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

# Machine Learning-Based Tennis Match Performance Prediction Model

Yiyang Zhou, Xingyu Chen, Zijin Wang *

International School, Beijing University of Posts and Telecommunications, Beijing, China, 100876

* Corresponding Author Email: WZJ2004324@bupt.edu.cn

**Abstract.** The paper mainly builds a machine learning model for tennis match score prediction and conducts significance analysis on the effects of concomitant conditions using PSM. Firstly, the paper constructed a set of new metrics system, including whether the athlete is on the serve side, his/her personal skill, the level of fatigue, and his/her mentality during the match, tested the significant effect of these metrics on the model prediction by binary logistic regression, and used various machine learning model such as XGBOOST, SVC, LGBM to build the prediction model. Then, the paper trained an LGBM model based on the tennis match data set and improved the metrics system to realize the dynamic evaluation of players' real-time performance, i.e., "momentum." Then, in order to prove that players' fluctuations and successes in the game are not random and that the role of "momentum" in the game really exists, the paper analyzes the concomitant conditions of scoring fluctuations using a counterfactual analysis framework and a propensity score matching algorithm and argues that the essential conditions have a significant effect on players' performances by comparing the T-stat values of different core explanatory variables, which ultimately proves and further explains the importance of the role of "momentum" in the game's outcome.

**Keywords:** Momentum, Tennis Match, Machine Learning, Statistics, Counterfactual Framework.

## 1. Introduction

This paper mainly studies two parts: the first part is to build a machine learning model based on tennis match score prediction, and the second part is to study Significance Analysis on the effect of concomitant conditions through PSM.

In the study of the first part, the paper emphasized and deeply explored the factors not limited to whether one is on the serving side. The factors include the player's individual technical ability, fatigue level, and real-time mental state of the match. Based on these factors, a more comprehensive metric system was constructed to assess the real-time performance of players on the court [1]. The paper first cleaned and labeled the data and used the LOGISTIC algorithm to determine the significance of the current metrics for model prediction. After obtaining metrics with significant predictive effects, the paper chose various machine learning algorithms to calibrate them. The paper computed these metrics and the labels of whether a player received a score, which provided data support for feature selection in the modeling process. The paper then employed multiple machine learning algorithms, including MLP, SVC, XGBOOST, and LGBM, to evaluate the performance of the models through 5-fold cross-validation, using the f1, accuracy, recall, roc, auc, and precision curve metrics to determine their The differences and advantages and disadvantages of LGBM provide data support for the final selection of LGBM, and also ensure the generalization ability and robustness of the model [2]. The paper found that the LGBM can capture the momentum changes during the game and also evaluate the real-time performance of the players, which improves the accuracy of predicting the outcome of the game.

In the second part of the study, the paper uses this predictive model along with a counterfactual analytic framework to argue for the existence of a "momentum" effect. The dictionary definition of momentum is "energy gained through motion or a series of events," and the existence of momentum depends on the existence of key conditions that generate or diminish it. Therefore, the model only needs to focus on the conditions that accompany fluctuations in a player's scoring trend. Analyzing

the correlation between the resulting condition and the score of the game can determine whether the accompanying condition has a significant effect on scoring performance [3].

## 2. Machine Learning-Based Tennis Match Performance Prediction Model

### 2.1. Statistical LOGISITIC test based on the index system

The paper first analyzes and discusses the influencing factors of the player's scoring in a tennis match and builds the index system according to whether he is a server, fatigue level (mileage of running map), real-time scoring, mental state (number of errors and netting) and other factors (shown in table 1) [4],as shown in Table 1.

