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

A Study on Predicting Tennis Player's Competition Direction Based on Logistic Regression and LSTM Model

Yan Ding, Chi Zhang *

Jinan University-University of Birmingham Joint Institute, Jinan University, Guangzhou, China, 511436

* Corresponding author: daju1802@gmail.com

Abstract. A tennis player's performance during a match affects the outcome of the match. The study of the issue can provide advice to tennis players on how to prepare for the tournament, and to coaches and their teams on how to develop training programs that will help athletes get better results in the tournament. Initially, according to the entropy weight method to assess the player's performance on the field. Two factor layers were used to finally get a score measuring the performance at each point. Next, based on the score and making use of conditional probabilities in logistic regression models to predict the winning rate in each point and then in each game. The result shows high accuracy in the logistic regression in both points and games. Ultimately, by assuming the existence of average effect of the players' performance in all points of one game, deep learning with long short-term memory is used to predict game fluctuations and identify factors that contribute to them.

Keywords: Entropy Weight; Logistic Regression; Long Short-term Memory; Performance Measure.

1. Introduction

Tennis has become increasingly professional and competitive, taking its place among the world's professional sports events in terms of prize money, influence, spectacle, and commercial operations [1] A player's clinical performance in a match affects the outcome of the match to a certain extent. Technical actions such as player errors while the game seem to be a big difference between winning and losing teams [2]. Analyzing the important match data of professional tennis players to find the winning factors, can help us to further analyze the winning rules of tennis matches and provide a reference basis and development direction for the competitive tennis career [3-4]. In this paper, the entropy weighting method is used to evaluate the performance of the player at a certain point in the competition, followed by logistic regression to predict the winning percentage, and finally, the long short-term memory method is used to obtain the factors affecting the player's performance.

2. Entropy weight method of analyzing athletic performance

2.1. Structure of the evaluation system

The entropy-weighted evaluation model is a method for evaluating objectives based on entropy weights [5], the indicators for evaluating the performance of the athlete at a given moment of the game are selected according to this method. Entropy weight evaluation assumes that the lower the degree of variability of an evaluation indicator, the less information it reflects, and the lower the corresponding entropy weight it has.

To avoid the risk of chance, data with a small probability of occurrence, affecting the objectivity of the evaluation and the validity of the predictions, we have categorized the factors affecting an athlete's on-field performance into the following categories as shown in Fig 1, based on the actual situation, and conduct the following discussion to select the evaluation indicators to measure each factor.

