Finding Factors Influencing Tennis Match Trends Based on AHP-TOPSIS and LightGBM

Qiang Fu *

School of Mathematical Sciences, Ocean University of China, Qingdao, China, 266100

* Corresponding Author Email: fuqiang@stu.ouc.edu.cn

Abstract. The paper conducts a comprehensive analysis of the 2023 Wimbledon men's singles tournament in an attempt to quantify momentum to look at swings in play and runs of success, and to make predictions about flows of momentum. The paper introduces an innovative approach to analyzing tennis match dynamics by integrating the Analytic Hierarchy Process-Technique for Order Preference by Similarity to Ideal Solution (AHP-TOPSIS) with the LightGBM algorithm. The AHP-TOPSIS model quantifies player momentum, identifying key factors such as serving ability, strength, mentality, and in-game state that influence match outcomes. A novel aspect of this study is the application of the CUSUM algorithm to detect significant turning points within matches, providing a granular insight into the flow of play. Subsequently, the LightGBM model, known for its efficiency and accuracy, is employed to predict these turning points, offering a comprehensive predictive framework that enhances the understanding of match dynamics. The paper's findings underscore the predictive power of the serving factor and the state factor in determining match swings. This research contributes to the field by offering a more nuanced and data-driven perspective on tennis match analysis, with practical implications for both players and coaches in strategizing for competitive advantage.

Keywords: Momentum, AHP-TOPSIS, Turning points, LightGBM.

1. Introduction

Tennis, a sport with a rich history dating back to the 19th century in England, has always been a subject of interest for both sports enthusiasts and researchers [1]. The game's dynamics, characterized by intense rallies and strategic plays, have led to a plethora of studies aimed at understanding and predicting match outcomes [2-3]. One of the central concepts in tennis analytics is "momentum," which is often credited for the dramatic shifts in a player's performance during a match.

However, with the advent of data science and machine learning, there has been a paradigm shift in how momentum is quantified and utilized in match analysis. Recent studies have employed various statistical models and machine learning algorithms to predict match outcomes and analyze player performance [4-5]. For instance, some researchers have used decision tree-based models to predict match results, while others have applied clustering techniques to identify patterns in player performance.

Despite these advancements, there remains a gap in the literature regarding a comprehensive model that not only quantifies momentum but also integrates it with other critical factors to predict match dynamics effectively. This paper contributes to the field by offering a multi-dimensional analysis of momentum, integrating it with other performance indicators, and employing advanced machine learning techniques for prediction. The innovative use of AHP-TOPSIS for momentum quantification and the application of LightGBM for turning point prediction set this study apart from existing research, offering a more granular and data-driven perspective on tennis match analysis. (Data source: The Mathematical Contest in Modeling (MCM)® and The Interdisciplinary Contest in Modeling (ICM)® - COMAP)

2. Capturing the Flow of Play

2.1. Feature Extraction

The paper choses nine factors from the data table to quantify the performances of players, which is shown in Table.1.

Table.1. The features and calculation methods

Features	Calculation method
Number of leading	The game number of winning – The game number of losing
Untouchable shot	Statistic
Rate of 1st serving error	$\frac{\text{The number of 2nd serving}}{\text{The number of total serving}}$
Distance_run	The distance a player runs in this point
Rally_count	The number of returns by the player
Speed_mph	Observed value
Serve or not	Statistic
Got previous point or not	Statistic
Number of double faults	Statistic

2.2. Establishment of the Feature Reduction Model Based on Factor Analysis

The paper uses the Feature Reduction Model based on Factor Analysis to classify more representatively: a new combination is created by analyzing the variance contribution between variables [6]. The method of the model is as follows:

$$x_1 = a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m + e_1 \quad (1)$$

$$x_2 = a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m + e_2 \quad (2)$$

$$x_p = a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pm}F_m + e_p \quad (3)$$

Here, $X = (x_1, x_2, \cdots, x_p)^T$ is an observable random vector, satisfying the continuity, and have $E(X) = 0, COV(X) = \sum$. $F = (F_1, F_2, \cdots, F_m)^T (m < p)$ is a common factor, also an unmeasurable random vector. Here has $E(F) = 0, Cov(F) = 1$, and the vectors are independent of each other. $\varepsilon = (e_1, e_2, \cdots, e_p)^T$ is independent of F and there is $E(e) = 0$. Each e is also independent of each other.

