65-67.

[7] Ouyang Gaohuang, LI Pinfang, Chen Jiancai. Analysis of Winning Factors of Different sets of Men's Singles Tennis Players Based on Logistic Regression Model [J]. Sports science and technology literature bulletin, 2022, 30 (6): 13-16.

[8] 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.

[9] Almarashi M A, Daniyal M, Jamal F. A novel comparative study of NNAR approach with linear stochastic time series models in predicting tennis player's performance. [J]. BMC sports science, medicine & rehabilitation, 2024, 16 (1): 28-28.

[10] Li L. Analysis of Winning Factors in Tennis Match based on Fuzzy Judgment [J]. International Core Journal of Engineering, 2024, 10 (5): 445-450

[11] Yang Yuqiang, He Daoyuan. Tennis tournament in the data analysis and application research [J]. Journal of contemporary sports science and technology, 2020, 10 (29) : 222-224