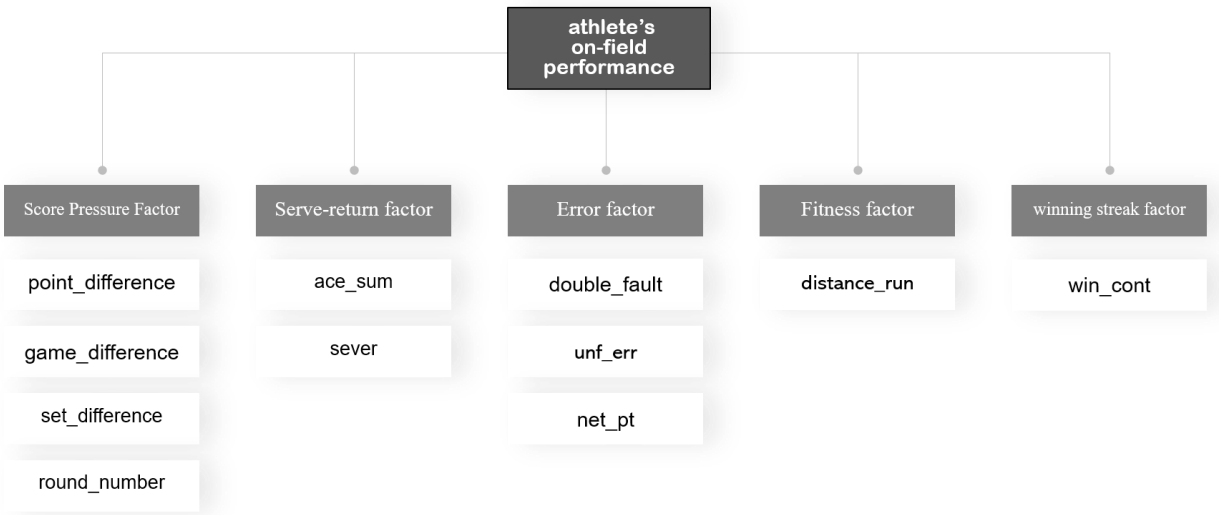


Figure 1. Factors and evaluation indicators of athletes' on-field performance

On the one hand, Observation from the data shows that players generally have a higher probability of winning this point on serve. On the other hand, If the returning player succeeds in taking the breakpoint, it will greatly increase the returning player's self-confidence, which will have a positive impact on the state of play in the following game. Similarly, Winning streaks can also boost morale and motivate the body to get into the game quickly. Moreover, the difference in scores to a certain extent indicates the gaps in the athletes' strengths, and conditions, also when a point is awarded to determine a winning point or as a game gets closer to the end of the season, most athletes' mental pressures increase [6].

from the other side, the error factors can be classified into the following three categories: the player missed both serves and lost the point, the player made an unforced error, and the player did not win the point while at the net. Errors may cause players to have more negative emotions and create an imbalance in their mentality, leading to an incapacity to concentrate physically and mentally. The running of players during the point will lead to fatigue of players to a certain extent, and the distance run by players will reflect the level of excitement of players to some degree [7].

2.2. The analysis of athletic performance

After deleting some outliers, two matches(1301 and 1315) were chosen as the test data. Applying the entropy weight method, the weighting of each evaluation indicator in measuring an athlete's on-field performance is obtained. The results are shown in Table 1.

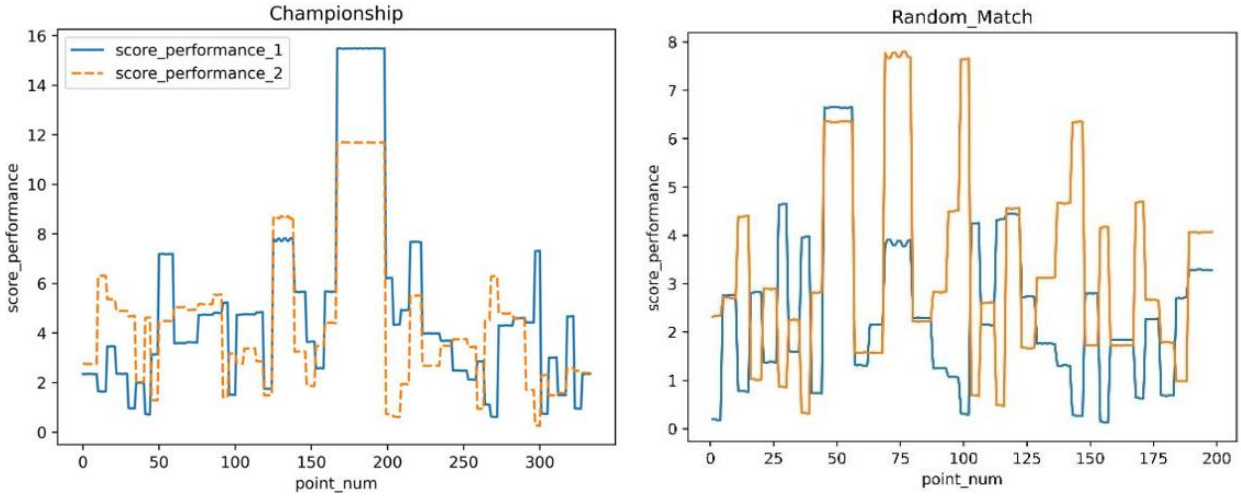
Table 1. Weighting of each evaluation indicator in the player's on-field performance

Factor	Weight(%)
Serve-return Factor	47.30%
Winning Streak Factor	44.82%
Score Factor	4.91%
Error Factor	1.64%
Fitness Factor	1.33%

Thus, from this table, get the expression for measuring an athlete's on-field performance (i.e. Sp'), then min-max standardize Sp' to get Sp, so that the range of Sp is controlled within $[0, 1]$ for subsequent analysis.

$$Sp' = 0.4730Serv + 0.4482Winn + 0.0491Score + 0.0164Erro + 0.0133Fitn$$
$$Sp \xleftarrow{\min-\max\ standarlization} Sp' \tag{1}$$

