Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 13, No. 1, 2024

Competitive Sports Research Based on Data Analytics and Machine Learning Models

Shuting Hou[1, *], Xile Lan[1], Huiyao Zhang[1], Xingying Luo[1]

[1]School of Computer and Electronic Information, Guangxi University, Nanning, China
[*]Corresponding author: 910800618@qq.com

Abstract: In contemporary competitive sports, data analysis and modeling applications have become indispensable tools for optimizing training and improving athletic performance. To effectively handle the variety and changing nature of matches, match data is combined to create models that can manage intricate match situations. In this paper, the data are normalized, missing values are filled by interpolation, and the outlier detection based on box diagram is used to complete the preprocessing. Furthermore, to assess player scores, 16 metrics are implemented, conventional machine learning models underwent training, and the top-performing LightGBM model is chosen, indicating the impact of each metric on the score. The results show that the LightGBM model is highly robust, precise, and accurate.

Keywords: Machine Learning, LightGBM, Regression Modeling.

1. Introduction

Developing and executing an effective training regimen is vital in contemporary sports. In the highly efficient sport of tennis, it's crucial to provide the athlete with a precise method at the appropriate moment. The complexity of this challenge escalates with the variation in weather patterns and the involvement of adversaries. As data gathering technology and computer science progress, the era of intelligence has emerged, characterized by the advent of data, algorithms, and computational capabilities[1]. Computer vision enables more precise tracking of athletes' movements, analysis of game footage, assessment of athletic abilities, and the provision of immediate feedback[2]. Through machine learning, sports can create tailored training advice for athletes. In the era of information technology, the fusion of digital media and sports is set to pave the way for novel avenues in sports, with the continual integration of artificial intelligence technology in the realm of sports [3].

2. Machine Learning

Machine learning involves an algorithm that assesses and forecasts an athlete's performance in matches through data-driven learning and enhancement. Through model training, a machine is capable of discerning patterns and consistencies in an athlete's tennis performance using past data, subsequently applying this insight to forecast and scrutinize upcoming games. Widely utilized algorithms in machine learning encompass Support Vector Machines, Random Forests, Neural Networks, and others. Such algorithms are adaptable to diverse tennis datasets, including elements influencing the game, its outcomes, the status of the player's match, strategic evaluations, and so on, to enhance precision in assessment and forecasting. Machine learning models, in contrast to conventional statistical models, possess parallel computing capabilities, enhancing the speed of handling extensive sports data and lessening the training duration and resources required for the model. Furthermore, machine learning models can utilize feature selection and dimensionality reduction methods to simplify data dimensions when dealing with diverse and multivariate data, thereby simplifying the training process. Furthermore, machine learning models are capable of identifying and handling complex, nonlinear data configurations, like building neural networks to dynamically investigate the nonlinear interplay among dimensional data, thereby enhancing the model's predictive precision by fine-tuning the possible change patterns in the data. The quintet of traditional regression frameworks includes the LightGBM Classifier, XGB Classifier, SVC, MLP Classifier, and Logistic Regression models [4-7]. This document aims to explore the application of machine learning algorithms in tennis games to evaluate the impact of various metrics on the scores and to offer valuable resources and guidance for training choices.

3. Results

3.1. Data processing

After the data are obtained, the outliers of the quantitative data are mapped. Replace the outlier with the average of the upper and lower data layers. Data conversion and standardization to better meet the requirements of machine learning algorithm. The standardization formula is as follows:

$$Z_{i,j} = \frac{x_{i,j} - u_j}{\sigma_j} \quad (1)$$

Where $Z_{i,j}$ are the normalized values, x_j is the original value, u_j is the mean value of the j-th feature, and σ_j is its standard deviation.

At the same time, standardization is based on the mean and standard deviation of each feature, so it is not affected by outliers. This makes us more robust when dealing with outliers in data, and will not skew the scaling results due to extreme values.

