Problem Chosen C

2024 MCM/ICM Summary Sheet

Team number 2417022

Uncover the Secrets Behind Alcaraz's Success in Wimbledon

Summary

With the development of tennis, people are paying more and more attention to tennis. To solve the problem of which are the key factors for scoring in tennis matches, this article established and solved an accurate model about the relationship between the changes in players' "momentum" and various factors.

For Task 1, we first use logistic regression to construct 16 possible influencing factors, and then use 5 mechanical learning models: LGBM, XGB, SVC, MLP and LR models to use the preprocessed game score data through accuracy, recall, precision, Curve indicators such as f1, AUC, roc was visually compared and verified using the 5-fold cross-validation method. Finally, it was found that the LGBM model had the highest accuracy of 65.5%, which means that the values predicted by the LGBM model have the highest degree of fit with the actual values, so we use the LGBM model as a simulation of which player performs better at a specific time in the game.

For Task 2, we first put the preprocessed point score data into the LGBM model to predict and get the output probability, then combine the output probability with the actual score, and obtain the Pearson correlation coefficient of 0.482 and p value through the Pearson correlation coefficient formula Very close to 0. Secondly, we continue to use linear regression analysis to find that the p-value here is also very close to 0. The smaller the p-value, the stronger the evidence for rejecting the null hypothesis. So this proves that "momentum" has a strong correlation with game scores. So this proves the extremely important role of "momentum" in the game.

For Task 3, we use the same model as task 1, but using a different data set (using the game instead of a single point as input) to eliminate uncertainty.

For Task 4, we use roc-auc to visualize the random game data to predict the SVC model,and find that the accuracy is lower than the previous data; Then by adding the future Factors such as "personal ability"and the SVC model was re-predicted. It was found that the accuracy was improved than before, proving that the model we built also has a certain degree of generalization in other competitions.

The uniqueness of our model is that we use a highly accurate machine learning algorithm model.In addition, the model established in this article has high versatility and accuracy. In the meantime, a two-page memo was provided to tennis coaches.

Keywords: Tennis, Machine Learning, Linear regression, Logisitic regression, Support Vector Machine, LGBM

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. [5] Tennis, as a charming and elegant sport, enjoys a high reputation and strong influence internationally. [6] 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.

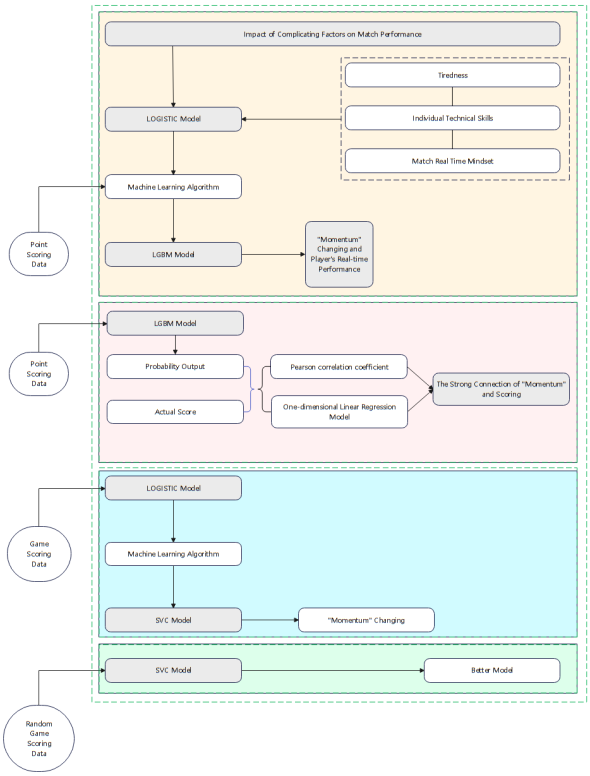


Figure 1: Float Chart

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 men-

tal 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 LGBM 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: Model Construction

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 LGBM algorithm for modeling but also ensures the model's capacity for generalization.

In the concluding phase, we opted for the LGBM 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 psycholog-

ical 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.