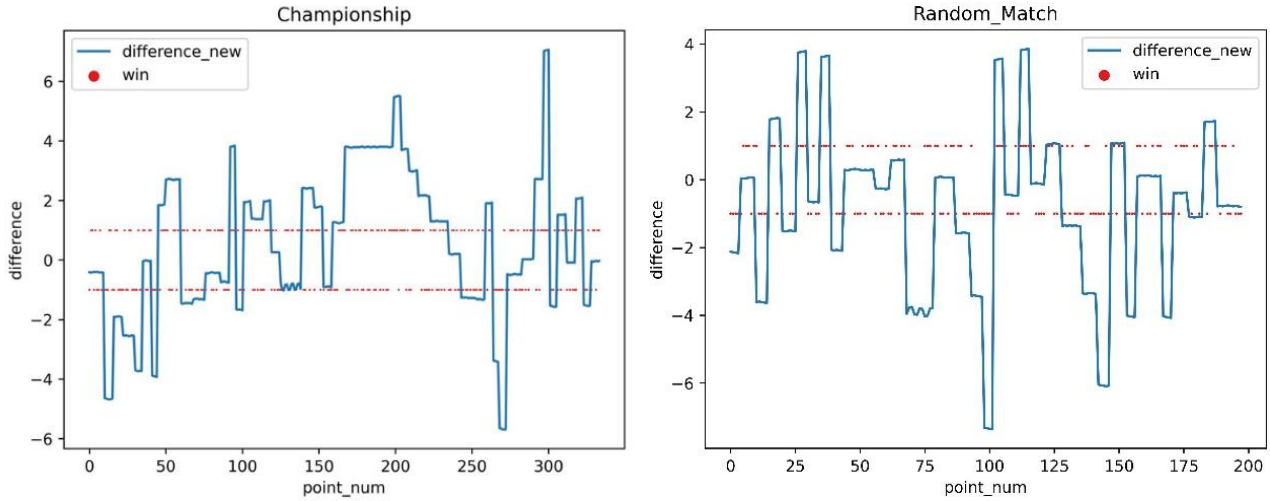
Next, select the championship match and one other match (third round, fifteenth match) which is not included in the test data to test the generalization ability of evaluation model. Setting the y-axis as the performance score Sp' and the x-axis as point_num to visualize and analyze to show the flow of the game. The Fig 2.a of the Championship match shows the change in the performance score of two players, which is largely the same as the one described above. The Fig 2.b also shows the relationship of the corresponding data for this match.



(a) Visual analytics for championship match (b) Visual analytics of the other match

Figure 2. Visual analytics of these two matches

To visualize this relationship more intuitively, denote the difference in performance scores between the two players in a match as *SDp*, using one of the players as the benchmark, and specify that an actual take of a point is denoted as 1, while a loss is denoted as -1, and make a scatter plot with parallel x-axis as shown in Fig 3.



(a) More intuitive visual analytics in a championship match

(b) More intuitive visual analytics in the other random watch

Figure 3. More intuitive visual analytics in two matches

Observation of the above two Figs shows that shifts in the difference of performance score often occur at the beginning of a game or set, suggesting that the player with the highest performance score is likely to play a crushing game to win the match.

3. logistic regression for predicting the winning rate

Logistic Regression is a statistical model for binary dependent variables. Binary logistic regression can be used to investigate the significance of the effect of different factors on the object of analysis [9]. The winner of the points was chosen to be dependent variable Y, using P1 as the analysis object, if P1 won the point in the data, Y=1, if P2 won the point, Y=0. Using *SDp* as the predictor X, input a given value of *SDp*, the Logistic Regression can give the probability:

$$p = P(P1 \ win \ the \ point \mid SDp) \tag{2}$$

In Logistic Regression: The log-odds is given by

$$\log \frac{p_2}{1-p_2} = \beta + \alpha SDp \tag{3}$$

Solving for p given $p = S(\beta + \alpha SDp)$, where S is the Sigmoid function

$$S(\zeta) = \frac{\exp(\zeta)}{1+\exp(\zeta)} \tag{4}$$

By the MLE (Maximum Likelihood Estimation), Suppose that $\mathcal{D}$ is the observed data, the MLE estimates are given by

$$\hat{\beta}, \hat{\alpha} = arg \ \max_{\beta,\alpha} \{l(\beta, \alpha; \mathcal{D})\} = arg \ \min_{\beta,\alpha} \{-l(\beta, \alpha; \mathcal{D})\} \tag{5}$$

Function $-l(\beta, \alpha; \mathcal{D})$ is the negative log-likelihood function.

