Predicting Final Result of Football Match Using Poisson Regression Model

IOP Conf. Series: Journal of Physics: Conf. Series **1108** (2018) 012066 doi:10.1088/1742-6596/1108/1/012066

Predicting Final Result of Football Match Using Poisson Regression Model

H R Azhari[1], Y Widyaningsih[2] and D Lestari[3]

[1,2,3] Departement of Mathematics, Universitas Indonesia, Kampus Baru UI, Depok, 16424, West Java, Indonesia.
(Email: lubishafis666@gmail.com, widyaningsihyekti@gmail.com, dian.lestari@sci.ui.ac.id)

Abstract. In any sport competition, there is a strong interest in knowing which team shall be the champion at the end of the championship and one of the sports is football. Football match predictions are of great interest to fans and sports press. In the last few years, it has been the focus of several studies. In this paper, the researchers propose Poisson regression model to predict the final result of football matches. The researchers predicted the average goals scored by each team by assuming that the number of goals scored by a team in a match followed a univariate Poisson distribution. Poisson regression model was formulated from four covariates: the goal average in a match, the home-team advantage, the team's offensive power, and the opponent team's defensive power. The methodology was applied to the 2017-2018 English Premier League. The results obtained using this model had a fairly good accuracy.

1. Introduction

Football is one of the most popular sports in the world which has its own charm compared to other sports. Football fans enchant in the tense and exciting moments of goals, especially for the last-minute goals, the final result, the match intensity, and the final rank. For example, the dark horse team Leicester City has won the 2014-2015 English Premier League. They surprisingly beat all the other strong teams. Real Madrid have also won three champions league title in a row, where no team has been able to do it before. This phenomenon reflects one of the most charming part of football that is complexity which makes the game result hard to be predicted.

Several papers are found in literature considering football score prediction applied to championship leagues such as the English Premier League ([1], [2], [3]), the Norwegian Elite Division [4] , the Brazilian Championship [5]. Lee [1] considered a Poisson regression to predict the number of goals from football team, where the average reects the strength of the team, the quality of the opposition and the home advantage (if it is the home team). The independence between the goals scored by the two teams was assumed and the methodology was applied to the 1995-1996 English Premier League. Brillinger [5] modeled the probabilities of win, tie and loss through an ordinal-value model and applied the model to the Brazilian Series A championship. Karlis and Ntzoufras [3] applied the Skellam's distribution to model the difference of goals between home and away teams. The authors illustrated the model using the 2006–2007 English Premier League. Koopman and Lit [6] developed a statistical model to predict the games of the 2010–2011 and 2011–2012 English Premier Leagues, assuming a bivariate Poisson distribution with coefficients that stochastically changed the intensity over time. Koopman and Lit [6] developed a statistical model to predict the games of the 2010–2011 and 2011–2012 English

Premier Leagues, assuming a bivariate Poisson distribution with coefficients that stochastically changed the intensity over time. An issue dealing with the papers cited above is that none of them considers the home team factor to calculate the probabilities of interest. Dixon and Coles [7] presented the result that 46% of the matches was won by the home team, 27% were draws and in 27% the home team lost.

In this paper, the researchers modeled the number of goal scored by each team in a match by a Poisson distribution, whose average reflected the strength of the attack and defense of the team and effect of being playing at home. The model was applied to the 2017-2018 English Premier League. The Definetti measure (DeFinetti [8]) was used to quantify the model predictive quality.

2. Model Construction

In this paper, the researchers assumed that the number of goals in each match followed the Poisson distribution so they can start constructing the regression models based on the assumption. The Poisson regression formula for this paper can be represented as:

