Revealing the Close Connection Between Tennis Match Momentum and Victory Through Machine Learning Methods

Guolin Yu [1, *], Xinyu Wang [2], Bohan Dong [1]

[1] Department of Electronic Science and Technology, Harbin Institute of Technology, Weihai, China, 264200

[2] Department of Vehicle Engineering, Harbin Institute of Technology, Weihai, China, 264200

* Corresponding Author Email: 2315641817@qq.com

Abstract. In competitive sports, especially in tennis matches, momentum is considered a key factor influencing the outcome. However, there is widespread controversy surrounding the definition of momentum, its quantification methods, and its impact on match results. We compared different models (LightGBM, SVM, XGBoost, MLP, logistic regression) using five-fold cross-validation to predict changes in momentum in tennis matches based on a given dataset. The results showed that LightGBM performed best with an accuracy of 77.6%. We selected LightGBM to predict the momentum score for each point of player 1 in the final match. It was found that the average momentum scores in the second, third, and fifth sets were greater than 0.5, which aligned with the actual outcomes. Furthermore, through Pearson correlation analysis and OLS validation of the relationship between momentum changes and match results, we obtained a p-value below 0.05, a Pearson correlation coefficient of 0.59, and an F-statistic of 376.6 in the OLS model, indicating a strong positive correlation between the two.

Keywords: Momentum forecast, LightGBM, Pearson correlation analysis, OLS.

1. Introduction

In the field of sports competition, particularly in sports like tennis that heavily rely on skills and strategies, momentum is widely acknowledged as one of the key factors influencing match outcomes. However, despite the widespread acceptance of momentum's impact in practice, there remains significant controversy in the academic community regarding the definition, quantification, and specific effects of momentum on athlete performance and match results.

For this study, we aim to investigate momentum changes in tennis matches using machine learning techniques and explore the close relationship between momentum shifts and match outcomes. Machine learning is a multidisciplinary field that encompasses knowledge from probability theory, statistics, approximation theory, and complex algorithms. It utilizes computers as tools to simulate human learning processes in real time and effectively enhance learning efficiency by organizing existing knowledge into structured frameworks [1]. Many researchers have previously contributed to predictive modeling through data analysis. For example, Hu et al. utilized machine learning to predict the mechanical properties of metal in additive manufacturing [2]. Tang et al. compared traditional methods with machine learning algorithms for predicting the adaptive weighting combination of rock mechanics parameters in the Mahu Depression Fengcheng Formation [3]. Wang et al. identified eight crucial genes for diagnosing and preventing gastric cancer using machine learning and established an optimal prediction model [4]. These studies demonstrate the versatility of machine learning across different fields and its potential application in tennis-related research.

The data for this study was obtained from https://www.comap.com/contests/mcm-icm. We initially used traditional logistic regression analysis to predict momentum in tennis matches. Subsequently, we compared four different machine learning algorithms (LightGBM, XGBoost, SVM, MLP) with logistic regression analysis using five-fold cross-validation and plotted ROC-AUC curves for each model. Finally, we selected the LightGBM model to predict momentum changes in tennis matches. To investigate the correlation between momentum changes and match outcomes, we

employed Pearson correlation analysis to examine the relationship between these two variables and further validated it using Ordinary Least Squares (OLS) regression.

2. Athlete Performance Prediction Research

2.1. Data Preparation

The cleanliness of the data directly affects the accuracy and robustness of the final model, as well as the resulting conclusions. To ensure the usability of the data, pre-processing is necessary before conducting data analysis. According to the rules of the game, scores can only be 0, 15, 30, 40, or AD. So, we removed outliers in the form of abnormal scores, such as scores of 1, 2, 3, etc., for player 1 and player 2 in the current game. Additionally, we observed missing values in the tennis serve speed data, denoted as NA, and addressed them accordingly.