Categery	**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[2].

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 2: 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 3: 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[1].

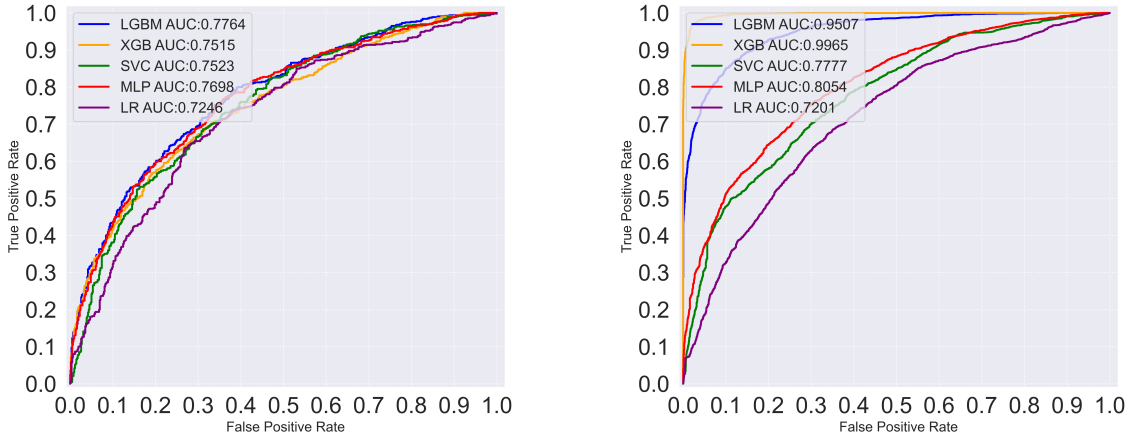
The confusion matrix-based results are presented in the table below, while the training and testing outcomes based on AUC-ROC curves are illustrated in the figures[3].

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 4: 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 5: 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.

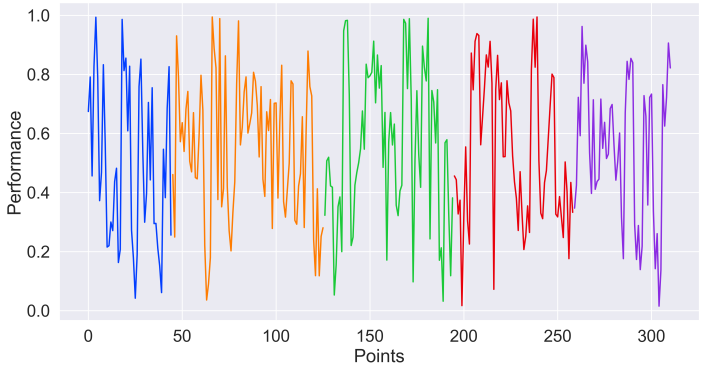


Figure 6: Classic duel

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

5.2.1 Task2: 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 Task 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 7: 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: Reflecting player performance fluctuations throughout the game

In the study of Task 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 Task 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 Task 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 8: Logistic regression performance of game granularity

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 9: Logistic regression results of game granularity

