Research On the Momentum Scoring System Model

Songtao Che

Marine Engineering College, Dalian Maritime University, Dalian, 116086, China

Abstract. In the 2023 Wimbledon men's singles final, 20-year-old Spanish rising star Carlos Alcaraz defeated Grand Slam player Novak Djokovic to claim the title. In sports, reversals and reversals often occur, while in tennis, this shift is even more pronounced due to the presence of hold and break games, a trend known as momentum. The purpose of this paper is to study the internal mechanism of momentum, understand the causes and influencing factors of momentum, so as to help players and coaches understand the game format from another perspective, so as to improve the level of competition. In response to problem one, a model was developed to capture the flow of the game when scoring occurs and apply it to one or more games. To solve this problem, build a time-series-based model, the Momentum Scoring System model, which will evaluate each point in the match, taking into account the advantage of serve and the impact of consecutive points. The goal of the model is to determine which player is performing better at any point in the game and to quantify the degree of dominance in that performance. In response to question 2, in order to evaluate the tennis coach's skepticism about the role of "momentum", a random process simulation was used to compare the actual match data with the randomly generated match results, and the randomness of the momentum transition was evaluated from the actual momentum score calculation and the random match simulation. Firstly, the momentum score is calculated based on the actual game data, then the distribution of momentum scores is generated by simulating a large number of random games, and finally the statistical hypothesis test is used to judge whether the momentum transition in the actual game is significantly different from the random case by using the null hypothesis (H0) and the alternative hypothesis (H1). In response to problem 3, in order to predict the change of momentum in a match (i.e., when the flow of the game will change from one player to another), a machine learning model is used to analyze the match data and identify possible turning point indicators. Firstly, a gradient boosting tree classifier was trained to predict the transition point of momentum in the game, and then the accuracy of the model on the test set was calculated, and the prediction performance of the model was visualized through the confusion matrix. The confusion matrix provides a comparison between the true class and the model prediction class, helping to understand how the model is performing at the point where the momentum of the prediction is turning. The goal of this paper is to evaluate the generalization ability of the previously developed model on other competition data and explore possible directions for improving the model. In terms of evaluating the generalization ability of models, we focus on data preparation, feature engineering, model testing, and performance evaluation. At the same time, the potential directions for improving the model are also analyzed: 1. More complex characteristics: player fatigue, psychological pressure at key moments of the game, etc. 2. Model tuning: Adjust model parameters according to different types of datasets, or try different machine learning algorithms. 3. Cross-motion generalization: Analyze the applicability of the model in different ball sports, and explore whether it can capture the general momentum transformation law in cross-motion.

Keywords: Tennis; Four Grand Slams; Technical and tactical measures; Prediction model; BP neural network.

1. Introduction

1.1. Problem Background

In the 2023 Wimbledon Gentlemen's final, 20-year-old Spanish rising star Carlos Alcaraz defeated 36-year-old Novak Djokovic. The loss was Djokovic's first at Wimbledon since 2013 and ended a remarkable run for one of the all-time great players in Grand Slams\(^1\). The match itself was a remarkable battle. Djokovic seemed destined to win easily as he dominated the first set 6-1 (winning 6 of 7 games). The second set, however, was tense and finally won by Alcaraz in a tie-breaker 7-6.
The third set was the reverse of the first, Alcaraz winning handily 6-1. The young Spaniard seemed in total control as the fourth set started, but somehow the match again changed course with Djokovic taking complete control to win the set 6-3[2]. The fifth and final set started with Djokovic carrying the edge from the fourth set, but again a change of direction occurred and Alcaraz gained control and the victory 6-4. The data for this match is in the provided data set. It can be seen all the points for the first set when Djokovic had the edge using the “set_no” column equal to 1[3-5]. The incredible swings, sometimes for many points or even games, that occurred in the player who seemed to have the advantage are often attributed to “momentum.” “match_id” of “2023-wimbledon-1701”. You can see all the points for the first set when Djokovic had the edge using the “set_no” column equal to 1. The incredible swings, sometimes for many points or even games, that occurred in the player who seemed to have the advantage are often attributed to “momentum.”

