Problem Chosen	2025 MCM/ICM Summary Sheet	Team Control Number
C		240718

Quantifying momentum, grasping victory in tennis

Summary

Physics defines where momentum as "the strength or force gained by motion or by a series of events that keeps an objective moving." In tennis, it is the psychological and physical effects of momentum that determine the direction of a match. The aim of this study is to **investigate the impact of momentum in tennis matches through a data-driven approach, and to develop a model to predict changes in momentum in matches.** Using the Wimbledon 2023 men's singles tournament as a case study, we analyzed match data to quantify momentum in matches and assess its impact on match outcomes.

Firstly, we defined a series of momentum metrics based on factors such as "score, aces, and double faults". Using these metrics, we utilized the **Random Forests Algorithm** to develop a dynamic model capable of tracking and evaluating a player's performance during a match in real time. The model takes into account the higher probability of the **serving team winning points** in a tennis match, weights the momentum score, and visualizes the flow of the match.

Secondly, with respect to the role of momentum, our model challenges the conventional wisdom that the effect of momentum on the outcome of a match is random. To this end, we designed a model that reflects the percentage of players' momentum and used a **logistic regression model** to make predictions. Some new factors were defined to quantify momentum, which referred to **"service score rate, break failure rate, net scoring rate and so on",** proving that momentum can indeed predict the outcome of a match, and that the accuracy of our model can reach up to **90% and more**.

Thirdly, inspired by the **Sliding Window Algorithm,** we designed a new quantitative model for momentum and examined its effectiveness in predicting match outcomes. We then utilized the use of historical data to identify key factors that lead to changes in momentum and predict shifts in momentum in future games. We also proposed some **model-based advice** for players going into a new match accordingly. After testing, our model can predict momentum shifts with a success rate of **71% percent**.

Finally, we applied the model to data from other games to test the model's ability to generalize. Although the model performed poorly in some cases, this prompted us to identify and suggest additional factors that may need to be included in future models, such as the **physical condition of the players, weather conditions,**as well as **psychological stress**.

Through this study, we have provided coaches and players with data-based insights to better understand and apply momentum shifts in matches, providing them with strategic advice going into new matches. The results of our study are not only applicable in tennis, but also **informative for other sports** that require an understanding of dynamic competitive states.

Keywords:

Momentum Analysis; Predictive Modeling; Random Forest; Sliding Window; Logistic Regression; Data Visualization; Generalization Capability

Contents

1 Introduction

1.1 Background

“Tennis more than any other sport, is a game of momentum. The absence of a clock to do the dirty work of finishing off an opponent, and a scoring system based on units used, makes the flow of the match much more important than any lead that has been established."

——Chuck Kriese

Physics defines where momentum as “the strength or force gained by motion or by a series of events that keeps an objective moving.” [1] In tennis, it is the psychological and physical effects of momentum that determine the direction of a match. A player seemingly in the ascendancy during a match is often said to “have the momentum”. Momentum in tennis can swing wildly from point to point, game to game, set to set. Swings in momentum are referred to as turning points. These can be obvious: players switching tactics after losing a set; a brilliant winner went on the ropes in a rally or an untimely double fault causing a opponent tightening up.

However, sometimes momentum can be so small as to be imperceptible, it is difficult to measure and it is not readily apparent how various events during the match act to create or change momentum if it exists. By understanding and tapping momentum, players can employ methods and tactics in games to ensur e they are in control of momentum rather than a victim of it.

1.2 Restatement of the Problem

Through in-depth analysis and research on the background of the problem, combined with topic specific constraints and requirements given, the restate of the problem can be expressed as follows:

- **Construct a model to capture the flow of play as points occur**. Identifying which player is performing better at a given time in the match, as well as how better they are performing. A visualization based on the model is required to depict the match flow. It is also noteworthy that the player to serve are supposed to be factored in to the model.
- **Use the model to assess whether “momentum” plays any role in a match**, as well as swings in play and runs of success by one player are random.
- **Identify indicators of the changing flow of play from favoring one player to the other**. Use the data provided to develop a model that predicts these swings in the match and try to probe into the most related factors. Advise a player going into a new match against a different player with the differential in past match “momentum” swings.
- **Test the developed model on other matches and identify factors that might need to be added.**
- **Produce a report of no more than 25 pages with the above findings and include a one-to two-page memo**,summarizing the results with advice for coaches on the role of “momentum”, and how to prepare players to respond to events that impact the flow of play during a tennis match.

