Problem Chosen	2024 MCM/ICM Summary Sheet	Team control number
C		2417022

Uncover the secrets behind Alcaraz's success in Wimbledon

Abstract

Abstract goes here.

Contents

1 Introduction

1.1 Background

Tennis originated in the 13th century in France. As early as the 16th to 17th centuries, French missionaries often played a game similar to tennis in the corridors of churches, using their hands to strike a small ball, providing a diversion from the monotonous church life. Now, modern tennis has formally developed and quickly gained popularity in Europe and America, becoming a widely loved sport. [1] Tennis, as a charming and elegant sport, enjoys a high reputation and strong influence internationally. [2] With the continuous development of tennis, the fluctuations in the scores of opponents have become increasingly scrutinized during tennis matches. Recently, in the men's singles final at Wimbledon in 2023, 20-year-old Spanish rising star Carlos Alcaraz defeated 36-year-old Novak Djokovic, ending Djokovic's winning streak at Wimbledon since 2013. The twists and turns in the score and the changing dynamics of the match attracted significant attention. Therefore, effectively exploring the impact of the ”momentum” on the score during the game is crucial.

1.2 Problem restatement

1. Develop a mathematical model to quantify and visualize the progression of tennis matches, identifying key performance metrics for players during specific time intervals, with a particular focus on the service advantage.

2. Construct a statistical framework to assess the impact of momentum within a match, challenging the notion that players' swings in performance are random, and providing empirical evidence for or against this assertion.

3. Formulate a predictive model capable of indicating potential shifts in match dynamics, utilizing data from previous matches to identify indicators that signal when the flow of play may change in favor of one player over another.

4. Execute a comprehensive testing protocol for the aforementioned models across a series of matches, evaluating the predictive capabilities with respect to different match conditions, and identifying additional variables that may enhance future model iterations. Analyze the generalizability of the models to other match formats and sporting contexts, such as women's tennis, various tournaments, court surfaces, and racket sports akin to table tennis.

1.3 Our Work

Firstly, the choice of model should be grounded in the context of the problem, taking into account the importance of momentum in tennis matches and its impact on player performance. Therefore, the model needs to capture the dynamic changes during a match and reflect the performance of players at different points in time. Factors such as scoring probability, fatigue levels, technical skills, and the mental state during the match are crucial indicators of momentum change, so the chosen model should integrate these factors to predict the scoring probability of players. Given that the dataset provides detailed match scores and other relevant statistics, machine learning algorithms like LGBM and XGBOOST are suitable choices because they can handle a large number of features and are applicable to classification problems. Moreover, considering the sequential nature of scoring in tennis matches, time series analysis or Hidden Markov Models could be used to capture the momentum changes throughout the match. Taking all these factors into account, opting for gradient boosting algorithms (such as LGBM or XGBOOST) for modeling is a wise decision. These algorithms perform well in dealing with complex relationships and nonlinear patterns and can manage large datasets efficiently. Additionally, incorporating sequence models to capture temporal momentum changes should also be considered. The final model selection will be based on predictive accuracy (metrics such as accuracy, recall, precision, F1 score, AUC, and ROC curves) and the model's interpretability. Such a model will be able to assess the impact of ”momentum” on the probability of scoring for players and provide insights into the presence of non-random fluctuations in matches.

2 Problem Analysis

In the first study, multiple key factors affecting the scoring abilities of tennis players were comprehensively considered. This included not only whether they were serving but also factors like their level of fatigue, technical skills, and psychological state during the match. A comprehensive evaluation system was designed to deeply analyze these variables, such as individual technical levels, degrees of fatigue, and real-time mental states during matches. By employing statistical Logistic regression, the proposed system of indicators was tested for its correlation with players' scoring, and its effectiveness was validated on a dataset of tennis matches. Machine learning techniques, including LGBM, XGBOOST, support vector machines, perceptron networks, and logistic regression, were applied along with five-fold cross-validation and multi-dimensional performance metrics. The gradient boosting algorithm LGAM was ultimately chosen to dynamically assess players' real-time performance, or what is referred to as match "momentum."

