

Quantifying and Predicting Tennis Match Momentum: An LGBM-Based Analysis and Visualization Model

Zhenrui Li^{1,*,#}, Qisheng Zhang^{1,#}, Yixiao Wang^{2,#}

¹School of Economics, Shanghai University of Finance and Economics, Shanghai, China, 200433

²School of Mathematics, Shanghai University of Finance and Economics, Shanghai, China, 200433

*Corresponding author: Lzray1@163.com

#These authors contributed equally.

Abstract. In any competitive games, the most exciting thing for the audience is the change of the scores of both players, and quantifying the trend of the scores of both players can help coaches provide timely guidance to the players during the games and also help the players in the preparation for the games. The author firstly obtains the data of the players' psychological state, physical reserve and athletic skills during the game, and then compares the LGBM, SVC, MLP and LR models, analyzes the accuracy and regression rate of each model, and then selects the optimal LGBM model, and then calculates the weights of each parameter through the selected model, and sums them up to obtain the quantitative value of "momentum". Finally, the fluctuating images of the two sides' "momentum" in the whole game are drawn, through which the scoring trend can be visualized and the turning points of the momentum can be marked. The use of such methods is a kind of innovation for sports events, which can scientifically help coaches to make decisions, and can bring more exciting sports events to the audience to fully reflect the athletes' style.

Keywords: Momentum, LGBM, Multi-model comparison.

1. Introduction

In the 2023 Wimbledon Gentlemen's final, 20-year-old Spanish rising star Carlos Alcaraz versused 36-year-old Novak Djokovic. It was a great match, the two were neck and neck, and the match was very tight. Finally, Alcaraz gained the victory 6–4. The most memorable aspect of the match was the way the scores fluctuated incredibly, with players who appeared to be dominant one moment at a disadvantage the next, which begs the question of whether such "momentum" can be predicted?

In the discussion of momentum in tennis a number of scholars have come to a large number of conclusions, Dietl H, Nessler C[1] used binary logistic regression to show that consecutive points are an important factor affecting momentum. Philippe Meier et al.[2] in their study of psychological momentum from strategic momentum demonstrated that the break point is a factor that influences the change in total momentum by affecting mood and hence, momentum. In earlier studies, Den Hartigh Ruud J R et al.[3] demonstrated the significant influence of team psychology on momentum In these studies, only one to three factors were used to analyze psychological fluctuations, rather than analyzing and predicting the momentum of the entire match from a global perspective using a variety of physical and psychological factors. Therefore, the author has developed a mathematical model that captures the "momentum" of scoring changes and applies it to one or more games. The model also identifies which athletes are performing better at a given time in the game, and how well they are doing. It also visualizes the course of the game and predicts when the momentum of the game shifts from one player to another.

2. Materials and Methods

2.1. Data Acquisition and Preprocessing:

The data required for the mathematical model were obtained from the real-time data from the 2023 Wimbledon Men's Singles Final, and the obtained data were cleaned before data analysis by removing the negative values and blanks, and assigning the tie-break "AD" to the value of "50" for quantification and finally the data were normalized^[4].

2.2. Introduction to the model

LGBM is a machine learning model based on the gradient boosting framework. It employs a decision tree-based learning algorithm to generate a powerful predictive model by continuously iterating the training set^[5]. Compared with traditional GBM, LightGBM has higher accuracy and faster training speed. Its principle consists of two main aspects: GOSS (Gradient-based One-Side Sampling) and EFB (Exclusive Feature Bundling). GOSS speeds up training by sampling the absolute value of the gradient and retaining the important samples. EFB, on the other hand, reduces the training speed by bundling the features to reduce the memory footprint and computational complexity^[6].

SVC is a commonly used supervised learning model whose principle is based on finding the optimal hyperplane that can effectively divide different classes of data. One of the advantages of SVM is that it performs well in high-dimensional spaces and is suitable for dealing with complex datasets^[7]. It can effectively handle the case of linear indivisibility by introducing the kernel function, which improves the generalization ability of the model. In addition, SVM performs well on small-sample datasets and still maintains good performance for the case where the feature dimension is larger than the number of samples^[8].

