Analysis of Race Momentum and Its Influential Factors Based on Logistic Regression and Chi-Square Test

Xia Zhang *, Jiajun Li, Ziheng Zhou

Department of Information Engineering, WuHan College, WuHan, China, 430212

* Corresponding Author Email: 13409870810@163.com

Abstract. With the rise of sports big data and advanced statistical methods, there is a burgeoning interest in understanding the psychological and tactical impacts of game momentum on match outcomes. This study presents a model that integrates logistic regression with the chi-square test to analyze sports game trends. The logistic regression model, chosen for its effectiveness in binary outcome prediction, was employed to identify key factors influencing momentum, such as serving status and unforced errors. These factors were quantified through a momentum calculation function that reflects player performance in real time. The model's validity was confirmed by a chi-square test, which provided a p-value of 0.002, indicating a statistically significant correlation between momentum and match outcomes. This analytical framework not only enhances our understanding of sports dynamics but also offers practical insights for coaches and strategists, equipping them with data-driven tools to optimize game tactics and leverage momentum effectively.

Keywords: Logistic Regression, Chi-Square Test, Momentum, Sports Analytics.

1. Introduction

In recent years, tournament momentum, as an important psychological and tactical factor affecting the game's outcome, has been a hot spot in sports science research. Momentum [1] is not only reflected in the scoring changes during the game but also in the psychological state and performance of athletes and teams. How to quantitatively analyze the relationship between momentum and tournament results, as well as how to assess its influencing factors statistically, is a challenge in current sports statistics research. With the rise of big data in sports and the development of statistical methods, researchers are increasingly inclined to use quantitative models to analyze and predict [2] the outcomes of sports events.

For this study, a logistic regression algorithm [3] is utilized to analyze the factors that impac-t momentum and to determine which factors have the most significant impact. The logistic regression algorithm [3] was proposed by the Belgian mathematician Pierre-Franc, ois Verhulst. As a well-established algorithm in statistics and machine learning, the algorithm is now wide-ly used in classification problems in modern machine learning.

In this study, to determine whether the factors affect momentum, a logistic regression model was built to make the judgment. Then, to make the factors reflect the degree of influence on momentum more intuitively, we established a function to calculate the magnitude of momentum. Subsequently, based on this function, we proved that momentum did affect the game through the chi-square test [4].

2. Data Preprocessing and Preparation

The data for this study was obtained from https://www. comap.com/contests/mcm-icm. To facilitate the subsequent model construction, extensive coding work was undertaken on the substantial dataset. As depicted in Table 1, due to the presence of a significant amount of textual data in the original dataset, additional steps were taken to make the subsequent modeling work more convenient.

Table 1. Data Encoding Table.

Data name	Encoding	Data name	Encoding	Data name	Encoding
AD	50	W	1	CTL	1
AD	50	BW	2	NCTL	2
0	1	BC	3	D	1
F	2	C	4	ND	2
B	3	B	5	NA	0

3. Analysis of Match Momentum Based on Logistic Regression

To determine which of the two athletes performed better at a certain point in the game and provide a visual graphic describing the trend of the game, the paper first created a new metric called Momentum to measure changes in athlete performance. Then, after conducting a data rationality analysis, the paper established a logistic regression model to obtain the important factors affecting Momentum. Ultimately, by establishing the functional relationship between Momentum and significant indicators, the paper determined the Momentum value and produced a graph.

3.1. Data Analysis

Before establishing the model, the study first conducted a chi-square [4] test to test the validity of the model. Ultimately, the study obtained a p-value of the chi-square test to be 0.002, which is less than 0.05, indicating that the model is effectively established. Consequently, the logistic regression [5] model was utilized to analyze the impact of each game indicator on player momentum. Throughout the statistical analysis process, a p-value of less than 0.05 was considered indicative of a statistically significant [6] difference.

3.2. The Establishment of Logistic Regression Model

The study aims to identify the most significant indicators that influence player momentum and to establish a calculation formula for momentum, using the magnitude of the momentum value to reflect the performance of the player. To obtain the significant indicators that affect momentum, the study initially considered the regression model. Since the impact of each indicator on momentum is binary, either yes or no, the study decided to construct a two-class logistic regression model. The dependent variable y is used as a dichotomous variable, and the independent variable x represents the influencing factor variable. The formula for the final probability P is as follows:

$$P(y=1|x)=\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} \tag{1}$$

After performing logic transformation on the probability P, the linear logistic regression [7] Expression of the probability P is obtained as follows:

$$\ln\left(\frac{P}{1-P}\right)=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_mX_m \tag{2}$$

After calculating the formula, this study obtained the significance P value of each indicator and visualized it as shown in Figure 1.