The second study focused on exploring whether the "momentum" effect during matches has statistical significance, that is, whether the fluctuations in players' scoring during matches are non-random. The optimized machine learning model developed in the first study was applied to a test dataset, with predicted scoring probabilities compared to actual scoring records of athletes, and validated through Pearson correlation analysis. Subsequently, a univariate linear regression model was established, using predicted "momentum" as the independent variable and players' scoring as the dependent variable, to quantify the impact of "momentum" on scoring.

The third study addressed the challenges of model construction, as the model from the first study primarily focused on predicting the scoring probability of each serve and shot, rather than volatility prediction. To bridge this gap, the predictive focus was adjusted from individual shots to the outcome of entire games, analyzing fluctuations in players' performances across individual sets and the entire match. By aggregating features, a predictive model was built using statistical Logistic regression, with the support vector machine chosen as the final model. Model performance was evaluated using LGBM, XGBOOST, recall, precision, F1 score, AUC, and ROC curves. The information gain method was employed to determine the importance of various variables in the indicator system, leading to recommendations for athletes' future performance in matches.

In the final study, by selecting four matches as a validation set and using other match data as a training set, the visualization method of the ROC-AUC curve showcased the model's performance differences across various matches. Considering the uncertainty of the "volatility" indicator, machine learning evaluation standards were used for performance evaluation. In tests across these four matches, the constructed model demonstrated significant performance differences. Given this, expandable research directions were proposed, emphasizing the importance of individual athlete abilities and proving the applicability and strong generalizability of the proposed indicator system in other competitions

3 Model Assumption

1. The server has a greater advantage over the receiver.This assumption posits that the player initiating the serve in a game is more likely to gain an upper hand.

2. Players are affected by the "momentum" of the match.It's assumed that the success or failure of players during a match influences their future performance, suggesting that momentum plays a critical role in the outcome of the game.

3. A player's technical skill has a significant and direct impact on their scoring performance.This assumption highlights the importance of individual skills and how they translate into points on the board.

4. Fatigue affects players during the match, impacting their performance in the later stages.This suggests that as the match progresses, players' energy levels and their ability to maintain peak performance diminish.

5. The psychological state of players, including their ability to handle mistakes and respond to critical moments, significantly influences the match.This assumption underscores the importance of mental toughness and psychological preparedness in determining a player's success in competition.

4 Symbol Description

Symbol	Description
$S1$	The number of game winning in the current set
$S2$	The leading scores in the current game
$S3$	Whether it is the server
$S4$	Whether score in the last point
$S5$	The score lead progress of this match
$S6$	Whether the serve is scored (no contact)
$S7$	Whether to score on a return kick (no touch).
$S8$	No touch score on the backhand
$S9$	Is there a double fault in this game?
$S10$	Whether there were unforced errors in this game.
$S11$	The ratio of the number of net to the number of times score by net
$S12$	The ratio of the chance of scoring when the opponent serves to the number of points actually scored.
$S13$	Total mileage in this match
$S14$	The total mileage in the last three points
$S15$	Mileage chart from last point
$S16$	Serve real-time pace

Table 1: Symbol Description

5 Model Construction and Analysis

5.1 Task1's Model

5.1.1 Data Preprocessing

In the initial phase of data preprocessing, the dataset is imported from a .csv file, and the Pandas library is employed to structure the data into a DataFrame. Subsequent to this, the dropna function is applied to eliminate records with missing values. Following the initial data preparation, the script undertakes a comprehensive feature engineering process, introducing novel features pertinent to the distinct characteristics of tennis matches.

In the subsequent phase, the dataset is readied for modeling through a meticulous iteration over each data point. This involves the extraction of features and the assignment of labels contingent upon the outcome of each point, specifically focusing on point victories. To enhance the robustness and suitability of the data for subsequent machine learning applications, the features undergo scaling. The MinMaxScaler from the scikit-learn library is employed for this purpose, ensuring that all variables are standardized within the [0, 1] range. This standardization is imperative to mitigate the undue influence of certain features during the modeling process, arising from variations in scale.