MLP is a deep learning model based on neural networks, which consists of an input layer, a hidden layer and an output layer. The hidden layer learns complex patterns and features in the data by nonlinearly transforming the input signal through an activation function. The MLP continuously adjusts the weight parameters through a back propagation algorithm, allowing the model to gradually approximate the optimal solution. Advantages include the ability to learn nonlinear relationships, applicability to a wide range of data types and problems, ability to handle large-scale datasets, and relatively little reliance on feature engineering^[9].

LR is a classical classification algorithm that classifies samples by mapping linear combinations of features to probability values between 0 and 1 via a Sigmoid function. The advantages of the LR model include simplicity of computation, ease of understanding and interpretation, fast training speed, and good handling of sparse data. In addition, the LR model is suitable for binary classification problems and maintains good performance with high feature dimensions^[10].

Random Forest (Random Forest) is an integrated learning method that consists of multiple decision trees. In constructing each decision tree, Random Forest is trained by randomly selecting features and samples, and then the prediction results of each tree are synthesized to finally output the overall prediction results^[11]. The advantages of Random Forest include strong processing ability for high-dimensional data and large-scale datasets, high accuracy and generalization ability, the ability to effectively deal with missing values and outliers, and not easy to overfitting^[12].

Logistic regression is a machine learning algorithm based on integrated learning and consists of multiple decision trees^[13]. In constructing each decision tree, Random Forest is trained by randomly selecting features and samples, and then integrating the prediction results of each tree to make the final prediction. The advantages of Random Forest include significant effect on high-dimensional data and large-scale datasets, high accuracy and generalization ability, strong robustness to missing values and outliers, and not easy to overfitting^[14]. In addition, Random Forest can assess the importance of features, which is helpful for feature selection and model optimization.

2.3. Model Evaluation and Selection

First calculate the model accuracy of the above LGBM, Random Forest and Logistic Regression models for selection, in which the LGBM model has the highest accuracy of 76.6%, so the LGBM model is selected among these three. Further among LGBM, SVC, MLP, LR models are applied using 5 fold cross validation method for selection i.e. by calculating accuracy, recall, precision, f1 and auc. Finally the accuracy, recall, precision, f1 and auc values of LGBM are 0.76, 0.77, 0.76, 0.76, 0.76, 0.84 are the highest and hence LGBM model is selected.

3. Model Establishment and Results Analysis

3.1. LGBM Model Prediction

After selecting the LGBM model for learning, this study set the training set to 80%. Finally, the importance of eight selected factors for the target factor in data preprocessing was obtained, as shown in Figure 1.

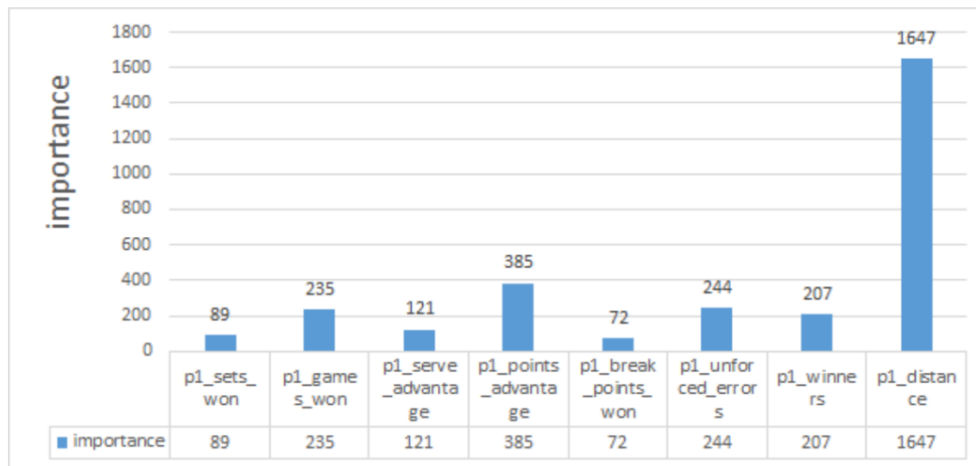


Figure 1. The importance of the selected factors

3.2. Weight Calculation Using LGMB Model

In this study, the importance indicators obtained from the LGBM model were weighted averaged to obtain the weights of different factors for the target factor. For example:

$$\text{Weight} = \frac{1647}{89+235+121+385+72+244+207+1647} \quad (1)$$

The weight for p1_games_won, p1_sets_won, p1_serve_advantage, p1_points_advantage, p1_break_points_won, p1_unforced_errors, p1_winners, p1_distance are 0.027, 0.070, 0.148, 0.114, 0.021, 0.072, 0.061, 0.487, with a total weight of 1.

3.3. Momentum Model Establishment

Subsequently, the model was established by setting the values of each factor as $\alpha_1, \dots, \alpha_8$, and their respective weights as $\beta_1, \dots, \beta_8$, defining the momentum value as M , and defining the disturbance term as μ_i . Then, the calculation formula was then derived:

$$M = \sum_{i=1}^8 \alpha_i \times \beta_i + \mu_i \quad (2)$$

The momentum calculation model was obtained thereafter.

3.4. Data Visualization

This study selected the Wimbledon men's singles final as a representative match for visualization, providing an intuitive understanding of the performance of both players at specific moments during

the match, and determining who performed better, thus completing the basic establishment of the model.

Figure 2 shows the momentum changes of both sides in the final, with the green side representing Carlos Alcaraz and the blue side representing Novak Djokovic.

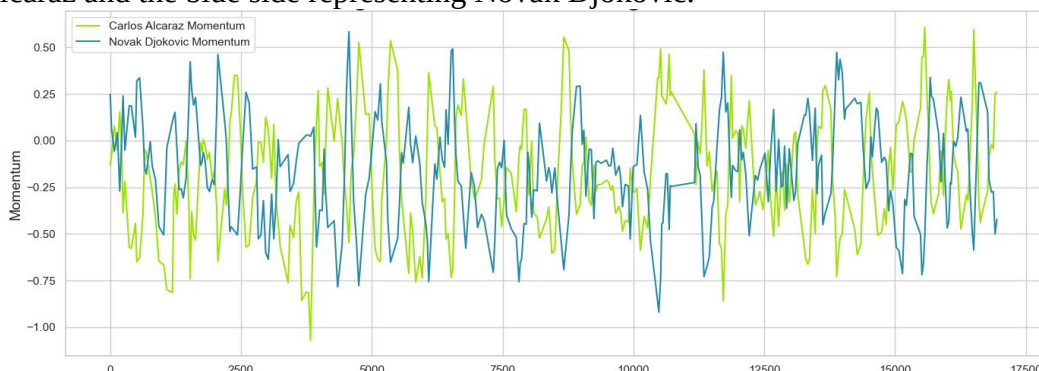


Figure 2. The change of momentum in 1701

From the figure, it is evident that when one side has higher momentum, they tend to perform relatively better. The magnitude of the values quantifies the performance of the players.

This study also analyzed data from multiple matches, obtaining similarly intuitive views, as shown in Figures 3 and 4.

