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

The Driving Force Behind Tennis Matches: Momentum

Summary

In competitive tennis, the notion of **momentum** is both pivotal and contentious, sparking debate among professionals and enthusiasts. Our study, leveraging data from the 2023 Wimbledon men's matches, aims to quantitatively analyze and model the concept of momentum to understand its impact on match outcomes. Through a detailed examination of game dynamics and player performance, we seek to offer insights into the patterns that signify shifts in momentum.

For Task 1, following data processing, we first identified features highly correlated with the outcome of each game using the Spearman correlation coefficient and Random Forest from the dataset. Subsequently, we propose the GBM-Logistic Regression Based on Voting Classifier model (**GLBVC**) for predicting match win probabilities. Thereafter, we determined the weight of each model within the voting classifier using a weight optimization method based on iterative greedy search, finalizing the weights as **[0, 0.515, 0.485]**. Ultimately, we utilized our proposed GLBVC model to predict the win probability trends of the two players in the fifth set of the final match, based on data changes during the match, serving as a criterion for evaluating player performance.

For Task 2, we initially quantified the momentum using the win probability and its rate of change as assessed by the proposed GBM-Logistic Regression model. We then attempted to interpret the performance of Alcaraz and Djokovic using the defined momentum. Following this, we conducted a randomness test using the Wald-Wolfowitz Runs Test, obtaining a Z-score of **4.71** and significant (biased) autocorrelation coefficients, thereby demonstrating that fluctuations in performance during the match were not random. Finally, through time series analysis under **ACF** measurement, we found significant autocorrelation even at lags = 10, leading to the conclusion that momentum has a certain "inertia," further refuting the hypothesis of randomness.

For Task 3, we employed Deep Feature Synthesis(**DFS**) for automatic feature engineering, combined with manual feature engineering, to generate effective features that are both reflective of the data characteristics and incorporate expert experience. We then used the **CatBoost** algorithm, based on GBDT, to train our prediction model, achieving an R^2 of **0.72110**. We specifically analyzed a match and offered recommendations for coaches, including reducing unforced errors, ensuring stable wins on serve, and encouraging more aggressive movement.

For Task 4, the predictive model tested on datasets from Wimbledon women's singles, European Open men's singles, and US Open men's singles showed varied prediction accuracies, highlighting the influence of gender differences, match formats, and playing surfaces on model performance. The analysis underscores the importance of tailoring the predictive model to account for these variations, including conducting performance evaluations, feature importance analysis, and integrating diverse data sources to improve accuracy. In addition, we explored how our model could be extended based on our model to better accommodate women's games and to accommodate different game formats and field types.

Keywords: Momentum; GLBVC; Iterative greedy search; CatBoost; DFS;

Contents

1 Introduction **3**
1.1 Background . . . 3
1.2 Restatement of the Problem . . . 3
1.3 Our Work . . . 4

2 Model Preparation **4**
2.1 Assumptions and Notations . . . 4
2.2 Data Preprocessing . . . 5

3 Task 1: GLBVC Model for Assessing Player's Performance **6**
3.1 Feature Selection . . . 6
3.2 Model Building . . . 8
3.3 Determining Weights for the Voting Classifier . . . 9
3.3.1 Weight Optimization Based on Iterative Greedy Search . . . 9
3.3.2 Result Analysis . . . 10
3.3.3 Match Flow Display and Player Assessment . . . 10

4 Task 2: Momentum Definition & Randomness Testing **11**
4.1 Definition of Momentum . . . 12
4.2 Wald-Wolfowitz Runs Test . . . 14
4.3 Time Series Correlation Analysis (ACF+PACF) . . . 16

5 Task 3: CatBoost-based Model for Swing Prediction **17**
5.1 Deep Feature Synthesis . . . 17
5.2 Introduction of CatBoost . . . 19
5.3 Result Analysis . . . 20

6 Task 4: Model Generalization & Sensitivity Analysis **22**
6.1 Model Testing . . . 22
6.2 Extension of Our Model . . . 23

1 Introduction

1.1 Background

The 2023 Wimbledon Gentlemen's final has concluded. Carlos Alcaraz, a 20-year-old Spanish prodigy, emerged victorious over the experienced 36-year-old champion, Novak Djokovic, putting an end to Djokovic's decade-long reign at Wimbledon. The match was a thrilling rollercoaster, with Djokovic initially dominating Alcaraz with a 6-1 victory in the first set. However, the match took several dramatic turns as Alcaraz fought back and won the second set in a tiebreaker. He continued to dominate in the third set, but Djokovic made a comeback in the fourth set. Ultimately, Alcaraz secured his triumph with a 6-4 victory in the final set.

This remarkable final emphasized the intriguing challenge of quantifying momentum—a term frequently mentioned but seldom measured. In sports, momentum is generally described as a surge in a player's strength or force, leading to a series of successful plays. However, capturing this concept in a data-driven model is complex, especially in a sport like tennis where psychological and tactical factors intertwine with physical play.

The dataset from Wimbledon 2023 presents an opportunity to develop a model that captures the ebb and flow of gameplay and evaluates the impact of momentum. This model aims to predict when and why shifts in gameplay occur, offering valuable insights for players and coaches. Through meticulous analysis, this paper will delve into the predictive power of such a model, its applicability to other matches, and its generalizability across different players, surfaces, and sports. The findings aim to shed light on the enigmatic nature of momentum, equipping coaches with strategies to prepare for and respond to dynamic changes during a match.

1.2 Restatement of the Problem

- ***Develop a Model:*** The model must quantify the flow of play and the relative performance of players throughout the match. It should account for the fact that serving players typically have a higher probability of winning points, which should be integrated into the analysis logically.

- ***Momentum vs Randomness:*** Address the skepticism around momentum by analyzing whether shifts in play and successful runs by players are purely random or if they exhibit patterns that can be attributed to momentum.

- ***Predicting Swings:*** Examine the existence of indicators that can help determine when the game situation shifts from one player to the other. The models created help coaches determine when the game situation is likely to change and strategize accordingly.

- ***Generalization:*** Test the effectiveness of the developed model in other competitions. We can apply the models we develop to other matches, including different matches, different court surfaces, and other sports (e.g., women's matches, and other ball sports) to assess the applicability and accuracy of the models. By comparing the results with those of actual matches, we can assess the predictive power of the model and identify factors that may need to be improved or adjusted.

1.3 Our Work

The work we have done in this problem is mainly shown in the following Figure 1.

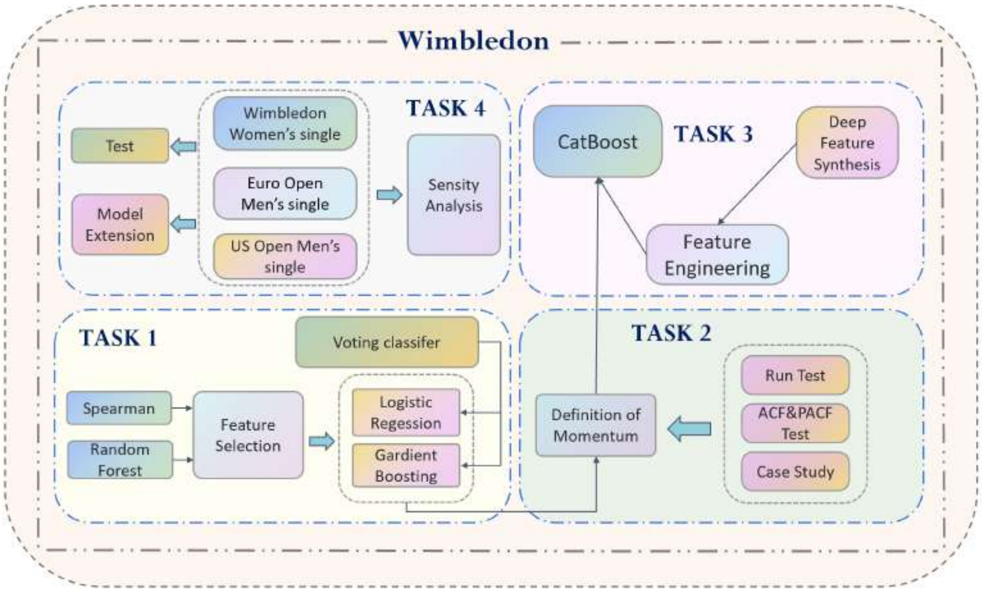

Figure 1: Our Work

2 Model Preparation

2.1 Assumptions and Notations

To simplify the given problems, we make the following basic assumptions, all of which are properly justified.

- ***Player Performance Homogeneity:*** Assume that each player's performance level can be quantified in a homogeneous manner across different matches, despite variations in opponents, playing conditions, and psychological factors.
- ***Momentum Persistence:*** Momentum, once gained, has a persistent influence on the immediate subsequent games within the match unless interrupted by an external event (e.g., a game break or an exceptional play).
- ***Match Conditions Constancy:*** The playing conditions (e.g., weather, surface wear) remain constant throughout the match, not affecting the probability of winning differently over the course of the game.
- ***Randomness of Point Outcomes:*** We assume each point's outcome as a Bernoulli trial, impacted by underlying player performance levels and serving advantage.
- ***Independence of Matches:*** Each match is considered an independent event, not influenced by the outcomes or performances of previous matches.