The finalization of the preprocessing sequence involves the preservation of the processed and standardized dataset in an Excel file. These methodical steps collectively address challenges associated with missing values, outliers, and standardization, culminating in the generation of a refined dataset poised for efficacious utilization in machine learning endeavors.

5.1.2 Model Analysis

The analysis of player performance in tennis is a multifaceted task that requires a comprehensive approach, blending traditional statistical methods with advanced machine learning techniques. In this section, we delve into the intricacies of our model, examining the preprocessing steps, model training, evaluation metrics, and the overall performance against various benchmarks.

Initially, we employed logistic regression to assess the statistical significance of the relationship between our established indicator system and the corresponding labels. Leveraging a tennis match dataset, we computed these indicators and labels indicating player scoring. This step not only aids in comprehending how various factors impact a player's score but also lays a crucial groundwork for feature selection in the modeling phase.

Subsequent to this, we utilized a range of machine learning algorithms, including LGBM, XGBoost, Support Vector Machines, Perceptron Networks, and Logistic Regression. We compared their effectiveness using metrics like accuracy, recall, precision, F1-score, AUC, and ROC curves. This step not only robustly supports our final choice of the LGAM algorithm for modeling but also ensures the model's capacity for generalization.

In the concluding phase, we opted for the LGAM algorithm for modeling. This algorithm not only dynamically evaluates a player's performance but also captures shifts in "momentum" during a match. This dynamic assessment model facilitates more accurate predictions of a player's performance, enhancing the precision of predicting match outcomes. Through this research, we have taken a noteworthy stride in comprehending and appraising player tactics and scoring, establishing a strong basis for more in-depth analysis and prediction.

Concerning the creation of the indicator system, we analyzed potential factors influencing a tennis player's scoring. The system was formulated based on elements like player fatigue, real-time and historical scoring situations, and the player's psychological state. This thorough analysis contributes to a nuanced understanding of player performance, paving the way for further exploration and prediction.

5.1.3 Index creation

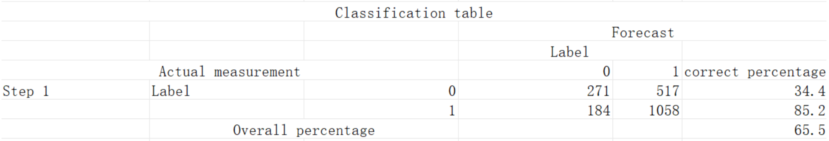
we analyze the factors that might correlate with the player's score. According to the player's fatigue and mental state etc., make an index system as follows.

Category	**Index**
Fatigue	Total mileage in this match
Fatigue	The total mileage in the last three points
Fatigue	Whether score in the last point
State	Whether there is a double fault
State	Whether there were unforced errors in this game.
State	Serve real-time pace
State	Whether it is an interactive item for the server's serving pace
State	The ratio of the number of net to the number of times score by net
State	The ratio of the chance of scoring when the opponent serves to the number of points actually scored.
Ability	Whether the serve is scored (no contact)
Ability	Whether to score on a return kick (no touch).
Ability	No touch score on the backhand
Ability	The score lead progress of this match
Ability	The leading scores in the current game
Ability	The number of game winning in the current set
Ability	Whether score in the last point
Situations	Whether it is the server

Table 2: Symbol Description

5.1.4 Statistical Logistic Regression Analysis Based on Indicator System

This study employs a comprehensive indicator system to assess the impact of various factors on the real-time scoring of players in sports matches. The indicator system allows for the calculation of scores for each point in every set of every match played by a participant, with corresponding assessments of the player's current performance indicators. To validate the significance of the indicators within this system concerning point scoring, a statistical regression analysis is conducted. Given the binary nature of point scoring outcomes, a binary logistic regression model is selected, and the analysis is executed using SPSS software.