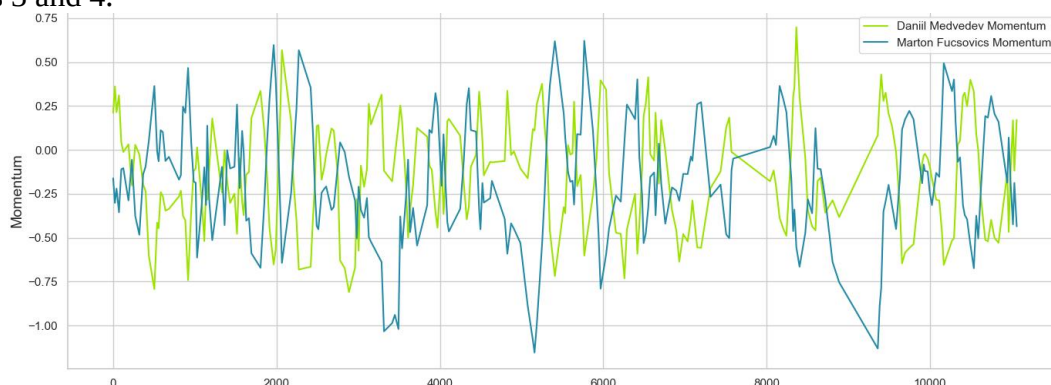


Figure 3. The change of momentum in 1305

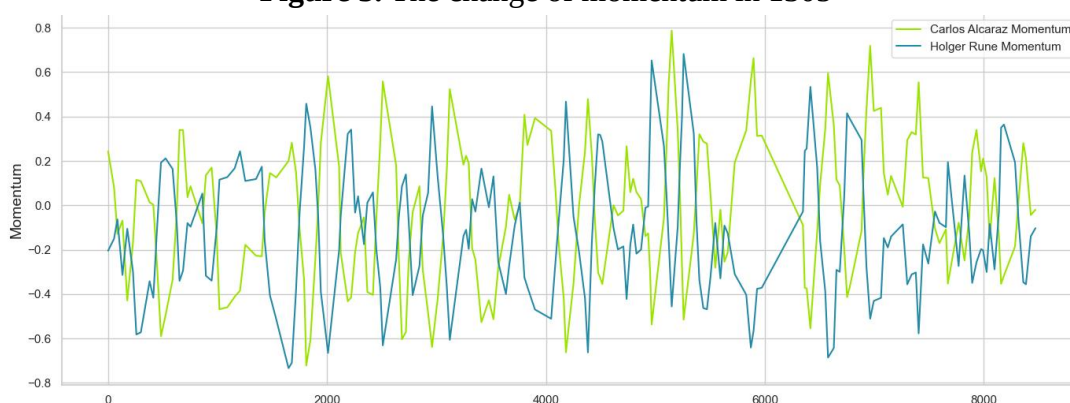


Figure 4. The change of momentum in 1501

3.5. Visualization Data Smoothing

To address the issue of significant fluctuations in momentum during matches, this study used a program to smooth the visualization, making the changes in momentum clearer. Figure 5 shows the smoothed view of the final match.

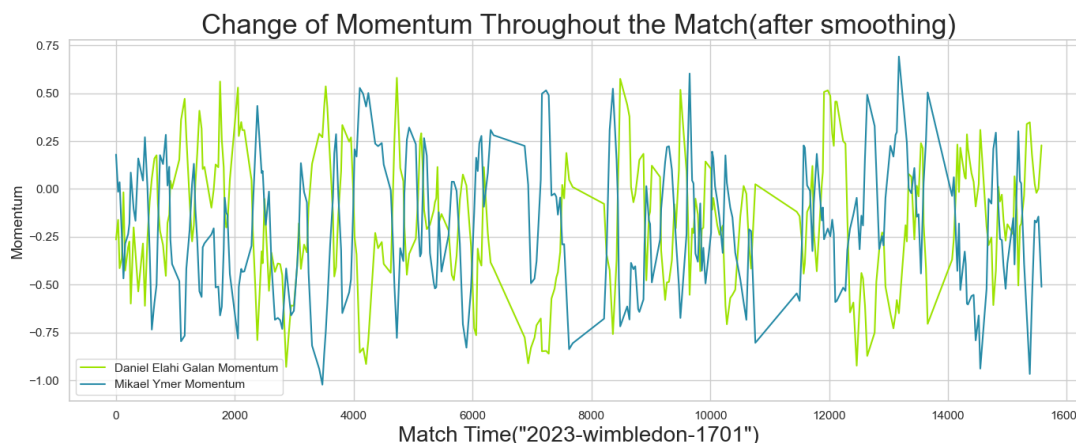


Figure 5. Change of Momentum in 1701(after smoothing)

3.6. Match Outcome Prediction Model

Using the momentum calculation model, this study calculated the momentum values of both players at the same moment during the match based on player data, comparing which side had a relatively larger momentum value. The player with greater momentum should theoretically win the point. This prediction was then compared with the actual score results to determine the accuracy of predicting whether a player scored.