The gradient descent method was used to solve the minimization problem and get the following result.

After removing some extremely high or low Sp of both players, 90% of the data were chosen as the training set and validated the results on the testing set, the specific results obtained from binary classification logistic regression are shown in the Table 2.

Table 2. Result of Logistic Regression

Experimental group = 1.0	Regression coefficients	standard error	P	Accuracy	AUC
constant	-4.046	0.124	0.000***		
SDp	7.476	0.223	0.000***	0.692	0.753
constant	16.84	1.109	0.000***	0.914	0.97
SDgm	-31.275	2.053	0.000***		
Dependent variable: point_victor					
Note: ***, **, and * represent the significance levels of 1%, 5%, and 10%, respectively					

From Table 2, both p-values are less than a pre-set significance level (usually 0.05 or 0.01) in logistic regression means that the coefficient of the independent variable is statistically significantly not equal to zero, i.e., the effect of the independent variable on the dependent variable is statistically significant.

Table 2 also indicates that the logistic regression results in points and games have high accuracy. The accuracy of predicting the winning rate of games is higher, which confirms to some extent the

reliability of the player performance evaluation model established in the previous section. Next, the hypothesis that the average effect of the winning rate per game (Gw) on Pw will be constructed with the result.

4. Long Short-Term Memory for catching the factors affecting competitions

4.1. Assumption and its verification

Assuming the average probability of a player winning points in each game can be represented by *Gw*, however, in actuality, events that affect the trend of the game can occur in a game that can disturb the true probability of a player scoring at a specific point (*Pw*). Define this relationship in the following equation:

$$\nu_i = Gw_j - Pw_{i,} \ for\ all\,(i,j)\ \ where\ point_i \in game_j \tag{6}$$

Where, v_i denotes the residual error, which is affected by some factors in the process of a match, of the true probability Pw_i and average probability Gw_j.

To verify the assumption, the two variables of *match 1301*(test data) were depicted in Fig 4.

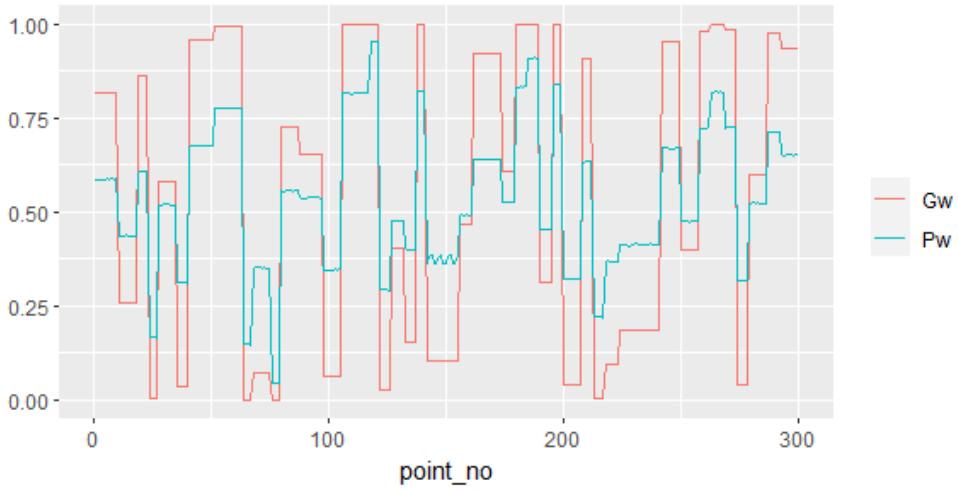


Figure 4. Visualisation of trends of Gw and Pw in 1301 match

According to Fig 4, the similar trend shared by the two variables Pw_i and Gw_j, can be observed. Thus, to catch which factor could significantly affect v_i and then give pricise pridiciton on the winning rate of a specific point, The Long Short-Term Memory was used with sensitive analysis of factors.