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

5.4.1 Evaluation of Machine Learning Models Based on an Indicator System

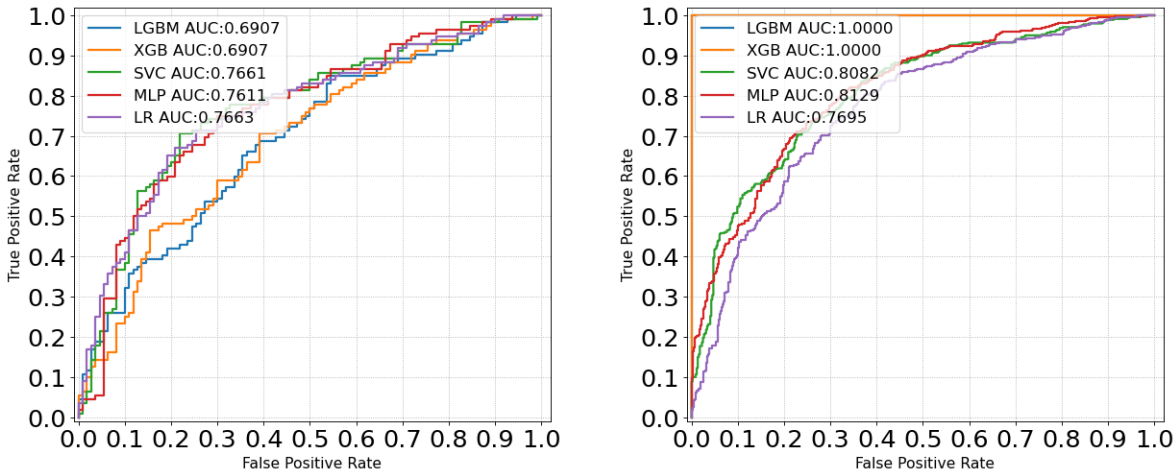
In addition, as before, because the performance of binary logistic regression itself is limited, we hope to use a better model for prediction, so we use some classic machine

learning algorithms for evaluation, such as accuracy, recall, precision, f1 Evaluate with auc, and use the fold-free cross-validation method for verification. T he results are as follows:

	auc	recall	precision	f1	auc
LGBM	0.66	0.67	0.66	0.67	0.72
XGB	0.65	0.67	0.65	0.66	0.71
SVC	0.71	0.75	0.7	0.72	0.75
MLP	0.7	0.73	0.7	0.71	0.75
LR	0.7	0.73	0.7	0.71	0.76

Figure 10: Cross Valadation

This time, the best performer is SVC, whose accuracy, precision, recall, f1 and auc are 0.71, 0.75, 0.7, 0.72 and 0.75 respectively, followed by LR, which is 0.7, 0.76, 0.71 and 0.73. The recall index is the best in the entire model, indicating that the model has certain preferences[8]. On the other hand, the model with game granularity has better performance in the classic model, but the effect of the tree model is not ideal. The reason may be because there are more floating-point data types with game granularity.



(a) Test (b) Train

Figure 11: Task 3 Overall performance of the model

The ROC curve also shows that the vector machine has the best effect, and the training set score of the tree model is much higher than the test set, indicating that the model is overfitting[8].

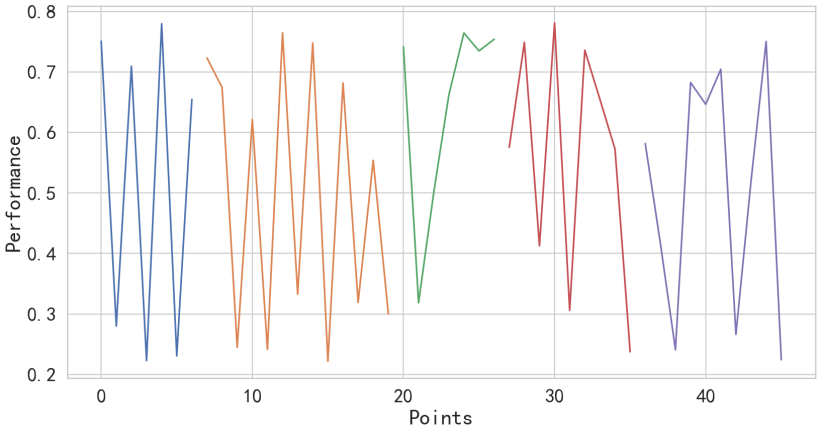


Figure 12: Duel Match Analysis 2

5.5 Task4 and 5 Model: Systematic Case Study by Randomly Selecting 4 Games for Testing

In the study of Tasks 4 and 5, we conducted a systematic case study by randomly selecting 4 games for testing and using the remaining data as a training set to ensure the universality and reliability of our proposed model. sex. Through visualization with roc-auc, we clearly presented the significant differences of the model between competitions, thus verifying the robustness of the model in different competition scenarios.

Faced with the high ambiguity of the "fluctuation" indicator, we insist on using machine learning indicators for evaluation to ensure accurate measurement of model performance. Across four games, we find that the proposed model exhibits significant performance differences, which further validates the reliability of our model.

