ISSN 2616-5775 Vol. 7, Issue 6: 88-94, DOI: 10.25236/AJCIS.2024.070614

Momentum Analysis Based on LightGBM and SHAP Methods

Yanwen Chen[1,*], Guoxuan Sun[1], Lei Ge[1]

[1]*College of Business, Hohai University, Nanjing, China*
[*]*Corresponding author: 2263310120@hhu.edu.cn*

Abstract: *Carlos Alcaraz sparked the tennis world's attention when he defeated the legendary Novak Djokovic in the men's singles final at Wimbledon 2023. This study utilized a unique data processing method to construct a LightGBM regression model to successfully predict the match outcome. Through SHAP analysis, it was found that the distance a player moves is crucial to the match outcome. The study demonstrated that proper utilization of technical and strategic resources can enhance player performance. This study provides new perspectives for understanding momentum changes in matches and emphasizes the potential of data-driven predictive models in sports competitions. For policy makers and researchers in tennis and other fields, this study provides useful insights and methods.*

Keywords: *Momentum, LightGBM, SHAP Analysis*

1. Introduction

Momentum is a key factor in sports competition and can make the difference between victory and defeat in a rapidly changing match [1]. The dramatic ending of the men's singles final at Wimbledon 2023 was a turning point in tennis history, as a young Carlos Alcaraz defeated the legendary Novak Djokovic, ending the latter's decade-long reign. The match not only exemplified the unpredictability common in sports competition, but also highlighted the importance of momentum in the game [2].

This study is dedicated to utilizing data science and machine learning techniques to delve into the impact of momentum on match outcomes. Through unique data processing methods and modeling techniques, we constructed a regression model based on LightGBM to successfully predict the outcome of the men's singles match at Wimbledon. The key feature analysis reveals that the distance a player moves on the court is crucial to match performance, triggering deeper thinking about technology and strategic resources.

With this study, we hope to bring new insights and approaches to the field of competitive sports, revealing the underlying factors behind momentum changes and providing better decision support for athletes and coaches. This study has important implications for understanding dynamic changes in competition and optimizing athlete performance, providing useful insights for sports science and data-driven decision making.

2. Data analysis and data preprocessing

2.1 Processing of missing values

We found that there are missing values in the data, and the column names and numbers of the filtered missing values are as shown in Table 1.

Table 1: The column name and number of missing values

Column Name	Number
Speed_mph	752
Serve_width	54
Serve_depth	54
Return_depth	309

Assuming that most of the missing values are caused by non-scene factors, and that the missing values are determined by the serving side, then the missing values can be filled in with the mode of the serving

ISSN 2616-5775 Vol. 7, Issue 6: 88-94, DOI: 10.25236/AJCIS.2024.070614

side in the column where the missing values are located.

Screen out all the contestants.

Filter each player's speed_mph, serve_width, serve_depth, and return_depth as the serving side, discard rows with null values, and save the results separately.

The saved content is calculated to obtain the mean, mode, and median speed_mph, serve_width, serve_depth, and return_depth of each player as the serving side.

Deal with the category column first. For rows with null values in original data, determine who is the server, and filter the mode corresponding to player and fill in the corresponding position. Then handle the missing value of the continuous variable of serving speed.

2.2 Processing of outliers

After the above data processing, a box plot is drawn to filter for outliers in each column. As shown in Figure 1.

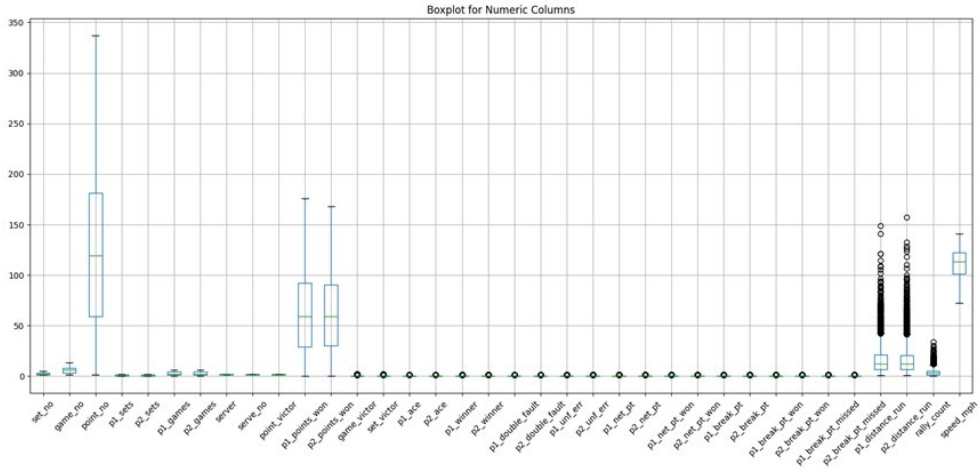

Figure 1: Box plot of all data list

As can be seen from the box plots of all the above data, there are several columns of data with outliers that exceed the upper edge of the box plot, indicating that there are outliers. We took out the columns with outliers separately and created a new box plot for better observation and analysis.

