A PLS-SEM Approach for Composite Indicators: An Original Application on the Expected Goal Model

Mattia Cefis(✉)

University of Brescia, Contrada Santa Chiara 50, Brescia, Italy
mattia.cefis@unibs.it

Abstract. In the field of football analytics, the goal is to improve (in terms of prediction performance) one of the emerging tools: the expected goal (xG) model. With this final aim, data from different sources have been merged: tracking data, match event data and some players' performance composite indicators obtained using a Partial Least Squares - Structural Equation Model (PLS-SEM) approach. Using a sample of match data relying to season 2019/2020 of the Italian Serie A, composed by 1 outcome variable (i.e. the GOAL) and 22 features, a logistic regression model was applied on different scenarios for sample balanced techniques. Results seem to be interesting in terms of sensitivity, F1 and AUC metrics, compared with a benchmark. In addition, some original performance composites and tracking variables introduced are significant for the classification model.

Keywords: Expected Goal · Logistic Regression · Imbalanced Sample · PLS-SEM

1 Introduction

Football (also called soccer, in North America, Australia and New Zealand) is one of the most popular sports in the world. Due to its appeal, football teams are treated more and more as firms: as a consequence, decisions about their management (coach, technical staff and scouting) are becoming strategic, and, for this reason, in the last few years football has been moving towards a data-driven approach [4]. To predict the final score of a football game is one of the cases where a data driven approach is more adopted. The final score might be predicted by summing the probability of score a goal every time a player shoots the ball. This involves the adoption of many features; this aspect has to do with artificial intelligence, as, by matching a proper method with a set of determinant features, might permit to predict accurately the result of a game. By this work, the idea is to refine and improve, in terms of prediction accuracy, the well-known expected goal (xG, [1]) model that surpasses the most basic and frequently used

Supported by Math&Sport.

M. Clayton et al. (Eds.): INTETAIN 2023, LNICST 560, pp. 127–135, 2024.
https://doi.org/10.1007/978-3-031-55722-4_10

metric in football to summarise the team performance: the shot, which does not consider the quality of the goal-scoring opportunity from which it arises.

2 Literature Overview and Data Employed

The main idea of the xG model is to assign a value between zero and one to each shot; this value represents the probability of a shot resulting in a goal, using a machine learning probabilistic classifier [17]. During the last years, the xG model has become increasingly popular, and it is used more and more in the football world as a proxy for measuring players' finalisation performance and teams' offensive strength during a match [9]. For this reason, some studies and websites have treated this topic: for example, someone [16,19] examined shots, taking in consideration only distance and angle to goal, whereas another one [9] made a spatial analysis of shots of the Mayor League Soccer, using a logistic regression. Another recent work [18] tried to quantify the effectiveness of defensive playing styles in the Chinese Football Super League by using xG. The main deficiency is that current xG models are based just on event data and do not take into account players' features. As an innovation, the objective is to improve the xG model by adding some composite indicators related to the players' performance and obtained by a Partial Least Squares Structural Equation Model (PLS-SEM, [10]), in order to take into account shooters' and goalkeepers' features [5,6] (Fig. 1). These composite indicators have been validated by a Model Invariance COMposite (MICOM) approach [11,12].

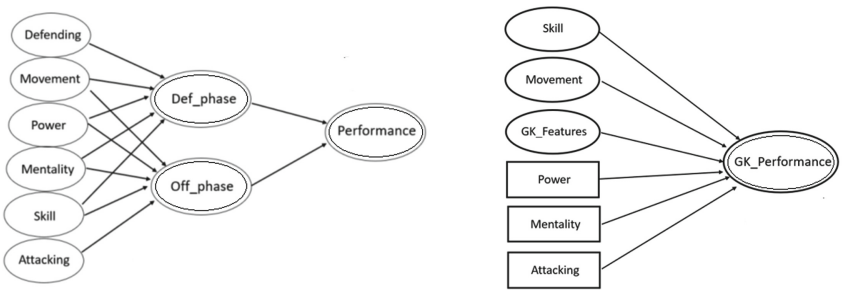


Fig. 1. Players' movement and Goalkeepers models used for the PLS-SEM approach.

