A dynamic online nomogram to predict match outcome in the UEFA Champions League: more than meets the eye

Shaoliang Zhang (✉ ZSL.INEF@gmail.com)  
Tsinghua University

Jianyang Hu  
Soochow University

Qing Yi  
Shanghai University of Sport

Ke Deng  
Tsinghua University

Haifeng Wang  
Tsinghua University

Carlos Lago  
Universidade de Vigo

Research Article

Keywords: athletic performance, team sports, football performance analysis, nomogram

Posted Date: May 16th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2915417/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Background

Recently, the application of interdisciplinary research methods to sports performance analysis has become a clear trend. These methods can enhance analytical techniques and provide a deeper understanding of the matching outcome.

Purpose

This study aimed to develop and validate a predictive model to predict match outcomes by transferring an analytical technique common to modern medicine to sports performance analysis. We would like to identify whether interdisciplinary research methods are applicable to predicting match outcomes based on historical data and what factors may affect match outcomes.

Methods

A nomogram was generated based on lasso-logistic regression analysis to identify the potential predictors associated with match outcomes. The predictive model was built based on a nomogram, and its performance was evaluated for discrimination, calibration, and clinical utility.

Results

The nomogram is an effective tool for predicting match outcomes in elite soccer, owing to its higher overall performance, discrimination, and calibration of the current model. Meanwhile, the current predictive model also highlights that counterattacks, shots on target, long balls, short passes, and fouls are positively associated with match outcomes, whereas crosses and yellow cards are negatively associated with match outcomes in the UEFA champion league. A nomogram with these variables had good predictive accuracy (Brier score: 0.21, calibration slope: 1.05, c-index: 0.84).

Conclusion

The nomogram model showed a good predictive accuracy and discriminatory ability. The current predictive model also highlighted that counterattacks, shots on target, long balls, short passes, and fouls are positively associated with match outcomes whereas crosses and yellow cards are negatively associated with match outcomes in elite soccer. Therefore, a nomogram may be an effective tool for analyzing soccer matches. More visualization of predicting match outcome can be checked on this website (https://athletic-performance-and-data-science-lab.shinyapps.io/DynNomapp/)

1. Introduction

Prediction, which involves estimating a certain level or possibility, has been studied in many applications [1, 2]. Through a well-processed performance prediction procedure, a production unit can be able to
predict future performance, analyze the abnormal situation, and consequently, take measures to prevent performance from deteriorating [3]. Indeed, Herold et al., (2019) suggested that elite sports practitioners gradually utilize data to monitor training and match processes related to athletes’ health and performance. A novel application of sports analytics that draws on the data analytics frameworks of business organizations is essential to gain valuable insights and recommendations that can guide decision-making in sport science [4]. Therefore, future analytic techniques in sport science need to transfer from descriptive and diagnostic analytics to predictive and prescriptive data analytics. Notably, although a test or tool with such 100% predictive accuracy is highly unlikely, those that can identify more athletes who go on to achieve or incur a specific outcome given a certain profile would provide confidence that implementing that specific test or tool would provide acceptable predictive ability on which to base confident recommendations [5].

Although there are a few predictive and prescriptive data analytics have been applied in team sports, interest in match outcome prediction has grown with the increased availability of sport-related data online and with the emergence of online sports betting [1, 3, 6]. In recent years, sports scientists intend to improve the accuracy of the predictive model, and algorithms from cross-disciplinary learning are likely to extend sports scientists’ critical thinking and metacognitive skill through novel perspectives generated by the interaction with similar quantitative sciences [7, 8]. For example, Duarte et al. [9] explored how sporting teams could be viewed as “superorganisms,” similar to how ecologists view aggregated organisms, such as flocks of birds, given that athletes are likely to base movement decisions on environmental information extracted from the relative positioning of both their opponents (predator) and teammates (organism aggregate). Considering players and sporting teams in such a nuanced way can provide novel insights into collective behaviors and patterns in play. On the other hand, the analytical technique also needs to be updated and applied to new theories (e.g. complex systems in team sports) which can adequately reveal nonlinear behavioral patterns. Thus, the examination of match-related data may require alternative or “outside the box” approaches adopted from other disciplines [8].