$$Y \sim Poisson(\lambda) \quad (1)$$

$$Y = X\beta \quad (2)$$

From the formula, it can be seen that Y is a vector of dependent variable that consists of the home goals and away goals in games, X is a matrix of explanatory variables that records the home and away teams corresponding to the games, β is a vector containing the parameters, Offence and Deffence of the model. There were 20 teams participating in the 2017-2018 English Premier League and each of 20 teams had its offence parameter and defense parameter. Meanwhile, each of the times appeared as either a home team or an away team. Thus, it can be said that $Y = (y_{a,b}^{1}, y_{b,a}^{1}, y_{a,b}^{2}, y_{b,a}^{2}, \ldots, y_{a,b}^{n}, y_{b,a}^{n})^T$ *and* $\beta = (O_{afcbou}, O_{ars}, \ldots, O_{westham}, D_{afcbou}, D_{ars}, \ldots, D_{west}, \delta)^T$*, where* $y_{a,b}^{i}$ *is the number of goals scored by team a versus team b in game-i,* O_j *and* D_j *stands for the offence and defense parameter of team j and δ explains home advantage. The status of a team in a variable incidence takes value 1 if it participates in the i-match and takes value 0 if it does not participate. For example, Arsenal versus Westham in Emirates Stadium with score 2-1. The vector* $Y = [2\ 1]^T$ *and first row of matrix X is [0 1 … 0 0 0 … 1 1] because arsenal's score with home advantage and the second row is [0 0 … 1 0 1 … 0 0] . The reserachers considered* $\beta = (O_{afcbou}, O_{ars}, \ldots, O_{westham}, D_{afcbou}, D_{ars}, \ldots, D_{west}, \delta)^T = (\beta_1, \beta_2, \ldots, \beta_p)$ *and this time p equally to 41. Then the log-likelihood function for β is given as follows:*

$$l(\beta) = \sum_{i=1}^{n} \left\{ y_i \sum_{j=1}^{p} \beta_j x_{ji} - exp\left(\sum_{j=1}^{p} \beta_j x_{ji} \right) - log(y_i!) \right\} \quad (3)$$

From the function, it can be seen that y_i *is the number of goals scored in the-i observation,while* x_{ji} *is the component in matrix X in the j-column and the i-observation.*

Predictions

Consider that a game between teams a and b will occurs at home stadium of team a. Denote the probability of win, draw, and defeat (loss) of team a by P_w*,* P_d *and* P_l *respectively. These probabilities are given as follows:*

$$P_w = P(Y_a > Y_b) = \sum_{i=1}^{\propto} \sum_{j=0}^{i-1} P(X = \lambda_a) P(Y = \lambda_b) \quad (4)$$

$$P_d = P(Y_a = Y_b) = \sum_{i=0}^{\propto} P(X = \lambda_a) P(Y = \lambda_b) \quad (5)$$

$$P_l = P(Y_a < Y_b) = \sum_{j=1}^{\propto} \sum_{i=0}^{j-1} P(X = \lambda_a) P(Y = \lambda_b) \quad (6)$$

Similiarly to Bastos and da Rosa [9], and Suzuki et al. [10], the researchers in this study calculated the de Finetti distance in order to measure the beneficence of a prediction. This distance was given by the Euclidean distance between the point corresponding to the real outcome and the one corresponding to the prediction. For this case, just assumed that the set of all possible forecasts is given by the simplex set $S = \{(P_w,\ P_d\ , P_l) \epsilon \mathbb{R}^3 :\ P_w + P_d + P_l = 1.\ P_w \geq 0, P_d \geq 0, P_l \geq 0\}$. *The possible real outcome, including the win, draw, and loss are represented by the points (1,0,0),(0,1,0) and (0,0,1), respectively.*

The de Finetti measure (df) which is defined as:

$$df = (P_w - b_1)^2 + (P_d - b_2)^2 + (P_l - b_3)^2 \qquad (7)$$

where $(b_1, b_2, b_3) \in \{(1,0,0), (0,1,0), (0,0,1)\}$. *For example, if the prediction for the game between teams a and b is (0.2,0.65,0.15) and the real outcome is (0,1,0), the de Finetti distances is* $df = (0.2 - 0)^2 + (0.65 - 1)^2 + (0.15 - 0)^2 = 0.185$.

For the equiprobable case, $P_w = P_d = P_l = 1/3$, *with the win of the home team (1,0,0), the de Finetti measure is given by* $df = (1/3 - 1)^2 + (1/3 - 0)^2 + (1/3 - 0)^2 = 2/3$. *This value is accepted as a threshold value in order to classify the prediction as acceptable or not (see the example in Suzuki et al. (2010). If* $df < 2/3$, *the predictions are considered acceptable; otherwise, If* $df > 2/3$, *the predictions are considered poor.*

3. Application

3.1 Single Match Prediction

In this section, the researchers presented the predictions of the 32th round of the English Premier League (EPL). Table 1 showed that the probabilities of win, draw, and loss for the games, the score, and de Finetti measure if the method correctly indicated the winner team as the team with higher probability of a win.

