Problem Chosen	2024 MCM/ICM Summary Sheet	Team Control Number
C		2414434

Foresee the Swings: Modeling Swings and Momentum in a Tennis Match

Summary

The glamour of Wimbledon between the blue skies and the grass was on full display with the spinning tennis balls in the air, the athletes' dexterity and the occasional cheers from the spectators. The game was fluid and both players seemed to have incredible fluctuations attributed to '**momentum**'. In order to capture this fluctuation, we developed **a quantitative evaluation model of situation and momentum** to determine and analyze the interaction between momentum and situation.

First, after data preprocessing and feature selection, we define and calculate Relative Advantage (RA) as a composite score of the game situation based on **entropy weighting** and **TOPSIS**, and quantitatively evaluate the prediction of Momentum based on **a non-linear regression model** with **factor analysis** and **AHP**. This model assesses the on-field situation and "momentum" of a player at a given time, and allows for **visualization** of changes in the situation.

Then, we analyzed the correlation between RA (situation) and MS (momentum) using the **Pearson correlation coefficient** and found a correlation coefficient of **0.564453**, indicating a certain level of correlation between the two. This indicates that momentum plays an important role for athletes on the field.

Next, we developed an **LGBM-based match fluctuation prediction model**, which predicts changes in the direction of the match by predicting which player has **a higher probability of scoring**. We predicted the classic Carlos Alcaraz vs. Novak Djokovic tennis match and found that the results were consistent with the actual situation, with a very high accuracy of **99.79%**. This model will help coaches and players to anticipate the direction of the match and adjust their strategies and techniques in time. Additionally, we also predicted **women's tennis** with an accuracy of **92.54%**, which indicates that our model is generalizable.

Furthermore, using **the characteristics of the LGBM model**, we analyzed **the importance of factors** influencing momentum. The results revealed that the **time of the match had the highest importance**, while the importance of unforced errors was the lowest. This model will assist coaches and athletes in predicting the swing of the match and making timely adjustments to their strategies and techniques.

Finally, we conducted **sensitivity analysis** on the model, which demonstrated its universality and robustness. This indicates that the model can deal with various situations in a match efficiently. Based on these findings, we wrote a memo to the coach, providing valuable suggestions for improving the performance of tennis players.

Keywords: Match Swing Momentum AHP Entropy Weight TOPSIS NLReg LGBM

Contents

1 Introduction

1.1 Background and Literature Review

Tennis is referred to as the second largest ball sport in the world, known for its high entertainment value and intense and thrilling match proceedings, which have gained popularity among people worldwide. Particularly noteworthy is the men's singles final of the 2023 Wimbledon Championships. In this match, 20-year-old Spanish rising star Carlos Alcaraz defeated the 36-year-old veteran Novak Djokovic, bringing an end to the remarkable achievements of this greatest player in the history of Grand Slam tournaments.

The Wimbledon Championships follow a best-of-five sets format, where a set consists of a collection of games, and a game consists of a collection of points. The extended duration of the match introduces various uncertainties, leading to remarkable fluctuations even in seemingly dominant players, occurring at the level of points and sometimes even games. Such fluctuations are often attributed to "momentum." The uncertainty of momentum poses challenges for match predictions.

In previous research, Huang utilized a BP neural network model to predict tennis match outcomes [1], but it did not consider the influence of momentum and failed to provide a comprehensive explanation of the underlying mechanisms. Taylor, J., and Demick, A., analyzed momentum [2], but their analysis remained at a qualitative level, lacking quantitative analysis. To quantitatively analyze the impact of momentum in matches, it is necessary to establish a new model for evaluation.

1.2 Problem Restatement

Considering the background, we need to establish a model to address the following issues:

- **Task 1:**Establish a model to quantify the situation and momentum, in order to describe the flow of play as points occur and assess how players perform at specific times. Provide visualized match descriptions based on this analysis.

- **Task 2:**Develop a model to demonstrate whether momentum affects the match flow and, if so, what kind of impact it has.

- **Task 3:**Develop a model to predict the flow of play based on factors of the game. Test our model on one or more matches and demonstrate its performance. Then, we test our model on other kinds of matches like women's tennis matches to show how generalizable is the model.

- **Task 4:**Analyzing the relationship between different factors and athlete performance to identify which factors appear to be most correlated. Then, offer some suggestions on how to better his performance when he goes into a new match against other players based on our model.

- **Task 5:**Summarize our model results on momentum in a two-page memo to provide coaches and athletes with appropriate recommendations to navigate potential events during the flow of a tennis match.

1.3 Our Work

The flowchart of our work is shown in Figure 1. We first establish a model to quantify the situation and momentum, in order to describe the flow of play as points occur and assess how players perform at specific times. Then, we develop a model to demonstrate whether momentum affects the match flow and, if so, what kind of impact it has. We also develop a model to predict the flow of play based on factors of the game. We test our model on one or more matches and demonstrate its performance. Then, we test our model on other kinds of matches like

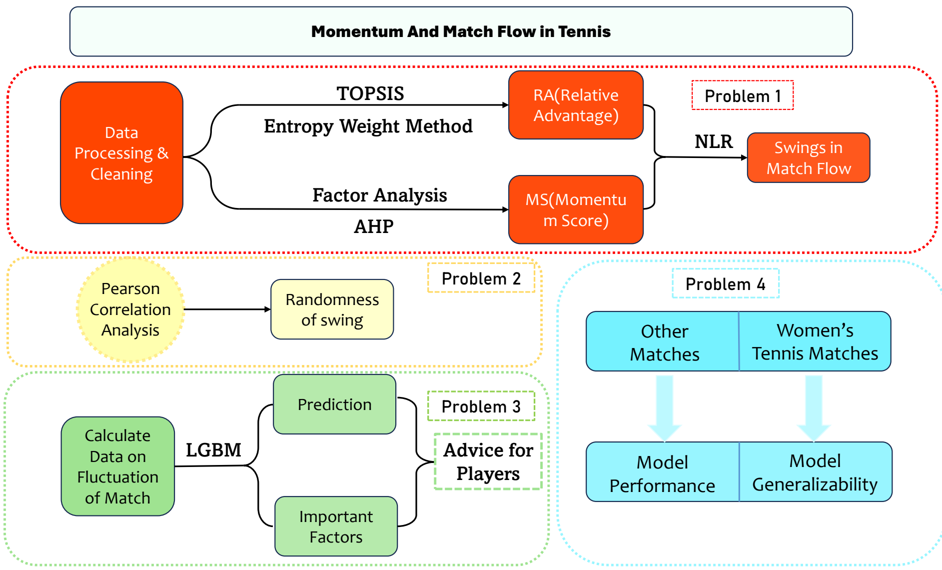


Figure 1: The flowchart of our work

2 Assumptions and Notations

2.1 Model Assumptions

Considering the conditions required for modeling, we make the following assumptions to simplify the model establishment, and each is properly justified.

- **Assumption1:** Neglecting the influence of external factors
  Justification: In a tennis match, external factors such as weather, atmosphere, and the playing venue can also affect the performance of players. However, these influencing factors are complex and cannot be incorporated into the model. Hence, we disregard the impact of these factors.

- **Assumption2:** Disregarding halftime rest for athletes' physical recovery
  Justification: The halftime break is relatively short and has a minimal impact on the athletes' physical recovery and mental adjustment. Therefore, we choose to ignore it.

- **Assumption3:** Assuming players start the match in their optimal condition
  Justification: At the beginning of a match, players have undergone pre-match adjustments

and warm-up exercises, which should have put them in their optimal condition. Therefore, we consider both players to be starting the match in their optimal state.

2.2 Notations

Symbol	Definition
x_{ij}	Raw data
z_{ij}	Standardized and positively transformed data
$\overline{z}$	Ideal optimal solution
$\underline{z}$	Ideal worst solution
D^+	Distance from ideal optimal solution
D^-	Distance from ideal worst solution
S_i	Comprehensive evaluation score
p_{ij}	Data proportion
e_j	Factor information entropy
w_j	Factor weights
RA_i	Relative advantage
$Cov(x,y)$	Covariance
MS_i	Momentum score
CI	Consistency index
RI	Average random consistency index
CR	Consistency ratio
RM_i	Relative momentum
θ_i	Regression model parameters
ϵ_i	Regression model random influencing factors

3 Data Preprocessing

3.1 Data Cleaning

The data provided is not perfect, and there are some missing values, as well as those that do not fit the analysis. We need to clean the data before we can use it. We use the following methods to clean the data:

1. **Completing the missing values**: The missing value NA occurs in the four metrics speed_mph, serve_width, serve_depth, and return_depth.We need to fill in the missing values. The results are presented in the Table 1.
   - **Speed_mph**: For the continuous variable speed_mph, we use the mean interpolation method, replacing NA with the mean value of the dataset.

