Exploring momentum in tennis using LightGBM-MLR fusion modeling

Hongwei Liu*, Qi Jing, Xufeng Wang

Department of Faculty of Robot Science and Engineering, Northeastern University, Shenyang, China, 110819

*Corresponding author: neu_fengnan@126.com

Abstract. Tennis Performance Prediction is a highly regarded study published daily in the 《Journal of Sports Science》. Its simplicity of prediction methodology and depth of insight have made it widely popular. In this paper we identify the factors influencing momentum in tennis, materialize the vague concept of momentum, and based on the results obtained we advise players on how to improve their performance. Initially, we meticulously scrubbed the raw data and narrowed down eight critical factors influencing player performance. Leveraging the gradient boosting tree, we formulated a LightGBM-MLR Evaluation Model to ascertain the significance of these eight factors. Post normalization, we employed a Multiple Linear Regression Model to compute a comprehensive player performance score. We then evaluated the model's accuracy, analyzed the outcomes to offer targeted recommendations for player performance enhancement, and compared the fit of five machine learning models, ultimately concluding that LightGBM offered the most optimal fit. This research has helped to improve player performance and prediction of game trends, and it can be extended to other sports.

Keywords: Tennis, momentum, Light Gradient Enhanced Machine, Multivariate Linear Regression.

1. Introduction

In the Wimbledon 2023 men's singles final, Spain's up-and-coming Carlos Alcaraz clinched victory over renowned Grand Slam champion, Novak Djokovic. Despite Djokovic's initial advantage, he lost the match, with momentum cited as a key factor. In sports, teams or athletes often perceive momentum shifts during a match, a phenomenon known as "momentum"[1]. However, measuring this accurately is challenging, and it's unclear how in-game events influence this momentum[2]. Therefore, to identify the determinants of momentum and its impact on match outcomes, it's crucial to analyze the correlation between player performance and momentum using match-time data[3].

Gradient Boosting Decision Tree (GBDT) is a popular machine learning algorithm and has quite a few effective implementations such as XGBoost and pGBRT[4]. Neural network models play an important role in tennis performance prediction. These models can capture complex nonlinear relationships, extract key features from a large amount of historical data, and build accurate prediction models. In this work, LightGBM is the core of its leaf-wise strategy and higher training efficiency[5]. The LightGBM model has been utilized in recent years in the prediction of various scenarios, such as Machine learning algorithms have attained widespread use in assessing the potential toxicities of pharmaceuticals and industrial chemicals[6], Prediction of volatility and incorrectness of wind power generation using LightGBM modeling[7], The improved LightGBM model based on the Bayesian hyper-parameter optimization algorithm achieves a mean square error of 0.5961 in blood glucose prediction[8], propose a forecasting method of train passenger load factor of high-speed railway based on LightGBM algorithm of machine learning[9], etc. This shows that the LightGBM model has very good predictive power and can be applied to many different scenarios, and scholars are constantly improving the model[10].

Despite the wide range of applications of the LightGBM model, predicting player performance in tennis is a complex problem that is affected by multiple factors. Existing research on LightGBM modeling does not deal well with multivariate and time-dependent issues. So we use the LightGBM-

MLR combined model for the analysis of this problem. We compare the fitting effect for multiple machine learning models.

2. The fundamentals of LightGBM-MLR

2.1. The Structure of the Light Gradient Boosting Machine

Light Gradient Enhanced Machine (LightGBM) uses a histogram-based algorithm to discretize continuous eigenvalues into k features to improve computational efficiency and prevent overfitting. Unlike traditional Gradient Boosting Decision Tree (GBDT), it uses a pre-ranking-based decision tree technique and voting parallel strategy for model training. In the optimization process, Light Gradient Enhanced Machine (LightGBM) uses a leaf-based algorithm to construct the decision tree to reduce unnecessary computations. Overall, a Light Gradient Enhanced Machine (LightGBM) is a gradient-enhanced decision tree algorithm with fast training speed, high accuracy, low memory consumption, etc., which is suitable for handling massive data. The Histogram-Based Decision Tree Algorithm is shown in Figure 1.