1.2. Our Work

Firstly, we will read the data. Additionally, in order to ensure the accuracy of the data, preprocessing operations will be conducted to analyze whether there are any missing values in the match data. When analyzing matches, it is generally preferable to select matches with fewer missing values to avoid their impact. Secondly, for the selection of evaluation indicators, a total of 10 indicators were quantified to provide a specific and detailed description of each player's performance. Based on this, we define "momentum" to analyze the performance trend of players over time. Furthermore, by using first-order differences, we obtain the differential values of the player's "momentum". Based on the differential values of momentum, peak nodes are determined using the slope method. On this basis, statistical tests are employed to analyze and demonstrate that "momentum" and a player's success are not random but have a certain correlation. Subsequently, important factors related to fluctuations in player performance during matches are identified through the established model. The model can effectively predict fluctuations in matches and provide reasonable and effective recommendations for players to make optimal adjustments when facing new opponents. Finally, by applying our established model to other matches, the results further confirm the high accuracy and good generalization ability of our proposed model.

2. Preparation of the Model

This mathematical modelling problem requires us to analyze the match data between Spanish rookie Carlos Alcaraz and Novak Djokovic in the 2023 Wimbledon men's singles final. It was observed that Djokovic had won the match by 6-1, but due to the multiple momentum variations in the subsequent matches, Alcaraz won ultimately by 6-4 in the fifth set. For this reason question arises and requires investigation to use the provided dataset (match_id=2023-wimbledon-1701) and to analyze the changes in momentum during the competition. (a): Description and quantification of the momentum: It is noted that momentum in sports competitions typically states the "power" or "momentum" gathered by a series of events, but it is hard to directly measure and is just an intellectual concept. We are inspired by this concept and now the question arises of how to explore as well as how to compute such momentum by using competition data and analyze it for different events which affect the variations in the momentum. (b): Data analysis: A vital step is to extract information on each scoring point and gather data. Such as the momentum data of the competition 2023 Wimbledon men's singles competition (except for the first two rounds) and other information about the players or game data that may need to be considered for precise quantification and analysis. This data further helps to quantify such momentum during competition, but it needs to build a model to evaluate the contribution of each scoring point of the momentum of the players. Moreover, it is worth noting that supplementary complex factors could be taken to adjust and calculate this momentum, for instance, the impact of winning streak points, important scores (such as breaking serve points), and external factors including the physical and psychological condition of the player. These factors should be explored throughout the game which are responsible for the variation in the momentum and the trends.
of the key events need also be analyzed that lead to the changes in momentum. It should be noted that the above models and formulas are a simplified example of momentum quantification. In practical applications, more complex statistical models and machine learning methods may be needed to accurately capture and predict momentum changes in competitions, including time series analysis, regression analysis, and other methods. To further analyze these issues, we can adopt a comprehensive model that combines multiple factors to quantify momentum, predict turning points in the game, and verify the universal applicability of the model.

3. Model I

Develop a model to capture the game process of scoring and apply it to one or more matches. To solve this problem, we can build a time series based model that will evaluate each scoring point in the game and consider the advantages of serving rights. The goal of the model is to determine which player performs better at any moment in the competition and quantify the degree of advantage of this performance. We can build a simplified momentum scoring system that takes into account the following factors: Scoring point: For each point earned, increase momentum score. Serving rights: The serving party has a higher probability of winning in exchange points, therefore, in the serving game, the points obtained are given a higher weight. Continuous scoring: Winning points continuously will increase additional momentum points, reflecting an overwhelming advantage in the game. Let Mt be the momentum score at time point t, Pt be the score obtained at that point (winning is 1, losing is 0), St be the serving weight (serving is 1.2, receiving is 1.0), C be the bonus of consecutive points (the first consecutive win is 1, and for each subsequent consecutive win, 0.2 is added), then:

\[
M_t = M_{t-1} + (P_t \times S_t \times C_t)
\]