- **Serve_width and serve_depth**: For the categorical variable serve_width and serve_depth, we use the mode interpolation method, replacing NA with the mode value of the dataset.
- **Return_depth**: For the indicator return_depth, its data missing values accounted for 17.97% of the dataset, which is too much missing for the data to be usable, so this indicator was not considered.

Table 1: Missing value interpolation

Missing value	Imputed Data Contents
speed_mph	112.41
serve_width	C
serve_depth	NCTL

2. **Digitization of indicators**The values of the indicators serve_width, and serve_depth are alphabetic, which is not suitable for data calculations, for this reason, we rewrite the data to make it numeric, which facilitates subsequent calculations.The digitization is shown in Table 2.

Table 2: Digitization of indicators

Indicator	Value	Digitized value
serve_width	B	1
	C	2
	W	3
	BW	4
	CW	5
serve_depth	D	1
	ND	0

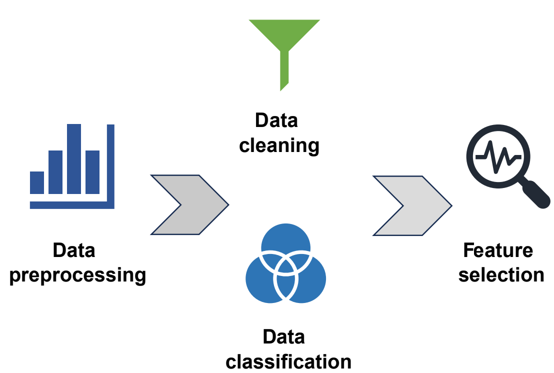


Figure 2: Data Preprocessing Process

3. **Data Abbreviation**The original data contains a large number of variables, and the variable names are too long, which is not conducive to subsequent data analysis. We use the abbreviation method to abbreviate the variable names, and the results are shown in Table 3.

Table 3: Abbreviation for Variables

variables	Abbreviations	variables	Abbreviations
match_id	MID	p1_break_pt_won	P1_BPW
player1	P1	p2_break_pt_won	P2_BPW
player2	P2	p1_break_pt_missed	P1_BPM
elapsed_time	ET	p2_break_pt_missed	P2_BPM
set_no	SN	p1_distance_run	P1_DR
game_no	GN	p2_distance_run	P2_DR
point_no	PN	rally_count	RC
p1_sets	P1_S	speed_mph	SM
p2_sets	P2_S	serve_width	SW
p1_games	P1_G	serve_depth	SD
p2_games	P2_G	return_depth	RD
p1_score	P1_S	p1_comboscore	P1_CS
p2_score	P2_S	p2_comboscore	P2_CS
server	SR	score_p1mp2	SP1M2
serve_no	SRN	set_p1mp2	SP1M2
point_victor	PV	games_p1mp2	GP1M2
p1_points_won	P1_PW	pt_p1mp2	PP1M2
p2_points_won	P2_PW	p1_combosets	P1_CS
game_victor	GV	p2_combosets	P2_CS
set_victor	SV	p1_combogames	P1_CG
p1_ace	P1_A	p2_combogames	P2_CG
p2_ace	P2_A	relative_advantage	RA
p1_winner	P1_W	p1_totalrun	P1_TR
p2_winner	P2_W	p2_totalrun	P2_TR
winner_shot_type	WST	p1_rscore	P1_RS
p1_double_fault	P1_DF	p2_rscore	P2_RS
p2_double_fault	P2_DF	p1_momscore	P1_MS
p1_unf_err	P1_UE	p2_momscore	P2_MS
p2_unf_err	P2_UE	relative_momentum	RM
p1_net_pt	P1_NP	p1_win	P1_W
p2_net_pt	P2_NP	p2_win	P2_W
p1_net_pt_won	P1_NPW	p1_tdf	P1_TDF
p2_net_pt_won	P2_NPW	p2_tdf	P2_TDF
p1_break_pt	P1_BP	p1_tbpm	P1_TBPM
p2_break_pt	P2_BP	p2_tbpm	P2_TBPM

4 Problem 1 Model: TOPSIS Evaluation Model based on Entropy Weight Method

4.1 Model Establishment

The scoring situation at a certain moment in a tennis match may not enough to indicate the flow of play as points occur. Therefore, it is necessary to establish a model to evaluate the match flow and momentum based on factors such as the score, the player's state, and the serving situation at a specific moment.

TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) is a commonly used multi-objective multi-attribute decision-making method that evaluates the current solution by calculating its similarity to the ideal solution. However, this model has the drawback of subjectivity. To overcome this drawback, we propose an improved approach using the entropy weight method to determine the weights and then evaluate the solutions using TOPSIS to obtain objective and accurate evaluation values.

4.1.1 Evaluation Method based on the TOPSIS Model

Firstly, data standardization is required. In the factors influencing match flow, some are positive indicators while others are negative indicators, and the data have different units, which poses challenges for quantitative analysis. To address this, we use formula (1) to standardize the positive indicators and formula (2) to standardize the negative indicators, thus unifying the data into a standardized positive indicator range of 0-1.

$$z_{ij} = \frac{x_{ij} - \min_{1\le i\le m} x_{ij}}{\max_{1\le i\le m} x_{ij} - \min_{1\le i\le m} x_{ij}} \tag{1}$$

$$z_{ij} = \frac{\max_{1\le i\le m} x'_{ij} - x'_{ij}}{\max_{1\le i\le m} x'_{ij} - \min_{1\le i\le m} x'_{ij}} \tag{2}$$

Where m is the number of samples, n is the number of indicators, x_{ij} represents the initial data of positive indicators, x'_{ij} represents the initial data of negative indicators, and z represents the standardized data.

Based on this, we need to determine the ideal solution as a reference for evaluating a particular sample. We define the positive ideal solution as

$$\overline{z} = (\overline{z_1}, \overline{z_2}, \overline{z_3}, ..., \overline{z_n}), \text{ where } \overline{z_j} = \max_{1\le i\le m} z_{ij}$$

And the negative ideal solution as

$$\underline{z} = (\underline{z_1}, \underline{z_2}, \underline{z_3}, ..., \underline{z_n}), \text{ where } \underline{z_j} = \min_{1\le i\le m} z_{ij}$$

To evaluate the indicator values of sample i and avoid the shortcomings of traditional linear summation, we calculate the relative distance between the sample and the positive and negative ideal solutions. The positive distance is defined by formula (3) and the negative distance is defined by formula (4).

$$D_i^+ = \sqrt{\sum_{j=1}^{n} (z_{ij} - \overline{z_j})^2} \tag{3}$$

$$D_i^- = \sqrt{\sum_{j=1}^{n} \left(z_{ij} - \underline{z_j}\right)^2} \tag{4}$$

Then, we use the formula (5) to obtain the relative evaluation criterion.

$$S_i = \frac{D_i^-}{D_i^+ + D_i^-} \tag{5}$$

S_i reflects the relative level of the distance between sample i and the positive and negative ideal solutions and can be used as the evaluation criterion for sample i.As S_i approaches 1, it indicates a higher evaluation score, whereas as S_i approaches 0, it indicates a lower evaluation score.

However, such a model cannot provide a relatively objective evaluation of the samples. In formula (3) and formula (4), each indicator adopts the same weight in the distance calculation, which fails to properly represent the fact that different indicators have different levels of influence. After all, the relationship between scoring on serves and player mentality is much more significant than the relationship between running speed and player mentality. To better measure the different weights of different influencing factors, we improve the TOPSIS model using the entropy weight method.

4.1.2 Model Improvement based on the Entropy Weight Method

Undoubtedly, factors with greater fluctuations contain more uncertainties and should not be considered as more important in evaluating momentum and match flow, with higher weights assigned to them. The entropy weight method extracts information utility from the uncertainty of the data and determines the weights of the influencing factors, effectively solving this problem.

