Problem Chosen	2024 MCM/ICM Summary Sheet	Team Control Number
C		2400228

Momentum Mastery: Unveiling Match Flows

Summary

In sports, momentum is the incredible swing that appears in a player who appears to have an advantage.

In the 2023 Wimbledon Gentlemen's final, before Carlos Alcaraz defeated Novak Djokovic, the course of the match had been in flux, with both men constantly fluctuating in momentum. To explore the changes and role of momentum in ball games, we conducted an in-depth and close study of this topic from multiple perspectives and levels, based on the data set of the 2023 Wimbledon Gentlemen's match.

First, to capture the flow of the game when the score occurs, we build a momentum calculation model. We summarize seven characteristic indicators to participate in the quantification of momentum and use the random forest model to analyze the weight of their influence on momentum, to establish the momentum calculation formula. Using the momentum calculation formula, we get the momentum time series data of each player in the match flow and predict the change of momentum between adjacent score time nodes based on the ARIMA model. In addition, we use scatter plots and line plots to visualize the flow of the match as described by the momentum calculation model.

Next, we use CUMSUM algorithm to detect the turning point in the original momentum time series not processed by ARIMA model and perform run tests on the obtained momentum turning point time series and the original momentum time series to judge the randomness of the two sequences. The results show that the change of momentum and the turn of momentum are non-random events, that is, the success of the player and the swing of the game are non-random events.

In addition, we trained four models based on machine learning using game-specific metrics and selected the most effective XGBoost regression model to predict players' momentum flow during the game after comparison. We then used the SHAP model to analyze the degree of influence of each metric in predicting momentum flow, statistically identified the six most influential factors, and made recommendations for players based on these factors.

Further, we evaluated the effectiveness of the XGBoost regression model in predicting momentum flow and confirmed that it has a small error. We also summarize the derived factors that should be considered for inclusion in future models when the model is ineffective and verify that the inclusion of these factors is reliable by retraining the XGBoost regression model and comparing the accuracy. In addition, we applied the constructed XGBoost regression model to multiple table tennis matches and calculated the accuracy of the predictions, and the validated results showed that the model also works well when applied to different ball games.

Finally, we wrote a memo to document the main findings and provide some recommendations to coaches and players based on the findings.

Keywords: ARIMA; CUMSUM; the Run Test; XGBoost regression model; SHAP model

Contents

1 Introduction

1.1 Background

Tennis is a sport enjoyed by millions of people around the world. Among them, one-on-one singles is a common way of tennis matches, it tests the athletes' physical and mental strength, and requires very good skills. In addition to the players' technical, physical, psychological, and tactical factors, one of the important factors that may affect the outcome of the game is momentum.

Momentum refers to the power or force that a player or team gains during a match through movement or a series of events, which can affect the pace and atmosphere of the game and even change the direction and outcome of the match.

In the 2023 Wimbledon Gentlemen's final, 20-year-old Spanish rising star Carlos Alcaraz defeated 36-year-old Novak Djokovic, ending the remarkable run of one of the greatest players in Grand Slam history. The match was a remarkable battle, with the two players engaged in a fierce scramble and their momentum fluctuated several times as they played. However, momentum phenomena are difficult to measure, and it is not easy to figure out how various events in a race generate or change momentum.

By studying the role of momentum and other factors in this game, we can find indicators that affect the fluctuation of game flow advantage between players. In addition, by analyzing the flow of the game when the score occurred, conducting statistics and visualization, we can identify the outstanding players and their performance in a specific period. The analysis of momentum and other influencing factors can also be applied to other sports, providing some useful references and suggestions for athletes and coaches.

1.2 Restatement of the Problem

We need to analyze all the data after the first two rounds of the 2023 Wimbledon Gentlemen's match and finish the following tasks:

- Task 1: Develop a model to identify players who excel during specific time periods in the flow of the game when scoring occurs and provide a visualization of the flow of the game.
- Task 2: Construct an evaluation model to assess the role of momentum in the race process.
- Task 3: Build another predictive model to predict fluctuations in advantage between players in the flow of a match, find the factors most associated with the fluctuations, and make recommendations for players who will be playing a new match with a different player.
- Task 4: Test the developed model in one or more other matches, identify factors that need to be incorporated into future models if the model performs poorly, and assess how well the model generalizes to other types of matches.

- Task 5: Record a memo summarizing the findings of the study and making recommendations for coaches and players regarding the role of momentum and how to prepare players to deal with factors that affect the course of a tennis match.

1.3 Data Processing

1.3.1 Data Cleaning

Observing the provided dataset, the problems that were identified and the corresponding treatments are as follows.

In the dataset, there are 752 missing values for the speed_mph attribute, 54 missing values for the serve_width attribute, 54 missing values for the serve_depth attribute, and 1309 missing values for the return_depth attribute. These missing values affect the training and prediction accuracy of the model, and data processing was performed to address these issues.

Since speed_mph has a mean of about 112.41 mph and a standard deviation of about 12.86 mph, it indicates some fluctuation in the distribution of its data, with a median of 115 mph, which means that the data is slightly skewed toward higher speeds. Given this distribution, for speed_mph, we use the mean of the columns to fill in the missing values because it is less sensitive to extreme values. For the categorical variables serve_width and serve_depth, we use the respective plurality to fill in the data. For return_depth, since it has more missing values, we use probability-based interpolation to fill in the data with random sampling, and the filling process is shown in Figure (1).

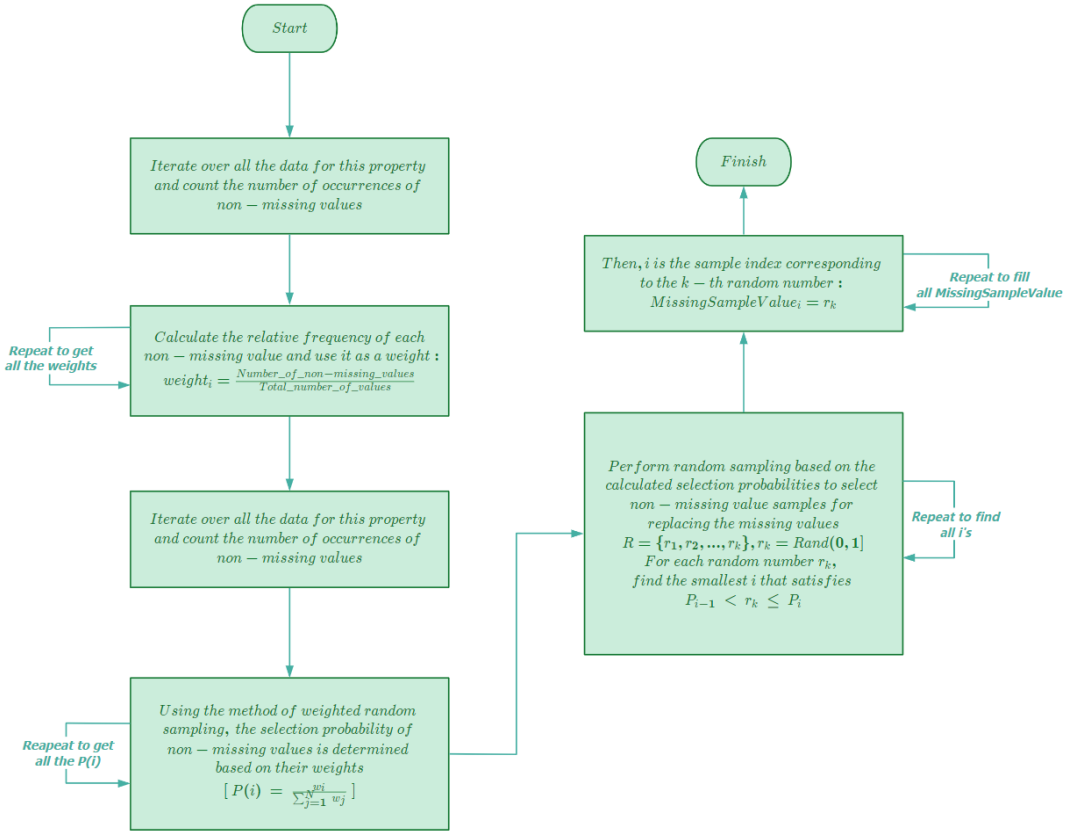

Figure 1: Flow of Completing the Missing Values of return_depth

1.3.2 Conversion of Non-numeric Text

We use solo thermal encoding to convert non-numeric data into numeric data.

1.3.3 Normalization

The goal of the normalization operation is to eliminate differences in numerical range and numerical magnitude between features to ensure that they have the same importance in the analysis and modeling process. We use linear normalization, also known as Min-Max normalization. It normalizes the data by mapping it to the interval [0, 1][1]. Specifically, we first find the minimum and maximum values of the dataset or feature and use the following formula to calculate the normalized values:

$$x_i = \frac{(x - x_{min})}{(x_{max} - x_{min})} \tag{1}$$

In the formula, x represents the original value, x_i represents the normalized value, and x_{max} and x_{min} represent the maximum and minimum values of the dataset or feature to which the value belongs, respectively.

1.4 Our Work

The work we have done on this problem C is shown in Figure (2) below.

