Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 6
ISET 2024

Prediction of tennis match results based on logistic regression

Xinyue Wang*

Information School of YNUFE, Yunnan University of Finance and Economics, Kunming, China, 650032

*Corresponding author: wangxinyue8449@sina.com

Abstract. More and more people like tennis, and the demand for the accuracy of predicting the results of tennis matches in sports betting is getting higher and higher. In this paper, a model is built to predict the results of tennis matches. In this paper, we first use the random forest model to analyze the factors that will have an impact on the outcome of the game, and the authors have selected out five factors that may have an impact on the outcome, which are: break of serve, Serve, Score, Net points , And they were assigned proportionally and the distribution was Break of serve 34.10%, Serve 24.30%, Score 38.00%, Net points 3.60% and using these proportions the influencing factors were calculated numerically, in this paper, the calculated value is named “momentum”. Then Pearson's correlation coefficient was used to analyze the relationship between momentum and match results, and the analysis found that the correlation between momentum and match results was as high as 0.807. Finally, a logistic regression model was used to predict and test the results of tennis matches. The final test results found that the model established in this paper has a high accuracy rate, and the research results also provide a new idea for tennis betting industry.

Keywords: Prediction Model, Logical Regression, Random Forests.

1. Introduction

The Wimbledon Tennis Championships has always enjoyed a high status in the tennis game. The men's singles final of the tournament has been highly anticipated and watched by spectators for years. In the 2023 men's singles final, 20-year-old Spanish star Alcaraz defeated Novak Djokovic to end the tennis great's record. The match had many twists and turns as the two players traded the lead before Alcaraz narrowly won the match. It was a scorching and intense match, and the superb mental strength displayed by both men during the race was admirable. In sports, there is a term used to describe the force or forces gained through motion or a series of events called "momentum". We can analyze the factors that have an effect on momentum and then materialize the momentum and calculate the value of the momentum. Discussing and understanding a player's momentum over the course of different games is important for understanding the outcome of the game, the athlete's own adjustments, and the coach's assigned tactics.

Wilkens, Sascha found that all models improve the tennis rankings of both players as the only indicator of match prediction, but fail to go beyond simple speculative predictions of betting odds, where most of the relevant information is embedded in the betting markets, and the addition of other match- and player-specific data does not lead to any significant improvement [1] .Gollub, Jacob switched the usual prediction methods and used Markov hierarchical modeling of athletes' performance to predict the outcome of the game, greatly improving the accuracy of the predictions [2]. Gao Zijian, Kowalczyk, Amandac used the random forest model to analyze the results of club tennis matches, the environment, and the physical fitness of the mobilizers, and finally identified the strength of the serve as a key predictor of the results of tennis matches [3]. Aisha Fayomi, Rizwana, Majeed, Ali Algarni, Sohail Akhtar, Farrukh Jamal, and Jamal Abdul Nasir evaluated the probabilistic predictions using the Bradley-Terry model with four performance indicators. The results showed that the playing surface had a significant impact on player performance, and finally the study found that the fitted model was more accurate for red clay field data [4]. Ying Zhu and Ruthuparna Naikar believe that the first serve, is very important in professional tennis. Therefore, a machine learning

method was developed to predict the first serve direction of professional tennis players. Through feature engineering, this method can predict the direction of the player's serve more accurately, which provides a greater help for later prediction [5]. Alberto Arcagni, Vincenzo Candila and Rosanna Grassi proposed a metric based on the centrality of feature vectors, and then the generated ratings were used as covariates in a simple logit model, and this method also achieved positive betting results [6]. Tristan Barnett, Stephen R. Clarke uses statistics to predict the serve statistics to be obtained when two given players meet. These statistics are then used in a spreadsheet model to predict the outcome of further matches, and these calculations can be updated as the match progresses, substantially increasing the accuracy of the predictions [7]. Leighton Vaughan Williams EMAIL logo, Chunping Liu, Lerato Dixon and Hannah Gerrard used quasi-Elo ratings, surface-specific Elo ratings, and a weighted synthesis of these ratings to add weight to the wider use of Elo in player ranking methods [8].