In order to implement the above model and provide visualization, we first need to extract competition data from the dataset, then apply the model to calculate the momentum score at each time point, and finally generate a visual chart of momentum changes.

```python
In [1]: import pandas as pd
import matplotlib.pyplot as plt

In [2]: df[‘momentum_score’] = 0
momentum_score = 0
consecutive_wins = 0
for index,row in df.iterrows():
    P_t = 1 if row[‘winner’] == row[‘server’] else 0
    S_t = 1.2 if row[‘server’] == ’player name’ else 1.0
    if P_t == 1:
        consecutive_wins += 1
    else:
        consecutive_wins = 0
    C_t = 1 + consecutive_wins * 0.2
    momentum_score += (P_t * S_t * C_t)
    df.at[index, ‘momentum_score’] = momentum_score
plt.plot(df[‘point_no’], df[‘momentum_score’], label=’Momentum Score’)
plt.xlabel(’Point Number’)
plt.ylabel(’Momentum Score’)
plt.title(’Momentum Flow’)
plt.legend()
plt.show()
```

**Figure 1.** Competition data
4. Model II

4.1. Establish Model II

First of all, it should be noted that the athlete's momentum is likely to be related to the outcome of the game, which means we need to refute the tennis coach's viewpoint. Through analysis, it is known that his key is to doubt the role of momentum in the game, and he believes that the fluctuations of the game and the success of the players are random. Consequently, to confirm his conclusion, it is necessary to prove whether the fluctuations of the game are related to the success of the players. For this proof, it carries a strong statistical flavor.

4.2. Analysis & Results

Firstly, it is necessary to organize and obtain relevant data, as shown below

![Figure 2. Relevant data](image1)

Using first-order difference to obtain the difference in athlete momentum. Based on the difference value of momentum, the slope method is used to obtain the peak node, and a turning point image is drawn, as shown.

![Figure 3. The difference in athlete momentum](image2)

On this basis, add the turning points of p1 and p2 to the table as new variables, as shown below
Furthermore, visualize the variables of the above content as follows.

On this basis, statistical tests are used for analysis to determine whether the variables are correlated, as shown below.

```python
chi2_statistic, p_value = stats.chisquare(match_1000_data['p1_v1'], match_1000_data['point_victory'])
if p_value < 0.05:
    print('There is a significant difference between Player 1 Momentum and Point Victory.')
else:
    print('There is no significant difference between Player 1 Momentum and Point Victory.')

chi2_statistic, p_value = stats.chisquare(match_1000_data['p2_v2'], match_1000_data['point_victory'])
if p_value < 0.05:
    print('There is a significant difference between Player 2 Momentum and Point Victory.')
else:
    print('There is no significant difference between Player 2 Momentum and Point Victory.')
```

Figure 4. The turning points of p1 and p2

Figure 5. Visualize the variables

Figure 6. Statistical tests
Therefore, it can be concluded that there is not a random relationship between the two, but a significant difference. In addition, analysis can also be conducted from the perspectives of proving the randomness of the two and providing a regression relationship between the two. And for the difference in momentum, other indicators can also be used to represent it.

5. Model III

5.1. Establish Model III

Using the data provided for at least one match, develop a model that predicts these swings in the match. Firstly, by analyzing the third question step by step, it can be concluded that this question requires exploring whether it is possible to construct an indicator to analyze how the progress of the competition process is beneficial to which player. For this issue, it is actually a repetition of the previous one, because we previously defined the momentum indicator momentum, which is dynamically changing. Therefore, I personally think this indicator can evaluate the dynamic process of players in the competition process.

Because the sentence "when the flow of play is about to change from favoring one player to the other" asks when, what we did before was that dynamic momentum assessment is particularly relevant to the meaning of the question.

5.2. Analysis & Results

According to previous data, there is a significant difference between the format of the American competition and the format of the national competition. The national competition usually presents a model for each issue, and the degree of correlation between multiple issues is relatively low. For the American competition, the number of questions is similar to that of the American competition, but the American competition focuses on integration, and multiple questions may share the same model, so it is normal for multiple questions to be connected in one model. For the dependent variable, the problem discusses the fluctuations in the competition. Personally, this fluctuation is defined as the trend of a player's momentum in the competition, and can be solved by using the difference function in Python. In addition, for the independent variable X, analysis of the dependent variable indicates that some character variables need to be removed (I personally feel that the overall prediction accuracy improvement is not very high, and they are all character variables, and assignment also needs to be careful. What I basically do is cover all variables.)