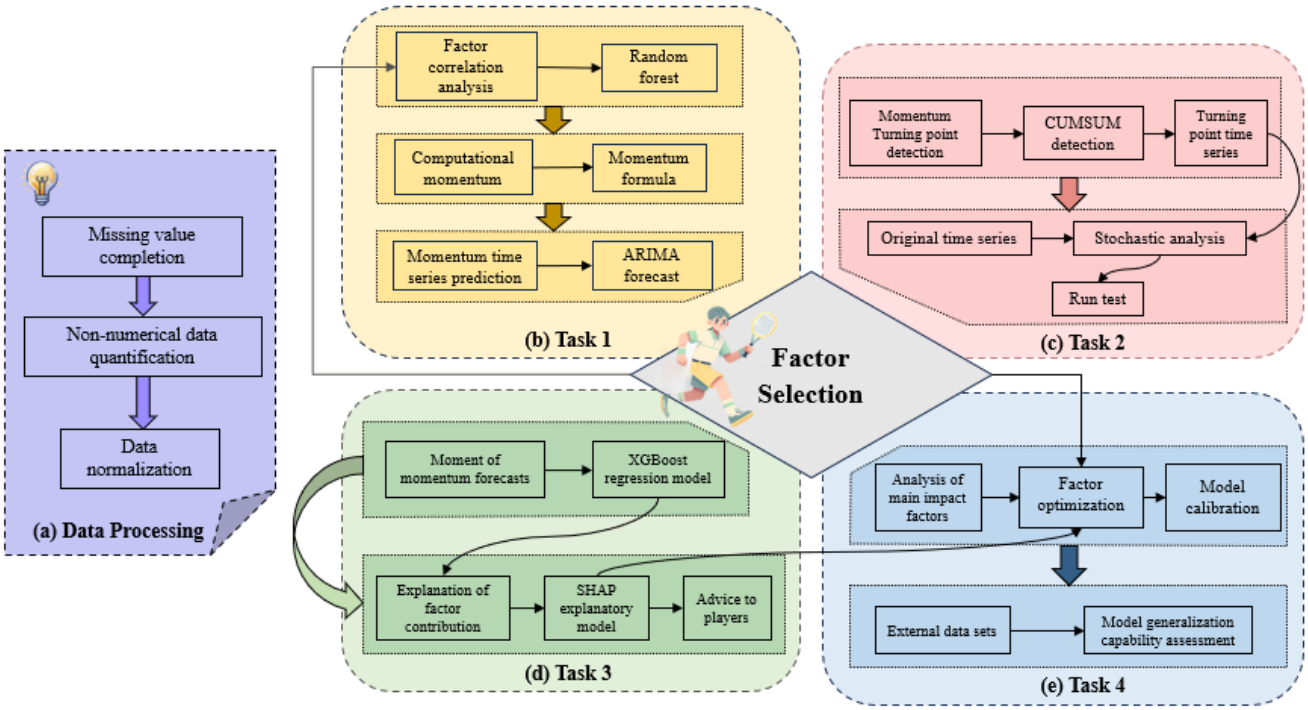

Figure 2: Our Work

2 Assumptions

- **The momentum of a player in a match is observable.** Although a player's momentum is not directly observable, it is reasonable to assume that these states are indirectly measured through observable outcomes of the match, such as games won, sets played, unforced errors, and match-winning points.

- **The data provided for modeling is complete and accurate.** It is assumed that the data collected for the competition is complete and accurate and reflects the reality of the competition.

3 Notations

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

Table 1: Notations

Symbol	Description
M_t	Momentum time series
$m,\ m_i$	Momentum value (m does not have a specific meaning)
fac_i	Value of indicator i (normalized)
w_i	Weight corresponding to the i-th indicator

4 Task 1: Momentum Calculation Model

4.1 Feature Selection

To identify a player's performance in a tennis match, we need to construct a momentum computation model capturing the flow of the match when scoring occurs. We define momentum as a measure of a player's current performance, relative to the deviation from the average or expected performance during the match. To quantify this situation, which reflects a player's scoring advantage relative to his opponent in a given time window, in this section we summarize seven characteristic metrics to participate in the quantification of momentum:

- **Scoring Advantage**: indicates the difference in scoring between two players in a given time window, reflecting the difference in scoring between players.
- **Games won**: indicates the number of games won by each of the two players in a given time window, reflecting a player's progress in a set.
- **Sets Won**: Indicates the number of sets won by each of the two players at a given time, reflecting how dominant a player is in a match.
- **Serve Situation**: Indicates the serve advantage and the percentage of points won by the serving side of both players at a given time, reflecting a player's ability to serve and control the serve set. The serve advantage refers to the difference between the points won by the serving side and the points won by the receiving side, and the percentage of points won by the serving side refers to the ratio of points won by the serving side to the total number of points in the service game.
- **Break Points**: indicates the number of break points won by each of the two players at a given time, reflecting a player's ability to attack and defend on the opponent's serve.
- **Unforced errors**: indicates the number of unforced errors committed by each of the two players in a given time window, reflecting player consistency and accuracy.
- **Winning Runs**: indicates the number of winning runs scored by each of the two players at a given time, reflecting the player's aggressiveness and initiative.

4.2 Computational Model for Momentum

4.2.1 Random Forest Based Weight Calculation

In the first stage of model building, we will rely on the seven feature metrics summarized in Section 4.1, which provide a comprehensive picture of player performance during a match. In order to quantify the relative contribution of each feature to the momentum of the match and to determine their weights, we adopt a random forest-based weighting method to quantitatively analyze the key feature metrics in tennis match data.

The Random Forest model was chosen to determine the weights of these metrics because of its ability to effectively handle high-dimensional data and assess the importance of features[2]. Random Forest improves the model's ability to fit the data and generalize by constructing multiple decision trees and combining their predictions. During the construction of each tree, the model randomly selects a subset of features, which not only increases the diversity of the model, but also allows us to assess the impact of each feature on the prediction results in different contexts. In particular, we focus on the ability of features to reduce impurity in the decision tree as a measure of their importance. The reduction in impurity indicates the effectiveness of the feature in differentiating the dataset, i.e., the degree to which the feature contributes to the change in momentum predicted by the model. By calculating the average reduction in impurity due to features in all decision trees, we obtained the weight of each feature.

After inputting these parameters into the model, we use the table to show the weights of each feature. In addition, considering that the serving side usually has a higher winning point rate in tennis matches, we pay special attention to the feature of serving situation. The model results show that serve situation is an important factor that affects the momentum of the match, and its weight is significantly higher than other features, verifying the importance of serve advantage in the match. The high value of the weights occupied by point advantage and break points indicates that these two characteristics have a large impact on the player's momentum among the seven indicators.

Table 2: Calculated Weights of the Seven Characterization Indicators

Feature	Weight
Serve Advantage	0.2500
Points Advantage	0.1800
Sets Won	0.1600
Break Points Won	0.1500
Winners	0.1300
Unforced Errors	0.0700
Games Won	0.0600

Weight broken down by Feature. Color shows Weight.

Weight

0.0600 0.2500

4.2.2 Momentum Formula

After determining the weights of each indicator, we defined the formula for calculating momentum:

$$m = w_1 fac_1 + w_2 fac_2 + w_3 fac_3 + w_4 fac_4 + w_5 fac_5 + w_6 fac_6 + w_7 fac_7 \quad (2)$$

In the formula, m_i indicates the current momentum value of the player, fac_1 to fac_7 indicate the value of each indicator, and w_1 to w_7 indicate the respective weights of the indicators.

4.2.3 ARIMA-based Momentum Timing Prediction

In order to accurately capture the change in each player's momentum at various moments in the game, we used an ARIMA model to construct time-series forecasts of momentum. This forecasting method is particularly well suited for analyzing trends and patterns in time-series data in order to predict future changes in player momentum during a match. The ARIMA model, or Autoregressive Integrated Moving Average Model, combines both autoregressive (AR) and moving average (MA) methods, and smoothes the data by differencing (I)[3]. The core of the model is to identify the autoregressive properties of the time series and the moving average properties of the error term as a means of predicting future data points. In this study, the ARIMA model is formulated as shown below:

$$M_t = c + \varphi_1 M_{t-1} + \varphi_2 M_{t-2} + \cdots + \varphi_p M_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t \quad (3)$$

In this equation, M_t is the momentum time series data we are considering. φ_1 to φ_p are parameters of the AR model to characterize the relationship between the current value and the value at the past p time points. θ_1 to θ_q are parameters of the MA model to characterize the relationship between the current value and the error at the past q time points. ϵ_t is the error term at point in time t, c is a constant term.

In constructing the model, we ensure the smoothness of the data by performing first-order differencing on the momentum time series data. Immediately after that, ARIMA(1,1,0) was chosen as the model parameter configuration by observing the autocorrelation function (ACF) and partial autocorrelation function (PACF) and discovering the first-order lag characteristics. This indicates that the model uses first-order autoregressive term and first-order difference without moving average term to capture the core dynamics of the time series.

Changes in momentum predicted by the model can visualize the flow of the game and further understanding of key turning points in the game and player performance.

