Problem Chosen	2024 MCM/ICM Summary Sheet	Team Control Number
C		2427586

The MCM Thesis of Team 2427586

Summary

In tennis matches, there seems to be a momentum effect where winning or losing streaks can significantly influence the outcome of subsequent games. Using the 2023 Wimbledon matches data, we examined the data for each point, considering player on-court performance metrics such as running distance, serving speed, etc. **This approach utilized cross-sectional data for each point, preserving the "performance" indicators without being contaminated by time-series effects.** Using the Analytic Hierarchy Process and machine learning, we built an evaluation model to construct a player performance assessment system. **This system determined the relative advantage of Player 1 at each moment, achieving a high model fitting accuracy of 95%**.

Next, we investigated the impact of "momentum" on winning points. **Momentum was defined as the weighted average of "player 1's winning ratio in the last five points" and "player 1's consecutive winning points"** (if consecutively losing x points, then the value is -x). Overall, **momentum had a significant positive impact** on whether a player could win the next point, with an information coefficient of 17%. Hypothesis testing confirmed that swings in play and runs of success by one player were not randomly distributed, and momentum was a key factor influencing scoring **(t-value of 62, p-value close to 0)**.

We then studied indicators that could help determine when the flow of play is about to change from favoring one player to the other. We attempted to build a more universal model to predict the future trend of the game using existing data but failed after trying 9 models.

Instead, we created a "swing" indicator defined by the change of flow of play. We constructed tendency factors. Through calculating information coefficients and hypothesis testing, we found that the most effective indicators were **momentum, the ratio of scoring and the number of consecutive misses**. This suggests that players and coaches need to carefully consider the "momentum" effect during the match and be prepared for different psychological, tactical, and other adjustments based on different momentum situations.

Finally, we tested the predictive performance of our model on two other types of tennis matches. We chose a French Open from a decade ago and a recent US Open. In these match datasets, **we confirmed that momentum, the ratio of scoring, the number of consecutive misses (in the last 10 points) were the three most effective indicators for predicting turning points in the game**. To further improve prediction of swings, we incorporated all factors into consideration. We conducted feature selection using Principal Component Analysis (PCA) and correlation analysis. By removing Momentum, our model achieved further improvement, with **the accuracy of predicting turning points reaching around 86%**.

Memo

In tennis matches, there appears to be a momentum effect, wherein a player who has won several matches is highly likely to continue winning the next match, and vice versa. **We confirmed that momentum has a highly significant positive impact on whether a player can win the next point, and it is a key factor influencing scoring.**

We found that the most effective indicators for change of flow in matches are **momentum, the ratio of scoring (in the last 5 points), and the number of consecutive misses.** Additionally, the effect of momentum reversal is highly significant, **indicating that players who have consecutively won many games are more likely to lose the next point.** This suggests that players and coaches need to handle the "momentum" effect with caution and **make different psychological, tactical, and other preparations and adjustments for different momentum situations**. Further recommendations will be provided in subsequent memos.

For coaches and players looking to grasp and leverage the momentum effect in a match, we propose the following steps:

1.To study the impact of different events on winning points, characterize the flow of the match first to observe which player is currently performing better and to what extent.

Use AHP and TOPSIS methods to decompose player performance into three aspects: Serving Factor, Processing Factor, and Hitting Factor. Consider these ten factors as indicators of performance and use machine learning to build an evaluation model, assigning weights to the ten factors and constructing a player performance assessment system. This will allow you to determine the relative advantage of Player 1 at each moment and plot the flow of the match.

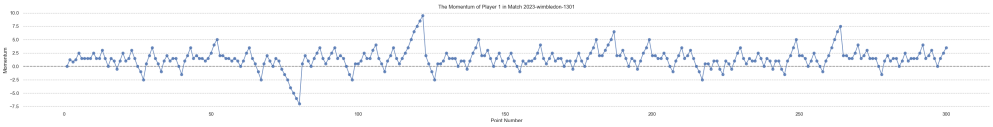

Figure 1: An example of how the Flow of Match looks

2. Characterize momentum by observing the duration and characteristics of the momentum effect for both opponents. Calculate the weighted average of "player 1's winning ratio in the last five points" and "player 1's consecutive winning points" (if consecutively losing x points, then this indicator is -x). Since the psychological impact of the latter is significantly greater than the former, assign twice the weight to the change rate of the latter compared to the change rate of the former.

Observing the momentum change chart reveals that, **on average, tennis players often lose the momentum effect after winning 4 consecutive points,** indicating that 4 points are a crucial turning point. Players who have won 4 consecutive points need to remain calm and focused, avoiding distractions in the next game. However, this tendency varies from person to person. Therefore, **it is necessary to characterize the duration and fea-**

tures of the momentum effect for both your own team and the opposing team, formulate match strategies tailored to these characteristics, and respond flexibly to momentum changes in the match.

3. Characterize turning points and study the impact of different events on turning points in the match.

Define turning points as follows: when the game shifts from favoring player 2 to player 1, the turning point factor is 1; conversely, the turning point factor is -1. **Turning points characterize the moment when a player "loses" the momentum effect**. After constructing the turning point indicator, categorize all player wins/losses, receiving serves, missing serves, hitting aces, hitting winners, net plays, etc., into "ratio" and "consecutive" types. For the former, calculate "the ratio of times player 1 made such a move in the last 10 points," and for the latter, calculate "the consecutive number of times player 1 made such a move." **By calculating information coefficients and conducting significance tests on regression equation coefficients, identify the three most significant indicators**. For regular matches, the three most effective indicators are **momentum (positive effect), the ratio of scoring (in the last 5 points) (negative effect), and the number of consecutive misses (negative effect)**.

This suggests that players **who perform well in the last five points need to be wary of the reversal of the momentum effect**; they need to stay calm, focused, and adjust tactics flexibly. **Players with consecutive misses in the last ten points need to be cautious about the continuation of the momentum effect**; they need to actively respond, identify weaknesses in the opponent's tactics, and improve their own tactics to reverse the momentum in the next game.

Indicator	Information Coefficient
Scoring_Ratio	-22.05%
break_pt_won_Ratio	-11.05%
Momentum	7.40%

Table 1: The Information Coefficients of the Top 3 Most Important Indicators

Coaches and players can use SVM Classifier to **predict turning points based on these factors** and raise prediction accuracy using feature selecting. Once turning points in the match are predicted, strategies can be designed and implemented accordingly.

Coaches and players should consider the individual characteristics of each player. The best approach is to conduct the above turning point analysis for both your own team and the opposing team in past matches. **The indicators for predicting turning points may vary for different players**; for example, for some players, the momentum continuation effect of "hitting untouchable shots" may be greater than the consecutive number of consecutive misses in the last ten points. **Additionally, it's important to consider the impact of different indicators on matches in different directions. Based on this, combine and plan a comprehensive, three-dimensional, and integrated strategy to effectively respond to the rapidly changing tennis court.**

Contents

1 Introduction

In tennis matches, there seems to be a momentum effect where winning or losing streaks can significantly influence the outcome of subsequent games. Momentum effect refers to the likelihood that a player who has won several matches is more likely to continue winning the next match, while a player who has lost several matches is more likely to continue losing.

Momentum refers to the psychological advantage that one player gains over another by winning a few consecutive points or games. Momentum is often discussed in matches, and its impact on game outcomes is debated.[1] **The aim of our study is to build a model that captures the swings of play during tennis matches, identifies which player is performing better and quantifies the extent of their advantage.** Additionally, the study seeks to assess the influence of momentum in a match. Furthermore, our research intends to develop a model that can predict the flow of the match and identify relevant factors, thus providing valuable insights for coaches and players.

