ISSN: 2456-4419
Impact Factor: (RJIF): 5.18
Yoga 2019; 4(1): 339-342
© 2019 Yoga
www.theyogicjournal.com
Received: 19-11-2018
Accepted: 22-12-2018

Amritashish Bagchi
Symbiosis School of Sports Sciences, Symbiosis International (Deemed University), Pune, Maharashtra, India

Nirmal Salvi
University Sports Board, Symbiosis International (Deemed University), Pune, Maharashtra, India

Shiny Raizada
Symbiosis School of Sports Sciences, Symbiosis International (Deemed University), Pune, Maharashtra, India

Correspondence
Amritashish Bagchi
Symbiosis School of Sports Sciences, Symbiosis International (Deemed University), Pune, Maharashtra, India

Predicting the outcome of FIFA world cup matches

Amritashish Bagchi, Nirmal Salvi and Shiny Raizada

Abstract

The aim of this study was to develop a prediction model to predict the outcome of FIFA World Cup matches. It may help the team captain, coaches or managers to change the tactics accordingly for the second half of the match and it will also help coaches to prepare practice sessions according to this specificity and to be ready to control these variables in competition. The data was collected from 2018 FIFA World Cup. A total 63 match data were recorded, out of which 12 matches were draw and therefore not included in the study. The dependent variable selected for this study was Match Outcome (Win/Loss). Total Shots Taken, Shots On-Target, Shots Blocked, Fouls, Corner Kick, Attempts from Free Kick, Penalty Kick, Penalty Converted, Offside, Ball Possession, Actual Playing Time and Half Time Score were selected as the predictor variables. For the purpose of this study only the first half data was used and in statistical technique Binary Logistic regression was used to predict the outcome of a match (Win/Loss). The result indicates that the developed Logistic regression Model was significant. According to the statistical significance of the predictor variables, they were numerically weighted and can be used to predict the match outcome. The predictor variables such as half time score, attempts from free kick and shots taken were included in the prediction model with coefficient of determination (R2) of.223 (Cox & Snell) and .297 (Nagelkerke). The classification matrix shows that 69.6% of match results were correctly classified by the model.

Keywords: Football, FIFA world cup, prediction model, win and loss

Introduction

Football is the most popular sport in the world. According to Nielsen World Football Report 2018, more than 700 million people across America, Europe, the Middle East and Asia shown interest in Football [12]. There are around 108 professional soccer leagues located in 71 countries around the world. Due to which there are millions and trillions of data available in the internet, which is being used by researchers, statisticians and data analysts to analysis the data, interpret the data and make some conclusions [10].

Now a days historical results are used for the prediction of future games results for which simple statistical algorithms have been applied by the researchers and statisticians. These algorithms are mostly developed by comparing the strengths and weakness of the teams in order to make a prediction. Usually, the longer the historical data, the more accurate are the results. However, predictions based on these simple statistical algorithms may not be very accurate when the two teams have not competed with each other [11].

The football prediction model has become increasingly popular in the last few years and many different approaches of prediction models have been proposed. Due to immense popularity, a lot of people have come up with statistical models of predicting games based on different parameters. In one study the researcher proposed that the Poisson model, used in predicting football match outcomes is mainly based on how much ball possession each team has, in respect to scoring when with the ball and not conceding when not with the ball; not necessarily about keeping the ball for a certain amount of time. Unlike the Poisson Model, there are other variables which have an impact on match outcomes [5]. Another study stated that Home Advantage plays an important role, where the main element is the crowd size and density. Two independent variables that can be taken is the team’s recent form home and away matches [14]. Haaren & Davis has developed ELO rating system to predict matches where it gives points to a team based on a win/draw/loss. But it also included an interesting variant that rewards points on goal difference.

So if one team wins 3-0 and the other one wins 2-1, the former would be rewarded with more points because of the superior goal different [7]. But this method was useful in encoding information of past results. In another study Simulation based methodology was used during the 2010 & 2014 world cup matches where they adopted number of goals of two teams whose mean is proportional to the relative technical level of the opponents. To measure this, FIFA ratings were considered and expert opinions were recorded to construct prior distribution of parameters [16]. Logistic regression is one of the statistical methods used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables [3, 15]. The main purpose of this statistical technique is to predict the outcome (binary or multinominal) on the basis of predictor variables selected by the researcher. To the best of our knowledge none of the studies have used the half-time variables to predict the final match outcome in football. The purpose of this study is to develop a prediction model where a match result can be predicted on the basis of Half-Time score.