3.7. Actual Prediction and Model Verification

3.7.1 Using Match Outcome Prediction Model

Continuing with the final match, this study utilized the match outcome prediction model, achieving a high prediction accuracy of 72.4%, successfully predicting 242 out of 334 scoring points, as shown in Figure 6(a).

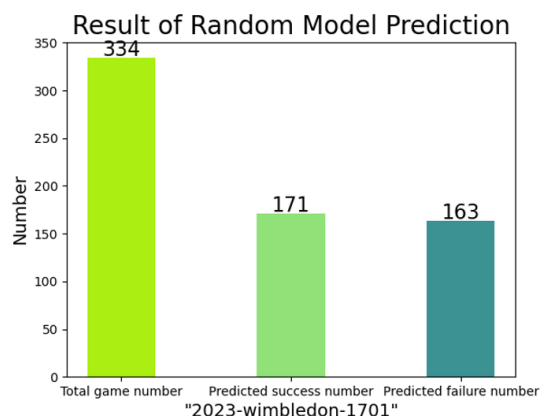
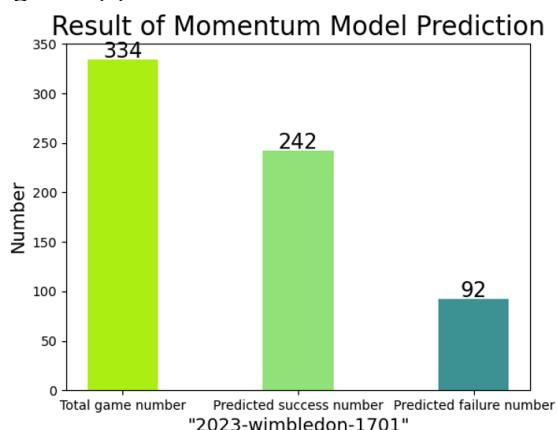


Figure 6. Result of Momentum Model Prediction(a) and Result of Momentum Model Prediction(b)

3.7.2 Using Random Prediction Model

Next, the study employed a random prediction model, using a 0-1 random distribution to predict the match outcome, resulting in an accuracy of 51.1%, successfully predicting 171 out of 334 scoring points, as shown in Figure 6(b).

3.7.3 Comparison with Other Match Models

Additionally, this study researched two other matches, with the results shown in the figures 7 and 8.

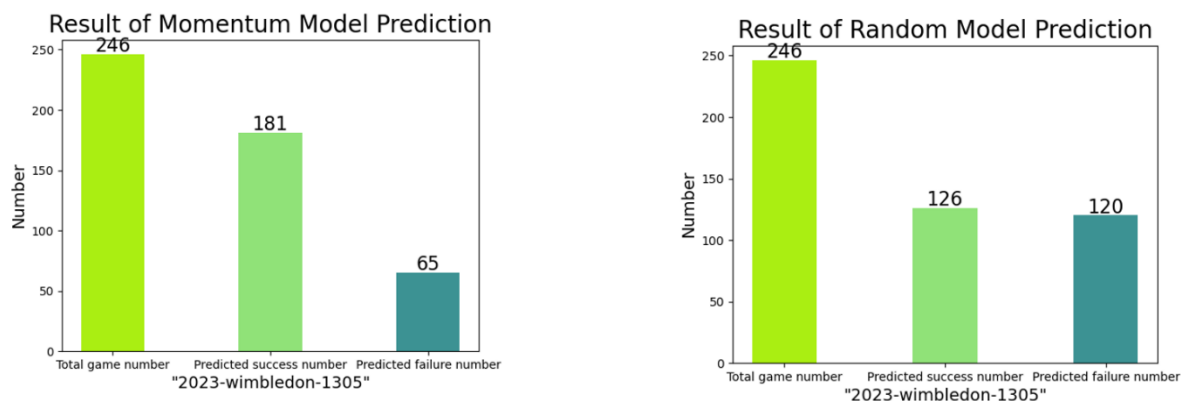


Figure 7. Result of Momentum Model Prediction and Result of Momentum Model Prediction