**Table 1:** Player Score Factor Index System.

| Indexes | Categorization |
|---|---|
| The number of games won in the current set | Individual athletic ability |
| Scoring lead progress in this game | Individual athletic ability |
| Is it the person who serves the ball | Real-time situation |
| Whether the previous point was scored or not | Real-time situation/Individual athletic ability |
| set lead progress for this match | Individual athletic ability |
| whether or not the game serves(untouched) | Individual athletic ability/Athletic status |
| Whether or not the game is returned for a score(untouched) | Individual athletic ability/Athletic status |
| Is there a double fault in this game | Athletic status |
| Whether or not there was an unforced error in this game | Athletic status |
| The number of net approaches and the percentage of net points won | Level of fatigue/athletic status |
| Ratio of scoring opportunities to points scored on opponent's serve to points scored on confidant's serve in this set | Individual athletic ability/athletic status |
| Total map miles run within this match | Level of fatigue |
| Total map miles run within the last three points | Level of fatigue |
| Previous Point's Run Mileage | Level of fatigue |
| Real-time pace of serve | Athletic status |
| Whether or not it's a real-time speed matching interaction term for the golfer and the serve | Athletic status |

The paper tests the correlation about the impact of the score situation based on the statistical regression approach, to verify that the impact of each index in the whole index system on the score is important. Noting that the point score situation is dichotomous, the paper uses a dichotomous LOGISITIC regression model. The basic form of the logistic regression model is:

$$P(Y=1|X)=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n)}} \quad (1)$$

The output using SPSS software is shown in Table 2:

**Table 2:** The output using SPSS.

| Classification Table[a] | | | | | |
|---|---|---|---|---|---|
| | Observed | | Predicted | | |
| | | | label | | Percentage Correct |
| | | | 0 | 1 | |
| Step 1 | label | 0 | 271 | 517 | 34.4 |
| | | 1 | 184 | 1058 | 85.2 |
| | Overall percentage | | | | 65.5 |
| a. The cut value is .500 | | | | | |

It can be observed that the accuracy of the binary LOGISITIC regression in SPSS is 65.5, the model classification accuracy is 85.2 regarding the samples where the player has genuinely earned a score, and 34.4 for the samples where the player has indeed not scored a score, which is much lower than the scored samples. It can be concluded that the current model prefers to categorize the samples to the actual score, which is the case with label 1. The current LOGISITIC regression analysis is only a prospective qualitative analysis to test how the index affects the player's score,as shown in Table 3.

**Table 3:** Variables in the Equation.

| Variables in the Equation | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | B | S.E. | Wald | df | Sig. | Exp(B) |
| Step 1a | x1 | 0.016 | 0.176 | 0.009 | 1 | 0.926 | 1.017 |
| | x2 | -1.141 | 0.363 | 9.881 | 1 | 0.002 | 0.319 |
| | x3 | 0.689 | 1.286 | 0.287 | 1 | 0.592 | 1.992 |
| | x4 | 0.170 | 0.166 | 1.048 | 1 | 0.306 | 1.186 |
| | x5 | 0.032 | 0.203 | 0.024 | 1 | 0.876 | 1.032 |
| | x6 | 0.689 | 0.127 | 29.400 | 1 | 0.000 | 1.991 |
| | x7 | 0.456 | 0.128 | 12.697 | 1 | 0.000 | 1.577 |
| | x8 | 0.472 | 0.153 | 9.491 | 1 | 0.002 | 1.603 |
| | x9 | -0.535 | 0.109 | 24.287 | 1 | 0.000 | 0.586 |
| | x10 | 0.914 | 0.124 | 54.371 | 1 | 0.000 | 2.493 |
| | x11 | 0.073 | 0.140 | 0.273 | 1 | 0.601 | 1.076 |
| | x12 | 0.120 | 0.288 | 0.174 | 1 | 0.676 | 1.128 |
| | x13 | -0.164 | 0.482 | 0.116 | 1 | 0.734 | 0.849 |
| | x14 | -1.445 | 0.657 | 4.842 | 1 | 0.028 | 0.236 |
| | x15 | 0.511 | 0.626 | 0.668 | 1 | 0.414 | 1.667 |
| | x16 | -0.876 | 1.693 | 0.268 | 1 | 0.605 | 0.416 |
| | Const | -0.072 | 0.346 | 0.043 | 1 | 0.835 | 0.931 |
| a. Variable(s) entered on step 1: x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11, x12, x13, x14, x15, x16. | | | | | | | |

Table 3 shows that half of the p-values in the data are less than 0.05, which suggests that x significantly affects y. Therefore, the independent variables that can affect the scores in real time in this index system include x2, x6, x7, x9, and x14. Therefore, it is concluded that personal skill, fatigue, and psychological state all affect the scores.