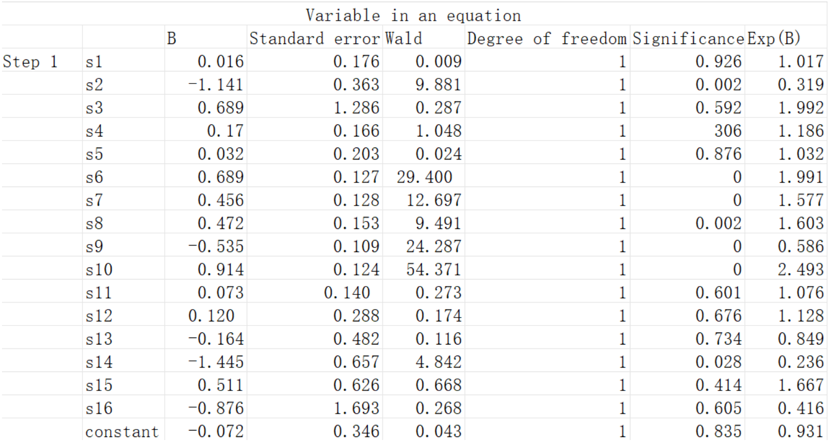


Classification table					
			Forecast		
			Label		
	Actual measurement		0	1	correct percentage
Step 1	Label	0	271	517	34.4
		1	184	1058	85.2
	Overall percentage				65.5

Figure 1: logistic regression performance

The logistic regression analysis in SPSS yielded an overall accuracy of 65.5%. The model exhibited a classification accuracy of 34.4% for instances where players genuinely did not score, while achieving a significantly higher accuracy of 85.2% for instances where players did score. This suggests a preference of the model to classify samples into situations where players actually scored (label 1).

However, it is crucial to note that this logistic regression analysis provides a prospective examination. Its purpose is to investigate whether the constructed indicators significantly influence the actual scoring situation of players. The results of the logistic regression analysis are presented in the following table.



Variable in an equation							
		B	Standard error	Wald	Degree of freedom	Significance	Exp(B)
Step 1	s1	0.016	0.176	0.009	1	0.926	1.017
	s2	-1.141	0.363	9.881	1	0.002	0.319
	s3	0.689	1.286	0.287	1	0.592	1.992
	s4	0.17	0.166	1.048	1	306	1.186
	s5	0.032	0.203	0.024	1	0.876	1.032
	s6	0.689	0.127	29.400	1	0	1.991
	s7	0.456	0.128	12.697	1	0	1.577
	s8	0.472	0.153	9.491	1	0.002	1.603
	s9	-0.535	0.109	24.287	1	0	0.586
	s10	0.914	0.124	54.371	1	0	2.493
	s11	0.073	0.140	0.273	1	0.601	1.076
	s12	0.120	0.288	0.174	1	0.676	1.128
	s13	-0.164	0.482	0.116	1	0.734	0.849
	s14	-1.445	0.657	4.842	1	0.028	0.236
	s15	0.511	0.626	0.668	1	0.414	1.667
	s16	-0.876	1.693	0.268	1	0.605	0.416
	constant	-0.072	0.346	0.043	1	0.835	0.931

Figure 2: Logistic regression results

As the figure table shows, the p-values for the majority of the independent variables are less than 0.05, indicating a significant influence of these variables on the dependent variable. Consequently, in this indicator system, variables such as s2, s6, s7, s8, s10, and s14 have been identified as significantly affecting a player's real-time scoring situation. This suggests that both a player's individual abilities and factors like fatigue or psychological state can exert a significant impact on their real-time scoring performance.

5.1.5 Evaluation of Machine Learning Models Based on an Indicator System

The model construction phase involves the selection of high-performing ensemble tree models, specifically LightGBM (LGBM) and XGBoost, alongside classical machine learning algorithms, namely Support Vector Machine (SVM), Perceptron Neural Network, and Logistic Regression, serving as comparative benchmarks. The models are validated using k-fold cross-validation, and performance metrics such as accuracy, precision, recall, F1 score, and AUC-ROC curve are derived from confusion matrices.