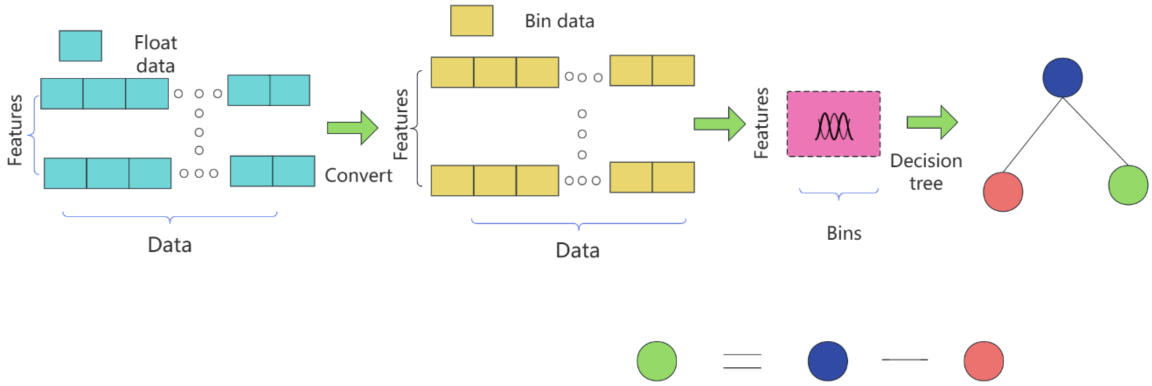

Figure 1. Histogram-Based Decision Tree Algorithm

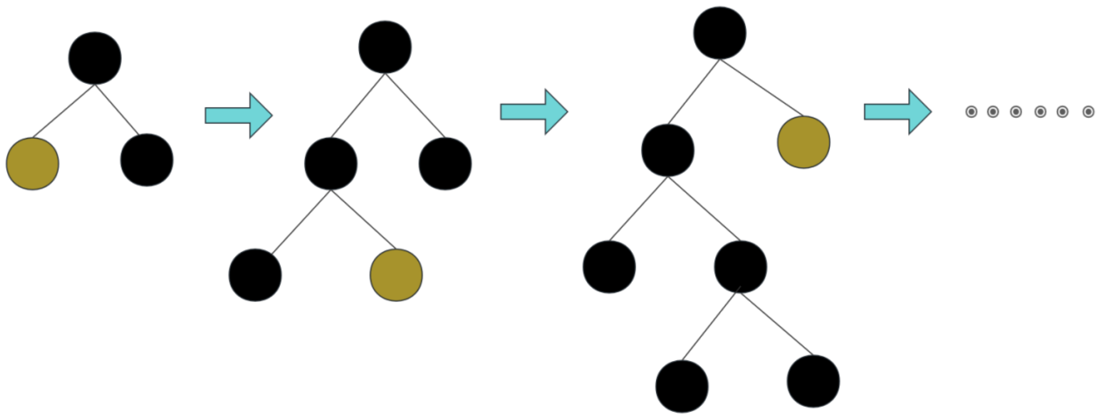

Figure 2. Schematic Diagram of Leaf-Wise Tree Growth

Light Gradient Enhanced Machine (LightGBM) employs a histogram-based decision tree algorithm, whose core idea is to discretize the continuous eigenvalues into k values and generate a histogram of corresponding width k. When traversing the samples, the Light Gradient Enhanced Machine (LightGBM) uses the discretized values as indexes, which significantly improves the computational efficiency. Through one traversal, the histogram accumulates the required statistics, and the discrete values can be used to find the optimal splitting point for the splitting operation. The Schematic Diagram of Leaf-Wise Tree Growth is shown in Figure 2.

Another Light Gradient Enhanced Machine (LightGBM) feature is the use of a more efficient leaf growth strategy, i.e., a leaf growth strategy with depth restriction, also known as leaf orientation. Before making a split, the strategy first traverses all the leaves in the tree and then finds the leaf with the largest split gain to split again and repeats this operation. Experiments have shown that using the leaf direction growth strategy results in higher accuracy for the same number of splits and adds a maximum depth limit to prevent overfitting.