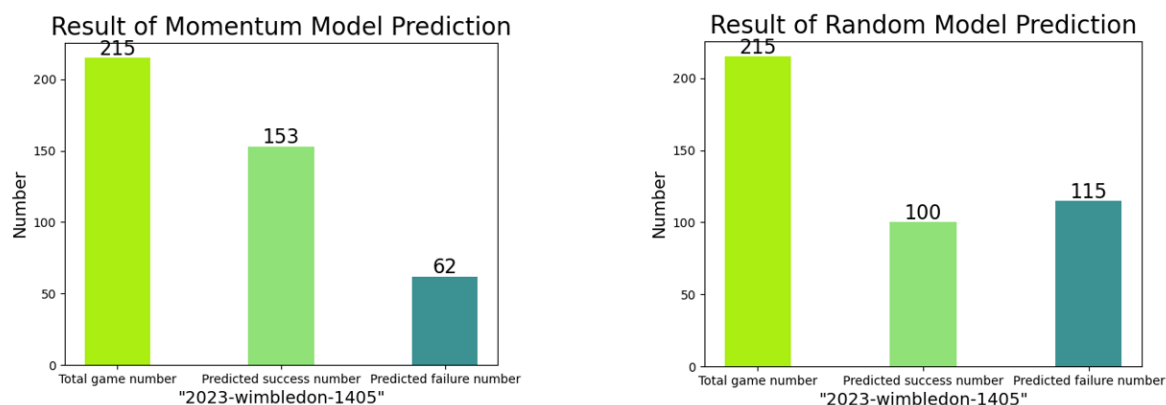


Figure 8. Result of Momentum Model Prediction and Result of Momentum Model Prediction

3.7.4 Summary

It is evident that the use of the momentum model achieved a prediction success rate of over 70% for the entire match, while the use of the random model resulted in a success rate of around 50%. This demonstrates the effective predictive ability of the momentum calculation model in forecasting the scoring process of tennis matches, indicating a significant correlation between scoring and various factors.

3.7.5 Secondary Validation

Furthermore, this study conducted a secondary validation of the model using LGBM machine learning. After learning from the training set data, LGBM formed a set of predicted data through fitting, which was then analyzed against the true data in the test set to calculate accuracy, precision, recall, and F1, as shown in Figure 9.

```
The accuracy of the model is: 0.8141410674446542
The precision of the model is: 0.8140635564570656
The recall of the model is: 0.8187691261475688
The f1 of the model is: 0.8164095609425326
The confusion matrix result:
```

Figure 9. value of the test

To examine the statistical situation of the predicted values and true values, a confusion matrix was created, as shown in Figure 10.