The matrix form is as follows:

$$X = AF + \varepsilon \quad (4)$$

Here, $A = (a_{ij})_{p*m}$ is called factor loading matrix. So, factor score model can be demonstrated as the following formulas:

$$F_1 = \omega_{11}x_1 + \omega_{12}x_2 + \cdots + \omega_{1p}x_p \quad (5)$$

$$F_2 = \omega_{21}x_1 + \omega_{22}x_2 + \cdots + \omega_{2p}x_p \quad (6)$$

$$F_m = \omega_{m1}x_1 + \omega_{m2}x_2 + \cdots + \omega_{mp}x_p \quad (7)$$

2.3. Solving the Factor Analysis Model

KMO and Bartlett Sphericity Test were used on the data, and result is shown in Table.2.

Table.2. Result of KMO and Bartlett Sphericity Test

Variable		Value
Statistical magnitude of KMO		0.638
Bartlett Sphericity Test	Approximatechi-square	2098.708
	Degree of freedom	36
	P-value	0.000

From Table.2, statistical magnitude of KMO is greater than 0.5, illustrating the nine features are relevant. P-value in Bartlett Sphericity Test is smaller than 0.005, which rejects null hypothesis, also showing the nine variables are relevant. So, Factor Analysis is valid.

Calculate common factor variance and variance explanation is shown in Table.3. Here, four factors' eigenvalues are greater than 1, and the cumulative variance explanation rate is 80.41%.

From Fig, the fifth factor is the inflection point of this plot, indicating factors before and after the fifth one is main factors and minor factors respectively. Therefore, the paper maintains four factors.

Table.3. Variance explanation table

Factor	Eigenvalue	Variance explanation rate	Cumulative variance explanation rate
1	2.961	32.905	32.905
2	1.893	21.031	53.936
3	1.35	15	68.935
4	1.033	11.474	80.41
5	0.898	9.977	90.387
6	0.597	6.629	97.016
7	0.157	1.75	98.765
8	0.101	1.12	99.885
9	0.01	0.115	100

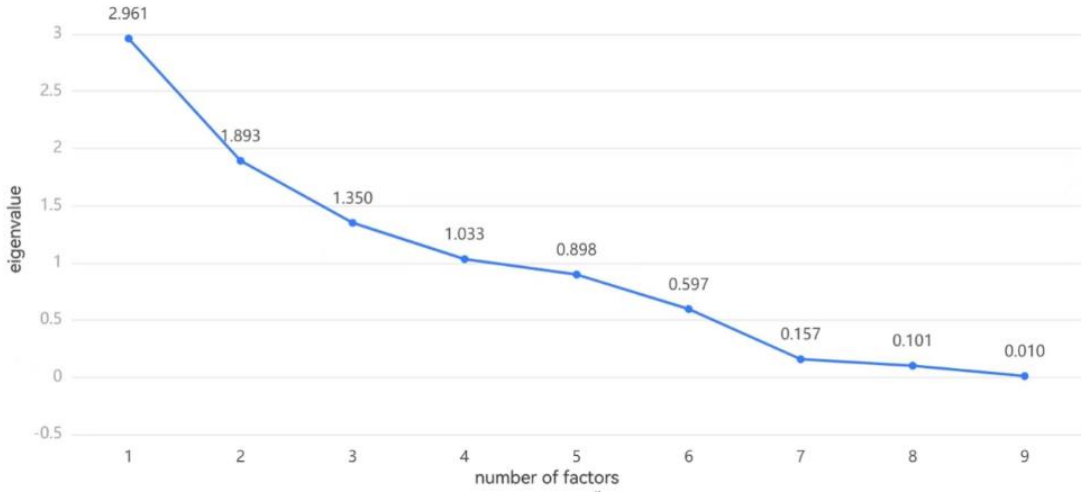


Fig.1. Scree plot of Factor

Factor loading coefficient table is created in Table.4. In F1 of Table.4, "speed_mph", "Serve or not" and "Rate of 1st serving error" have large loads relatively, so F1 is defined as "Serving factor". Similarly, "Rally_count" and "Distance_run" have large loads relatively, so F2 is defined as "Strength factor"; "Number of double faults" and "Number of leading" have large loads relatively, so F3 is defined as "Mentality factor"; "Got previous point or not" has a large load relatively, so F4 is defined as "State factor".