Starting from the standardized data, we need to proceed with calculating the ratios of each indicator for every sample, which is given by the formula (5).

$$p_{ij} = \frac{z_{ij}}{\sum_{i=1}^{m} z_{ij}} \tag{6}$$

Where p represents the relative ratios and z represents the normalized data. We use the concept of information entropy to characterize the uncertainty of the influencing factors. The calculation method for the information entropy of indicator j is as follows.

$$e_j = -\frac{1}{\ln m} \sum_{i=1}^{m} p_{ij} \ln p_{ij} \tag{7}$$

Where e is the information entropy, and thus the information utility value $d_j = 1 - e_j$. Based on the information utility values of different indicators, we obtain their weights by formula (8).

$$w_j = \frac{d_j}{\sum_{j=1}^{n} d_j} \tag{8}$$

The weights provided by the entropy weight method are used to improve the distance calculation in formula (3) and formula (4) in TOPSIS. The distance calculation formula now becomes:

$$D_i^+ = \sqrt{\sum_{j=1}^{n} w_j \left(z_{ij} - \overline{z_j}\right)^2} \tag{9}$$

$$D_i^- = \sqrt{\sum_{j=1}^{n} w_j \left(z_{ij} - \underline{z_j}\right)^2} \tag{10}$$

By substituting this distance into formula (5) and performing the calculation, we obtain a quantifiable value S_i, which is used to comprehensively evaluate the indicator values of a particular sample.

4.1.3 Evaluation of Match Flow

To evaluate and analyze the match flow, we have selected several evaluation indicators for the match flow and assessed them using the TOPSIS model based on the entropy weight method.

The indicators we have chosen include: P1_S, P2_S, P1_G, P2_G, P1_S, P2_S, SR, SRN, P1_PW, P2_PW, P1_A, P2_A, P1_W, P2_W, P1_UE, P2_UE, P1_BP, P2_BP, P1_BPW, P2_BPW, P1_BPM, P2_BPM, SP1M2, SP1M2, GP1M2, PP1M2.

Using these indicators and applying the model code for computation, we obtain the score of the situation for p1 and p2 at each moment. Then, we utilize the formula (11) to calculate the Relative Advantage (RA) as points occur as a comprehensive score for the match flow. By computing RA at different moments and different scores, we can predict the flow of the match, thereby visualizing the progression of a match.

$$RA_i = \frac{score_i - \min\limits_{1\le i\le m} score}{\max\limits_{1\le i\le m} score - \min\limits_{1\le i\le m} score} - 0.5 \tag{11}$$

For simulation, we have selected the classic tennis match between Carlos Alcaraz and Novak Djokovic. Using the total points as the x-axis and RA as the y-axis, we visualize the match flow, as shown in Figure 3.

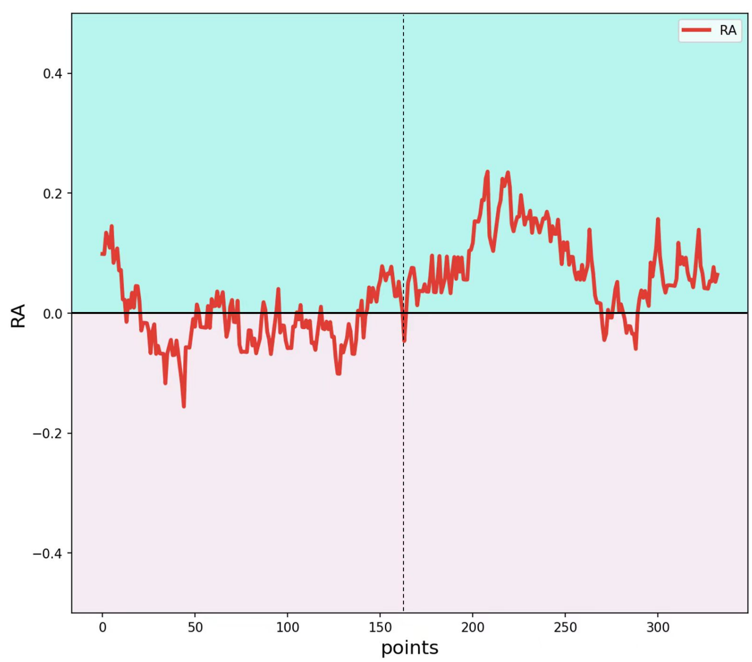


Figure 3: Match Flow of Carlos Alcaraz vs Novak Djokovic

4.2 Nonlinear Regression Model Based on Factor Analysis and AHP: Evaluation and Prediction of Momentum

Undoubtedly, momentum in a match is much more complex than relative advantage. The influencing factors of this metric are not only diverse but also interdependent, posing significant challenges in quantifying momentum. To address this issue, we first applied factor analysis to analyze the principal components of selected influencing indicators, abstracting numerous influencing factors into three composite factors. We then constructed a nonlinear regression model based on the Analytic Hierarchy Process (AHP) to quantify momentum and predict its value at a given game point.

4.2.1 Data Dimensionality Reduction Using Factor Analysis

We initially calculated the Kaiser-Meyer-Olkin (KMO) value of the data as 0.687. Since the KMO value exceeds 0.6, it indicates a strong correlation between the influencing factors. Additionally, the Sig value of Bartlett's test of sphericity was 0.000, which is less than 0.005, indicating the presence of interrelated variables. Therefore, factor analysis is able to be conducted.

We employed the formula (12) to perform data dimensionality reduction.

$$X = AF + \varepsilon \tag{12}$$

Where

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}, A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mp} \end{bmatrix}, F = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix}, \varepsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{bmatrix}$$

X represents the matrix of normalized influencing factor data.

A denotes the loading matrix, where a_{ij} reflects the relationship between the original influencing factor i and the common influencing factor j.

F represents the common factor matrix, indicating the common components of the influencing factors, with each common factor being uncorrelated.

ε, signifies the specific factor matrix, where ϵ represents the impact components in the influencing factors that cannot be explained by the common factors.

To achieve data dimensionality reduction and obtain the common factors, we need to solve the loading matrix. Principal component analysis is utilized. First, we define the covariance matrix

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}$$

where $r_{ij} = Cov\left(x_i,\ x_j\right)$, and $Cov\left(x_i,\ x_j\right)$is the covariance between the influencing factors i and j.

Next, we compute the eigenvalues $\lambda_1, \lambda_2, ..., \lambda_p$ and the corresponding unit eigenvectors $e_1, e_2, ..., e_p$

By calculation, we obtain a cumulative explanatory variance of 50.77%. The eigenvalues decrease significantly after the third component, indicating that the first three components are the main factors, while the subsequent components are minor factors. Hence, we retain three factors.

Thus, we have

$$a_{ij} = \sqrt{\lambda_i \lambda_j} e_i e_j^T$$

and the loading matrix is depicted as a heatmap in Figure 4.

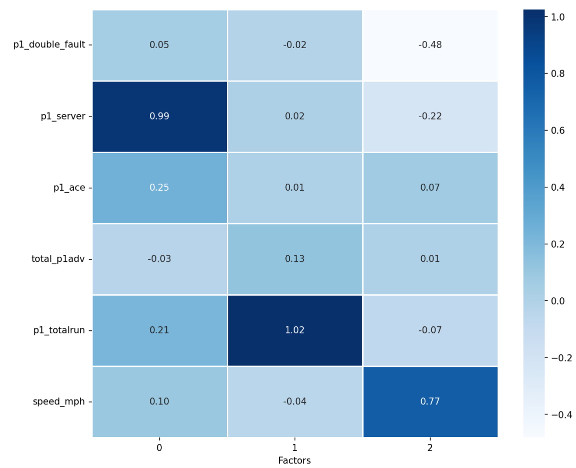


Figure 4: Loading matrix heatmap

From the loading matrix, we can infer that the first-factor variable has a high loading for winning points on serve and whether serving, defining it as the serving factor.

The second-factor variable has a high loading for running_distance and leading_position, defining it as the process factor.

The third-factor variable has a high loading for double_faults and serve_speed, defining it as the performance state factor.