- ***Data Completeness:*** The dataset provided contains accurate and complete records of all relevant match events.
- ***No Injuries or External Disruptions:*** Assume players do not experience any injuries or external disruptions (e.g., crowd interference) that could significantly impact match performance unaccounted for in the dataset.

The primary notations used in this paper are listed in Table 1.

Table 1: Notations

Symbols	Significance
CPW	Player's current number of consecutive points won
ADV	Player's current advantage: own current score − opponent's score
DRP	Player's distance ran during a point (meters)
λ	Whether the player serves this time
f_{ace}	The ratio of players hitting an untouchable winning serve in a game
f_{win}	The ratio of players hitting an untouchable winning shot in a game
η	Break of server
β	The number of double faults made by the player
γ	The number of unforced errors by the player

2.2 Data Preprocessing

The dataset, named Wimbledon featured matches.csv, encompasses comprehensive information regarding the Gentlemen's singles matches at Wimbledon 2023, specifically after the second round. Each data entry corresponds to a specific match and includes essential details such as match_id, player's name, elapsed time, and specific match statistics. Remarkably, there exists a discrepancy in the elapsed time values between rows 586 and 636, as they appear to be longer by a duration of 24 hours. To rectify this inconsistency, we subtract 24 hours from the affected data points, restoring their proper timings, as shown in Table 2.

Table 2: Corrected Elapsed Time

Row	**Original Elapsed Time**	**Corrected Elapsed Time**
586	24:56:34	00:56:34
587	24:57:00	00:57:00
...	...	...
636	25:37:54	01:37:54

However, it is worth mentioning that an anomaly is observed in the match linked to the match_id "2023-wimbledon-1403." In this particular match, only two sets of match data are available, prompting us to infer that some data points are missing. As a consequence, we have made the decision to exclude the data from this match in our subsequent analysis and training processes to ensure the accuracy and integrity of our findings.

After conducting a thorough statistical analysis and eliminating outliers, we obtained accurate statistics, as shown in Figure 2. The dataset comprises a total of 30 matches, encompassing 115 sets and 1273 games. On average, each match consists of approximately 42.43 games, which translates to an average of 238.77 points. Remarkably, each rally contributes approximately 0.42% of the total points tallied in a match. By leveraging the win-loss outcome of each rally as a performance indicator, we are able to account for the dynamic nature of match play, where the momentum can shift rapidly between players.

Figure 2: Accurate Statistics

Please note that the aforementioned statistics have been calculated based on the processed dataset, ensuring that any anomalies or irregularities have been effectively addressed.

3 Task 1: GLBVC Model for Assessing Player's Performance

In our assessment of athletes' performance fluctuations, we have chosen to use the outcome of each point as a reflection of the player's current performance level. This approach is based on the assumption that if Player 1 wins a point against Player 2, it indicates that Player 1 has outperformed Player 2 in that particular exchange. To some degree, these variations in performance can also depict the ebb and flow of the match. Thus, our use of the win/loss outcome of each rally as a performance metric is consistent with the dynamic nature of the game.

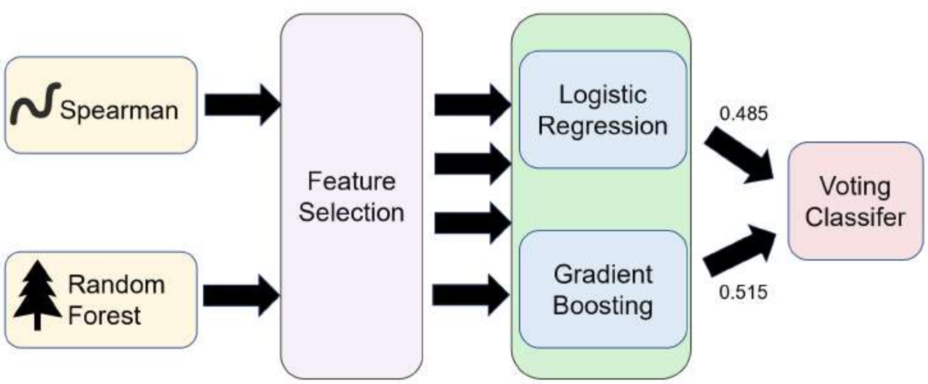

Figure 3: Work Flow of Task1

3.1 Feature Selection

In our endeavor to capture the game's direction when a player wins a point and to predict the potential existence of a "momentum" factor, we consider various elements that may influence a player's momentum. Utilizing Spearman's correlation coefficient analysis along with the Random Forest algorithm, we examined the correlation of these factors with the probability of a player winning, which is shown in Table 3.

Table 3: Correlation Coefficients

Indicators	CPW	λ	β	DRP	f_{ace}	λ	ADV	f_{win}	η
Correlation	0.778	0.347	0.086	0.025	0.341	0.132	0.192	0.403	0.303

Spearman's correlation coefficient analysis offers a thorough examination of the monotonic relationships between our proposed features and the winning outcomes, irrespective of the linearity of these relationships. This nonparametric measure is instrumental in determining how well the relationship between variables can be described using a monotonic function. In addition, we employ the complete set of features in our analysis with the Random Forest algorithm. When a feature was removed, if the predictive accuracy decreased, the feature was deemed significant and retained. If the predictive accuracy increased upon removal, the feature was considered irrelevant and thus discarded [3]. Whilst Spearman's analysis is particularly useful for assessing linear correlations, the Random Forest algorithm is advantageous for identifying features with non-linear correlations.

Some of the more influential factors we have identified include:

- ***Serve:*** The serving player or team dictates the game's tempo with their serve. They have the discretion to choose the speed, angle, and placement of the serve, exerting pressure on the opponent or altering the game's rhythm. This dominance in serving allows them to take the initiative and steer the direction of the match, thereby enhancing their probability of winning.
- ***Number of Points Won:*** Each point secured translates into an edge within the game. This advantage is not just reflected in the scoreboard but also imparts a positive psychological effect conducive to winning. Players or teams with a higher tally of points won are more likely to maintain control and secure victory.
- ***Distance Run:*** The total distance covered by a player is indicative of their activity level and engagement in the game. An increase in the distance run suggests vigorous pursuit of the ball, signaling a greater degree of competitiveness and dedication. Such fervor and commitment can positively influence game performance and, consequently, the likelihood of emerging victorious.
- ***Consecutive Points Won:*** Securing a series of points consecutively can create momentum within the game, psychologically bolstering the winning side. A run of multiple points can instill confidence and potentially demoralize the opposition. This psychological upper hand often translates into increased determination and a higher chance of winning for the team or player in the lead.
- ***Break of Server:*** Achieving a break of serve, where the receiver wins the game against the server, is a pivotal event in tennis. It is indicative of the receiver's capability to triumph in a

game where they do not have the initial advantage of serving. This not only impacts the scoreboard but also delivers a substantial psychological setback to the opponent. A break of serve can decisively shift the match's momentum, offering the receiver a significant psychological edge.

In our final model, we selected seven features as inputs: serve, the number of points won, the distance run, consecutive points won, break of server, the current set score, and hitting an untouchable winning shot. Table 4 displays the coefficients or importances of these features in logistic regression and random forest when the target is set to whether a game is won or lost:

Table 4: Coefficients or Importances of Features

Feature	Logistic Regression	Random Forest(percentage)
Serve	1.556	0.27
Points won	0.899	0.19
The distance run	0.028	0.2
Consecutive points won	0.17	-
Break of server	0.764	0.02
Untouchable winning shot	0.04	-
Current game score	0.058	0.1

As can be seen from the table above, the server plays a dominant role in both methodologies, leading us to infer that being the server has a significant impact on the victory of the game. The distance run holds considerable importance in the random forest but has a coefficient close to zero in logistic regression (where coefficients closer to zero indicate a smaller impact), suggesting a possible nonlinear relationship with game victories. In contrast, the break of server has a high coefficient in logistic regression but low importance in the random forest, indicating a higher linear correlation with game victories.

3.2 Model Building

Given the numerous uncertainties present on the court, such as weather conditions, audience presence, the psychological impact of breaking serve or winning streaks, and so forth, the resulting variations are often highly unpredictable and it is challenging to classify the system as purely linear or nonlinear. Consequently, we need to establish a model that encompasses both linear and nonlinear characteristics. Therefore, we have decided to employ a Voting Classifier (as shown in Figure 4) that combines the outcomes of multiple models to capture the relationship between linear and nonlinear elements. The Voting Classifier can perform ensemble learning, which improves overall prediction accuracy and robustness by aggregating different models' predictions. We anticipate that it will consolidate models that are adept at identifying linear and nonlinear relationships to achieve more accurate predictions and mitigate the risk of overfitting. Moreover, the Voting Classifier can be configured for soft voting in Python, allowing for different weights to be assigned to each model, providing us with greater flexibility.