Table.4. Factor loading coefficient table

Factor	Factor loading coefficient				Common factor variance
	F1	F2	F3	F4	
Number of leading	-0.149	0.144	0.799	-0.043	0.683
Untouchable shot	0.362	0.098	0.024	-0.461	0.353
Rate of 1st serving error	0.926	0.08	0.045	0.06	0.87
Distance_run	-0.182	0.948	-0.121	0.013	0.946
Rally_count	-0.03	0.968	-0.1	-0.021	0.949
Speed_mph	0.975	0.059	0.07	0.012	0.959
Serve or not	0.973	0.065	0.065	0.031	0.957
Got previous point or not	0.075	0.055	0.023	0.902	0.822
Number of double faults	-0.109	0.098	0.822	0.022	0.697

Table.5. Factor score coefficient table

Variable	Element			
	F1	F2	F3	F4
Number of leading	-0.05	0.076	0.592	-0.042
Untouchable shot	0.122	0.052	0.018	-0.446
Rate of 1st serving error	0.313	0.042	0.034	0.058
Distance_run	-0.061	0.501	-0.09	0.013
Rally_count	-0.01	0.512	-0.074	-0.02
Speed_mph	0.329	0.031	0.052	0.011
Serve or not	0.329	0.034	0.048	0.03
Got previous point or not	0.025	0.029	0.017	0.873
Number of double faults	-0.037	0.052	0.609	0.021

Calculate factor score coefficient matrix by regression estimation method, obtain Table.5. after calculation. Factor Analysis Model can be written as follows:

$$F_1 = -0.05x_1 + 0.122x_2 + 0.313x_3 - 0.061x_4 - 0.01x_5 + 0.329x_6 \quad (8)$$

$$F_2 = 0.076x_1 + 0.052x_2 + 0.042x_3 + 0.501x_4 + 0.512x_5 + 0.031x_6 \quad (9)$$

$$F_3 = 0.592x_1 + 0.018x_2 + 0.034x_3 - 0.09x_4 - 0.074x_5 + 0.052x_6 + 0.048x_7 + 0.017x_8 + 0.609x_9 \quad (10)$$

$$F_4 = -0.042x_1 - 0.446x_2 + 0.058x_3 + 0.013x_4 - 0.02x_5 + 0.011x_6 + 0.03x_7 + 0.873x_8 + 0.021x_9 \quad (11)$$

2.4. Momentum quantization based on AHP-TOPSIS [7-8]

The fundamental premise of TOPSIS is that the best solution has the shortest distance from the positive ideal solution, and the longest distance from the negative ideal one. Alternatives are ranked with the use of an overall index calculated based on the distances from the ideal solutions.

However, it is difficult to estimate the factors' weights by using TOPSIS to calculate and analyze momentum. Therefore, it is useful to introduce AHP to give weight to factors. The paper combines AHP with TOPSIS to evaluate player's performance.

The weights of each index are calculated by AHP in Table.6.

Table.6. Weights of each index

Index	Weight
Serving factor	0.5376
Strength factor	0.2431
Mentality factor	0.1494
State factor	0.0699

The final first level index data of characteristic factors is obtained by comprehensive evaluation based on Table.6 as:

$$\text{Momentum}=\sum W_i F_i \tag{12}$$

In order to show the results more intuitively, the paper chooses match: 2023-wimbledon-1701 to visualize the result. The visualization is shown in Fig.2, 错误!未找到引用源。.