As regards to the game of tennis itself, it is played on a rectangular court divided by a net. Here are the fundamental rules of tennis:

Scoring System: A tennis match is typically divided into sets, and each set consists of games. To win a game, a player or team must score at least four points and have a two-point advantage over their opponent(s). The scoring sequence is 15, 30, 40, and then the game. If the score reaches 40-40, it is called deuce, and a player or team must win two consecutive points to win the game. A set is won by the first player or team to reach six games with a two-game advantage. If the set reaches a 6-6 tie, a tiebreaker is usually played to determine the set winner.

Service: Players take turns serving, and a serve must land within the opponent's service box on the diagonal side of the court. The server gets two chances (first and second serve). If the ball goes out on both attempts, it is a double fault, and the opponent wins the point.

Tiebreaker: In the event of a tied set (6-6), a tiebreaker is played to determine the set winner. The first player or team to reach seven points with a two-point advantage wins the tiebreaker.

2 Data Description

We primarily used the data from each point of the 2023 Wimbledon tennis match for modeling and analysis. This original dataset has 7285 observations. Every observation is based on every point from all Wimbledon 2023 men's matches after the first two rounds, encompassing players and 46 features like scores, speed, distance, servers, errors, break points, net points and so on. In the final section, we employed the data from each point of the 2013 French Open and the 2022 US Open tennis matches as a validation set to test the generalizability of our model. Choosing these two matches for the validation dataset takes into account variations in court types, match durations, clothing regulations, match atmospheres, and other differences, providing stronger support for the robustness of our model.

Among all the data, there exists missing data in four features *speed_mph*, *serve_width*, *serve_depth* and *return_depth*, as shown in Table 2.

	Unique Values	Null Values
speed_mph	66	752
serve_width	5	54
serve_depth	2	54
return_depth	2	1309

Table 2: Missing Data Statistics

3 Problem One

3.1 Definition of performance

First of all, we define performance as the probability of player 1 winning or scoring relative to player 2, which ranges from 0 to 1. The better Player 1 performs, the more likely he/she is to win, and the closer the value of performance gets to 1.

We do not use the ratio of scores to assess performance because they are time-series data across games, containing factors influenced by momentum. In fact, our model is based on in-game performance features for each scoring event, using cross-sectional data.

3.2 Data Pre-processing

3.2.1 Missing Value Handling

In order to fill in the missing values in the data, **we perform data imputation using a feedforward neural network (BP neural network) to improve data completeness and quality.** The process involves the following steps:

Firstly, we clean the data and split the data into a training set (80%) and a test set (20%).

Then we build a BP neural network model which is created with one hidden layer containing 10 nodes in each layer. The neural network model is trained using the standardized input and output of the training set, and a maximum number of failed attempts is set to control the training process.

Next, the neural network model is tested using the standardized input of the test set, and predictions (test_out) are obtained.

Ultimately, in the data imputation phase, the code selects rows from data that require imputation, specifically those with -1 values. It uses the trained neural network model to predict values for these rows and stores the imputed values back into the corresponding positions in data.

3.2.2 Uniform as Numerical Variables

To facilitate the construction of our model , we convert non-numeric variables into numeric variables using assignment, such as *serve_width* and other features, with values as

shown in Table 3.

Non-numeric variables	*serve_width*						*serve_depth*			*return_depth*		
	BC	B	BW	W	C	NA	NCTL	CTL	NA	ND	D	NA
Numeric variables	1	2	3	4	5	-1	1	2	-1	1	2	-1

Table 3: Numerical Transformation of *serve_width*, *serve_depth* and *return depth*.

3.2.3 Data Summary

After data pre-processing, we selected features related to performance on the tennis court and get a summary as shown in Table 4.

We have plotted a heatmap (Figure 2) for all the features and find that there's no significant correlation between any two variables. Thus, it's reasonable to use all of these variables to construct the *Performance* indicator, **without worrying strong multicollinearity problems.**

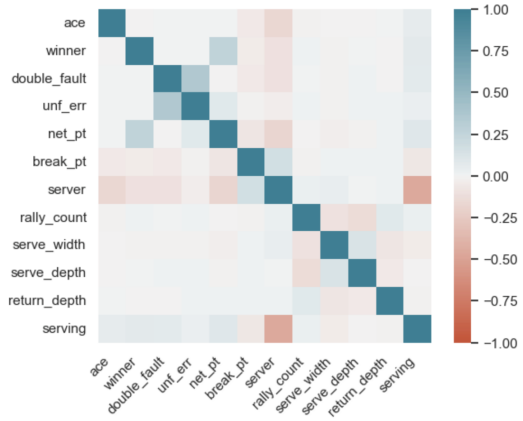

Figure 2: Heatmap Correlation

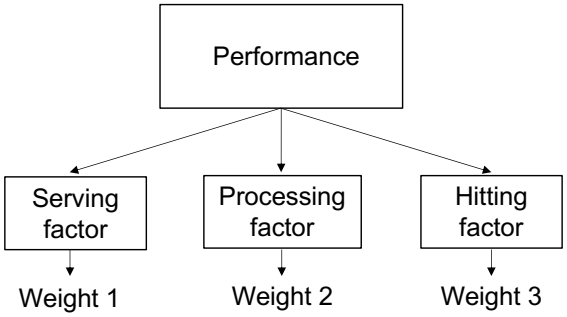

Figure 3: AHP

3.3 Methodology

We utilize the AHP method (Figure 3) for assigning weights and standardization to calculate the weights of factors. Then, we use the trained model to compute the performance of Player 1 relative to Player 2 at each point.

3.3.1 Dimensionality Reduction and Merging

We analyze the factors for dimensionality reduction and merging.

Firstly, for features 'ace' and 'winner', we have observed that when a player scores an ace point, his/her winner feature must be 1. In this case, these two features overlap. **Therefore, we have modified the definition of winner to exclude points scored through ace. Specifically, when ace=1, winner=0.** This adjustment helps to eliminate the influence of ace on feature winner and avoid high correlation between these two factors.

Statistics	Count	Mean	Std	Min	25%	50%	75%	Max
set_no	7284	2.47	1.19	1	1	2	3	5
game_no	7284	5.91	3.41	1	3	6	8	13
point_no	7284	125.87	80.29	1	59	118	182	337
p1_sets	7284	0.77	0.81	0	0	1	1	2
p2_sets	7284	0.70	0.73	0	0	1	1	2
p1_games	7284	2.50	1.85	0	1	2	4	6
p2_games	7284	2.42	1.81	0	1	2	4	6
server	7284	1.51	0.50	1	1	2	2	2
serve_no	7284	1.36	0.48	1	1	1	2	2
point_victor	7284	1.49	0.50	1	1	1	2	2
p1_points_won	7284	63.49	41.52	0	29	58	93	176
p2_points_won	7284	62.35	39.35	0	30	59	90	168
game_victor	7284	0.24	0.58	0	0	0	0	2
set_victor	7284	0.02	0.19	0	0	0	0	2
p1_ace	7284	0.05	0.21	0	0	0	0	1
p2_ace	7284	0.04	0.20	0	0	0	0	1
p1_winner	7284	0.17	0.38	0	0	0	0	1
p2_winner	7284	0.16	0.37	0	0	0	0	1
p1_double_fault	7284	0.02	0.13	0	0	0	0	1
p2_double_fault	7284	0.02	0.13	0	0	0	0	1
p1_unf_err	7284	0.13	0.33	0	0	0	0	1
p2_unf_err	7284	0.14	0.34	0	0	0	0	1
p1_net_pt	7284	0.10	0.31	0	0	0	0	1
p2_net_pt	7284	0.13	0.33	0	0	0	0	1
p1_net_pt_won	7284	0.07	0.26	0	0	0	0	1
p2_net_pt_won	7284	0.09	0.28	0	0	0	0	1
p1_break_pt	7284	0.04	0.19	0	0	0	0	1
p2_break_pt	7284	0.03	0.17	0	0	0	0	1
p1_break_pt_won	7284	0.01	0.12	0	0	0	0	1
p2_break_pt_won	7284	0.01	0.10	0	0	0	0	1
p1_break_pt_missed	7284	0.02	0.15	0	0	0	0	1
p2_break_pt_missed	7284	0.02	0.14	0	0	0	0	1
p1_distance_run	7284	14.00	13.49	0	4.98675	9.97	18.9285	148.723
p2_distance_run	7284	13.87	13.61	0	4.90175	9.782	18.4205	156.856
rally_count	7284	3.13	3.19	0	1	2	4	34
speed_mph	6532	112.41	12.86	72	103	115	123	141