The general model of a Light Gradient Enhanced Machine (LightGBM) consists of four basic elements, which are:

(1) Tree growth strategies: Light Gradient Enhanced Machine (LightGBM) uses gradient cueing to grow decision trees, which employs a small batch stochastic gradient descent algorithm to minimize the loss function and thus maximize model accuracy

(2) Feature Selection: Light Gradient Enhanced Machine (LightGBM) uses an objective function-based feature selection algorithm that can reduce the number of features in the training data while maintaining performance.

(3) Data loading: Light Gradient Enhanced Machine (LightGBM) can read data in different formats, such as CSV, libsvm, etc., and use the concurrent loading technique to load larger data.

(4) Normalization Processing: Light Gradient Enhanced Machine (LightGBM) provides an automatic normalization processing function, which can intelligently help data deflation and make the model more robust.

The formula for the LightGBM model is as follows:

$$Obj(t) = L(t) + \Omega(t) + c \tag{1}$$

where $\Omega(t)$ denotes a regular function, $L(t)$ denotes the loss function, C denotes an additional parameter,t indicates sampling time.

2.2. The structure of Multivariate Linear Regression

Multiple linear regression (MLR), also known as multivariate regression, is a statistical technique that uses multiple explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and the response (dependent) variables. Essentially, multiple regression is an extension of ordinary least squares (OLS) regression because it involves multiple explanatory variables.

Simple linear regression is a function that allows an analyst or statistician to make predictions about one variable based on known information about another variable. Linear regression is only used when there are two continuous variables (independent and dependent). The independent variable is the parameter used to calculate the dependent variable or outcome. Multiple regression models extend to multiple explanatory variables.

The Multiple linear regression (MLR) is based on the following assumptions, which are:

There is a linear relationship between the dependent and independent variables

The correlation between the independent variables is not too high

Yi observations are selected independently and randomly from the population

Residuals should be normally distributed

We normalize the data in different dimensions:

$$v_s = \frac{v - v_{\min}}{v_{\max} - v_{\min}} \tag{2}$$

where V_s denotes the normalized value, V is the original value, and V_{max} and V_{min} represent the maximum and minimum values of the eight attributes respectively.

$$s_i = \sum_{j=1}^{m} w_j P_i,\ i = 1, \cdots, n \tag{3}$$

where S_i represents player performance score, W_i indicates weights, and P_i represents normalized values of factors.

For the completeness and reliability of the assessment phase, the assessment criteria used three methods: score assessment, point prediction assessment, and error assessment. The score metrics used in the score assessment method include i.e., coefficient of determination (R2), mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), and maximum error (ME). The error measures were calculated using Scikit-learn with the following parametric equations.

Mean absolute percentage error(R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2}, \tag{4}$$

Mean square error（MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i)^2, \tag{5}$$

Average absolute error(MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - y_i\right|, \tag{6}$$

where y_i denotes the composite score, $\overline{y_i}$ indicates the average of the composite scores, y_i indicates the predicted value of the composite score.

We first standardized the relevant data, and then analyzed the Light Gradient Boosting Machine (LGBM) machine learning model to obtain the weight of importance of different features, and then constructed a multiple linear regression model to change the weight of different factors into coefficients, and weighted the average to obtain the value of players' real-time performances.

$$\left(p1 \quad p2 \quad ... \quad p8\right) \xrightarrow{LightGBM} \left(w1 \quad w2 ... \quad w8\right) \xrightarrow{\text{Multiple linear regression}} \left(score\right) \tag{7}$$

where P_x represents 8 factors affecting momentum, W_x represents the corresponding weight

3. Results

3.1. Numbering of different data

We get eight factors that affect the potential energy in tennis(shown in Table 1), numbered to make it easier to follow the lines of the article.

Table 1. Notations used in this paper

Symbol	Description	data name
P_1	Whose side is serving	*server*
P_2	Which side is leading in scoring	*lead _ score*
P_3	players hit an untouchable winning serve	p_x _ *ace*
P_4	players hit an untouchable winning shot	p_x _ *winner*
P_5	players missed both serve and lost the point	P_x*_double_fault*
P_6	player made an unforced error	P_x*_unf_err*
P_7	players won the point while at the net	P_x*_net_pt_won*
P_8	Previous wins	*Previous_point_victor*