1.3 Our Work

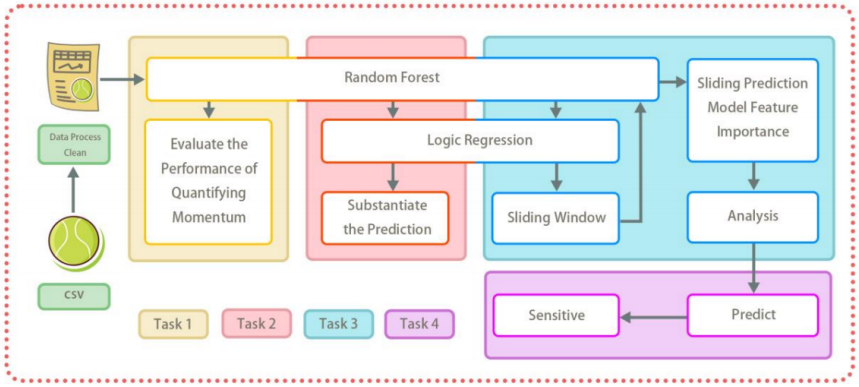


Figure 1: Work Process

2 Assumptions and Justification

- **Athletes will not be affected by the results of previous matches while playing the current match**. Supposing that the athlete is always in a good state of mind and that his or her performance in the previous game does not affect the outcome of the current game.
- **Athletes have a fixed interval between each score**. Supposing the interval between goals can be equated to the concept of time. In this way the time cost of scoring is quantifiable and relatively stable in the tennis match and we can analyze game progression and athletes' performance from a new perspective.

3 Notations

Table 1 shows the necessary notations and signs used in this paper. Other notations and signs will be declared or defined when using.[2]

Table 1: Notations

Symbols	Descriptions
S_{P1i}	player 1's momentum integral value in one match
S_{P2i}	player 2's momentum integral value in one match
A_r	service score rate
F_r	service failure rate
B_r	break success rate
B_{fr}	break failure rate
N_r	net scoring rate
E_r	error ratio
T_s	the total number of serve in the set
M	momentum

4 Data Preprocessing

4.1 Data Cleaning

The data cleaning processes are as follows.

Overview of data sets: This dataset is derived from the featured races of the Wimbledon Championship and contains detailed race statistics. It is worth noting that there are a number of missing values in the dataset, particularly in the $speed_mph$ (752 missing), $serve_width$ and $serve_depth$ (54 missing each), and $return_depth$ (1309 missing) fields.

Missing value handling: In dealing with missing values in the dataset, special attention is paid to the NA values in the speed column. In an initial check, 752 missing values were found in the speed column. These missing values can occur for a variety of reasons, including data entry errors or omissions during the data collection process. There were even instances in matches 1310 and 1311 where the whole bureau had unrecorded $rally_count$ as well as speed values.

Further analysis showed that some of the missing speed values were associated with a $rally_count$ of 0, which reflected a specific scenario of a no-ball exchange during the match, not missing data, as shown in **Figure 1**. In contrast, the missing data for complete matches may stem from technical problems or human negligence in the recording process.