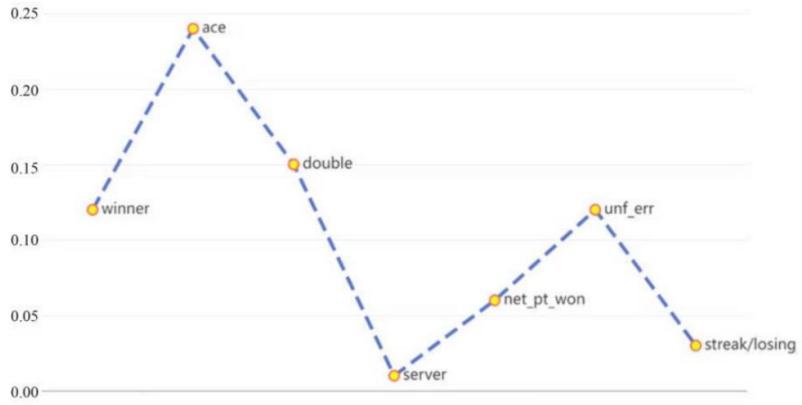


Figure 1. Analysis of the significance P value of each indicator.

From the above analysis, it can be determined that Server, Streak/losing, Net points won, and Unforced error are important indicators that affect momentum. This study establishes the functional relationship between the aforementioned indicators and momentum. The established momentum calculation function is as follows:

$$Mt = Mt - 1 + \left(Pt \times St \right) + \eta_1 Ct + \eta_2 Ut + \eta_3 Nt \tag{4}$$

In the formula, Mt is the momentum score at time point t, Pt is the score obtained at that point (win is 1, loss is 0), St is the serve weight (The serving side is 1.2, the receiving side is 0.8), Ct is the bonus for consecutive points (the first winning streak is 1, and each subsequent winning streak adds 0.2; in contrast, the first losing streak is -1, each subsequent losing streak decreases by 0.2).

3.3. Results

The above analysis allows this study to derive the change in momentum of the two players during the game. Since the magnitude of the value of the momentum indicator can be seen as an indicator of how well the players performed, this study considers it a reflection of their in-game performance. At the same time, recognizing that the good and bad performances of the two players are opposite to each other, this study only presents the momentum change of one player. As shown in Figure 2, this study illustrates the momentum [3] change of Djokovic during the match.