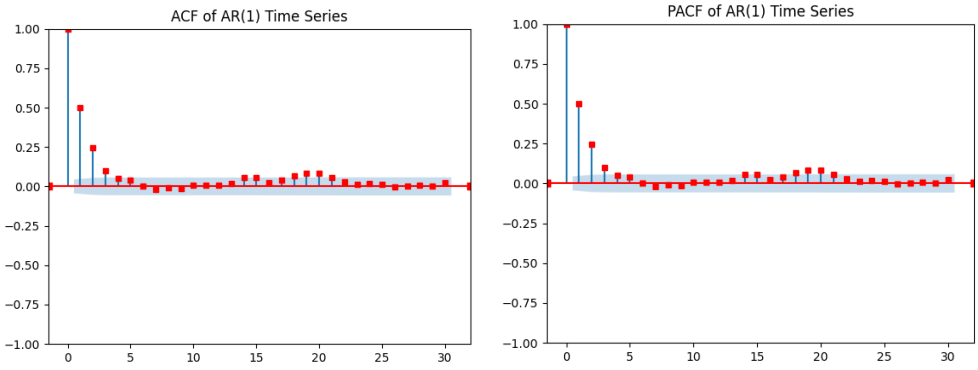

Figure 3: Plot of Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) for the Time Series of the Momentum

4.3 Visualization of the Match Process

Taking the final match between Carlos Alcaraz and Novak Djokovic as an example, we calculated the momentum values of each of the two players as the moments of the match changed according to the momentum formula, and the magnitude of the values directly reflected the degree of player dominance. We constructed a scatterplot to depict the momentum of each player over a specific time period in the flow of the match, with the color of the scatterplot reflecting the magnitude of the momentum value.

The following scatterplot reflects the momentum values of the two players during the flow of the game. It can be observed that between the 0th and 2500th second of the match flow, i.e., in the first game, Djokovic has more momentum and performs better, and around the 14,000th second, Djokovic performs the best. However, in the middle and end of the match, Alcaraz's momentum clearly outweighed Djokovic's, which set the stage for his eventual win.

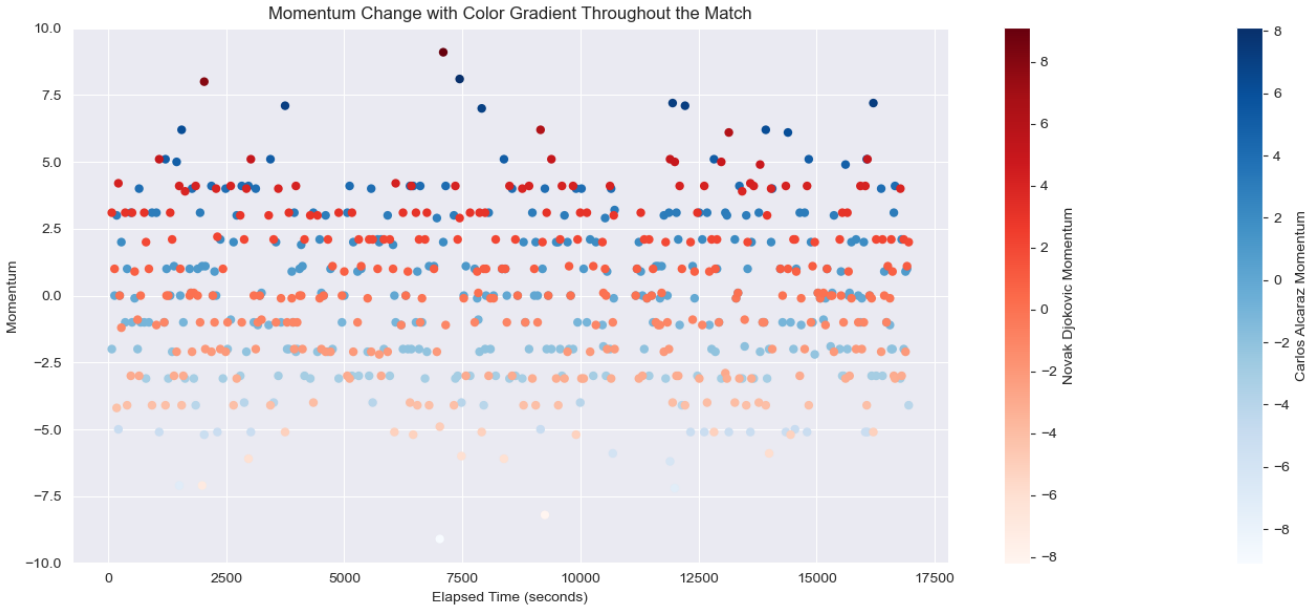

Figure 4: Scatterplot of momentum values for Alcaraz and Djokovic in the Final

We constructed an ARIMA(1,1,0) model to predict changes in momentum between adjacent scoring time nodes. In order to visualize the change in momentum for each player during the flow of the match, we build line graphs to visualize the predictions. The figure below plots the predicted change in momentum for the players in the first quarter of the period of the final match between Carlos Alcaraz and Novak Djokovic. To make it easier to see, we include the difference in momentum between the two players, which means the difference in momentum between Alcaraz and Djokovic, in the plot. The line graph depicts the momentum of the two players according to the time, and the larger the absolute value of the momentum, the more intense the match was. When the momentum value is greater than 0, we consider the momentum to be "positive", which means the player is in a favorable position. On the other hand, the momentum is negative, which means the player is in an unfavorable situation. The difference in momentum between the two players is shown by the green curve. As the game progresses.

The green curve gradually rises or falls, indicating that one player is gradually gaining momentum, the match is under his control, and the offense and defense are changing shape.

Figure 5: Momentum Swings for Alcaraz and Djokovic in the Match

5 Task 2: Momentum Role Assessment Model

5.1 CUMSUM Detection Algorithm

To explore the stochastic nature of players' momentum flow during a match, we apply the CUMSUM (Cumulative Sum) algorithm[4] in this subsection to identify significant change points, i.e., turning points in momentum, in the momentum time series. First, we utilize the seven characteristic indicators in the fourth subsection to quantify the momentum of each player during the game, and accordingly construct a momentum time series reflecting the dynamic changes of the game. The sequence not only reflects the dynamic changes during the game, but also provides a data base for assessing the randomness of players' success.

Next, we set out to identify momentum turning points - that is, significant change points in the momentum time series. Difference series computation is the first step, and for a given momentum time series $M_t = \{m_1, m_2, \dots, m_n\}$, its difference series $D = d_1, d_2, \dots, d_n$ is computed by the following equation:

$$d_i = m_{i+1} - m_i \tag{4}$$

Immediately following the computation of the cumulative sum, the cumulative sum sequence $C = c_1, c_2, \dots, c_n$ of the difference sequence D can be computed by the following equation:

$$c_i = \sum_{j=1}^{i} d_j \tag{5}$$

These two processes are combined into the following formula for the combined calculation of cumulative sums:

$$c_i = \sum_{j=1}^{i} (x_{j+1} - x_j), \mathrm{i} \geq 1 \tag{6}$$

Immediately thereafter, we identify turning points by the following expressions, which

mark the locations in the cumulative sum sequence where values change from positive to negative or negative to positive:

$$c_i * c_{i-1} < 0 \tag{7}$$

If $c_i * c_{i-1} < 0$, then position i is identified as a turning point.

Through the application of the CUMSUM algorithm, we successfully detected turning points from the original momentum time series and further constructed the turning point time series, thus accurately revealing the key moments of the momentum flow of the race.

5.2 Evaluation Model Based on Run Tests

In the established computational model of momentum, a player's momentum is determined by match-related metrics. In order to confirm whether a player's swing and success in a match is random, we use the run test method. This method was applied to two key sequences: the original momentum time series and the momentum turning point time series identified by the CUMSUM algorithm. The purpose of the run test is to assess the randomness of the changes in these sequences[5].

First, the sequence is binarized, where each element is compared to the median of the sequence: if the value of the element is greater than the median, it is marked as 1; otherwise, it is marked as 0. This process converts the sequence into a series of binary values for subsequent analysis.

Immediately after, we compute the number of trips in the sequence, defined as the length of a sequence of identical values (1 or 0). The total number of trips (denoted R) is computed by traversing the binary sequence and incrementing each time the value changes.

Next, we calculated the expectation $E[R]$ and variance $Var[R]$ of the number of tours based on the number of 1's and 0's in the binary sequence, based on the following equation:

$$E[R] = \frac{2n_1n_2}{n} + 1 \tag{8}$$

$$Var[R] = \frac{2n_1n_2(2n_1n_2-n)}{n^2(n-1)} \tag{9}$$

In this equation, n_1 and n_2 are the number of 0s and 1s in the binary sequence, respectively, and n is the total length of the sequence.

We then calculated the statistic Z and the corresponding probability p. The Z statistic is the difference between the actual number of trips (R) and the desired number of trips, divided by the standard deviation of the number of trips:

$$Z = \frac{R-E[R]}{\sqrt{Var[R]}} \tag{10}$$

$$p = 2\left(1 - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{|Z|} e^{-t^2/2}dt\right) \tag{11}$$

In this equation, $\Phi(|Z|) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{|Z|} e^{-t^2/2}dt$ is the Cumulative Distribution Function (CDF) of the normal distribution about Z.