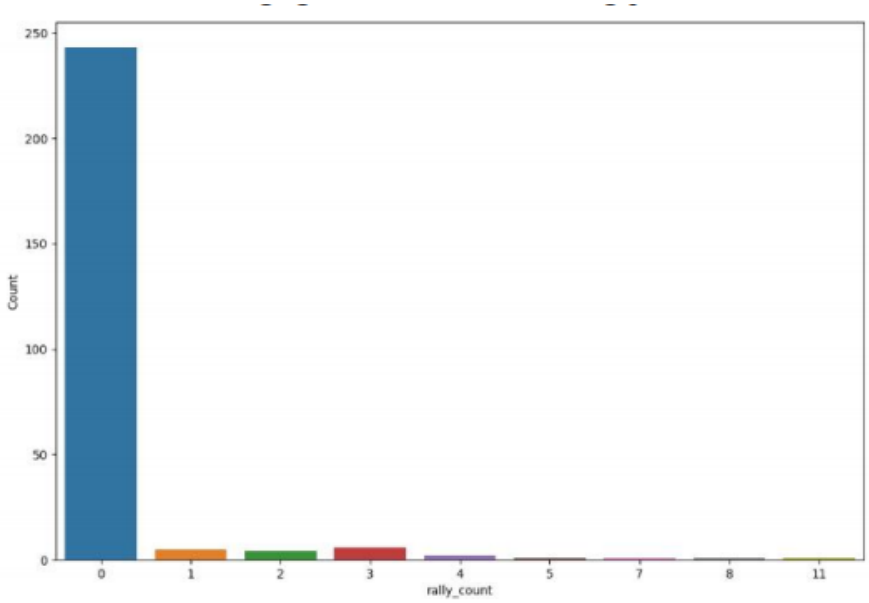


Figure 2: $SpeedNA_Count_{BY}{}_{R}allyCount$

In the case of unrecorded data for an entire race, we decided to exclude these records from the dataset given the non-direct impact of this data on model training. Data where a $rally_count$ of 0 resulted in a speed of 0 was considered a valid record and retained to accurately reflect the reality of the race.

Through this careful missing value handling strategy, we ensured the accuracy and reliability of the data analysis and laid a solid foundation for the subsequent data analysis work.

Abnormal value progress: In the dataset, there are 235 rows of records showing that when double fault is 1, the serve_width and serve_depth fields are still recorded. Theoretically, in the event of a double serve fault, the ball did not successfully cross the net and therefore the width and depth of the serve should not be recorded.

4.2 Data Cleaning

We find that serve_no and speed are related. The box plot shows the distribution of serve speeds according to the number of serves (first and second). The following points can be observed from the graph:

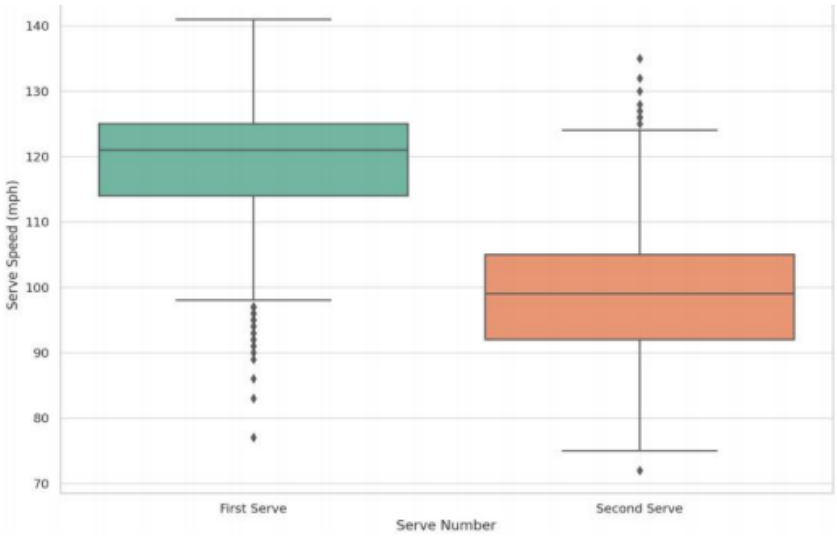


Figure 3: Boxplot of Serve Speed By Serve Number

First serve speeds were typically higher than second serve speeds, reflecting a common strategy in tennis whereby players tend to use more powerful serves on the first serve in an attempt to score points outright or to gain a favorable position, while they tend to use more secure serves on the second serve to avoid double faults.

The wider distribution of speeds on the first serve and the higher outliers at the top suggest that players' serve speeds vary more when attempting more powerful serves.

The relatively more concentrated distribution of speeds and fewer outliers on the second serve may be due to the fact that players focus more on accuracy and consistency on the second serve to minimize the risk of double faults.

These observations are consistent with the conventional strategy in tennis where the first serve is more focused on aggressiveness while the second serve is more focused on insurance. This analysis helps us to gain a deeper understanding of players' serving strategies and their potential impact on match outcomes.

5 Model Construction

5.1 Discrete Performance Evaluating Model Based on Random Forest

5.1.1 Model Preparation

According to the requirements of problem 1, this paper needs to build a model to obtain the scoring points occurring in the match. **The Random Forest model** is chosen to determine the weights of the indicators, therefore we are able to identify the better performance of the players visually, then we apply the model to as many games as possible.

- **Firstly, we acquire data information such as "ace" or "net_pt"** , which refer to a not-served shot and a player's position separately from the provided data set, result data of each

set or game can also be found.