One particular analytical and visualization approach commonplace or the study of prediction is Nomograms [10]. Nomograms are pictorial representations of a complex mathematical formula. Medical nomograms use biological and clinical variables, such as tumor grade and patient age, to graphically depict a statistical prognostic model that generates a probability of a clinical event, such as injury risk or death, for a given individual [11]. There are two primary ways nomograms are used [11]. One is pictorially where each variable is listed separately, with a corresponding number of points assigned to a given magnitude of the variable. Then, the cumulative point score for all the variables is matched to a scale of the outcome. Alternatively, the formula is contained in a computer or smartphone-based calculator, where specific variables are entered and the likelihood of an event is computed [12]. This analytical cross-disciplinary learning transfer from medicine to the sports sciences may enable the emergence of novel data visualization techniques, while simultaneously increasing the sophistication of research questions regarding athlete and team behaviors.
To achieve this aim, the aim of this study is to develop and validate a commonly used nomogram from modern medicine to predict and visualize match outcomes in elite football. Ultimately, this may provide sports coaches or sporting administrators with greater objectivity to support decision-making.

2. Methods

2.1 Sample and selected variables

This retrospective observational study was conducted using the dataset from match-related data of teams in the UEFA Champions League from season 2009–2010 to 2020–2021 (n = 1486 matches). Match data were obtained from a public-accessed football statistics website “whoscored.com” (https://www.whoscored.com). The data resource of the website is the sports analytics company OPTA Sports (London, UK). The reliability of OPTA in coding match actions and events has already been successfully tested as highly reliable [13].

Twenty-two performance-related match actions or events were chosen as variables in the present study and were divided into four groups based on previous studies. Definitions of these variables can be found in previous studies [14–16]. Match-related statistics were analyzed considering four situational variables: (1) match location: home and away; (2) team quality: teams that qualified for the knockout stage and teams that didn’t qualify for the knockout stage; (3) opponent quality: opponents that qualified into the knockout stage and opponents that didn’t qualify into the knockout stage; and (4) match outcome: win and no win (e.g. draw and lose).

2.2 Statistical analysis

2.2.1 Feature selection and data preprocessing

Predicting the match outcome belongs to the dichotomous without standardization, and the factors interact with the dependent variables accompanied by collinearity. We selected the least absolute shrinkage and selection operator (LASSO) regression analysis which is a shrinkage and variable selection method for minimizing high-dimensional data. In order to obtain the subset of predictors, the LASSO regression analysis minimizes prediction error for a quantitative response variable by imposing a constraint on the model parameters that cause regression coefficients for some variables to shrink toward zero [17]. Variables with a regression coefficient equal to zero after the shrinkage process are excluded from the model while variables with nonzero regression coefficients are most strongly associated with the response variable [18]. Based on the type measure of -2 log-likelihood and binomial family, the LASSO regression analysis running in R software runs 10 times K cross-validation for centralization and normalization of included variables and then picks the best lambda value. “Lambda. Ise” gives a model with good performance but the optimal number of independent variables [19]. The LASSO method was used to analyze the data to select the optimal predictors, and then, the insignificant variables are unswervingly dropped, avoiding data Overfitting. Features with nonzero coefficients were screened out by running the LASSO method.
2.2.2 Model development

The data sample was randomly split into two datasets: a training (80%) cohort, which was used to train the model with hyperparameters, and an internal validation (20%) cohort, which was used to test the developed models on unseen data. The characteristics of training and testing sets are presented as the mean (SD) for data that were normally distributed and the frequencies and percentages for categorical variables. The Kolmogorov-Smirnov test was used to inspect the normality and the homogeneity of variance of all the data was tested by the Levene test. Then, a prediction model was established by binary logistic regression analysis, combined with the nonzero coefficient features screened by the LASSO regression model (Kidd et al., 2018). The level of statistical significance was stationed bilateral and the odds ratio (OR) as a P value was 95% confidence interval (CI). Following the definition of the OR for each variable evaluated independently for the probability of winning, a nomogram was developed as the graphical representation of our final model. The nomogram has a reference line on the top for scoring points for each predictor from 0 to 100. The predictive variables are displayed below with bars that scale their effect size, demonstrating visually the relative weight of each variable and allowing for points to be assigned to each significant clinical characteristic. The summation of points from each predictor and the corresponding predicted probability of winning can be read from the bottom 2 lines [11]. The nomogram outperforms the traditionally used staging systems because it considers multiple commonly available prognostic variables simultaneously, including the identification, calibration, and stratification of key factors according to a particular outcome [10, 20]. Consequently, the outcomes could be analyzed more effectively to support match strategies.