Finally, for the obtained p -value, if it is less than 0.05, the null hypothesis of sequence

randomness is rejected, indicating that the variation in the sequence is not completely random. We label this result as a sequence $T = t1, t2, t3, \ldots, tn$, where:

$$t_i = \begin{cases} 0, & p_i < 0.05 \\ 1, & p_i \geq 0.05 \end{cases} \tag{12}$$

If the value of t_i is 1, it indicates that the event is considered non-random in our analytical framework. With this approach, we can accurately assess the stochastic nature of player momentum changes during a match, providing a new perspective for understanding momentum flow during a match.

5.3 Results of the Model

Using an evaluation model based on run testing, we performed a comprehensive analysis of the provided tennis match dataset. This involved performing a run test on the raw momentum time series of each match and its momentum turning point time series, resulting in a range of evaluation results. For example, in the match 2023-wimbledon-1301, p1_momentumisRand equals to 1, indicating that the momentum flow of the p1 player in this match was not random, but was correlated with certain factors. Therefore, we can observe from the obtained result data that the momentum changes of the players in all matches are non-random events. For example, Table (3) reflects the evaluation result data of the model:

Table 3: Results of the Run Test for the original momentum time series and the momentum turning point time series

Match Id	p1 momentumisRand	p1 turning pointsisRand	p2 momentumisRand	p2 turning pointsisRand
2023-wimbledon-1301	1.000	1.000	1.000	1.000
2023-wimbledon-1302	1.000	1.000	1.000	1.000
2023-wimbledon-1303	1.000	0.000	1.000	0.000
2023-wimbledon-1304	1.000	1.000	1.000	1.000
2023-wimbledon-1305	1.000	1.000	1.000	0.000
2023-wimbledon-1306	1.000	1.000	1.000	0.000
2023-wimbledon-1307	1.000	0.000	1.000	1.000
2023-wimbledon-1308	1.000	1.000	1.000	0.000
2023-wimbledon-1309	1.000	1.000	1.000	1.000
2023-wimbledon-1310	1.000	0.000	1.000	1.000
2023-wimbledon-1311	1.000	0.000	1.000	1.000
2023-wimbledon-1312	1.000	1.000	1.000	0.000
2023-wimbledon-1313	1.000	0.000	1.000	1.000
2023-wimbledon-1314	1.000	1.000	1.000	1.000
2023-wimbledon-1315	1.000	0.000	1.000	0.000
2023-wimbledon-1316	1.000	0.000	1.000	1.000
2023-wimbledon-1401	1.000	1.000	1.000	1.000
2023-wimbledon-1402	1.000	0.000	1.000	1.000
2023-wimbledon-1403	1.000	1.000	1.000	1.000
2023-wimbledon-1404	1.000	1.000	1.000	1.000
2023-wimbledon-1405	1.000	0.000	1.000	0.000
2023-wimbledon-1406	1.000	1.000	1.000	1.000
2023-wimbledon-1407	1.000	1.000	1.000	1.000
2023-wimbledon-1408	1.000	1.000	1.000	1.000
2023-wimbledon-1501	1.000	1.000	1.000	0.000
2023-wimbledon-1502	1.000	1.000	1.000	1.000
2023-wimbledon-1503	1.000	1.000	1.000	1.000
2023-wimbledon-1504	1.000	1.000	1.000	0.000
2023-wimbledon-1601	1.000	0.000	1.000	1.000
2023-wimbledon-1602	1.000	1.000	1.000	1.000
2023-wimbledon-1701	1.000	0.000	1.000	0.000

P1 momentumisRand, p1 turning pointsisRand, p2 momentumisRand and p2 turning pointsisRand broken down by Match Id. Color shows p1 momentumisRand, p1 turning pointsisRand, p2 momentumisRand and p2 turning pointsisRand.

Measure Values

0.000 1.000

Statistically analyzing the results of Table (3), we obtain the results shown in Figure (6). Based on these data, we can conclude the following: if the credibility of considering the

change in momentum as a non-random event is set to 100% and the turning point of momentum is considered to be similarly non-random, the credibility that the change in momentum of a player located at position p1 exhibits non-randomness in all the analyzed matches is 54.83%, whereas that of a player located at position p2 is 61.29%. We determined that an event is plausible when its confidence level exceeds 50%. Based on this criterion, we conclude that both the change in momentum and its turning point are non-random events, which means that neither the player's success nor the momentum shift in the match occurred randomly.

INDEX	p1_momentumisRand	p2_momentumisRand	p1_turning_pointsisRand	p2_turning_pointsisRand
25%	1.00	1.00	0.00	0.00
50%	1.00	1.00	1.00	1.00
75%	1.00	1.00	1.00	1.00
count	31.00	31.00	31.00	31.00
mean	1.00	1.00	0.65	0.68
std	0.00	0.00	0.49	0.48

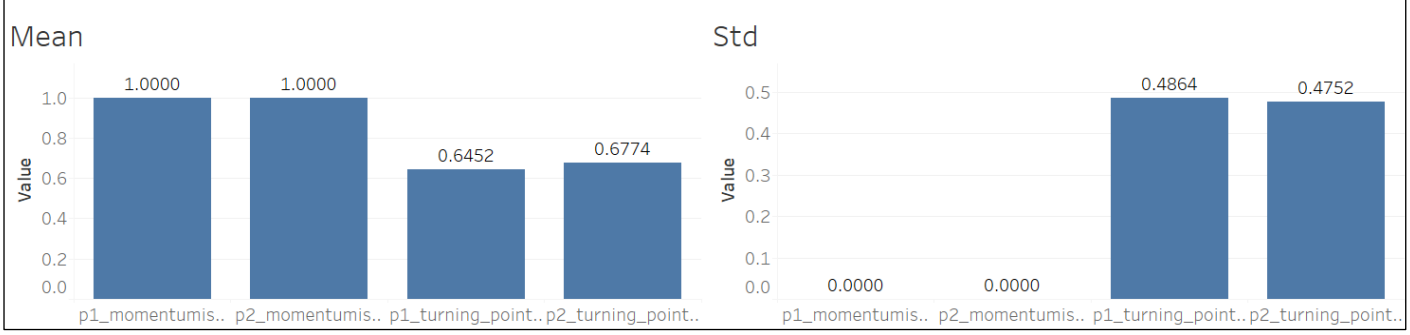

Figure 6: Statistics on the Results of Run Tests

6 Task 3: Machine Learning Models and SHAP Model

6.1 XGBoost Regression Model

To accurately predict when the momentum of a match will shift from one player to the other, four machine learning models were selected as suitable for the prediction task: linear regression, GBDT regression, neural networks, and the XGBoost regression model. These models were considered suitable for the task because they can effectively handle regression, i.e., predicting numerical outputs, which is crucial for identifying moments of shifting momentum in a match. Each model has the ability to handle specific data features to learn and predict momentum shifts from specific metrics in the match.

In Section 4, we process and normalize all raw data with missing values and compute all momentum values for all matches. In this section, we construct two key data structures: a numerical matrix X, which contains all the relevant metrics; and a first-order difference sequence of the momentum time series Y. Taking X and Y as the input parameters for the model training, their formulas are expressed as follows:

$$\begin{cases} X = [f_{ac_1}, f_{ac_2}, \dots, f_{ac_n}] \\ Y = [d_1, d_2, \dots, d_{n-1}] \end{cases} \tag{13}$$

The performance of the model is evaluated by the R^2（R-squared）score, where a value

closer to 1 indicates that the model is more capable of interpreting the data. The results obtained after model training are as follows:

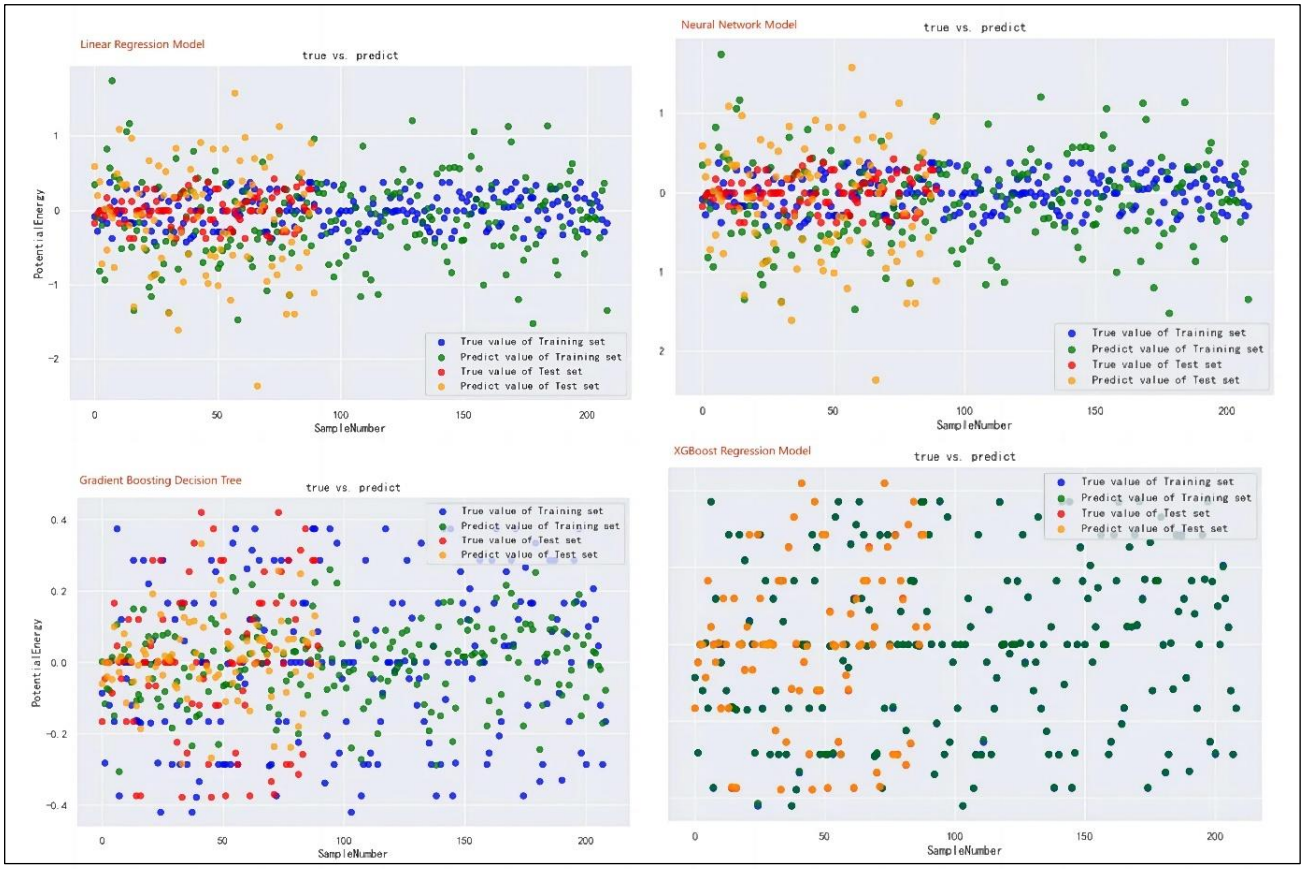

Figure 7: Scatterplot of Training Results for Four Machine Learning Models Applicable to the Task of Prediction

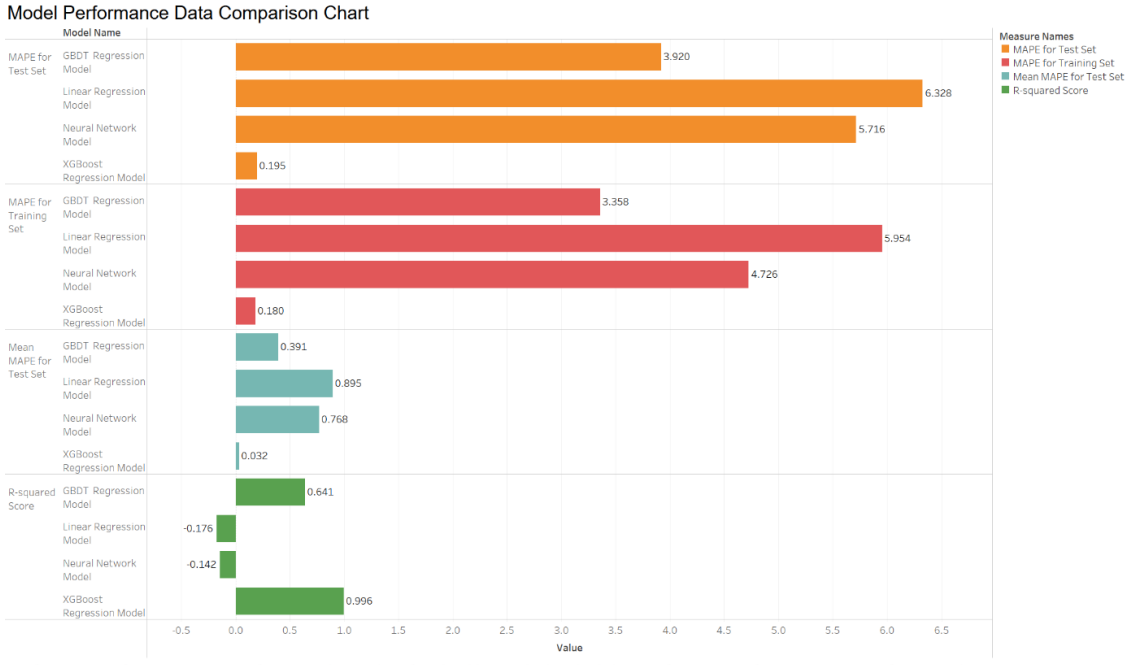

Figure 8: Error and R2_Score Bars of Training Results for Four Machine Learning Models

After a comparative analysis of the four models, the XGBoost regression model demonstrated the best performance based on the R^2 score. XGBoost outperforms the other models mainly due to its ability to efficiently handle sparse data, its built-in regularization mechanism to reduce overfitting, its parallel computation to enhance training efficiency, and its ability to be highly customized to fit specific problems. These features enable XGBoost to provide accurate predictions when dealing with complex data structures and large-scale datasets, thus effectively solving the problem of this task, i.e., accurately predicting the moments of shifts in the momentum of a match[6].

6.2 SHAP Model

After confirming the use of the XGBoost regression model to predict momentum shift moments, we used the SHAP model to investigate the contribution of each factor in the XGBoost regression model predictions[7].

To construct the SHAP model, first, we preprocess the dataset to determine the category labels of each relevant factor in X, which are used to group the relevant factors. Next, we use the SHAP interpreter to calculate the SHAP value of each factor in the input matrix F of the XGBoost model with respect to the model prediction, which can be expressed as:

$$\phi_i = \sum_{S \subseteq X \setminus \{i\}} \frac{|S|!(|F|-|F|-1)!}{|F|!} \left[f_{\text{xgb}}(S \cup \{i\}) - f_{\text{xgb}}(S) \right] \tag{14}$$

Here, ϕ_i denotes the SHAP value of feature fac_i for the model prediction, S is the subset of features without feature fac_i, $|S|$ is the number of elements in the set, $|F|$ is the number of all features, $f_{xgb}\ (S \cup \{i\})$ and $f_{xgb}\ (S)$ are the model predictions of the XGBoost with and without feature i, respectively, $\frac{|S|!(|F|-|F|-1)!}{|F|!}$ is the number of features i added to the subset S in all possible orders as a proportion of all possible orders.

We then took the absolute values of the SHAP values for each factor and calculated the average of these absolute values.

$$\overline{|\phi_i|} = \frac{1}{m} \sum_{j=1}^{m} |\phi_{ij}| \tag{15}$$

Here, ϕ_{ij} denotes the SHAP value of feature fac_i on sample j, m is the number of samples, and $\overline{|\phi_i|}$ is the average of the absolute values of SHAP values of feature fac_i over all samples.

Finally, to understand the overall impact of the different categories of factors on the model XGboost model predictions, we categorized the factors by their labels in order to summarize the SHAP values for the different categories. The following SHAP charts demonstrate the impact of each factor on the model predictions.

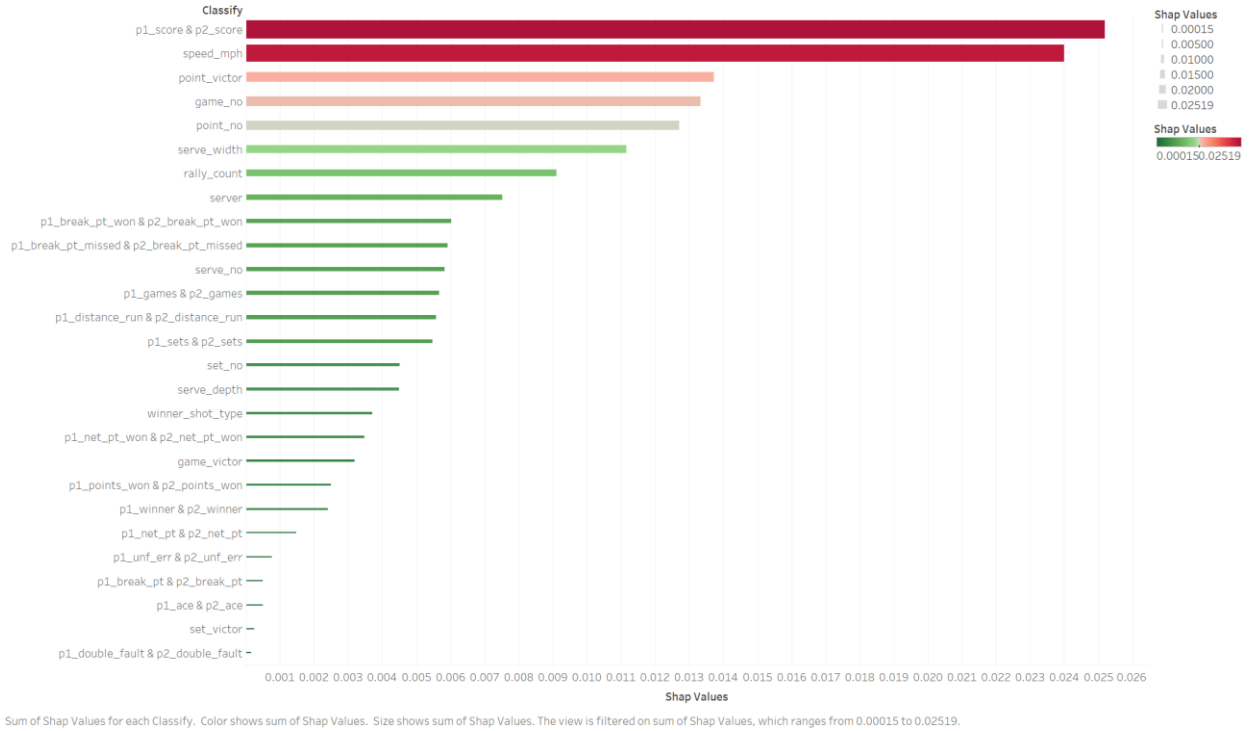

Figure 9: Bar Chart of the Effect of Each Factor on the Model's Predictions

We summarize all SHAP values in a table as shown below.