To predict the momentum changes of athletes throughout the entire match, we need to establish a model that takes into account the serving player's winning probability. Momentum is an abstract term that is difficult to measure but reflects the advantage demonstrated by a player in winning a game. Therefore, we define the dependent variable label as whether player 1 scores at the current point. We assign the value of 1 when player 1 scores and 0 when player 2 scores. We define "momentum" as the probability of player 1 scoring at the current point, using point changes to visualize momentum changes.

Next, we analyzed each column of the dataset and extracted 16 factors that may affect the prediction of whether a player scores at the current point. To avoid the dominance of a single variable due to its larger magnitude, we standardized the data using the following formula:

$$x' = \frac{x - x_{mean}}{\sigma} \tag{1}$$

Where x represents any data point, x_{mean} represents the mean value of the column, and σ represents the standard deviation of the column. The standardized data satisfies a normal distribution with a mean of 0 and variance of 1. After standardization, the training data is provided in the appendix.

2.2. Logistic Model

Based on observations, the dependent variable y_{label} consists of 0 and 1. Therefore, we chose the Logistic prediction model.

2.2.1 Introduction to the Logistic Model

Linear regression is a statistical analysis method used in regression analysis in mathematical statistics to determine the quantitative relationship between two or more variables. When there is only one independent variable and one dependent variable, it is called simple linear regression. As the number of independent variables increases, it is referred to as multiple linear regression [5].

Similar to linear regression, logistic regression is used to estimate the relationship between a dependent variable and one or more independent variables. However, its purpose is to predict categorical variables and continuous variables. Logistic regression is a generalized linear regression analysis model that has the advantages of simplicity, efficiency, fast computation speed, interpretable output results with probability values, the ability to handle linearly separable and non-linearly separable problems, and the use of regularization methods to prevent overfitting.

2.2.2 Model Establishment

The logistic regression model combines the independent variables through a linear combination and passes the result into the Sigmoid function, obtaining an output value that can be interpreted as the probability of the event occurring. For binary classification, if the output value is greater than or equal to 0.5, it is classified as 1; otherwise, it is classified as 0. The Sigmoid function is defined as follows:

$$\sigma = \frac{1}{1+e^{-x}} \tag{2}$$

The mathematical expression of the logistic regression model is as follows:

$$\begin{cases} \log it(p) = \dfrac{1}{1+e^{-p}} \\ \ln\left(\dfrac{p}{1-p}\right) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n \end{cases} \tag{3}$$

p is the probability of the dependent variable being 1. $\log it(p)$ is the dependent or response variable. $x_1, x_2, \cdots, x_n$ represents the independent variables, and $b_0, b_1, \cdots, b_n$ is the regression coefficient [6].

Using the standardized training data extracted in the previous section, we input it into the logistic regression model using SPSS. The model represents whether player 1 scores at the current point, with 1 indicating player 1 scoring and 0 indicating player 2 scoring. The 16 indicators that may affect the momentum prediction are used as independent variables. The predicted results are shown in Table 1.

Upon analyzing the variables in the equation, it was found that more than half of the independent variables have a significant impact on the model's prediction. Among them, "The margin number of points won in the current game", "Whether player 1 hit an untouchable winning serve", "Whether player 1 hit an untouchable winning shot", "Whether player 1 missed both serves and lost the point", "Whether player 1 made an unforced error", "The ratio of the times at the net to points won at the net", and "Player 1's distance ran during the point" have significance levels lower than 0.05, indicating a strong correlation with the dependent variable. These variables have a greater influence on whether Player 1 scores at the current point. On the other hand, "The number of games won in the current set", "The margin of set scores in the current match", and "Player 1's distance ran during the latest three points" have greater significance levels, indicating a weaker correlation with the dependent variable and a smaller impact on whether player 1 scores in the current point.