### 2.2. Machine learning model and test based on metrics system

The results are evaluated using 5-fold cross-validation and confusion matrix-based accuracy, procession, recall, auc, roc, f1 curves comparing MLP, SVC, XGBOOST, LGBM algorithms [5]. The

results based on the confusion matrix are shown in Table 4 based on the auc and roc training set, as shown in Table 4.

**Table 4:** Results Based on Confusion Matrix.

| | acc | recall | precision | f1 | auc |
|---|---|---|---|---|---|
| LGBM | 0.69 | 0.69 | 0.7 | 0.69 | 0.77 |
| XGB | 0.67 | 0.68 | 0.68 | 0.68 | 0.75 |
| SVC | 0.67 | 0.65 | 0.7 | 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 |

The test set results are shown in fig.1 and fig.2, respectively.

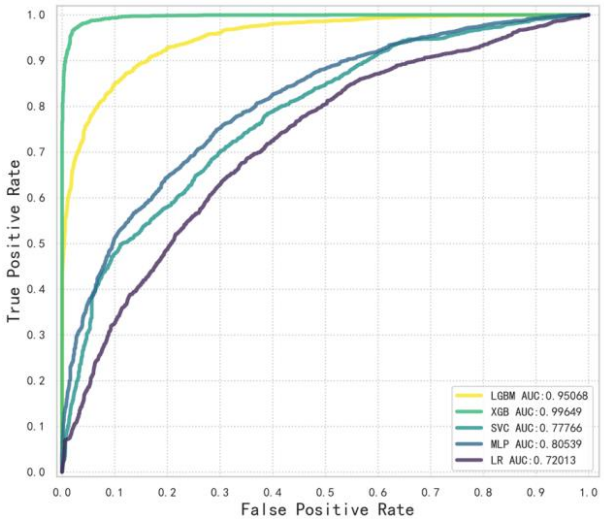


**Fig 1:** Training set ROC Curve.

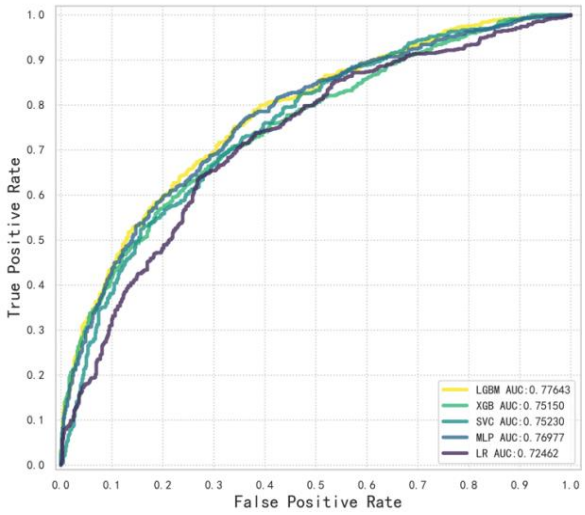


**Fig 2:** Test set ROC curve.

It can be seen that the LGBM has the best effect where the results of accuracy, procession, recall, auc, roc and f1 are 0.69, 0.69, 0.7, 0.69 and 0.76, respectively. The difference between several indexes is very small, which indicates that the model does not have obvious positive and negative preferences, proving that the LGBM is qualified. The ROC curve reflects the changes in precision and recall metrics for different thresholds, and through the ROC curve, it is also verifiable that LGBM is the most effective algorithm. So, the paper re-trained the model, selecting the real-time data from the 2023 Wimbledon final for real-time performance visualization, selecting the player Carlos Alcaraz's results in fig.2. In fact, there is an upper limit for the accuracy of the score point prediction, which is not counter-intuitive because the real situation of the player whether to score the impact factor is very complex, so the model can only judge 0.7 players' real scoring situation. According to the data analysis, the paper finds that in this game the sender and the scorer are consistent with the player's scoring situation, and the momentum is higher when Carlos Alcaraz wins and lower when he loses the victory, which can prove that the robust model is reliable and valid [6].