Our rationale for the selection of base models is as follows:

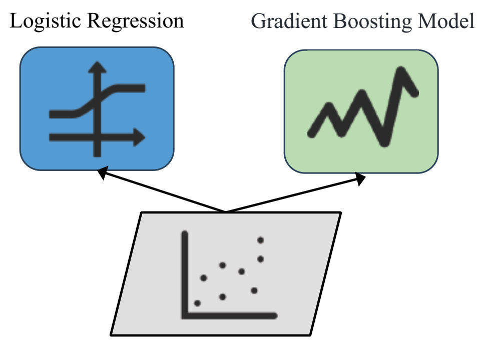

Figure 4: Voting Classifier

1. ***Logistic Regression:*** Logistic regression models are comparatively straightforward and quick for training and prediction. They perform well with linearly correlated data. Among the features we have extracted, server, points won, and break of serve can be considered linearly related to winning a game. Thus, selecting logistic regression as one of the base models within the voting classifier is a judicious choice.

2. ***Random Forest:*** The Random Forest model has high robustness, and it mitigates the risk of overfitting by constructing multiple decision trees, thereby enhancing the model's generalization capability. Additionally, Random Forest can assess the degree of impact of each feature on the prediction outcome, which aids us in identifying features with high importance. Compared to logistic regression, Random Forest has a superior performance in handling nonlinear relationships, making it suitable for modeling the relationship between the distance run and the success in winning a game.

3. ***Gradient Boosting Machine (GBM):*** The GBM can be seen as an improved version of the Random Forest, building models sequentially to correct errors from the preceding model, providing more precise predictions. Although its interpretability may not be as strong as logistic regression, GBM is more capable of handling non-linearly correlated data. Therefore, we have also included GBM in our voting classifier.

3.3 Determining Weights for the Voting Classifier

Based on the introduction above, the next important task is to determine the respective weights of the three base models within the voting classifier. Let's denote the weights for logistic regression, random forest, and gradient boosting machine as $[w_1, w_2, w_3]$. We divided all the cleaned and processed data into a training set and a test set with a ratio of 7:3 to train the above three models.

3.3.1 Weight Optimization Based on Iterative Greedy Search

Having trained the three models, we used an iterative greedy search to accurately determine the weights of each classifier within the voting classifier. The essence of this method is to refine and narrow the range of weights iteratively, which indirectly achieves a pruning effect, ensuring both speed and accuracy. We used cross-validation scores as the criterion for evaluating the weights' efficacy. Unlike traditional evaluation metrics such as accuracy, precision, and recall, cross-validation divides the dataset into several smaller subsets and uses parts of it as the test set and the rest as the training set, repeatedly, to assess model performance. By evaluating the model across multiple different training and testing sets, cross-validation reduces the chance of random errors in the evaluation results, providing better comprehensiveness, robustness, and unbiasedness.

Initially, we set the weight space to (0,1,11), meaning we evenly selected 11 numbers between 0 and 1. Using a greedy traversal approach, we preliminarily determined the best weights to be [0.5, 0, 0.5]. After this step, we concluded that the random forest model did not contribute to the final outcome, so we discarded it. Next, we only solved for $[w_1, w_3]$, updating the weight space to (0.45, 0.55, 11) and repeating the above operation. After continuous iterative greedy search, we finally found the appropriate weights.

3.3.2 Result Analysis

The optimal weights we obtained were [0.485, 0, 0.515], with a cross-validation score of 0.861, indicating high accuracy. However, we also noticed an interesting phenomenon: the weight for the random forest model was determined to be 0. We suspect this might be because GBM and random forest are both tree-based models and may have some redundancy between them, and they are both adept at handling nonlinear relationships, leading to a competitive relationship. In our current dataset, GBM was able to completely cover the information provided by the random forest, making the latter less important in the combination.

Moreover, we observed that GBM took a slight advantage in weight, suggesting that it provided more accurate predictions on our dataset and dealt well with the complex nonlinear relationships. We should not overlook the strong performance of the logistic regression model; its nearly equal weight with GBM indicates that a significant proportion of linear relationships also exist in the dataset, which the logistic regression model captured well.

3.3.3 Match Flow Display and Player Assessment

We ultimately decided to apply our trained model to analyze the fifth set between Carlos Alcaraz and Novak Djokovic, a set that would determine the champion of the Wimbledon Tennis Championships. The players' conditions and the atmosphere of the match had reached their peak, making this set destined to be repeatedly appreciated and watched by audiences. Therefore, we believe that analyzing this set is particularly meaningful.

Our model calculated the winning probabilities for each player at every scoring moment within the game. We used this as an indicator to evaluate the performance of the two players in the current game. Furthermore, we plotted the cumulative scores of both players in the current set as another reference, with the final results shown in Figure 5. The dashed lines in the figure indicate the end of each game. Novak Djokovic began the match serving, and from the figure, it is evident that serving significantly impacts winning probabilities. At the start of each game, our model gives a higher winning probability to the server, and in most cases, players managed to maintain a high probability and finish the game on top. Unless unexpected events occurred, Novak Djokovic was predicted to win the final victory. However, a twist happened in the third game when Carlos Alcaraz scored three points consecutively on his opponent's serve, successfully breaking serve, and then swiftly ended the match by scoring four points in the fourth game. At this moment, our model also gave the highest winning probability assessment for Carlos Alcaraz throughout the set. Despite some fluctuations in winning probabilities in the following games, both players steadily won their serving games. That break of serve laid the foundation for Carlos Alcaraz's winning probability.

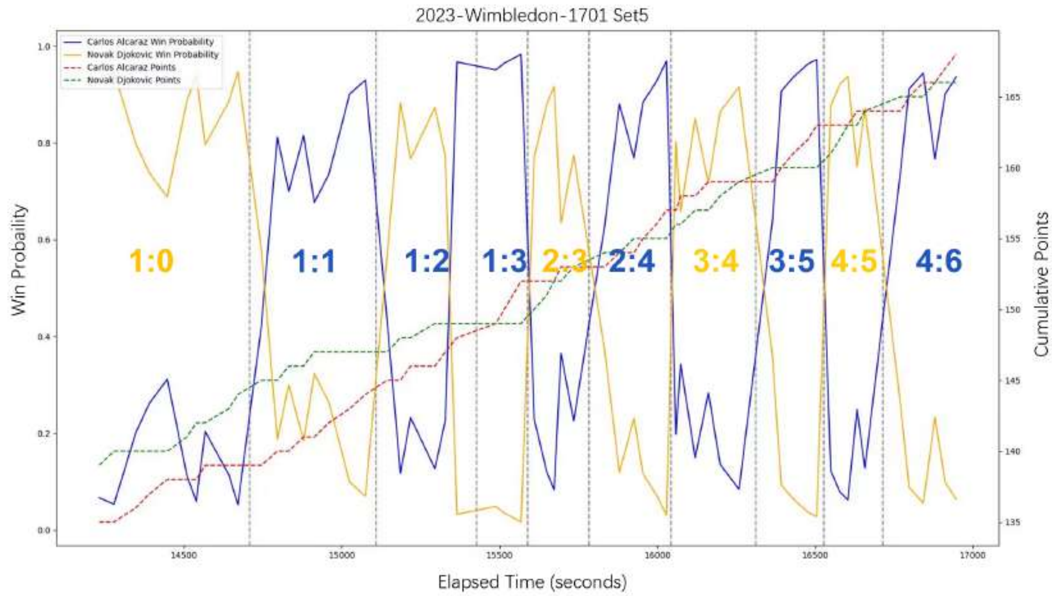

Figure 5: Winning Probability and Points of Two Players

The key factors influencing the players' performance in the match are shown in Figure 6. Throughout the entire set, despite the tense situation, Novak Djokovic only lost one serving game. However, after summarizing the performance of both players in the match, we found that concerning the key factors determined to influence the game, Novak Djokovic was almost entirely outperformed by Carlos Alcaraz. This explains why he lost the match and also validates the accuracy of our model's feature selection.

Figure 6: Key Factors Affecting the Game

4 Task 2: Momentum Definition & Randomness Testing

To investigate whether "momentum" plays a role in competition outcomes, we initially quantified momentum in Section 5.1 using players' current win rates and the change in win rates. Subsequently, we illustrated the momentum and win rate variations of Carlos Alcaraz and Novak Djokovic during the second set to visually demonstrate the impact of momentum on match dynamics. Finally, in Section 5.3, to determine whether match volatility and players' success were random or correlated with momentum, we employed the Wald-Wolfowitz runs test and the Autocorrelation Function Test (ACF) for analysis. Our findings conclusively demonstrated that momentum indeed influences match outcomes and is not random, thereby validating that performance fluctuations and consecutive successes of a player are not due to chance.