Table 1: Predicted Results

Item		Predicted Value		Correct Percentage
		0	1	
Original Value	0	271	517	34.4
	1	184	1058	85.2
Total Percentage		-	-	65.5

Table 1 shows that the logistic regression model used for predicting whether player 1 scores at the current point has a correct percentage of 34.4% for correctly predicting the dependent variable when its true value is 1, a correct percentage of 85.2% for correctly predicting the dependent variable when its true value is 0, and an overall correct percentage of 65.5%. The logistic regression model demonstrates good predictive performance in determining point-scoring changes.

2.3. Machine Learning Models

The logistic regression model achieved an accuracy of 65.5% in predicting whether the current point would score, indicating good performance and the ability to comparatively accurately predict score changes. To determine the best model for this problem among all prediction algorithms, we further improved the algorithms. We used 16 influencing factors as independent variables and employed five-fold cross-validation. We compared the results of logistic regression with LightGBM, XGBoost, SVM, and MLP models. Finally, we concluded that the LightGBM model had the best training performance in this dataset.

2.3.1 Model Introduction

The LightGBM algorithm is based on the gradient boosting decision tree (GBDT) framework. Compared to the popular XGBoost algorithm, LightGBM has faster training speed, lower memory consumption, and higher accuracy [7]. LightGBM introduces Histogram, Goss, and EFB algorithms based on XGBoost, significantly reducing the time required to construct a leaf. It supports efficient parallel training and can handle massive data swiftly. [8]. XGBoost is a machine learning algorithm used for regression, classification, and ranking tasks. It can operate on large-scale datasets and exhibits strong generalization capability. Its core lies in using multiple weak learners to construct a strong learner by iteratively optimizing the loss function [9]. SVM is a data classification algorithm originating from statistics. It can train finite data by fitting functions and obtain training results accordingly. SVM can train both linear and nonlinear data, requiring a small number of samples. It also demonstrates strong feature extraction capabilities [10]. The MLP model is a fully connected neural network that adjusts the weights of neurons to minimize prediction errors, thereby achieving model training for result prediction [11].

2.3.2 Five-Fold Cross-Validation

Step 1: Initially, we randomly divided the dataset into five subsets.

Step 2: Sequentially, one subset was chosen as the testing set, while the remaining four subsets were used as the training set. The model was trained and evaluated on the testing set. This process was repeated five times to ensure that each subset was used as the testing set once.

Step 3: LightGBM, XGBoost, SVM, MLP, and Logistic Regression algorithms were employed to train the training sets.

Step 4: For each training iteration, accuracy, recall, precision, F1 score, and the area under the ROC curve (AUC) were calculated to evaluate the performance of the models on the testing set.

The obtained prediction results are presented in Table 2 below:

Table 2: Five-Fold Cross-Validation Results

Algorithm	Accuracy	Recall	Precision	F1	AUC
LGBM	0.69	0.69	0.70	0.69	0.77
XGB	0.67	0.68	0.68	0.68	0.75
SVC	0.67	0.65	0.70	0.67	0.75
MLP	0.69	0.66	0.71	0.68	0.76
LR	0.67	0.69	0.67	0.68	0.72

From Table 2, it can be observed that LightGBM achieved the best prediction performance, while Logistic Regression has room for improvement. The ROC curves of the training and testing sets are depicted in Figure 1:

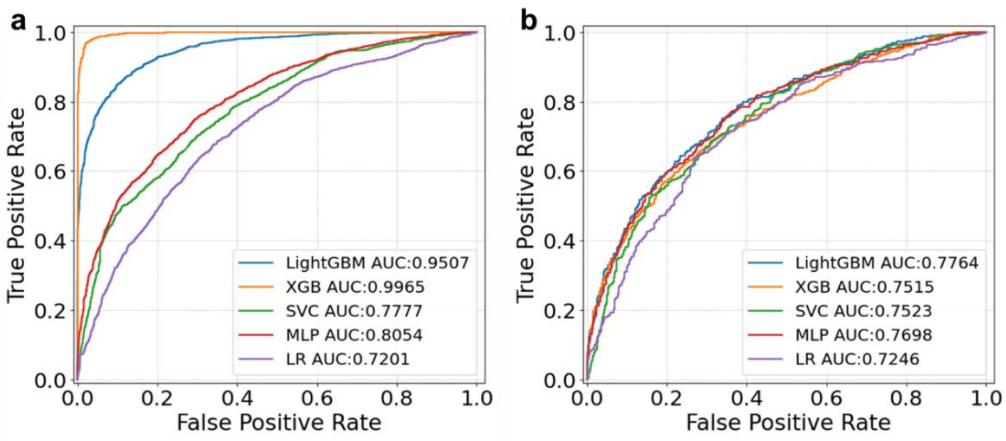

Figure 1: ROC curves of the five models. (a) ROC curve of the training set. (b) ROC curve of the test set

The figure depicts the ROC-AUC curve for the training set in (a) and the testing set in (b). In Figure 1, the area enclosed by the ROC curve and the x-axis represents the AUC, indicating the quality of the model algorithm. Hence, from Figure 1, it is evident that the AUC of the testing set's LightGBM ROC curve is the largest, indicating its superior performance in this case.

2.4. Model Results

Finally, we selected the LightGBM model to predict whether player 1 would score at the current point. The predicted result was utilized to calculate the momentum using the following equation:

$$y_{label} = \ln\left(\frac{p}{1-p}\right) \tag{4}$$

Here, p denotes the probability of player 1 scoring at the current point, which we consider as "momentum". Using equation (4) to calculate the momentum changes in continuous point changes. Taking the final match as an example, the fluctuation curve of Player 1's momentum for each point in the five sets is displayed in Figure 2:

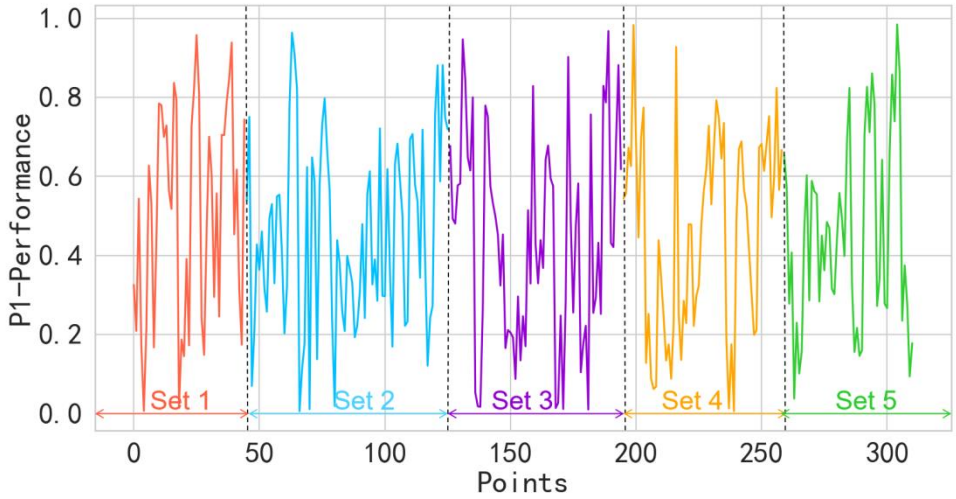

Figure 2: Fluctuation curve of player 1's momentum in the first match

Figure 2 reveals that most of the time, player 1's momentum is above 0.5, indicating that player 1 is more likely to score the current point. Furthermore, we calculated the average momentum of player 1 for each set in the final match, as shown in Table 3 below:

Table 3: Average Momentum of Player 1 in Each Set of the Final Match

Set 1	Set 2	Set 3	Set 4	Set 5
0.4689	0.5214	0.5045	0.4893	0.5226

From Table 3, it can be observed that the average momentum values for sets 2, 3, and 5 are greater than 0.5, indicating a higher probability of player 1 scoring. On the other hand, sets 1 and 4 show a higher probability of player 2 scoring, aligning with the actual outcomes.