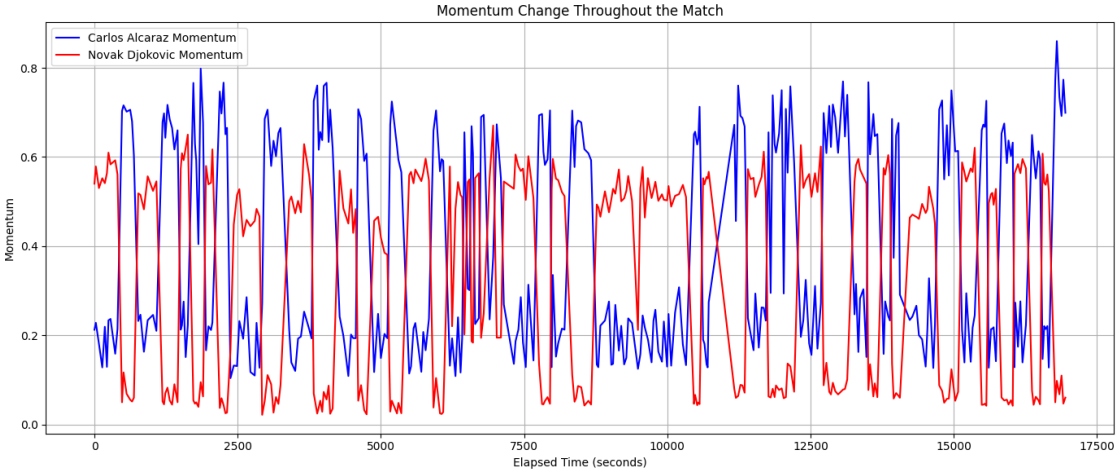


Fig.2. Momentum changes throughout the match

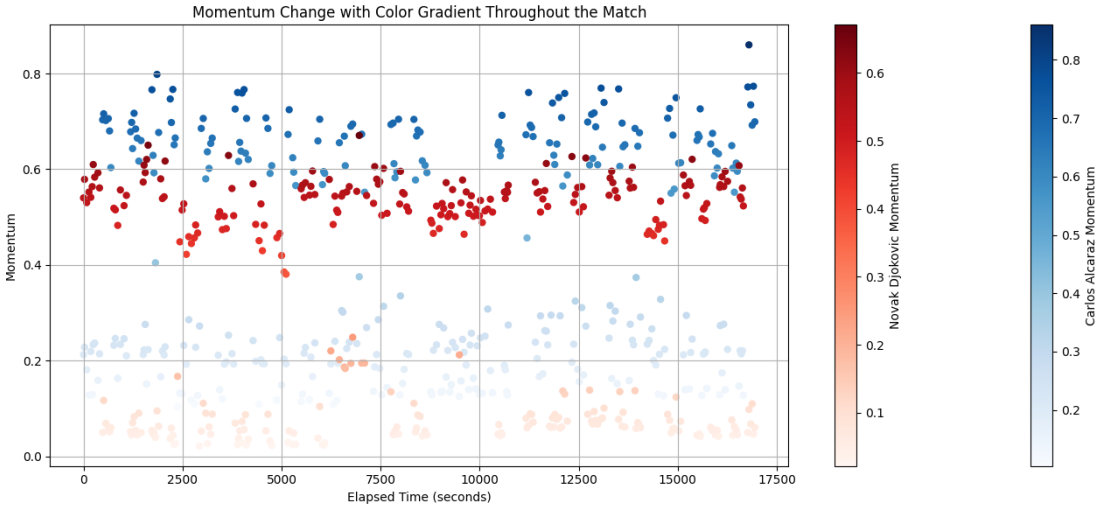


Fig.3. Momentum changes with color gradient throughout the match

During the 2023 Wimbledon match identified as 2023-wimbledon-1701, Carlos Alcaraz accumulated a total of 13 points, surpassing Novak Djokovic, who secured only five points, between the time frames of 14,796 and 15,610 seconds. As depicted in Fig.2, Alcaraz's momentum consistently surpassed that of Djokovic throughout the specified interval.

Between the timestamps of 15,365 and 15,610 seconds, Alcaraz achieved a consecutive score of six points, during which his momentum peaked, as illustrated in Fig.2.

The integrated analysis, underpinned by quantitative data and complemented by visual representations, yields profound insights into the match's dynamics. Players exhibiting heightened momentum are observed to consistently secure a greater number of points within a given match segment. Furthermore, the analysis reveals a correlation between increased momentum and the propensity for consecutive scoring, suggesting a direct impact on the flow of the game. This trend underscores the likelihood of players with superior momentum emerging victorious, highlighting the pivotal role of momentum in determining match outcomes.

3. Predicting the Flow of Play

3.1. Finding the Turning Points by CUSUM Algorithm

Select the data of 2023-wimbledon-1301 to plot the momentum changes of two players. The momentum changes are shown in Fig.4 and Fig.5.

