Statistical Modelling for The Prediction of Football Matches in German Bundesliga

RACHEL WANGECI MACHARIA[1], VINCENT NYONGESA MARANI[2], JOSEPH OUNO OMONDI[3]
[1, 2] Kibabii University, Bungoma, Kenya
[3] Masai Mara University, Narok, Kenya

Abstract- In this paper we have modeled football results using the Elo and Logistic regression ratings by incorporating the aspect of removal of players for the German Bundesliga for 2017/18 and 2018/19 seasons, estimated the expectations of teams' wins or losses and compared the model developed with Zacchary Andrews's model which was developed in 2019.

Indexed Terms- Statistical Modelling, Prediction of Football Matches

I. INTRODUCTION

The model incorporating the removal of player through issuance of red card(s) was developed as

$$e^H = \frac{1}{1+c^{(A_r - (H_r + \omega) - R_t)/d}} \quad (3.9)$$

where R_t is the difference in the removal of players for the away team and the home team given by:

$$R_t = R_A - R_H \quad (3.10)$$

Where R_A is the number of red cards given to the away team while R_H is the number of the red cards given to the home team. Similarly the away team rating is thus given by:

$$e^A = 1 - e^H \quad (3.11)$$

The expectations for the various teams were obtained from the equation 3.9 above. The ratings for various teams were obtained so that their expectations to be obtained. The elo ratings for various teams in the German Bundesliga were calculated. The results for the research paper were obtained and stored in Comma Separated Value (CSV) file containing predictions from each game.

II. EXPECTATIONS

After completion of the model, the outcomes were kept in a CSV (Comma Separated Value) file, some of which can be seen in table 4.2 below where records of teams playing at home, teams playing away, home win percentage, away win percentage and the actual observed result of the game were all included in the CSV file:

home team	away team	Home win (%)	Away win (%)	actual result
Borussia Dortmund	Darmstadt 98	90.71699362	9.283006385	home win
M'gladbach	Werder Bremen	82.7178017	17.2821983	home win
Augsburg	Mainz 05	48.2737476	51.72625244	away win
Augsburg	Darmstadt 98	75.23149936	24.76850064	home win
Borussia Dortmund	Hertha BSC	75.44540419	24.55459581	draw
M'gladbach	Hamburger SV	78.30774434	21.69225566	draw
Eintracht Frankfurt	Bayern Munich	2.303452269	97.69654773	draw
Hamburger SV	Eintracht Frankfurt	85.69591393	14.30408607	away win
Bayer Leverkusen	Hoffenheim	90.32184043	9.67815957	home win
Darmstadt 98	Wolfsburg	36.25916233	63.74083767	home win
Mainz 05	Ingolstadt	72.22345336	27.77654664	home win
Augsburg	Bayern Munich	7.358755612	92.64124439	away win
Köln	Hamburger	67.12405808	32.87594192	home win
Hertha BSC	M'gladbach	50.71950817	49.28049183	home win

Table 4.2 part of the CSV file that was generated to store our data

Few observations were made concerning games that resulted in draws because the model did not predict draw with high accuracy. For the model to indicate a draw then the expectation should be equal to 50.0000% which was not seen from the results. From the table 4.2 above, the team with the highest expectation was considered to be the predicted winner for each game in the testing dataset. From the table 4.2 above also it was found out that when Borrusia Dortmund and Darmstat 98 were playing the model had predicted that the Borussia Dortmund would win by 90.72% against its opponent 9.28%. It was noted that the draws were missing from this model for it was unable to predict draws with high accuracy. Considering also the match between Hertha BSC and M'gladbach, it was noted that the predictions were 50.7 and 49.3 respectively. Critical thinking will place this result more of a draw than a home win since the two teams had almost equal opportunity of winning.

III. MODEL ACCURACIES

The model's general prediction precision when forecasting a win, loss or draw was 56.32%. This was achieved by dividing the accurate scores for the data by the total observations then multiplying by one hundred to convert it into percentage. In other words, how many correct predictions against the total predictions. There were crucial values or accuracies that were got by running the model which can be better shown in table 4.3 below:

PARAMETER	VALUE
Home Field Advantage	80
Overall Model Accuracy	56.32%
Accuracy of Home Win	83.72%
Accuracy of Away Win	54.17%
Accuracy Draw	0.00%

Table 4.3 showing the parameter and their value.

From the table 4.3 above, the home wins, away wins and draws specific accuracies from the model was shown. The model was able to predict 83.72% of all games that resulted in a home win accurately. This accuracy for the home win was calculated by dividing the correct predictions for the home wins by the total number of actual home wins. The model also correctly predicted 54.17% of all the away wins that resulted in away wins. However, the model was unable to show the games that resulted in draws as shown in the table 4.3 above.