The confusion matrix-based results are presented in the table below, while the training and testing outcomes based on AUC-ROC curves are illustrated in the figures.

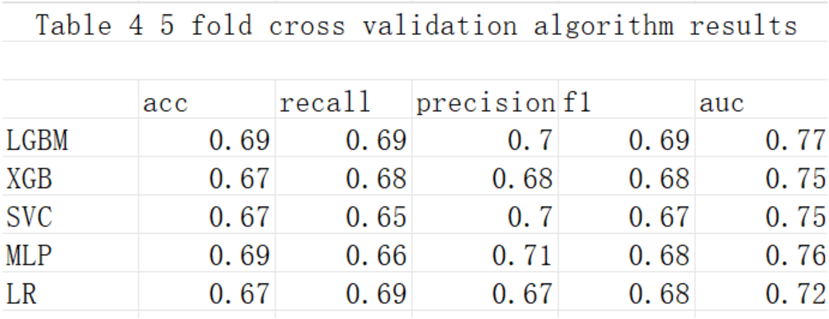
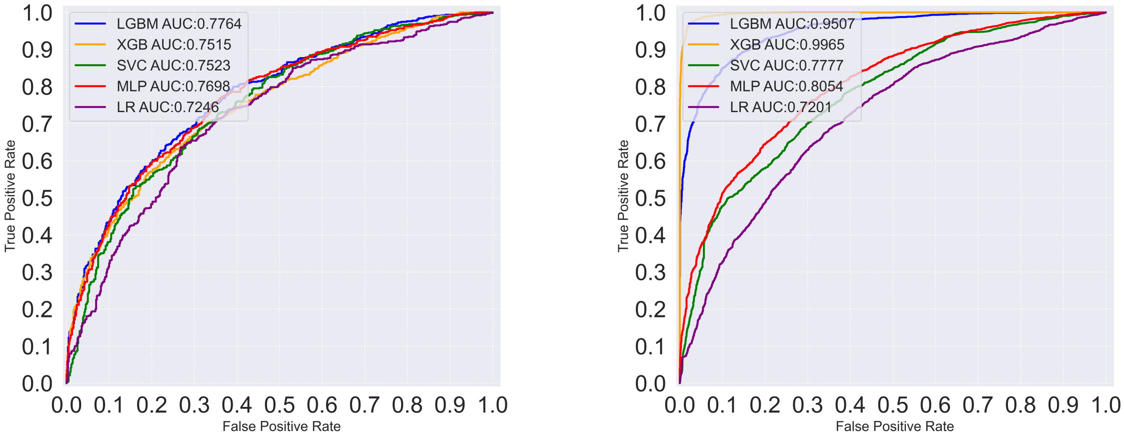


Table 4 5 fold cross validation algorithm results

	acc	recall	precision	f1	auc
LGBM	0.69	0.69	0.7	0.69	0.77
XGB	0.67	0.68	0.68	0.68	0.75
SVC	0.67	0.65	0.7	0.67	0.75
MLP	0.69	0.66	0.71	0.68	0.76
LR	0.67	0.69	0.67	0.68	0.72

Figure 3: Cross Valadation

Among the models, LGBM exhibits superior performance with accuracy, precision, recall, F1 score, and AUC values of 0.69, 0.69, 0.77, 0.69, and AUC_L respectively. Neural Network follows closely, demonstrating minimal discrepancies in metrics (0.69, 0.66, 0.71, 0.68, AUC_NN), indicating a lack of pronounced bias in positive or negative sample discrimination. Consequently, the effectiveness of LGBM is affirmed.



(a) Test

(b) Train

Figure 4: Task 1 Overall performance of the model

The ROC curves illustrate variations in precision and recall metrics at different threshold settings. Through ROC analysis, it further validates LGBM as the most effective algorithm.