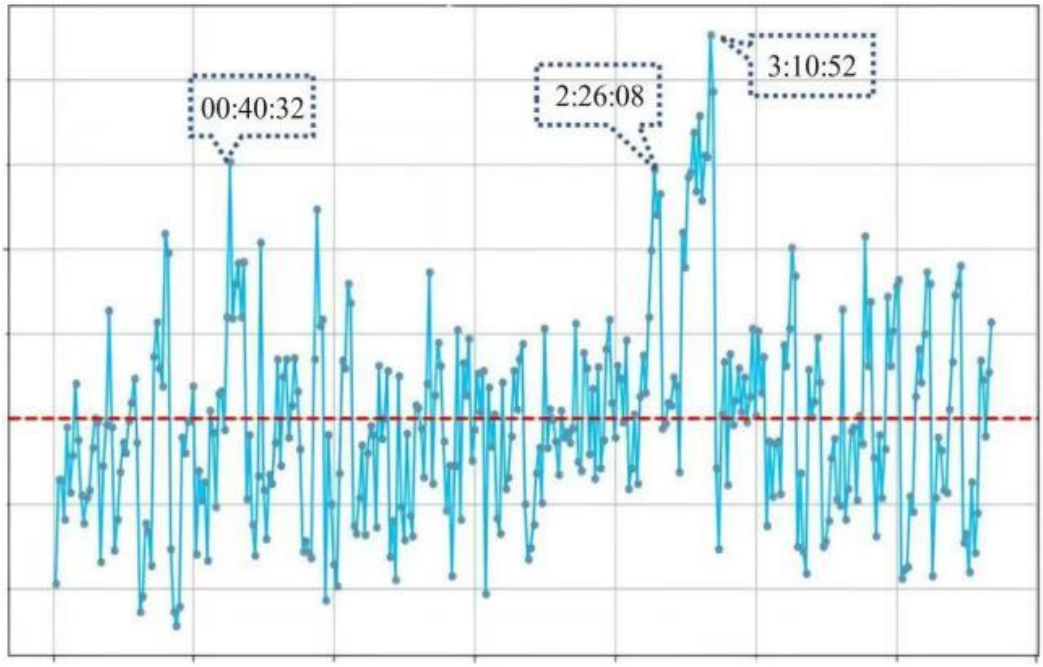


Figure 2. Momentum changes for player 1

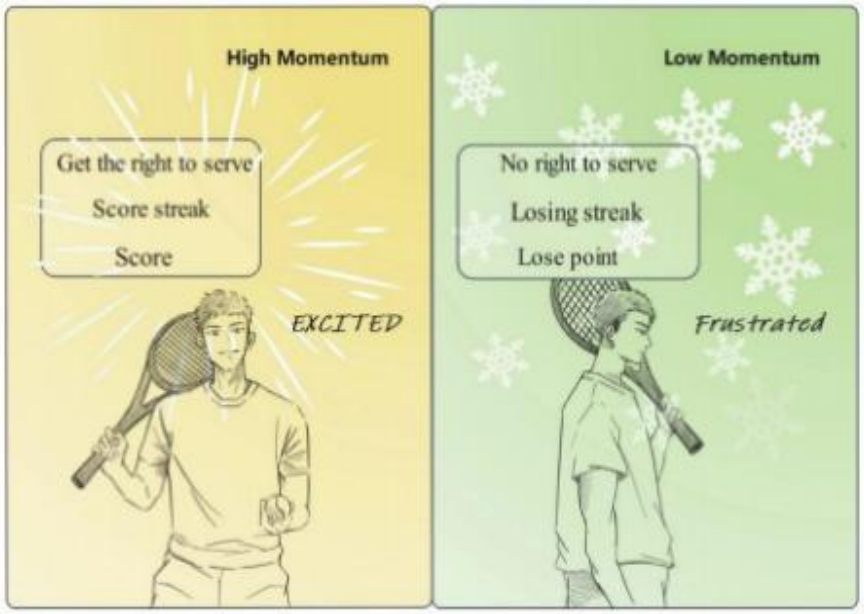


Figure 3. Relationship between Player Performance and Momentum Indicators.

By analyzing the combination of Player 1's momentum changes and the game situation, the study found that Player one ended up winning when his momentum was at a high value. The study randomly selected three-time points during the game when player one's momentum was high for further analysis. For example, the diagram above marks: when the time node is 00:40:32, player one wins the first small game in the second set; when the time node is 2:26:08, player one wins the fourth game of the third set; and when the time node is 3:10:52, player one wins the first game of the fourth set. The study considers the situation to be in player 1's favor when player 1's momentum value increases. Conversely, when Player 1's momentum value decreases, the situation is favorable for Player 2. Finally, the study visualizes the relationship between player performance and momentum indicators for intuitive understanding. As shown in Figure 3 above.

4. Statistical Validation of Momentum Effects

4.1. Description of the Algorithm for Correlation Analysis Based on Chi-square Test

Before establishing the chi-square test model, the study proposes the null hypothesis, assuming no correlation between the two variable factors. The study measures the deviation between observed and expected frequencies using residuals, which are squared and summed to normalize their impact. This normalization accounts for the relativity of residuals; larger expected frequencies diminish their apparent impact, while smaller ones amplify it. The study uses this method to evaluate the correlation strength: a significant deviation and a large normalized value suggest a correlation, leading to the rejection of the null hypothesis. Conversely, a small deviation and normalized values uphold the null hypothesis, indicating no correlation.

From the above analysis, this paper can obtain the mathematical definition formula for the correlation analysis of factors using the chi-square test as follows:

$$X^2 = \sum_{j=1}^{r} \sum_{j=1}^{c} \frac{\left(f_{ij}^0 - f_{ij}^e \right)^2}{f_{ij}^e} \tag{4}$$

In the above formula, the number of rows in the contingency table, and the number of columns in the contingency table indicate the frequency of observation and indicates the desired frequency. The calculation formula for the expected frequency is as follows:

$$f^e = \frac{RT}{n}\frac{CT}{n}n = \frac{RTCT}{n} \tag{5}$$

The chi-squarestatistic quantifies the difference between observed and expected frequencies, indicating the correlation between variables. If the observed frequency matches the expected, the statistic is zero, suggesting independence. A greater divergence suggests a stronger correlation. The test's degrees of freedom, calculated by multiplying the difference of one from each factor's count, contextualize the results.

Finally, after this study finds the chi-square statistic and the degrees of freedom between factors, this study can determine the P-value of the chi-square statistic. If the P-value is less than or equal to the significance level, this study can determine that there is a correlation between the factors.