Table 4: SHAP Values for the Effect of Each Factor on Model Predictions

Classify (group)	
p1_score & p2_score	0.02519
speed_mph	0.02399
point_victor	0.01372
game_no	0.01334
point_no	0.01270
serve_width	0.01116
rally_count	0.00911
server	0.00753
p1_break_pt_won & p2_break_pt_won	0.00603
p1_break_pt_missed & p2_break_pt_..	0.00592
serve_no	0.00583
p1_games & p2_games	0.00567
p1_distance_run & p2_distance_run	0.00558
p1_sets & p2_sets	0.00547
set_no	0.00450
serve_depth	0.00450
winner_shot_type	0.00371
p1_net_pt_won & p2_net_pt_won	0.00348
game_victor	0.00320
p1_points_won & p2_points_won	0.00249
p1_winner & p2_winner	0.00241
p1_net_pt & p2_net_pt	0.00147
p1_unf_err & p2_unf_err	0.00077
p1_break_pt & p2_break_pt	0.00050
p1_ace & p2_ace	0.00049
set_victor	0.00026
p1_double_fault & p2_double_fault	0.00015

Sum of Shap Values broken down by Classify (group). Color shows sum of Shap Values. The view is filtered on sum of Shap Values, which ranges from 0.00015 to 0.02519.

Shap Values

0.000150.02519

6.3 Results of the Model

Using the established XGBoost regression model, we can predict the shift in momentum values during a match. Through the analysis of the SHAP model, we found that the top six factors that have the greatest influence on the shift of momentum in a match are: score (the current score of each player), speed_mph (the speed of hitting the ball), point_vector (whether or not the serve is broken), game_no (the number of games that have been played in the match), point_no (the current total number of points scored by each player), serve_width (the width of the serve).

The following swarm diagram shows the degree of influence of each factor. Of these, each player's current score is the most relevant factor.

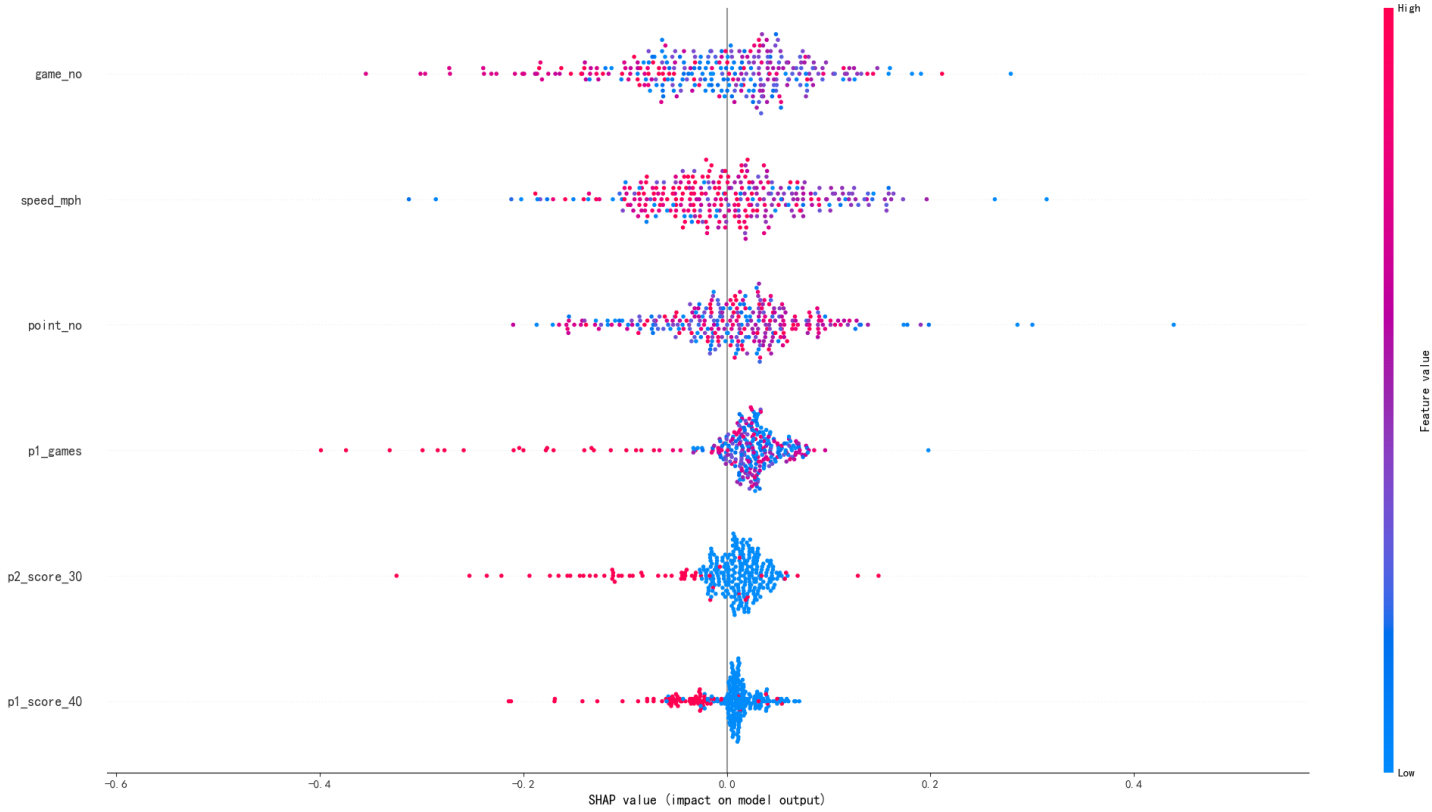

Figure 10: the Swarm Plot of the Influence of the Top Six Factors that have the Greatest Impact on Shifts in Momentum during a Match

6.4 Advice for Players Entering a New Match with Different Players

Considering the above results for the factor analysis, we make recommendations for players in several ways.

In a match, the player's current score (score) is an important basis for measuring their performance in the game, it is the highest correlation with the momentum value conversion in the game, the score will directly affect the player's confidence and decision-making. Therefore, players should try their best to gain a scoring advantage. The number of games played (game_no) reflects the progress of the game, so during the game we recommend that players adjust their strategy according to the number of games played, for example, focusing more on defense towards the end of the game. The current point_no reflects the overall situation of the

match, so players need to adjust their mindset according to the point_no, stay calm and not be swayed by temporary wins and losses. Speed_mph is an important indicator of a player's attacking power, which directly affects the pace of the game and the pressure on the opponent's defense. Players need to adjust their strategy according to their own speed and the opponent's reaction. At the same time, maintaining a stable speed will help to stabilize the player's mindset. Whether or not to break serve successfully (point_vector) is a key point in the match. Successful breaks of serve will not only increase one's own score, but also put pressure on the opponent and affect his/her mindset. Players need to pay close attention to the chances of breaks of serve, and at the same time, they also need to prevent themselves from being broken. Serve_width is an important part of the serve strategy, which can affect the start of the match and the pace of the rest of the match. By adjusting the width of the serve, the player can try to break the opponent's defensive balance and create an advantage.

In conclusion, we recommend players to consider the above factors before entering a new match against a different player and adjust their mindset and strategy according to the responses we have provided in order to better cope with the match and improve their competitiveness.

7 Task 4: Evaluation of the Effectiveness of Model and Analysis of the Degree of Generalization

7.1 Evaluation of Model Effects

In Section 6.1, we finalized the XGBoost regression model for predicting and explaining the data. Therefore, we tested the XGBoost regression model using the validation set and found that the average absolute error was around 6.7%, which means that the model worked excellently.

