Dr. Deepak Paswan
Assistant Professor, Belda
College, Belda, West Bengal,
India.

A binary logistic model for predicting the win probability of Liverpool FC

Dr. Deepak Paswan

Abstract
Football is known to be one of the most popular games played in the world, and when we talk about football, one can't simply ignore the name of Liverpool FC from it, from past 2-3 years the performance of the team has gone up very significantly. There can be many factors which are responsible behind making Liverpool FC successful. So, for the purpose of the study the data of total of nine independent variables were taken form 100 matches the club played across all competitions, the variables were Shots taken on opponent, Shots taken on target, Ball possession, Passes completed, Accuracy of passes, Corners taken, Goal scored in first half, Ground (home or away), Goals scored by Mohammed Salah. Logistic regression (binary logistics) to predict the win probability of Liverpool FC, after the analysis it was found that variables like shots on target, ball possession, passes completed, goal scored in 1st half and goal scored by Mohammed Salah were included in the model for prediction of win probability of Liverpool FC.

Keywords: Goals scored, Liverpool FC, binary logistic

Introduction
"In Life, as In a Football Game, the Principle to Follow is "Hit the Line Hard"
-Theodore Roosevelt

You may wonder while watching a football match, why football is so popular across the globe? The last football world cup was played in 2018 at Russia and it was viewed by more than 4 billion people and world's population is near 8 billion, which means almost half of the world population glued themselves in front of their television screen and witnessed the leviathan doom of all sporting event. (Chanis, 2018) [4].
Liverpool FC is one of the most successful football clubs which participates in English Premier League, and from the past three seasons the team is simply unstoppable, from the season 2017-18, 2018-19 and 2019-20 the team has played two champions league final and won at one occasion. (Begley, 2020) [3].
In Premier League, the year 2018-19 was astonishing as Liverpool FC finished 2nd on the chart just 1 point behind from winners Manchester City FC. (EPL, 2019) [8].
The whole football fraternity got shocked with the consistency the red devils showed, their golden form continued from the last season to 2019-20 and they are standing 25 points clear from the 2nd position holder of the league loosing and drawing just 1 game each (EPL 2019) [8].
This doesn't ends here, along with Champions League in 2019 they also won the Super Cup and were crowned as the world champions after winning the FIFA Club World Cup in the same year and became the first football club in the world to win a Continental Treble. (Erozden, 2019) [9].
The numbers are certainly amazing, but the main thing which is astonishing is the consistency of their performances and the factors behind it. So, to analyze and predict the reason behind their huge success present study was taken.

Material and methods
For the purpose of the current study, a total of 100 match facts were analyzed from 17th march 2018 to 12th march 2020 which Liverpool FC played. The matches of Premier League,

Corresponding Author:
Dr. Deepak Paswan
Assistant Professor, Belda
College, Belda, West Bengal,
India.

Champions League, FA cup, UEFA Super Cup, Carabao cup and FIFA Club World cup were taken into consideration. So, to predict the win probability the independent variables chosen in the study were as follows:

- Shots taken on opponent
- Shots taken on target
- Ball possession
- Passes completed
- Accuracy of passes
- Corners taken
- Goal scored in first half
- Ground (home or away)
- Goals scored by Mohammed Salah

So, a total of nine independent variables were taken to predict the win probability of Liverpool FC. The data of a total of 100 matches were taken from the official site of Liverpool FC for the purpose of the study. (Liverpool, 2019) [12].

The dependent variable was considered and named "result", as the result of each match would be dichotomous in nature i.e. win or lose. As data were collected from a valid source (official site of Liverpool FC) hence, it was considered to be reliable.

Statistical Procedure

For the purpose of predicting the win probability of Liverpool FC on the basis of the independent variable taken Logistics Regression (binary logistics) was employed by IBM SPSS 23 and the level of significance was set at 0.05.

Results & Discussion

Table 1: Shows unweight cases n percent

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	100	100.0
	Missing Cases	0	.0
	Total	100	100.0
Unselected Cases		0	.0
Total		100	100.0

Findings of Table no1 reveals the number of included analysis in the study, as the total number of matches taken for the purpose of the study was 100, hence total selected cases analyzed and further processed in the study is 100 (N=100).

Table 2: Shows dependent variable encoding

Dependent Variable Encoding	
Original Value	Internal Value
win	0
loss	1

From Table no.2 one can find out the coding of the dependent variable done in the logistic model i.e. 0 for win and 1 for loss

Table 3: Shows categorical variables coding

Categorical Variables Coding			
	Frequency		Parameter coding (1)
ground	home	49	1.000
	away	51	.000

From the finding of Table no. 3 the categorical coding of the categorical variables used in the study is done, by home games it refers to matches played at Anfield stadium and away refers to the matches played outside Anfield stadium.

Block 0: Beginning Block

Table 4: Classification Table [a, b]

Observed			Predicted		
			Result		Percentage Correct
			Win	Loss	
Step 0	Result	Win	83	0	100.0
		Loss	17	0	.0
	Overall Percentage				83.0

a. Constant is included in the model.
b. The cut value is .500

The finding of Table no. 4 states that if nothing is known about the independent variables in the beginning block the model predicts the win probability by 83%.