Building upon these findings, a new training model is established utilizing the top-performing LGBM model. Real-time performance visualization is conducted on the classic match between 20-year-old rising Spanish star Carlos Alcaraz and 36-year-old Novak Djokovic in the 2023 Wimbledon Men's Singles Final. The result is shown as follows.

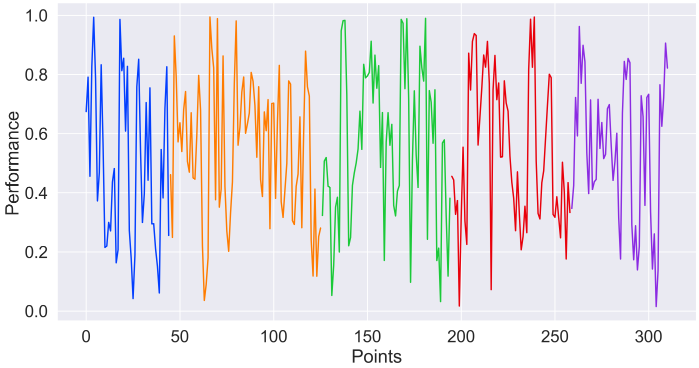


Figure 5: Classic duel

In fact, the predictive accuracy for scoring points has no upper limit, which aligns with our intuition. The complexity of factors influencing whether a player can score is highly intricate. Without considering these complex elements, the noise in a player's performance becomes significantly pronounced. As a result, the model can accurately discern the real scoring situations of only approximately 70% of the players. We observe that in this classic match, the server and scorer precisely align with the actual scoring situations of the players. When Carlos Alcaraz secures victory, the momentum is high, and it is relatively low when facing defeat. This indicates that the model remains effective, as it reflects the consistency between the classic battle outcomes and the real-time scoring situations of the players.

5.2 Task2's Model

5.2.1 Task2's Model Analysis

In the second phase of our research, we are dedicated to gaining a deeper understanding of the practical role of "momentum" in sports competitions. The aim is to unveil whether the fluctuations and successes exhibited by players during a match demonstrate a non-random trend. To achieve this objective, I initially applied the optimal machine learning model established in question one to the test set, generating probability outputs for player scoring. Through Pearson correlation tests between these outputs and the actual player scoring situations, we attempt to identify the degree of association between the output probabilities and the real scoring outcomes.

Furthermore, we employ a univariate linear regression model, where the predicted "momentum" serves as the independent variable and the player's actual scoring as the dependent variable. Through this modeling process, we not only track the impact of momentum on real-time scoring but also endeavor to uncover the interpretable components behind the "momentum." This in-depth analytical process aims to reveal the exact mechanisms through which "momentum" influences player scoring, providing a more comprehensive understanding of the dynamic characteristics of player performance.

Within this research framework, we anticipate concluding that "momentum" is not merely a simple random phenomenon but indeed exerts a noticeable and meaningful impact on player scoring. Through these efforts, we aim to offer a more detailed and profound insight into the practical significance of momentum in sports competitions, providing specific recommendations and guidance for player performance during matches.

5.2.2 Pearson Correlation Test

Principles of Pearson Correlation Test

The Pearson correlation coefficient is a statistical measure used to assess the linear relationship between two variables, with values ranging from -1 to 1. Specifically:

- A value of 1 indicates a perfect positive correlation: when one variable increases, the other variable increases correspondingly.

- A value of -1 indicates a perfect negative correlation: when one variable increases, the other variable decreases.
- A value of 0 suggests no linear relationship.

The calculation formula for the Pearson correlation coefficient (often denoted by the symbol r) is as follows:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}$$

Here, x_i and y_i are the corresponding data points, $\bar{x}$ and $\bar{y}$ are the respective means of the variables. The numerator represents the sum of the product of the deviations of each data point from their respective means, while the denominator signifies the square root of the product of the sums of squared deviations for both variables.

A computed r value closer to 1 or -1 indicates a stronger linear relationship between the two variables. Conversely, an r value close to 0 suggests a lack of linear relationship between the variables.

Pearson Correlation Test Results

As the goal is to demonstrate whether the momentum output by the model can influence the players' actual performance, it is necessary to establish the significance of their correlation.