2.2.3 Model validation

To reduce the overfitting bias, bootstrapping using 500 repetitions was used for internal validation of our model and to obtain bias-corrected predictive accuracy measures of the final model. Furthermore, the predictive accuracy of the final model was assessed using (1) Brier score for overall performance, which assesses the difference between observed and predicted values with values closer to 0 indicating better predictive ability; (2) calibration slope for calibration, which assesses the agreement between observed and predicted values with values closer to 1 indicating better performance; and (3) c-index for discrimination, which assesses how well the model distinguishes between those with and without the outcome of interest with values of 0.5 indicating a noninformative model and 1 indicating perfect discrimination. In addition to these numeric measures, we also used the calibration plot and receiver operating characteristic curve to present the calibration and discrimination aspects of our final model.

2.2.4 implementation

All analyses were conducted using the statistical programming environment R (version 4.1.2) and Python (version 3.9). Specifically, the R package “glmnet” statistical software was used to perform the LASSO regression. Subsequently, variables identified by LASSO regression analysis were entered into logistic regression models that were performed using the “autoReg” package. The nomogram was performed using the “rms” package. The "skit-learn" was used to plot and validate the model. Furthermore, an online
web application facilitating the use of nomograms was developed using the “DynNom” package in the R software. All probabilities are two-tailed. P < 0.05 was considered statistically significant.

3. Result

3.1 Characteristics of training and validation sets

We randomly divide 2378 of these matches (80%) into a development cohort and 594 matches (20%) into an independent validation cohort. The detailed information is listed in Table 1. No significant differences were found between the training and testing set.
Table 1
Descriptive statistics of all matches in the training and testing set.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Training set (N = 2,379)</th>
<th>Testing set (N = 593)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Match outcome</td>
<td></td>
<td></td>
<td>0.99</td>
</tr>
<tr>
<td>No win</td>
<td>1,444 (61%)</td>
<td>360 (61%)</td>
<td></td>
</tr>
<tr>
<td>Win</td>
<td>935 (39%)</td>
<td>233 (39%)</td>
<td></td>
</tr>
<tr>
<td>Team quality</td>
<td></td>
<td></td>
<td>0.14</td>
</tr>
<tr>
<td>Strong</td>
<td>1,441 (61%)</td>
<td>379 (64%)</td>
<td></td>
</tr>
<tr>
<td>Weak</td>
<td>938 (39%)</td>
<td>214 (36%)</td>
<td></td>
</tr>
<tr>
<td>Opponent quality</td>
<td></td>
<td></td>
<td>0.97</td>
</tr>
<tr>
<td>Strong</td>
<td>1,458 (61%)</td>
<td>363 (61%)</td>
<td></td>
</tr>
<tr>
<td>Weak</td>
<td>921 (39%)</td>
<td>230 (39%)</td>
<td></td>
</tr>
<tr>
<td>Match location</td>
<td></td>
<td></td>
<td>0.61</td>
</tr>
<tr>
<td>Away</td>
<td>1,195 (50%)</td>
<td>291 (49%)</td>
<td></td>
</tr>
<tr>
<td>Home</td>
<td>1,184 (50%)</td>
<td>302 (51%)</td>
<td></td>
</tr>
<tr>
<td>Shots</td>
<td>13.2 ± 6.0</td>
<td>13.4 ± 5.7</td>
<td>0.25</td>
</tr>
<tr>
<td>Shots on target</td>
<td>4.76 ± 2.84</td>
<td>4.76 ± 2.73</td>
<td>0.83</td>
</tr>
<tr>
<td>Open play</td>
<td>9.0 ± 4.6</td>
<td>9.2 ± 4.4</td>
<td>0.18</td>
</tr>
<tr>
<td>Set piece</td>
<td>3.02 ± 2.16</td>
<td>3.18 ± 2.22</td>
<td>0.13</td>
</tr>
<tr>
<td>Counter attack</td>
<td>0.44 ± 0.76</td>
<td>0.41 ± 0.75</td>
<td>0.58</td>
</tr>
<tr>
<td>Possession</td>
<td>50 ± 12</td>
<td>50 ± 12</td>
<td>0.76</td>
</tr>
<tr>
<td>Passes</td>
<td>509 ± 140</td>
<td>508 ± 137</td>
<td>0.95</td>
</tr>
<tr>
<td>Pass success</td>
<td>81 ± 7</td>
<td>80 ± 7</td>
<td>0.47</td>
</tr>
<tr>
<td>Short passes</td>
<td>428 ± 137</td>
<td>428 ± 135</td>
<td>0.89</td>
</tr>
<tr>
<td>Long balls</td>
<td>60 ± 14</td>
<td>59 ± 14</td>
<td>0.20</td>
</tr>
<tr>
<td>Crosses</td>
<td>18 ± 11</td>
<td>18 ± 9</td>
<td>0.75</td>
</tr>
<tr>
<td>Through balls</td>
<td>2.34 ± 2.81</td>
<td>2.17 ± 2.41</td>
<td>0.71</td>
</tr>
<tr>
<td>Dribbles won</td>
<td>9.6 ± 4.8</td>
<td>9.8 ± 4.8</td>
<td>0.15</td>
</tr>
<tr>
<td>Aerial success</td>
<td>50 ± 14</td>
<td>50 ± 13</td>
<td>0.82</td>
</tr>
<tr>
<td>Corners</td>
<td>4.98 ± 3.01</td>
<td>5.01 ± 2.94</td>
<td>0.72</td>
</tr>
</tbody>
</table>
### 3.2 Feature selection and model development