Table 4: Data Summary

Secondly, we assert that the server has an impact on performance. Therefore, we have created a new variable called 'serving' to reflect this effect.

$$Serving = \begin{cases} 1, & \text{if } Player1 \quad serves, \\ -1, & \text{if } Player2 \quad serves. \end{cases}$$

Thirdly, for features 'double_fault', 'unf_err', 'net_pt' and 'break_pt', we use interaction terms to represent Player 1's performance relative to Player 2's performance.

$$Performance = \beta \cdot F + \alpha \cdot Serving + \delta \cdot F \cdot Serving$$

where F is the features such as 'double_fault', 'unf_err', 'net_pt' and 'break_pt', Serving is 1 if player 1 is serving and -1 if player 2 is serving.

This process is equivalent to a **logistic regression with fixed effects**, where a fixed effect for "whether Player 1 serves". As long as the dependent variable is probability of victory, the absolute value of this fixed effect will be the same for both sides.

In conclusion, we use features {***double_fault, unf_err, net_pt, break_pt, p1_distance_run, p2_distance_run, server, rally_count, serve_width, serve_depth, return_depth, speed_mph, ace, winner***}.

3.4 Model Building

Because the dependent variable is binary, we choose 4 models: **Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, and Support Vector Machine (SVM)** to fit. Then we evaluated them based on four criteria: precision, recall, F1-score, and support. After comparison, we find that Random Forest had the best fit, as shown in Figure 5.

3.5 Result

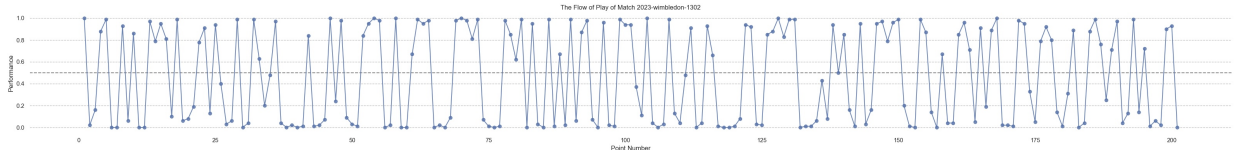

Figure 4: Performance

The flow of play as points occur is depicted in Figure 4.

Performance greater than 0.5 means that Player 1 performs better, while performance less than 0.5 indicates that Player 2 performs better.

The performance value represents the degree of good performance. **The closer performance is to 1, the greater Player 1 is. The closer it is to 0, the better Player 2's performance is.**

Random Forest has the best performance in these four models

	precision	recall	f1-score	support
-1	0.94	0.92	0.93	729
1	0.92	0.94	0.93	728
accuracy			0.93	1457
macro avg	0.93	0.93	0.93	1457
weighted avg	0.93	0.93	0.93	1457

Figure 5: Random Forest Model Evaluation

4 Problem Two

From the first part, we have the flow of the players' performance in each match. Now, the question is that does the performance of a tennis player change in random or it follows a certain pattern?

4.1 The definition of Momentum

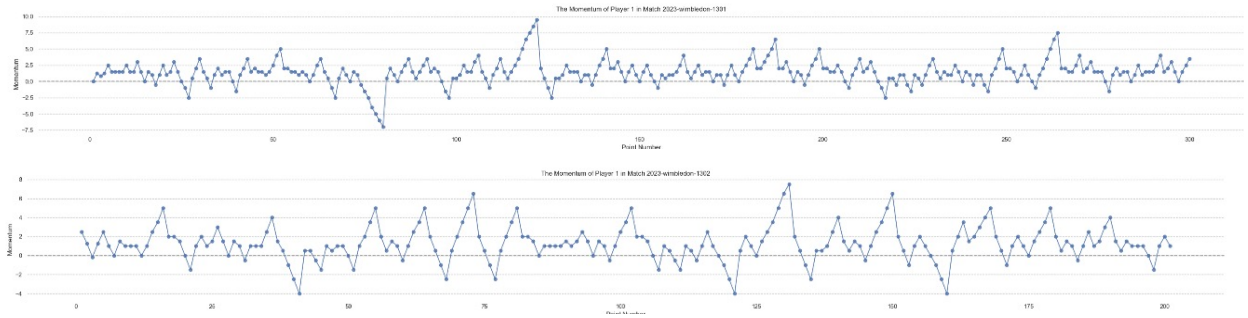

Figure 6: Momentum

Momentum means that a player benefits from a psychological and/or physiological boost.[2] Our definition is in line with theirs, t**hat is the increment of the players' performance during the tennis match.** With regards to our analysis, we will give the exact equation for Momentum in 4.2.

The charts Figure 6 depict the momentum of the first player in match-Wimbledon-1301 and match-Wimbledon-1302 respectively.

4.2 The Regular Pattern of Momentum

We investigated the relationship between the momentum of the player and two factors. The first one, Scoring_Ratio, looks into the player's performance in the recent few sets. **We define it as the winning rate of the player in the last five sets. The second factor, Consecutive_Scores, describes how many consecutive points have the player won.** These two factors comprehensively consider the player's recent performance and overall effectiveness, **so we consider them as two effective measurements of the players' overall condition.** The pseudocode of the two factors are as follows in Figure 7.

```
Procedure Scoring_Ratio (playerScores)
        totalScore = Sum(recentFiveScores)
        scoringRate = totalScore / 5
End Procedure

Procedure Consecutive_Scores(playerScores)
    maxConsecutive = 0
    currentConsecutive = 0
    For each score in playerScores
        If score > 0 Then
            currentConsecutive = currentConsecutive + 1
            If currentConsecutive > maxConsecutive Then
                maxConsecutive = currentConsecutive
            End If
        Else
            currentConsecutive = 0
        End If
    End For
End Procedure
```

Figure 7: PseudoCode

To find out whether momentum is related to the two factors, we first assigned weights to the two factors and combine them into a composite factor. **The composite score is defined as the following:**

$$Momentum = w_1 \times Scoring\_Ratio + w_1 \times Consecutive\_Scores$$

$$Scoring\_Ratio = \frac{Scores \quad of \quad Player1 \quad in \quad Last \quad 5Points}{5}$$

$$Consecutive\_Scores = \begin{cases} -Consecutive \quad Scores \quad of \quad Player2, & \text{if } Player2 \quad is \quad winning, \\ 0, & \text{if } Nobody \quad is \quad winning \quad consecutively, \\ Consecutive \quad Scores \quad of \quad Player1, & \text{if } Player1 \quad is \quad winning. \end{cases}$$

We take $w_1 = 2.5$ and $w_2 = 1$. This is because, subjectively speaking, the momentum psychological effect of Consecutive_Scores is greater. In practical terms, for each scored point, Scoring_Ratio changes by 0.2, while Consecutive_Scores changes by 1. **Therefore, after weighting, Scoring_Ratio changes by 0.5, precisely half of the change rate of Consecutive_Scores.**

4.3 Scatter Plot of Momentum and Performance

Firstly, **we observe a scatter plot of momentum and performance**, and it is evident that there is a strong positive correlation between momentum and performance.(Figure 8 and Figure 9)

4.4 Information Coefficients

Next, **we calculate the information coefficient between the momentum factor and the outcome of the next winning point.**

Information coefficient (IC) is a statistical measure used in finance and investment to assess the predictive power of a model or investment strategy. It quantifies the degree of correlation between a certain contributing factor x and the actual future y. The information coefficient is expressed as a value between -1 and 1, where 1 indicates a perfect positive correlation, 0 indicates no correlation (the model's predictions are no better than random chance), and -1 indicates a perfect negative correlation. **Normally, an IC value of above 10% is enough to claim that the factor is overwhelmingly significant.**

The IC of Momentum is about 61%, indicating significant impact of momentum on winning a game.