For the vacant data, we use difference filling, and get the vacant data by averaging the upper data and the lower data, so that we can better preserve the distribution and trend of the data. At the same time, when we fill in the missing values, we will keep the changing trend of the original data as much as

possible, so it has good fidelity to the time series data. Of course, we don't need to introduce additional parameters, but fill in the missing values through the trend of the original data, which also reduces the possibility of introducing noise or uncertainty.

3.2. Modeling

A model was created to capture the flow of play for every score in the game. The core goal of the model is to identify which player performs better in the game and evaluate their performance level. We need to first establish a system that can quantify the player's performance and game dynamics. This system should comprehensively consider the conventional score, serve advantage, continuous scoring and other factors.

In addition to being the server, the key reasons related to the player's victory are also related to the player's physical condition, immediate psychological state and personal level. In order to reflect the real-time scores of players, we introduced sixteen indicators, which are as Table.1

Table 1. 16 indicators and corresponding symbols

Three-level index	Symbol
The number of games won in the current set	X1
The score of this game is ahead of schedule	X2
Whether it is the server	X3
Whether the last point was scored	X4
The score of this match is ahead of schedule	X5
Serve or not (no touch)	X6
Return score (no touch)	X7
No touch score forehand and backhand	X8
Whether there is an error in the game	X9
Whether there was an unforced turnover in the game	X10
The ratio of the number of touches to the score	X11
The ratio of the chance to score when the opponent serves to the point actually scored	X12
Total mileage in the match	X13
Total chart miles in the last three points	X14
Last point mileage	X15
Serve real time pace	X16

3.3. Comparison of model results

Firstly, we quoted five classic regression models of machine learning to train them, and then compared the performance parameters of LGBM Classifier, XGB Classifier, SVC, MLP Classifier and Logistic Regression models to choose the most suitable model.

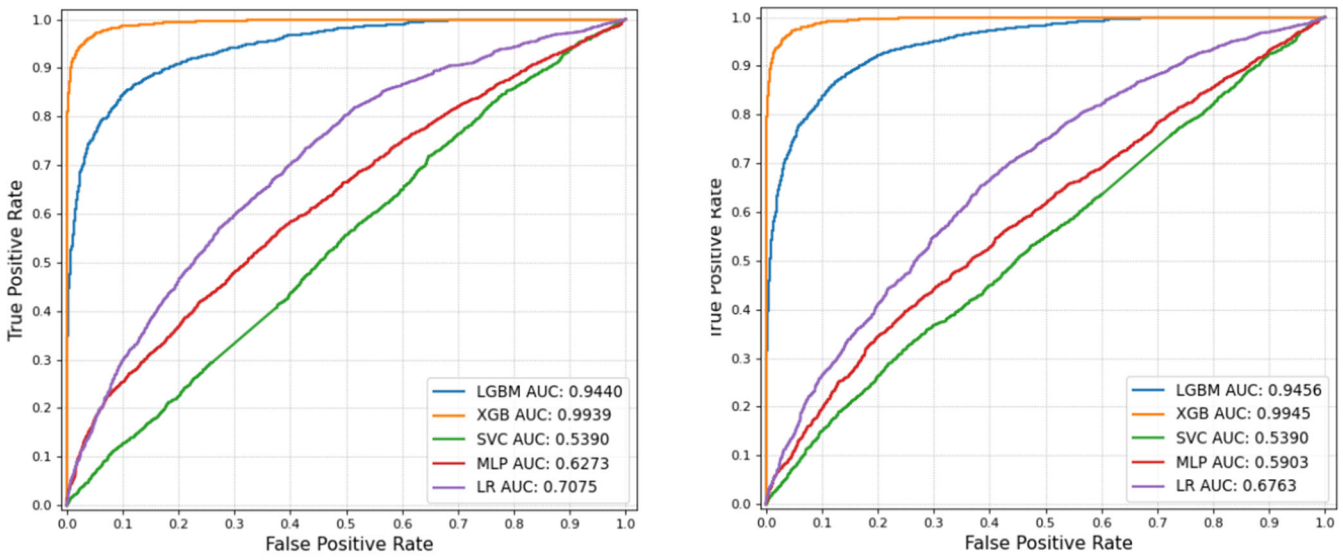

Figure 1. ROC curves for Player 1 and Player 2

In Figure 1, we list the ROC curves of these five models, which show the trade-off relationship between true positive rate and false positive rate under different thresholds. AUC in the figure represents the area under ROC curve, which is an important index to evaluate the performance of classifier. Under our assumption, the closer the AUC value is to 1, the better the model performance. Through the value of AUC in Figure 1, we will evaluate and compare LGBM and XGB.