In previous studies, the prediction results may not be accurate enough and the occurrence of the turning point of the game is not clear, etc. In our study, based on the analysis of the athlete's "momentum", we clearly calculated the probability of winning the game at each point in time and the possible turning point of the game, which makes the prediction results clearer and more explicit.

2. Model

2.1. The structure of Random Forest Feature Importance Analysis

Random forest is an integrated learning method that consists of multiple decision trees, each of which is independently trained by random sampling and random feature selection. The Random Forest Importance Analysis model is an important component of the Random Forest model and can be understood as the interpretation and analysis of the Random Forest model. Random Forest Importance Analysis is the process of evaluating the importance of individual features of the evaluation model to understand which features have a greater impact on the results, so that the features can be filtered and subsequently analyzed.

There are two basic principles of Random Forest importance analysis, one is average impurity reduction and the other is replacement feature importance. The one we will focus on is replacement feature importance.

The steps of displacing feature importance are:

(1) disorganize the data of feature values to destroy the connection between features and targets

(2) Run the model on the disrupted data and calculate the performance metrics.

(3) If the model performance decreases significantly after the replacement, it means that the feature is very important to the model prediction; if the performance does not change much after the replacement, the feature has less impact on the model.

$$I_F = P_{original} - P_{permuted} \tag{1}$$

where $P_{orginal}$is the performance of the undisplaced test data, where I_F is a characterization of the test data, where $P_{permuted}$ is the performance of the trained random forest model.

$$\bar{I_F} = \frac{1}{N}\sum_{i=1}^{N}\left(P_{original} - P_{permuted_i}\right) \tag{2}$$

where N is the number of calculations， where $P_{permuted_i}$ is the performance after the is the substitution.

2.2. Prediction of tennis match results based on logistic regression modeling

Logistic regression modeling is a widely used method of regression analysis in statistics, which is a model for predicting the probability of a binary target variable (e.g., yes/no, success/failure) based on a categorical model of logistic functions. Logistic regression accomplishes this by passing the output of a linear regression to a logistic function, which maps the predicted values to probabilities between 0 and 1.

The steps of logistic regression modeling are:

(1) The logistic regression model first requires the computation of the weighted sum of the characteristics (independent variables)

(2) Determine a decision threshold (usually 0.5) to decide the classification. If P(Y=1)>0.5, then predict Y=1; otherwise predict Y=0.

(3) The parameters of the logistic regression model are computed by estimating to calculate the parameters of the logistic regression model through maximum likelihood estimation (MLE), a process that seeks to find the set of parameters that maximizes the probability of occurrence of the observed sample data.

(4) The calculated parameters are brought into the logistic function formula, and the output of the linear regression is converted into a probability (between 0 and 1) by means of the logistic function.

(5) The model is evaluated and optimized to prevent overfitting of the model.

$$Z=\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n \tag{3}$$

where Z is the weighted sum of the features, $\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n}$ are independent variables (characteristic),$\beta_0,\beta_1,\beta_2,\ldots,\beta_n$ are the model parameters, where β_0 is the intercept and β_1 to β_1 are the coefficients of each feature.

$$L(\beta)=\prod_{i=1}^{m} P_i^{y_i}(1-P_i)^{1-y_i} \tag{4}$$

where L(β) is the joint probability of occurrence of all sample data observations for maximum likelihood estimation.

$$\ell(\beta)=\sum_{i=1}^{m}\left[y_i \log(p_i)+(1-y_i)\log\left(1-p_i\right)\right] \tag{5}$$

where L(β) and ℓ(β) are the likelihood and log-likelihood functions, respectively, used for parameter estimation, where ℓ(β) is the logarithmic form of L(β), which is usually used to simplify the calculation, m is the sample size, where y_i is the actual category label of the sample (0 or 1), where P_i is the probability that the model predicts y_i to be 1 given X_i.