4.2.2 Determination of Main Factor Weights Based on AHP

Based on personal experience and evaluations of the importance of the aforementioned main factors in the literature, we assessed the degree of influence for the three main factors in the criterion layer and constructed the judgment matrix

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 0.5 & 1 & 2 \\ 0.333 & 0.5 & 1 \end{bmatrix}$$

We obtained the maximum eigenvalue λ_{max} of the judgment matrix A, and the corresponding unit eigenvector

$$e = \begin{bmatrix} 0.467532 \\ 0.389610 \\ 0.142857 \end{bmatrix}$$

represents the weights of the three factors.

Next, we conducted a consistency test on the AHP model. Firstly, we calculated the consistency index CI using the formula (13).

$$CI = \frac{\lambda_{\max} - n}{n - 1} \tag{13}$$

Then, referring to the table of random consistency index (RI), we found that the RI value for three factors is 0.58.

Subsequently, we calculated the consistency ratio CR using the formula (14).

$$CR = \frac{CI}{RI} \tag{14}$$

The consistency ratio CR of the model was found to be 0.0046, which is very close to 0 and significantly less than 0.10. This indicates that the comparison matrix is relatively consistent, and the weight determination effect of the model is good.

Therefore, we defined

$$MS = \sum_{i=1}^{p} w_i x_i$$

as the quantitative indicator of the momentum score of the athlete at a certain moment during the match, where w_i represents the weight of the main factor, and x_i represents the value of the main factor.

By applying the model code for calculation, we can obtain the MS for p1 and p2 at each moment. Then, similar to the formula (11)using the formula (15), we calculate the Relative Momentum (RM) as an evaluation indicator for relative momentum.

$$RM_i = \frac{MS_i - \min\limits_{1 \leq i \leq m} MS}{\max\limits_{1 \leq i \leq m} MS - \min\limits_{1 \leq i \leq m} MS} - 0.5 \tag{15}$$

We applied this model to the classic tennis match between Carlos Alcaraz and Novak Djokovic, and the results are shown in Figure 5.

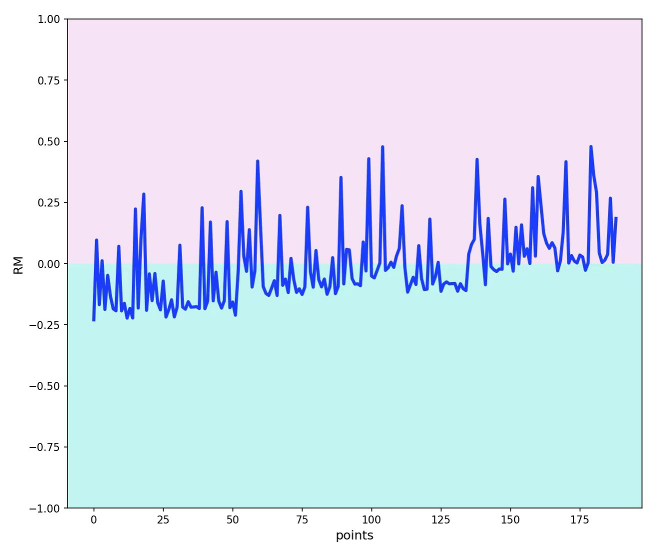


Figure 5: Momentum of Carlos Alcaraz vs Novak Djokovic

From the figure, we can see that the relative momentum fluctuates dramatically between the two players, indicating a closely contested match. However, as the match progresses, the relative momentum gradually favors the young Spanish rising star, seemingly foretelling Carlos Alcaraz as the ultimate winner of the match.

4.2.3 Momentum Prediction and Evaluation Model Based on Nonlinear Regression

In previous predictions, we evaluated matches based on the data available in the table, achieving excellent results. However, when given data falls outside the range of the dataset, the evaluation effectiveness of the model becomes uncertain. Therefore, it is necessary to establish a prediction model that can evaluate players' momentum for any given match data.

Considering the limitations of linear regression, we adopted polynomial regression, a form of nonlinear regression.

Considering that polynomials of degree three or higher significantly reduce interpretability, we use formula (16) to define a third-degree polynomial regression function.

$$y_i = f\left(x_i,\ \theta_i\right) + \epsilon_i,\ i = 1, 2, ..., n \tag{16}$$

Where, y_i represents the predicted variable, i.e., MS; f denotes the polynomial function; x_i is the independent variable; θ_i represents the model parameters; and ϵ_i represents the random component in the model.

By fitting the data from the first 30 matches in the dataset, we obtained the model parameters. Using this model, we made predictions for matches and calculated the mean squared error (MSE) between the predicted results and actual data, which was found to be 4.57×10^{-30}. This value is extremely close to zero, indicating the exceptional fitting performance of the model.

Additionally, we calculated the coefficient of determination R as 1.0, indicating that the model can fully explain the variance of the target variable. In other words, the model perfectly fits the data.The model is solved using Python code, and the detailed code is attached in Appendix 9.1.

5 Problem 2 Model: Pearson Index-Based Correlation Model

5.1 Establishment of the Correlation Model

To measure the correlation between momentum and match flow, we have established a correlation model based on the Pearson coefficient. This model calculates the Pearson correlation coefficient between momentum and match flow to reveal their relationship.

The linear correlation between two sets of variables $X = (x_1, x_2, x_3, \ldots, x_n)$ and $Y = (y_1, y_2, y_3, \ldots, y_n)$ is given by the Pearson coefficient, which is computed using the formula (17).

$$r_{xy} = \frac{Cov(X, Y)}{\sqrt{D(X)}\sqrt{D(Y)}} \tag{17}$$

Where the covariance of variables X and Y is

$$Cov(X, Y) = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

And the standard deviation of variable X (or Y) is

$$\sqrt{D(X)} = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where $\bar{x}$ and $\bar{y}$ are the means of X and Y, respectively. The Pearson coefficient ranges from -1 to 1, with 1 indicating a perfect positive linear relationship, -1 indicating a perfect negative linear relationship, and 0 indicating no linear relationship.

Using these formulas, we can calculate the Pearson correlation coefficient between the two variables and determine their correlation relationship. A Pearson coefficient closer to 1 indicates a stronger positive correlation between the variables, while a coefficient closer to -1 indicates a stronger negative correlation. A coefficient closer to 0 suggests a weaker correlation between the variables. The relationship between the absolute value of the Pearson coefficient and the strength of the correlation can be seen in Table 4.

Table 4: Pearson Coefficient and Correlation Strength

Range of Pearson Coefficient	Degree of correlation	Direction of correlation
$0.5 < r \leq 1$	Strong correlation	Positive correlation
$0.3 < r \leq 0.5$	Moderate correlation	Positive correlation
$0 < r \leq 0.3$	Weak correlation	Positive correlation
$r = 0$	No correlation	No correlation
$-0.3 \leq r < 0$	Weak correlation	Negative correlation
$-0.5 \leq r < -0.3$	Moderate correlation	Negative correlation
$-1 \leq r < -0.5$	Strong correlation	Negative correlation

5.2 Correlation Analysis of Momentum and Match Flow

Using the correlation analysis model based on the Pearson coefficient, we can compute the correlation coefficient between momentum and match flow, thereby inferring their relationship based on Table 4.

By calculating the RA and MS values at different time points in the data table, we obtain two sets of variables, RA and MS. First, we plot a scatter plot of the two variable sets, as shown in Figure 6.

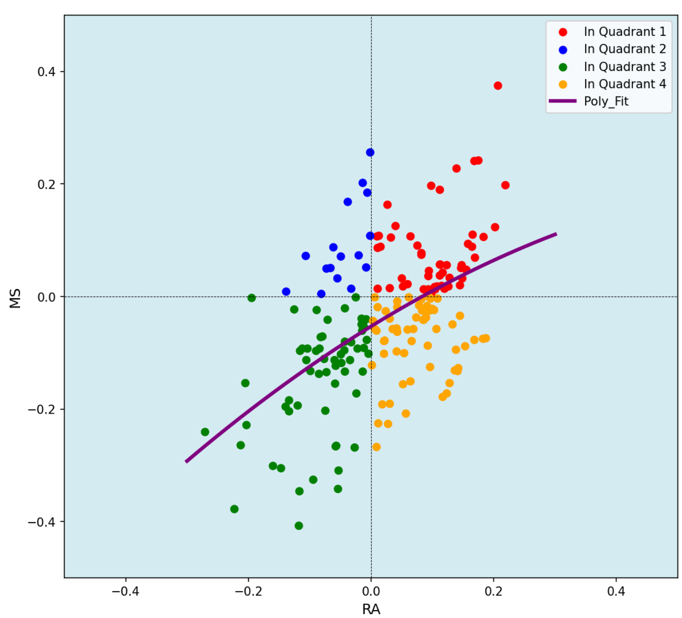


Figure 6: Scatter plot of RA and MS

From the scatter plot, it is evident that there are more data points distributed in quadrants 1 and 3 than in quadrants 2 and 4, suggesting a possible correlation. Additionally, we plot the Poly_Fit curve, represented by the purple curve in the figure, indicating a likely positive correlation.