4.1 Definition of Momentum

In quantifying players' momentum, we opted to derive momentum directly from the win rate (P) evaluated by the model developed in our previous task. This approach is justified because the win rate P itself is derived from numerous features used during model development, such as consecutive scoring, breaks of serve, and the server, among others. We believe that assessing momentum from the following two aspects offers a more comprehensive insight:

- Player's Current Win Rate (P) The current win rate serves as a basic metric for assessing a player's present performance, acting as a static indicator of momentum. A higher win rate indicates excellent performance in the recent past, suggesting strong competitiveness and possibly high momentum. Maintaining a high win rate over time often signifies continuous scoring by the player, at which point we consider the player to be in a state of high momentum.
- Change in Win Rate over Time($\frac{dP}{dt}$) The rate of change in the current win rate over time reflects the trend of win rate variation, serving as a dynamic indicator of momentum. It highlights the trend in a player's performance over a period. Significant changes in win rates of players in a short duration suggest impactful events in the match (e.g., breaks of serve, service rotations). If a player's current win rate is increasing, it indicates improving performance and positive momentum. Conversely, a declining win rate may imply deteriorating form and negative momentum. Observing the rate of change in the current win rate can effectively determine a player's development trend and potential competitiveness.

Taking into account both the player's current win rate and the rate of change of the win rate over time allows for a more comprehensive assessment of a player's momentum. High win rates combined with a positive rate of change typically indicate strong momentum. Conversely, low win rates and a negative rate of change may suggest that a player is facing challenges and issues, indicating poor momentum. Therefore, after detailed statistical analysis of match data and multiple adjustments to the model training parameters, we define momentum as follows to achieve optimal evaluation effectiveness:

$$M = \alpha \cdot \ln(P + 1) + \beta \cdot \frac{dP}{dt} + \gamma \tag{1}$$

Here, M represents the player's momentum, P is the player's current win rate, and $\frac{dP}{dt}$ is the rate of change of the win rate. We also posit that as the win rate increases, the additional momentum gained from an increase in win rate diminishes. This implies that momentum is more sensitive to lower values of the win rate. Conversely, a decrease in win rate should impose a heavier penalty on momentum. The term γ represents the impact of some unknown factors.

Furthermore, to more accurately depict the rate of change of the win rate $\frac{dP}{dt}$ from discrete win rate values P and to further calculate momentum, it is necessary to interpolate the win rate P. In our model, we opted for cubic interpolation to obtain a smooth trend of win rate changes. This trend closely mirrors the continuous changes in players' performances in real scenarios, thereby providing a basis for tactical and technical adjustments. The parameters were ultimately set to $\alpha = 0.6$, $\beta = 15$ and $\gamma = 0.3$.

In Figure 7, we selected the second set from the match between Alcaraz and Djokovic to calculate the momentum values we defined. By marking the pivotal points of momentum on the graph, we identified their correlation with the players' conditions. The turns marked in red circles represent unforced errors made by Alcaraz, those in yellow circles indicate unforced errors made by Djokovic, and the blue circles denote an ace served by Alcaraz. We observed that Alcaraz is more prone to making errors after a series of wins and more likely to deliver strong plays like aces after a series of losses. Additionally, Alcaraz's own mistakes do not significantly negatively impact his momentum, whereas his opponent's mistakes have a positive effect on his momentum, suggesting that Alcaraz is confident in himself and believes in his ability to win the match. This set was extremely tense, and entering the tie-break, we could see that mistakes from both sides significantly impacted the momentum, which is related to the players' psychological pressure and the audience's reaction, greatly affecting the players' performance.

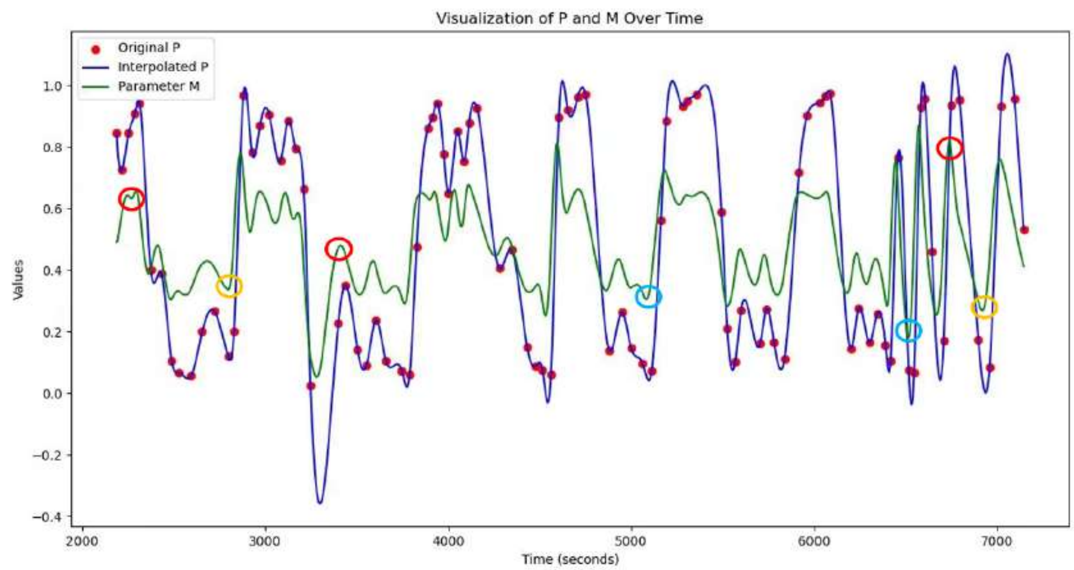

Figure 7: Carlos Alcaraz's Momentum Swings in Final Set 2

To evaluate whether the performance fluctuations and consecutive successes of a player in a tennis match are random, we used the Wald-Wolfowitz runs test and tools from time series analysis, ACF (Autocorrelation Function), and PACF (Partial Autocorrelation Function), for analysis and validation.

4.2 Wald-Wolfowitz Runs Test

The Wald-Wolfowitz runs test is a non-parametric statistical test method used to assess whether a sequence of observed data is random or exhibits some pattern. A "run" is defined as a sequence of adjacent repeated events. The essence of the test is to compare the observed number of runs with the expected number of runs under random conditions. If the observed number of runs is significantly less or more than expected under randomness, we can infer non-randomness in the data sequence. Therefore, in analyzing tennis matches, especially in exploring the impact of momentum on match dynamics, the runs test is very useful. Momentum in a match may lead to non-random sequences of winning points (runs). For example, a player may win several consecutive points after gaining an advantage. Through the runs test, we can determine whether sequences of victories (or points) in a match are merely random events or if factors like momentum significantly impact player performance.

We analyzed the trend of consecutive "momentum" values for Carlos Alcaraz over time using the runs test, as shown in Figure 8.

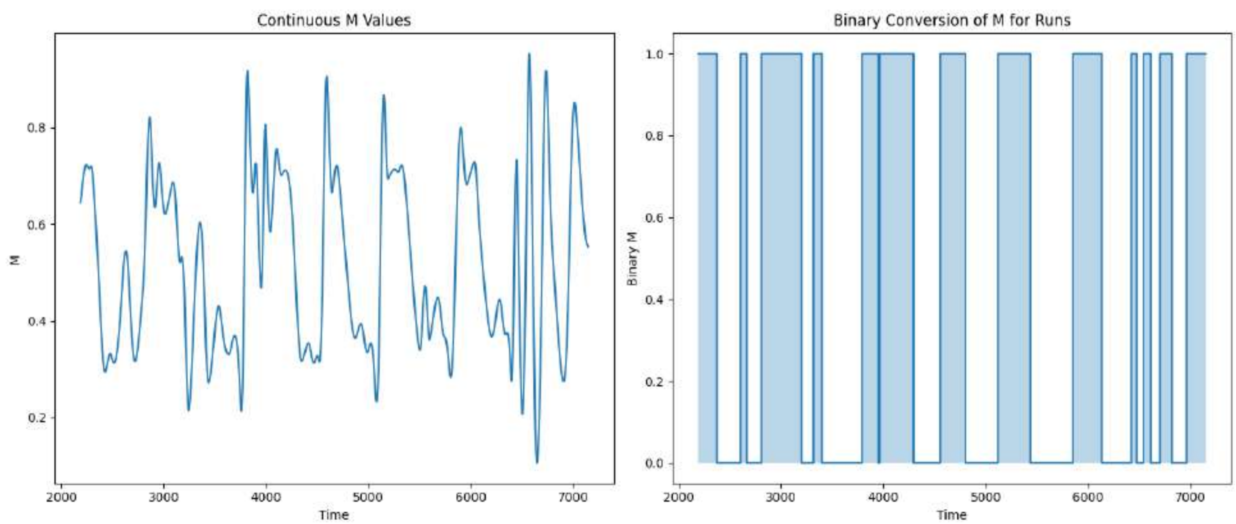

Figure 8: Change in Momentum and Binary Conversion of Momentum for Player 1

Figure 8 on the left displays the continuous M values (momentum) and their binary conversion over time. The graph indicates that M values fluctuate over time, allowing us to preliminarily infer that "momentum" may have influenced the match dynamics. Detailed statistical measurement data provided below further illustrate this pattern. The right graph in Figure 8 shows the binary conversion of M values relative to their median. Specifically, M values above the median are assigned 1 (True), and those below the median are assigned 0 (False). By shading the areas where M values exceed the median, we can visually understand the alternating pattern of runs.