3. Impact of Momentum on Athlete Performance

To explore the correlation between the fluctuation of momentum during a match and the success of the athlete, we consider using Pearson correlation analysis to calculate the correlation between momentum and the actual performance of the athlete, followed by Ordinary Least Squares (OLS) regression for further validation.

3.1. Pearson Correlation Analysis

The Pearson coefficient measures the ratio of the covariance of two variables to the product of their respective standard deviations. For two variables X Y, the calculation formula for the Pearson coefficient ρ_{XY} is as follows:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \tag{5}$$

Where, $Cov(X,Y)$ represents the covariance between X and Y, σ_X represents the standard deviation of X, and σ_Y represents the standard deviation of Y.

After analyzing the data for player 1, it was found that the p-value of the Pearson correlation analysis for momentum and actual performance is extremely small, almost close to 0. It is generally considered that a p-value less than 0.05 indicates a very high level of significance in the correlation. Therefore, the high significance level of the correlation between these two variables is established. The Pearson correlation coefficient is 0.59, indicating a high degree of correlation between the variables. The results are presented in Table 4:

Table 4: Pearson analysis parameters

Pearson Coefficient	p-value
0.59	0.0013

3.2. OLS Regression

Following the Pearson correlation analysis, to further validate the degree of correlation between momentum and performance, we utilized Ordinary Least Squares (OLS) regression for linear fitting analysis. We regarded the momentum (p) of player 1 as the independent variable, and the performance (y_{label}) as the dependent variable. The fitted results are presented in Table 5.

In the model fitting process, to assess the goodness of fit of the model, we established the F-statistic. The formula for F-statistic is as follows:

$$F = \frac{ESS/k}{RSS/(n-k-1)} \tag{6}$$

Here, RSS represents the residual sum of squares, ESS the explained sum of squares, k the number of predictors (independent variables), and n the sample size.

Since the Total Sum of Squares (TSS) remains constant regardless of changes in the model, there exists a strict negative correlation between ESS and RSS. If ESS reaches its minimum value, RSS will reach its maximum value, and the ratio of RSS and ESS will also reach its maximum value. Therefore, the larger the F-statistic, the better the model fit. In our case, the calculated F-statistic is significantly large, indicating a good model fit.

To compute the confidence interval for the coefficient term, we propose the following hypothesis:

Null Hypothesis H_0: The error term follows a normal distribution.

Alternative Hypothesis H_1: The error term does not follow a normal distribution.

The obtained result shows that the possibility of rejecting the null hypothesis is almost zero, indicating that the error term follows a normal distribution. The coefficient term for the independent variable is 0.9231, implying that a one-unit increase in momentum leads to an increase of 0.9231 in scores. At a 95% confidence level, the confidence interval for the effect of a one-unit increase in momentum on scores is estimated to be (0.830, 1.106).

3.3. Summary

The results of the Ordinary Least Squares regression model are presented in Table 5:

Table 5: OLS Regression Results

Parameter	F-statistic	Prob(F-statistic)	Coefficient Term (X1)	Possibility (P>|t|)	Confidence Interval at a 95% Confidence Level
Value	376.6	1.74e-73	0.9231	0.000	(0.830, 1.106)

In this section, we explore the relationship between the swings of momentum and runs of success in a match. Using Pearson correlation coefficients and Ordinary Least Squares (OLS) linear regression analysis, the study found a moderate positive correlation between momentum and actual scores. The statistical data obtained strongly suggests that swings in momentum and the runs of success for a player in a match are not random but significantly correlated.