- **Next, we try to design a model to quantify "momentum"**. Momentum can be perceived as strength or force during a match, since it is difficult to quantize, so we attempt to incorporate the Calculation of short-term indicators, transforming scoring concepts into momentum indicators.
- **Then, we discover that momentum is usually reflected as a change in game performance over a short period of time**. For example, consecutive scores can be seen as a direct reflection of momentum. By calculating short-term changes in scoring for each point (e.g., consecutive points, short-term serve success and break rates, etc.), these short-term performances can be quantified as indicators of momentum.

5.1.2 Random Forest Model

Although the ultimate goal is to predict the outcome of the game, the impact of momentum changes on the outcome of the game can be revealed by analyzing the relationship between short-term momentum indicators and the final outcome of the game. Short-term momentum indicators can be trained as features and match results (win/lose) as labels in the model to determine the association between momentum and match.

We can begin to quantify indicators by virtue of the above idea, but it is not yet possible to determine the specific levels of quantification. We then use a random forest model to analyze the importance of the characteristics of these indicators.

Random Forest is a supervised algorithm that uses an integrated learning method consisting of numerous decision trees. It's able to handle large and complex datasets, including multiple types of variables, which is useful for analyzing the impact of various factors in tennis matches. What is more, it can provide an effective method for assessing the importance of individual features in predicting outcomes, which allows to identify which factors have the greatest impact on player momentum and match results.[4]

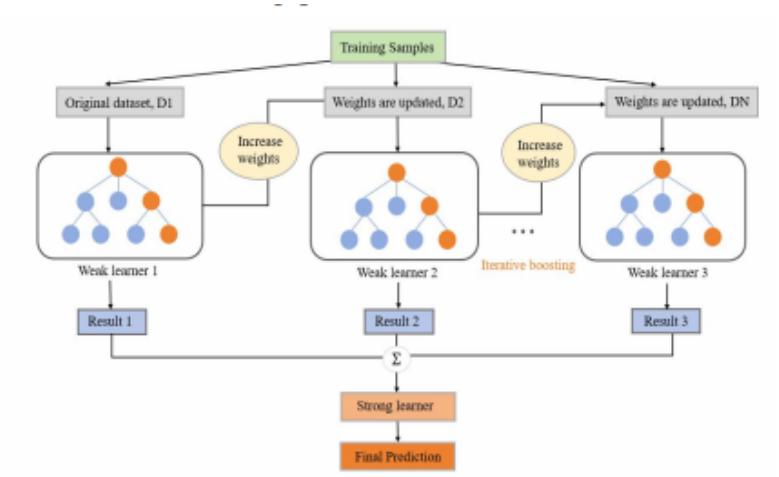


Figure 4: Concept of Random Forest

Therefore we try to describe the method by means of detail principles of the Random Forest model, including how it handles large numbers of input variables and evaluates the importance of features. We can explain the source of the dataset, the variables chosen (e.g., score, serve success rate, unforced errors, etc.), and how the data were prepared for use in the analysis accordingly.

We focus on the characteristic importance scores of each factor's impact on the outcome of the match. Specially, we weight the "serve" artificially low.

The results of the Random Forest Model analysis are in the following graphs

Table 2: Feature Importance Ranking in Random Forest Model

Order	Features	Weights
1	break_pt	0.275942
2	point_victor	0.168837
3	unf_err	0.139420
4	winner	0.108685
5	ace	0.095252
6	double_fault	0.077130
7	net_pt	0.074068
8	net_pt_won	0.039238
9	break_pt_missed	0.021428