Under this premise, it is possible to use X variables to predict Y variables by constructing a prediction model between X and Y, which completes what we call these fluctuations in the prediction competition. In addition, we need to explore which factors seem to be most relevant (if any). For this problem, it can actually be solved by constructing models such as random forests and decision trees, which have inherent importance in output and feature correlation. It is necessary to analyze the relationship between X and Y, that is, to train the relationship between X and Y. We can build models through machine learning methods. In addition, then the model structure can be trained through this model. In this model structure, Y can be mapped by the value of the X variable. So, here, we can use the SHAP library (a machine learning visualization library) to obtain the relationship between each feature X and Y. You can take shortcuts and use heuristic algorithms to analyze the samples in SHAP, in order to search for the corresponding combination of X values (including the values of players p1 and p2) for the highest Y situation (where player p1 has the highest momentum fluctuation).

Therefore, by obtaining the values of p1 and p2 corresponding to the specific combination, corresponding adjustments can be made in the highest case of Y.

6. Sensitivity Analysis

3.1 What factors seem most related (if any)?
3.2 Given the differential in past match “momentum” swings how do you advise a player going into a new match against a different player?

4.1 Test the model you developed on one or more of the other matches. How well do you predict the swings in the match?

4.2 If the model performs poorly at times, can you identify any factors that might need to be included in future models?

4.3 How generalizable is your model to other matches (such as Women’s matches), tournaments, court surfaces, and other sports such as table tennis.

Regarding the question, analysis shows that there is a high correlation between the fourth and third questions. Test your developed model on one or more other matches. How well do you predict the fluctuations in the competition? The fourth question requires implementing the model of the third question in the fourth question, so the third and fourth questions are related. So, our analysis shows that combining the fourth and third questions can be done directly together. Can you identify any factors that may need to be included in future models if the model sometimes performs poorly in the following section of question 4, 4.2? In terms of this question, it is actually asking whether the model can collect other additional indicators to improve the accuracy of the model, which makes sense to analyze. How can it be promoted to other competitions (such as women’s competitions), tournaments, court surfaces, and other sports (such as table tennis). For this issue, it is impossible to gather those data sources. Therefore, for this issue, we can simply generalize our momentum model, demonstrate the differences between tennis matches and other sports, or replace our model.

The third question 3.1 and 3.2 are actually building a prediction model, while the fourth question consists of three small questions. 4.1, 3.1, and 3.2 are combined together. In 4.2 and 4.3, it can be used as theoretical exploration, namely text expansion and model expansion. So, now we will implement sections 3.1, 3.2, and 4.1 of the model, and provide some relevant references in sections 4.2 and 4.3.

Firstly, the data set we have chosen is as follows:

In (32): import pandas as pd

In (36): # 洗數據的結果

In (37): # 使用dataframes的尾數的平均和中位數，並取平均數

In (38): # 使用dataframes的尾數的平均和中位數，並取平均數

Figure 7. The data set

In addition, calculate the dynamic momentum and store the results in a table.

Figure 8. The dynamic momentum

Further data analysis to prepare for constructing datasets for X and Y:
Perform data detection.
On this basis, the data is processed to complete the construction of the X and Y datasets.
Figure 12. Data processing

On this basis, multiple models are constructed for comparative analysis to predict the relationship between X and Y (3.1) and analyze which factors are relevant (3.2). Provide the prediction results of the model (4.1).

Build a dataset as follows:

```
import pandas as pd

# Load the dataset
data = pd.read_csv('dataset.csv')

# Further process the data
...
```

Figure 13. Dataset

Make predictions for multiple models as follows:

```
from sklearn.linear_model import LinearRegression

# Prepare the training and testing data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Fit the model
model = LinearRegression()
model.fit(X_train, y_train)

# Predict the target variable
predictions = model.predict(X_test)
```

Figure 14. Predictions for multiple models

Further visualize the above results by drawing visual images.