## 3. Significance Analysis on the effect of concomitant conditions through PSM

The paper proposes 16 concomitant conditions that may affect the score distribution and fills the cleaned data into the labeled items of the corresponding concomitant conditions. For the given reference data and the large number of score distributions generated by the prediction model, the paper used a counterfactual analysis framework to analyze whether the effects of the concomitant conditions on the game results are significant [7]. As an important theory in causal prediction techniques, the counterfactual analysis framework can make causal predictions more explanatory, comparing the output differences of the hypothesized results horizontally. This approach reduces the

dimensionality of covariates and reduces the number of independent variables. By adjusting for individual differences through the propensity score, all confounders are equally comparable except for the exposure/treatment factor and the outcome variable, which are unevenly distributed [8].

The propensity score matching (PSM) shown in fig.3 analyzes the effect of the concomitant conditions (x1~x16) on the match score (label), the paper used Stata 17 to match the data with propensity scores, and the t-stat values for the different core explained variables are shown in fig.3 When 1.64 <= |T-stat| <= 1.95, we consider it to be significant at the 10% level when 1.95 <= |T-stat| <= 2.58, we consider it to be significant at the 5% level, and when |T-stat| >= 2.58, we consider it to be significant at the 1% level. Therefore, it is reasonable to infer that x12, x13, x14, and x15 are extremely significant for score performance, x1, and x2 are more significant for score performance, and the rest of the concomitant conditions are not so significant for score performance [9].

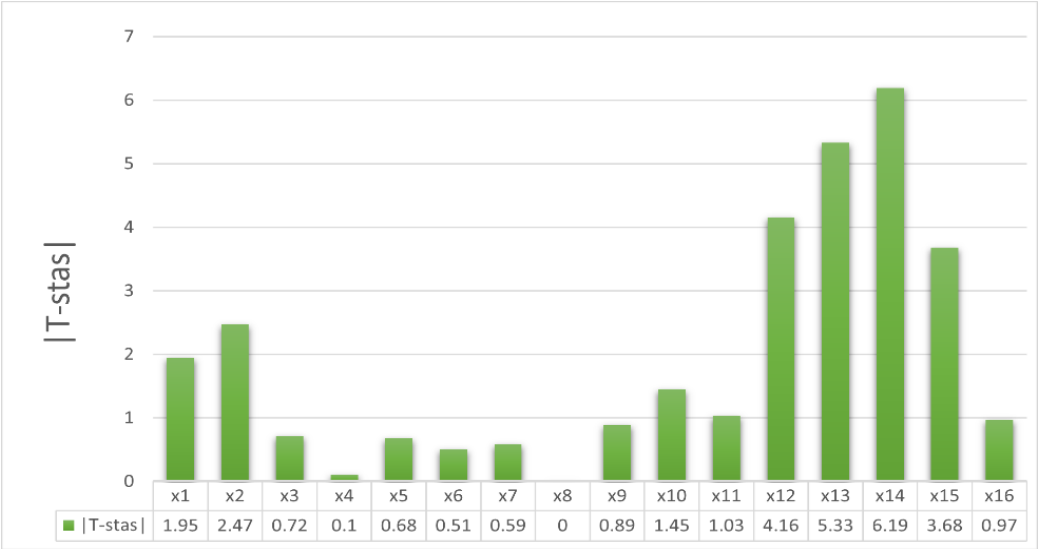


**Fig 3.** PSM performing on the data and the T-stat values under different core explained variables.

To summarize, the idea that players'scoring fluctuations in a game are completely random is incorrect. The emergence of key conditions often signals a change in momentum, while a change in momentum often signals a change in scoring performance. Mileage in the current match, mileage in the last 3 points, mileage in the last point, and real-time pace of serve have an extremely significant impact on the score of a match. These concomitant conditions were likely to give players a substantial degree of momentum. The number of games won in the set, and the progress of the score lead in a single game significantly affect the score on a certain level. These concomitant conditions were likely to generate a moderate degree of momentum on the player, while the remaining concomitant conditions were likely to have little effect on the player [10].