In order to reach the aim, a merging between data coming from different sources (e.g. Understat[1] for event data, Math&Sport for tracking data and *Sofifa*[2] for building the players' performance indicators) was done. The final dataset was composed of a sample of 600 shots, 23 variables (e.g. 1 binary outcome and 22 regressors, Table 1) coming from 50 official football matches of the

[1] www.understat.com.
[2] www.sofifa.com.

season 2019/2020 (Italian Serie A league). In particular, there are three binary features: GOAL (i.e. the outcome, 1 = goal and 0 = no goal), previous dribbling before the shot (1 = yes, 0 = no) and favourite foot, i.e. if the player shoots on goal with his favourite foot or not: 1 = yes, 0 = no. In addition, the variable that counts the number of opponents around the shooter was converted into two dummies: the first one (i.e. D1.OpponentsPlayer) is equal to 1 when the opponents number is greater than 0, whereas the second one (D2.OpponentsPlayer) is active when that number is greater than 1; the other features are continuous.

Table 1. Statistics of the variables for the sample of 660 shots from the 50 matches of the Italian Serie A 2019/2020.

Variable description	Source Dataset	Mean	Std.	Q1	Q2	Q3
GOAL (Yes/No)	Understat	0.10	0.30	-	-	-
x (shooter coordinate, %)	Understat	83.22	7.64	77.00	83.00	89.00
y (shooter coordinate, %)	Understat	49.98	15.06	37.00	50.00	63.00
Favourite foot (Yes/No)	Understat	0.77	0.42	-	-	-
Previous dribbling (Yes/No)	Understat	0.32	0.47	-	-	-
Angle of shot (degree)	Understat	37.39	20.81	20.22	39.42	52.80
Previous ball distance (%)	Tracking	14.43	8.41	8.60	14.33	17.09
Possession duration (sec.)	Tracking	6.49	4.81	5.68	6.49	6.49
D1.OpponentsPlayer	Tracking	0.13	0.33	-	-	-
D2.OpponentsPlayer	Tracking	0.05	0.21	-	-	-
GK x coordinate (%)	Tracking	96.74	5.25	96.55	97.67	98.41
GK y coordinate (%)	Tracking	50.08	5.06	47.37	50.04	53.07
Defending	Sofifa (PLS-SEM)	0.06	0.99	−0.62	0.14	0.61
Mentality	Sofifa (PLS-SEM)	0.38	1.00	−0.28	0.41	1.17
Movement	Sofifa (PLS-SEM)	0.41	1.03	−0.27	0.32	1.00
Power	Sofifa (PLS-SEM)	0.49	1.06	−0.16	0.53	1.10
Skill	Sofifa (PLS-SEM)	0.47	0.91	−0.15	0.49	1.12
GK Attacking	Sofifa (PLS-SEM)	−0.03	1.01	−0.78	0.2	0.77
GK Features	Sofifa (PLS-SEM)	0.68	0.78	0.32	0.82	1.12
GK Mentality	Sofifa (PLS-SEM)	0.26	0.98	0.08	0.43	0.82
GK Movement	Sofifa (PLS-SEM)	0.55	0.73	0.12	0.53	1.02
GK Power	Sofifa (PLS-SEM)	0.45	0.92	0.10	0.45	1.19
GK Skill	Sofifa (PLS-SEM)	0.06	0.79	−0.51	−0.19	0.52

3 The Frameworks Developed

From a methodological point of view, since the target, namely the Goal, is a rare event [16], a logit model (LM) with parameters estimated by maximum likelihood

[14] was applied on different samples for three situations: the basic Imbalance sample and two machine learning sample-balanced techniques, such as random oversampling examples (ROSE, [15]) and the synthetic minority oversampling technique (SMOTE, [7]). The benchmark adopted was the xG value provided by Understat. From a practical point of view, it was developed a routine in *R* by using a stratified 3-Fold cross validation for evaluating the model fit and computing the performance measures, with 5,000 replications (*R* packages *ROSE* and *RSBID*). The LM model was preferred because of its easy interpretation concerning the regressors effects and since the real focus was to introduce new predictors in the xG model, in order to improve the goal probability estimation. In the context of the xG, this model lets to estimate the conditional probability of goal for any shots and their set of features values $\mathbf{X}$, and estimate parameters $\hat{\beta}$ in the next Eq. (1). Note that the regression coefficients are estimated by maximum likelihood [14]. Typical classification metrics have been used to assess the models performance [13].

$$xG = P(Goal|\mathbf{X}) = \frac{e^{\mathbf{X}\hat{\beta}}}{1 + e^{\mathbf{X}\hat{\beta}}} \tag{1}$$

In addition, balancing a dataset thanks to some techniques like ROSE or SMOTE, we must take into account the effects of our modifications to the training data [2,8]. Due to the Bayes' Theorem, posterior probabilities are proportional to the prior ones, which can be estimated as the relative frequencies in the respective categories. Therefore, the estimated posterior probability (expected goal) obtained using artificially balanced data set can be corrected (calibrated, xG^*) using the following formula (2):

$$xG^* = \frac{\frac{0.1}{0.5}xG}{\frac{0.1}{0.5}xG + \frac{(1-0.1)}{(1-0.5)}(1 - xG)} \tag{2}$$

Take into account that in this case 10% and 50% are respectively the real and the artificial (balanced) sample proportion of the rare class.