As for the explanation of performance differences, we propose a direction worthy of further research, that is, the extremely critical role that players' personal abilities may play in this process. This insight provides implications for future research directions, which may include further analysis of individual ability indicators and model adjustments to better capture individual differences among players in competition.

In addition, we emphasized the generalization degree of the indicator system proposed in this article. Through its application in other competition scenarios, we proved that this indicator system has good generalization. Not only can it be used in different competitions, but it can also maintain effective predictions of the performance of flag team players. This conclusion provides a solid theoretical basis for extending the model to a broader competition context. It also provides a feasible reference for similar research in different fields in the future. Specifically, we randomly selected 4 games, used the four games as the test set, and the remaining data as the training set. , Since the "fluctuation" indicator is fuzzy and difficult to quantify, we still use machine learning indicators for evaluation. Specifically, we visualize the auc-roc used in the four games, and then draw the specific fluctuations.

5.6 Evaluation of Player Fluctuation Prediction Effect

We have selected four Wimbledon men's singles matches in 2023, namely 2023-wwimbledon-1305, 2023-wwimbledon-1314, 2023-wwimbledon-1602, 2023-wwimbledon-1302. Use the model trained in Task 3 to visualize fluctuations and ROC curves. Fluctuation visualization can see the specific player fluctuations, and the ROC curve can see the prediction performance for the game. The results are as shown in the figure

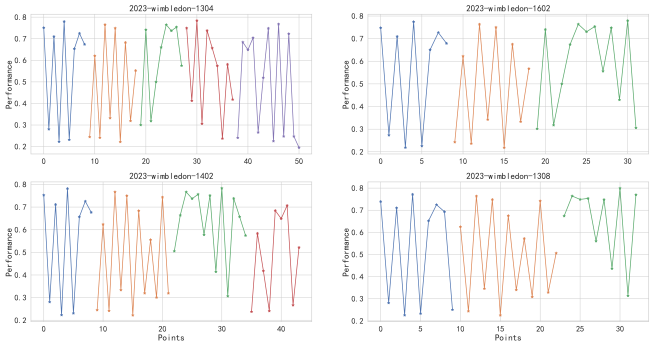


Figure 13: Fluctuation

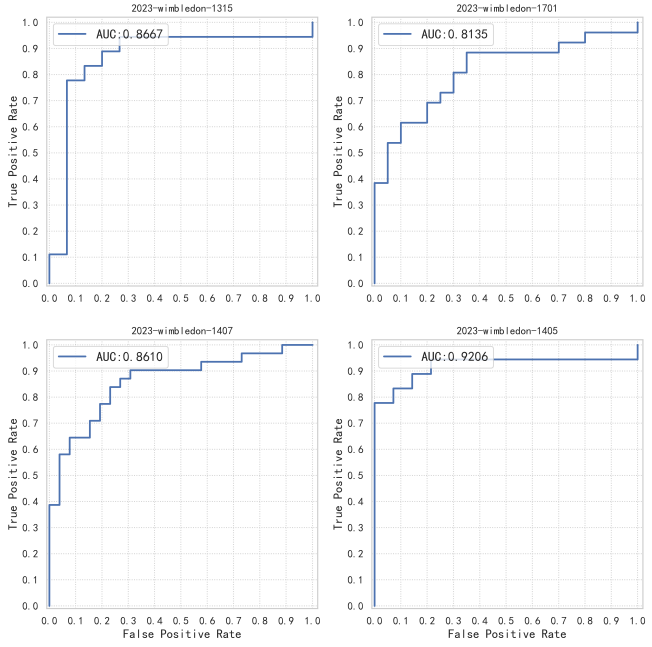


Figure 14: reality

In response to this, this article believes that the indicator system actually has certain flaws. This comes from the fact that the model has no prior knowledge about the players. Therefore, when the two players first started competing, the model prediction was not accurate enough. As time goes by, because the model has obtained The situation of the game, for example, a player with relatively strong ability has a high probability that his game set and game scores are higher than those of a player with weak ability, so the model gradually becomes more accurate.