Twenty-two contextual and technical variables were included in the original model and later reduced to 14 potential predictors using the LASSO regression model (Fig. 1).

To develop a predictive model for predicting match success, binary logistic regression analysis was performed based on the aforementioned 14 variables selected by the LASSO regression technique. The results of the binary logistic analysis are presented in Fig. 2.

The detailed results were as follows: Shots on target (OR: 1.48; 95% CI 1.40 to 1.56; P < 0.001), SFCA-Shots from counterattacks (OR: 1.51; 95% CI 1.29 to 1.75; P < 0.001), short passes (OR: 1.00; 95% CI 1.00 to 1.00; P < 0.002), Long balls (OR: 1.01; 95% CI 1.01 to 1.02; P < 0.001), Crosses (OR: 0.94; 95% CI 0.93 to 0.95; P < 0.001), Aerial success (OR: 1.01; 95% CI 1.01 to 1.02; P < 0.001), Fouls (OR: 1.06; 95% CI 1.03 to 1.09; P < 0.001), Yellow cards (OR: 0.84; 95% CI 0.77 to 0.92; P < 0.001). Furthermore, team quality (OR: 0.22; 95% CI 0.17 to 0.29; P < 0.001), opponent quality (OR: 3.44; 95% CI 2.74 to 4.31; P < 0.001), and match location (OR: 2.23; 95% CI 1.78 to 2.80; P < 0.001).

### 3.3 Validation of the predictive model

The overall performance of this nomogram had good predictive accuracy (Brier score: 0.21, calibration slope: 1.05, c-index: 0.84). The predictive accuracy of the nomogram in the training cohort, the AUC was 0.838 (Fig. 4A), and the calibration curve was close to the ideal diagonal line (Fig. 4B). Furthermore, 297 matches were used to test the nomogram. The AUC was 0.847 (Fig. 4C), reflecting a good accuracy of the nomogram. Meanwhile, the model had a good consistency and the calibration curve of the validation cohort was also close to the ideal diagonal line (Fig. 4D).