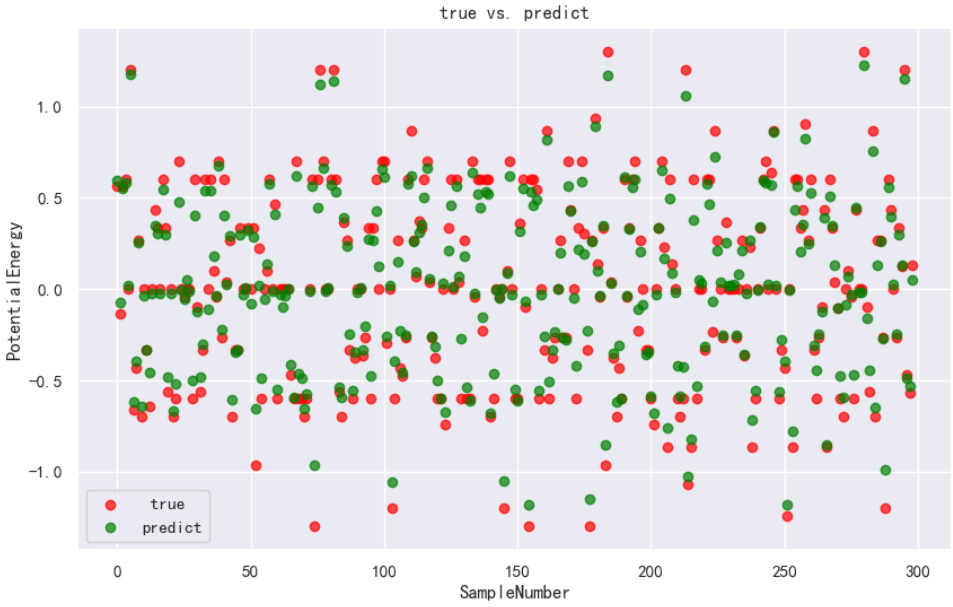

Figure 11: Results of the Tests on the XGBoost Model

7.2 Considerations for Inclusion in Future Models

In Section 6.2, the SHAP model we built yields the contribution of various factors to momentum. Based on the contribution degrees, we summarized the six factors that are most relevant to momentum shifts when the model underperforms. Therefore, for models that may underperform in the future, by deriving these six factors, we conceptualized the following six metrics[8]:

- **Score Difference**: This is a derivative of the current score (score), which reflects the difference in scores between players.
- **Game Progress Percentage**: This is a derivative of game_no, which reflects the progress of the game.
- **Point Rate**: This is an indicator of the total number of points scored by both players (point_no), which reflects how efficiently the players are scoring.
- **Average Speed_mph**: This is a derivative of speed_mph, which reflects a player's average bat speed.
- **Break Point Success Rate**: This is a derivative of point_vector, which reflects the probability of a player breaking successfully.
- **Average Serve Width**: This is a derivative of serve_width, which reflects the player's average serve width.

These 6 derived factors were added to the original factors as new input parameters for the XGBoost model and trained again. At the same time, we calculated the SHAP values of the 6 derived factors.

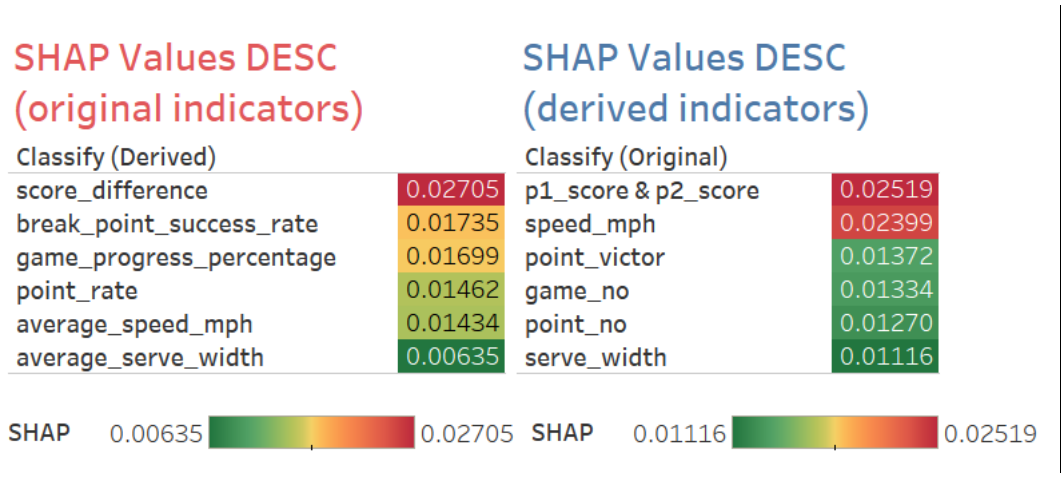

Figure 12: SHAP Values for 6 Factors and Their Derivatives

After completing the training of the new XGBoost model, we selected the validation set to validate the new model and found that the accuracy of its model was improved, with an average absolute error of 0.0428. Since our model was trained based on a large amount of match data, each factor was diluted, and the accuracy of the model was improved with the addition of the above 6 derived factors. Thus, we believe that for a particular match, the accuracy of the model can also be improved after adding the above 6 derived factors. Therefore, the above 6 derived factors can be included in the future model to improve the accuracy when the model performs poorly.

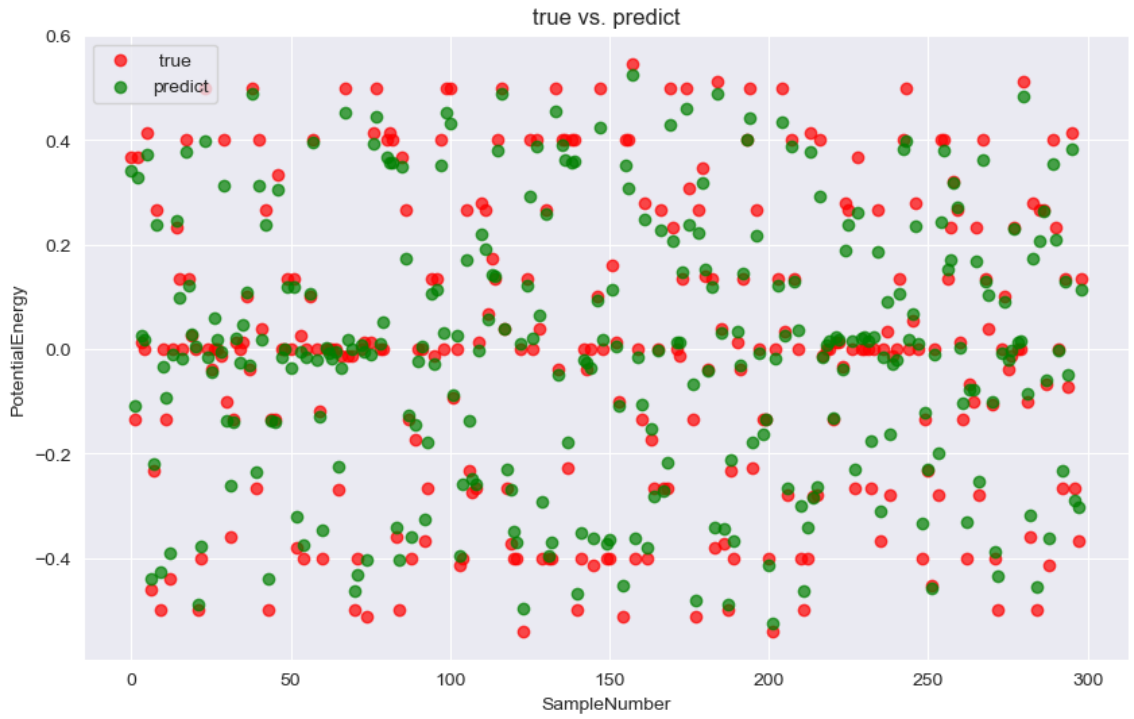

Figure 13: Test Results of the Retrained XGBoost Model

7.3 Ability of Models to Generalize

Table tennis, badminton and even other games where balls are hit with rackets or similar tools have many similarities with tennis. In order to test the degree of generalization of our XGBoost prediction model, we consider applying the model to table tennis matches. In table tennis, the outcome of the match is also affected by various factors, and since the factors affecting table tennis are not exactly the same as those affecting tennis, we consider using several factors common to both table tennis and tennis for prediction, and the factors used are as follows:

Table 5: Factors Common to Tennis and Table Tennis

Factors
p1_sets & p2_sets
p1_score & p2_score
p1_games & p2_games
p1_points_won & p2_points_won
p1_winner & p2_winner
p1_net_pt & p2_net_pt
set_victor
point_victor
game_no
point_no
rally_count
server

We selected six datasets from the men's singles matches of the Asian Table Tennis Championships in 2009, 2012, 2013, 2015, 2017 and 2019 as the validation sets to validate the model and obtained the effect of "momentum" prediction of table tennis matches as shown in the figure below.