4.5 Hypothesis Testing

Last but not least, we examine the significance of momentum by using Hypothesis Testing. The basic idea is to construct a regression equation of performance on momentum to test whether the regression coefficient (β) is equal to zero. If β equals zero, it indicates that momentum has no impact on performance. If significantly different from zero, it

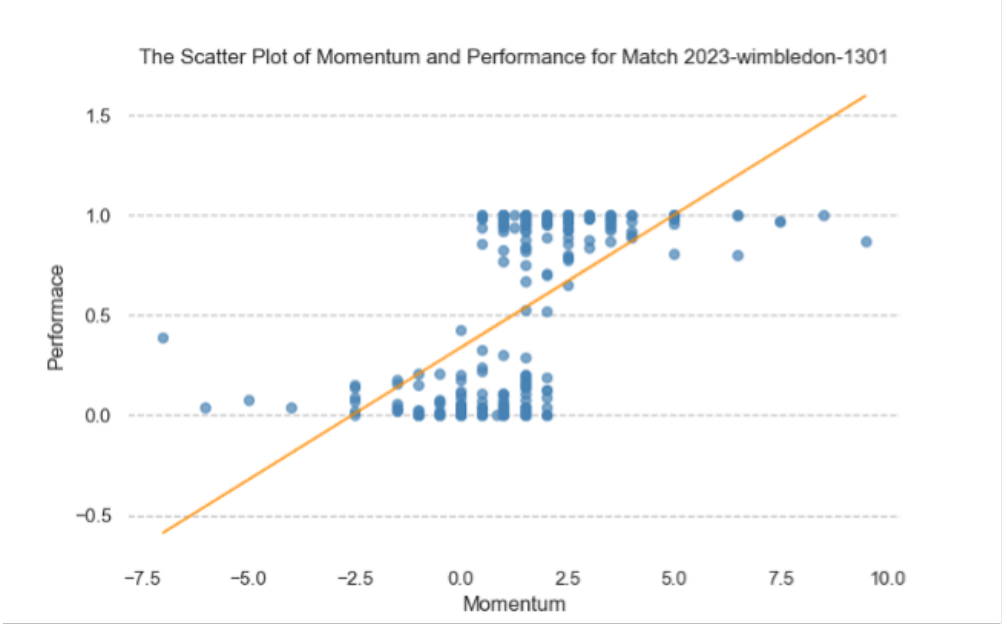

Figure 8: Relationship between Momentum and Performance

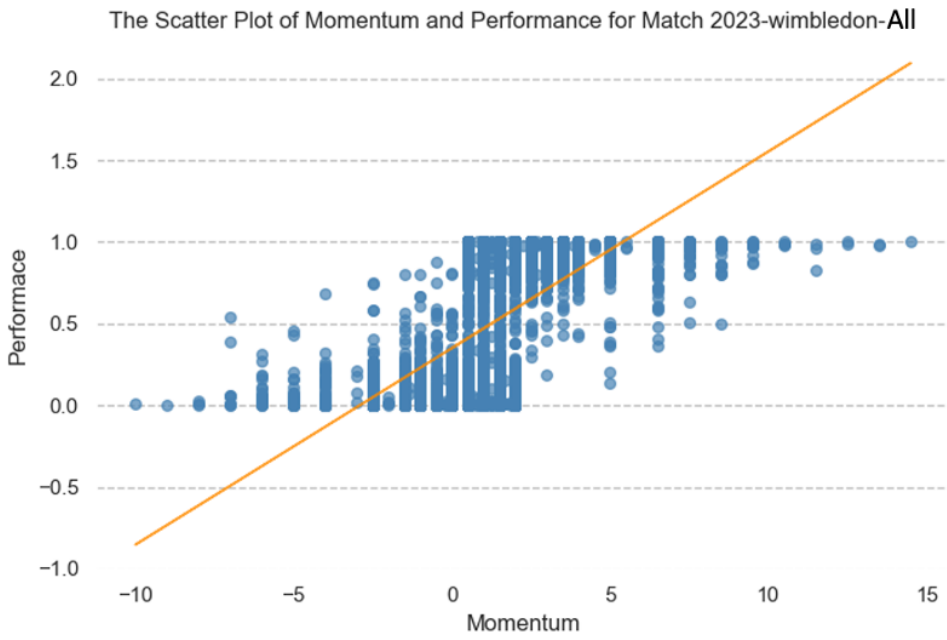

Figure 9: Relationship between Momentum and Performance

suggests that momentum affects performance. If the regression coefficient is positive, it indicates a significant positive effect of momentum; otherwise, it signifies a significant negative effect.

Null Hypothesis (H0): Momentum has no effect on the flow of the match. Alternative Hypothesis (H1): At least one of these factors has some significant effect on the flow of the match.

We use the data to fit a linear regression model. The model equation is of the form:

$$Performance = \beta \cdot Momentum + \epsilon,$$

where β is the regression coefficients and F is Momentum. Then we conduct a hypothesis test for the regression coefficient to determine whether β is significantly different from 0.

The outcoming result Figure 10 shown below displays that momentum is highly correlated with the composite factor.

It means that if a player performs well in several sets, chances are that he would play better in the following sets. Indeed, momentum does impact.

5 Problem Three

5.1 Swing Definition

We define swing by consecutive scores. **There is a swing if a player beats his opponent's consecutive scores.** For example, if player1 breaks player2's consecutive scores, the value

OLS Regression Results

Dep. Variable:	point_victor	R-squared:	0.352
Model:	OLS	Adj. R-squared:	0.352
Method:	Least Squares	F-statistic:	3957.
Date:	Sun, 04 Feb 2024	Prob (F-statistic):	0.00
Time:	20:03:06	Log-Likelihood:	-3704.5
No. Observations:	7284	AIC:	7413.
Df Residuals:	7282	BIC:	7427.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.3284	0.006	59.359	0.000	0.318	0.339
Momentum	0.1365	0.002	62.905	0.000	0.132	0.141

Omnibus:	48829.059	Durbin-Watson:	2.111
Prob(Omnibus):	0.000	Jarque-Bera (JB):	657.847
Skew:	-0.022	Prob(JB):	1.41e-143
Kurtosis:	1.528	Cond. No.	3.13

Figure 10: OLS Regression Result

of swing is noted as 1. Vice versa, if player2 breaks player1's consecutive scores, the value of swing is noted as -1. We show the results as follows in Figure 11. The orange marker shows the turning points.

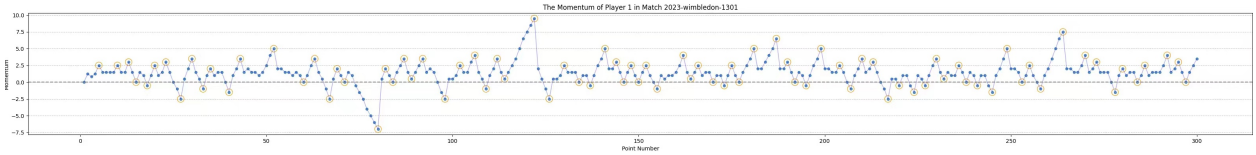

Figure 11: Swing

5.2 Data Pre-processing

We define a new variable ratio, which means the percentage of goals scored in the last ten goals. The ratio variables we use are Scoring_ratio, ace_ratio, winner_ratio and so on.