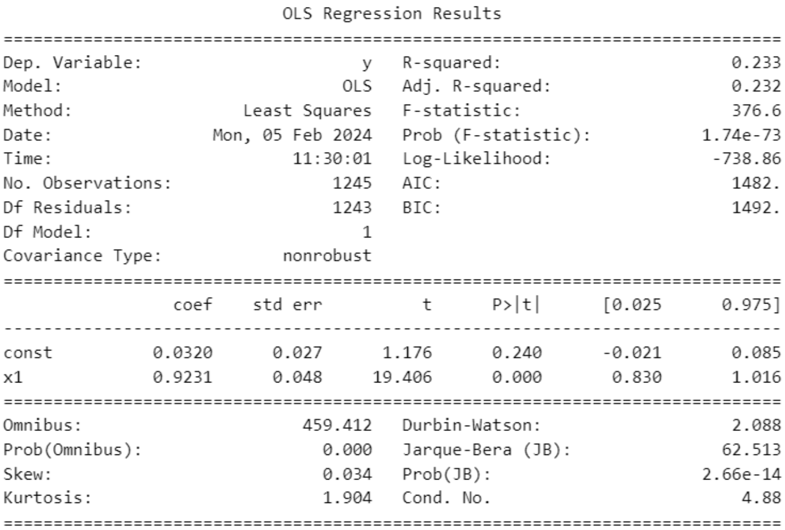
Correlation coefficient	P Value
0.482	0.000

Table 3: Pearson result

It is observed that the p-value is less than 0.05, indicating a significant correlation between momentum and player performance. Furthermore, this relationship is characterized as a positive correlation, signifying that a larger momentum corresponds to better player performance and an increased likelihood of scoring.