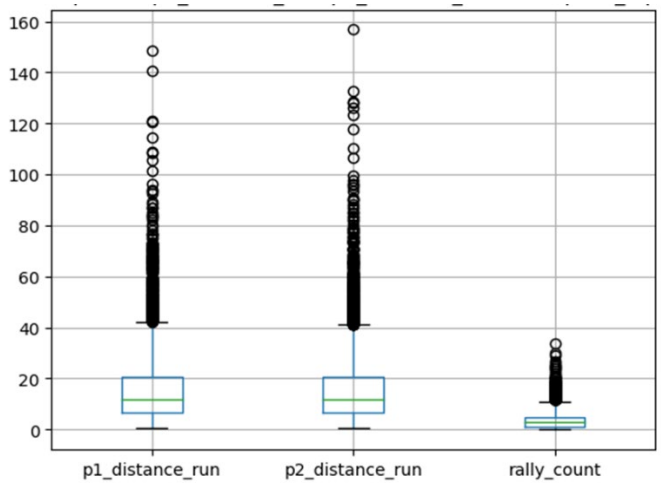

Figure 2: Box plot of data list with outliers

Looking at the above Figure 2, it is found that there are abnormalities in the three columns of "p1_distance_run", "p2_distance_run", and "rally_count". Based on the authenticity of the data, there may be a large gap in the physical fitness and playing strategy of the players, so it is reasonable that there is a significant difference in running distance. The number of catches is affected by both players, and the uncertainty is greater, so these outliers can also be con-sidered reasonable. So, we take the approach of

ISSN 2616-5775 Vol. 7, Issue 6: 88-94, DOI: 10.25236/AJCIS.2024.070614

preserving outliers.

2.3 Splitting of datasets

However, considering that each round of each game in each game has a serving side, and whether a player serves or not affects the description of the player's behavioral attributes, in tennis, the strategy and performance of the serving and receiving sides are usually different, so the two players are divided into two independent data sets, one specifically for the serving side (server_data) and the other for the receiving side (returner_data).

To this end, we have carried out the following:

Determine public and unique columns: The columns (common_columns) common to both datasets contain basic information about the tournament, such as tournament ID, player name, match time, etc., which are the same for both the serving and receiving sides.

Columns (server_unique_columns) unique to the serve, i.e., information related only to the serve, such as speed, direction and depth of the serve. There is only one column (returner_unique_columns) unique to the receiver in this scene, which is the return depth.

Divide dynamic columns: Dynamic columns (dynamic_columns) are those that change depending on the side served, such as scores, winning points, etc. The data for these columns depends on whether the current serving side is No. 1 or No. 2. While processing the data, I check the server column in each row to determine the current tee. If the value of server is 1, it means that player 1 is the serving side, and I will keep all the columns that start with p1_ distance_run and add the data from those columns to the data set of the serving side. Correspondingly, the data of the columns starting with the p2_ distance_run is added to the receiver's dataset. If the value of server is 2, it means that player 2 is the serving side, and the processing logic is reversed.

Construction of datasets and output of results: Traverse each row of the original dataset, extract the corresponding data based on the serve, and build two new DataFrames to store the data for the serving and receiving sides. During the construction process, in order to ensure the consistency of the column names, the prefixes of p1_ distance_run or p2_ distance_run were removed to make the final dataset clearer.

3. Prediction modeling

3.1 Introduction of LightGBM

LightGBM is a novel GBDT (Gradient Boosted Decision Tree) algorithm proposed by Ke in 2017 [3]. One of the characteristics of LightGBM is the use of a Histogram-based decision tree algorithm，its principle is shown in Figure 3. Another feature of LightGBM is to adopt a more efficient leaf growth strategy, namely the leaf growth strategy with depth limitation, its principle is shown in Figure 4. Therefore, LightGBM has the advantages of fast training speed, high accuracy, low memory consumption, and support for parallel computing, which can be used to solve the problems encountered by GBDT in massive data processing [4].