Methodology

A total 63 match data were recorded, out of which 12 matches were draw and therefore not included in the study. All the data were collected from the website FIFA.com [1]. The dependent variable selected for this study was Match Outcome (Win/Loss).

Total Shots Taken, Shots on Target, Shots Blocked, Fouls, Corner Kick, Attempts from Free Kick, Penalty Kick, Penalty Converted, Offside, Ball Possession, Actual Playing Time and Half Time Score were selected as the predictor variables. For the purpose of this study only the half time data was used. Data is presented as mean with standard deviations. The statistical technique Binary Logistic Regression was used to develop the prediction model. For this purpose Statistical Package for Social Science (SPSS) version 24.0 was used. The level of significance was set at 0.05.

Results and Discussion

Logistic regression does not require many of the key assumptions to be fulfilled, such as linearity, normality, homoscedasticity, and measurement level [2]. Therefore, only the descriptive statistics (i.e. mean, standard error of mean, standard deviation) was used to see the nature of data and the correlation matrix was used to check the assumption of high multicollinearity among the variables.

Table 1: Descriptive Statistics

	Total Shots Taken	Shots On Target	Shots Blocked	Fouls	Corner Kick	Attempts from Free Kick	Penalty Kick	Penalty Converted	Offside	Ball Possession	Actual Playing time	Half time Score
Mean	5.61	1.66	1.72	6.38	2.10	.29	.11	.09	.66	49.137	13.58	.44
Std. Error of Mean	.266	.135	.174	.254	.141	.051	.037	.031	.081	1.1882	.343	.076
Std. Deviation	2.689	1.368	1.760	2.560	1.425	.519	.370	.318	.814	12.0000	3.460	.765

Table 2: Correlation

	Total Shots Taken	Shots On Target	Shots Blocked	Fouls	Corner Kick	Attempts from Free Kick	Penalty Kick	Penalty Converted	Offside	Ball Possession	Actual Playing time	Half time Score
Total Shots Taken	1	.606**	.434**	-.132	.305**	.154	.103	.041	-.171	.245*	.181	.075
Shots On Target	.606**	1	.062	-.146	.068	.102	.328**	.344**	-.107	.102	.034	.468**
Shots Blocked	.434**	.062	1	-.224*	.153	-.048	-.105	-.132	-.096	.165	.232*	-.082
Fouls	-.132	-.146	-.224*	1	.109	.056	-.054	-.005	.021	-.005	-.174	-.001
Corner Kick	.305**	.068	.153	.109	1	-.080	.017	-.019	-.056	.296**	.318**	-.095
Attempts from Free Kick	.154	.102	-.048	.056	-.080	1	-.064	-.039	-.181	-.097	-.085	-.031
Penalty Kick	.103	.328**	-.105	-.054	.017	-.064	1	.929**	.091	.093	-.041	.565**
Penalty Converted	.041	.344**	-.132	-.005	-.019	-.039	.929**	1	.042	.093	-.047	.652**
Offside	-.171	-.107	-.096	.021	-.056	-.181	.091	.042	1	.123	.004	.150
Ball Possession	.245*	.102	.165	-.005	.296**	-.097	.093	.093	.123	1	.689**	.043
Actual Playing time	.181	.034	.232*	-.174	.318**	-.085	-.041	-.047	.004	.689**	1	-.090
Half time Score	.075	.468**	-.082	-.001	-.095	-.031	.565**	.652**	.150	.043	-.090	1

The correlation matrix between sets of variables is shown in the above table. The correlation matrix was used to check the assumption of multicollinearity. Although there is a significant correlation between few of the variables but the severity of the correlation (VIF Value) was not high except for the correlation between Penalty kick and Penalty Converted. The researcher has used eta value to check the relationship of each variable (Interval) with the dependent variable (Nominal). The variable Penalty kick was removed for the further analysis, as the eta value (.186) of the Penalty kick and match results was found to be less as compared to eta value (.217) of Penalty Converted (.217) and match results. Variance Inflation Factor (VIF) was used to check severity of multicollinearity. For all the variables the VIF value was near by 1, which means the multicollinearity between the independent variables was low.