4.2. Long Short-Term Memory on v_i

AThe LSTM model (Long Short-Term Memory) is a recurrent neural network (RNN) model for deep learning. It effectively avoids the problem of gradient vanishing or explosion problems in traditional RNNs when dealing with long-term dependencies by introducing internal memory units called "cell states"[10].

Next a maximum of 500 iterations and a learning rate of 0.005 were chosen, using 90% of the samples as the training set, to carry out LSTM deep learning. The root mean square error (RMSE) is the evaluation metric for the prediction effectiveness of the LSTM, which is formulated as follows.

$$RMSE = \left(\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i) \right)^{\frac{1}{2}} \tag{7}$$

The higher the accuracy of the prediction model, the smaller the RMSE. By the training result after the 150th iteration, the RMSE gradually converges to 0.08, and the indicating model is more accurate. The following Fig 5 shows the visualization of the prediction we obtained by LSTM prediction.

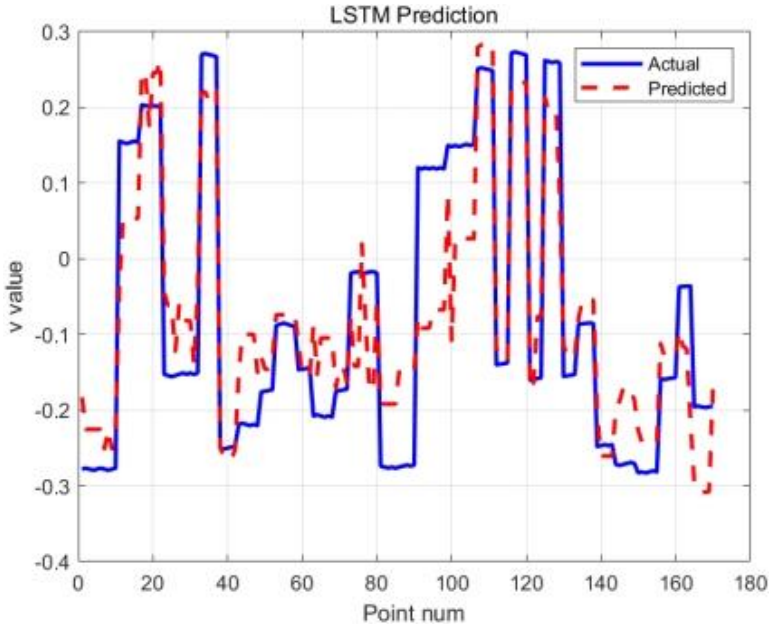


Figure 5. the visualization of the prediction from obtained LSTM prediction

In the Fig 5, the experimental results show that the LSTM model has good prediction effect, and also shows good prediction accuracy after using the data in match1407 for experiments.

4.3. Sensitivity analysis on factors

By varying the original data by 0.05 for the perturbation, the difference between the predictions before and after the perturbation is used as the measure of the influence of the selected factors on the game flow, and the factors sensitive to value changes in Table 3 are also used as the main factors that can affect the game flow.

Table 3. Selected factors for LSTM prediction

Index	Name	Index	Name
1	SDp	5	game_start
2	SDgm	6	error_factor
3	SDsm	7	fitness_factor
4	server	8	winning_streak

Fig 6 shows a visualization of this sensitivity.