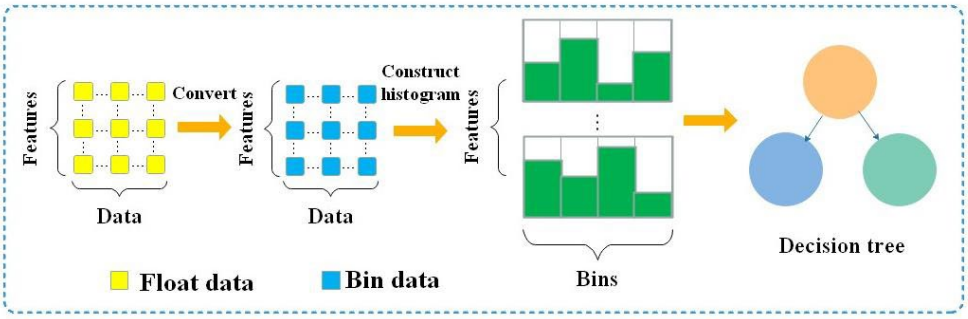

Figure 3: Histogram-based decision tree algorithm

ISSN 2616-5775 Vol. 7, Issue 6: 88-94, DOI: 10.25236/AJCIS.2024.070614

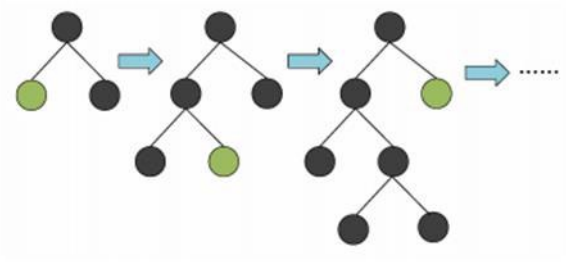

Figure 4: Schematic diagram of leaf-wise tree growth

3.2 Data description and preprocessing

The main objective of this model is to identify whether technical indicators in a tennis match affect the volatility of the match and to explore the mechanism of their influence. Therefore, the model is labeled with the volatility of the match at the current time, and the volatility we quantify by the difference between the player's momentum at the current time and the momentum at the previous point, denoted as pct_t.

Considering the technical indicators of the previous momentum difference and the lagged effect of the label itself on the current label, our feature sequence consists of the 1-period to 5period lagged terms of all the attribute indicators involved in the dataset, as well as the current value of the momentum difference and its 1 -period to 5 -period lagged terms. Denote all the attribute indicators of the given momentum difference in period t as X, then the model inputs include: $pct_{t-1}, pct_{t-2}, \dots, pct_{t-5}; X_t, X_{t-1}, \dots, X_{t-5}$.

In order to ensure the validity and reliability of the model, the dataset needs to be preprocessed. A very important step is to process the categorical variables in the attribute metrics are processed with unique hot coding, which converts each value of these metrics into a new binary column. It is also necessary to normalize the value of each attribute and map the data to $[0,1]$. The methodology is as follows:

$$v_s = \frac{v - v_{\min}}{v_{\max} - v_{\min}} \tag{1}$$

Where v_s is the normalized value, v is the original value, and $v_{\max}$ and $v_{\min}$ are the maximum and minimum values of the attribute, respectively. The normalized data is collectively referred to as $Data_{input}$.

3.3 Model results and evaluation

We input a sequence of features and train the model with a ratio of 7:3 between the training set and the test set. Then, through a large number of historical data correlation calculations, different weights are assigned to different features, and these weights are used to output estimates (or predictions) of the momentum difference.

In order to verify the reliability of the LightGBM model, we use MAPE (the mean absolute specific error) and the R2-Score scores on the training and test sets as a measure of model accuracy, and the results of each metric within the training and test sets are calculated as shown in Table.2.

Table 2: LightGBM model accuracy test results

	MAPE	R2_Score
Training set	1.579	

Taken together, the MAPE for both the training and test sets indicates that the model has low prediction errors on both the training and test data, while the r2_score for the test set indicates that the model also explains the target variables relatively well. We can easily assert that the model results match the actual data very well, and this conclusion can also be drawn from the scatter plot of predicted versus observed values below, as shown in Figure 5.