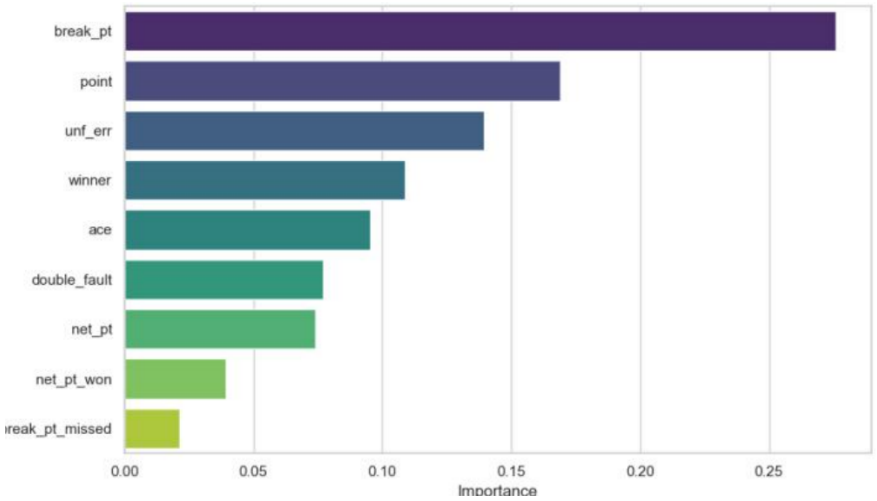
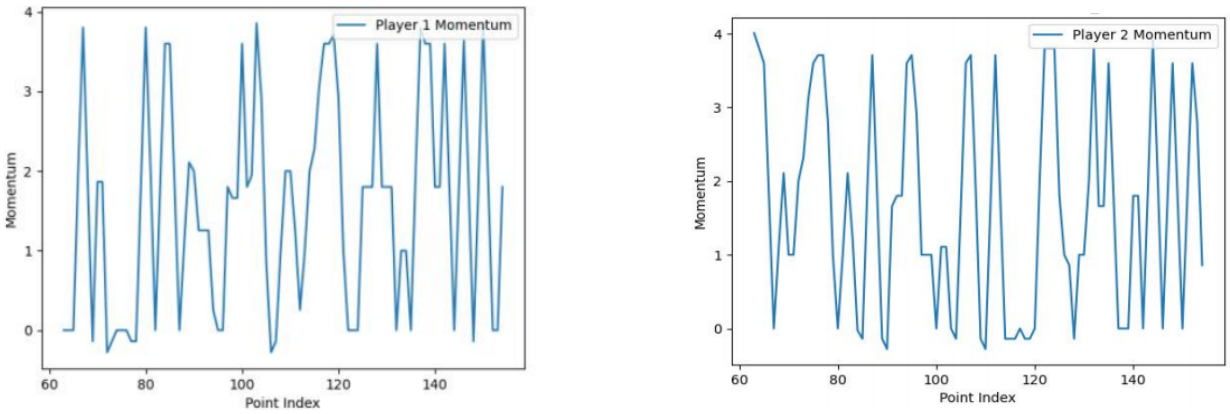


Figure 5: Feature Importance in Random Forest Model

According to the chart presented above, we define the following weights: $W = (0.275942, 0.8 \times 0.168837, -0.13$

(1)

$$M = \sum_{i=1}^{9} (W_i \times \alpha_i) \tag{2}$$



(a) First Image (b) Second Image

Figure 6: Player 1/2's Momentum in this set_no

These two images above show the performance for the players themselves:

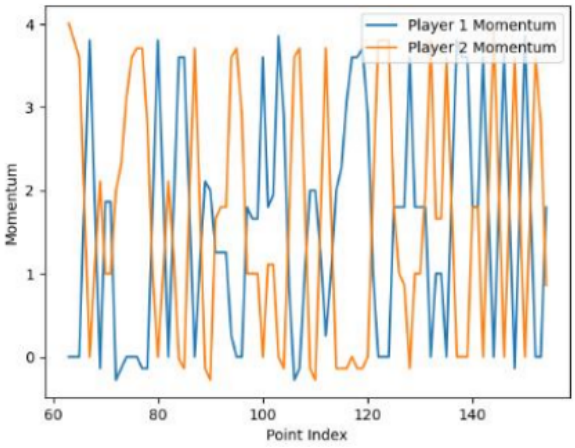


Figure 7: Player 1's vs Player 2's Momentum Comparison in the set_no

This picture above reflects the fact that the momentum values of the two players are in a state of waxing and waning, which basically corresponds with players' actual state in a real game.[3]