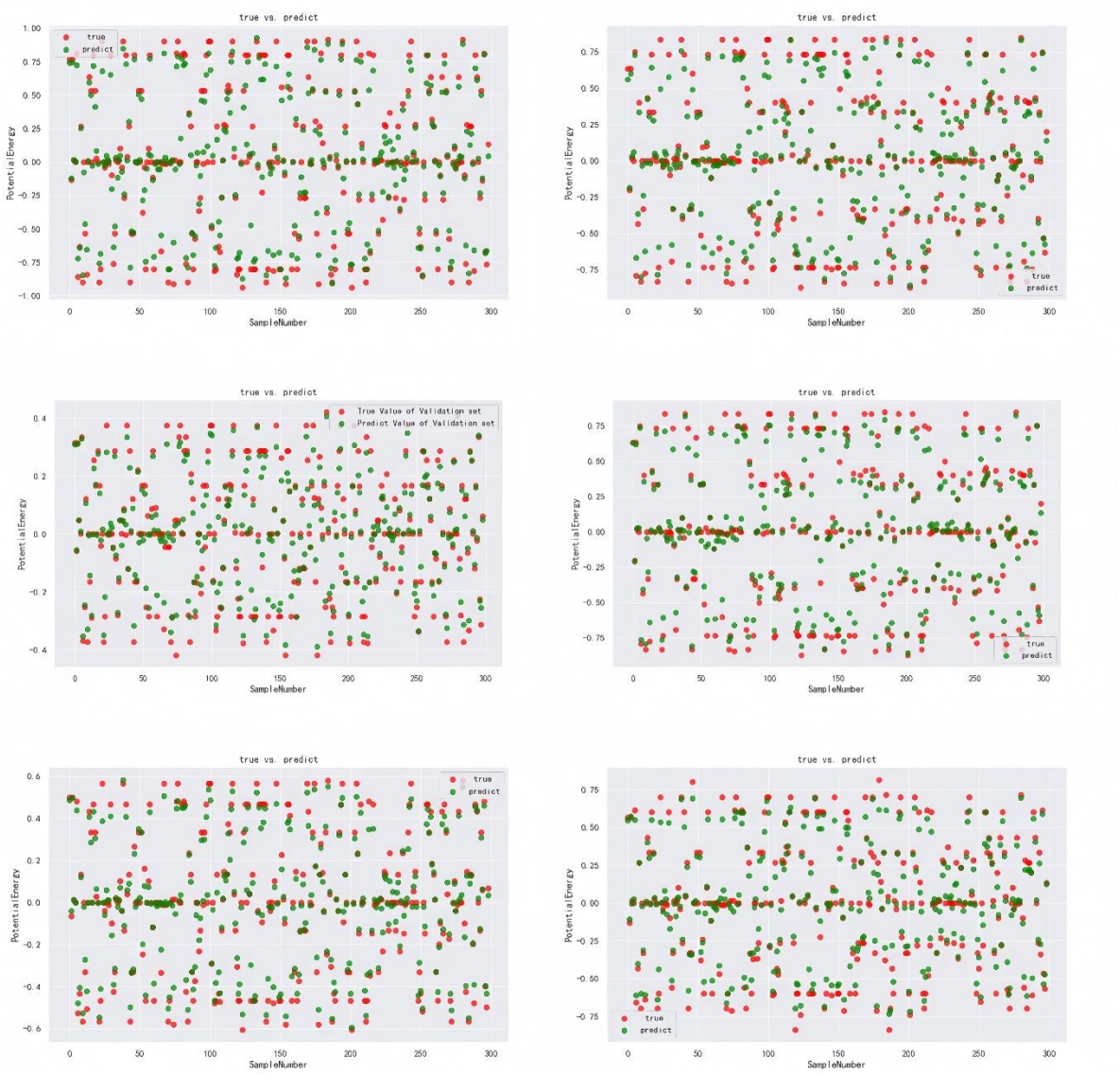

Figure 14: Effect of Momentum Prediction on 6 Table Tennis Datasets

After the statistics, the prediction accuracy of our model reached 73.41% on average accuracy and 84.59% on average R2_Score, which indicates that it is effective for momentum prediction of table tennis matches. Thus, although we did not test other ball sports, our model is effective for momentum prediction of table tennis matches and the model generalizes well, so we can infer that the model is also effective for momentum prediction of other ball sports.

8 Sensitivity Analysis

In optimizing the XGBoost prediction model for Task 3, we performed a sensitivity analysis to assess the effect of feature weight adjustments in the model on prediction accuracy. We used the model's coefficient of determination, R^2, to indicate the model's fitting effectiveness.

In this analysis, we increased or decreased the weights of a set of features individually, changing the weights of only one feature at a time while keeping the weights of other features unchanged. The weights were adjusted by 10% to observe the change in the R^2 score of the model.

Based on the provided data, we observe that the R^2 score of the model does not change much and remains higher than 90% even after weight adjustment, which indicates that our XGBoost model is insensitive to changes in feature weights and has relatively stable predictive performance. This higher stability makes our model more reliable when applied to new data or for future prediction.

Table 6: XGBoost Model Sensitivity Analysis Results

Processing of weights	Feature	R2 Score of the model
Increase 10%	p1_score & p2_score	92.871
	point_num	94.467
	point_victor	93.689
Decrease 10%	game_no	94.135
	serve_width	96.178
	speed_mph	96.059

R2 Score of the model broken down by Processing of weights and Feature.

9 Strengths and Weaknesses

9.1 Strengths

- **Multi-dimensional analysis**: By combining momentum computing, time series analysis, turning point detection, and machine learning prediction, we provide a comprehensive analytical framework to capture and interpret the dynamics of a game from multiple perspectives.
- **Real-time prediction capabilities**: We use ARIMA models and XGBoost regression models to predict momentum changes, helping coaches and players analyze game momentum and predict future changes in real time.
- **Factor Impact In-depth analysis**: Through SHAP model analysis, our research reveals the main factors that influence momentum change, providing insight into understanding key turning points in the game.
- **Model validation and application**: We not only verified the validity of the model in tennis, but also applied the model to table tennis, showing the wide applicability of the model in different sports.

9.2 Weaknesses

- **Data dependency**: The effectiveness of the model is highly dependent on high quality and sufficient data. Challenges in data collection and processing can affect the accuracy and reliability of models.
- **Risk of overfitting**: Although XGBoost models could handle overfitting, in highly complex models, there is still a risk of overfitting training data and resulting in reduced generalization.

10 MEMO

From: Team 2400228, ICM 2024
To: International Tennis Coaches Council
Date: February 5, 2024
Subject: Quantitative Analysis of Tennis Momentum and Strategic Recommendations

Dear International Tennis Coaches Council, we have the honor to present to you our research findings on the momentum of the tennis game and the summarized recommendations offered to coaches and players.

In our study of momentum, we built a momentum calculation model using seven key metrics designed to capture the flow of the game and identify dominant players in a given period. We also used a random forest model to calculate the weights of the metrics, establish a momentum calculation formula, and predicted momentum changes using an ARIMA model. In addition, we verified the non-randomness of momentum changes using the CUMSUM algorithm and tour tests. By training four different models, we selected the XGBoost regression model to predict the momentum transition moments between players and analyzed the influence degree of each indicator using the SHAP model. Finally, we assessed the effectiveness of the XGBoost model and outlined directions for model improvement.

Our empirical research and data analysis all point to a common conclusion: momentum has a non-negligible impact on in-game wins and players' scoring streaks. With this in mind, we make the following application recommendations to coaching teams:

- Analyze players' momentum characteristics: Coaches can use the model to understand in detail the ups and downs of each player's momentum in a game, and accurately identify their strengths and weaknesses. The model reveals the key factors affecting the momentum, such as psychological stress management, physical condition, and game tempo, etc., which can help coaches develop personalized training plans for players. These drills can be targeted to improve players' ability to seize opportunities at critical moments, thus effectively improving their skills in controlling the momentum of the game.
- Predicting game momentum shifts: By utilizing the model's prediction function, coaches can more accurately foresee the momentum shifts in the game, and thus formulate tactical responses. For example, when the model predicts that the opposing player is about to experience a rise in momentum, the coach can instruct the players to adjust their serving strategy or change the pace of hitting the ball to disrupt the opposing player's pace and inhibit the growth of their momentum.
- Coping with the psychological challenges of the game: Every point scored in a game and the outcome of each set may have an impact on the psychological state of the players. Coaches should encourage players to develop mental toughness and teach them how to stay focused and calm when facing a deficit or increased pressure.

We believe that tennis coaches should emphasize the role of momentum in the game, and incorporate the predictive results of our model to tailor coaching programs for their players to help them respond more effectively to the various events that may occur during a match.

Finally, we hope that the results of these analyses will provide you with strong data to help you better utilize the concept of momentum in your daily training and match strategy development, thus improving your team's overall performance and competitiveness.

Yours Sincerely,
Team # 2400228

References

[1] Patro, S. G. O. P. A. L., & Sahu, K. K. (2015). Normalization: A preprocessing stage. arXiv preprint arXiv:1503.06462.

[2] Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. The Stata Journal, 20(1), 3-29.

[3] Wagner, B., & Cleland, K. (2023). Using autoregressive integrated moving average models for time series analysis of observational data. The BMJ, 383, p2739.

[4] Granjon, P. (2013). The CuSum algorithm - a small review. HAL, hal-00914697.

[5] Bujang, M. A., & Sapri, F. E. (2018). An application of the runs test to test for randomness of observations obtained from a clinical survey in an ordered population. The Malaysian Journal of Medical Sciences: MJMS, 25(4), 146.

[6] Abbasi, R. A., Javaid, N., Ghuman, M. N. J., Khan, Z. A., Ur Rehman, S., & Amanullah. (2019). Short term load forecasting using XGBoost. In Web, Artificial Intelligence and Network Applications: Proceedings of the Workshops of the 33rd International Conference on Advanced Information Networking and Applications (WAINA-2019) 33 (pp. 1120-1131). Springer International Publishing.

[7] Meng, Y., Yang, N., Qian, Z., & Zhang, G. (2020). What makes an online review more helpful: an interpretation framework using XGBoost and SHAP values. Journal of Theoretical and Applied Electronic Commerce Research, 16(3), 466-490.

[8] Marcílio, W. E., & Eler, D. M. (2020, November). From explanations to feature selection: assessing SHAP values as feature selection mechanism. In 2020 33rd SIBGRAPI conference on Graphics, Patterns and Images (SIBGRAPI) (pp. 340-347). Ieee.