4 Results and Discussion

For what concern the output of the logit, the regression coefficients are presented in Table 2: take in mind that significant coefficients are emphasized by asterisks (e.g. * = 5%, ** = 2.5%, *** = 1%); the main significant estimated regressors (for all the approaches) are the shooter position (in particular, the x coordinate), the Angle of shot, and some new variables introduced: the D2.OpponentsPlayer, the goalkeeper position (GK x coordinate and Gk y coordinate) and some performance features like the shooter Movement ability, GK_mentality and GK_skill related the goalkeeper.

Table 2. The Logistic Regression coefficients estimates after 5,000 replications for the 50 matches of the Italian Serie A 2019/2020.

Regressor	Coeff. ROSE	Coeff. SMOTE	Coeff. Imbl.
x	0.94***	2.09***	1.74***
y	0.02	−0.05	−0.09***
Favourite foot	−0.16	0.10	0.17***
Previous dribbling	0.21*	0.33*	0.46***
Angle of shot	−0.53***	−1.40***	−1.09***
Previous ball distance	−0.09*	−0.22*	−0.06**
Possession duration	−0.18	−0.25*	−0.22***
D1.OpponentsPlayer	−0.74	−1.02*	−0.71
D2. OpponentsPlayer	−5.69***	−6.13***	−5.68***
GK x coordinate	−0.33***	−0.59**	−0.33***
GK y coordinate	0.27***	0.48***	0.46***
Defending	0.14*	0.13	0.16***
Mentality	0.03	−0.23	−0.40***
Movement	0.12*	0.40*	0.27***
Power	−0.01	−0.07	0.14***
Skill	0.08	0.23*	0.21***
GK_Attacking	0.02	0.05	0.03*
GK_features	0.03	0.32*	0.22***
GK_Mentality	−0.15*	−0.44**	−0.33***
GK_Movement	0.10*	0.15	0.07**
GK_Power	−0.17*	−0.20	−0.17**
GK_Skill	0.10*	0.27**	0.20***

The second part of the results is focused on the performance metrics: in order to do this, we must take into account that for binary classification problems usually the probabilistic classifier (xG) is transformed in the categorical one (Goal, NoGoal) using the threshold 0.5. In Table 3 are proposed all the classification performance metrics and their average scores (5,000 replications), comparing them with the benchmark (directly provided from Understat). Take in consideration that asterisks in Table 3 must be interpreted as a value statistically significant different (e.g. * = 5%, ** = 2.5%, *** = 1%) from the benchmark; the metrics that outperform the benchmark are emphasized in bold. Both ROSE and Imbalance significantly outperform in terms of specificity and precision the benchmark (Understat) whereas SMOTE seems able to better detect the goals (sensitivity equals to 0.36 vs 0.16 of the benchmark) and to improve Understat in terms of F1 (the armonic mean between precision and sensitivity) and precision. The Area Under the Receiver Operating Characteristic Curve (AUC) metric is

similar for all the situations, except for the Imbalance, that significantly outperforms Understat (0.74 vs 0.72).

Table 3. The performance classification metrics compared with the benchmark (classification threshold = 0.5).

Metric	ROSE	SMOTE	Imbalance	Understat
Accuracy	0.89*	0.86***	0.90	0.91
Sensitivity	0.14*	**0.36*** **	0.15	0.16
Specificity	**0.98***	0.91***	**0.98***	0.96
Precision	**0.35***	**0.32****	**0.51*** **	0.22
F1	0.16	**0.33*** **	0.23	0.19
AUC	0.72	0.73	**0.74***	0.72

Now, in order to emphasize the importance of the new regressors introduced in the model, two real situations will be proposed, comparing the expected goal for each framework and introducing some variation, in order to better understand how the xG changes.