### 4. Discussion

To the best of our knowledge, this is the first study to develop and validate a predictive model to evaluate match outcomes by transferring an analytical technique common to modern medicine into sports performance analysis. The nomogram model showed a good predictive accuracy and discriminatory ability. Eleven indicators were included in the nomogram model to evaluate match outcome (win or loss). Notably, the current predictive model highlighted that counterattacks, shots on target, long balls, short passes, and fouls are positively associated with match outcomes whereas crosses and yellow cards are
negatively associated with match outcomes in the UEFA champions league. Furthermore, contextual variables should be considered because they have a significant impact on match outcome.

The current study importantly highlights that counterattacks are the most significant indicator associated with match success. This result is in line with a previous study that found that direct attacks and counterattacks were three times more effective than elaborate attacks for producing score-box possession [21]. Likewise, Tenga et al. (2010b) found that a counterattack is an effective way to be successful because it exploits imbalances in the opponent’s defense to achieve penetration. Although whether “possession play” or “direct play” is more effective has long been disputed in football science, the majority of studies demonstrated that winning teams increased their use of direct play and counterattack, and decreased the use of maintenance, build-up, and sustained threat [15, 22]. This is mainly because winning teams’ reduction in these styles could be a focus on maintaining the advantage through defending, which results in reduced possession time. Moreover, increasing the use of direct play and counterattack when winning allows the team to keep players close to their own goal and take advantage of the advanced position of opposing teams combining with long balls to easily start a fast break [23]. On the other hand, losing teams decreased the use of direct play and increased the use of build-up and sustained threats to try to maintain the attack close to the goal of the opposition [23].

Crosses are not related with match success. This finding was also found by [24]. According to this study, crosses often are labeled as an inefficient method to create good scoring opportunities due to an airborne delivery of the ball into the opponent’s penalty area. Furthermore, crosses were more frequent for losing teams, which might suggest that losing teams employ this tactic to create more goal-scoring opportunities when attacking [23]. However, the current study also found that short passes were also the one that most distinguished win and loss in build-up plays leading to goals scored. This is in line with the available literature showing passing as a key skill underlying successful performance in football [25]. Surprisingly, a negative effect was found with respect to crosses. Indeed, some studies suggest that the last technical action was a cross, and the probability of offensives sequences ending successfully was 2.8 times higher compared to using a short or medium pass [25]. Consequently, the present study suggests that successful teams did not strictly explore the wings for creating overloads in less congested sub-areas of play. Apparently, they were also able to perform more ‘penetrating passes’ into vital sub-areas of play, affording advantageous spatial-temporal conditions (e.g., wider angles) to shoot on goal.

Fouls without cards are positively associated with match success, but yellow cards may decrease winning probability during match-play. This result is in agreement with previous studies. Fouls can break down the match pace and fix tactics in the attacking progress of the opponent [26]. However, it is worth noting that fouls with yellow cards will have a negative effect on match success which is supported by the previous studies that unsuccessful teams received a greater number of yellow cards in the UEFA champions league [27]. A possible explanation is that a yellow card is a precursor to a red card, which can lead to playing with a conservative attitude after the yellow card and the accumulating yellow cards will be negative especially if the sending-off occurs at the end of the match [28].
In addition, contextual variables have a significant impact on match success. This has been well documented by previous studies, match location, team quality, and opponent quality have the same importance for match success and the impact of each variable is interactive [15, 25, 28, 29].

The nomogram can provide better-individualized prediction match outcomes in an intuitive and visual way (Fig. 3). The current nomogram had excellent discrimination, with an AUC of 0.840. Therefore, this nomogram has outstanding clinical transformation value. The nomogram also showed good calibration, thus a convenient tool with clinical value.

Bootstrapping using 500 repetitions was used for internal validation of our model and to obtain bias-corrected predictive accuracy measures of the final model. The overall performance of the model is close to the perfect model. An example of a dynamic online nomogram is from Real Madrid versus Liverpool in the 2020–2021 UEFA champions league. Online visualization presented that the probability of winning for Real Madrid is approximately 86.4% (95% CI of 81.9–89.9%) based on historical match-related statistics. More examples in the UEFA Champions League from season 2009–2010 to 2020–2021 can be checked on the following website: https://athletic-performance-and-data-science-lab.shinyapps.io/DynNomapp/ (Fig. 5).