## 4. Conclusions

The research introduced and implemented a sophisticated machine learning model, with Light Gradient Boosting Machine (LGBM) being determined as the most effective algorithm for predicting tennis match scores based on a comprehensive set of metrics. These metrics were meticulously designed to encapsulate various dimensions of a player's performance, including technical skills, fatigue level, and mental state throughout the match. The results have proven that the LGBM model, with its outstanding performance as measured against key indicators like accuracy, precision, recall, and the ROC curve, emanates a strong predictive power which surpasses that of its counterparts, such as XGBOOST, SVC, and MLP models.

The essence of "momentum" within tennis matches, an often speculated but rarely quantified aspect, has been significantly illuminated by this research. Through the employment of a counterfactual analysis framework combined with Propensity Score Matching (PSM), it was discovered that not all

score fluctuations are random. Instead, certain critical conditions, notably the mileage covered by players during a match, and the speed of serve, among others, significantly influence scoring outcomes and can either generate or diminish a player's momentum. This insight challenges the prevailing notion of randomness in scoring fluctuations and highlights the predictability and influence of momentum on match outcomes.

Additionally, the analysis delineated the varying degrees of impact different conditions have on momentum and, consequently, match scores. It found that factors such as cumulative distance covered in the match or specific segments thereof, as well as the serving speed, hold a profound influence on a player's performance. In contrast, other examined conditions, such as the progression of score leads within games or sets, were found to exert a more moderate impact. This tiered understanding of momentum's drivers underscores the multifaceted and nuanced nature of performance dynamics in tennis.

In conclusion, this paper has pushed the boundaries of predicting tennis match outcomes by ingeniously integrating a machine-learning approach with a deep dive into the concept of momentum, debunking the notion of scoring randomness. The successful application of the LGBM model validates its potential as a robust tool for real-time match performance analysis, as showcased by the empirical testing with data from the 2023 Wimbledon final featuring Carlos Alcaraz. This research not only enhances our understanding of key performance indicators in tennis but also opens avenues for further explorations into the predictive modeling of sports outcomes, leveraging machine learning technologies.

## References


[1] Zhang Rong. Tennis match player prediction and analysis [D]. Yunnan University, 2023.

[2] Whiteside D, Reid M. Spatial characteristics of professional tennis serves with implications for serving aces: A machine learning approach [J]. Sports Sci. 2017, 35(7): 648–54.

[3] Forcher L, Beckmann T, Wohak O, Romeike C, Graf F, Altmann S. Prediction of defensive success in elite soccer using machine learning-Tactical analysis of defensive play using tracking data and explainable AI [J]. Sci Med Footb. 2023 Aug 4: 1-16.

[4] Liu Mengxin, Yuan Ruowei. Summary of the application of artificial intelligence in the tennis technical movement analysis [J]. Journal of contemporary sports science and technology, 2023, 13(33): 20-22.

[5] Robert Seidl, Patrick Lucey. Live Counter-Factual Analysis in Women's Tennis using Automatic Key-Moment Detection [DB/OL]. [2022-03].

[6] Jiang, H.W.; Zou, B.; Xu, C.; Xu, J.; Tang, Y.Y. SVC-Boosting based on Markov resampling: Theory and algorithm [J]. Neural Netw. 2016, 131, 276–290.

[7] C. Floyd, M. Hoffman, E. Fokoue. Shot-by-shot stochastic modeling of individual tennis points, Journal of Quantitative Analysis in Sports [J], 16(1), 57-71, 2020.

[8] Zhang Lifei, Zhu Yang, Wu Xuan, et al. Analysis of Winning Factors of Chinese Elite Men's Singles Tennis Players Driven by Data -- A case study of Zhang Zhizhen [C]// Chinese Society of Sport Science. Abstract Collection of the 13th National Sports Science Conference -- Wall Newspaper Exchange (Sports Training Science Branch) (3). [Publisher unknown], 2023:3.

[9] Tian Qian, Wu Jian, Zhao Dong. Prediction and trend evaluation of ancient landslide deformation based on neural network and multi-scale feature analysis [J]. Geodesy and Geodynamics, 2022, 42 (10): 1056-1062

[10] Huang Zhiying. Research on prediction model of tennis match results based on BP neural network [D]. Fujian Normal University, 2022.