Problem Chosen	2024 MCM/ICM Summary Sheet	Team Control Number
C		2408868

On the Doubt about 'Momentum' Explanation of Swings in Tennis Matches

Abstract

In tennis, the flow of play may change as points occur. We design a model to capture the flow of play in tennis matches, and analyze the most related factors.

First, we develop the **Flow of Play Prediction Model** based on **Random Forest**, a machine learning algorithm that can be applied for classification and regression tasks. We describe the flow of play from two aspects. On the one hand, we use the **winner of the incoming point** to describe the current flow of play, which is a short-term description; on the other hand, we introduce the **winning percentage of next 5 points** as a long-term description. Our proposed model takes the information of an ongoing tennis match as the input, and predict the winner of the next point along with the winning percentage of the next 5 points of each player. Hence, our proposed model captures the flow of play. We then determine the most related factors that influence the flow of play by incorporating **Shapley Value** of the learned random forest, which measures the feature importance of random forest.

Random forest is a kind of **Machine Learning** algorithm. Specifically, it is an **Ensemble Learning** algorithm that takes the average of diverse **Decision Trees**. We first select features from the data file provided, and train two random forests based on the data. One is a classifier that predicts the winning of the incoming point, while the other is a regressor that gives the estimation of the winning percentage of the next 5 points.

Shapley value is a famous tool to analyze the **Feature Importance** of machine learning algorithms, especially for random forests. To investigate the most related factors that influence the flow of play, we analyze the Shapley value of each feature of the trained random forest classifier and regressor.

For Question I, we conduct extensive experiments to verify that **our proposed model captures the flow of play well**. Our model identify which player is performing better by predicting the winner of the incoming point along with the winning percentage of the next 5 points. We present a detailed visualization of our model to depict the flow in a match.

For Question II and III, we obtain the result from Shapley value which demonstrate that **the Server of Previous and Incoming Point**, **the Number of Points each Player has Won in Current Game**, **the Winning Percentage of each Player since the Beginning of Current Game** and **the Winning Percentage of each Player in 2023 Season** have great impact on the flow of play. By further analyze the Shapley values of moments when swings occur, we find that the most important factors that influence the flow of play remain unchanged. Hence, we conclude that the **'momentum' has little impact on the swings** in a match. Based on the analysis of feature importance, we advise a player that

- no matter how the flow of play goes, focus on the game since 'momentum' does not play an important role in a match;
- try to earn points when you serve, since one of the most related factors that influence the flow of play is the server;
- it is also important to improve the technique since the winning percentage also plays an important role.

For question IV, we conduct experiments to verify that our proposed models can generalize to other tennis matches. For other sports like table tennis, our model can be easily extended since the workflow of our model does not restrict the type of sports, and it will work once provided enough data. However, our model does perform poorly on some matches, and we summarize that this poor performance is caused by the strategies of the players, which is difficult for our model to predict.

Key Words: Tennis; Match Prediction; Machine Learning; Shapley Value; Random Forest

Contents

1 Introduction

1.1 Background

In competitive sports, the display of power between the two sides often takes two forms. One is that there is a certain gap in the strength of the two sides, that means one side is significantly stronger than the other. The other is that the two sides are close in strength, and it is difficult to predict the winner of the battle. However, observations and studies have shown that not only the gap in strength between the two sides could influence the tendency of the match, the flow of play of the game is also affected by a substance that called momentum.

Momentum is the product of the mass and velocity of a moving object in physics, and it can be used in competitive sports to describe the movement of a player or team. **In a match, when one player seems to win the game exactly, there is tends to be positively influenced on a mental, technical and psychological level, creating a momentum.** This momentum can be demonstrated by scoring consecutive points, counter-attacking quickly, or taking the initiative in the game. The momentum is a relatively subject and vague parameter, since the only evidence of the existence of momentum is the personal feeling of being energetic, packed with strength and feeling like being in strong advantage to win the match.

Various examples have included momentum as a potential impact factor on the result of matches. In the final Wimbledon 2023, Djokovic had a big lead in the first set. However, in the following matches, the opponent Alcaraz gradually found his form and tied the score in the second set. At third set, the momentum began to turn to Djokovic. In the adversity,however, Alcaraz quickly took control of the situation and eventually won the 4th and 5th sets to become the new king. This is a shocking news for the whole world, since Djokovic has been a common gold-medal winner of Wimbledon since more than ten years ago from this contest. People wonder why this could happen. Specifically, why Alcaraz would change the flow of the game and win the final medal.

Figure 1: Alcaraz and Djokovic in Wimbledon Final 2023 [8]

This change in the situation is quite interest, we are willing to know **whether there exist a momentum to play a role in the flow of play or not. And if exists, how the momentum works.** This is a problem that combines physics and psychology. There has long been discussed about how an athletes is performed ‘mentally’. If a player is said to be not as status or is said to be in negative mood, some coaches would consider it as an important issue that would determine the result of coming matches. Similar in other fields, ‘not in the mood’ or ‘loss in momentum’ is often viewed as a mysterious factor to be paid extra attention or being used as excuses. Due to this reason, it is of significant meaning to find an explanation to the parameter ‘momentum’ and discuss the role it plays in competitive sports, especially Men’s single tennis.

In order to figure out this conjecture, it is necessary to comprehensively consider the player’s performance, technique, tactics, psychological quality and other factors. The role of momentum in Wimbledon can be explored in more depth through data statistics during the match, performance and evaluation of players. We now restate the questions and try to seek for their answers while establishing our mathematical model to study the impact of momentum on the results of the matches.

1.2 Restatement of the Questions

1. Construct a model that assesses which player performed better at a given point in the game, and by how much.
2. A coach doubts the legitimacy of momentum. Prove whether it is real or just a random performance of the game.
3. Help coaches know when the game will shift from favoring one player to favoring another:
   (a) Using the data provided for at least one match, develop a prediction of these fluctuations in the game. What factors, if any, seem to be most relevant.
   (b) Given the differences in past matches, the "momentum" fluctuates, how would you suggest that players will play new matches against different players.
4. Sensitivity Analysis:
   (a) Test your developed model on one or more other matches. Anticipate fluctuations in the game.
   (b) If the model is sometimes underperforming, identify factors that might need to be included in the model in the future.
   (c) How it could be generalized to other competitions (e.g. women's competitions, tournaments, court surfaces and other sports such as table tennis).
5. Complete a 1-2 page memo summarizing the results and providing advice to the coach on the role of "Momentum" and how to prepare players for events that affect the flow of playing in tennis.

1.3 Overview of Our Work

Our work mainly achieves the target of designing a model to capture the flow of play in tennis matches, and analyzing the most related factors.

First, we develop the **Flow of Play Prediction Model** based on **Random Forest**, a machine learning algorithm that can be applied for classification and regression tasks. We describe the flow of play from two aspects. Our proposed model accurately captures the flow of play.

We then determine the most related factors that impact on the flow of play by implying **Shapley Value**, which is capable of measuring the feature importance of random forest. Combining the Flow of Play Prediction Model and the Shaply Value approach, we construct the overall mathematical model to solve the give questions.

To answer Question I is given by conducting extensive experiments to verify that **our proposed model captures the flow of play well**. Our model identify which player is performing better by predicting the winner of the incoming point along with the winning percentage of the next 5 points.

To answer Question II and III, we obtain the result from Shapley value which demonstrate that **the Server of Previous and Incoming Point**, **the Number of Points each Player has Won in Current Game**, **the Winning Percentage of each Player since the Beginning of Current Game** and **the Winning Percentage of each Player in 2023 Season** have great impact on the flow of play. We conclude that the **'momentum' has little impact on the swings** in a match. We then advise a player that no matter how the flow of play goes based on the analysis of feature importance. Then, we give suggestions to the athletes and explain corresponding reasons about focusing on the game since 'momentum' does not play an important role in a match, trying to earn points when you serve since one of the most related factors that influence the flow of play is the server, and improving the technique since the winning percentage also plays an important role.

To answer question IV, we conduct experiments to verify that our proposed models can generalize to other tennis matches. For other sports such as table tennis, our model can be easily extended due to that the workflow of our model does not restrict the type of sports, and our model has a strong ability in making generalizations. It will provide accurate results once it works under enough data provided as learning samples.