Next we perform a dimensionality reduction analysis of the swing factors. We observe that several factors are highly correlates, namely scoring ratio and consecutive scores, net_point_Ratio and net_ pt_won_ratio, net_pt_consecutive_ration and net_ pt_won_ consecutive_ratio.

To solve this problem, we made synthetic factors. Momentum was synthesized using scoring ratio and executive scores, and a weighted average of equal weights was taken for the net_point_Ratio and net_ pt_won_ratio factors, i.e.,

$$net\_Ratio = net\_point\_Ratio + net\_pt\_won\_Ratio$$

Looking again at the heat map between the new factors, as shown below, **we find that the correlation between the factors is significantly reduced.**

Figure 12: Original Heatmap

Figure 13: Heatmap with Synthetic Factors

5.2.1 Data Summary

After visualizing the data, we find that the data of each feature all conform to a normal distribution (as shown below). **Among them, 'Scoring_ratio' is a clear left-skewed distribution, and the data of the rest of the features are more in line with the standard normal distribution.**

5.3 General Model

We tried to calculate the IC(information coefficient) between each factor and the performance to determine if there is a significant correlation between indicators and performance, so that we can get more general results. We used 9 different models to observe the fit, but the regression was not satisfactory. **This means it might not be a good idea to develop a model that can predict all the trends of the match. So instead, we just focus on predicting the turning points.**

5.4 Model Optimization

5.4.1 Information Coefficient(IC)

By definition, we pick out the swings in performance and perform a correlation test between the preprocessed data and the swings.

5.4.2 Hypothesis Testing

Null Hypothesis (H0): All of these factors has no effect on the flow of the match.

Alternative Hypothesis (H1): At least one of these factors has some significant effect on the flow of the match.

First we use the data to fit a linear regression model. The model equation is of the form:

$$Swing = \beta \times F + \epsilon$$

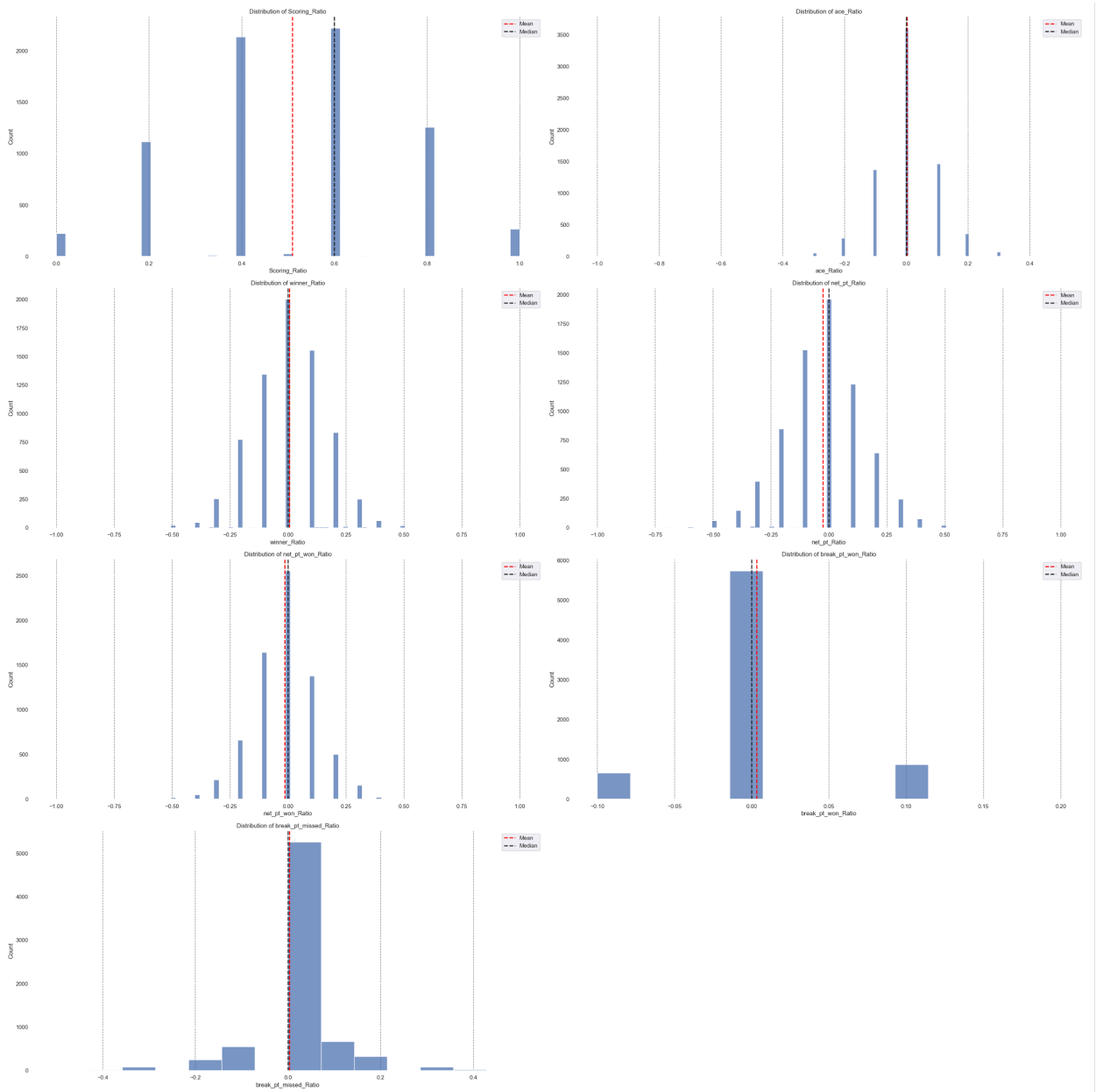

Figure 14: Distribution of the Factors

Table 5: Data Summary

Statistics	Count	Mean	Std	Min	25%	50%	75%	Max
Scoring_Ratio	7284	0.5097	0.2287	0	0.4	0.6	0.6	1
Consecutive_Scores	7284	0.0592	1.7232	-10	-1	0	1	12
ace_Consecutive_Scores	7284	0.0000	0.0000	0	0	0	0	0
ace_Ratio	7284	0.0026	0.1000	-1	0	0	0.1	0.5
winner_Consecutive_Scores	7284	0.0021	0.0562	-1	0	0	0	2
winner_Ratio	7284	0.0060	0.1592	-1	-0.1	0	0.1	1
net_pt_Consecutive_Scores	7284	-0.0010	0.0511	-2	0	0	0	1
net_pt_Ratio	7284	-0.0260	0.1735	-1	-0.1	0	0.1	1
net_pt_won_Consecutive_Scores	7284	-0.0001	0.0262	-1	0	0	0	1
net_pt_won_Ratio	7284	-0.0144	0.1369	-1	-0.1	0	0.1	1
break_pt_won_Consecutive_Scores	7284	0.0000	0.0000	0	0	0	0	0
break_pt_won_Ratio	7284	0.0031	0.0461	-0.1	0	0	0	0.2
break_pt_missed_Consecutive_Scores	7284	-0.0003	0.0234	-1	0	0	0	1
break_pt_missed_Ratio	7284	0.0035	0.0909	-0.5	0	0	0	0.5

name	IC
Momentum	0.590611
break_pt_won_Ratio	0.141623
winner_Ratio	0.14134
ace_Ratio	0.082825
net_Ratio	0.057947
winner_Consecutive_Scores	0.042713
net_Consecutive	0.019819
break_pt_missed_Ratio	-0.005733
break_pt_missed_Consecutive_Scores	-0.02144

Table 6: IC for Each Factor

model	mae	mse	r2
Decision Tree	0.290203	0.201545	0.040122
KNN	0.302613	0.153252	0.270028
Lasso Regression	0.453007	0.210032	-0.000529
Linear Regression	0.333091	0.136831	0.348227
Random Forest	0.271894	0.146156	0.303878
Ridge Regression	0.333098	0.13683	0.348227
SVR	0.280716	0.125081	0.404261
XGBoost	0.271809	0.128592	0.387532

Table 7: Regression Result

where β is the regression coefficients matrix and F is the vector of factors. Then we conduct a hypothesis test for the regression coefficient to determine whether $\beta 1$, $\beta 2$,..., βn is significantly different from zero. We also examine what the factor loadings are. The larger and more significantly different from zero the factor loadings are, the greater the

name	IC
Scoring_Ratio	-0.22052
break_pt_won_Ratio	-0.110483
ace_Ratio	-0.039632
winner_Ratio	-0.035778
break_pt_missed_Consecutive_Scores	-0.0353
net_Ratio	-0.018617
break_pt_missed_Ratio	-0.006585
winner_Consecutive_Scores	0.024785
net_Consecutive	0.028043
Momentum	0.069393

Table 8: Modified IC for Each Factor

contribution of the factor.