ISSN 2616-5775 Vol. 7, Issue 6: 88-94, DOI: 10.25236/AJCIS.2024.070614

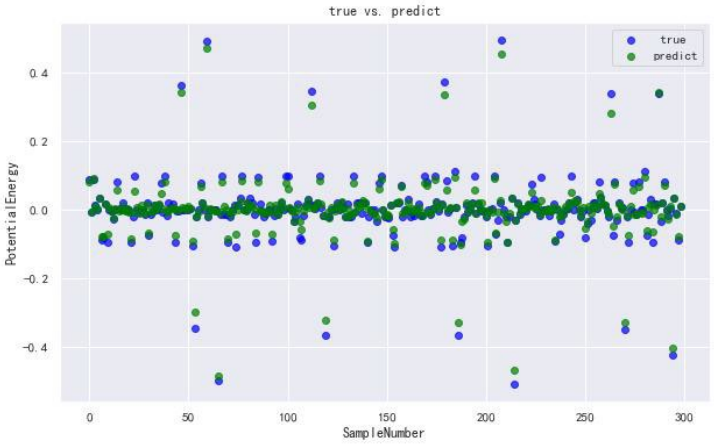

Figure 5: Scatter plot comparing true and predicted values

Due to the reliability of the model, we can use the weight of each feature to determine whether the player's technical indicators affect the volatility of the game. However, the feature importance of the LightGBM model may not be stable enough at times due to certain errors. This is because the gradient boosting model is a tree-based integration method, and its results are affected by the tree generation process and randomness. This stochasticity may lead to different structures of the generated trees in different runs, thus affecting the feature importance, and there may also be strong correlation between different technical metrics, so we chose to utilize explanatory tools such as SHAP to provide more stable and interpretable feature importance.

3.4 Interpreting model predictions using the SHAP model

SHAP builds a linear model based on the game-theoretically optimal Shapley value, which can be used to explain the output of any machine learning model [5]. For each sample, the machine learning model gives a predicted value, SHAP considers all features as "contributors", and the Shapley value is the value assigned to each feature in that sample.

We can understand this by assuming that the $j\text{-}th$ feature of the $i\text{-}th$ sample takes the value of $x_{i,j}$, the predicted value of the machine learning model for the $i\text{-}th$ sample is $\hat{y}_i$, and the base value of the model is $\phi_{i,j}$, then the following equation holds:

$$\hat{y}_i = \phi_0 + \sum_{j=1}^{m} \phi_{i,j} = \phi_{i,1} + \phi_{i,2} + \phi_{i,3} + \cdots + \phi_{i,m} \quad (2)$$

The basic idea of SHAP value is to consider the mean value of the marginal contribution of a feature to the model's predicted value when it is added to the model [6]. The reason why it is a mean value is that for m features, we have 2^m ways of combining the features, and for all of the ways of combining the features, we need to compute the marginal contribution of the feature to the predicted value of the model when it is added to the model.

We interpret and analyze the prediction results of the trained LightGBM model based on the SHAP interpretation method.

Step 1: For the interpretation of individual samples, we picked the example in the dataset with sample index 3860. For each prediction sample, the model produces a prediction value, and the SHAP value is the value assigned to each feature in that sample. We can understand this in such a way that the SHAP value of each feature variable reflects its contribution to the prediction result, and each feature variable drives the model prediction result from the baseline to the final prediction result through its SHAP value. We visualize the results as shown in Figure 6.

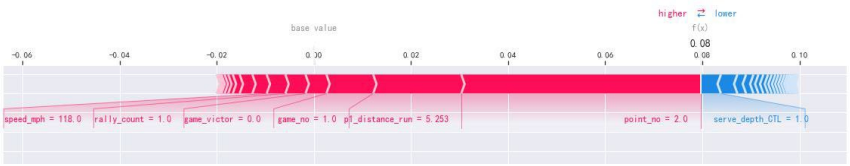

Figure 6: One sample interpretation of the graph

ISSN 2616-5775 Vol. 7, Issue 6: 88-94, DOI: 10.25236/AJCIS.2024.070614

Analyzing this chart specifically, speed_mph significantly decreases the model's predictions, which is the negative feature that has the most impact. point_no and serve_depth_CTL are the features that have the most positive impact, significantly increasing the model's predictions.

Step 2: Based on the single-sample SHAP values, the SHAP of global features can be aggregated to obtain the SHAP of the uniquely heat-coded metrics. The following Figure 7 shows a graph of feature importance analysis with aggregated SHAP values, which shows the importance of each feature in the model decision. The length of the bars in the chart indicates the importance of the feature, with more frequent lengths indicating that the feature has a greater impact on the model's predictions.