Table 3: Omnibus Test of Model Coefficients

	Chi-square	df	Sig.
Step	4.226	1	.040
Block	25.734	3	.000
Model	25.734	3	.000

The developed model is significantly better fit than the null model. The omnibus test of model coefficients shows a significant decrease in the -2 Log Likelihood value (i.e. 115.668), as compared to -2 Log Likelihood value (i.e. 141.402) of the null model.

Table 4: Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
3	115.668^a	.223	.297
a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.			

From the above table it can be seen that the value of Nagelkerke R^2 is .297 in the third model developed in binary logistic regression, but the value of Cox & Snell R-square is found to be .223. The Nagelkerke R^2 value was considered for the developed model because the Cox & Snell R-square is based on the log likelihood for the model compared to the log likelihood for a baseline model. However, even for a "perfect" model with categorical outcomes, it has a theoretical maximum value of less than 1. Nagelkerke R^2 is the adjusted version of the Cox & Snell R-square that adjusts the scale of the statistic to cover the full range from 0 to 1 [9]. The value of Nagelkerke R^2 is .297 which means 29.7% of the variability in the dependent variable is explained by the selected independent variables.

Table 5: Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
3	5.764	8	.674

The Hosmer-Lemeshow test (HL test) is a goodness of fit test for developed logistic regression model. The null hypothesis of Hosmer-Lemeshow test is that the fitted model is correct, which means that higher the p – value better is the model. In the above table, the p – value of Hosmer and Lemeshow test is .674 which is insignificant. Hence the model fit is good, in other words the observed event rates match the expected event rates in population subgroups.

Table 6: Classification Table

Step	Observed		Predicted		
			Match Result		Percentage Correct
			Loss	Win	
Step 3	Match Result	Loss	36	15	70.6
		Win	16	35	68.6
	Overall Percentage				69.6

The above table shows the summary of correct and wrong classification of the subjects in match Outcome (i.e. Loss or Win) on the basis of the developed regression model. It can be seen from the table that 71 (Loss 36 and Win 35) matches were correctly classified from 102 matches. Overall 69.6% of matches were correctly classified on the basis of selected independent variables.

Table 7: Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 3^c	Total Shots Taken	.214	.091	5.498	1	.019	1.239
	Attempts from Free Kick	.914	.468	3.817	1	.051	2.493
	Half time Score	1.426	.442	10.418	1	.001	4.164
	Constant	-2.011	.611	10.820	1	.001	.134
a. Variable(s) entered on step 1: Half time Score.							
b. Variable(s) entered on step 2: Attempts from Free Kick.							
c. Variable(s) entered on step 3: Shots Taken							

The above table provides the regression coefficient (B), the Wald statistic (used to test the significance of individual coefficients in the model) and the all-important Odds Ratio (Exp (B)). "B" coefficients are also known as unstandardized coefficients and are used to develop the regression equation [4]. A total of three variables (i.e. half time score, attempts from free kick and total shots taken) out of twelve variables were selected by the model. All these variables are important in predicting the match outcome of a Football match. It may help the team captain, coaches or managers to change the tactics accordingly for the second half. Coaches can use this information to design the training sessions and matches [13]. But it should be taken into consideration that although the variables included in the model is highly significant and it can classify upto 69.6% of cases correctly, still it only explain 29.7% of the variability in the dependent variable. It means 70.3% of the variability is explained by some other variables which were not included in the study.

Previous studies have shown that frequency and effectiveness of shots on goal and passing are essential parameters that differentiate the winning and the losing teams [6]. Hughes & Franks showed that there were differences between successful and unsuccessful teams in converting possession into shots on goal, with the successful teams having the better ratios [8]. The results from the present study indicate that winning teams made more shots, attempts from free kick and score in first half than losing teams.

Regression Equation

Using regression coefficients (B) of the model shown in the table 7, the regression equation was developed which is as follows:

Logit = -2.011 + 1.426 (Half Time Score) + .914 (Attempts from Free Kick) + .214 (Total Shots Taken)

Odds = e^{logit} = $e^{(-2.011 + 1.426 \text{ (Half Time Score)} + .914 \text{ (Attempts from Free Kick)} + .214 \text{ (Total Shots Taken)}}$

P(Y) = $\frac{odds}{1+odds}$

The above regression equation can be used to predict the match outcome (i.e. Win/Loss) of the future FIFA World Cup matches on the basis of three predictor/ independent variables (i.e. Half Time Score, Attempts from Free Kick and Total Shots Taken) of the first half data.