Figure 15. Visual images
### 7. Index Importance Ranking

#### Table 1. Index Importance Ranking

<table>
<thead>
<tr>
<th>Table</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1.p1_distance_run</td>
<td>0.092041</td>
</tr>
<tr>
<td>Table 2.p1_m</td>
<td>0.090838</td>
</tr>
<tr>
<td>Table 3.p2_distance_run</td>
<td>0.082944</td>
</tr>
<tr>
<td>Table 4.p1_S</td>
<td>0.082196</td>
</tr>
<tr>
<td>Table 5.p2_B</td>
<td>0.071096</td>
</tr>
<tr>
<td>Table 6.p1_B</td>
<td>0.068353</td>
</tr>
<tr>
<td>Table 7.p2_S</td>
<td>0.064535</td>
</tr>
<tr>
<td>Table 8.p2_m</td>
<td>0.043718</td>
</tr>
<tr>
<td>Table 9.p2_T</td>
<td>0.043633</td>
</tr>
<tr>
<td>Table 10.p2_score</td>
<td>0.041032</td>
</tr>
<tr>
<td>Table 11.p1_T</td>
<td>0.032407</td>
</tr>
<tr>
<td>Table 12.p1_score</td>
<td>0.030478</td>
</tr>
<tr>
<td>Table 13.point_no</td>
<td>0.026334</td>
</tr>
<tr>
<td>Table 14.game_no</td>
<td>0.023792</td>
</tr>
<tr>
<td>Table 15.p1_games</td>
<td>0.023387</td>
</tr>
<tr>
<td>Table 16.p1_points_won</td>
<td>0.021057</td>
</tr>
<tr>
<td>Table 17.p2_points_won</td>
<td>0.020811</td>
</tr>
<tr>
<td>Table 18.rally_count</td>
<td>0.019119</td>
</tr>
<tr>
<td>Table 19.p2_games</td>
<td>0.016263</td>
</tr>
<tr>
<td>Table 20.p2_winner</td>
<td>0.011043</td>
</tr>
<tr>
<td>Table 21.p1_break_pt_missed</td>
<td>0.010997</td>
</tr>
<tr>
<td>Table 22.p2_ace</td>
<td>0.008494</td>
</tr>
<tr>
<td>Table 23.point_victor</td>
<td>0.007967</td>
</tr>
<tr>
<td>Table 24.serve_no</td>
<td>0.007713</td>
</tr>
<tr>
<td>Table 25.p1_ace</td>
<td>0.007498</td>
</tr>
<tr>
<td>Table 26.p2_net_pt_won</td>
<td>0.007371</td>
</tr>
<tr>
<td>Table 27.game_victor</td>
<td>0.005959</td>
</tr>
<tr>
<td>Table 28.p2_net_pt</td>
<td>0.005870</td>
</tr>
<tr>
<td>Table 29.p2_unf_err</td>
<td>0.005837</td>
</tr>
<tr>
<td>Table 30.p1_net_pt</td>
<td>0.005553</td>
</tr>
<tr>
<td>Table 31.server</td>
<td>0.003905</td>
</tr>
<tr>
<td>Table 32.p1_winner</td>
<td>0.003198</td>
</tr>
<tr>
<td>Table 33.set_no</td>
<td>0.002967</td>
</tr>
<tr>
<td>Table 34.p1_sets</td>
<td>0.002274</td>
</tr>
<tr>
<td>Table 35.p2_sets</td>
<td>0.001704</td>
</tr>
<tr>
<td>Table 36.p1_unf_err</td>
<td>0.001669</td>
</tr>
<tr>
<td>Table 37.p1_break_pt</td>
<td>0.001569</td>
</tr>
<tr>
<td>Table 38.p1_net_pt_won</td>
<td>0.000967</td>
</tr>
<tr>
<td>Table 39.p2_break_pt</td>
<td>0.000714</td>
</tr>
<tr>
<td>Table 40.p2_break_pt_won</td>
<td>0.000613</td>
</tr>
<tr>
<td>Table 41.p2_double_fault</td>
<td>0.000565</td>
</tr>
<tr>
<td>Table 42.p1_double_fault</td>
<td>0.000543</td>
</tr>
<tr>
<td>Table 43.p2_break_pt_missed</td>
<td>0.000477</td>
</tr>
<tr>
<td>Table 44.set_victor</td>
<td>0.000438</td>
</tr>
<tr>
<td>Table 45.p1_break_pt_won</td>
<td>0.000061</td>
</tr>
</tbody>
</table>
References