By inputting the data of RA and MS into the model, we calculate the Pearson coefficient value between the two variable sets to be 0.564553. Referring to Table 4, it becomes evident that the two sets of data have a strong correlation, indicating that match flow is indeed related to momentum and that momentum does have an influence on the course of the match.

6 Problem 3 Model: LGBM-Based Match Fluctuation Prediction Model

6.1 Establishment of the LGBM Model

A model needs to be established to predict the fluctuation of a match at a certain moment. However, predicting the fluctuation of a match situation is often not easy. It is challenging to foresee the course of a match midway, especially in matches where the strengths of both sides are comparable. To address this, we made a transformation by predicting the likelihood of scoring the next significant point rather than directly predicting the match fluctuation. Through this transformation, the model becomes easier to understand and solve. We can summarize patterns from past matches and use the current match's data to predict which side is more likely to score the next point, thereby making predictions about the current state of the match.

This inspired us to use the LGBM model, which is a prediction and classification model based on the gradient-boosting decision tree (GBDT). It has advantages such as efficient handling of large-scale data, good handling of categorical features, and high prediction accuracy. The process of establishing the model is described as follows.

First, to predict the scoring situation of the next point, we need to select the features that the model will use, as these features will affect the scoring. We selected xxxx as the feature indicator for the model evaluation, using these features to predict which side will score the next point.

Then, the dataset needs to be divided. We need to split the dataset into a training set and a test set, where the training set is used to train the model and construct the decision tree of the model. The test set is used to evaluate the model to verify its accuracy.

Upon this foundation, we constructed the LightGBM model. For the model parameters num_leaves, max_depth, and learning_rate, we employed a cross-validation approach to obtain initial parameters and build the primitive model. Subsequently, through iterations, we fed the training set into the model, constructing the most optimized decision tree each time, thereby training an optimized LightGBM decision tree model. This model can ultimately be utilized to predict the scoring outcome of the next point based on various data points at the current game point.

6.2 Predicting Match Fluctuation using the LGBM Model

Using the model, we made predictions on the classic tennis match between Carlos Alcaraz and Novak Djokovic. We input various match data as points occur into the model to predict which side is more likely to win at each crucial point. Based on this, we generated a match fluctuation prediction graph using score predictions, as shown in Figure 7.

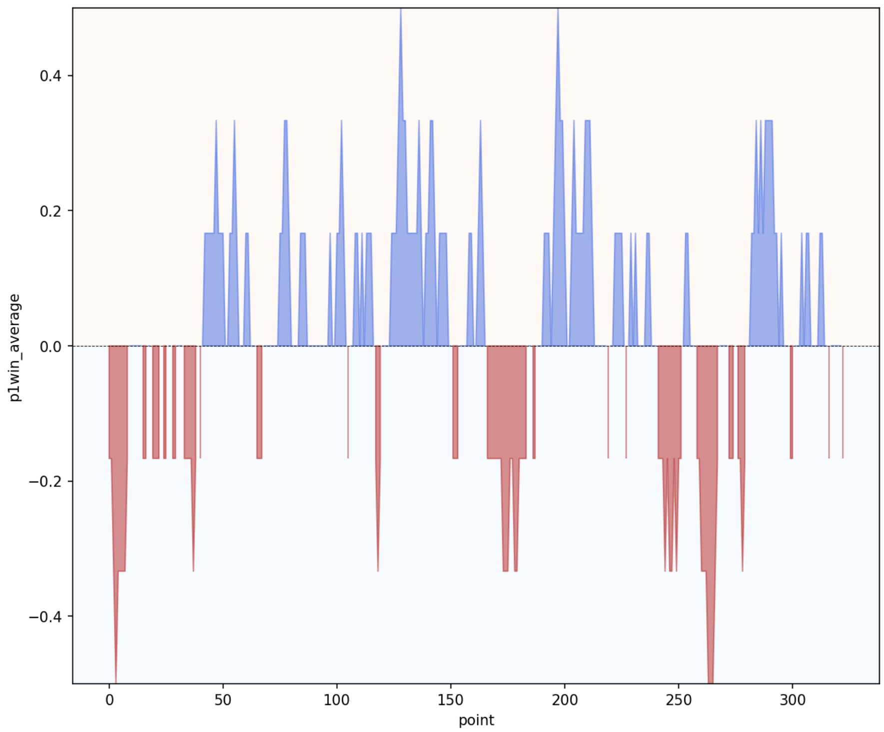


Figure 7: Match Fluctuation Prediction Graph

The classification results of the model indicated a 100% accuracy rate, which was indeed the case. Figure 7 effectively depicts the overall flow of the match. Initially, Novak Djokovic had a clear advantage and almost controlled the entire match. However, as the match progressed, the young Spanish rising star gradually turned the tide in his favor, shifting the momentum towards Carlos Alcaraz. With the match progressing, the veteran's advantage diminished. However, in the third set, Novak Djokovic gradually regained momentum and achieved a certain scoring advantage, but he still lost to the Spanish newcomer in that set. In the fourth set, the veteran's momentum further increased, seemingly regaining control of the situation and securing victory in that set. However, in the fifth set, Carlos Alcaraz's momentum surged again, reclaiming the lead and ultimately winning the match.

In Figure 7, the area of the upper part of the graph is greater than that of the lower part, which aligns with Carlos Alcaraz's victory over Novak Djokovic.

Additionally, we used the LGBM model to make predictions on the test set in the dataset. Out of the 1457 data points tested, 1454 were correctly predicted, and 3 were incorrect, as shown in Figure 8.

The results demonstrate an accuracy rate of 99.79% in the predictions. This high accuracy indicates the excellent performance of the model. The LGBM model can effectively predict the scoring side for the next point based on the current match's data, thereby predicting the match fluctuation.

6.3 The application of the LGBM model in other matches

We utilized the constructed LGBM model to predict Gentlemen's matches at Wimbledon, yielding notably excellent forecasting results. However, the performance of this model in predicting other

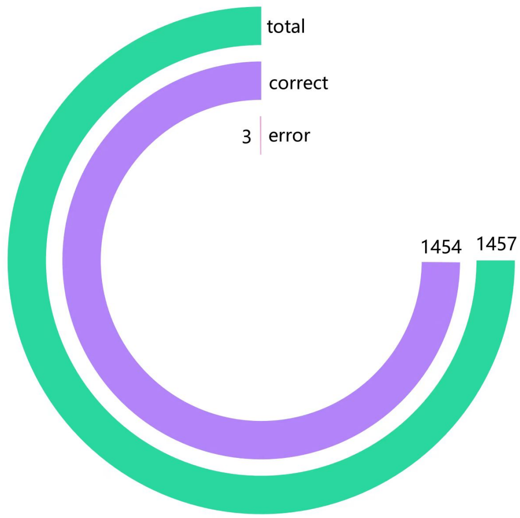
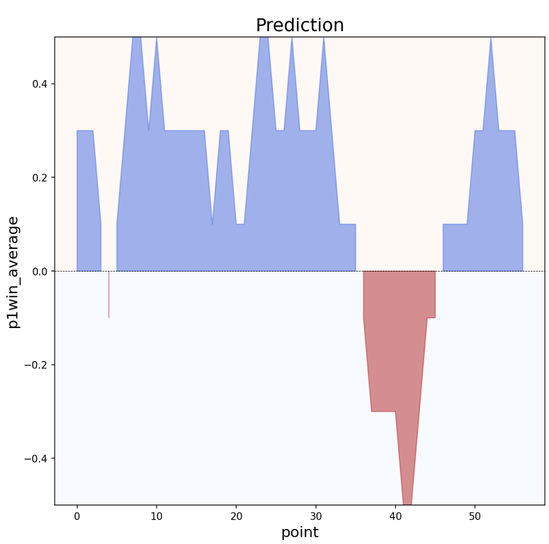