5.4.3 Model Fitting Result

We find that SVM Classifier is most effective when training the model after eliminating the invalid feature 'net_Ratio' and 'break_pt_missed_Ratio'.

After training the model, we can feed the data of the past points into the model to **predict whether there is a swing in the next game**: swing = 1 for a shift from favoring player 2 to favoring player 1, and swing = -1 for vice versa.

OLS Regression Results

Dep. Variable:	swing	R-squared:	0.266
Model:	OLS	Adj. R-squared:	0.265
Method:	Least Squares	F-statistic:	263.8
Date:	Mon, 05 Feb 2024	Prob (F-statistic):	0.00
Time:	15:16:09	Log-Likelihood:	-4149.0
No. Observations:	7283	AIC:	8320.
Df Residuals:	7272	BIC:	8396.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.7639	0.018	43.405	0.000	0.729	0.798
Scoring_Ratio	-2.0139	0.042	-48.486	0.000	-2.095	-1.933
ace_Ratio	0.0383	0.052	0.739	0.460	-0.063	0.140
winner_Consecutive_Scores	0.1072	0.090	1.193	0.233	-0.069	0.283
winner_Ratio	0.0590	0.036	1.649	0.099	-0.011	0.129
break_pt_won_Ratio	-0.3407	0.121	-2.808	0.005	-0.579	-0.103
break_pt_missed_Consecutive_Scores	-0.5559	0.214	-2.593	0.010	-0.976	-0.136
break_pt_missed_Ratio	-0.0062	0.058	-0.107	0.915	-0.120	0.107
Momentum	0.1940	0.004	45.906	0.000	0.186	0.202
net_Ratio	-0.0087	0.018	-0.470	0.638	-0.045	0.028
net_Consecutive	0.1177	0.074	1.596	0.111	-0.027	0.262

Omnibus:	3.451	Durbin-Watson:	2.691
Prob(Omnibus):	0.178	Jarque-Bera (JB):	3.643
Skew:	-0.006	Prob(JB):	0.162
Kurtosis:	3.109	Cond. No.	113.

Figure 15: Modified OLS regression

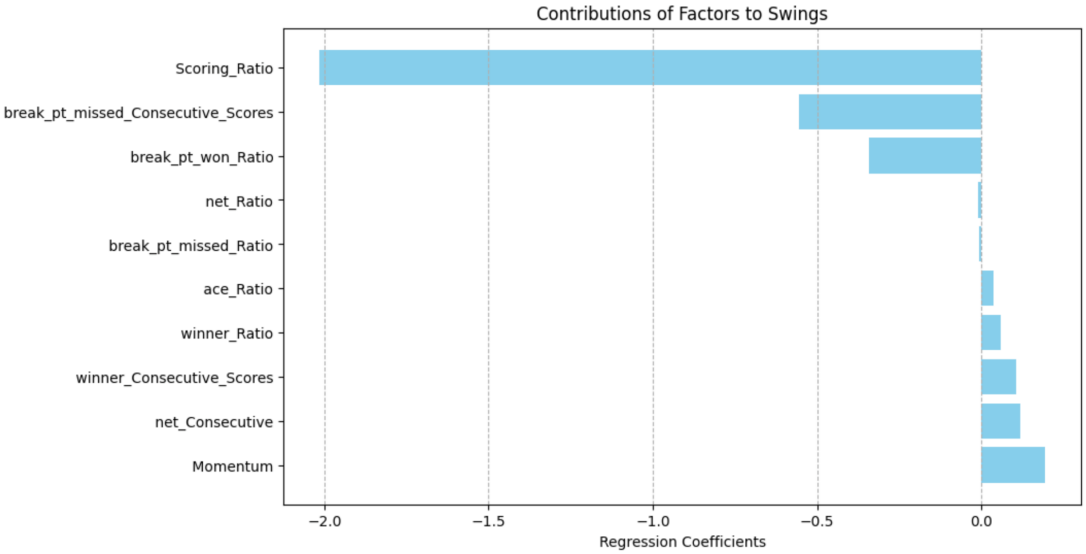

Figure 16: Contributions of Factors to Swings

SVC has the best performance in these four models

	precision	recall	f1-score	support
-1.0	1.00	0.35	0.51	179
0.0	0.82	1.00	0.90	1088
1.0	1.00	0.37	0.54	190
accuracy			0.84	1457
macro avg	0.94	0.57	0.65	1457
weighted avg	0.87	0.84	0.81	1457

Figure 17: Fitting Result

Momentum, Scoring Ratio, break_pt_won_Ratio and break_pt_missed_Consecutive_Scores are the most effective 4 factors.

In addition, the coefficient of Scoring Ratio indicates **after a lot of consecutive scores, it is highly likely that there will be a lost score.** [3] In general, this 'a lot of' in 2023 Wimbledon Tennis Championships means **4 points in a row**.

5.5 Advice

Given the importance of momentum in a tennis match and the momentum reversal effect, we have following suggestions so that players can employ methods and tactics to ensure they are in control of momentum rather than a victim of it.[4]

Momentum is against a player: It's advised to slow down, be careful and follow rituals to reverse momentum.

Momentum is turning against a player: It's suggested to boost energy level and play more aggressively.

Momentum is neutral: It's recommended for players to utilize their primary strategies when serving and returning to take control of momentum.

Momentum is in favor of a player: It's significant to rationalize victories and avoid a false sense of security.

Momentum is totally with a player: Players are supposed to seek to diversify their game to maintain the momentum.

Specifically, coaches and players can get sufficient guides on what to look for by analyzing what factor is contributing.

For example, for regular matches, the three most effective indicators are momentum (positive effect), the ratio of scoring (in the last 5 points) (negative effect), and the number of consecutive misses (negative effect). This suggests that **players who perform well in the last five points need to be wary of the reversal of the momentum effect**; they need to stay calm, focused, and adjust tactics flexibly. **Players with consecutive misses in the last ten points need to be cautious about the continuation of the momentum effect**; they need to actively respond, identify weaknesses in the opponent's tactics, and improve their own tactics to reverse the momentum in the next game.(Figure 18)

6 Problem Four

6.1 Test Our Model On Other Matches In 2023 Wimbledon

Before training the model, we set aside 20% of the dataset for testing purposes. The evaluation of the model's performance in the previous sections was based on its performance on the test data. **Therefore, as for the effectiveness of the model in predicting the swings in the 2023 Wimbledon tournament, we have already discussed it in the previous section and will not dwell on it further.**

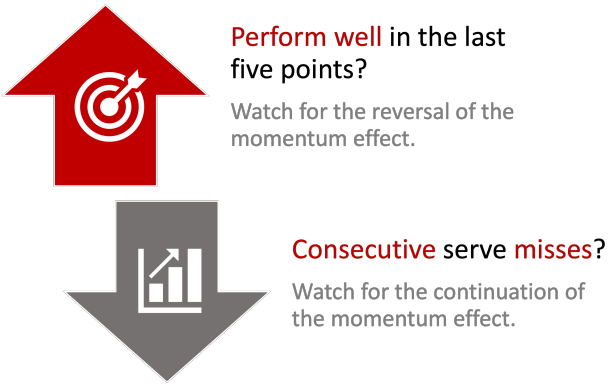

Figure 18: An Example of Advise

6.2 Test Our Model On Other Matches In 2022 US Open and 2013 French Open

We used the data available online (from Visit the website), specifically the data from the 2013 French Open Tennis and the 2022 US Open Tennis tournaments, as test data to assess whether the model can be used in different matches. We take the temporal differences (as tennis conditions may vary over a decade), as well as the type of tennis matches into consideration, so the two different scenarios can provide robust evaluation of our model's applicability under different conditions.