IV. COMPARISONS OF THE MODELS

Comparing the accuracies of formulated model with the Zacchary's model we found out that

PARAMETER	MY MODEL VALUE	ZACCHARY MODEL VALUE
Model Accuracy	56.32%	53.03%
Home Win Accuracy	83.72%	73.84%
Away Win Accuracy	54.17%	58.04%
Draw Accuracy	0.00%	0.00%

Table 4.4: comparison between the Zacchary's model and model developed

From the comparison above, it can be observed that adding the aspect of removal of players through red cards reduces the chances of a team winning in a football match. Therefore the issuance of red cards during football matches cannot be overlooked since it affects the outcome or result of our matches. We compared Zacchary's model with our model since we found his work was most recent and had added the aspect of home field advantage. His model had incorporated the home field advantage unlike other models formulated earlier which didn't add the aspect of home field advantage.

V. BOX PLOTS ANALYSIS

Analysis of the expectations of the home teams and away teams was also done and came up with the Box plots fig 4.2 below. The Box plots are usually plotted to assist a researcher to learn the distributional

characteristics of a group of scores as well as the level of scores. In a box plot we have five computations which we have discussed in detail. From the box plots below, the minimum values for both the teams at home and away were slightly different where the values were 0.023034523 and 0.030482845 respectively. Also the median for the two sets of data differed in such a way that the median for the teams at home was 0.625351393 while that of teams playing away was 0.374648607 which account for the comparatively overestimated home wins. The interquartile range for both home and away teams were equal, which was found to be 0.946482632. The data was also not normally distributed as the ones for the home teams was skewed to the left while for the away teams was skewed to the right. From the fig 4.2 below, we found out that size of the box plots differed. The difference could have been brought about by the over predicted home wins and can assume that our model predicted draws as home wins and therefore the box plot for the home team was bigger than that of the away team.

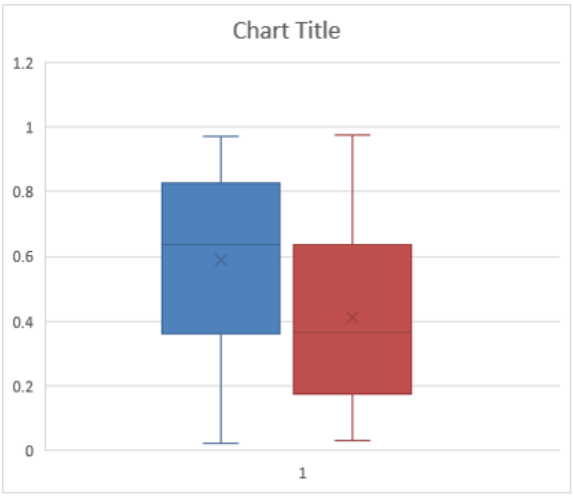

Fig 4.2 Box plot showing the expectation of the home and away teams respectively

CONCLUSION

The bias test results showed that β_1 was 1.737313203 and β_2 was -1.737313203. the bias then was zero showing that the formulated model was unbiased hence the results. From the calculations it was noted that the model was unbiased hence it was concluded that the estimates were unbiased hence the model automatically gives unbiased results.

The test for the validation of the model was determined by using the ranked probability score (RPS) which is a measure of how good forecasts matches the observed outcomes when expressed as probability distributions. The RPS formula was used below:

$$\text{RPS} = \frac{1}{r-1}\sum\left(\sum_{j=1}^{i}(p_j - e_j)\right)$$

Where r is the number of outcomes (that is win, loss or draw), P_j is the forecasted probabilities of outcome j and e_j is the actual probability of outcome j. from our data we were able to obtain the ranked probability score for both the team playing at home and the team playing away from its home field as summarized below. RPS for the home team was 0.5053 while RPS for away team was 0.2757 showing that the model was valid. The model's validity is great when forecasting the away expectations as compared to when forecasting the home team.

REFERENCES

[1] Arntzen, H., & Hvattum, L. M. (2020). Predicting match outcomes in association football using team ratings and player ratings. *Statistical Modelling*, 1471082X20929881.

[2] Hvattum L. M. and Arntzen H. (2010). Using ELO ratings for match result prediction in association football," International Journal of Forecasting, vol. 26, pp. 460{470, Jul. 2010.

[3] Zacchary A. (2019). Comparing Predictive Models for English Premier League Games.