The results of the runs test, indicating run lengths and run values, are presented in the Table 5.

The encoding results show four different runs alternating above and below the median. The variation in run lengths indicates the data's volatility and changing pattern. Longer runs signify that the trend of momentum remains consistent within these intervals, whereas shorter runs indicate

Table 5: Lengths and Values of Runs Test

No.	1	2	3	4	5	6	7	8	9	10	11	12	13
Run Lengths	37	47	12	29	79	23	17	79	32	3	66	53	49
Run Values	True	False	True	False	True	False	True	False	True	False	True	False	True
No.	14	15	16	17	18	19	20	21	22	23	24	25	-
Run Lengths	64	64	84	57	58	10	14	14	18	24	28	39	-
Run Values	False	True	False	True	False	True	False	True	False	True	False	True	-

rapid transitions in momentum trends. These lengths directly reflect the frequency and duration of M values alternating above and below the median.

In randomness testing, the number and distribution of run lengths are key indicators we use to assess the impact of ”momentum”. Theoretically, if the fluctuations in a match and a player’s success are random, the expected distribution of run lengths should be close to uniform or follow a specific statistical distribution. Additionally, other key statistical indicators such as the runs’ variance VAR, and the standard normal Z-score are needed to assess randomness. The specific analysis process is as follows:

- **Actual Number of Runs R:** The total number of runs R is 25, with the number of True runs $n_1 = 13$ and the number of False runs $n_2 = 12$.
- **Expected Number of Runs (ERL):** ERL refers to the expected length of consecutive identical values in a random sequence. It provides a benchmark for comparing the actual number of runs against the expected number of runs in a random sequence. By comparing the actual number of runs with the expected number of runs, one can determine if the sequence exhibits statistically significant non-random characteristics. The ERL is calculated using the formula:
$$ERL = \frac{2n_1 \cdot n_2}{n_1 + n_2} + 1 \tag{2}$$
- **Variance of the Number of Runs (VAR):** VAR measures the variability of run lengths in a random sequence. A larger VAR indicates higher variability of run lengths within the sequence, possibly suggesting the presence of non-random patterns or structures. Conversely, a smaller VAR suggests lower variability of run lengths, closer to the characteristics of a random sequence. The VAR is calculated using the formula:
$$VAR = \frac{2n_1 n_2 \left(2n_1 n_2 - n_1 - n_2\right)}{\left(n_1 + n_2\right)^2 \left(n_1 + n_2 - 1\right)} \tag{3}$$
- **Standard Normal Distribution Z-score:** The Z-score is used to measure the degree of deviation between the actual number of runs and the expected number of runs. By standardizing this deviation, comparisons across different samples become more feasible. A significant deviation of the actual number of runs from the expected number leads to a Z-score deviating

from zero, where a larger absolute value indicates a more significant difference. Calculating the Z-score and comparing it to critical values allows for a runs test, determining if the sequence exhibits statistically significant non-random characteristics. A Z-score exceeding critical values suggests the presence of non-random patterns or structures in the sequence. The Z-score is calculated using the formula:

$$Z = \frac{R - ERL}{\sqrt{VAR}} \tag{4}$$

- **p-value:** The p-value measures the significance of the runs statistic. It represents the probability of observing the given result, or more extreme, under the null hypothesis. Generally, if the p-value is less than a predetermined significance level (e.g., 0.05), the null hypothesis can be rejected, suggesting the sequence exhibits significant non-random characteristics. Note that the p-value does not directly indicate the strength of non-randomness or the size of the effect; it only provides probability information about the observed statistic under the null hypothesis.

In the analysis conducted with the Wald-Wolfowitz runs test, key statistical measures are presented as shown in Table Table 6. We observe that the Z-score is significantly greater than 1.96, the common critical value at the 95% confidence level, indicating a significant difference between the observed total number of runs and the expected number of runs for a random sequence. This suggests the sequence is not random. Therefore, based on the analysis, there is reason to suspect that the data sequence does not conform to the randomness assumption. In other words, "momentum" indeed plays a role in competitions, and the fluctuations and success of a participant are not random.

Table 6: Key Statistics in the Wald-Wolfowitz Runs Test

Statistic	R	n_1	n_2	ERL	VAR	$Z-score$	$p-value$
Value	25	13	12	13.48	5.97	4.71	≈ 0

4.3 Time Series Correlation Analysis (ACF+PACF)

The Autocorrelation Function (ACF) is employed to measure the degree of correlation between a time series and its past values, identifying randomness and periodic characteristics within the data. Specifically, if ACF values range between -1 and 1, values close to 1 indicate strong positive correlation, while values close to -1 denote strong negative correlation. The ACF test is particularly suitable for analyzing dependency within time series data, making it reasonable to use for assessing the impact of momentum in competitions. If a time series is completely random (i.e., white noise), its ACF plot will show significant autocorrelation only at zero lag ($lag0$), with autocorrelations near zero for other lags. Significant correlations at non-zero lags in the ACF plot would indicate non-randomness, such as consecutive wins driven by momentum.

We calculated the ACF and PACF for momentum (M values) using 20 lags, as shown in Figure Figure 9. It is important to note that, as mentioned earlier, due to the irregular and discrete nature of each time point in the given data, we interpolated both the win rate P and time T, meaning each lag here represents 5 seconds. In the ACF plot, we observe that the bar graphs of autocorrelation coefficients exceed the shadowed area of the blue confidence interval at multiple lag values, indicating significant autocorrelation at these lags. The autocorrelation coefficients gradually decrease with increasing lag values but do not drop to insignificant levels even at higher lags (e.g., 10), suggesting the series is non-random and has certain correlations with its past values. In the PACF plot, partial autocorrelation coefficients exceed the confidence interval before lag 5, indicating significant partial autocorrelation at these lags. Combining the ACF and PACF plots, we can preliminarily infer that momentum might play a role in competitions, possessing certain ”inertia,” and that the fluctuations and victories of a contestant are not random.

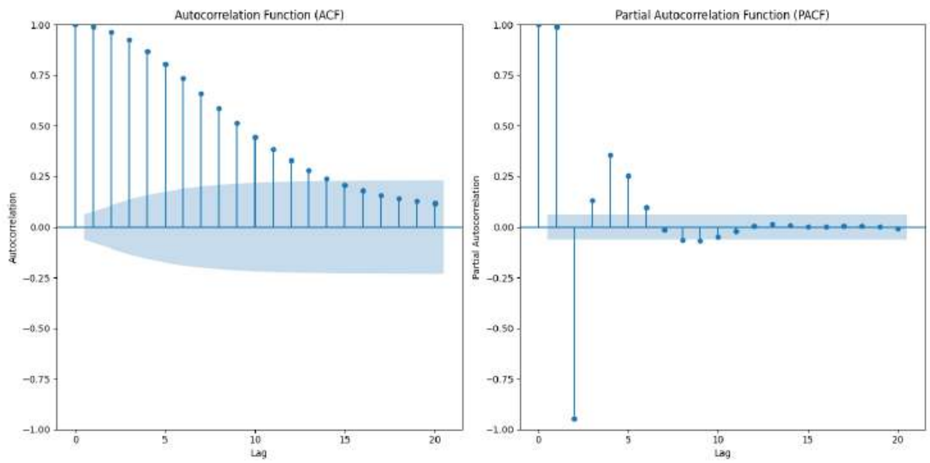

Figure 9: ACF and PACF of Momentum

5 Task 3: CatBoost-based Model for Swing Prediction

5.1 Deep Feature Synthesis

In competitive sports, predicting changes in the flow of a match holds significant importance for coaches and athletes alike. The ability to anticipate shifts in the game allows coaches to adjust tactics timely, providing athletes with a competitive edge. Previous research has demonstrated that the dynamics of sports competitions largely depend on complex interaction patterns, which are often difficult to capture with traditional data analysis methods. In recent years, with the advancement of machine learning technologies, particularly in automatic feature engineering, researchers have begun to explore more advanced methods for analyzing and predicting changes in the flow of sports competitions. Featuretools, as an advanced tool for feature synthesis, has shown great potential in such research. This study employs Deep Feature Synthesis (DFS) to explore and identify key factors affecting changes in the flow of the game [2].

Deep Feature Synthesis (DFS) is an automated method for constructing new features from a raw dataset. These new features are generated by leveraging various attributes of the existing data and derived statistics, aiming to uncover complex patterns in the data that may benefit model performance.

Forward relationships refer to scenarios in relational databases where a record in a parent table is associated with multiple records in a child table. In contrast, backward relationships refer to situations where multiple records in a child table are associated with a single record in a parent table. Entity features (EFEAT) are derived by calculating the value of each entry, with direct features (EFEAT) applied to forward relationships, and relational features (RFEAT) applied to backward relationships. Both RFEAT and DFEAT features can be synthesized independently, while EFEAT features depend on both RFEAT and DFEAT features. The relationship between them is shown in Figure 10.