We finally sketch a memorandum to inform a professional coach of our latest progress. We summarized our work, made suggestions towards coaches and players on what is novel to be paid with extra cautions in training and actual matches. We appeal for more scientific approaches done to ensure that physical activities throughout training, competing and educating has reasonable scientific explanations. We also appealed for more future researches to be done.

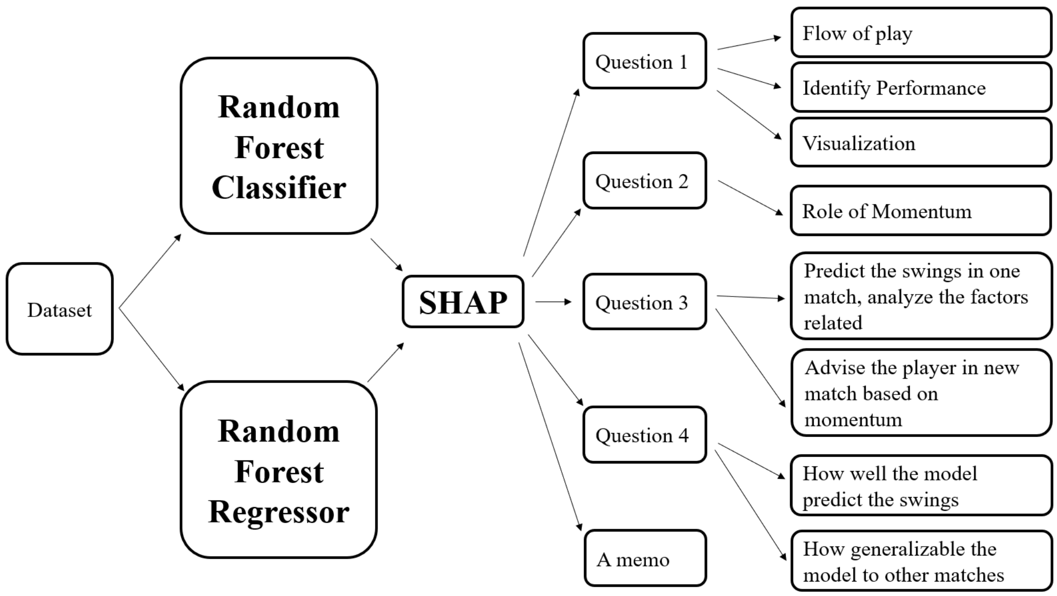

Figure 2: Alcaraz and Djokovic in Wimbledon Final 2023

2 Assumptions and Notations

This section introduces all the assumptions and notations that are about to be used to construct our models. The following assumptions ensures that the problems we are dealing is solved scientifically and effectively. The following notations convenient us to introduce our mathematical tools with less difficulty.

Assumptions

To simplify the questions and best simulate practical situations, we make the following basic assumptions and the corresponding justifications of each assumption.

- **Assumption 1:** Historical behaviors of athletes could be analyzed and used for prediction.

  Justification: The analysis and prediction of the results should be based on the known events, such as matches and coming results of smaller scales such as points or games. The previous incidents occurred in the current match should reflect useful data of the status of players. Also, the basic information of the players should act as an initial and overall parameter for future evaluation.

- **Assumption 2:** Both major and minor incidents in the matches are considered in construction of models.

  Justification: We assume that every detailed event in the matches act as a factor that would impact on the probability of winning the point and the overall ratio of achieving victory. Once the match is going on, all kinds of events must be take into consideration.

- **Assumption 3:** We assume that feature importance could be quantified in matches.

  Justification: Certain instances of tennis matches plays greater importance and others plays lesser importance. Once the features have been given out by the dataset, it will be take into account in our model preparation. We will finally make a choice depend on how important the feature means to us.

- **Assumption 4:** Momentum reflects in short periods when abnormal phenomenon appears.

  Justification: We assume that momentum plays its role when prediction fails to match with actual behaviors of a player while the model is not malfunctioning, i.e., when irregular behaviors occurs. Obviously, the professional tennis athletes won't let the things like 'momentum' influence them for a long time.

- **Assumption 5:** Environment influence and playing field doesn't take count in our model's predictions.

  Justification: We assume that the environment influence can be ignored. And all the playing field will not influence the performance of the tennis players. We don't take this kind of accident that break the match into account,because it gives no benefit for us to study the flow of play and 'momentum'.

- **Assumption 6:** The coaching team doesn't influence the player's performance in the competition.

  Justification: We assume that the coaching team will not let the player has huge change of his performance during the whole match.We don't take any strategies into consider because it will distribute us learning the features from data.

- **Assumption 7:** All participating players are in good physical condition.

  Justification: We assume that all players are in good physical condition and will not fall ill during the entire match. While we cannot guarantee the health status of every player, we generally presume that they are in good shape, without variation in until the match ends.

Notations

We introduce the notations in Table 1 in this work for simplicity.

Table 1: Notations in this work

Notation	Explanation
lowercase letters $a, b, y_1, \cdots$	scalars
bold lowercase letters $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{w}, \cdots$	vectors
$S = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_n, y_n)\}$	dataset of size n
$\mathbb{I}(\cdot)$	indicator function
$f_{m,s}(\cdot)$	random forest classifier with m trees
$g_{m,s}(\boldsymbol{x})$	random forest regressor with m trees
$\phi(i, f)$	Shapley value of function f and the i-th feature
$E_{\boldsymbol{x}\sim\mathcal{D}}[f(\boldsymbol{x})]$	the expectation of $f(\boldsymbol{x})$ over distribution $\mathcal{D}$

3 Our Proposed Approach

This section introduces our proposed approach to capture the flow of play in a match, and analyze the most related factors that influence the prediction of the swings in the match.[1] To capture the flow of play at a moment in a match, we use Random Forest to predict the winner of next point and the winning percentage of the next 5 points. We propose to use Shapley value to identify most related factors that influence the flow of play.[5]

Table 2: Features of each instance used to predict the winner of next point.

	Feature	Type
match information	match identification	categorial
	sets winned of each player by each player	numerical
	games winned of each player in current set	numerical
	scores winned of each player in current game	numerical
point information	server of previous 5 points	categorial
	winner of previous 5 points	categorial
	speed of previous 5 serves	numerical
	direction of previous 5 points	categorial
	depth of previous 5 serves	categorial
	server of the incoming point	categorial
player information	winning percentage of players	numerical

Table 3: Comparisons of the test accuracies of SVM, Random Forest and Neural Network on each Wimbledon 2023 men's match after the first 2 rounds. The best accuracy on each match is bolded. [9] [2] [3]

Match id	1301	1302	1303	1304	1305	1306	1307	1308	1309	1310	1311
Random Forest	**0.6281**	**0.7684**	**0.6992**	**0.6541**	**0.6653**	**0.7034**	**0.7450**	**0.7175**	0.6509	0.5574	**0.6335**
SVM	0.6211	0.7401	0.6842	0.6258	0.6490	**0.7034**	0.7050	**0.7175**	**0.6557**	**0.5709**	0.6211
Neural Network	0.5544	0.6441	0.6015	0.5409	0.6041	0.6207	0.6300	0.6723	0.6604	0.5270	0.5404

Match id	1312	1313	1314	1315	1316	1401	1402	1403	1404	1405	1406
Random Forest	**0.5730**	**0.6160**	**0.6821**	**0.7043**	0.6623	0.6473	**0.7056**	0.6116	**0.6861**	**0.6749**	**0.6546**
SVM	**0.5730**	0.6008	0.6763	0.6882	**0.6753**	**0.6786**	0.6976	**0.6198**	**0.6861**	0.6700	0.6495
Neural Network	0.5474	0.5095	0.5838	0.6129	0.5584	0.6339	0.6492	0.5702	0.6533	0.5222	0.5619

Match id	1407	1408	1501	1502	1503	1504	1601	1602	1701
Random Forest	0.7368	**0.7645**	**0.6966**	**0.6679**	**0.6458**	**0.6528**	0.6013	**0.6684**	**0.6038**
SVM	**0.7401**	0.7521	0.6685	**0.6679**	0.6406	0.6435	**0.6076**	0.6474	0.5723
Neural Network	0.6809	0.6281	0.6404	0.5941	0.6250	0.6111	0.5570	0.5684	0.5881

3.1 Point-Winner Prediction

In this section, we build a model to predict the winner of each point based on Random Forests classifier. We begin with the process of data preparation and then introduce how to build our classifier.