6.2.1 Data Screening

As before, we defined the swing of the game. Additionally, we integrated the performance of the players, such as the untouchable winning serve, untouchable winning shots, break point of the players into two categories: Ratio and Consecutive.

First we examine the contribution of the factors using hypothesis testing.

Null Hypothesis (H0): All of these factors has no effect on the flow of the match.

Alternative Hypothesis (H1): At least one of these factors has some significant effect on the flow of the match.

As for the 2013 French Open, the outcome is shown in Figure 19 and Figure 20.

As for the 2022 US Open, our outcome is shown in Figure 21 and Figure 22.

From above, we find that most of the contribution come from score_Ratio and Momentum, which corresponds with the conclusion we have drawn. **We confirmed that momentum (positive effect), the ratio of scoring (negative effect), the number of consecutive misses (negative effect) were the three most effective indicators for predicting turning points in the game.**

As explained before, this suggests that players who perform well in the last five points need to be wary of the reversal of the momentum effect; they need to stay calm, focused, and adjust tactics flexibly. Players with consecutive misses in the last ten points need to be

OLS Regression Results

Dep. Variable:	swing	R-squared:	0.031
Model:	OLS	Adj. R-squared:	0.031
Method:	Least Squares	F-statistic:	97.09
Date:	Mon, 05 Feb 2024	Prob (F-statistic):	1.69e-218
Time:	17:09:26	Log-Likelihood:	-23025.
No. Observations:	33248	AIC:	4.607e+04
Df Residuals:	33236	BIC:	4.618e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0010	0.003	-0.376	0.707	-0.006	0.004
scores_Ratio	-0.3451	0.016	-21.610	0.000	-0.376	-0.314
ace_Consecutive	0.0142	0.045	0.316	0.752	-0.074	0.102
ace_Ratio	-0.0688	0.039	-1.778	0.075	-0.145	0.007
winner_Consecutive	0.0048	0.011	0.449	0.654	-0.016	0.026
winner_Ratio	0.0014	0.018	0.078	0.938	-0.033	0.036
break_pt_won_Consecutive	-7.5e-17	4.7e-17	-1.596	0.110	-1.67e-16	1.71e-17
break_pt_won_Ratio	-0.2116	0.074	-2.877	0.004	-0.356	-0.067
break_pt_missed_Consecutive	0.1501	0.017	9.048	0.000	0.118	0.183
break_pt_missed_Ratio	-0.0401	0.029	-1.395	0.163	-0.096	0.016
Momentum	0.0304	0.002	16.948	0.000	0.027	0.034
net_Ratio	-0.0155	0.012	-1.310	0.190	-0.039	0.008
net_Consecutive	0.0072	0.012	0.585	0.559	-0.017	0.031

Omnibus:	698.612	Durbin-Watson:	2.025
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1559.954
Skew:	-0.002	Prob(JB):	0.00
Kurtosis:	4.061	Cond. No.	3.18e+17

Figure 19: 2013 French Open Coefficients Significance

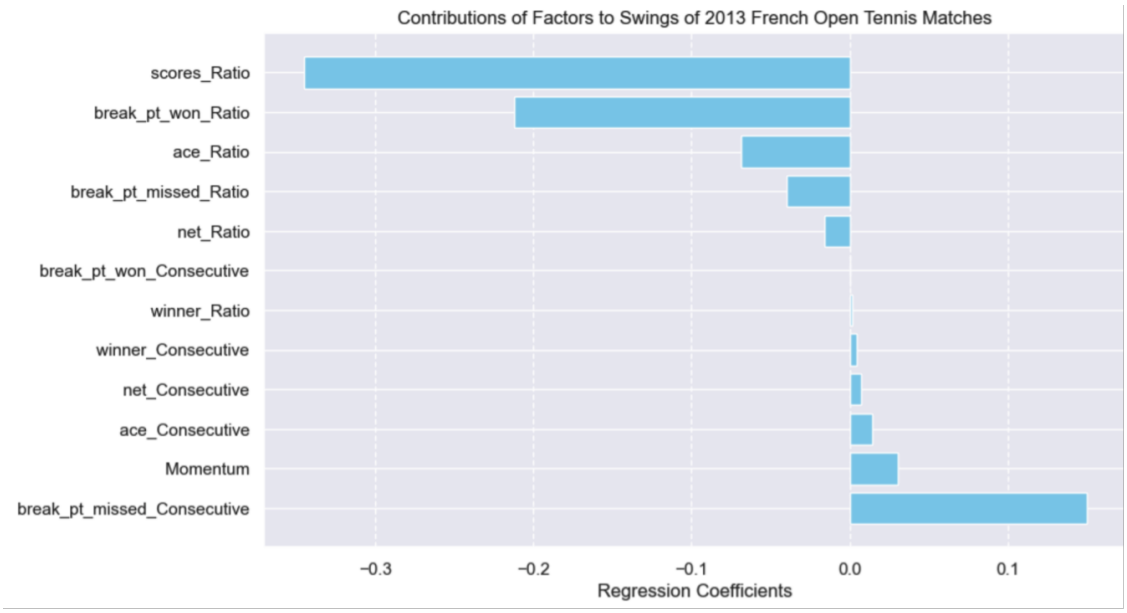

Figure 20: 2013 French Open Coefficients

OLS Regression Results

Dep. Variable:	swing	R-squared:	0.034
Model:	OLS	Adj. R-squared:	0.034
Method:	Least Squares	F-statistic:	150.7
Date:	Mon, 05 Feb 2024	Prob (F-statistic):	0.00
Time:	17:48:08	Log-Likelihood:	-32462.
No. Observations:	47243	AIC:	6.495e+04
Df Residuals:	47231	BIC:	6.505e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0007	0.002	0.302	0.763	-0.004	0.005
scores_Ratio	-0.4292	0.015	-28.944	0.000	-0.458	-0.400
ace_Consecutive	0.0109	0.027	0.407	0.684	-0.042	0.063
ace_Ratio	-0.0737	0.027	-2.722	0.006	-0.127	-0.021
winner_Consecutive	-0.0014	0.011	-0.129	0.897	-0.023	0.020
winner_Ratio	0.0182	0.017	1.072	0.284	-0.015	0.051
break_pt_won_Consecutive	-1.147e-17	9.04e-17	-0.127	0.899	-1.89e-16	1.66e-16
break_pt_won_Ratio	0.0363	0.067	0.540	0.589	-0.096	0.168
break_pt_missed_Consecutive	0.1861	0.014	13.622	0.000	0.159	0.213
break_pt_missed_Ratio	-0.0356	0.025	-1.445	0.148	-0.084	0.013
Momentum	0.0336	0.002	21.965	0.000	0.031	0.037
net_Ratio	-0.0275	0.010	-2.735	0.006	-0.047	-0.008
net_Consecutive	0.0157	0.009	1.764	0.078	-0.002	0.033

Omnibus:	1042.821	Durbin-Watson:	2.031
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2384.979
Skew:	0.005	Prob(JB):	0.00
Kurtosis:	4.101	Cond. No.	3.80e+17