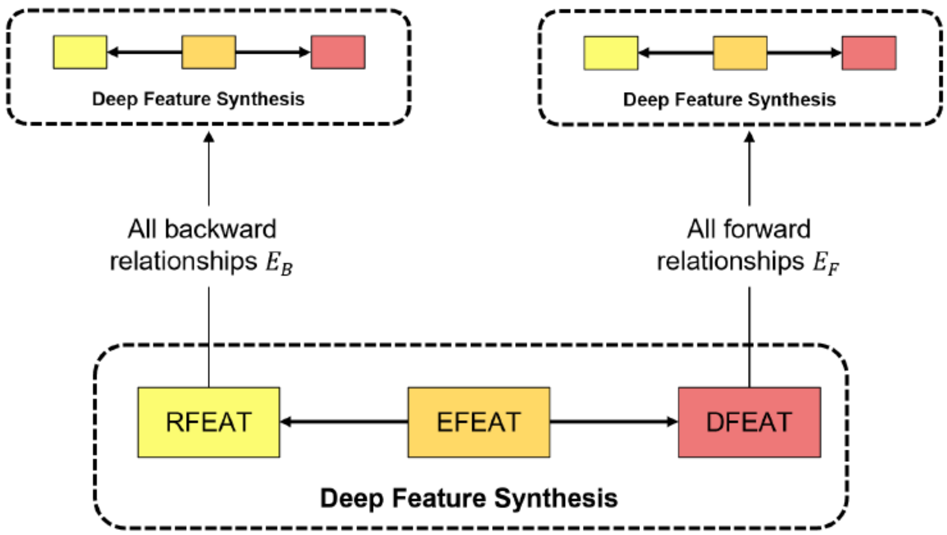

Figure 10: Relationships Among the Three Features

DFS primarily relies on two basic operations: aggregation and transformation. These operations are typically used in the context of relational data, where data is organized in related tables.

- **Aggregation operations:** Aggregation operations calculate statistical data from a subset, such as count, sum, average, median, and standard deviation.
- **Transformation operations:** Transformation operations perform actions on columns within a single table, such as addition, subtraction, multiplication, division, date differences, and string concatenation.

This study utilizes the Featuretools library for Deep Feature Synthesis to construct a predictive model capable of identifying and forecasting when a change might occur in the performance of an athlete or team. Initially, we preprocess the collected data to clean it and format it suitably for feature synthesis. Subsequently, we define an entity set that reflects the main elements in a game, such as players and scoring events, and connect these entities through logical relationships. DFS constructs features by applying feature operators to the entity relationships within the entity set. We chose the momentum difference ΔM as the target, where ΔM is defined by the equation:

$$\Delta M = M_t - M_{t+1} \tag{5}$$

Despite the broad set of features provided by DFS, we observed that the inclusion of numerous features might unnecessarily increase the model's complexity. Thus, we adopted a manual feature optimization approach, eliminating those features considered to have a minor contribution to the predictive goal, using domain knowledge and feature selection techniques. After manually adjusting the feature set, the model's predictive performance significantly improved. Specifically, improvements were observed in [insert specific performance metrics, such as accuracy, F1 score, AUC, etc.]. This indicates that reducing the number of features and focusing on the most relevant information can enhance the model's generalization ability and predictive accuracy.

The effectiveness of manual feature engineering may stem from several aspects. Firstly, by eliminating redundant and irrelevant features, we reduced the risk of model overfitting. Secondly, the manual screening process might have excluded noise features, improving the signal-to-noise ratio of the remaining features. Lastly, this process also helped us integrate domain experts' knowledge and intuition into feature selection, ensuring the model's inputs are more aligned with the actual application scenario's needs.

Table 7: Correlation of Features

Feature	correlation
DIFF(PERCENTILE(win_streak(p1)))	0.440255
DIFF(p1_winner)	0.238503
DIFF(p1_break_pt_won)	0.272255
DIFF(p2_unf_err)	0.153662
p1_score	0.080232
p1_net_pt_won	0.073064
point_victor	-0.144533
DIFF(server)	-0.469462

5.2 Introduction of CatBoost

CatBoost is an innovative Gradient Boosting Decision Tree (GBDT) algorithm that incorporates the functional characteristics of Gradient Boosting and Decision Trees [1]. It boasts advantages such as effective training outcomes, resistance to overfitting, rapid training speed, high accuracy, low memory consumption, and support for parallel computing. GBDT is capable of addressing issues encountered in processing massive datasets.

CatBoost utilizes a histogram-based optimized decision tree learning algorithm, which is shown in Figure 11. The core idea of this algorithm is the pre-discretization of continuous feature values into a finite number of discrete values. Specifically, it divides the continuous feature domain into k intervals of equal width, using the endpoints of these intervals as the values for discretization. This approach creates a histogram of width k, which is used during training as a means to quickly calculate the distribution of features and gradient information. Rather than using the original continuous feature values directly during sample traversal, the algorithm employs the discretized indices corresponding to these values, significantly reducing memory usage and enhancing computational efficiency. After a single traversal, the histogram for each feature accumulates necessary statistical information, such as gradients and second-order derivatives. The algorithm then traverses the discrete values of the histogram, utilizing the accumulated statistical information to identify the optimal splitting point for each feature.

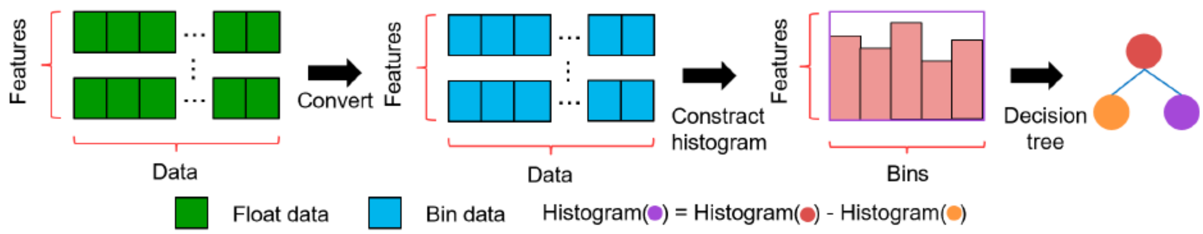

Figure 11: Histogram-Based Decision Tree Algorithm

CatBoost has implemented an efficient tree growth strategy known as the Level-wise strategy, which is shown in Figure 12. This strategy ensures uniform growth at each level of the tree, evaluating and splitting all leaf nodes in the current layer simultaneously during each round of splitting. This approach emphasizes balanced growth across the entire layer, rather than solely optimizing the leaf with the maximum split gain. To control model complexity and prevent overfitting, CatBoost's Level-wise strategy also imposes a depth limit on growth. This means the model's growth halts upon reaching a specific depth, even if some leaves within the current layer could potentially achieve higher split gains, thereby ensuring the model's generalization capability.

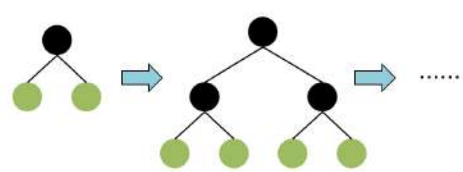

Figure 12: Diagram of Level-Wise Tree Growth

5.3 Result Analysis

Based on the features extracted using Deep Feature Synthesis, we divided the dataset into 80% for the training set and 20% for the test set. We evaluated the model's performance using metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2), as detailed below:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \tag{6}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2} \tag{7}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right| \tag{8}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2} \tag{9}$$

Simultaneously, we have plotted a graph comparing the model's predicted values against the actual values for an intuitive comparison, as shown in Figure 13.

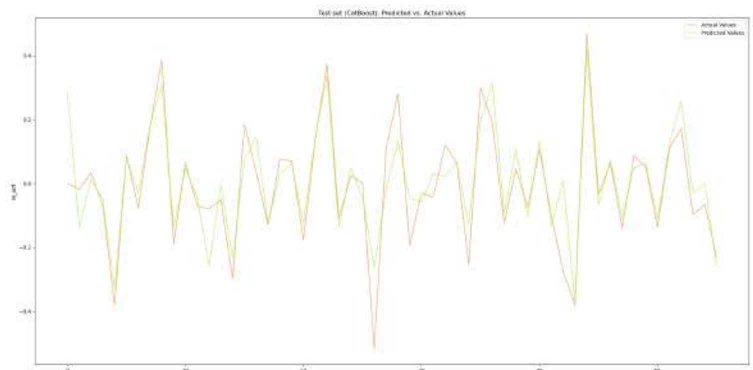

Figure 13: Test Set (CatBoot): Predicted vs. Actual Values

Following this, we selected the first set of the ”2023 Wimbledon-1304” match to use our model in predicting swings within the match and to conduct an analysis. The results are illustrated in the Figure 14. When the absolute value of a swing is high, it typically correlates with a significant change in momentum, further validating the accuracy of our model. The grey circles in the figure denote the fluctuations in momentum, or swings, caused by the change of server at the end of each game. The points encircled in red, upon comparing with the original data, indicate that a negative swing value represents a player making an unforced error, while a positive value indicates the opponent making an unforced error. The points circled in black signify moments when the opponent won a point at the net, and those circled in blue represent successful breaks by the opponent.