Conclusion

The purpose of this study was to develop a prediction model to predict the outcome of FIFA World Cup matches. It may help the team captain, coaches or managers to change the tactics accordingly for the second half of the match and it will also help coaches to prepare practice sessions according to this specificity and to be ready to control these variables in competition. The developed Logistic regression Model was found to be significant. According to the statistical significance of the predictor variables, they were numerically weighted and were used to predict the match outcome. The variables which are selected in the prediction model are Half time Score, Attempts from free kick and Total Shots Taken all together explaining only 29.7% of the variability in the dependent variable. 69.6% of match results were correctly classified by the model. Further study could be done by including more variables that significantly contribute to the match outcome. So that the remaining variability can be explained and the model fit can be improved for more correct prediction along with high probability.

References

1. FIFA World Cup Russia™. FIFA.com. www.fifa.com. https://www.fifa.com/worldcup/statistics/. Published, 2018. Accessed February 24, 2019.
2. Assumptions of Logistic Regression. Statistics Solutions. Statistics Solutions. https://www.statisticssolutions.com/assumptions-of-logistic-regression/. Accessed February 24, 2019.
3. Bagchi A, Raizada S, Mhatre A, Augustine A. Forecasting the winner of pro kabaddi league matches. International Journal of Physiology, Nutrition and Physical Education. 2019; 4(1):383-386.
4. Bewick V, Cheek L, Ball J. Statistics review 14: logistic regression. Critical Care. 2005; 9(1):112-8.
5. Boldrin B. Predicting the Result of English Premier League Soccer Games with the Use of Poisson Models. Deland, Florida, 2017, 1-66. http://www2.stetson.edu/~efriedma/research/boldrin.pdf. Accessed January 1, 2017.
6. Grant AG, Williams AM, Reilly T. Analysis of the goals scored in the 1998 World Cup. Journal of Sports Sciences. 1999; 17:826-827.
7. Haaren J, Davis J. Predicting The Final League Tables Of Domestic Football Leagues. 5th ed. Loughborough, United Kingdom: Math Sport International, 2015, 1. https://lirias2repo.kuleuven.be/bitstream/handle/123456789/495623/msi15-paper.pdf?sequence=1. Accessed January 24, 2019.
8. Hughes M, Franks I. Analysis of passing sequences, shots and goals in soccer. J Sport Sci. 2005; 23:509-514.
9. IBM Knowledge Center. Ibm.com. https://www.ibm.com/support/knowledgecenter/en/SSLVMB_24.0.0/spss/tutorials/cslogistic_bankloan_rsq.html. Accessed February 24, 2019.
10. Igiri C, Nwachukwu E. An Improved Prediction System for Football a Match Result. IOSR Journal of Engineering. 2014; 04(12):12-20.
11. Leung C, Joseph K. Sports Data Mining: Predicting Results for the College Football Games. Procedia Computer Science. 2014; 35:710-719.
12. World Football Report. 2018. https://nielsensports.com/wp-content/uploads/2014/12/Nielsen_World-Football-2018-6.11.18.pdf. Accessed February 24, 2019.
13. Peñas C, Ballesteros J, Dellal A, Gómez M. Game-related statistics that discriminated winning, drawing and losing teams from the Spanish soccer league. Journal of Sports Science and Medicine. 2010; 9:288-293.
14. Ponzo M, Scoppa V. Does the Home Advantage Depend on Crowd Support? Evidence from Same-Stadium Derbies. Journal of Sports Economics. 2016; 19(4):562-582.
15. Raizada S, Bagchi A, Menon H, Nimkar N. Predicting the outcome of ICC cricket world cup matches. International Journal of Physiology, Nutrition and Physical Education. 2019; 4(1):119-122.
16. Saraiva E, Suzuki A, Filho C, Louzada F. Predicting football scores via Poisson regression model: applications to the National Football League. Communications for Statistical Applications and Methods. 2016; 23(4):297-319.