Table 5: Variables in the equation

	B	S.E.	Wald	df	Sig.	Exp (B)
Step 0 Constant	-1.586	.266	35.476	1	.000	.205

The result of table no.5 reveals the significance of Wald statistics.

Block 1: Method= Forward Stepwise (Likelihood Ratio)

Table no. 6: Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	72.639[a]	.169	.283
2	54.435[b]	.307	.514
3	50.164[b]	.336	.562
4	44.879[b]	.371	.620
5	40.279[b]	.399	.667

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.
b. Estimation terminated at iteration number 20 because maximum iterations have been reached. Final solution cannot be found.

The finding of Table no. 6 reveals the values of pseudo R^2 which describes the goodness of fit of the model.

Table 7: Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	1.204	2	.548
2	1.375	5	.927
3	4.066	8	.851
4	2.415	8	.966
5	1.195	8	.997

Table no. 7 reveals the efficiency of test by the value of chi-Square.

Table 8: Classification Table [a]

Observed			Predicted		
			Result		Percentage Correct
			Win	Loss	
Step 1	Result	Win	83	0	100.0
		Loss	17	0	.0
	Overall Percentage				83.0
Step 2	Result	Win	74	9	89.2
		Loss	4	13	76.5
	Overall Percentage				87.0
Step 3	Result	Win	79	4	95.2
		loss	8	9	52.9
	Overall Percentage				88.0
Step 4	result	win	82	1	98.8
		loss	6	11	64.7
	Overall Percentage				93.0
Step 5	result	win	80	3	96.4
		loss	4	13	76.5
	Overall Percentage				93.0

a. The cut value is .500

In the above Table no.8 it can be seen the observed predicted values of dependent variables in all five models.

Table 9: Variables in the Equation

		B	SE	Wald	DF	Sig	Exp (B)
Step 1[a]	goal scored in 1st half	-1.770	.520	11.595	1	.001	.170
	Constant	-.398	.359	1.231	1	.267	.672
Step 2[b]	goal scored in 1st half	-1.652	.571	8.369	1	.004	.192
	goal scored by salah	-19.422	4261.392	.000	1	.996	.000
	Constant	.259	.419	.380	1	.538	1.295
Step 3[c]	ball possession	-.063	.032	3.794	1	.051	.939
	goal scored in 1st half	-1.924	.624	9.514	1	.002	.146
	goal scored by salah	-19.658	4089.272	.000	1	.996	.000
	Constant	4.066	2.029	4.017	1	.045	58.324
Step 4[d]	ball possession	-.194	.074	6.808	1	.009	.824
	passes completed	.014	.007	4.727	1	.030	1.014
	goal scored in 1st half	-2.053	.677	9.206	1	.002	.128
	goal scored by salah	-19.703	4083.284	.000	1	.996	.000
	Constant	3.523	2.106	2.798	1	.094	33.885
Step 5[e]	shots on target	-.375	.197	3.626	1	.057	.687
	ball possession	-.238	.093	6.526	1	.011	.788
	passes completed	.018	.008	5.149	1	.023	1.018
	goal scored in 1st half	-1.749	.656	7.105	1	.008	.174
	goal scored by salah	-19.981	3958.650	.000	1	.996	.000
	Constant	5.498	2.706	4.129	1	.042	244.117

a. Variable (s) entered on step 1: goal scored in 1st half
b. Variable (s) entered on step 2: goal scored by salah
c. Variable (s) entered on step 3: ball possession
d. Variable (s) entered on step 4: passes completed
e. Variable (s) entered on step 5: shots on target

The finding of Table no. 9 reveals the values of regression coefficients *B* along with the Wald statistics.

Results
The results of the binary logistics reveals that a total of 100 cases (matches) were processed in the model and from table no. 4 it can be seen that without adding the independent variables in the model, it can predict the win probability by 83%, from table no.5 the value of Wald statistics can be seen which states that the Wald statistics is significant at 0.05 level. The values of pseudo R^2 can be seen from Cox & Snell R Square and Nagelkerke R square which states the goodness of fit of the model which is 39.9% and 66.7%. It was also seen that the value of Chi-square is highly insignificant (.997) so, it can be concluded that the model if efficient. From table no.8 reveals the observed predicted model and it can be concluded that the variables depicts 93% of the overall win percentage. After the analysis of all the independent variables it was seen that the variables like shots on target, ball possession, passes completed, goal scored in 1st half and goal scored by Mohammed Salah were included in the model for prediction of win probability of Liverpool FC.

$$In\left(\frac{p}{1-p}\right) = Z = B_0 + B_1X_1 + B_2X_{2+}\dots\dots\dots B_nX_n$$

Hence, the developed equation for predicting the win probability is as follows;

Z= 5.498+ (-.375)*shots on target+ (-.238)*ball possession + (.018)*passes completed+ (-1.749)* goal scored in 1st half+ (-19.981)* Goal scored by Mohammed Salah.