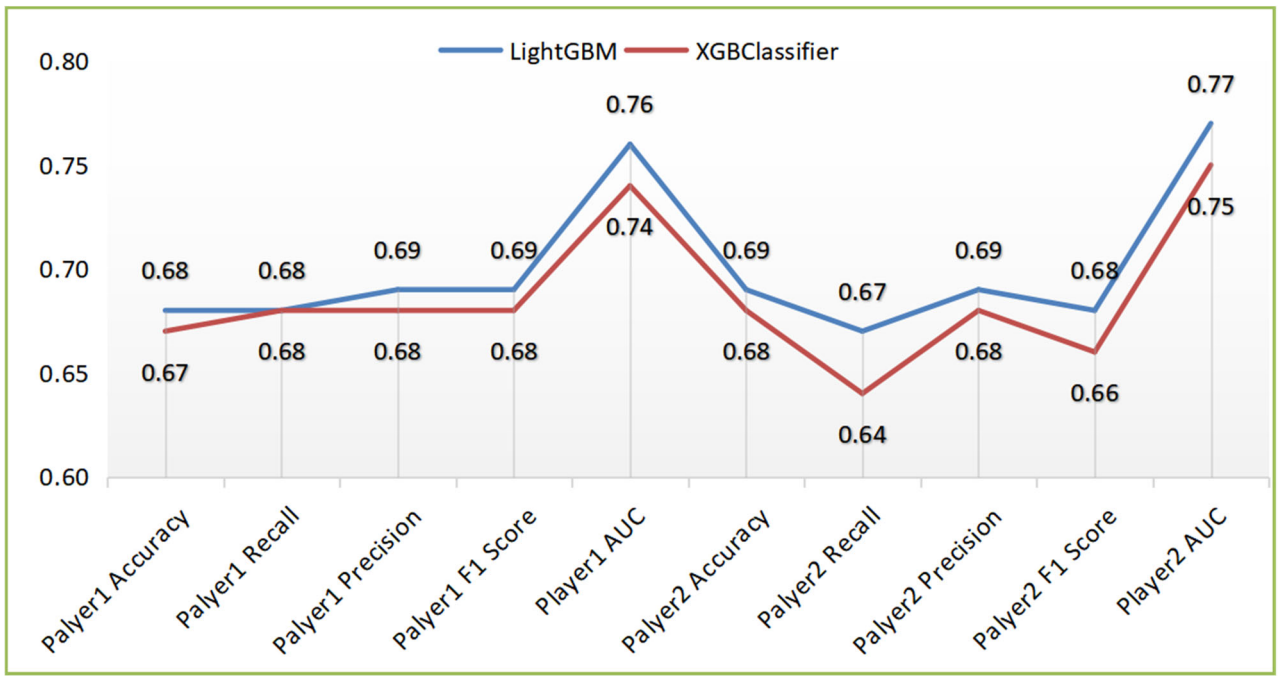

Figure 2. Comparison of X and Y indicators

In the line chart of Figure 2, we compare the accuracy, recall, accuracy, F1 score and AUC coefficient of Player 1 and Player 2, and through analyzing the relevant indicators of the above model, we finally adopt the LGBM with the best effect. LightGBM is a fast and distributed gradient lifting framework based on gradient lifting framework.

Suitable for large-scale data sets, LGBM Cassifier is a model for classification tasks in LightGBM library. (LGBM and XGB) For data sets with thousands of features, it can better handle and utilize the data information, which allows us to use a user-defined loss function to adapt to different problems and tasks, and increases the flexibility of the model. It can also be trained in parallel, and it is very advantageous for processing large-scale data sets by using multithreading to accelerate the learning process of the model. The model we trained is applied to the match of Bourdon gentleman's final, and the dynamic real-time scores of Player1 and Player2 are displayed in the match, which is convenient for us to analyze the competition process.