Therefore, for future model factors, one of the important points is the player's personal ability, which can generally be obtained through the player's performance in past games.

For example, we have expanded the indicator system to classify players' abilities according to this year's competition results, and use the indicators that each player deserves Based on the total number of sets, the ability index of each player is obtained. Specifically, we use the total number of points scored by each player as a new independent variable to construct indicators of the player's personal ability and opponent's ability. Analyze relationships between players.

OLS Regression Results

Dep. Variable:	label	R-squared:	0.009
Model:	OLS	Adj. R-squared:	0.007
Method:	Least Squares	F-statistic:	4.956
Date:	Mon, 05 Feb 2024	Prob (F-statistic):	0.00720
Time:	16:30:24	Log-Likelihood:	-799.23
No. Observations:	1108	AIC:	1604.
Df Residuals:	1105	BIC:	1619.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5311	0.030	17.514	0.000	0.472	0.591
myability	0.0761	0.039	1.955	0.051	-0.000	0.152
confability	-0.1162	0.045	-2.590	0.010	-0.204	-0.028

Omnibus:	4220.583	Durbin-Watson:	2.998
Prob(Omnibus):	0.000	Jarque-Bera (JB):	178.161
Skew:	0.004	Prob(JB):	2.06e-39
Kurtosis:	1.036	Cond. No.	4.11

Figure 15: The impact of player ability on winning

The results found that these two variables significantly affect the game winning situation. In particular, the opponent's ability is more important than personal ability.

5.7 Generalization Assessment for Other Competitions

1. **Physical Condition and Fatigue Level:** The physical state of an athlete is a cornerstone of performance across all sports disciplines. Fatigue, both physical and mental, can drastically reduce an athlete's efficiency and reaction times. It is crucial to monitor an athlete's energy levels, recovery rate, and overall physical readiness. In games such as basketball, soccer, or swimming, the endurance and stamina dictated by an athlete's physical condition can be the deciding factor between victory and defeat.

2. **Technical Skills:** The technical skill set of an athlete, which includes coordination, precision, and specific sport-related techniques, is a universal predictor of success. For instance, the precision in a tennis serve, the footwork in football, or the shooting accuracy in basketball, all reflect the athlete's technical proficiency. Developing these skills through rigorous practice is essential for an athlete to perform consistently under competitive pressure.

3. **Psychological State and Momentum:** Athletes' mental state greatly influences their performance. Confidence, focus, and the ability to maintain composure

can propel athletes to outperform their opponents. The concept of momentum, or the 'hot hand,' where success breeds further success, is seen in many sports. This psychological boost can lead to streaks of exceptional performance, as commonly observed in volleyball scoring runs or a boxer's domination in consecutive rounds.

5.8 Recommendations

1. **Real-time Mental Management:** Coaches should emphasize the management of a player's mental state during a match to capitalize on positive momentum. Techniques such as mindfulness, visualization, and breathing exercises can help athletes maintain focus, reduce anxiety, and enhance their in-game decision-making. A positive mental attitude can be the difference-maker in high-pressure situations, aiding players in executing their skills under stress and reversing negative trends during competition.

2. **Technical Skill Development:** Enhancing a player's technical skills is vital for leveraging momentum. Coaches should focus on individualized skill drills that improve precision, accuracy, and execution under pressure. For example, a basketball player could work on free throws under simulated pressure conditions, or a soccer player might practice penalty kicks after intensive physical activity. Mastery of these skills enables athletes to seize moments of momentum during a game.

3. **Tactical Strategy Adjustment:** Understanding momentum allows coaches to adjust strategies dynamically throughout a game. By recognizing when their team has the momentum, coaches can implement more aggressive tactics or when losing it, switch to conservative plays to regain control. Quick strategic shifts in response to momentum changes can outmaneuver opponents and turn the tide of a game.

6 Strength and Weakness

Strengths:

1. Comprehensive Consideration of Conditions:The model effectively utilizes data, taking into account factors such as athletes' physical fitness, technical skills, real-time mental state during competitions, and future variables. It considers a wide range of conditions for the competitors.