3.2. Analysis of experimental results

We use the eight sets of data, input the feature sequences, and train the model with the ratio of training set: test set = 7:3 to get the feature importance of different factors for the composite score. Then, using the multiple linear regression model, the feature importance is weighted and averaged to get the player's performance as good or bad, which reflects the player's performance at a specific time in real time. Table 2 shows the characteristic importance of the eight factors:

Table 2: Characteristic importance

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
Importance	0.16	0.336	0.092	0.113	0.00	0.123	0.071	0.104

We found that point leads, which side of the serve, and player errors have the greatest impact on momentum, and athletes can use this as a basis for adjusting their form and building momentum to play at a better level. The pictures below visualize the data:

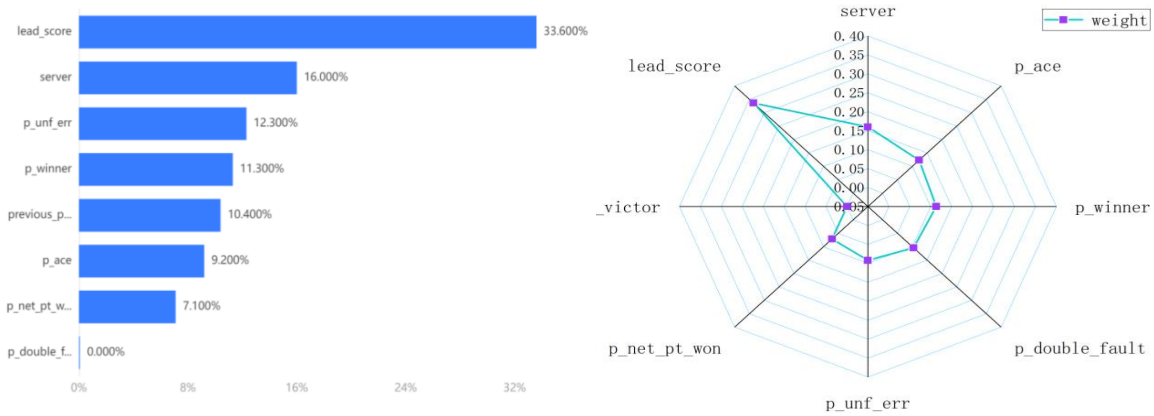

Figure 3. Eight factors with different impacts

Bringing in the data, we arrive at the result in Table 3:

Table 3: Model Accuracy Test Results

	MSE	RMSE	MAE	R^2
Training Set	0.149	0.386	0.299	0.602
Validation Set	0.04	0.199	0.173	0.841

The smaller the value of MSE, RMSE, MAE, ME means the smaller the model error and the higher the accuracy; the closer R^2 is to 1, the better the model fits the data.

It is easy to see that the LightGBM-MLR Evaluation Model has a good analytical ability to inidentify how good or bad a player's performance is at a particular time of the game, and the absolute value of $R^2$is close to 1 indicating that the model has a high degree of accuracy in capturing the whole process of the game. Based on the analyses in Table 3, we conclude that our model is comprehensive and effective.

3.3. Advice for players

(1) Let players clearly understand what "momentum" is, and that having "momentum" in a game will make it easier for players to score. In training, the performance characteristics of the players to do a comprehensive ability to score, the feeling of "momentum" visualization, you can consciously guide the players to feel the existence of "momentum", and through positive ways to accumulate and use this "momentum". Momentum".

(2) When players are at a disadvantage, remind them to stay calm and not to give up easily. Encourage them to look for opportunities in adversity and gradually turn the situation around. And when they are at an advantage, remind them to stay focused and not to slack off because of a momentary lead.

(3) In the course of the game, according to the change of "momentum", adjust the application of technology at the right time. For example, when the "momentum" is in favor of the serving side, you can consider using a more aggressive serve; on the contrary, when the receiving side masters the "momentum", you should pay more attention to the defensive counterattack.

(4) According to the process of the game and the "momentum" of the change, flexible adjustment of tactical strategy. For example, when one side has obvious "momentum", the other side may need to take a more aggressive offensive strategy to reverse the situation.