$$P=\sigma(Z)=\left(\frac{1}{1+e^{-z}}\right) \tag{4}$$

2.3. Improvement of the Model Based on Regularization

Regularization is a technique to reduce model overfitting by adding a penalty term to the loss function to constrain the complexity of the model. In logistic regression, regularization usually involves adjusting the loss function by adding a term that is related to the size of the model weights (parameters). The two common forms of regularization are:

(1) Lasso regularization, ridge regularization, Elastic Net regularization.

(2) In mathematical modeling process we use Ridge regularization.

Ridge regularization works by adding the sum of squares of the weights. This method tends to distribute the weights evenly across all features rather than eliminating some weights altogether. This helps in handling those cases where there is a high degree of correlation between features. The loss function for Ridge regularization can be expressed as

$$L(\theta)+\lambda \sum_{j=1}^{n} \theta_j^2 \tag{5}$$

where λ controls the strength of the regularization. Ridge Regularization without d1 has the advantage that it does not lead to sparse solutions, it does not make the weights zero. In our regular regression, we set λ to 0.8, the effect of regularization is stronger, the model will be more inclined to keep the weights smaller, thus reducing the model complexity, which can be very effective in order to prevent overfitting.

3. Results

3.1. Selection of eigenvalues

First of all, the factors that have the potential to have an impact on the outcome of the tennis match are analyzed, and this paper has selected five aspects, which are as follows feature name, break of serve, Serve, Score, Net points, for the following reasons:

First of all, the advantage that the serve can establish in a tennis match is significant. Whether it is a flat serve on the first serve or a topspin serve on the second serve, it puts more pressure on the opponent, thus favoring the player in the match.

Accordingly, if a player fails to hold his/her own serve and is broken by his/her opponent, and the immediately following game is the opponent's serve, then the player will be under more pressure in that game.

Secondly, whether a point is scored or not also has a great impact on the outcome of the match.

Finally, due to the diversity of tennis rules, when a player adopts an aggressive net playing strategy, it is equivalent to actively compressing the opponent's reaction and hitting time, so that the opponent can not be ready and lose points, causing pressure on the opponent, so the net scoring is also included in the analysis.

So, in this paper, these four influencing factors are selected.

3.2. Significance analysis results

We have built a model to synthesize the various influences in order to better predict the outcome of the matches, and to predict the outcome of each match based on the final synthesized result. We define the final synthesized result as momentum.

We subjected the individual influences to a randomized forest significance analysis. The results are shown in Table 1.

Table. 1. Proportion of each influencing factor

Feature name	Feature Importance
Break of serve	34.10%
Serve	24.30%
Score	38.00%
Net points	3.60%

3.3. Analyze the momentum calculation model and the results of the calculation

First of all, we have to determine whether there is a relationship between momentum and the result of the game. In this paper, we first randomly select a few games data as an example, use Pearson correlation coefficient, calculate the momentum of their games and analyze the correlation between momentum and the result of the game, and then visualize the result of their analysis by using a heat map. The results are shown in Figure 1.