3.4. Analysis of experimental results

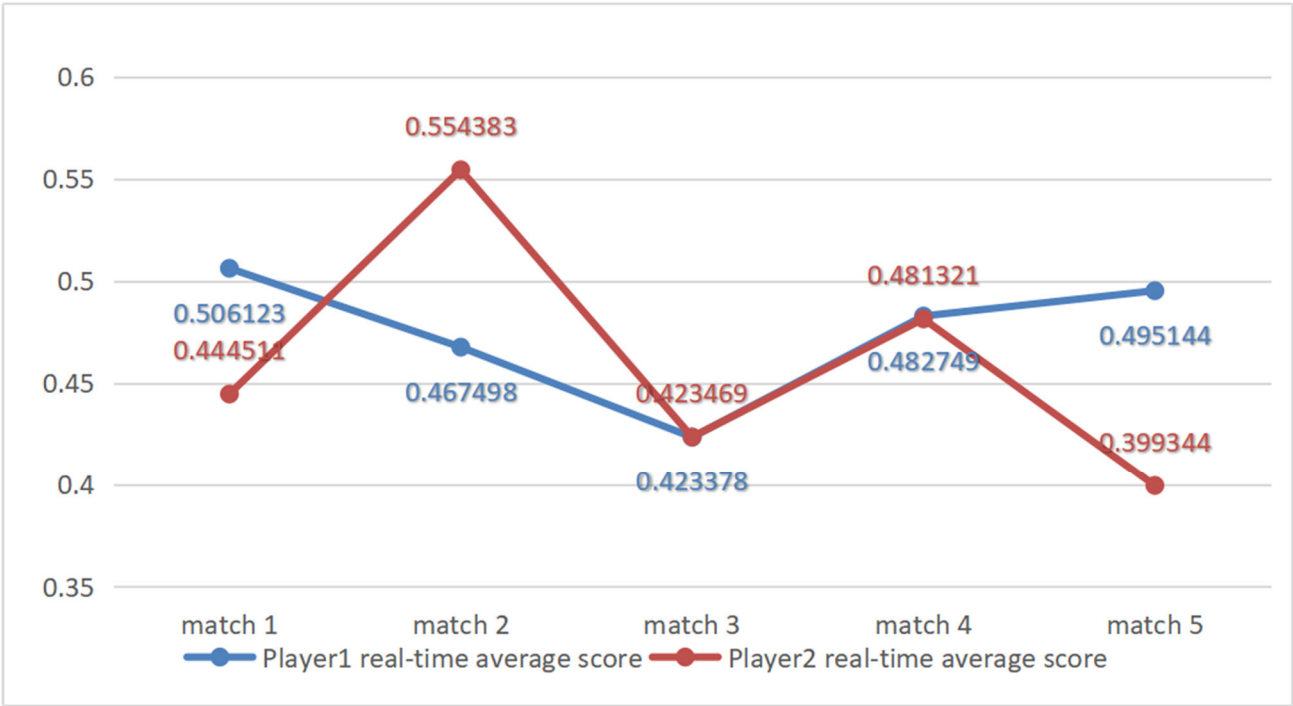

Figure 3. Average score of the two-player game

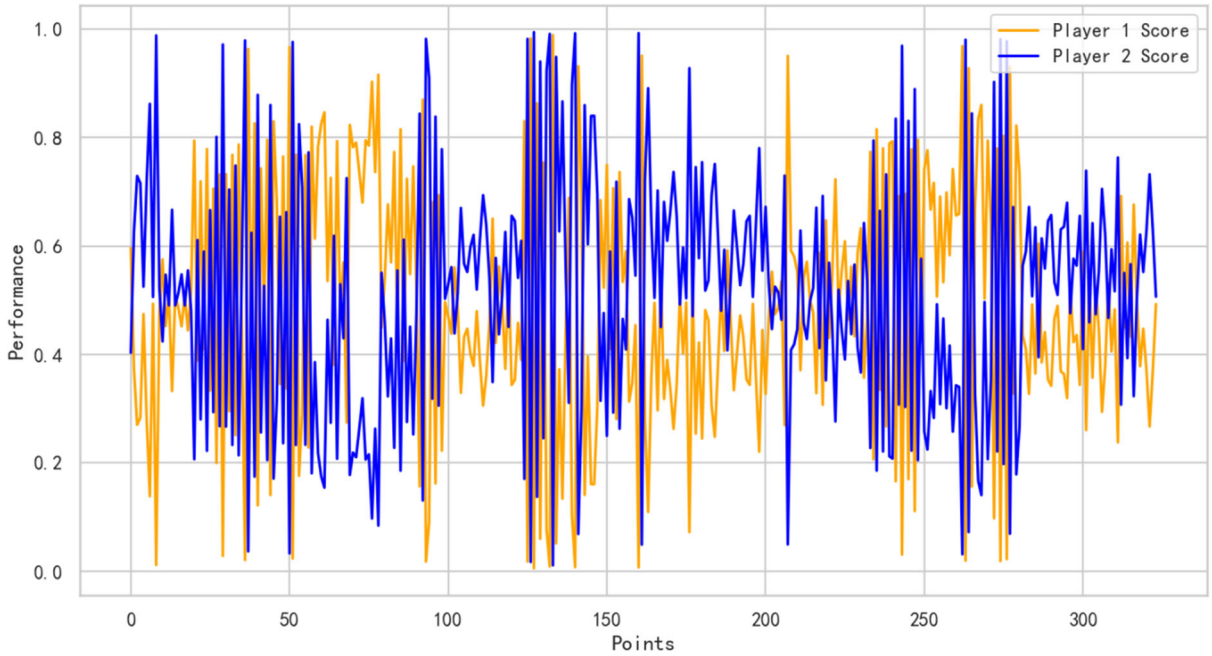

Figure 4. Real-time scores per game for Player 1 (left) and Player 2 (right) (one color represents one game)

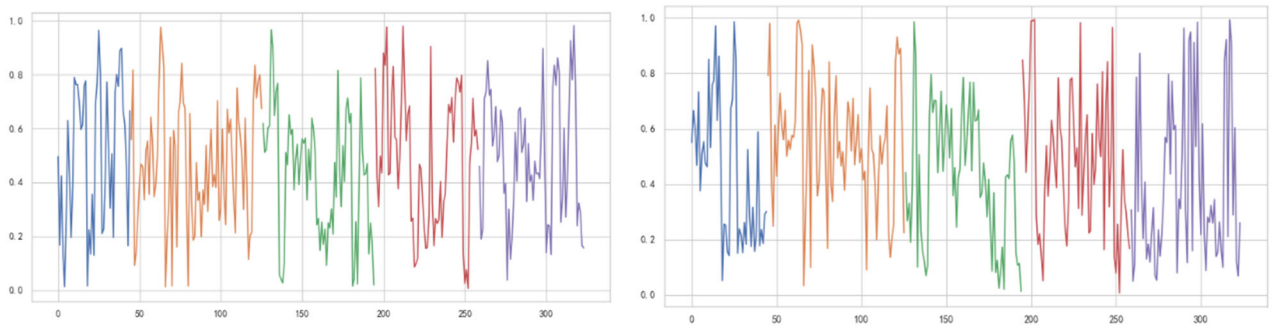

Figure 5. Dynamic real-time scores of the two players

Through the above three pictures, we can clearly see the score changes of the two players and their score performance trends. At the same time, according to the real-time score analysis of the two players, we also get the priority degree of the sixteen indicators in Table 1. From Figure 5, we can see that the total mileage in the last three point has the greatest influence on the real-time score of the players, while the forehand and backhand of the touchless score have the least influence, so we can make the following analysis according to this indicator.