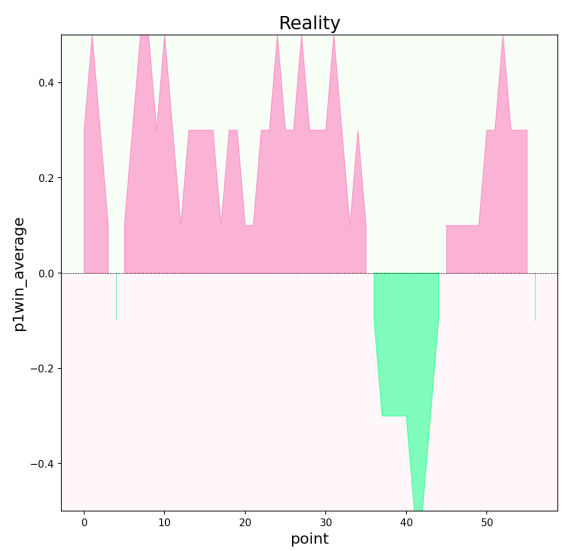
Figure 8: Prediction Accuracy of the LGBM Model

matches remains uncertain. In order to enhance the model's generalizability, it was applied to forecast women's tennis matches, followed by result validation and assessment.

We obtained several records of women's tennis match data from Kaggle datasets for testing and evaluation purposes. We predicted the outcome of a single tennis match, with the prediction results depicted in Figure 9(a). Concurrently, we plotted the actual results in Figure 9(b).



(a) Predicted results



(b) Real outcomes

Figure 9: Model prediction compared to actual outcomes

The comparison between predicted and actual outcomes was strikingly similar. The accuracy rates are displayed in Figure 10, where out of 62 test samples, 57 were correctly predicted and 5 were incorrectly predicted, resulting in a model accuracy of 92.54%. This exceptional accuracy underscores the model's robustness and effectiveness in diverse match scenarios, showcasing its outstanding performance characteristics.

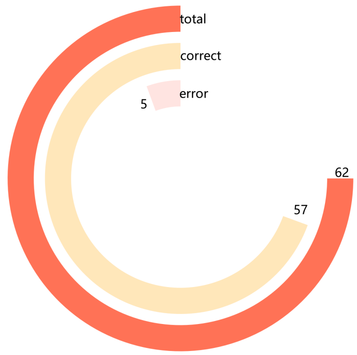


Figure 10: Prediction Accuracy of the LGBM Model

7 Problem 4 Model: LGBM-based evaluation model for assessing factor importance.

7.1 Evaluating the Importance of Factors Using the LGBM Model

In order to provide players with more practical decision advice, it is necessary to assess the importance of various factors that influence match fluctuations. In the LGBM model, we assess the importance of features based on their frequency of use in the model or the gain they bring from feature splits (i.e., the degree of improvement in model performance).

- Based on Split Count: We calculate the number of times each feature is used as a split node in all the trees of the model. If a feature is frequently used as a split node, it indicates its high importance for the model's decision-making.
- Based on Gain: We calculate the total gain from all the splits that each feature brings in the model. Gain can be understood as the performance improvement obtained by splitting on that feature.

By considering these two aspects, we analyze the importance of features, and the results are shown in Table 5.

Table 5: Feature Importance of the LGBM Model

Feature	Importance
ET	2052
P1_TR	2002
GN	834
P1_CB	686
SN	520

P1_W	392
PV	346
SV	330
P1_A	238
P1_DF	167
P1_UE	158

The importance of features is shown in the radar chart Figure 11.We can see that the importance of each feature is represented by the length of the corresponding spoke. The longer the spoke, the greater the importance of the feature.

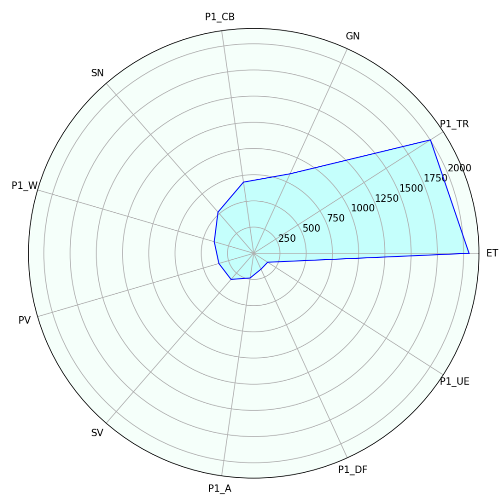


Figure 11: Radar Chart of Factor Importance

We also visualize the importance of features in a word cloud, as shown in Figure 12.

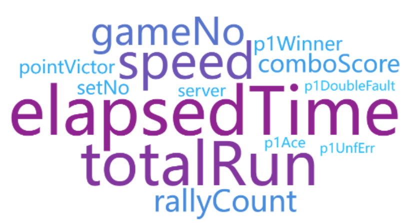


Figure 12: Word Cloud of Feature Importance

From the Table 5 and these two figures, it can be observed that match duration and total distance covered have the greatest impact on players' performance. This is quite intuitive, as with the progression of the match, the running distance continuously increases, posing an ever-growing challenge to the players' endurance, thus exerting the greatest influence on their overall performance.

On the other hand, while double faults and unforced errors do have an impact, it is not substantial, indicating that the influence of these two metrics is not as significant as initially assumed, and players need not overly concern themselves with these factors.

7.2 Recommendations for Players Based on Importance Results

In light of the importance analysis presented above, the following recommendations are proposed for players to better adjust their performance and achieve improved results in matches:

1. Match duration stands as one of the most crucial factors. Athletes are advised to focus on endurance and stamina training, encompassing aerobic exercises and high-intensity interval training. Moreover, it is essential to strategically manage physical exertion during matches to ensure consistent performance during critical moments.

2. The total number of sets won by athletes significantly impacts match outcomes. It is recommended for athletes to emphasize strategic planning and psychological resilience. By studying opponents, analyzing match scenarios, and devising appropriate tactical strategies, athletes can enhance their ability to win sets. Additionally, maintaining a positive mindset and self-confidence, persisting until the end without giving up easily, is crucial.

3. The number of games won by athletes in the current set notably influences their mental state. Athletes are advised to focus on decisive skills and strategies for each game. During pivotal moments, they should concentrate on every point, ensuring stability and confidence when it matters most. By identifying opponents' weaknesses and flexibly applying tactics and techniques, athletes can strive for victory in every crucial game.

4. The importance of double faults and unforced errors is relatively low. Athletes are advised not to overly dwell on mistakes during matches but instead focus on adjusting their state for upcoming competitions.

5. The number of aces and service winners is also crucial. Athletes are advised to focus on serving skills and strategies, as well as the ability to win points directly from serves. By improving serving accuracy and power, athletes can effectively control the game and gain an advantage.

Based on the importance results, we provide a suggestion for players, as shown in Figure 13.

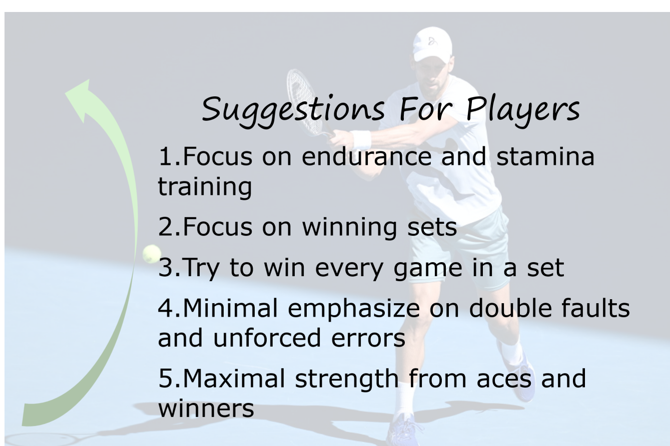


Figure 13: Suggestion for Players Based on Importance Results

8 Model Analysis

8.1 Strengths

1. In the model for Problem 1, we differentiate between the situation and momentum, giving a more complex definition of momentum, which is more important and variable in the game. As a result, we obtain more realistic outcomes.