3.4. LightGBM model fits well

In the first modeling question, we use two gradient boosting tree models, LightGBM and XGBoost, and five models, Support Vector Classifier, Multilayer Perceptron, and Logistic Regression, to analyse feature importance. The five models were used to analyze feature importance, and AUC (Area Under the Curve) was used to assess the performance of the classification model and to measure the predictive accuracy of the model. It can be seen that XGBoost overfitting, Support Vector Classifier, Multilayer Perceptron, and Logistic Regression models only have a fitting effect close to 0.7, while the LightGBM model has a fitting effect of 0.9417, which indicates that the fitting effect of using the model is better.

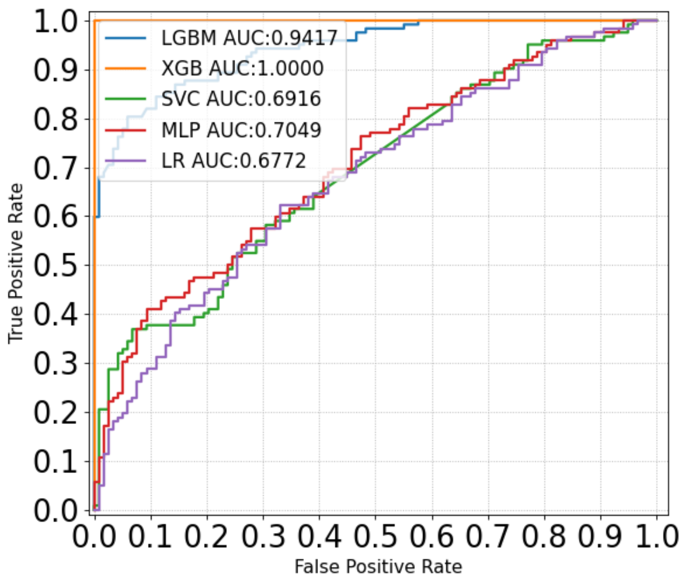

Figure 4. Fitting effects of five different models

4. Conclusions and outlooks

In this paper we identify the factors influencing momentum in tennis, materialize the vague concept of momentum, and based on the results obtained advice players on how to improve their performance. We first examined and cleaned the raw data and filtered out 8 factors relative to the performance of players. We constructed a LightGBM-MLR Evaluation Model based on a gradient boosting tree to obtain the characteristic importance of the 8 factors. We normalized the data and then used a Multiple Linear Regression Model to derive a comprehensive player performance score. We then examined the accuracy of the model, as well as analyzd the results to give appropriate recommendations to improve the player's performance, and finally compared the fit of the five machine learning models, proving that LightGBM fitted the best. While this research is useful for improving individual player performance and predicting outcome analysis, this tennis-based research can be extended to other sports to improve individual performance and predict game trends.

References

[1] Dietl H, Nesseler C. Momentum in tennis: Controlling the match[J]. UZH Business Working Paper Series, 2017 (365).

[2] Depken C A, Gandar J, Shapiro D. Strategic and Psychological Momentum in Professional Tennis[J]. Available at SSRN 3659495, 2020.

[3] Meier P, Flepp R, Ruedisser M, et al. Separating psychological momentum from strategic momentum: Evidence from men's professional tennis[J]. Journal of economic psychology, 2020, 78: 102269.

[4] Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree[J]. Advances in neural information processing systems, 2017, 30.

[5] Zhang Y, Zhu C, Wang Q. LightGBM-based model for metro passenger volume forecasting[J]. IET Intelligent Transport Systems, 2020, 14(13): 1815-1823.

[6] Zhang J, Mucs D, Norinder U, et al. LightGBM: An effective and scalable algorithm for prediction of chemical toxicity–application to the Tox21 and mutagenicity data sets[J]. Journal of chemical information and modeling, 2019, 59(10): 4150-4158.

[7] Ju Y, Sun G, Chen Q, et al. A model combining convolutional neural network and LightGBM algorithm for ultra-short-term wind power forecasting[J]. Ieee Access, 2019, 7: 28309-28318.

[8] Wang Y, Wang T. Application of improved LightGBM model in blood glucose prediction[J]. Applied Sciences, 2020, 10(9): 3227.

[9] Wang B, Wu P, Chen Q, et al. Prediction and analysis of train passenger load factor of high-speed railway based on LightGBM algorithm[J]. Journal of Advanced Transportation, 2021, 2021: 1-10.

[10] Al Daoud E. Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset[J]. International Journal of Computer and Information Engineering, 2019, 13(1): 6-10.