Figure 21: 2022 US Open Coefficients Significance

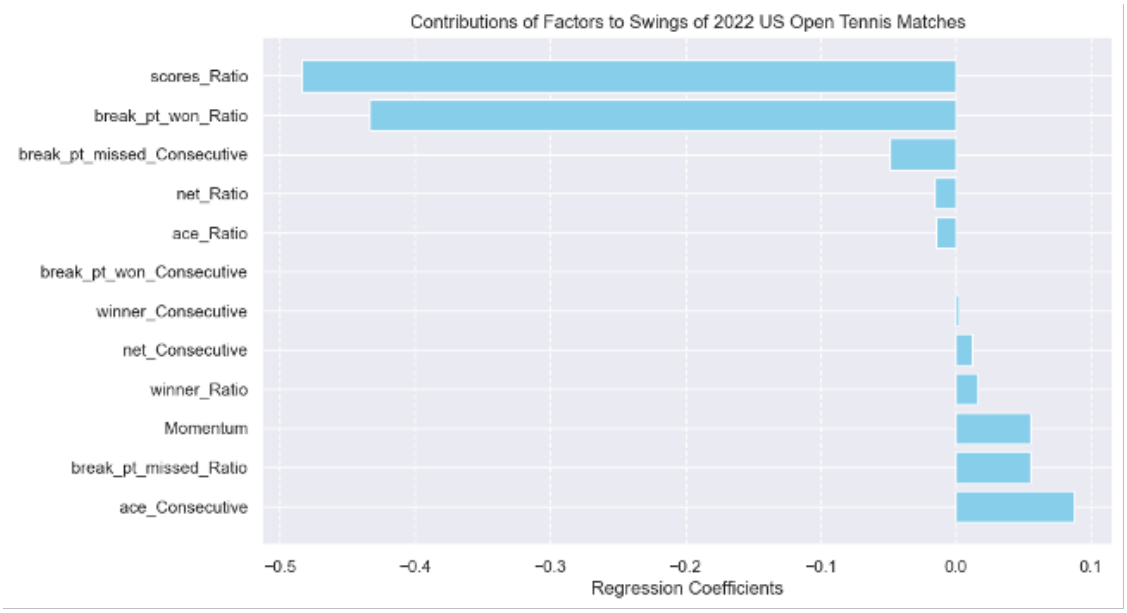

Figure 22: 2022 US Open Coefficients

cautious about the continuation of the momentum effect; they need to actively respond, identify weaknesses in the opponent's tactics, and improve their own tactics to reverse the momentum in the next game.

6.2.2 Swing of the Games

We used the model mentioned above to predict the swing in the 2022 match. However, the result isn't as good as expected. As shown in Figure 23, the precision of the prediction is only about 65%. Therefore, we think of ways to improve our model.

6.2.3 Improving Model: PCA

Consequently, we used PCA (Principal Component Analysis) to improve the model. During practice, we found that limited improvement is observed after applying PCA. As is shown in Figure 24, the precision of predicting the positive swing (the match changing from favoring player 2 to favoring player 1) decreases by about 20 percent.

	precision	recall	f1-score	support
-1	0.60	0.00	0.01	773
0	0.76	1.00	0.86	5049
1	0.50	0.00	0.00	828
accuracy			0.76	6650
macro avg	0.62	0.33	0.29	6650
weighted avg	0.71	0.76	0.66	6650

Figure 23: Evaluation of the Model Before CPA

	precision	recall	f1-score	support
-1	0.78	0.01	0.01	1110
0	0.76	1.00	0.86	7175
1	0.86	0.01	0.01	1164
accuracy			0.76	9449
macro avg	0.80	0.34	0.30	9449
weighted avg	0.77	0.76	0.66	9449

Figure 24: Evaluation of the Model After CPA

This makes sense. Since **PCA assumes that the data has a linear relationship**, if the structure of the data is more complex with nonlinear relationships, PCA may not be able to capture this complexity. PCA selects principal components by capturing the directions with the maximum variance, but variance is not always consistent with the importance of the data. **In this dataset, features like Momentum and Score_Ratio have relatively small variances, yet they are among the most important features. Their importance is underestimated by PCA,** leading to a model fit with PCA that does not show a significant improvement in performance. So we had our second trial: correlation analysis.

6.2.4 Improving Model: Correlation Analysis

By using the Heatmap(Figure 25) to test the relationship between the factors, **we found that the momentum, break_pt_won_Ratio and scores_Ratio are closely related to each other**, which may lower the precision of our model. As a result, we tried to adjust the composition of the factors. In our first attempt, we deleted the factors momentum and break_pt_won_Ratio, which means that when considering which factors would affect the swing of the match, we excluded the factors momentum and break_pt_won_Ratio.

After applying the models (Logistic Regression Classifier, K-Nearest Neighbors (KNN) Classifier, Random Forest Classifier, Support Vector Classifier (SVC)) to the dataset, **we could not see significant improvement in the outcome**, as the precision of the best prediction to the dataset is 75%, as shown in Figure 25.

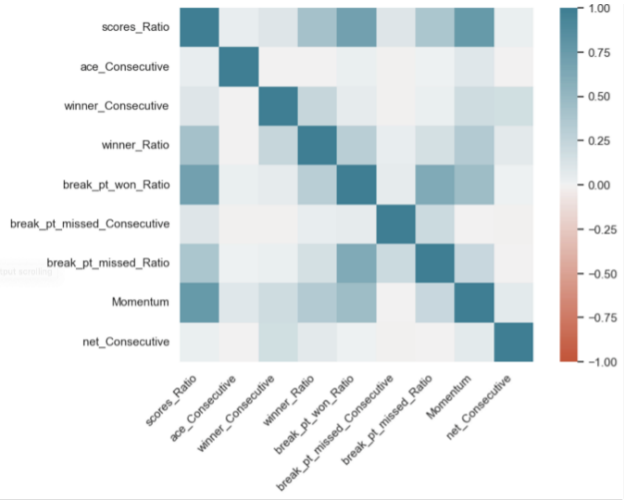

Figure 25: Heatmap of Correlation

Then we deleted the single factors: momentum. The result is as shown in Figure 27.

The result improves by another 5%, which is the best among all tests.

We apply the same to data of 2013 French Open, and the model also fits well, with predicting precision remarkably improved.

	precision	recall	f1-score	support
-1	0.75	0.01	0.01	1110
0	0.76	1.00	0.86	7175
1	1.00	0.00	0.01	1164
accuracy			0.76	9449
macro avg	0.84	0.34	0.29	9449
weighted avg	0.79	0.76	0.66	9449

Figure 26: Evaluation of the Model without Momentum and break_pt_won_Ratio

SVC has the best performance in these four models

	precision	recall	f1-score	support
-1	0.86	0.01	0.01	1110
0	0.76	1.00	0.86	7175
1	1.00	0.00	0.00	1164
accuracy			0.76	9449
macro avg	0.87	0.34	0.29	9449
weighted avg	0.80	0.76	0.66	9449

Figure 27: Evaluation of the Model without Momentum

References

[1] P. A. Richardson, W. Adler, and D. Hankes, "Game, set, match: Psychological momentum in tennis," *The Sport Psychologist*, vol. 2, no. 1, pp. 69–76, 1988.

[2] H. Dietl and C. Nesseler, "Momentum in tennis: Controlling the match," *UZH Business Working Paper Series*, no. 365, 2017.

[3] J. M. Silva, C. J. Hardy, and R. K. Crace, "Analysis of psychological momentum in intercollegiate tennis," *Journal of sport and exercise psychology*, vol. 10, no. 3, pp. 346–354, 1988.

[4] Baseline Tennis. "Understanding momentum in tennis." (2021), [Online]. Available: `https://www.youtube.com/watch?v=TDfBzQShh00`.