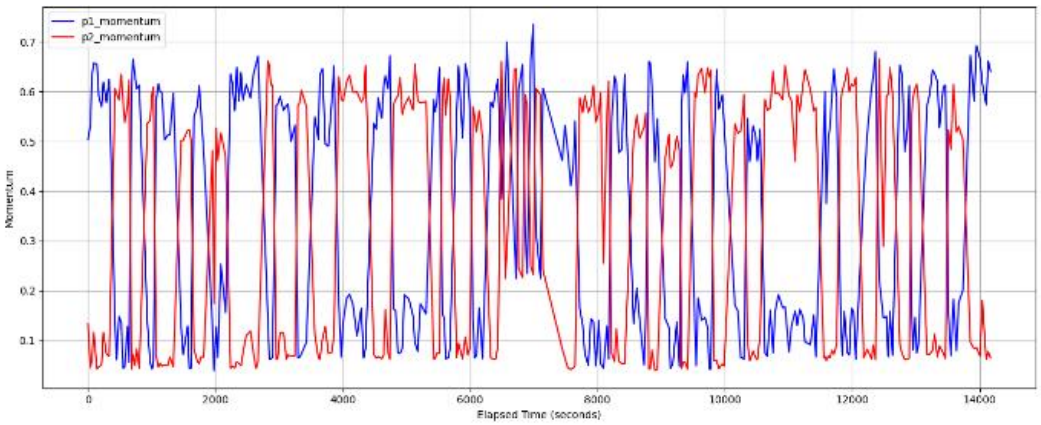


Fig.4. Momentum changes throughout the match

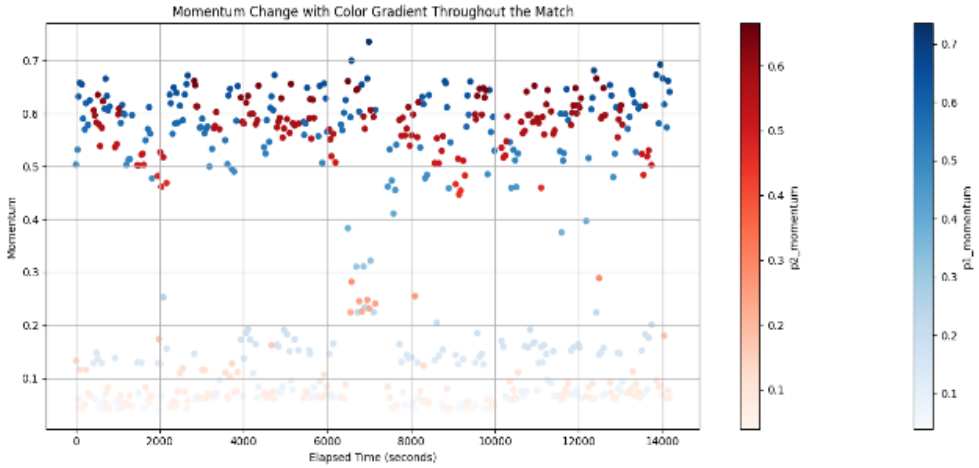


Fig.5. Momentum changes with color gradient throughout the match

Then find the significant change points in momentum change plots by differential inspection algorithm. Here, the paper chooses to use CUSUM (Cumulative Sum) algorithm. Then define a function in Python that uses CUSUM to test for turning points, implementing the CUSUM algorithm to detect turning points in a one-dimensional data series.The cumulative sum of the CUSUM algorithm can be expressed as:

$$S_t = \sum_{i=1}^{t} (X_i - \mu) \tag{13}$$

Among them:

S_t: is the cumulative sum at time point t.

X_i: the i_{th} observation.

μ: is the mean value of the process (in practice, this may be the initial mean of the sequence or the mean value obtained by some method of estimation).

The detection of turning points is based on the accumulation and sign change of the sequence S_t, i.e:

$$S_{t+1} \cdot S_t < 0 \tag{14}$$

This means that the sign of the cumulative sum changes between time points t and t+1. 错误!未找到引用源。 shows the 2023-wimbledon-1301 turning point obtained by the algorithm.