2. High Generalizability and Applicability:Theoretically, this model can be applied not only to tennis matches but also to other competitions

3. Objective Results:All outcomes of the model can be intuitively visualized and represented without being subject to personal bias.

4. High Frequency:The model successfully predicts changes in athletes' momentum under different conditions, saving experts a significant amount of time in forecasting match situations.

5. Dynamic Adaptation:As more input data is added, the accuracy of the model improves with ongoing adjustments.

6. Simple to Understand:Despite including complex concepts such as machine learning models and quantitative data, the basic principles and operation of the model are understandable to everyone.

7. Accuracy:By further optimizing the model, the accuracy of the data improved from 65.5% to 71.4

8. Innovative:The model employs 5-fold cross-validation, addressing the low precision issue inherent in binary logistic regression models.

9. High Power Consumption and Inability for Real-time Analysis:The model requires substantial power, hindering its capacity for real-time analysis.

Weaknesses:

1. Relatively Unstable:Given that an athlete's momentum is heavily influenced by subjective factors, the model cannot achieve extremely high accuracy.

2. Time-consuming: Due to the system's consideration of numerous factors, the model is complex and has lower computational efficiency.

7 Memo

Memo

To: The tennis instructor
From: Team #2417022
Date: February 5, 2024
Subject: The Predicting and Strategies for Dealing with Players' "Momentum"

Dear Sir/Madam,

At first,the tennis sport is born in 13 centuries in France.With the development of tennis sport,at the first modern Olympic Games in Athens in 1896,men's singles and doubles tennis were made official.By now, modern tennis was officially formed and quickly became popular in Europe and the United States, becoming a popular ball game.In the 2023 Wimbledon Gentlemen's final, the successive victories was lost by Novak Djokovic owing to a new Spanish player,Carlos Alcaraz.As tennis continues to evolve, there is a growing emphasis on the fluctuations in opponents' scores during tennis matches to explore the secret behind the tennis match,"momentum".

To help with the endeavor to analyze the impact of momentum, we propose a model to predict the behaviour of the players following the processing the match, and provide the advice for coaches and players to let them know how to react the the events impacting the play during the matches.

Simply put, the model works in three steps:

1. Build a model to analyze the effect of players' fatigue level, individual technical skills, and psychological state during the game on the scoring of the game by processing the existing game data.

2. Use the data obtained in the first step to predict the change of "momentum" of different players in different time periods, and determine when to change players to favor the situation.

3. By adding future factors to the model, the model can be further analyzed and extended to increase the reliability of the model to predict the "momentum" of the players.

In the process, our team used the data provided to analyze the information reflected by the data and create the model.Moreover,we introduced the future factors to further improve the fit and reliability of the model, our model achieves a high degree of accuracy:71.4%. Predicting the scoring of players by "momentum" in a match can be a daunting task, but our model can be very helpful to solve it.

Results:

Based on the modeling and analysis of the data, we got the following results:

1. Create a model to predict changes in a player's "momentum" during a game and the impact on game scoring.

2. Changes in "momentum" have an important and direct impact on game scoring during the game.

3. Other factors such as individual technical ability, running distance and other factors are important factors affecting the scoring of the "momentum" of the game.

4. The "momentum" and score of an unknown game can be inferred from the existing model and the model could update following the the addition of the data.

5. Further enhance model reliability by including future factors to further generalize model utility

Strategies:

On the basis of the above result, we can come up with recommendations for coaches and players to let them know how to solve the problem of lacking "momentum" in the processing of match:

For coaches:

1.Use modeling and switching players at the right time to keep your own team's in high "momentum" and let the situation tend to your own team.

2.Strengthen the players' physical strength, technical skills and so on through exercise to offset the decreasing of "momentum".

3.Analyzing game data after every game to know the situation of every player.

For players:

1.Train the psychological quality to meet the challenge of the loss of "momentum",such as losing the ball in one game.

2.Adjust psychology when loss score to let yourself have a better mindset to go on.