Let

$$S = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_n, y_n)\}$$

be a training sample, where $\boldsymbol{x}_t$ is the information about the current situation of a match, and y_t is the winner of the next point. More specifically, $\boldsymbol{x}_t$ consists of information about the previous 5 points and the winning percentage of each player in 2023, which is obtained from the provided data file and Pepperstone ATP Rankings [1]. For categorial features, we represent each category with natural number $0, 1, \cdots$, and we scale all features into the interval $[0, 1]$. The detailed information of each feature is summarized in Table 2.

the match identification, number of sets, games, and scores won by each player, the server of the previous 5 points and the incoming point, the winner of the previous 5 points, the speed, direction, and depth of the previous 5 serves, along with the winning percentage of each player in 2023. [6]

Given training sample S, we try to build a random forests classifier $f_{m,s}(\boldsymbol{x})$ to predict the winner of each point, where

[1]https://www.atptour.com/en/rankings/singles?DateWeek=2023-07-17

$f_{m,s}(\boldsymbol{x})$ takes a majority vote over m individual randomized trees $f_{S_1,\Theta_1}(\boldsymbol{x})$, $f_{S_2,\Theta_2}(\boldsymbol{x}), \ldots, f_{S_m,\Theta_m}(\boldsymbol{x})$, that is,

$$f_{m,S}(\boldsymbol{x}) = \mathbb{I}\left[\sum_{i=1}^{m} f_{S_i,\Theta_i}(\boldsymbol{x}) \geq \frac{m}{2}\right]. \tag{1}$$

For $1 \leq i \leq m$, S_i denotes the set of n instances obtained by independently sampling with replacement from S, and the i-th randomized tree classifier $f_{S_i,\Theta_i}(\boldsymbol{x})$ is constructed according to S_i. Random vectors $\Theta_1, \Theta_2, \ldots, \Theta_m$ are distributed identically and independently, and characterize the mechanisms of random selections of splitting dimensions during the construction of randomized trees. [12]

Formally, a random tree classifier can be constructed as follows. Each node is associated with a rectangular cell, and all leaves (external nodes) constitute a partition of $[0, 1]^d$ at each iteration of tree construction. The root of random partition is $[0, 1]^d$ itself. For simplicity, we introduce a structural list $\mathcal{P}$ to store leaves (or rectangle cells) for further splitting, which aims to keep the leaves split in successive layer. We initialize $\mathcal{P}$ by $\mathcal{P} = \{[0, 1]^d\}$, and the following procedure is repeated until the tree reaches the pre-defined depth h or the list $\mathcal{P}$ is empty.[10]

- The first leaf (or rectangular cell) is selected and removed from $\mathcal{P}$, and it will not be split if all training instances have the same label in the leaf (move to the next iteration).
- $\sqrt{d}$ candidate dimensions are randomly selected from $\{1, 2, \cdots, d\}$.
- A candidate split dimension and a split threshold are selected to maximize the *gini-index*.
- The leaf is split along the split dimension at selected threshold, and two resulting leaves are appended to $\mathcal{P}$.

Here, the gini-index serves as the criterion for the selection of the split dimension and threshold. For a given split, we use $S_L = \{(\boldsymbol{x}_1^L, y_1^L), (\boldsymbol{x}_2^L, y_2^L), \cdots, (\boldsymbol{x}_{n_L}^L, y_{n_L}^L)\}$ and $S_R = \{(\boldsymbol{x}_1^R, y_1^R), (\boldsymbol{x}_2^R, y_2^R), \cdots, (\boldsymbol{x}_{n_R}^R, y_{n_R}^R)\}$ to denote the set of training instances partitioned into left and right leaves, respectively, and gini-index could be given by

$$\text{Gini-index} = \frac{n_L}{n_L + n_R}\left(1 - \left(\frac{\sum_{i=1}^{n_L} y_i^L}{n_L}\right)^2 - \left(\frac{n_L - \sum_{i=1}^{n_L} y_i^L}{n_L}\right)^2\right) + \frac{n_R}{n_L + n_R}\left(1 - \left(\frac{\sum_{i=1}^{n_R} y_i^R}{n_R}\right)^2 - \left(\frac{n_R - \sum_{i=1}^{n_R} y_i^R}{n_R}\right)^2\right). \tag{2}$$

Intuitively speaking, Eqn. (2) measures the purity of the instances within the left and right leaves after the split. After the iterations above, we could use $C_1, C_2, \cdots, C_k$ to denote the rectangular cells associated with all k leaves, and $f_{S,\Theta}(\boldsymbol{x})$ can be further written as follows:

$$f_{S,\Theta}(\boldsymbol{x}) = \mathbb{I}\left[\sum_{j=1}^{k} \frac{\mathbb{I}[\boldsymbol{x} \in C_j] \sum_{(\boldsymbol{x}',y')\in S} \mathbb{I}[\boldsymbol{x}' \in C_j] \cdot y'}{\sum_{(\boldsymbol{x}',y')\in S} \mathbb{I}[\boldsymbol{x}' \in C_j]} \geq \frac{1}{2}\right].$$

Figure 2 demonstrates the detailed structure of random forest. We compare the performance of random forest with two popular machine learning algorithms, that is, SVM and neural network. As can be seen from Table 3, random forest generally takes better performance. Hence, we select to use random forest to predict the winner of each point and the winning percentage of the next 5 points to answer the questions. [4]

3.2 Winning Percentage Prediction

This section introduces our approach to predict the winning percentage of next 5 points based on Random Forest regressor. We construct our training sample based on the provided data file and Pepperstone ATP Rankings [14] [2], [16] and the detailed information of features can be found in Table 4. [13]

[2] https://www.atptour.com/en/rankings/singles?DateWeek=2023-07-17

Table 4: Features of each instance used to predict the winner of next point.

	Feature	Type
match information	match identification	categorial
	#sets winned of each player by each player	numerical
	#games winned of each player in current set	numerical
	#scores winned of each player in current game	numerical
	winning percentage of each player since match started	numerical
	winning percentage of each player since current set started	numerical
	winning percentage of each player since current game started	numerical
	winning percentage of each player as server	numerical
	winning percentage of each player as returner	numerical
point information	server of previous 5 points	categorial
	winner of previous 5 points	categorial
	speed of previous 5 serves	numerical
	direction of previous 5 points	categorial
	depth of previous 5 serves	categorial
	server of the incoming point	categorial
player information	winning percentage of players	numerical

Given training sample S, we build a random forests regressor $g_{m,s}(\boldsymbol{x})$ to predict the winning percentage of the next 5 points, which takes the average output over m individual randomized tree regressors $g_{S_1,\Theta_1}(\boldsymbol{x})$, $g_{S_2,\Theta_2}(\boldsymbol{x}), \dots, g_{S_m,\Theta_m}(\boldsymbol{x})$, that is,

$$g_{m,S}(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m} g_{S_i,\Theta_i}(\boldsymbol{x}). \tag{3}$$

For $1 \le i \le m$, S_i denotes the set of n instances obtained by independently sampling with replacement from S, and the i-th randomized tree regressor $g_{S_i,\Theta_i}(\boldsymbol{x})$ is constructed according to S_i. Random vectors $\Theta_1, \Theta_2, \dots, \Theta_m$ are distributed identically and independently, and characterize the mechanisms of random selections of splitting dimensions during the construction of randomized trees.