Conclusion

From the above findings it can be concluded that the above generated model has high goodness of fit and also highly efficient, after the analysis the acquired variables which were included in the model for prediction were shots on target, ball possession, passes completed, goal scored in 1st half and goal scored by Mohammed Salah.

So, for prediction of the win probability of Liverpool FC on the basis of the above acquired variables the following equation was used which is stated below.

$$In\left(\frac{p}{1-p}\right)=Z=B_0+B_1X_1+B_2X_{2+\ldots\ldots\ldots\ldots\ldots}B_nX_n$$

Now, $p = e^z/1+e^z = e5.498 +(-.375)^*$ shots on target $+(-.238)^*$ ball possession $+ (.018)^*$ passes completed $+(-1.749)^*$ goal scored in 1st half$+ (-19.981)^*$ goal scored by Mohammed Salah / $1+5.498 +(-.375)^*$ shots on target $+(-.238)^*$ ball possession $+(.018)^*$ passes completed $+(-1.749)^*$ goal scored in 1st half $+(-19.981)^*$ goal scored by Mohammed Salah

So, in a hypothetical match situation when Mohammed Salah scores 2 goals, Liverpool scores 1 goal in the 1st half, total passes completed by the team is 500. The ball possession kept is 60% and total shots taken on the target is 10 then it can be predicted that Liverpool has chances of winning the match by 69.13%.

Reference

1. Alves AM, Ramos TG, Sant'anna AP, Soares DE, Mello JCCB. Uma proposta de previsão de resultados para o campeonato brasileiro de futebol através do modelo logito, Anais do SPOLM. Rio de Janeiro, *2008*.
2. Bloyce D, Murphy P. Sports administration on the hoof: the three points for a win 'experiment" in English soccer, soccer and society. 2008; 9:95-113.
3. Begley E. *BBC Sport*. Retrieved, 2020, 4-7, from bbc.com
   https://www.bbc.com/sport/football/50688370
4. Chanis A. Master Soccer mind, 2018. Retrieved 04 06, 2020, from Mastersoccermind.com
   https://mastersoccermind.com/17-reasons-why-soccer-is-the-most-popular-sport-in-the-world/
5. Calôba GM, Lins MPE. Performance assessment of the soccer teams in Brazil using DEA *Pesquisa Operacional*. 2006; 26:521-536.
6. Courneya KS, Carron AV. The home advantage in sport competitions: a literature review, Journal of sport and exercise psychology. 1992; 14:13-27.
7. Craig LA, Hall AR. Trying out for the team: Do exhibitions matter? Evidence from the national football league, Journal of the American statistical association. 1994; 89:1091-1099.
8. EPL. Premier League, 2019, 2020, 4-7.
   Retrieved, from premierleague.com
   https://www.premierleague.com/tables?co=1&se=210&ha=-1
9. Erozden C. AA Sports Retrieved, 2019, 12-24 04 07, 2020, from aa.com
   https://www.aa.com.tr/en/sports/liverpool-enjoy-2019-with-continental-treble/1682323
10. Lawal HB. Modelling the 1984-1993 American League Baseball Results as dependent categorical data Math Scientist. 2002; 27:53-66.
11. Lawal HB, Sundheim AR. Modeling 2002-2003 English premier league results IMA Sport, 2007, 115-124.
12. Liverpool F. *LFC*, 2019. Retrieved 04 07, 2020, from liverpoolfc.com
    https://www.liverpoolfc.com/match/2019-20/first-team/fixtures-and-results
13. Lee AJ. Modeling scores in the premier league: Is Manchester united really the best? Chance. 1997; 10:15-19.
14. Nevill AM, Newell SM, Gale S. Factors Associated with home advantage in English and Scottish soccer matches, Journal of sports science. 1996; 14:181-186.
15. Pollard R. Home advantage in soccer: a retrospective analysis, Journal of sports sciences. 1986; 4:237-248.
16. Pollard R. Evidence of a reduced home advantage when a team moves to a new stadium, Journal of sports sciences. 2002; 20:969-973.
17. Pollard R. Worldwide regional variations in home advantage in association football, Journal of sports sciences. 2006; 24:231-240.
18. Roche FP. La planificación estratégica en las organizations deportivas, 4th ed. Editorial Paidotribo, Barcelona, 2007.
19. Sant'anna AP. Rough sets analysis with antisymmetric and intransitive attributes: Classification of Brazilian soccer clubs. Pesquisa Operacional. 2008; 28:217-230.
20. Sant'anna AP, Barboza EU, Soares DE, Mello JCCB. Classification of the teamsin the Brazilian Soccer Championship by probabilistic criteria composition. Soccer and Society. 2010; 11:261-276.
21. Scarff PA, Shi X. The importance of a match in a tournament computers & operations research. 2008; 35:2406-2418.
22. Stefani RTE, Clarke SR. Predictions and home advantage for Australian Rules football, Journal of applied statistics. 1992; 9:251-261.
23. Watson HJ, Blackstone JR JH. Computer simulation, 2nd ed, Wiley & Sons, New York, 1989.