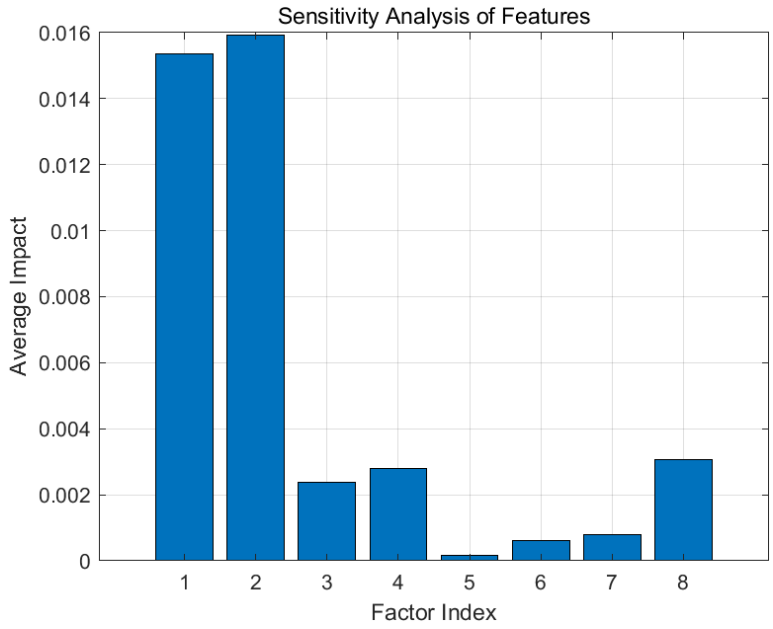


Figure 6. Visualization of sensitivity analysis.

Based on the above Fig 6, the main factors affecting game flow can be categorized in Table 4.

Table 4. The main factors

Index	Name	Index	Name
1	SDp	4	server
2	SDgm	8	winning_streak
3	SDsm		

Although serve was weighted twice in the EW model and did not have a significant impact on the player's performance, it showed a significant impact on game flow in the sensitivity prediction of this model. Combined with the hypothesis test made in the EW model, we cannot ignore the impact of player serving on changing game flow.

5. Conclusion

This paper focuses on whether a player's on-field performance is related to whether the player wins or loses the game. By utilizing the method of entropy weighting to evaluate the target, the EW model is established by selecting evaluation indicators to evaluate the player's on-field performance in a game, so that the player's performance score and the player's success have a certain linear relationship. Then, the winning percentage of the match was obtained by logistic regression and further confirmed the reliability of the players' match performance model. Finally, a long and short-term memory model was used to capture the factors affecting a player's performance, which included the psychological stress caused by the player's score difference at each point, game, and match, the first serve, and the winning streak, and it can be used to predict the direction of a player's game.

References

[1] He Yang, Xia Zhengbing, Ou Yueshan. Competition pattern and development trend of singles in the world's four major open tennis tournaments [J]. Sports Culture Magazine. 2013 (03): 80 - 82.

[2] Ofoghi Bahadorreza, Chenaghlou Milad, Mooney Mitchell, Dwyer Dan B, Bruce Lyndell. Team technical performance characteristics and their association with match outcome in elite netball [J]. International Journal of Performance Analysis in Sport. Volume 21, Issue 5. 2021. PP 700 - 712.

[3] Zhao Yue, Cui Yixiong, Liu Hongyou, et al. Analysis of winning factors of professional male tennis players based on their performance in Grand Slam tournaments [J]. Journal of Beijing Sport University. 2022, 45 (04): 78 - 90.

[4] Ma Jiacai, Chen Qingguo, Liang Cairong, et al. Model construction and research analysis of match-winning factors of professional tennis players [J]. Bulletin of Sports Science and Technology Literature, 2023, 31 (04): 66 - 68+118.

[5] Gan Langxiong, Zhang Huaizhi, Lu Talent, et al. Safety factors of water transportation based on the entropy weight method [J]. China Navigation, 2021, 44 (2): 53 - 58.

[6] Lv Sheng, Yu Chengcheng, Huang Puquan. Analysis and research on the mentality imbalance problem of badminton players when they are behind in score [J]. Sports Goods and Science and Technology, 2019 (11): 172 - 173.

[7] Brooks Edward R, Benson Amanda C, Fox Aaron S, Bruce Lyndell M. Movement intensity demands between training activities and competition for elite female netballers [J]. PloS one. Volume 16, Issue 4. 2021. PP e0249679 - e0249679.

[8] Zhu Lulu, Cui Yulong. Evaluation of landslide susceptibility in Liangshan Prefecture based on logistic regression model [J]. Shanxi Construction, 2024, 50 (05): 71 - 73+97.

[9] Li Syungha, Cai Xiaobo. Binary logistic regression analysis of sports participation of residents in the Yangtze River Delta [J]. Contemporary Sports Science and Technology, 2021, 11 (36): 98 - 102.

[10] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steinbrink and J. Schmid Huber, "LSTM: A Search Space Odyssey," in IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 10, pp. 2222 - 2232, Oct. 2017.