Based on the features extracted through Deep Feature Synthesis related to the ΔM correlation index, and the statistical analysis of events occurring at points of high absolute swing value from the previous section, we can infer some features most closely related to swings. First is whether one is serving; it is observed that a change in server significantly impacts the swing value, though this factor cannot be actively enhanced since the ideal scenario is for the game numbers of each server to converge towards a 1:1 ratio. While we cannot decide the ratio of serving, conversely, we can strive to increase the number of breaks. The analysis reveals that occurrences of breaking serve

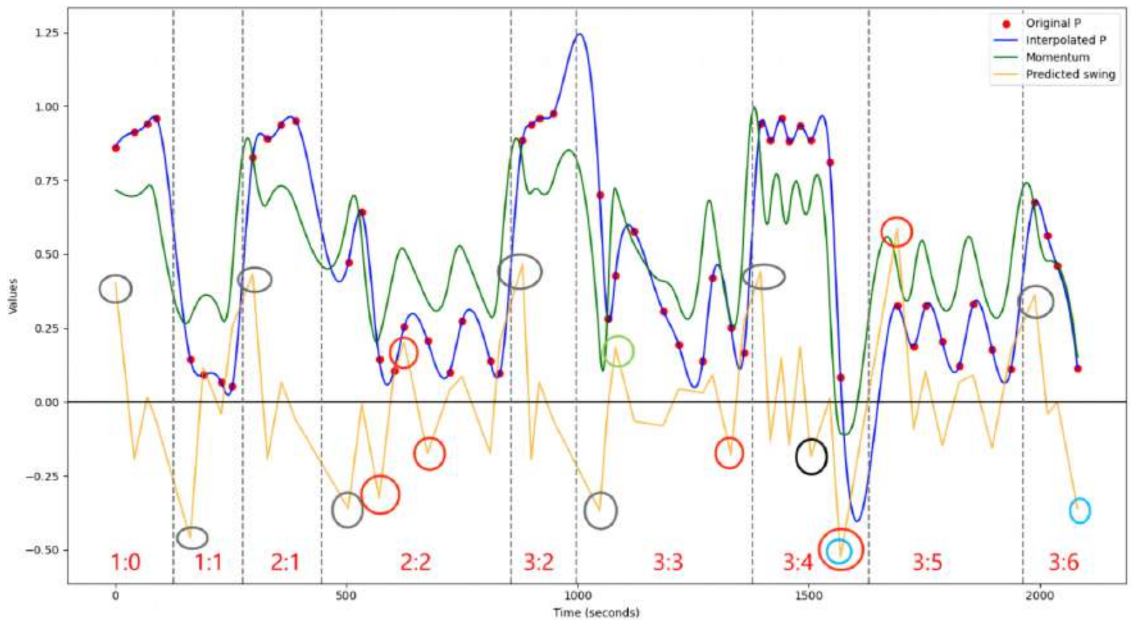

Figure 14: ΔM and Momentum

significantly affected the swing value, with one instance being the maximum value in the entire set. Next is the consecutive scoring streaks; in the feature extraction phase, consecutive scoring streaks showed a dominant correlation with ΔM, leading us to believe there is a strong correlation between consecutive scores and swings. Lastly, whether an unforced error was made is also considered; analysis shows that significant swing values often accompany a player making an unforced error, thus marking unforced errors as an important factor.

6 Task 4: Model Generalization & Sensitivity Analysis

6.1 Model Testing

As shown in Figure15-17, in testing our predictive model with datasets from the Wimbledon women's singles, European Open men's singles, and US Open men's singles, we observed that the prediction accuracy for women's singles data ($r^2 = 0.1059$) was significantly lower than that for the other two events (European Open men's singles $r^2 = 0.5960$ and US Open men's singles $r^2 = 0.4226$). This discrepancy may stem from inherent differences in physiological and technical characteristics between male and female tennis players, which could lead to gender differences in key match data, thereby affecting the accuracy of the predictive model. Additionally, the Wimbledon Championships exhibit a difference in match formats between men's and women's singles, specifically, men's matches are best of five sets while women's are best of three sets, which could be another key factor contributing to the model's predictive deviation.

In contrast, although the other two events showed better prediction outcomes, they still did not reach the prediction accuracy observed for Wimbledon men's singles datasets. Possible explanations include Wimbledon being played on grass surfaces, whereas the other two events are held on hard courts. Grass surfaces, compared to hard courts, cause significant differences in ball bounce

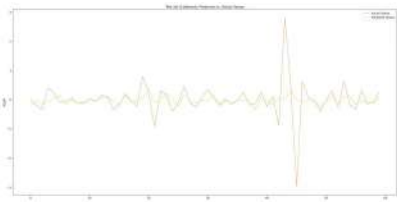

Figure 15: WC(Women)

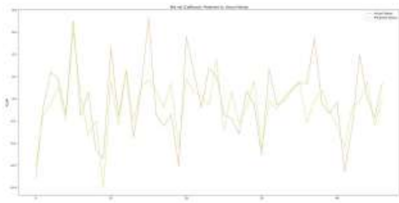

Figure 16: EURO Open(Men)

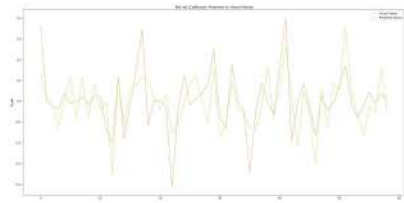

Figure 17: US Open(Men)

characteristics and player movement. Furthermore, Wimbledon's outdoor setting is subject to more unpredictable factors, such as weather conditions. The tournament format of Wimbledon, especially the final set without a tie-break rule, adds to the unpredictability of match outcomes, which might explain the variation in prediction results to some extent.

Our model is robust and precise, developed specifically for Wimbledon men's singles, yet the strategies applied are universal. When we identify poor performance in certain matches, our first step is to conduct a performance evaluation, quantitatively analyzing the model's performance metrics, such as accuracy, F1 score, ROC-AUC, etc. We then pinpoint the specific circumstances under which the model performs poorly, whether it be a particular round, against specific players, under certain weather conditions, or other specific situations. This is followed by an in-depth analysis of instances where the model predicted incorrectly, examining the differences between the model's predictions and the actual outcomes. We attempt to categorize errors to see if there are any commonalities.

Subsequently, we conduct a feature importance analysis using CatBoost's built-in feature importance tool to assess which features currently in the model have the greatest impact on the predictions. We analyze those features that contribute minimally to the model to determine whether they are unimportant or not utilized correctly. Then, we review feature engineering: revisiting the features used in the current model, especially those generated automatically through featuretools, as well as any manually constructed features. We evaluate whether important interactions or multivariate relationships were overlooked. Collaborating with experts in the relevant fields, we discuss factors that may influence match outcomes but have not been captured by the model, such as a player's psychological state, injury history, or pre-match preparation. We ensure that all critical domain knowledge has been transformed into features that the model can utilize. Finally, we consider integrating other data sources into the model, such as social media sentiment analysis, player training data, or more detailed match statistics, to identify the causes of the model's poor performance and achieve a more accurate model.

6.2 Extension of Our Model

Women's Matches: Women's matches significantly differ from men's in terms of match pace and power, which could affect the model's generalization ability. For instance, women's matches typically have fewer sets than men's, possibly necessitating different adaptability of the model to match length. Furthermore, the strategies and styles adopted by male and female players during matches also vary [4]; compared to men, women are more likely to make unforced errors at crucial moments. As the score becomes more critical, women may adopt more conservative and less aggressive strategies. Therefore, when developing models tailored to women's matches, we might focus more on indicators such as match pace, unforced errors, running distance, and serve speed.

Men's Matches in Other Competitions: The format and surface type (grass, hard, clay) of different tournaments, such as the Grand Slams, also influence matches. Our model may be overfitted to specific types of competition data and thus might require adjustments to adapt to other types of matches. However, as our previous tests have shown, while the surface type does not significantly impact prediction outcomes, the difference in tournament formats could lead to substantial variation in results. Additionally, the venues of different tournaments can vary greatly. Thus, in future model development, we might pay more attention to features induced by differences in tournament formats and the impacts of weather and geographical factors.

References

[1] Prokhorenkova, L. et al. "CatBoost: unbiased boosting with categorical features." NIPS (2018), 6639–6649.

[2] Kanter, J. M., and Veeramachaneni, K. "Deep feature synthesis: Towards automating data science endeavors." IEEE (2015), 1-10.

[3] Horvat, T., and Job, J. "The use of machine learning in sport outcome prediction: A review." WIRES DATA MIN KNOWL 10,5 (2020), e1380.