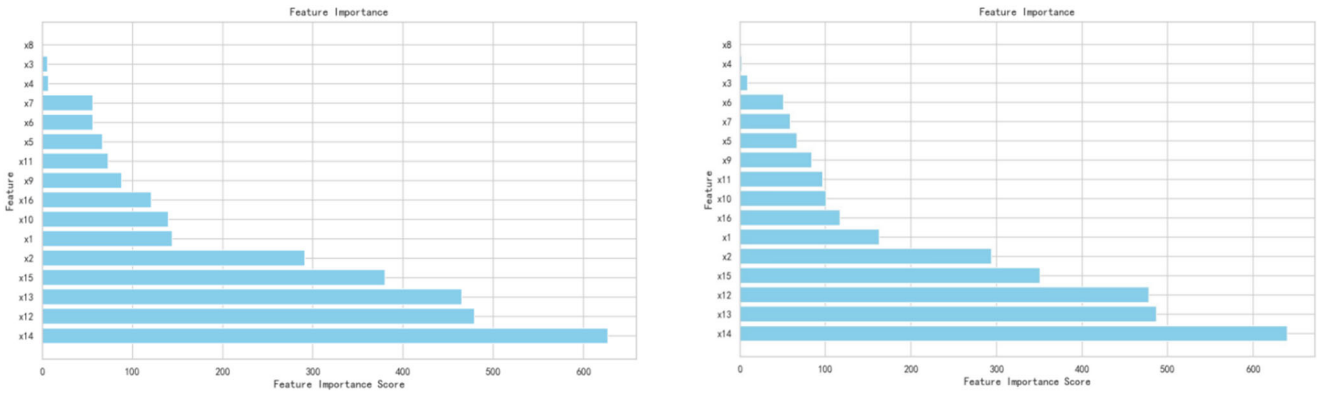

Figure 6. Influence level of each index for player 1 (left) and player 2 (right)

4. Conclusions

This research aimed to predict player performance by utilizing various metrics, including tennis players' scores and factors that influence these scores in the 2023 Wimbledon season. In order to achieve this, machine learning techniques were employed for data preprocessing, followed by training regression equations using code.

To begin, data on 16 metrics were collected and used to train traditional machine learning models. Among these

models, the LightGBM algorithm demonstrated the best performance and was subsequently selected for further analysis. Notably, LightGBM utilizes an efficient gradient boosting framework, which guarantees both high accuracy and scalability, making it ideal for handling large datasets. Additionally, the insights derived from this model can be leveraged to inform strategic adjustments for tournaments.

References

[1] Wei Mengli, Zhong Yaping, Gui Huixian, et al. A machine learning based sports injury warning model[J]. Chinese Journal of Tissue Engineering Research, 2025, 29 (02): 409-418.

[2] Kong Xiangshen, Chu Xiangtong, Gao Han, et al. Empowering Innovative Applications and Development of Sports with Digital Intelligence: A Review of the 14th International Symposium on Sports Computer Science[J]. Sports Research, 2023, 44 (06): 20-26.

[3] Gao Xuebo, Yang Jihong. Research on Swimming Training Method Based on Machine Learning[J]. Journal of Lanzhou University of Arts and Sciences (Natural Science Edition), 2024, 38 (03): 110-114.

[4] Aziz Rabia Musheer, Baluch Mohammed Farhan, Patel Sarthak, et al. LightGBM: a machine learning approach for Ethereum fraud detection[J]. International Journal of Information Technology, 2022, 14(7): 3321-3331.

[5] Sharma Sachin. Stroke Prediction Using XGB Classifier, Logistic Regression, GaussianNB and BernaulliNB Classifier [C]. 2023 International Conference on Circuit Power and Computing Technologies (ICCPCT). IEEE, 2023: 1577-1581.

[6] Mannino, Gaetano, et al. "Nonlinear and multivariate regression models of current and voltage at maximum power point of bifacial photovoltaic strings." Solar Energy 269 (2024): 112357.

[7] Song, Honglin, et al. "Using complex networks and multiple artificial intelligence algorithms for table tennis match action recognition and technical-tactical analysis." Chaos, Solitons & Fractals 178 (2024): 114343.