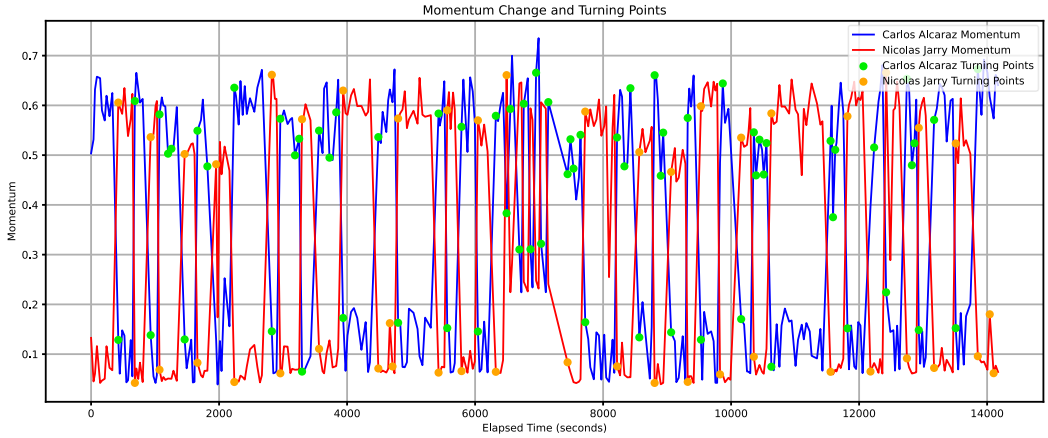


Fig.6. Momentum changes and turning points

3.2. Establishment of Prediction Model of Swings Based on LightGBM [9-10]

LightGBM represents a sophisticated Gradient Boosting-based machine learning framework, renowned for its exceptional performance in terms of both accessibility and scalability. This framework adeptly harmonizes the competing demands of precision and computational efficiency. It leverages an advanced gradient boosting technique, coupled with a histogram-based decision tree algorithm, to expedite the model training regimen.

The paper uses LightGBM Model's default parameters in training. Default parameters are in Table.7.

Table.7. Model's setting of default parameters

Parameter	Default setting
Learning_rate	0.1
n_estimators	10
max_depth	8
num_leaves	31
min_child_samples	20
min_sum__hessian_in_leaf	0.001
bagging_fraction	1
bagging_freq	0
max_bin	20
feature_fraction	1

To verify the reliability of the LingtGBM Model, we use Accuracy, Recall, Precision, F1 score and AUC-ROC as metrics for model precision. After parameter optimization, the calculation results of each index in the training set and test set are as Table.8. shows.

Table.8. Result of model accuracy test

	Accuracy	Recall	Precision	F1 score	AUC
Training set	0.842	0.593	0.869	0.612	0.741
Testing set	0.814	0.563	0.779	0.564	0.726

It can be considered that the model results are in good agreement with the real data because ROC is greater than 0.8. This conclusion can also be drawn from Fig.7.