2. In the model for Problem 2, we use a relatively simple model that is easy to understand, yet achieves good results.

3. The LGBM model for Problem 3 and Problem 4 is used for prediction. Not only does it have high accuracy in predicting outcomes for the given matches in the dataset, but it also achieves very high accuracy in predicting results for other matches.

4. The LGBM model we use is not only capable of making predictions but also able to assess the importance of various indicators based on the tree splits, resulting in two outcomes from a single model.

8.2 Weaknesses

1. The correlation analysis in Model 2 only provides correlations and does not directly establish causality. Therefore, further discussions may be needed on how momentum affects match flow.

2. We do not know the construction process of the LGBM decision tree in Model 3, which reduces its interpretability. We may need to employ additional methods to understand its internal workings.8.3 Sensitivity Analysis

8.3 Sensitive Analysis

Finally, to test the robustness of the model, we perturb the inputs to the model and observe the changes in the output results. We apply perturbations ranging from -0.5 to +0.5 to the normalized data and calculate the momentum based on this. The results are shown in Figure 14.

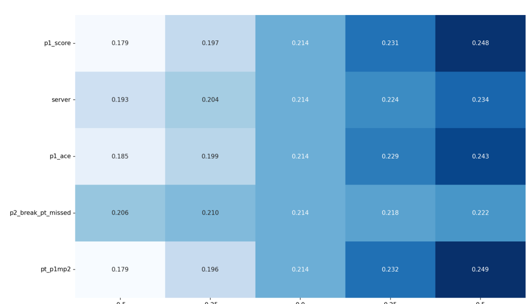


Figure 14: Sensitivity Analysis of the Model

We can see that the serve and the opponent's break point missed are the most sensitive to perturbations. However, overall, the fluctuations are still very small. This indicates that our model exhibits strong robustness and has good resistance to interference when faced with perturbations.

MEMORANDUM

FROM:TEAM# 2414434
TO:MCM
SUBJECT:MOMENTUM IN TENNIS

Tennis is such a passionate and energetic sport, every swing of the tennis game releases the fighting spirit of passionate blood, every spin and impact of the tennis ball is always showing the tension and excitement of the court. The situation changes rapidly and the momentum of the players seems to be unpredictable. Here we have created a mathematical model to assess the situation and momentum, to predict the changes and role of momentum in a match, and to give you and your tennis players some solid advice.

We have developed a model for evaluating player performance based on data from every point of all men's matches after the first two rounds of the 2023 Wimbledon Open, using factor analysis, hierarchical analysis, and non-linear regression to quantify the "momentum" of the players on the court. Further, we analyzed the correlation between the quantified situation and momentum, and concluded that momentum plays an important role in the game, and that the fluctuation of the game and the success of the players are not random. Then we use the LGBM model to predict the fluctuation of the situation of the game and analyze the side that is more likely to win at each point of the game accordingly. The model has been tested and performed well with very low errors, based on which we give some recommendations:

- Focus on endurance and persistence training, including aerobic and high intensity interval training.
  The model concludes that the time of the game and the running distance have a great weight on the situation, which indicates that the players must rationalize the use of physical strength during the game, and must not be impatient.
- Developing tough mental endurance, including the ability to stabilize mental qualities with the level of play. The analysis concluded that the number of leading games and the number of discs have a great influence on the situation, and the players should exercise the ability to play normally or even exceedingly well when they are behind, and always maintain stability and self-confidence.
- fully believe in their own strength level and play bravely and boldly. After analyzing, the importance of double serve errors and non-forcing errors is not high, players do not need to care too much about the errors in the match, but firmly believe in their own strength, adjust their mentality to maintain good play.

References

[1] Taylor, J., & Demick, A. (1994). A multidimensional model of momentum in sports. Journal of Applied Sport Psychology, 6(1), 51–70.

[2] Huang Zhiying. Research on the Prediction Model of Tennis Match Results Based on BP Neural Network [D]. Fujian Normal University, 2022. DOI: 10.27019/d.cnki.gfjsu.2022.000427.

[3] Gernigon, C., Briki, W., & Eykens, K. (2010). The Dynamics of Psychological Momentum in Sport: The Role of Ongoing History of Performance Patterns. Journal of Sport and Exercise Psychology, 32(3), 377-400. Retrieved Feb 5, 2024

[4] Perreault S, Vallerand RJ, Montgomery D, Provencher P. Coming from Behind: On the Effect of Psychological Momentum on Sport Performance. Journal of Sport and Exercise Psychology. 1998;20(4):421-436. doi:10.1123/jsep.20.4.421

[5] Giovanni Angelini, Vincenzo Candila, Luca De Angelis,Weighted Elo rating for tennis match predictions,European Journal of Operational Research,Volume 297, Issue 1,2022,Pages 120-132,ISSN 0377-2217,

[6] Baiget E, Fernández-Fernández J, Iglesias X, Rodríguez FA. Tennis Play Intensity Distribution and Relation with Aerobic Fitness in Competitive Players. PLoS One. 2015 Jun 22;10(6):e0131304. doi: 10.1371/journal.pone.0131304. PMID: 26098638; PMCID: PMC4476777.

[7] Zhang, T., & Yan, H. (2022). Optimization and application of multi-attribute decision-making algorithm based on entropy weight method. Journal of North China University of Science and Technology (Natural Science Edition), 44(01), 82-88.

[8] Jin, L., & Li, Y. (2019). Discrimination of several correlation coefficients and their implementation in R language. Statistics and Information Forum, 34(04), 3-11.

[9] Meffert D, O'Shannessy C, Born P, Grambow R, Vogt T. Tennis serve performances at break points: Approaching practice patterns for coaching. Eur J Sport Sci. 2018 Sep;18(8):1151-1157. doi: 10.1080/17461391.2018.1490821. Epub 2018 Jul 11. PMID: 29996725.

[10] Starczewski T. Simulation Research on the Relationship between Selected Inconsistency Indices Used in AHP. Entropy (Basel). 2023 Oct 19;25(10):1464. doi: 10.3390/e25101464. PMID: 37895585; PMCID: PMC10606368.

[11] Leal JE. AHP-express: A simplified version of the analytical hierarchy process method. MethodsX. 2019 Dec 4;7:100748. doi: 10.1016/j.mex.2019.11.021. PMID: 32021813; PMCID: PMC6993013.

9 Appendices

9.1 Appendix 1: Python Code for Predicting Momentum of Match Points Using AHP and Polynomial Nonlinear Regression.

```python
import pandas as pd
from factor_analyzer import FactorAnalyzer
import matplotlib.pyplot as plt
from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity
import numpy as np
from sklearn.preprocessing import PolynomialFeatures
from sklearn.linear_model import LinearRegression

# Read data and select columns
data = pd.read_csv('data.csv')
input_cols = ['p1_double_fault', 'p1_server', 'p1_ace', 'total_p1adv',
    'p1_totalrun', 'speed_mph']
output_cols = ['p1_break_pt_won', 'p1_winner']
input_data = data[input_cols]
output_data = data[output_cols]

# Factor analysis
fa = FactorAnalyzer()
fa.fit(input_data)
factor_loadings = fa.loadings_

# Normalize and add factors to data
for i in range(fa.n_factors_):
    fi = pd.Series()
    for rn, rd in input_data.iterrows():
        num = sum(rd[name] * factor_loadings[j][i] for j, name in
            enumerate(input_cols))
        fi[rn] = (num - num.min()) / (num.max() - num.min())
    data[f'f{i+1}'] = fi

data.to_csv("data.csv", index=False)

# Output communalities and explained variance ratio
communality = fa.get_communalities()
explained_variance = fa.get_factor_variance()
print("Communality:\n", communality)
print("Explained Variance Ratio:\n", explained_variance)

# Calculate AHP weights
comparison_matrix = np.array([[1, 2, 3], [1/2, 1, 2], [1/3, 1/2, 1]])
weights = np.mean(comparison_matrix, axis=1)

```

```
# Non-linear regression
X = np.column_stack([data[f'f{i+1}'].values for i in range(fa.n_factors_)])
y = sum(data[f'f{i+1}'].values * weights[i] for i in range(fa.n_factors_))
poly_features = PolynomialFeatures(degree=2)
X_poly = poly_features.fit_transform(X)
regression_model = LinearRegression().fit(X_poly, y)
y_pred = regression_model.predict(X_poly)

# Plot prediction vs actual scores
fig, ax = plt.subplots()
ax.plot(range(300), y_pred[:300], label='Prediction', color="orange")
ax.plot(range(300), y[:300], label='Actual', color="blue")

# Customize plot
ax.set_xlabel('point', fontsize=15)
ax.set_ylabel('score', fontsize=15)
ax.axhspan(-5, 5, color="#F5FFFA")
ax.set_ylim(0, 5)

ax.legend()
plt.show()
```

Artificial Intelligence Usage Report

In this mathematical modelling, we used ChatGPT based on GPT-3.5 architecture with September 2021 knowledge deadline. We used gpt to help us in refining our modelling ideas, correcting textual syntax errors, modifying code and other uses to complete our tasks.