4.2. Analytical Modeling Process Based on Chi-square Test

According to the basic concepts and algorithms of chi-square test correlation analysis, after checking that the data meets the modeling requirements of the chi-square test, two factors were selected for the chi-square test to determine their degree of correlation This study selects two attribute factors—victory or defeat outcomes and athletes' end-of-competition momentum values—and proposes a null hypothesis of their independence. By analyzing match data, this study calculates the cross-values for each attribute dimension to obtain RT and CT values. Using equation (5), this study computes the expected frequency for each row and columncrossover. The observed and expected frequencies are then aggregated to derive the chi-square statistic and its frequency values through equation (4).

With a predefined significance level of 0.01, this study establishes the critical chi-square value corresponding to the degrees of freedom. The study then assesses whether the calculated chi-square statistic is less than or equal to this critical value. If it is, the null hypothesis of no correlation between the two factors is rejected.

The flow chart of the algorithm for correlation analysis of the chi-square test is shown in Figure 4.

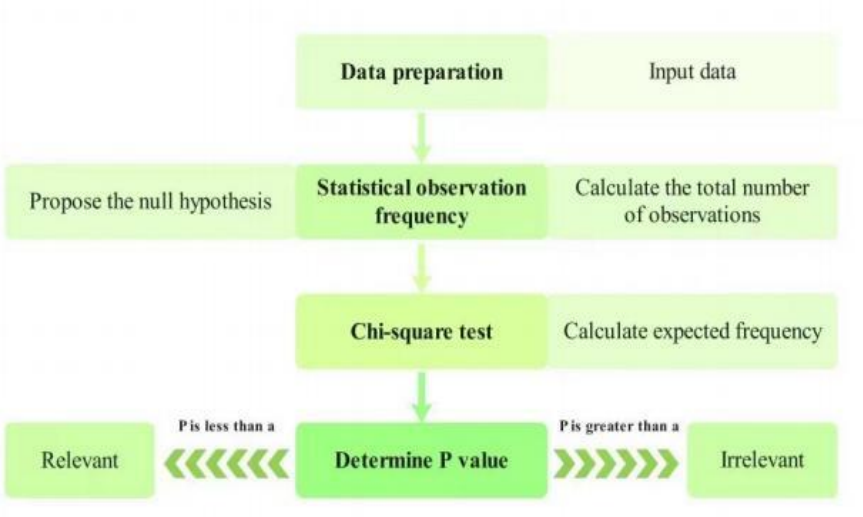


Figure 4. The process of modeling based on Chi-square test.

4.3. Calculation and Results of Factor Correlation Analysis Based on Chi-square Test

Referring to the modeling steps above, the study selected the momentum values of the athletes at the end of the race to perform a chi-square test with the results of the race. The specific calculations are as follows:

The level of significance and the critical value at which the hypothesis of predetermined judgment holds. When the degree of deviation between the calculated observed and expected frequencies is large, the calculated value exceeds a predetermined level of significance, or the observed value of the statistic is less than the critical value, the study rejects the previously proposed null hypothesis and

determines that the two variable factors are correlated with each other. When the deviation between the calculated observed frequency and the desired frequency is small, the calculated value falls below the predefined critical value. Or when the observed value of the statistic is greater than the critical value, then the null hypothesis previously proposed cannot be rejected and it is judged that there is no correlation between the two variable factors. The study analyzed and calculated for all 31 games and obtained the relevant data presented in the following table:

Table 2. Chi-square test analysis results.

P1_momentum	-6.0	-4.8	-3.6	-2.4	2.4	3.6	4.8	6.0	X^2	P
Win	0	0	1	0	4	5	3	1	23.753	0.00126
Fail	1	2	2	4	1	0	0	0		

Based on the data from Table 2, we perform a heat map analysis of the relationship between momentum and race results.

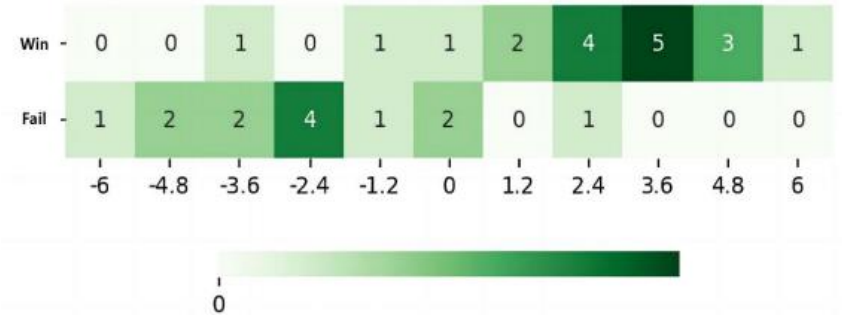


Figure 5. Heat map analysis of momentum and race results.