A random tree regressor $g_{S,\Theta}(\boldsymbol{x})$ is constructed in a manner similar to a random tree classifier in section 3.1, except for the choice of the splitting criterion. For a given split, we could use $S_L = \{(\boldsymbol{x}_1^L, y_1^L), (\boldsymbol{x}_2^L, y_2^L), \cdots, (\boldsymbol{x}_{n_L}^L, y_{n_L}^L)\}$ and $S_R = \{(\boldsymbol{x}_1^R, y_1^R), (\boldsymbol{x}_2^R, y_2^R), \cdots, (\boldsymbol{x}_{n_R}^R, y_{n_R}^R)\}$ to represent the set of training instances partitioned into left and right leaves, respectively, and we evaluate this split by the mean squared error as follows:

$$\text{Mean-squared-error} = \frac{n_L}{n_L + n_R}\left(\sum_{i=1}^{n_L}\left(y_i^L - \frac{1}{n_L}\sum_{j=1}^{n_L} y_j^L\right)^2\right) + \frac{n_R}{n_L + n_R}\left(\sum_{i=1}^{n_R}\left(y_i^R - \frac{1}{n_R}\sum_{j=1}^{n_R} y_j^R\right)^2\right).$$

After the iterations above, we could use $C_1, C_2, \cdots, C_k$ to denote the rectangular cells associated with all k leaves, and $g_{S,\Theta}(\boldsymbol{x})$ can be further written as follows:

$$g_{S,\Theta}(\boldsymbol{x}) = \sum_{j=1}^{k} \frac{\mathbb{I}[\boldsymbol{x} \in C_j] \sum_{(\boldsymbol{x}',y')\in S} \mathbb{I}[\boldsymbol{x}' \in C_j] \cdot y'}{\sum_{(\boldsymbol{x}',y')\in S} \mathbb{I}[\boldsymbol{x}' \in C_j]}.$$

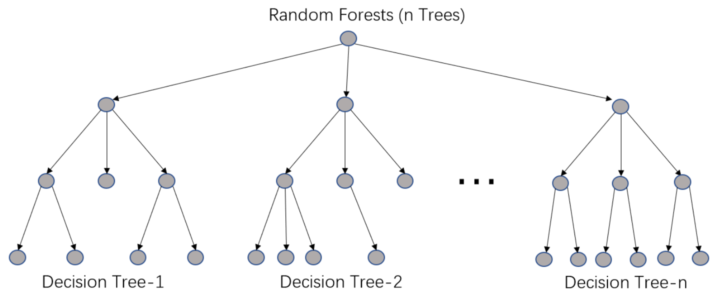

Figure 3: Schematics of Random Forest

3.3 Most Related Factors Analysis based on Shapley Values

Given target model f and instance $\boldsymbol{x}$, we try to analyze the importance for each feature of $\boldsymbol{x}$. Specially speaking, we apply the Shapley value to measure the importance of the i-th feature as follows: [10]

$$\phi(i, f) = \sum_{N \subseteq \{1,\cdots,d\} \setminus \{i\}} \frac{|N|!(d - |N| - 1)!}{d!} \left(v(N \cup \{i\}) - v(N) \right) , \tag{4}$$

where $v(N)$ is defined by

$$v(N) = E_{\boldsymbol{x}' \sim \mathcal{D}} [f(\boldsymbol{x}') \mid x'_i = x_i \text{ for } i \in N] .$$

In practice, we could estimate $v(N)$ based on training set S as follows:

$$v(N) \approx \frac{\sum_{(\boldsymbol{x}',y') \in S} f(\boldsymbol{x}') \mathbb{I}[x'_i = x_i \text{ for } i \in N]}{\sum_{(\boldsymbol{x}',y') \in S} \mathbb{I}[x'_i = x_i \text{ for } i \in N]} .$$

From a view of game theory, we can regard d individual features as d players, and these players are ordered uniformly at random. Eqn. (4) then essentially counts the average contribution of feature i to players that precede i in the orderings.

The Shapley value has several desirable properties as follows:

- *Linearity axiom.* For two independent models f, g, we have

  $$\phi(i, f + g) = \phi(i, f) + \phi(i, g) \text{ for any } i \in \{1, \cdots, d\} .$$

  This enables us to efficiently calculate Shapley values for ensemble models.

- *Dummy axiom.* Given feature $i \in \{1, 2, \cdots, d\}$, if $v(N \cup \{i\}) = v(N)$ for every $N \subseteq \{1, \cdots, d\} \setminus \{i\}$, then we have

  $$\phi(i, f) = 0 ,$$

  i.e., the Shapley value is zero for a feature with no marginal contribution.

- *Implementation invariance axiom.* Given two models f and g, if $g(\boldsymbol{x}) = f(\boldsymbol{x})$ for any $\boldsymbol{x}$, then we have

  $$\phi(i, f) = \phi(i, g) \text{ for any } i \in \{1, \cdots, d\} .$$

  This indicates that two equivalent models have identical feature importance, even with different implementations.

- *Efficiency axiom.* For individual feature $i \in \{1, \cdots, d\}$, we have

  $$\sum_{i=1}^{d} \phi(i, f) = f(\boldsymbol{x}) - E_{\boldsymbol{x}' \sim \mathcal{D}}[f(\boldsymbol{x}')] ,$$

  that is, the sum of importance of individual features is equal to the difference between $f(\boldsymbol{x})$ and $E_{\boldsymbol{x}' \sim \mathcal{D}}[f(\boldsymbol{x}')]$.

4 Answers to the Questions

In this section, we make use of our previous approaches and algorithms to answer the given questions. We also present visualized analytic results with figures that represents the characterization of the answers to the given questions.

4.1 Question I

We give our solution to Question I by presenting our mathematical model, providing the flow of game, and identifying the player in advantage at any moment as well as how much better the are.

The model required in Question I is given by the Point-Winner Prediction and the Winning Percentage Prediction approaches. This model is capable in capturing the flow or each game, we visualize our results in Figure 2. The flow of game is given by the oscillating curve shown in each diagram.

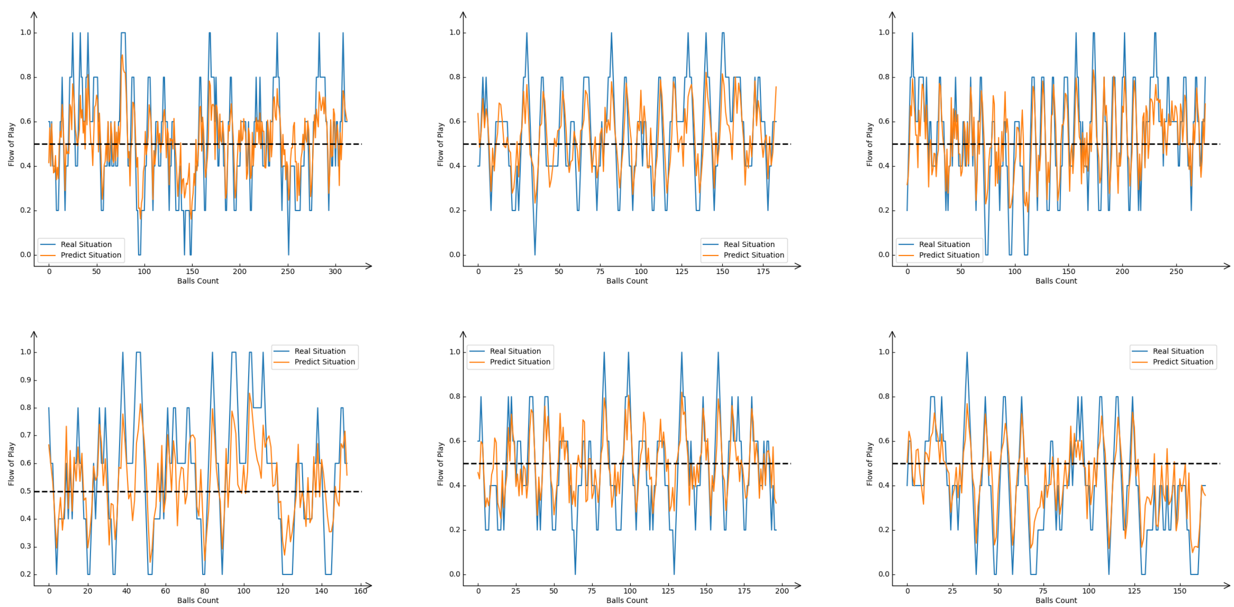

Figure 4: Visualization of the Flow of Game

This figure gives the predicted situation and the real situation. We explain them with the following terms:

- The horizontal line represents the number of balls played in one match. Time elapses as the curve travels to right hand side.
- The data on the vertical axis represents the flow of game. The more the curve approaches to +1, the more likely player 1 wins the the game. Similarly, player 2 tends to win the game as the curve approaches to -1. The middle line reflects the balancing point, at which the two players has equal percentages to win.
- The blue-colored curve represents the real situation while the red-colored curve represents the predicted situation. Hence, how well these two curves matches with each other represents how well our prediction is. The closer the red curve is to the blue curve, the better is our result.