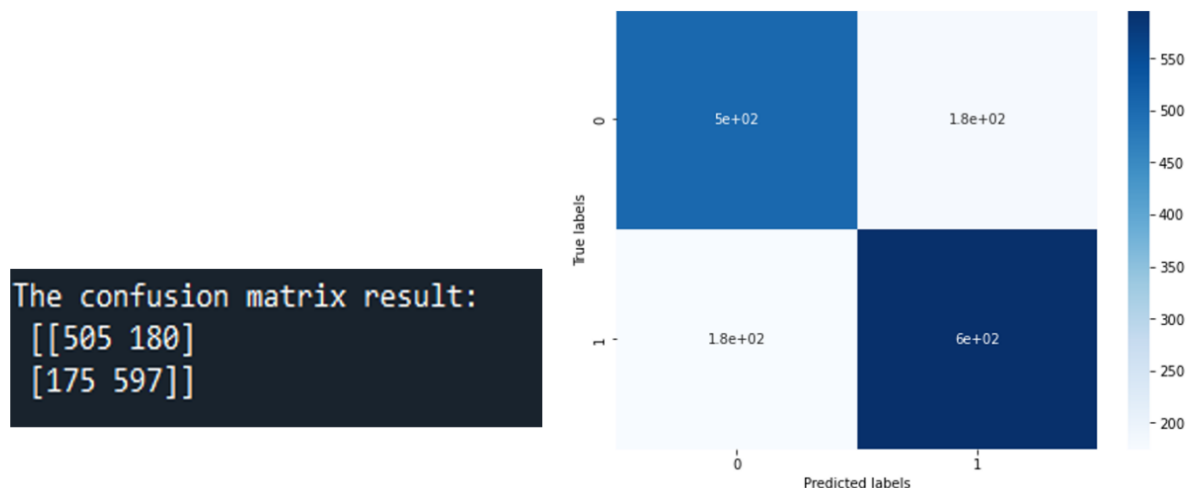


Figure 10. Confusion matrix and heat map of the confusion matrix

From the above, it is clear that the match outcome prediction model of this study achieved an accuracy of nearly 80% after machine learning, surpassing the probability achievable by randomness. Therefore, the model possesses predictive significance from a statistical perspective.

3.8. Turning Point Analysis

In analyzing when one player gains an advantage over the other during a match, this study attempted to extract moments when the relative momentum values of both players changed, as shown in Figure 11. Still using the data from the Wimbledon final, the analysis identified points in time where the momentum relationship between the two players changed.

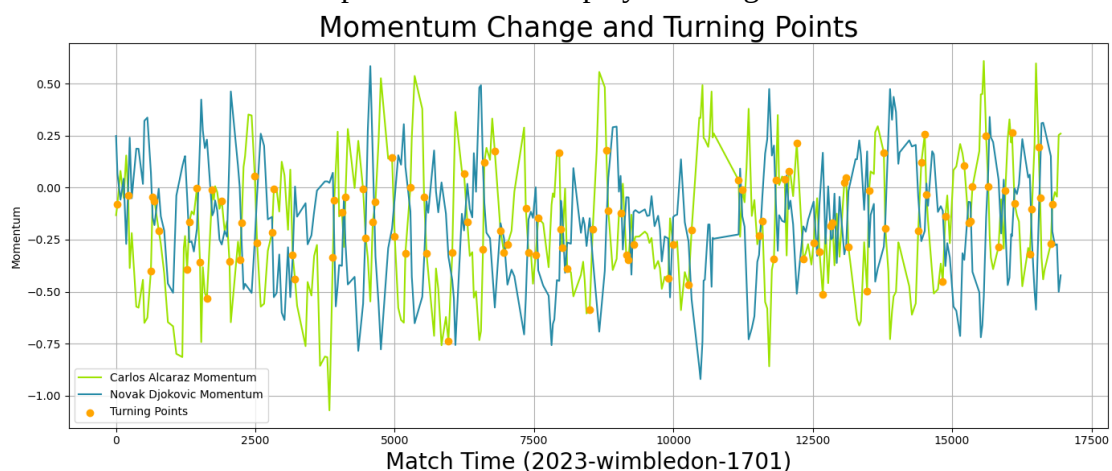


Figure 11. Momentum Change and Turning Points

This analysis of changes in momentum relationships can significantly assist coaches in determining when a match is about to shift from favoring one player to favoring another.

4. Discussion

This paper computationally proves that the momentum of a tennis player on the court can be computed and predicted using various machine learning models. This paper develops a mathematical model that can capture the momentum of score changes and apply it to one or more matches. The model can also determine which athletes are performing better at certain times during a match and how well they are performing. It can also visualize the course of a game and predict when the momentum of a game shifts from one athlete to another.

The application of machine learning in match result prediction has received a lot of attention, but previous articles usually use logistic regression and random forest for match result prediction. In this

paper, by comparing several machine learning models, it is concluded that the LGBM model has a higher prediction accuracy in tennis match result prediction. This result explains the feasibility of LGBM models for match result prediction. This finding increases the diversity of match result prediction methods and also expands the application areas of LGBM models.

This paper uses the match data of 2023 Wimbledon Men's Open as the research sample, establishes the basic tennis players' game momentum calculation model and prediction model, the numerical weights of different players' momentum calculation model represent the difference of individual ability or habitual strategy, according to the size of the weights to judge their own and their opponents' game styles, which provides a greater reference significance for the coaches to guide the players' actual game training.