5.2.3 Univariate Linear Regression Model

Similarly, we use linear regression to analyze whether momentum can significantly affect player scores. The result is as shown below:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.233
Model:                            OLS   Adj. R-squared:                  0.232
Method:                 Least Squares   F-statistic:                     376.6
Date:                Mon, 05 Feb 2024   Prob (F-statistic):           1.74e-73
Time:                        11:30:01   Log-Likelihood:                -738.86
No. Observations:                1245   AIC:                             1482.
Df Residuals:                    1243   BIC:                             1492.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0320      0.027      1.176      0.240      -0.021       0.085
x1             0.9231      0.048     19.406      0.000       0.830       1.016
==============================================================================
Omnibus:                      459.412   Durbin-Watson:                   2.088
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               62.513
Skew:                           0.034   Prob(JB):                     2.66e-14
Kurtosis:                       1.904   Cond. No.                         4.88
==============================================================================
```

Figure 6: Univariate regression

It can be seen that the p-value of momentum is also less than 0.05, and its weight is 0.9231. The impact of momentum x1 on the results is significant, and it can explain 0.233 of the results.

5.3 Task3's Model

In the study of question three, we deeply analyzed the limitations of the established model. This model mainly focuses on predicting the score of each player's serve and stroke, and is therefore limited to having high randomness. ,To overcome one of the limitations, we adopted a ,more nuanced approach designed to more accurately reflect ,player performance fluctuations throughout the game.

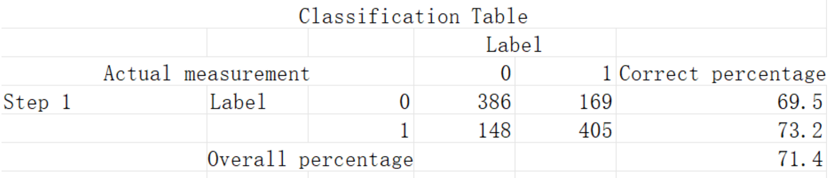
First, we upgraded the model in Question 1 to aggregate features and calculate the player's single game win. This new prediction goal provides us with more specific and accurate data, allowing us to better understand Player performance in the game. Subsequently, we again used statistical LOGISTI to build a model to determine the significant impact of the proposed indicator system on predicting game results.

Comparing the results of Question 1, we found that the indicator system proposed in the text has a more significant impact on the independent variables in game prediction. In the comparative analysis of multiple machine learning algorithms, especially LGBM, XGBOOST, support vector machine, perceptron network, and logistic regression, the support vector machine performed well and became the best model we selected. In order to further strengthen the model For interpretability, we used the information gain method to calculate the importance score of the indicator system, which helps us understand the specific contribution of each indicator to the players.

In the end, we not only provided an excellent support vector machine model for modeling player performance fluctuations, but also gained a deeper understanding of the importance of the indicator system through the analysis of information gain. At the end of the study, we provide performance recommendations for players when entering new games, aiming to help players achieve better performance in the game. These recommendations are not only based on our model results, but also incorporate the overall status of the players and specific situations. Deep insights provide players with more comprehensive support.

5.4 Logistic test with game as granularity

Similarly, we use statistical logistic regression for testing. The results are shown in the table

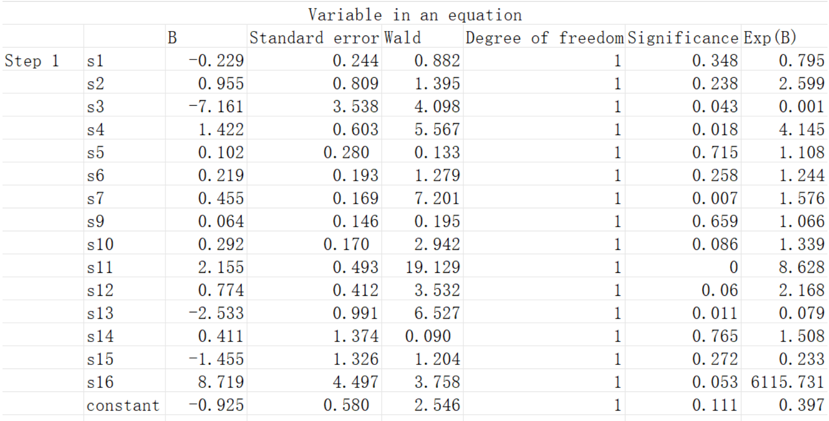


Classification Table

			Label		
Actual measurement			0	1	Correct percentage
Step 1	Label	0	386	169	69.5
		1	148	405	73.2
	Overall percentage				71.4

Figure 7: Logistic regression performance

It can be seen that compared to point-based logistic regression, the accuracy of game-based logistic regression has a clear value-added, from 65.5 to 71.4. At this time, the results of the model can better reflect the real fluctuation state of the players. Then, the results of the logistic regression test are as shown in the figure



Variable in an equation

		B	Standard error	Wald	Degree of freedom	Significance	Exp(B)
Step 1	s1	-0.229	0.244	0.882	1	0.348	0.795
	s2	0.955	0.809	1.395	1	0.238	2.599
	s3	-7.161	3.538	4.098	1	0.043	0.001
	s4	1.422	0.603	5.567	1	0.018	4.145
	s5	0.102	0.280	0.133	1	0.715	1.108
	s6	0.219	0.193	1.279	1	0.258	1.244
	s7	0.455	0.169	7.201	1	0.007	1.576
	s9	0.064	0.146	0.195	1	0.659	1.066
	s10	0.292	0.170	2.942	1	0.086	1.339
	s11	2.155	0.493	19.129	1	0	8.628
	s12	0.774	0.412	3.532	1	0.06	2.168
	s13	-2.533	0.991	6.527	1	0.011	0.079
	s14	0.411	1.374	0.090	1	0.765	1.508
	s15	-1.455	1.326	1.204	1	0.272	0.233
	s16	8.719	4.497	3.758	1	0.053	6115.731
	constant	-0.925	0.580	2.546	1	0.111	0.397

Figure 8: Logistic regression results

Compared with the results of Question 1, in the existing results, many variables that were originally insignificant became significant, including x3, x4, x11 and x13

References