These diagrams indicates the result of the first part of Question I that for any given match, the player in advantage is at the same side of the value of the blue-colored curve. The advantage is further quantified by the absolute value

of the output of the curve. Since the red-colored curves and the blue-colored curves are closely attached, our predictions can be viewed as accurate.

As Question I mentions, it is reasonable to make the hypothesis that the server is an important factor that impact the probability of obtaining the current score. To see how great is the importance of each factor during a match and seek out the most essential factor, we introduce the following results given Shapley value. Figure 3 present the results of our random forest predicting the final, the best prediction, the least prediction, two semifinals and a randomly picked match.

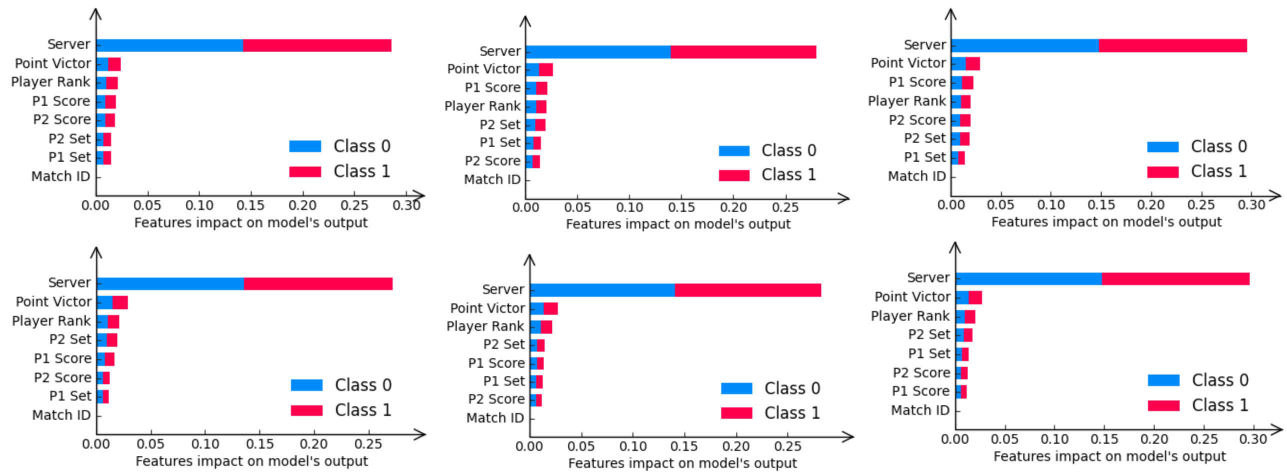

Figure 5: Importance of Factors

To explain our diagrams, the length of each bar represents the importance of the corresponding instance. **The longer a bar, the greater the importance of the instance.** The red-colored and blue-colored bar represents two classes of samples after division.

Figure 3 gives answer to the second half of Question I since the importance of each factor is reflected by the length of the corresponding bar. Also, the server instance is the factor that causes the greatest influence towards the result, since the first bar of each diagram as the greatest length. The followings have the order as point victor, player rank, p2 set, p1 score, p2 score and p1 set. **This result remains in all matches, which proves that our analysis holds great reliability.**

4.2 Question II

To reply the doubts of the tennis coach, we choose a strict way to analyze the influence of the role of momentum in tennis games based on Question I. Now we could predict the flow of play of every match as the time goes by, we pay more attention on the flow of play graphs that could help us know the momentum.

We find a common law in all matches: **If the flow of play suddenly shifts from one side to the other, it might be the effect of momentum**.But this is just a assumption we give out, to be more precise, we used SHAP to analyze our Random Forest Regression Model and we found some useful information below:

Table 5: Meaning of features in Figure 6

Features ID	Meanings
F 1	Winning percentage of player 1 in this season
F 2	Winning percentage of player 2 in this season
F 8	The number of points that player 1 has won in current game
F 9	The number of points that player 2 has won in current game
F 10	The server of the previous point
F 12	The winner of the point that just ended
F 22	The server of the incoming point
F 48	The winning percentage of player 1 from the beginning of current game
F 49	The winning percentage of player 2 from the beginning of current game
Others	Other features

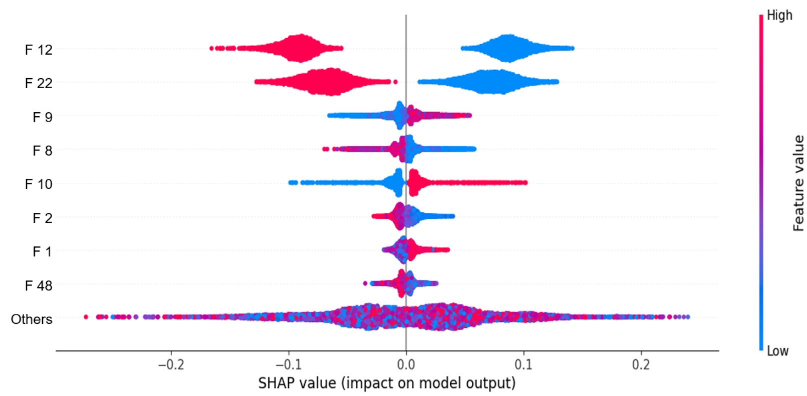

Figure 6: Most important features that influence the flow of play(Not focus on swings)

The table and figure above show the importance of all the features from highest to lowest.We could easily find out that **the winner of the ball that just ended** and **the server of the next ball** caused the biggest impact on the flow of play in the whole match. Not only it further proves the correctness of our model that server plays a irreplaceable role in tennis game but also shows that: if the turning point holds the same features, the influence of momentum is small.

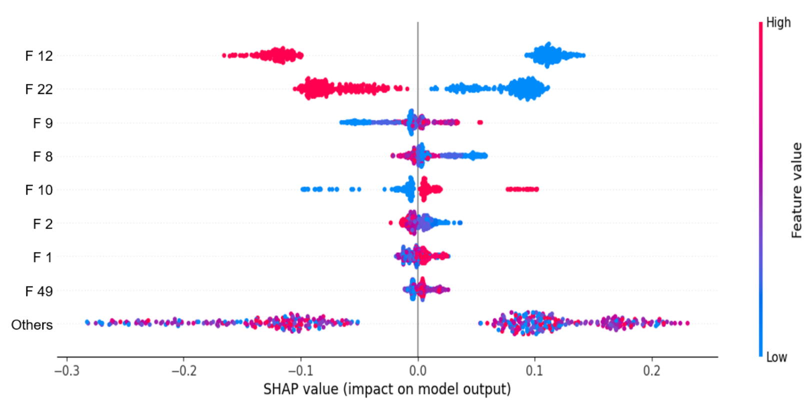

Figure 7: Most important features that influence the flow of play(Focus on swings)

The figure above shows the most important features that influence the flow of play on turning point. This holds the same features of Figure one in Question II.So the answer is: **The 'Momentum' does not play an important role in the game**. In other words: **'Momentum' doesn't outweigh other factors, and it doesn't change game expectations**. So as the tennis coach says, the swings in play and runs of success by one player are random.

4.3 Question III

We give our answer to Question III by finding the turning point of the flow of game, presenting a model that predicts the swings in matches and finding the most relevant factors.

4.3.1 Predict Swings in One Match

To answer this question, we use our two models to seek for the moment at which the flow of play is about to favor the opposite player. The ranking of factors in order of importance is then given by SHAP methods. We apply the final data into the model and obtain the prediction of the flow of play in Figure 8.