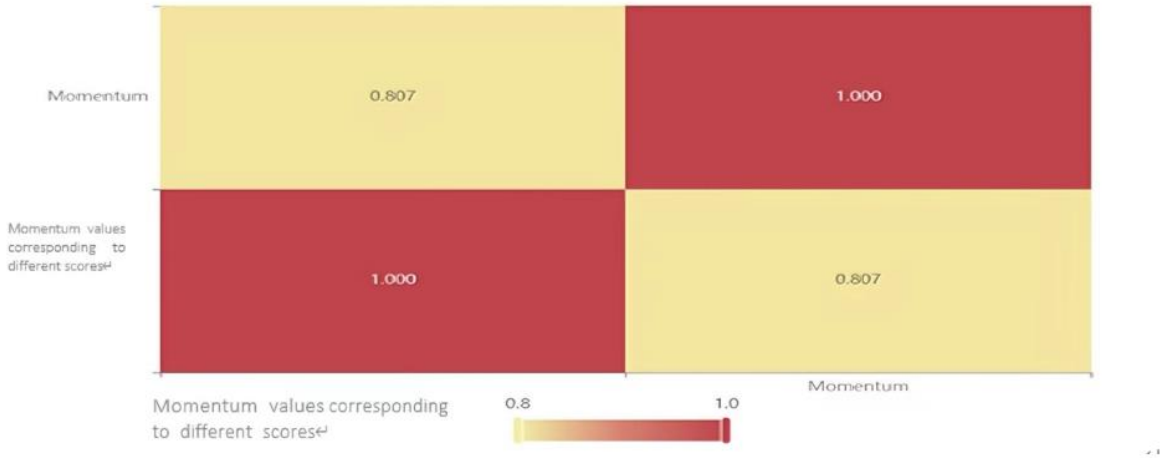

Figure 1 Momentum plotted against race results

As shown in Figure1, momentum and race results are highly correlated, and in the next step we will analyze whether they are positively or negatively correlated. We will still select a few races and use Matlab to fit the function. The fitting results are shown in Table 2.

Linear model Poly1: $\mathbf{f(x)} = P_1 * \mathbf{x} + P_2$

Table 2 Fitting results

parameters	estimated value	95% upper confidence limit	95% lower confidence limit
p1	-1.267	-1.742	-0.7933
p2	0.003069	0.002965	0.003173

Graphical Description:

The Figure 2 shows the results of fitting the model, including the estimation results of each parameter and the upper and lower confidence limits of the estimation (indicating that there is a 95% probability that the parameter falls in the interval [upper confidence limit, lower confidence limit]), from which the estimation values can be used to obtain the formula of the model.

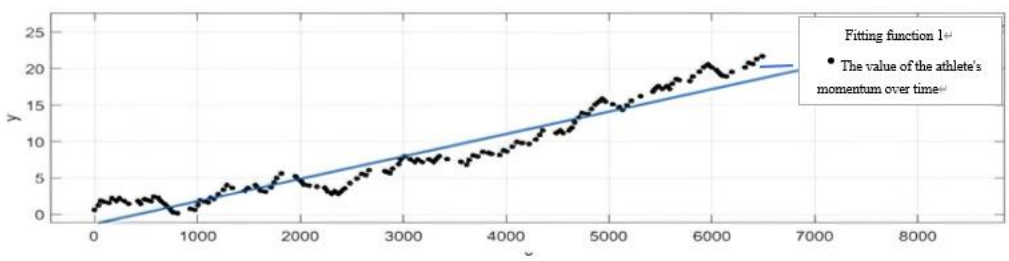

Figure 2 Fitting effect diagram

The Table 3 shows a plot of the fit of the model, if the original scatterplot are close to the fitted curve or the fitted surface, it means that the fit is better.

Table 3 the results of the fit evaluation indicators of the model

R^2	RMSE
0.9539	1.667

Chart Description:

Table 3 shows the results of the fit evaluation indicators of the model.

R^2: Coefficient of Multiple Determination, the closer the value is to 1, it indicates that the variables of the equation explain y better.

RMSE: Root Mean Square Error, the closer the value is to 0 the value indicates less error and better prediction.

Table 3 shows that the linear equation has a good fit.

3.4. Predictive Modeling of Tennis Match Outcomes with Logistic Regression Modeling

Using logistic regression models, we built a model to make predictions about the outcome of tennis matches. Figure 3 show the actual and predicted win probabilities, with the red dots representing turning points in the predicted match outcomes. This provides a clear visualization of the fit of the predicted values to the true values and demonstrates the accuracy of the model's predictions. In the model, we also use regularization to prevent overfitting of the model.