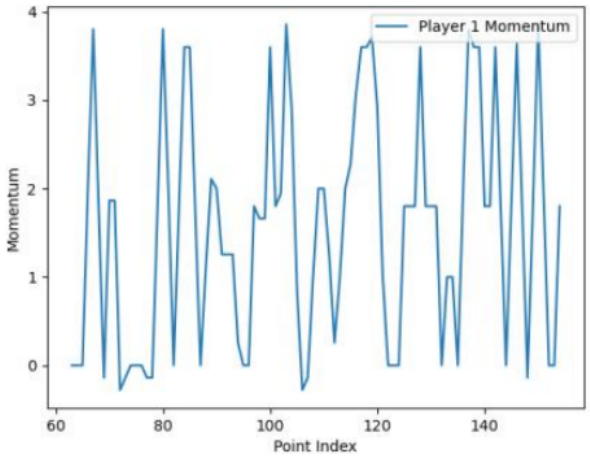


Figure 8: : Player 1-Player's Momentum Difference Comparison in this set_no

This picture above shows that in which time period which athlete performs better. For example, the red spot in the diagram indicates that at the 78-minute mark of the opening, Player 1's momentum minus player 2's momentum is negative, which represents clearly that at this time, player 2 performs better than player 1.

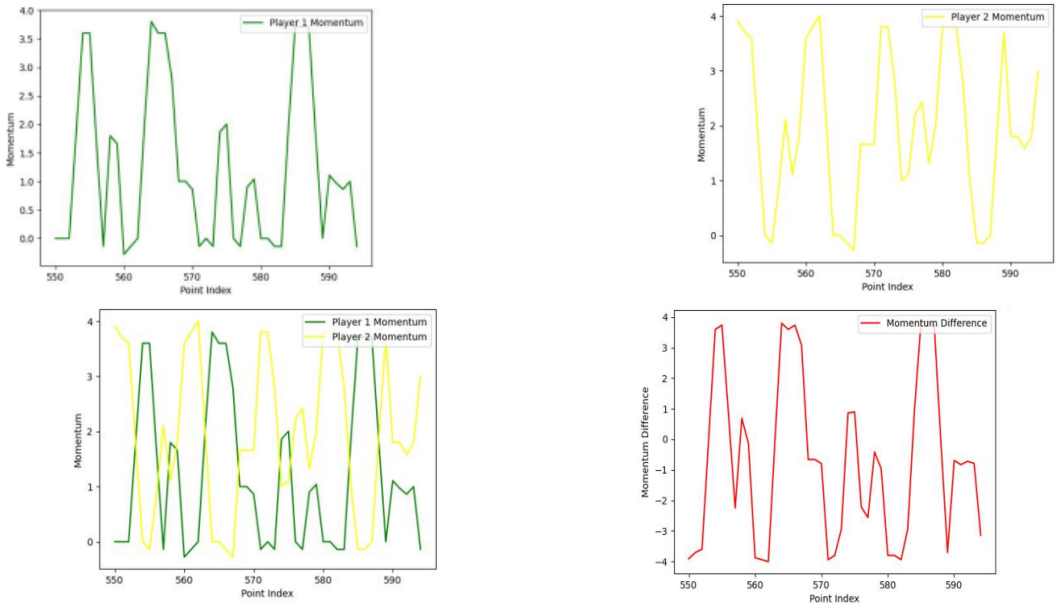


Figure 9: Player 1/2's Momentum in this set_no

5.2 Assessment Based on Logistic Regression

5.2.1 Assessment Preparation

According to the requirements of the problem, this paper needs to assess that "momentum" actually plays an role in a match and swings in play and runs of success by one player are not random but correlative. We choose **Logistic Regression in classification algorithms** to train the model, then we can obtain the training results with higher predicted value. Finally we are able to visualize the

description based on the model results to prove the crucial role "momentum" acts in the match and its high relevancy.

Through the weight we determine in the previous question, we are able to derive two chart with its horizontal axis coordinates are the number of points scored per game and the vertical axis coordinates are the momentum values for each player. Then we integrate the momentum value folds in the two charts separately

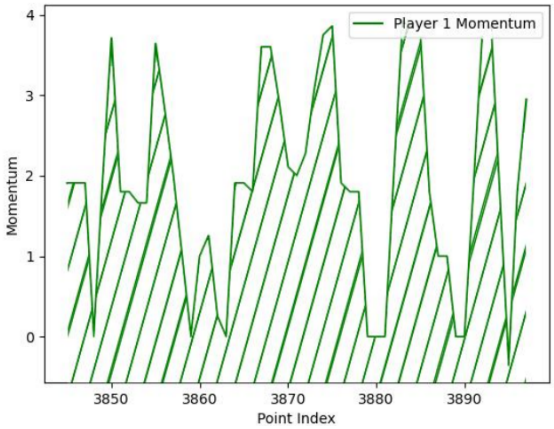


Figure 10: Player 1's Momentum in this set_no

Accordingly, we can acquire the total momentum value for each player. By analogy with the definition "work" in physics, we figure that momentum can also accumulate over time, so we define three values: S_{p1},S_{p2},Xi In the same way, we accordingly integral to get the cumulative value of the momentum of the two athletes in the current match, so we build this model:

$$Xi = \frac{S_{p1}}{S_{p2} + S_{p1}} \tag{3}$$

Since there are many rounds in the competition, the momentum performance of the athletes will vary accordingly, so there will be multiple X_i,Y_i represents the results of the set, we stipulate that :