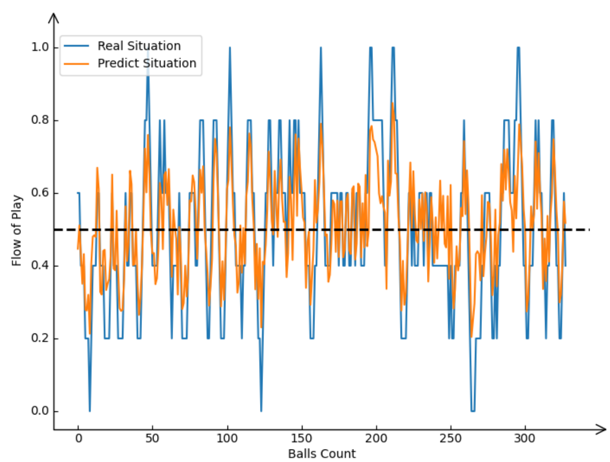

Figure 8: The flow of play of the final match

As Figure 8 indicates, Djokovic owns huge advantage in the opening game Our prediction also point to the advantage side for Djokovic at the start of the game with the result below 0.5. According to the final depiction, Djokovic performed as the underdog in the 2nd, 4rd and 5th games, while his opponent Alcaraz performed as the underdog in the 1st and 3th games. These truths fully in line with our predictions.

We apply the data from our prediction into SHAP to receive feature extraction and analysis. The following results are shown in Figure 11.

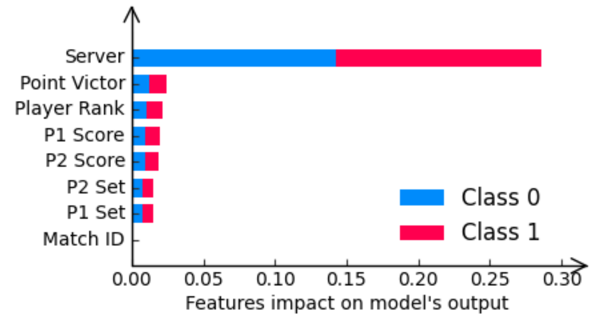

Figure 9: The final's feature exerction

The length of the bars represents how important the corresponding factor. Longer bars represent greater importance.

As indicated from Figure 11, the advantage of serving is indeed self-evident even for top tennis players. Meanwhile, the level of competition cannot be the main factor affecting winning or losing since their strengths are relatively similar. On the contrary, the current scoring situation of the players and the ratio of loss and win can affect the direction of the game even more than the previous factor.

4.3.2 Predict Swings of the Best-Fitted Match

We now test our model with the best-fitted match to see how well it responds to swings. We prove that our model could work the same good at the best competition. Here is the flow of play of the best-fitted match

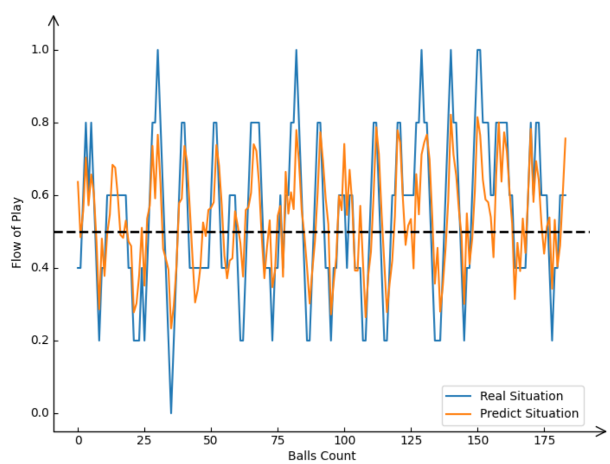

Figure 10: The best match's flow of play

Through the figure, we could find that our model predict most time of the best-fitted points and behaves well in the overall sense. Next we see the SHAP value of the best match:

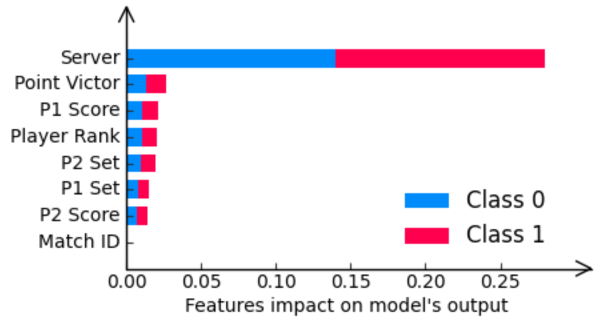

Figure 11: The best match's feature extraction

By analyzing the results given in Figure 11, we see that the server is the most important feature deciding the result of the current point. The following is the point victor with secondary importance. This test indicates that of the best-fitted match is enough reliable. We now have enough reason to deduce that **the ranking in order of importance of the impacting factors in actual cases is in such an order shown above**.

4.3.3 More Testings on Our Models

To further test on how well our models deal with random swings, we conduct the following tests on our models.

We randomly pick three matches from the sample set § and test our model with procedures of making predictions towards these testing samples. The results are reflected in the following Figure 12. Similarly, the perfection of match between the red-colored curve and the blue-colored cure represents the accuracy of our predictions.

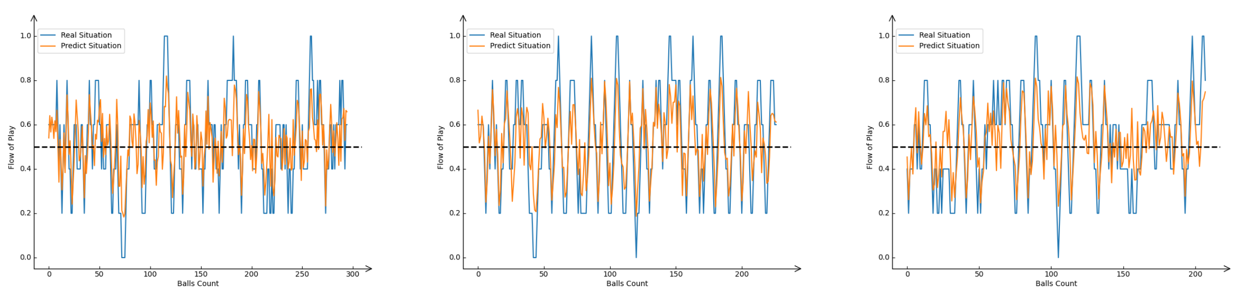

Figure 12: Prediction of flow of game on randomly picked matches

By analyzing the data from Figure 12, we concludes that our results are in false tolerance. The red-colored curves fits the blue-colored curves accurately so that practical cases could be predicted with such prediction curves.

4.3.4 Advice for player in new match

In light of all of the above analysis and the consideration of momentum in the second question, we can give the following 3 suggestions for players to better prepare for their next game:[11]

- Psychological factors such as momentum is not the main factor affecting the result of matches, so do not worry too much when facing strong opponents. Meanwhile, do not take weak opponents lightly. It is recommended to approach the game and encounter emergent circumstances calmly.

- Players should grasp the opportunity of their own service game. By analyzing the historical experiences, we see that the benefit of the first serve is owning a majority position. Thus, we should try our best to win the service game. Players are recommended to train their pointing percentage and service accuracy in the service game in advance.
- The result of matches of players often does not act as the key factor to affect the game when their rankings are relatively close. However, if the gap between the players is over large, it is more likely to cause the competitive level to become a key factor affecting the game. Players should also continue to improve their competitive level in order to minimize this gap and increase the probability of winning future matches.

4.4 Question IV

We give our answer to Question IV by testing our models on other cases, calculating errors when encountering swings and testing its ability in doing generalizations.

We first give the forecast of our model on other games. Then, by computing the mean-square averages of the errors, we estimate the accuracy of our estimations.