In the Figure 4, the top graph is the actual winning probability, the bottom graph is the predicted winning probability and the marked red points are the turning points in the game, i.e., where the change in the flow of the game occurs.

The data in Table 4 shows the accuracy, precision, recall and F1 score, i.e. the four evaluation metrics of the logistic regression model, from which we can conclude that the model predicts the matching data well.

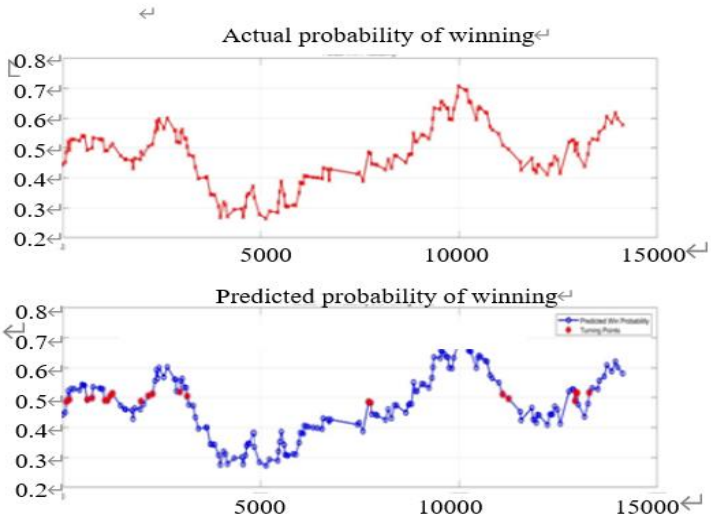

Figure 3 Comparison of actual results with projected results

Table 4 Model Accuracy Analysis

Accuracy	Recall	Precision	F1
0.92	0.831	0.912	0.841

4. Conclusions

This paper establishes a model to predict and analyze the tennis results. Firstly, this paper analyzes the factors that will have an important influence on the results of tennis matches by means of the logistic forest model, through the model analysis, four influencing factors are selected, and these influencing factors are ranked according to the size of their influence. Then a model is built to calculate the four influencing factors, and the result is called “momentum”. In this paper, a model for predicting tennis matches is built by logistic regression model, and the model is regularized in order to prevent the model from overfitting. Finally, through testing, the model is found to be able to predict the results of the match more accurately, providing an idea for the prediction of tennis matches in the sports betting market. However, since this paper uses data from men's tennis matches, the model may not be applicable to women's tennis matches or other ball games.

References

[1] Wilkens S. Sports prediction and betting models in the machine learning age: The case of tennis[J]. Journal of Sports Analytics, 2021, 7(2): 99-117.

[2] Gollub J. Forecasting serve performance in professional tennis matches[J]. Journal of Sports Analytics, 2021, 7(4): 223-233.

[3] Gao Z, Kowalczyk A. Random forest model identifies serve strength as a key predictor of tennis match outcome[J]. Journal of Sports Analytics, 2021, 7(4): 255-262.

[4] Fayomi A, Majeed R, Algarni A, et al. Forecasting Tennis match results using the Bradley-Terry model[J]. International Journal of Photoenergy, 2022, 2022.

[5] Zhu Y, Naikar R. Predicting tennis serve directions with machine learning[C]//International Workshop on Machine Learning and Data Mining for Sports Analytics. Cham: Springer Nature Switzerland, 2022: 89-100.

[6] Arcagni A, Candila V, Grassi R. A new model for predicting the winner in tennis based on the eigenvector centrality[J]. Annals of Operations Research, 2023, 325(1): 615-632.

[7] Barnett T, Clarke S R. Combining player statistics to predict outcomes of tennis matches[J]. IMA Journal of Management Mathematics, 2005, 16(2): 113-120.

[8] Vaughan Williams L, Liu C, Dixon L, et al. How well do Elo-based ratings predict professional tennis matches?[J]. Journal of Quantitative Analysis in Sports, 2021, 17(2): 91-105.