One of the strengths of this study is that it was done on a considerable number of football matches from the UEFA champions league and the use of the LASSO regression has helped to select the prominent independent variables influencing match outcome which clearly differentiates the previous studies that mainly use the PCA (principle component analysis) in sports science.

Several limitations should be highlighted, First, although we develop an effective tool to predict match outcomes, other statistical techniques may be used to complete the information provided by the nomogram. For example, clustering teams in the UEFA champions league into different levels would be interesting. Second, although the current predictive model presents higher predictive accuracy, future research is still recommended to add more variables such as event data and tracking data to optimize and further validate the accuracy of the current model and explore the rationality for applying it to sports performance analysis.

5. Conclusion

The nomogram model showed a good predictive accuracy and discriminatory ability. The current predictive model also highlighted that counterattacks, shots on target, long balls, short passes, and fouls are positively associated with match outcome whereas crosses and yellow cards are negatively associated with match outcome in the UEFA champions league. Finally, although the interdisciplinary research method presented higher accuracy to predict match outcome in current research, this method applied to sports performance analysis still needs more attention and validation in various football leagues.
Abbreviations

UEFA  Union of European Football Associations
AUC  Area under the curve
CI  Confidence interval
OR  Odds ratio

Declarations

Acknowledgements
Not applicable

Author contributions
Shaoliang Zhang and Jianyang Hu are responsible for model building and article writing. Qing Yi, Ke Deng, Haifeng Wang and Carlos Lago are responsible for data processing and article proofreading. All authors read and approved the final version of the submitted manuscript.

Funding
Not applicable

Availability of data and materials
The data that support the findings of this study are available from the corresponding author upon reasonable request.

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

References


Figures
Figure 1

Feature selection using the least absolute shrinkage and selection operator (LASSO) binary logistic regression model. (A) LASSO coefficient profiles for key features, each coefficient profile plot is produced via log(lambda) sequence. The dotted vertical line is set at the nonzero coefficients selected via 10-fold cross-validation, where 14 nonzero coefficients are included. (B) By verifying the optimal parameter (lambda) in the LASSO model, the partial likelihood deviance (binomial deviance) curve is plotted versus log (lambda). At the log (lambda) of the optimal values, where features are selected, dotted vertical lines are set using the minimum criteria and the one standard error of the minimum criteria. (C) The specific coefficient of each variable is presented by (LASSO) binary logistic regression. The red dot means the variable is excluded while the green dot means the variable is selected in the following study.
Figure 2

Independent key factors related to match outcome by multivariate logistic regression Analysis. Note: OR, odds ratio; CI, confidence interval, SFCA, shots from counterattack; SFSP, shots from set pieces.
Figure 3

Proposed nomogram for predicting the probability of winning during match-play. SFCA, shots from counterattack.
Figure 4

Receiver operating characteristic (ROC) curve for assessing the discrimination performance of the nomogram in the training sets (A) and testing sets (C); the area under the curve (AUC) was 0.838 (A) and 0.847 (C). The calibration curves of the nomogram in the training group (B) and the testing group (D). The calibration curves depict the calibration of the nomogram in terms of the agreement between the predicted outcome and actual outcome of no-reflow. The x and y axes are the predicted probability and true probability, respectively.
Figure 5

Example of a web-based dynamic nomogram to predict winning probability in the UEFA champions league. The probability of winning is approximately 73.7% (95% CI of 65.7% to 80.3%). Next, More examples are the following website, https://athletic-performance-and-data-science-lab.shinyapps.io/DynNomapp/. The procedure for using the dynamic nomogram is as follows. First, enter the value of each variable. Then, click on the “Predict” button and the result will be displayed on the right side of the page. After completing the prediction for the current patient, click the “Quit” button and repeat the above steps to start the prediction for the next patient. Note: If the “Quit” button is not clicked, the website may not work when someone else logs in and uses it again.