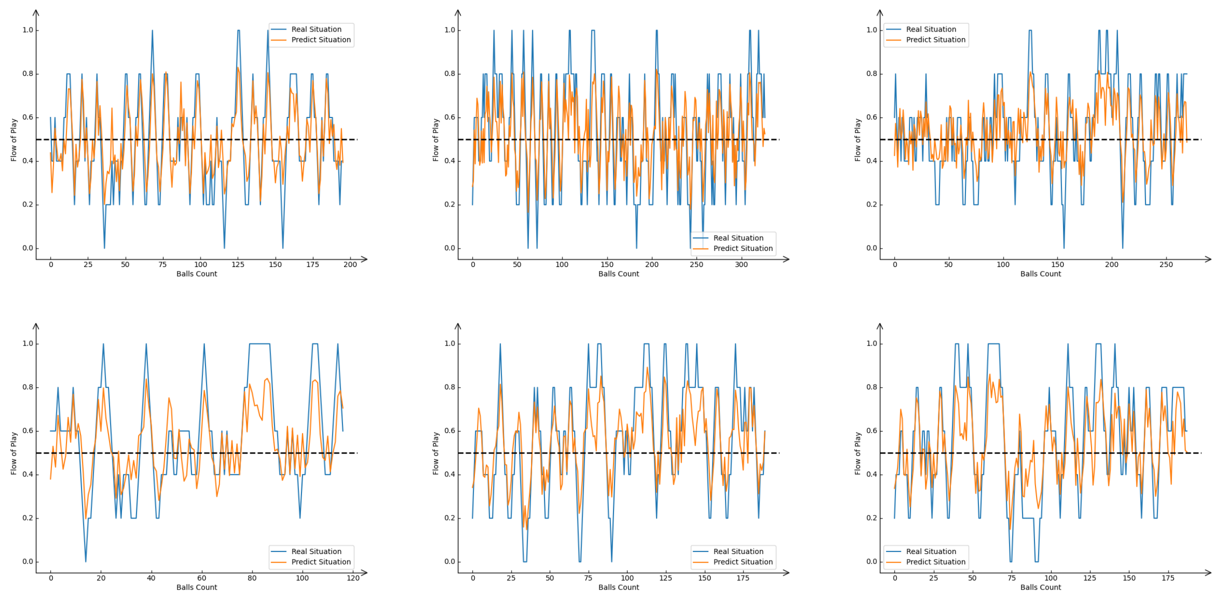

Figure 13: All 6 graphs are randomly chose from 31 matches.Figures 1-1,1-2 and 1-3 are the match 2,6,12 from 16th-finals.Figures 2-1,2-2 are the match 19,22 from 8th-finals.Figure 2-3 is the match 27 from semifinals

The graphs indicates that our model is universally reliable by having well performances in all 31 games. Also, the predicted trend was basically close to or having the same trend of the real game. The model acts accurate in predicting the overall process including crucial turning points.

As far as we have tested, the model always present satisfying results. However, if there happens to occur a situation where the model gives false prediction, **we adopt the method of adding more detailed data and dimensions of samples into consideration**. For example, if the fit degree around 80 balls of 2-1 in the figure is not accurate, we will then consider whether physical changes in the player at this time plays a greater role. Other possible examples including undetectable cases by data analysis such as sudden strategic changes of a player.

In terms of the generalization of our models, we proved that they are capable of predict other different competitions, such as Women's tennis, or table tennis, or soccer. Since our models fits both micro and macro aspects of sports games, our approach achieves maximum generalizing ability comparing with similar approaches. [7] [17]

We put our model into table tennis to see if we have the same precise on table tennis. We find that it could fit the table tennis also quite good, we use MSE to consider whether the model do well in the table tennis or not, here is the definition of MSE:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2 \tag{5}$$

Where the Y_i is the real value and $\bar{Y}_i$. The value of the sum of all the minus indicates the degree of fit. The less the MSE is, the better we fit our results. After we put our model onto the table tennis, we get the MSE in the form of

Table 6: Meaning of features in Figure 6

Value in sports	MSE Value
MSE in table tennis	0.034121723466857402
MSE in tennis	0.004522637485970819

Our model fits our results in a quite perfectly with few errors. There is no doubt that our model is not only successful in tennis, but also in other sports games like table tennis.

5 Strengths and Weaknesses

In this section, we review our previous work and list out the strengths and rooms for further improvement. Overall, the pros of our models greatly surpasses their cons.

Strengths

- Our model is competent of giving reliable predictions towards both points and matches results. We have sufficiently proved that our model is reliable when making historical analysis and future predictions. Further, we could give explicit figures representing the flow of game for any given match.
- Our model gives the most related factors and the quantified relative importance of these factors. The ranking of importance of the factors works universally for all samples, hence indicating the correctness of the resulting order of these factors.
- Our model applies approaches that has the best results. By making parallel comparisons with data from SVM and neural network, we see that our approach using the random forest has the best average ratio of accuracy in making predictions.

Weaknesses

- We have not fully considered the trivial differences of mental flux of players, since momentum is viewed as a minor impact factor comparing with other objective factors. This added potential risks to the result due to the accumulation of flux and the complexity of psychological synergistic effect.
- We could not fully consider the strategic factors that often occur in plays. Athletes often choose to give up on some certain scores and thrive for some other certain scores intentionally. Either decided by players themselves or by their coaches, these confidential strategies could not be well-predicted in practical cases.

- We have not found a universal method to cope with the problem of over-fitting. This would cause additional problems when complexity of samples varies drastically. The current method applying the random forest already lowers this chance to a relatively safe status.

6 Conclusions

- Our proposed model captures the flow of play. The flow of play is visualized by figures provided in the previous results. We give results of both short-term and long-term predictions of the player winning and losing the game and the ratio of how much one is at advantage of the other.
- Our proposed model captures the flow of play well. Our model identify which player is performing better by predicting the winner of the incoming point along with the winning percentage of the next 5 points, and all of them gives results that are with trivial errors.
- We see that the server is an important factor influencing the result of the points and matches. In fact, it is the greatest influencing factor, weighing over all the other kinds at a great percentage. This phenomenon holds for all the matches we observed.
- The 'momentum' has little impact on the swings in a match. Though psychological factors such as the abstract concept of momentum could influence the flow of the game, it weighs too little comparing with other objective factors with the server as an example.
- The flow of play is about to be changed when the prediction result turns it signal. If the signal turns from positive to negative, the corresponding result is that the favoring player is changing from the first respecting the positive to the second respecting the negative.
- After establishing our model to predict for the swings of the match, we see that the most related factor is the server of the current point. The server has a much greater ratio in winning this point than the opposing side. This fact is universal in all matches we tested.
- To give advice to coaches: No matter how the flow of play goes, focus on the game since 'momentum' does not play an important role in a match. Try to earn points when you serve, since one of the most related factors that influence the flow of play is the server.
- To give advice to athletes: Improve the technique since the winning percentage also plays an important role. The competitive strength is the only factor athletes are able to control and make improvements before the match begins. They could change the result ahead of time.
- Our model responds well to all kinds of situations and make reliable predictions. For other sports like table tennis, our model can be easily extended since the workflow of our model does not restrict the type of sports, and it will work once provided enough data.
- If our model performs poor to certain occasions, we adopt the method of adding more detailed data and dimensions of samples into consideration. In this way, our models could better cope with samples with adding diversity.
- Our model has strong ability in making generalizations towards other types of competitive sports. This is due to the strong approaches and models that we have chosen. The random forest has significant advantage in generalizing ability, which competes the other possible approaches.
- A memorandum is made to inform a professional coach of our latest progress. We summarized our work, made suggestions towards coaches and players on what is novel to be paid with extra cautions in training and actual matches. We also appealed for future researches.

References

[1] Tristan Barnett and Stephen R. Clarke. Combining player statistics to predict outcomes of tennis matches. *IMA Journal of Management Mathematics*, 16:113–120, 2005.

[2] Firas Bayram, Davide Garbarino, and Annalisa Barla. Predicting tennis match outcomes with network analysis and machine learning. In Tomáš Bureš, Riccardo Dondi, Johann Gamper, Giovanna Guerrini, Tomasz Jurdziński, Claus Pahl, Florian Sikora, and Prudence W.H. Wong, editors, *SOFSEM 2021: Theory and Practice of Computer Science*, pages 505–518, Cham, 2021. Springer International Publishing.

[3] R. Can, S. Kocaman, and C. Gokceoglu. A comprehensive assessment of XGBoost algorithm for landslide susceptibility mapping in the upper basin of Ataturk dam, Turkey. *Applied Sciences*, 11(11):4993, 2021.

[4] Andre Cornman, Grant Spellman, and Daniel Wright. Machine learning for professional tennis match prediction and betting. 2017.

[5] Wei Gao, Fan Xu, and Zhi-Hua Zhou. Towards convergence rate analysis of random forests for classification. *Artificial Intelligence*, 313:103788, 2022.

[6] Zijian Gao and Amanda Kowalczyk. Random forest model identifies serve strength as a key predictor of tennis match outcome. *ArXiv*, abs/1910.03203, 2019.

[7] Andreas Groll, Cristophe Ley, Gunther Schauberger, and Hans Van Eetvelde. A hybrid random forest to predict soccer matches in international tournaments. *Journal of Quantitative Analysis in Sports*, 15(4):271–287, 2019.