4. Conclusions

In tennis matches, momentum reflects the advantage demonstrated by an athlete at a specific time in winning the match. It can provide confidence and motivation to the athlete, while also potentially impacting the performance and psychological state of the opponent. To predict the performance of athletes at a specific point, 16 potential factors influencing momentum in the competition were extracted and analyzed. Initially, the accuracy of prediction using the Logistic model was discussed, revealing an accuracy of only 65.5%. To improve the accuracy of our prediction model, a comparison was made through five-fold cross-validation using Logistic, LightGBM, XGBoost, SVM, and MLP models, with LightGBM achieving the highest accuracy of 77.6%. Finally, LightGBM was selected to predict the performance of player 1 at each point of the final match, and a momentum-time variation curve was plotted, indicating that the average momentum values in the 2nd, 3rd, and 5th sets were greater than 0.5, consistent with the actual situation. Regarding the impact of momentum on athlete performance, the Pearson correlation coefficient and Ordinary Least Squares were used to analyze the correlation between "momentum in the competition" and "actual performance". The obtained p-value was less than 0.05, indicating a significant correlation, with a Pearson correlation coefficient of 0.59, representing a high degree of correlation. Subsequently, through OLS regression, a significant F-statistic of 376.6 was found, signifying a good fit for the model. It was also discovered that the possibility of rejecting the null hypothesis is almost 0, indicating a highly significant level of correlation between the two variables. The coefficient of the independent variable was 0.9231, showing a positive correlation between the two variables.

This paper provides a research framework for using the LightGBM model to predict the variation of athlete momentum in tennis matches, demonstrating the feasibility of applying machine learning in the field of sports science and sports psychology.

References

[1] Li Haopeng. Exploration of Intelligent Robots based on Machine Learning Methods[J]. Communications World, 2019, 26(04): 241-242.

[2] Hu Yanan, Yu Huan, Wu Shengchuan, et al. Research Progress and Challenges in Predicting Mechanical Properties of Additive Manufacturing Metals based on Machine Learning[J/OL]. Acta Mechanica Sinica: 1-25 [2024-02-14].

[3] Tang Junfang, Xiong Jian, Liu Xiangjun, et al. Adaptive Weighted Combination Prediction of Rock Mechanics Parameters in Mahu Sag Fengcheng Formation [J]. Petroleum Geophysics Exploration, 2024, 59(01): 1-11.

[4] Wang Zepeng, Li Kunpeng, Zhou Yu, et al. Selection of Key Genes and Construction of Prediction Model for Gastric Cancer based on Machine Learning [J]. Chinese Journal of Medical Physics, 2024, 41(01): 115-124.

[5] Zhao Luan. Bayesian Quantile Regression Analysis and Application of Binary Vector Autoregressive Model [M]. 2023. Changchun University of Technology, MA thesis.

[6] Gao Yan. Research on Logistic Modeling and Its Application[J]. Journal of Langfang Normal University (Natural Science Edition), 2016, 16(03): 12-15.

[7] Liu Bo, Wang Xiaotian, Xu Chen. Aggregated Off-site Delay Prediction of Airport based on LightGBM Algorithm[J]. Journal of Xi'an Aeronautical University, 2024, 42(01): 26-30.

[8] Wu Hairong, Li Zhenhua, Cheng Ziyi, et al. Invalid Data Cleaning Method for Audible Noise of Ultra-High Voltage Direct Current Transmission Lines based on Attention Mechanism and LSTM-LightGBM [J/OL]. Southern Power System Technology: 1-10 [2024-02-14].

[9] Fu Pei, Cui Lan, Li Shuo. Experiment on Differentiation of Photosensitive Ink Types Based on Hyperspectral Imaging[J/OL]. Chinese Journal of Inorganic Analytical Chemistry: 1-9 [2024-02-15].

[10] Xu Liangde, Guo Ting, Lei Caijia, et al. Algorithm Design for Load Prediction of Grid Integration based on Support Vector Machine[J]. Electronic Design Engineering, 2024, 32(03): 12-16.

[11] Ma Tianshou, Zhang Dongyang, Chen Yingjie, et al. Prediction Method of Horizontal Well Fracture Pressure based on Neural Network Model[J]. Journal of Central South University (Natural Science Edition), 2024, 55(01): 330-345.