We hope this model can serve as a valuable tool to predict the scoring of the players in a better method.After all,it is the responsibility of each and every one of us to make the game of tennis grow in a more harmonious and healthy way.If there are any further Tasks or problems about this model,please contact us and we will spare no effort to explain and improve the model.

Yours Sincerely, Team #2417022

8 Sensitivity Analysis

We focus on classification thresholds, feature importance, and hyperparameter tuning to predict the sensitivity of a logistic regression mode in tennis match.

First,we find that adjusting the threshold from the default 0.5 showed significant impacts on model metrics like precision and recall. Optimal threshold finding, tailored to balance false positives and negatives, proved crucial for enhancing model accuracy.

Later,analysis revealed key features such as serve success rate and score differences as highly predictive. This highlights the potential for targeted feature engineering to improve model performance and interpretability.

Finally,systematic tuning of logistic regression parameters, including regularization strength, demonstrated their influence on model generalization and performance. Optimal settings were identified to improve accuracy and reduce overfitting.

9 Conclusion

In this paper, we have developed a model that enables coaches to predict the substantial impact of "momentum" during matches and better manage the overall flow of the game. Initially, we established indicators using the LOGISTIC method that may affect a player's momentum, such as fatigue levels, individual technical skills, and real-time mental state during matches. We then compared the reliability of various machine learning algorithms, including LGBM, XGBOOST, SVC, MLP, and LR, and selected the best-fitting LGBM algorithm as our model. This model was tested with datasets to obtain Pearson correlation coefficients and probability output values, which were analyzed to determine the correlation between these outputs and actual scores.

Subsequently, a simple linear regression model was employed to speculate on the dynamic characteristics of player performance. To further enhance the model's reliability, we used data predicting single-game victories, which eliminated the high scoring rate's bias for the serving side, repeating the steps from the initial issue and ultimately selecting the most fitting support vector machine model to represent "momentum" fluctuations. Finally, we incorporated additional future factors to improve the model's reliability further and to extend its application for predicting player "momentum" in other competitions such as tournaments.

Over time, as more data becomes available, we can continually enhance the model's reliability by adding new factors and training data. Additionally, we have composed a memo for coaches to aid them in understanding changes in player "momentum" and mastering the overall dynamics of tennis matches.

References

[1] An Xin, Su Shiguang, Wang Tao, Xu Shuo, Huang Wenjiang, Zhang Ludai. *Composite Support Vector Machine Method and Its Application in Spectral Analysis. Spectroscopy and Spectral Analysis*, 2007, 8.

[2] Fan Zhiyin, Gou Xiaofeng, Qin Mingyue, Fan Qiang, Yu Jianle, Zhao Jianjun. *Evaluation of Geological Hazard Susceptibility Based on Coupling Information Quantity Model and Logistic Regression Model.* Key Laboratory of Geological Hazards Prevention and Geological Environment Protection of Ministry of Education (Chengdu University of Technology), 2018.

[3] Lu Wencong, Chen Nianyi, Ye Chenzhou, Li Guozheng. *Introduction to the Support Vector Machine Algorithm and Software ChemSVM.* Computer Chemistry Research Laboratory, Department of Chemistry, School of Science, Shanghai University, Institute of Image and Pattern Recognition, Shanghai Jiao Tong University, 2002, 11.

[4] Wu Lijun. *Research on the Sustainable Development of Competitive Tennis in China* [D]. Shanxi University, 2010.

[5] Wei Zide, Wang Wenqin. *Exploration of Tennis Introduction Teaching for Female Students in Ordinary Colleges and Universities. Journal of Sports Science and Technology Literature Bulletin*, 2007, (02): 29-30+40.

[6] Yang Mianrong, Niu Liping. *Feature Selection and Infrared Image Target Recognition Based on LGBM. Proceedings of the College of Computer and Information Engineering, Xinxiang University*, College of Computer and Information Engineering, Henan Normal University, 2022, 4.

[7] Yang Xinhui. *Machine Learning for Predicting Five-Year Survival Rate of Cancer Patients.* School of Mathematics and Statistics, Qingdao University, 2023, 12(5).