[8] Alexis Jones. Carlos alcaraz defeats 4-time consecutive wimbledon winner novak djokovic in thrilling men's singles final. https://people.com/carlos-alcaraz-defeats-novak-djokovic-wimbledon-mens-singles-final-7561683. Accessed 2-3-2024.

[9] Marius Köppel, Alexander Segner, Martin Wagener, Lukas Pensel, Andreas Karwath, and Stefan Kramer. Pairwise learning to rank by neural networks revisited: Reconstruction, theoretical analysis and practical performance. In Ulf Brefeld, Elisa Fromont, Andreas Hotho, Arno Knobbe, Marloes Maathuis, and Céline Robardet, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 237–252, Cham, 2020. Springer International Publishing.

[10] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles, 2019.

[11] Graham Pollard and Tristan Barnett. Fairer service exchange mechanisms for tennis when some psychological factors exist. *Journal of sports science & medicine*, 5 4:548–55, 2006.

[12] G. Shobana and M. Suguna. Sports prediction based on random forest algorithm. 2021.

[13] Tim D. Smithies, Mark J. Campbell, Niall Ramsbottom, and Adam J. Toth. A random forest approach to identify metrics that best predict match outcome and player ranking in the esport rocket league. *Scientific Reports*, 11(1):19285, Sep 2021.

[14] Demetris Spanias. Professional tennis : quantitative models and ranking algorithms. 2014.

[15] CFI Team. Random forest a combination of decision trees that can be modeled for prediction and behavior analysis. https://corporatefinanceinstitute.com/resources/data-science/random-forest/. Accessed 2-4-2024.

[16] ATP Tour. Pepperstone atp rankings. https://www.atptour.com/en/rankings/singles?DateWeek=2023-07-17. Accessed 2-3-2024.

[17] K.-C. Yang, O. Varol, P.-M. Hui, and F. Menczer. Scalable and generalizable social bot detection through data selection. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 1096–1103, 2020.

Appendix

1.Python core source code of Classifier:

```
if __name__ == '__main__':
X, Y = extract_feature('data.csv')
feature_idx = [0, 1, 2, 7, 8, 9, 10, 11, 12, 13, 14, 15, 46]
X = X[:, feature_idx]
for i in range(len(X[:,10:11])):
X[i,1:2] = 0
X = augment_feature(X)
unique_values = np.unique(X[:, 0])
need_unique_value = np.concatenate((unique_values[0:7],
unique_values[10:12], unique_values[13:15],
unique_values[22:26], [unique_values[29]]), axis=0)
neednt_unique_value = np.concatenate((unique_values[7:10],
[unique_values[12]], unique_values[15:22],
unique_values[26:29], [unique_values[30]]), axis=0)

X1_train = []
Y1_train = []
for i in  range(len(X)):
if X[i,0] in need_unique_value:
X1_train.append(np.concatenate((X[i, 1:2], X[i, 4:7],
X[i, 7:9], X[i, 11:13])))
Y1_train.append(Y[i])
X1_train = np.array(X1_train)
Y1_train = np.array(Y1_train)

# Define X_train and Y_train here
X_train, Y_train = X1_train, Y1_train
count = 0
X1_test = []
Y1_test = []
for i in range(len(X)):
if X[i, 1] == X[i, 10] and X[i, 0] == 0.0225:
if count%2 != 0:
concatenated_array = np.concatenate
((X[i, 1:2], X[i, 4:7], X[i, 7:9], X[i, 11:13]))
X1_test.append(concatenated_array)
Y1_test.append(Y[i])
count += 1
X1_test = np.array(X1_test)
Y1_test = np.array(Y1_test)
X_test,Y_test = X1_test,Y1_test

rf = RandomForestClassifier(n_estimators=100, max_depth=6)
rf.fit(X_train, Y_train)
train_accuracy = rf.score(X_train, Y_train)
y_pred = rf.predict(X_test)
print('Test accuracy: ', accuracy_score(Y_test, y_pred))
```

2.Python core source code of Regressor:

```
if __name__ == '__main__':
X, Y = extract_feature('Wimbledon_featured_matches.csv')
curr_feature_idx = [0, 1, 2, 3, 7, 8, 9, 10, 11, 12, 13,
14, 15, 16, 17, 39, 40, 41, 42, 43,     44, 45, 46,
47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60]
```

```
prev_feature_idx = [39, 40, 41, 42, 43, 44, 45, 47, 48, 49,
50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60]
X = augment_feature(X, curr_feature_idx , prev_feature_idx)
# X_train , X_test , Y_train , Y_test = train_test_split(X, Y)
X_train , Y_train = X, Y

rfr = RandomForestRegressor(n_estimators=100, max_depth=13)
rfr.fit(X_train , Y_train)
y_pred = rfr.predict(X_train)
print('MSE train: ', mean_squared_error(y_pred , Y_train))

# explainer = shap.TreeExplainer(rfr)
# shap_values = explainer(X_train)
# shap.plots.beeswarm(shap_values ,max_display = 9)

n = X_train.shape[0]
idx = []
for i in range(n):
if Y[i] == 0 or Y[i] == 1:
idx.append(i)
explainer = shap.TreeExplainer(rfr)
shap_values = explainer(X_train[idx , :])
shap.plots.beeswarm(shap_values ,max_display = 9)
```

3.Python core source code of standardization:

```
def scale_data(X):
min_f = X.min(axis=0)
max_f = X.max(axis=0)
d = X.shape[1]
for i in range(d):
if max_f[i] - min_f[i] < 1e-7:
max_f[i] = min_f[i] + 1.0
X = (X - min_f) / (max_f - min_f)
return X


def extract_feature(data_source):
num_features = 46
X = np.zeros((0, num_features))
Y = []

with open(data_source) as f:
f_csv = csv.reader(f)
for row in f_csv:
if row[9] == row[10] == '6':
continue
row = convert_features(row, num_features)
X = np.vstack((X, np.array(row)))
if row[15] == 1:
Y.append(1)
elif row[15] == 2:
Y.append(-1)

X = scale_data(X)
Y = np.array(Y)
server = X[1:, 13:14]
X = X[:-1, :]
Y = Y[1:]
```

```
X = np.hstack((X, server))

n = X.shape[0]
idx = []
# delete the last point
for i in range(0, n - 1):
if X[i][0] == X[i + 1][0]:
idx.append(i)
X = X[idx, :]
Y = Y[idx]

return X, Y
```

MEMORANDUM

To: Coach Smith

From: Tennis Fan Club

Subject: New Result on Tennis Prediction

Date: Monday, February 5, 2024

I am writing to inform you about our latest work on tennis matches predictions. We could now predict the flow of play in tennis matches the give the most related factors to the result of play.

Our study was inspired by the Wimbledon 2023 final, during which Alcaraz defeated previous world champion Djokovic and became the new world champion. This draws our interest on the question of whether the result is predictable, and what are the factors that cause Alcaraz to make this historical achievement.

Motivated by so, we developed mathematical models and captured the flow of play. We made thorough analysis on the most related factors. The greatest relative factor turns out to be the server of the point. The followings include winner of the previous point, win rate of both players in season, distance of ball flight and more.

Many coaches look highly on athletes' mental stability. They believe that 'momentum', the subjective feeling of having strength or advantage in a game, is the most important factor in winning. However We concluded that momentum has little impact on the swings in a match. The server factor and other objective aspects weigh much more than momentum. We advise coaches that memorandum should not be specifically viewed in training. What is needed for athletes is still the regular strengths and technical training.

As for athletes, we advise a player to cope with drastic changes of the flow of play with the following methods.

- No matter how the flow of play goes, focus on the game since 'momentum' does not play an important role in a match;
- Try to earn points when you serve, since it is one of the most related factors that influence the flow of play;

- Improve your own strengths and techniques since the winning percentage also plays an important role.

We believe that our novel results should be taken into consideration for more scientific training on professional athletes. Further, more relative research should by conducted to explore other vague areas of possible factors. Here are some suggestions. Prediction for other sports like table tennis. Impact of environmental influence and the playground condition. Influence of physical condition of players. Much room of topics are to be discovered and studied.

Should you need more information, I will be glad to send you a copy of our result sheet of detailed data, together with the overview of our work.