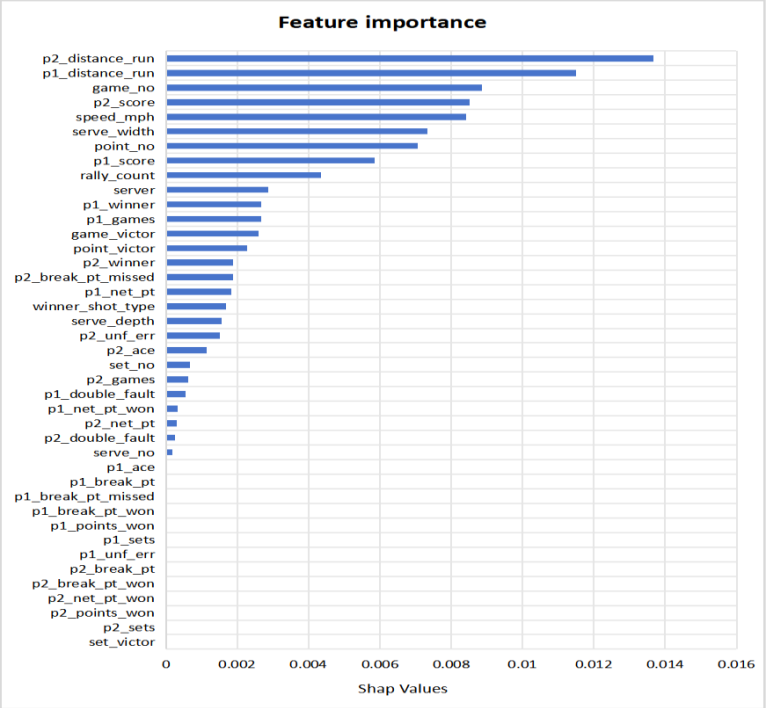

Figure 7: Feature importance bar chart

By observing the plotted histograms in Figure 7, it is clear that p2_distance_run (0.01368) and p1_distance_run (0.011499) are the two most important features, which indicates that the distance a player moves on the field has a significant effect on the predicted results. This may reflect the importance of momentum in the game, such as players' endurance and mobility. The game_no (0.008869), p2_score (0.008511) and speed_mph (0.008425) are also more important features. This suggests that the stage of the match in progress, the score of both sides as well as the speed of the serve also have a greater impact on the model's predictions.

4. Conclusion

Through this study, we have delved into the importance of momentum in sports competition, particularly the strategic turnaround embodied in the 2023 Wimbledon men's singles final. The successful construction and validation of machine learning models demonstrates the potential of data science in sports prediction. We found that the distance a player moves on the court has a significant impact on the outcome of a match, highlighting the criticality of the effective use of technology and strategic resources in winning and losing.

In addition, a deeper reflection on the complexity and difficulty of measuring momentum is presented. Changes in momentum are not only influenced by surface factors but may also be driven by deeper factors that are still unknown. Future research could further explore the influences behind momentum to improve the understanding and prediction of race outcomes.

Ultimately, our research provides athletes and coaches with a fresh perspective to help them optimize their training and competition strategies. By combining data science and sports competition, we can better understand the changes and trends in the game and provide more scientific support for decision making in the field of sports. This study opens a new direction for the field of sports science and provides useful

ISSN 2616-5775 Vol. 7, Issue 6: 88-94, DOI: 10.25236/AJCIS.2024.070614

references and insights for future research and practice.

References

[1] Briki W, Den Hartigh R J R, Markman K D, et al. How psychological momentum changes in athletes during a sport competition [J]. Psychology of Sport and Exercise, 2013, 14(3): 389-396.
[2] Ferreira A P. From game momentum to criticality of game situations[M]//Routledge handbook of sports performance analysis. Routledge, 2013: 270-282.
[3] 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.
[4] Zhang Y, Zhu C, Wang Q. LightGBM-based model for metro passenger volume forecasting[J]. IET Intelligent Transport Systems, 2020, 14(13): 1815-1823.
[5] Zhou W, Yan Z, Zhang L. A comparative study of 11 non-linear regression models highlighting autoencoder, DBN, and SVR, enhanced by SHAP importance analysis in soybean branching prediction[J]. Scientific Reports, 2024, 14(1): 5905.
[6] Prendin F, Pavan J, Cappon G, et al. The importance of interpreting machine learning models for blood glucose prediction in diabetes: an analysis using SHAP[J]. Scientific Reports, 2023, 13(1): 16865.