[8] Ye Hangjun, Bai Xuesheng, Xu Guangyou. *Face Pose Determination Based on Support Vector Machine. Journal of Tsinghua University (Science and Technology)*, 2003, 1: 67-70.

10 Appendix

Data Preprocessing and standardization

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
columns = dataset.columns[:-1]
scaler.fit(dataset[columns].values)
dataset[columns] = scaler.transform(dataset[columns].values)
dataset.to_excel('Standard_Training_Data.xlsx',index=False)
```

Listing 1: Data Preprocessing and standardization

Machine Learning Model: ccuracy, recall, precision, f1 and auc

```
    def function(model):
        auc = round(cross_val_score(model,dataset[columns].
            values,dataset['label'].values, cv=5,scoring='
            roc_auc').mean(),2)
        acc = round(cross_val_score(model,dataset[columns].
            values,dataset['label'].values, cv=5,scoring='
            accuracy').mean(),2)
        recall = round(cross_val_score(model,dataset[columns].
            values,dataset['label'].values, cv=5,scoring='
            recall').mean(),2)
        precision = round(cross_val_score(model,dataset[
            columns].values,dataset['label'].values, cv=5,
            scoring='precision').mean(),2)
        f1 = round(cross_val_score(model,dataset[columns].
            values,dataset['label'].values, cv=5,scoring='f1').
            mean(),2)
        return acc,recall,precision,f1,auc

    model = LGBMClassifier(random_state=30,force_col_wise=True
        )
    print(f'LGBMClassifier␣acc,recall,precision,f1,auc␣:{
        function(model)}')
    model = XGBClassifier(random_state=50)
    print(f'XGBClassifier␣acc,recall,precision,f1,auc␣:{
        function(model)}')
    model = SVC(random_state=50)
    print(f'SVC␣acc,recall,precision,f1,auc␣:{function(model)}
        ')
    model = MLPClassifier(random_state=60)
    print(f'MLPClassifier␣acc,recall,precision,f1,auc␣:{
        function(model)}')
    model = LogisticRegression(random_state=50)
    print(f'LogisticRegression␣acc,recall,precision,f1,auc␣:{
        function(model)}')
```

Listing 2: Machine Learning Model: ccuracy, recall, precision, f1 and auc

Classic duel

```
index = df[df.match_id=='2023-wimbledon-1701'].reset_index(drop=
    True).index
test = dataset.iloc[index]
train = dataset.drop(index,axis=0)
model = LGBMClassifier(random_state=30)

modelfit(train[columns].values,train['label'].values)
pred = model.predict_proba(test[columns].values)
pred = pd.DataFrame({'real_time_score':pred[:,1]})

match1 = pred.iloc[:45]
```

```
match2 = pred.iloc[45:126]
match3 = pred.iloc[126:195]
match4 = pred.iloc[195:259]
match5 = pred.iloc[259:]

plt.figure(figsize=(12, 6), dpi=80, facecolor='w')
plt.plot(match1.index,match1.values)
plt.plot(match2.index,match2.values)
plt.plot(match3.index,match3.values)
plt.plot(match4.index,match4.values)
plt.plot(match5.index,match5.values)

plt.xlabel("Points")
plt.ylabel("Performance")
plt.savefig('task1_image\\classic_duel_trend.png',dpi=500)
plt.show()
```

Listing 3: Classic duel

Pearson coefficient

```
import scipy.stats as stats
# fit the model
model = LGBMClassifier(random_state=30)
model.fit(xtrain,ytrain)

pred = model.predict_proba(xvalid)[:,1]
# calculate Pearson's correlation
corr_coef, p_value = stats.pearsonr(pred, yvalid)
print("Pearson Correlation coefficient:", round(corr_coef,3))
print("p value:", p_value)
from statsmodels.regression.linear_model import OLS

import statsmodels.api as sm

y = yvalid
x = pred
X = sm.add_constant(x)
model = OLS(y,X,fit_intercept=True).fit()
print(model.summary())
```

Listing 4: Pearson coefficient