[4] Paserman, M. D. "Gender differences in performance in competitive environments: evidence from professional tennis players." IZA Discussion Paper (2007), 2834.

[5] `https://theanalyst.com/eu/2022/03/capturing-momentum-in-tennis/`

[6] `https://www.atptour.com/en/`

Memo

Momentum, as quantified in our study, is captured through two primary metrics: the player's current win rate P and the rate of change in win rate over time $\frac{dP}{dt}$. The current win rate reflects a player's immediate past performance, serving as a static indicator of momentum. In contrast, the rate of change in the win rate offers a dynamic view, revealing trends in performance and potential shifts in momentum due to critical in-match events, such as breaks of serve or scoring sequences.

Strategic Insights for Coaches:

- ***Recognize Momentum Shifts:*** Coaches should train players to recognize the signs of momentum shifts, both in themselves and their opponents. This awareness can be developed through reviewing match footage, focusing on pivotal moments that led to significant changes in win rates.

- ***Mental Resilience Training:*** Players should be equipped with mental resilience strategies to withstand pressure during critical points. Techniques such as visualization, focused breathing, and positive self-talk can help maintain composure and counteract negative momentum.

- ***Error Management:*** Our analysis indicates that players' reactions to their own and their opponents' unforced errors significantly impact momentum. Coaches should emphasize error management, encouraging players to maintain a positive mindset and quickly recover from mistakes.

- ***Strategic Play Post-Error:*** Training should include strategies for immediately following up an error, whether it's an opponent's error (to capitalize on the momentum shift) or the player's own (to minimize negative impact). This includes focusing on high-percentage shots and possibly adjusting serve or return strategies to regain control.

- ***Adapting to Opponent's Momentum:*** Players should be trained to adapt their game plan based on shifts in their opponent's momentum. This involves being flexible with strategies, whether it means being more aggressive when sensing an opponent's momentum decline or being more conservative when the opponent is gaining momentum.

- ***Physical Conditioning:*** Physical endurance plays a crucial role in maintaining and combating momentum. A well-conditioned player is better equipped to execute strategies effectively throughout the match, despite the mental and physical toll of momentum swings.

Momentum is a tangible force in tennis that significantly impacts match outcomes. By understanding and measuring momentum, coaches can better prepare their players to navigate the complex dynamics of match play. Implementing the strategies outlined above will enable players to better manage and exploit momentum shifts, leading to improved performance and outcomes in competitive matches.

Report on Use of AI

1. OpenAI ChatGPT (Jan 10, 2024 version, ChatGPT-4)
   Query1: How can we perform interpolation and differentiation on discrete values in Python? Please provide an example.
   Output: To perform interpolation and differentiation on discrete values in Python, we can use libraries such as numpy for handling arrays and scipy for interpolation and differentiation functionalities. Here's a step-by-step example:

   Interpolation: Use scipy.interpolate module to create an interpolation function based on the discrete data points. This can be done using methods like interp1d for 1-dimensional interpolation.

   Differentiation: Once we have the interpolation function, we can differentiate it. For numerical differentiation, scipy.misc.derivative can be used to compute the derivative of the interpolation function at any given point.

   Let's look at a practical example:

```
import numpy as np
from scipy.interpolate import interp1d
from scipy.misc import derivative

# Sample discrete data points
x = np.array([0, 1, 2, 3, 4, 5])
y = np.array([0, 0.8, 0.9, 0.1, -0.8, -1])

# Create a linear interpolation function
f_linear = interp1d(x, y, kind='linear')

# Create a cubic interpolation function for a smoother curve
f_cubic = interp1d(x, y, kind='cubic')

# Now, to differentiate the cubic interpolation function at x = 2.5
# We need to define a small function that takes a single argument for scipy
    ↪ .misc.derivative to work
def cubic_interp(x_point):
    return f_cubic(x_point)

# Compute the derivative at x = 2.5
derivative_at_2_5 = derivative(cubic_interp, 2.5, dx=1e-6)

print("The derivative of the cubic interpolation function at x = 2.5 is:",
    ↪ derivative_at_2_5)
```

   In this example, we first interpolate the given discrete data points using both linear and cubic methods to get a smoother representation of the underlying function. Then, we calculate the derivative of the cubic interpolation function at a specific point (x = 2.5) using the derivative function from scipy.misc. This approach allows us to estimate the rate of change (derivative) of interpolated values between discrete data points.

2. OpenAI ChatGPT (Jan 10, 2024 version, ChatGPT-4)
   Query1: When I finally combined the three using a voting classifier and determined the best

weights through cross-validation score iteration, I found the results to be 0.485 for logistic regression, 0 for random forest, and 0.512 for GBM. What can you interpret from these results, and what could be the reasons for the random forest weight being 0?
Output: When you used a Voting Classifier to combine logistic regression, random forest, and GBM, and determined the best weights through cross-validation score iteration, the weight distribution you achieved (logistic regression: 0.485, random forest: 0, GBM: 0.512) revealed several key insights:

The Dominant Role of GBM: GBM received the highest weight (0.512), indicating that on your dataset, the GBM model provides more accurate predictions compared to the other models. GBM can handle complex non-linear relationships well and can build precise models by progressively reducing residuals, which is likely why it was assigned a higher weight.

The Contribution of Logistic Regression: Logistic regression got the second-highest weight (0.485), suggesting that even though it is a linear model, it still offers valuable predictions on your dataset. This might imply that there is some degree of linear separability in the dataset, which logistic regression can capture, contributing to the final prediction outcome.

The Zero Weight for Random Forest: Random forest receiving a weight of zero in this combination could have several reasons. Firstly, it might be because, on your specific dataset, the prediction performance of the random forest is not as good as GBM and logistic regression. As both random forest and GBM are tree-based models, there might be some redundancy between them, and in your case, GBM obviously performs better. Secondly, it could also be due to the interplay between models—GBM may have already captured all the useful information that random forest could provide, plus the logistic regression adds a complementary linear perspective, making the contribution of the random forest less significant in the final combination.

Conclusion This weight distribution result highlights the relative importance and contributions of different models on a specific dataset and problem. GBM received the highest weight due to its ability to handle complex non-linear relationships, logistic regression demonstrated the importance of linear relationships in the dataset for prediction results, and the contribution of the random forest in this particular combination is considered minimal. This finding suggests that understanding the performance of each base model on a specific dataset and their interactions with each other is very crucial when building an ensemble model that is both efficient and accurate.

3. OpenAI ChatGPT (Jan 10, 2024 version, ChatGPT-4)
Query1: Please help me polish the paragraph below. The polished paragraph should be formal and clear, characterized by accuracy, conciseness, logicality, and appropriate technicality

Develop a Model: The model must quantify the flow of play and the relative performance of players throughout the match. It should account for the fact that serving players typically have a higher probability of winning points, which should be integrated into the analysis logically. Momentum vs Randomness: Address the skepticism around momentum by analyzing whether shifts in play and successful runs by players are purely random or if they exhibit patterns that can be attributed to momentum. Predicting Swings: Examine the existence of indicators that can help determine when the game situation shifts from one player to the other. We can use the provided match data, especially when the match situation changes, to develop a model to

predict the change of situation in a match. We can analyze the relationship between different factors (e.g., point differential, consecutive points scored, percentage of serves won, etc.) and situation changes and identify the most relevant factors. Through this model, we can provide indicators or signals to help coaches determine when the situation in a match may change and adopt strategies accordingly. Generalization: Test the effectiveness of the developed model in other competitions. We can apply the models we develop to other matches, including different matches, different court surfaces, and other sports (e.g., women's matches, and other ball sports) to assess the applicability and accuracy of the models. By comparing the results with those of actual matches, we can assess the predictive power of the model and identify factors that may need to be improved or adjusted.
Output: Developing a Predictive Model: It is imperative to construct a model that precisely quantifies the flow of play and the relative performance metrics of players throughout a tennis match. This model must consider the increased likelihood of point wins by those serving and incorporate this element coherently into the evaluative framework.

Momentum versus Randomness: There remains skepticism regarding the concept of 'momentum' in sports. Our objective is to rigorously analyze whether observable shifts in play and sequences of successful outcomes for players are attributable to randomness, or if they indeed reflect discernible patterns consistent with momentum.

Predictive Indicators of Swings in Play: The project aims to identify reliable predictors that signal shifts in match dynamics from one competitor to another. By utilizing comprehensive match data, particularly during pivotal transitions in play, we will develop a sophisticated model capable of anticipating these shifts. We will investigate the correlation between various factors—such as point differentials, runs of consecutive points, and service game win percentages—and changes in match status to isolate those with the highest predictive value. This model will equip coaches with actionable insights to foresee and adapt to evolving match situations.

Model Generalization and Validation: To ascertain the robustness of our predictive model, we will extend its application to a variety of competitive contexts. This will involve testing the model across a spectrum of matches, different playing surfaces, and potentially other racquet sports, including women's competitions. By benchmarking the model's predictions against actual match outcomes, we will be able to evaluate its predictive accuracy and identify any elements that require refinement.