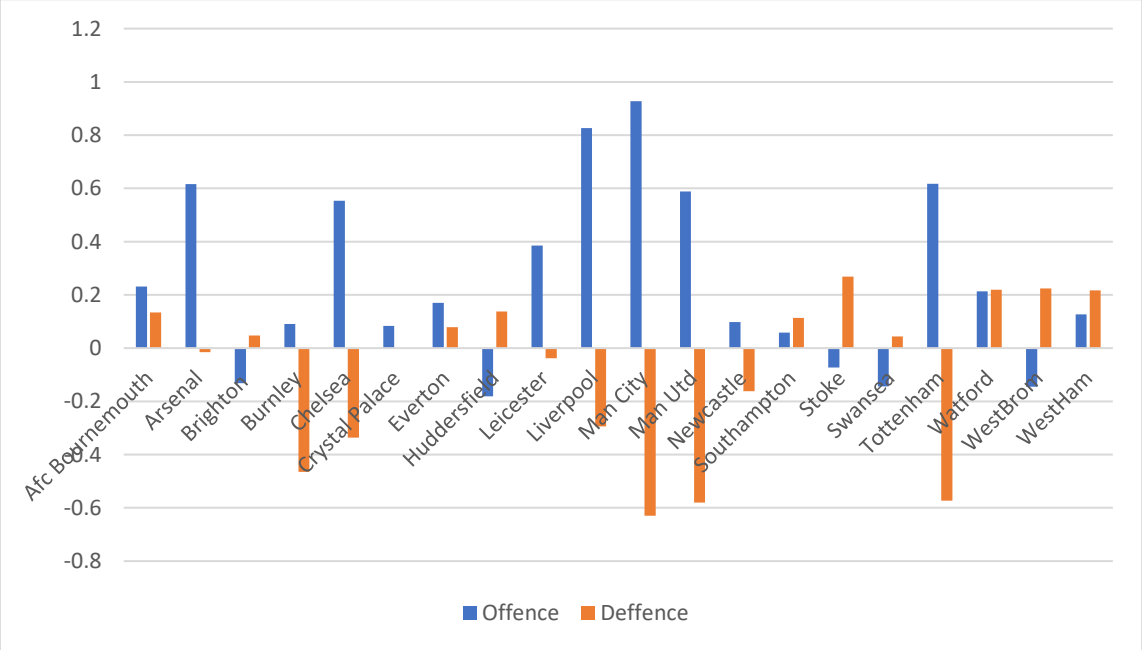

Figure 1. Offence and Defense effect

The proportion of correct prediction was 80%. The teams with an estimated probability of win higher than 0.5, were the actual winning team. Figure 1 displays the graphic of the attack and defense effect. In the graphic of Figure 1, it can be seen each attack and defense effect of each team. In offence effect,

the more positive the value, the stronger the effect of the attack is. Meanwhile, in the defensive effect, the value more negative then it indicates the persistence.

Table 1. Probabilities of win, draw, and loss for each match in 32th round.

Home	Away	Probability			Score	de Finetti	Correct
		Win	Draw	Loss			
Crystal Palace	Liverpool	0,166	0,189	0,644	1-2	0,191	Yes
Brighton	Leicester City	0,401	0,302	0,297	0-2	0,746	No
Manchester United	Swansea City	0,814	0,133	0,053	2-0	0,055	Yes
Newcastle United	Huddersfield	0,607	0,238	0,154	1-0	0,235	Yes
Watford	Bournemouth	0,453	0,222	0,325	2-2	0,916	No
West Brom	Burnley	0,225	0,278	0,497	1-2	0,381	Yes
West Ham	Southampton	0,463	0,238	0,299	3-0	0,434	Yes
Everton	Manchester City	0,083	0,046	0,87	1-3	0,026	Yes
Arsenal	Stoke City	0,817	0,112	0,071	3-0	0,051	Yes
Chelsea	Tottenham	0,27	0,227	0,502	1-3	0,372	Yes

3.2 Predictions for the whole Tournament

Manchester city has the greatest attack and defense effect, so it's only natural that this team won the EPL. The smallest defense effect decreased the expected number of goals of the opposing team. In the opposite, Stoke has the worst defense effect. The researchers estimated the number of points, the number of wins, draw and loss, the number of goals for and against each team. Table 2 presents these values and it is organized by the real number of points for each team. Note that the six teams with the highest-estimated number of points are really the six best teams of the championship.