Other factors, such as the type of field and the gender of the athlete, have been identified in many previous professional studies as having an impact on the athlete's play and strategy. However, these factors were not considered in the modeling process of this paper. Therefore, if the model in this paper is to be applied to other competitions, it may be necessary to consider the impact of different types of competitions on the weighting of the outcome prediction model building.

Due to the interoperability of ball game rules and techniques, the results established in this thesis have the potential to be extended to the game process of other ball games, such as table tennis and badminton, and thus fulfill their greater application value.

5. Conclusion

This paper takes the performance of tennis players in the game and the results of the game as the research object, and on the basis of the realistic data of the tennis game, compares the prediction accuracy of different machine learning methods, and selects the LGBM model as the method of calculating the weights, and establishes the calculation model of the tennis player's game momentum.

Compared to logistic regression and random forest, the LGBM model performs better in predicting outcomes. The accuracy of the model is as high as 76.6%. Based on the calculation results of the LGBM model, we get the eight parameters and their corresponding weights that have more influence on the prediction of the game, and complete the construction of the momentum model. Putting the real values and random values into the prediction model, we can find that the momentum has a correlation with the game situation, which is statistically significant for prediction.

The tennis players' momentum calculation and prediction model based on the LGBM model can effectively help athletes to formulate more scientific tactical strategies and training focuses during matches. However, the model constructed in this study also has certain limitations, and in subsequent studies, the sample data set can be further expanded, more factors can be considered, and more in-depth statistical analyses can be conducted to provide more accurate and targeted pre-match tactical advice.

References

- [1] M. H D ,Cornel N .Momentum in tennis: Controlling the match[J].INTERNATIONAL JOURNAL OF SPORT PSYCHOLOGY,2017,48(4):459-471.
- [2] 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(prepublish):
- [3] R J R H D ,Christophe G ,W N Y V , et al.How psychological and behavioral team states change during positive and negative momentum.[J].PloS one,2014,9(5):e97887.
- [4] <https://www.independent.co.uk/sport/tennis/novak-djokovicwimbledon-final-carlos-alcaraz-b2376600.html>
- [5] LI H.Statistic learning method(the second edition)[M].Beijing:Tsinghua University Press,2019.
- [6] Drin S. Forecast model of the price of a product with a cold start [R].2024.

- [7] El Mrabet M A, El Makkaoui K, Faize A. Supervised machine learning: A survey[C]//2021 4th International Conference on Advanced Communication Technologies and Networking (CommNet). IEEE, 2021: 1-10.
- [8] Ünal T. Predicting tennis match outcome: a machine learning approach using the SRP-CRISP-DM framework[D]. Middle East Technical University, 2023.
- [9] Naskath J, Sivakamasundari G, Begum A A S. A study on different deep learning algorithms used in deep neural nets: MLP SOM and DBN[J]. Wireless Personal Communications, 2023, 128(4): 2913-2936.
- [10] Gu W, Saaty T L. Predicting the outcome of a tennis tournament: Based on both data and judgments[J]. Journal of Systems Science and Systems Engineering, 2019, 28: 317-343.
- [11] Ingram M. A point-based Bayesian hierarchical model to predict the outcome of tennis matches[J]. Journal of Quantitative Analysis in Sports, 2019, 15(4): 313-325.
- [12] Randrianasolo A S, Pyeatt L D. Comparing Different Data Representations and Machine Learning Models to Predict Tennis[C]//Future of Information and Communication Conference. Cham: Springer International Publishing, 2022: 488-500..
- [13] Friligkos G, Papaioannou E, Kaklamanis C. A framework for applying the Logistic Regression model to obtain predictive analytics for tennis matches[J]. 2023.
- [14] El Mrabet M A, El Makkaoui K, Faize A. Supervised machine learning: A survey[C]//2021 4th International Conference on Advanced Communication Technologies and Networking (CommNet). IEEE, 2021: 1-10.