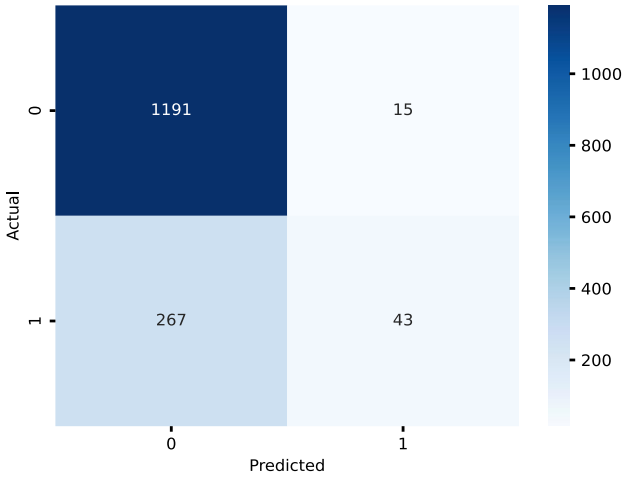


Fig.7. Confusion matrix heat map

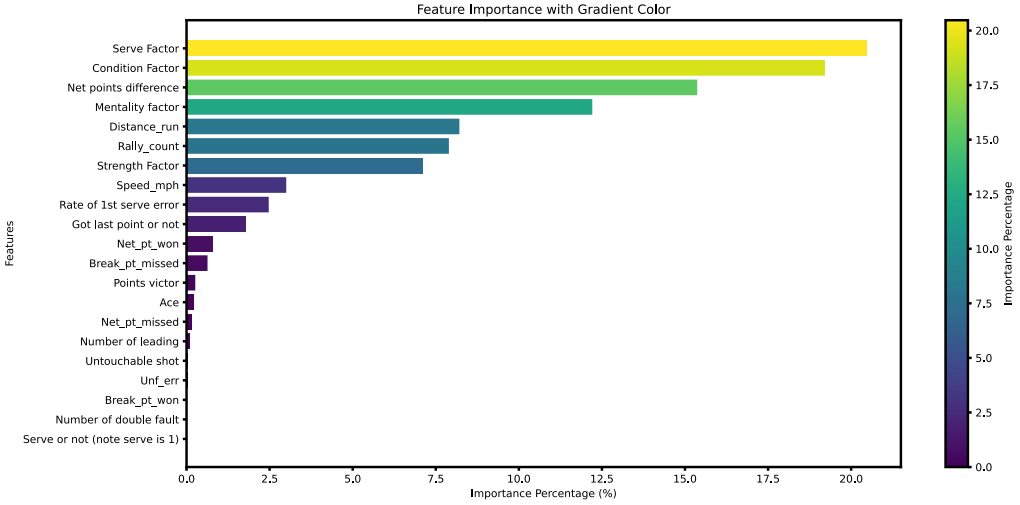


Fig.8. Relative importance of features

The robustness of the model permits the utilization of feature weights to evaluate the influence of various indicators on the prediction of match swings. Fig.8 reveals that the "State factor" and "Serving factor" are paramount, as they encapsulate the player's current form and the critical role of serving in shaping match dynamics. Additionally, the "Net point difference," "Mentality factor," "Distance_run," "Rally_count," and "Strength factor" exert notable effects, underscoring the significance of a player's endurance and mental fortitude in determining the course of the match. While attributes such as "serving speed," "Rate of 1st serving error," "Net_point_won," and "break_pt_missed" exert a lesser degree of influence on the model's predictive accuracy, their significance in the overall predictive framework should not be underestimated, as they contribute modestly yet meaningfully to the model's comprehensive assessment.

4. Conclusions

The paper develops a quantitative momentum model to assess tennis players' performance and a predictive LightGBM Classification Model for match turning points. By integrating multiple influential factors—serving, strength, mentality, and state—the model provides a comprehensive and universally applicable assessment system. The model's predictive performance was enhanced through rigorous parameter tuning and optimization, ensuring high accuracy across various datasets and match conditions.

Key findings indicate that higher momentum correlates with increased scoring potential. The Serving factor and State factor are identified as the most influential in match dynamics, significantly affecting players' wins and losses. The research suggests strategic implications for athletes: capitalize on every serve opportunity, refine serve speed control to minimize errors, and maintain or swiftly adjust mental and physical states to sustain or regain momentum. Additionally, attention to game rhythm and physical endurance is crucial, as these factors can impact match outcomes and may be exploited by opponents.

References

[1] Li Wei, Cui Yongmin, Wang Xiyue. Research on science and technology to help the development of tennis[J]. Sports Goods and Technology, 2024, (10):117-119.

[2] Li Meng, Li Weifeng. Utility analysis of different hitting zones of the world's outstanding male tennis players--Taking Djokovic and Federer as an example [J]. Sports Science and Technology Literature Bulletin,2024,32(04):59-67.

[3] Wen Yahu. The effect of shouting on hitting effect in tennis [J]. Tennis World, 2024, (04):61-63

[4] Zhu Haoyu, Xu Guodong. Application of posture assessment methods in tennis[C]//Chinese Society of Sports Science. Abstracts Collection of the 13th National Sports Science Conference - Special Report (Sports Engineering Division). Jianghan University, 2023.

[5] Gu Xiaoyang, Zong Zhiwei. Research on tennis motion target detection by fusing Kalman filter [J]. Industrial Control Computer,2023,36(12):48-50.

[6] Jiang Lanfang, Li Xiaoli, Cao Yong. Prediction of wheat quality by gluten aggregometer based on principal component analysis [J]. Food Science,2022,43(14):85-92.

[7] Guo Jinyu, Zhang Zhongbin, Sun Qingyun. Research and application of hierarchical analysis [J]. Chinese Journal of Safety Science, 2008, 18(5):148-153.

[8] Liu Jing, Sun Yuhe, Zhang Zhengdong. Research on early warning method of university public opinion based on AHP-TOPSIS method [J]. Big Data Times,2024,(03):27-30.

[9] Zhong Junting, Zhang Xiaoye, Gui Ke. Robust prediction of hourly PM (2.5) from meteorological data using LightGBM[J]. National Science Review,2021,8(10):115-126.

[10] Wang Mingyuan, Wang Zhenhong, Chen Songting. Prediction model of marine soil compression parameters based on LightGBM algorithm [J]. Journal of Zhejiang University of Technology, 2024, 52(01): 17-24.