Drawing from Table 2 and Figure 5, it is evident that when the momentum value is less than 0, only 3 games were won. Conversely, when the momentum value exceeds 0, only a single game is lost. This leads to the conclusion that momentum significantly influences the outcome of the match.

5. Conclusions

In this thesis, a comprehensive computational system was constructed to analyze the momentum effect in tennis matches and to predict the outcome of the matches. Calculations based on a logistic regression model found that factors such as serving, scoring, and consecutive scoring had a significant effect on momentum. Then the analysis of the chi-square test yielded a Ceammer's V value of 0. 875 for momentum on the outcome of the match, thus determining that momentum has a strong effect on the outcome of the match. Then in the study, we constructed an XGBoost [8-10] model based on the eigenvalue index to predict the trend of the game and combined it with the LSTM [11] model to increase the prediction accuracy to 92.

The study of momentum in sports is highly applicable to real-world coaching strategies, offering valuable insights into how coaches can organize game tactics to capitalize on a team's psychological and physiological states. Future research directions could focus on leveraging big data and machine learning to predict momentum shifts, which would allow for more strategic in-game adjustments. Additionally, exploring gender differences and cultural influences on momentum perception could lead to more personalized and effective coaching methods. The concept of momentum coaching also presents an opportunity for athletes to build resilience and maintain a competitive edge throughout their careers. Overall, the integrationo-f these research areas has the potential to significantly enhance sports performance and strategy development.

References

[1] Zhanneta Kozina et al. “General and individual factor structure of complex preparation of young tennis players of 10-12 years” [N]. In: Journal of Physical Education and Sport 20 (2020), pp. 1242–1249. 22

[2] Hartonen, tuomo, Jermy, Bradley, Sõnajalg, Hanna, Vartiainen, Pekk Krebs, Kristi, Vabalas, Andrius Metspalu, Andres Esko, tõnu Nelis, Mari Hudjashov, Georgi Leino, tuija Nohynek, Hanna Sivelä, Jonas Mägi, Reedik Daly, Mark Ollila, Hanna M Milani, Lili Perola, MarkusRipatti, Samuli Ganna, Andrea FinnGen Estonian, Biobank Research Team. Nationwide health, socio-economic and genetic predictors of COVID-19 vaccination status in Finland [J]. Nature Human Behaviour. 2023, 7(7):1069-1083.

[3] K., He, C., He. Housing Price Analysis Using Linear Regression and Logistic Regression: A Comprehensive Explanation Using Melbourne Real Estate Data[C]// IEEE International Conference on Computing (ICOCO), 2021:241-246.

[4] Qun Li, HongJian Xu, Jin Yang, LinKang Wang, JingYi Sun, GuangChen Zhang. Debris Flow Susceptibility Evaluation in the Bomi-Medog Area Based on Pearson Chi-Square Test Algorithm for Index Selection [J]. Bulletin of Geological Science and Technology, 2024:16.

[5] B., N B JS. YadhukrishnanM. A. A Comparative Study on Document Images Classification using Logistic Regression and Multiple Linear Regressions[C]//Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), 2023:1096-11042

[6] Argaw-Denboba, Ayele Schmidt, thomas S B Di Giacomo, Monica, Ranjan, Bobby Devendran. Paternal microbiome perturbations impact offspring fitness [J]. Nature, 2024, 629(8012):652-659.

[7] G., Khekare, K., Balaji, M., Arora, R., R Tirpude, B., Chahar, A., Bodhankar. Logistic and Linear Regression Classifier Based Increasing Accuracy of Non-Numerical Data for Prediction of Enhanced Employee Attrition[C]//2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2023:758-761.

[8] Zhang, Biao, Zhang, Ying, Jiang, Xuchu. Feature selection for global tropospheric ozone prediction based on the BO-XGBoost-RFE algorithm [J]. Scientific Reports, 2022, 1(12):9244.

[9] Y., MoY. MeiY. , Chen. Used Sailboat Price Prediction Based on Hybrid PSO-XGBoost Algorithm[C]. 2023 International Conference on Algorithms, Computing and Data Processing (ACDP), 2023:258-262.

[10] Jiang, Jiuzhou, Pan, Hao, Li, Mobai, Qian, Bao, Lin, Xianfeng, Fan, Shunwu. Predictive model for the 5-year survival status of osteosarcoma patients based on the SEER database and XGBoost algorithm [J]. Scientific Reports, 2021, 1(11):5542.

[11] A., AkandehF, M Salem, Slim LSTM NETWORKS: LSTM_6 and LSTM_C6[C]//2019 IEEE 62nd International Midwest Symposium on Circuits and Systems (MWSCAS), 2019:1558-3899.