Table 2. Predictions and real values.

Team	Points		Won		Draw		Lost	
	Est.	real	Est.	real	Est.	real	Est.	real
Manchester City	105	100	34	32	3	4	1	2
Manchester United	86	81	27	25	5	6	6	7
Tottenham Hotspurs	82	77	25	23	7	8	6	7
Liverpool	81	75	24	21	9	12	5	5
Chelsea	77	70	24	21	5	7	9	10
Arsenal	69	63	21	19	6	6	11	13
Burnley	61	54	17	14	10	12	11	12
Everton	49	49	14	13	7	10	17	15
Leicester City	60	47	17	12	9	11	12	15
Newcastle United	38	44	10	12	8	8	20	18
Crystal Palace	39	44	10	11	9	11	19	16
AFC Bournemouth	41	44	11	11	8	11	19	16
West Ham United	41	42	11	10	9	12	18	16
Watford	48	41	14	11	6	8	18	19
Brighton and Hove Albion	37	40	9	9	10	13	19	16
Huddersfield Town	33	37	9	9	6	10	23	19
Southampton	33	36	7	7	12	15	19	16
Swansea City	36	33	10	8	6	9	22	21

Stoke City	26	33	6	7	8	12	24	19
West Bromwich Albion	23	31	4	6	11	13	23	19

4. Final Remarks

Through this paper, the researchers proposed a simple method with good predictive quality. It has easy implementation and low computational effort for predicting match outcomes. The researchers developed a model to estimate the probabilities of win, tie and defeat in football games. In order to calculate these probabilities, they proposed Poisson regression model in which the average of goals scored reflected the strength of attack of the team, the strength of defense of the opposing team and the home team effect. The methodology was implemented in the software R. In this paper, the researchers estimated 100 games (29^{th}-38^{th} round), and the accuracy for our modelling was 61%. Although the model was applied to the 2017-2018 English Premier League. Principally, it was exible and it could be easily adapted to other different tournaments. However, it would be interesting if it can count every team's chances of winning the league or the chances of being degraded. It can be seen as a direct generalization of the researchers' proposed model and it may be investigated further by adding a significant parameter which can update the model so that it gets better in terms of its accuracy.

5. References

[1] Lee A J 1997 Modeling scores in the Premier League: is Manchester United really the best? *Chance* **10** pp 15–19
[2] Everson P, Goldsmith-Pinkham P 2008 Composite Poisson Models for Goal Scoring. *Journal of Quantitative Analysis in Sports* **4** (2)
[3] Karlis D and Ntzoufras I 2009 Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference *IMA Journal of Management Mathematics* **20** pp 133–145
[4] Brillinger D R 2006 Modelling Some Norwegian Soccer Data *Advance in Statistical Modelling and Inference (Ed. V. J. Nair.) World Scientic*pp 3-20
[5] Brillinger D R 2008 Modelling Game Outcomes of the Brazilian 2006 Series a Championship as Ordinal-valued *Brazilian Journal of Probability Statistics* **22** pp 89–104
[6] Koopman SJ and Lit R 2015 A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **178** pp 167–186
[7] Dixon MJ and Coles SG 1997 Modelling association football scores and inefficiencies in the football betting market *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **46** pp 265-280
[8] DeFinetti B 1972 *Probability, Induction and Statictics* John Wiley, London.
[9] Bastos L S and da Rosa J M C 2013 Predicting Probabilities for the 2010 FIFA World Cup Games Using a Poisson-Gamma Model *Journal of Applied Statistics* **40** pp 1533–44
[10] Suzuki AK, Salasar LEB, Leite JG, and Louzada-Neto F 2010 A Bayesian approach for predicting match outcomes: the 2006 (Association) Football World Cup *Journal of the Operational Research Society* **61** pp 1530–39

Acknowledgements

This research was funded by Directorate of Research and Community Service of Universitas Indonesia (DRPM UI) as a grant of PITTA (Publikasi Terindeks untuk Tugas Akhir) 2018, Number: 2336/UN2.R3.1/HKP.05.00/2018.