1. OpenAI *ChatGPT*（2021.9，ChatGPT-3.5）
   Query1: <*ValueError: all arrays must be same length This is a runtime error.*>
   Output: < This error usually occurs in Numpy array operations, and is caused by an array length mismatch.

   In the case of this Q-learning code, there are several possible reasons for this error.

   State and action definitions don't match and have different lengths. For example, if the state is (score1, score2), the action is defined as score1.

   The length of the state data does not match the length of the Q table structure, so the corresponding Q value cannot be found in the table.

   The length of reward definition does not match with other data.

   The length of state and action data used in train function to calculate q_predict, q_target is wrong.

   The length of the column (state) definition is wrong when the Q-table is initialised.

   To troubleshoot the cause, you can check.

   Whether the state and action space definitions match completely.
   Whether the data length is consistent with the Q-table structure.
   whether the column defined by reward exists
   whether the core computation in the train function uses the correct data.
   Q-table initialisation process
   Generally speaking, this error can be caused by length issues in all areas involving arrays or Series calculations.

   You need to debug with breakpoints or print key data to locate the mismatch and then adjust the definition to solve the problem.>。

2. OpenAI *ChatGPT*（2021.9，ChatGPT-3.5）
   Query1: <*Can you give me an example of how to use a grey predictive model and a Markov chain model for tennis match analysis?*>
   Output: < Here is a simple example of how to combine a grey prediction model and a Markov chain to analyse a tennis match.

   Defining the grey state space
   The score states use three equally spaced grey sets: {Near set, Tied set, Ahead set}, which represent the relative strengths of the players at the time.

   Construct the state transfer relation matrix
   Based on the historical match points, the probability of the next state change of a player in each grey state is counted.

   Initialise the a priori prediction matrix
   The confidence level at the beginning of the first game is {0.5,0.5,0}, indicating the initial player strength balance.

   Recursively calculate the a posteriori prediction
   Iteratively calculate the a priori matrix for each match point using the grey Markov model

formulation.

Output prediction results
Given the hopefulness of the players in each grey state, the expected outcome curve.

Backpropagation correction
The model is updated according to the actual scores and the error is minimised to make the next prediction more accurate.

This combines the state transfer properties of Markov chains with the strengths of grey systems in uncertainty reasoning. The quality of prediction is continuously improved through iterative calculations.>。

3. OpenAI *ChatGPT*（2021.9，ChatGPT-3.5）
Query1: <*Can you explain how the logistic regression model predicts point_victor?*>
Output: < Ordinary logistic regression models are used to make categorical predictions about nominal variables.
In this task, I utilise server as the main feature of prediction.
The logistic regression assumes that.
The probability of winning is higher for the server (serving side) than for the served side.
Other conditions (set_no, game_no) being equal, the server has the highest probability of winning.
Training procedure.
Encode point_victor as 0 (the served side wins) and 1 (the server wins).
Add a feature set to each sample: [accept, server, set_no, game_no].
Use maximum likelihood estimation to train the model parameters (weight vectors) so that the predicted probability closest to the true probability.
After obtaining the model, we can calculate the predicted category probability for each sample.

Prediction.

Calculate the predicted category probability based on the characteristics of the new sample

Select the category with high probability as the predicted value

So in this task, the model uses the a priori knowledge of the dominance of the server to predict the probability of different combinations of feature values.

Of course, this is only a simple demonstration, and more complex time-series models are needed to actually predict "momentum" shifts and so on.>。

4. OpenAI *ChatGPT*（2021.9，ChatGPT-3.5）
Query1: <*Can you explain how a finite state automaton can be embedded into a Markov chain state machine to improve accuracy?*>
Output: < Embedding a Finite State Automata (FSA) into a Markov chain state machine can improve the accuracy of tennis match prediction in this way.

FSA defines the legal state transfer path
According to the rules of tennis, define the legal trajectory of the score of each set and each match.

Initialise the FSA state machine
Use the number of match points as a time-driven state transfer.

Incorporate Markov model
Use FSA state as Markov chain state space, construct state transfer probability matrix.

Prediction of path constraints
Skip Markovian low probability but illegal state transfer paths based on FSA restrictions.

Iteratively calculate the probability
Considering both FSA path restriction and Markovian probabilistic transfer characteristics, update the state distribution.

Output prediction results
The winning percentage of a player in each legal state is given, forming a compact expected trend curve.

Backpropagation correction model
Update the FSA state and probability matrix according to the actual score to improve the next prediction performance.

In this way, through the path restriction of FSA, some unreasonable outputs of Markov model can be eliminated, while retaining its statistical learning effect. It helps to improve the overall prediction accuracy of the tennis schedule model.>。

5. OpenAI *ChatGPT*（2021.9，ChatGPT-3.5）
Query1: <*What is the appropriate size of the approximate chi-square for Bartlett's test of sphericity?*>
Output: < In the Bartlett's test of sphericity, we use an approximate chi-square distribution to compute a statistic that assesses whether the correlations between the variables in a factor analysis model are high enough for factor extraction. This statistic is calculated by comparing the observed matrix with the theoretical diagonal matrix.

When performing the Bartlett's test of sphericity, we want the observed matrix to be not significantly different from the theoretical diagonal matrix, i.e., the p-value of the statistic should be large enough. Usually, we choose a significance level (e.g., 0.05) as a judgement criterion, and if the p-value is greater than that significance level, then the observed matrix is similar to the theoretical diagonal matrix, and factor extraction can be performed.

Therefore, for the Bartlett's test of sphericity, the larger the p-value, the better, and a p-value close to 1 indicates that the observed matrix is very close to the theoretical diagonal matrix, and it can be assumed that the factor analysis model is suitable for the dataset.

It should be noted that Bartlett's test of sphericity is not the only method for assessing the applicability of factor analysis, but can also be combined with other indicators and methods to make a comprehensive judgement.>。

6. OpenAI *ChatGPT*（2021.9，ChatGPT-3.5）
Query1: <*I have thirty items in the dataset, it is about the variation of data over time, how to make time predictions without using the arima model*>
Output: < If you only have 30 time series data points and do not want to use an ARIMA model for time forecasting, here are some alternatives to consider:

Linear regression: if you believe that there is a linear relationship between the data, you can use a linear regression model for time forecasting. Linear regression can help you find the line of best fit to make predictions about future time data.

Moving Average: This method predicts future time data by calculating an average over a time window. You can predict the next point in time based on the average of the last few data points.

Exponential Smoothing: Exponential smoothing is a commonly used time series forecasting method that puts a higher weight on the newest data and gradually reduces the weight of older data. This can help capture trends and seasonal variations in the data.

Non-Linear Regression: If you think there is a non-linear relationship between the data, you can try using non-linear regression models such as polynomial regression or neural networks. These models can better capture complex data patterns.

Time series decomposition: You can decompose time series data into components such as trend, seasonality and residuals and model and forecast each component separately. This provides a better understanding of trends and seasonal patterns in the data.

It is important to note that when choosing an alternative method, base your choice on the characteristics of your data and forecasting goals. Also, for time series forecasting, more data usually leads to more accurate forecasts. If possible, try to collect as much data as possible to improve your forecasting ability.>。