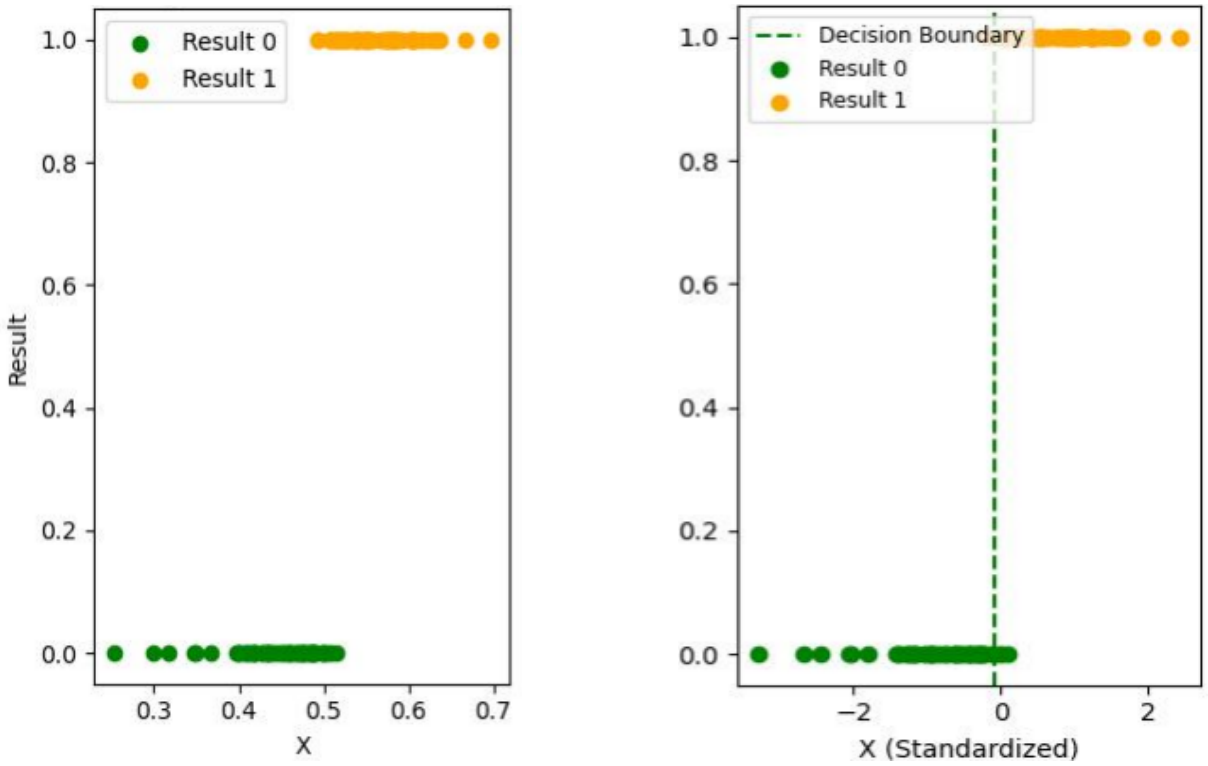
$$Yi = 0, P_1 lost the set \tag{4}$$

$$Yi = 1, P_1 won the set \tag{5}$$

So we can capture series of dataset (X_i,Y_i)

Through the original data scatter plot, we notice that the distribution of X_i is quite compact, so we choose to standardize X_i to make it more discrete, as it shown in the following diagram:

We then use **Logistic Regression:**



(a) Original Data Scatter Plot (b) Standardized Data Scatter Plot

Figure 11: Two Types of Data Scatter

$$\vec{W} \leftarrow \vec{W} + \eta \sum_{i=1}^{n} \frac{y_i \vec{x_i}}{1 + exp(y_i \vec{w} \cdot \vec{x_i})} \tag{6}$$