In the first real case (Fig. 2) a goal scored from a high distance was proposed, in a situation with a high number of opponents in front of the shooter. Then the expected goal for each situation (xG^* for ROSE and SMOTE) and others two scenarios (Table 5) were proposed: the first one putting a top player as shooter (Ronaldo, Juventus), whereas the second one leaving the same top player as shooter and considering a normal goalkeeper. Let's see how in the real scenario the balanced frameworks increase the estimated goal-scoring probability (higher xG than the benchmark); the xG for the imbalance approach is very similar. It's interesting to note that introducing firstly a top player as shooter (Scenario 1), then a normal goalkeeper (Scenario 2) the expected goal increases in both three situations, emphasizing the importance of introducing players' performance indices in the model, as innovation of this work (Table 4).

Table 4. The expected goal for each situation and different cases

Case	Shooter	GK	xG^* ROSE	xG^* SMOTE	xG Imbalance	xG Understat
Real	Soriano (Bol)	Handanovic (Int)	4.1%	2.3%	2.0%	2.1%
Scenario 1	Ronaldo (Juv)	Handanovic (Int)	5.4%	3.3%	3.1%	2.1%
Scenario 2	Ronaldo (Juv)	Skorupski (Bol)	7.1%	4.4%	4.5%	2.1%

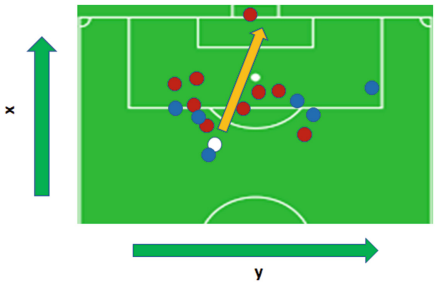
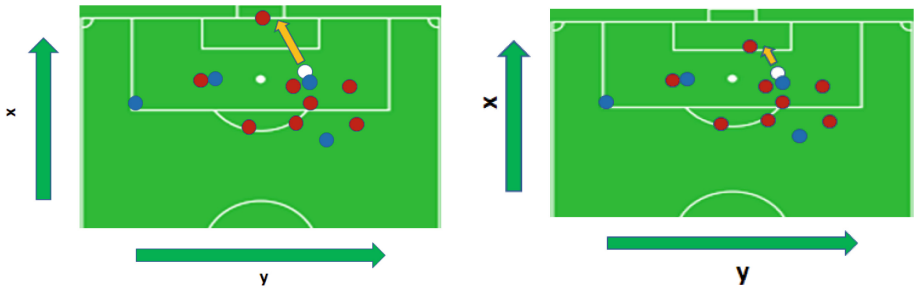


Fig. 2. Real case 1: goal from the distance

In the second real case (Fig. 3a), it has proposed a goal scored from a short distance in a very favourable situation (Atalanta vs Brescia, July 2020). Here, the alternative scenario considers the goalkeeper's position: whereas in the real case he is aligned at his own goal post, in the alternative scenario he is slightly out of the goal (Fig. 3b). Let's see in the real case 2 that the balanced frameworks increase the estimated goal-scoring probability (higher xG than the benchmark), with SMOTE and imbalance having a similar xG. It is interesting to note that moving the goalkeeper position outside the box (Fig. 3b) increases the xG of the shot.



(a) Real case 2: goal near the box. (b) Scenario 1: moving the GK position.

Fig. 3. The second real case and its alternative scenario on the pitch.

Table 5. The real case 2: expected goal for each framework and its alternative scenario.

Situation	Shooter	GK	xG^* ROSE	xG^* SMOTE	xG Imbl.	xG Understat
Real case 2	Pasalic (Ata)	Andrenacci (Bre)	14.6%	22.1%	21.7%	11.0%
Scenario 1	Pasalic (Ata)	Andrenacci (Bre)	30.0%	58.5%	52.2%	11.0%

5 Conclusion

This contribution proposes an improvement of the current expected goal (xG) model, one of the emerging tools in the field of football analytics. The main idea behind this model is to assign a quality metric (probability) for a goal for each shot. The main lack of the current xG frameworks is that they take into consideration only the classical event data: in order to overcome this weakness, the xG model has integrated some players' performance composite indicators obtained from a PLS-SEM approach and some tracking data, such as the goalkeeper's position and the number of opponents. As summary, the original approach presented in this work seems to suggest that some performance composite indicators and some tracking variables are helpful to detect the goals, refining the benchmark model (in the sense that the prediction in the extended model depends also on such additional features): by replacing a shooter with a better one (or reducing the abilities of the goalkeeper), one increases the estimated goal-scoring probability, which looks quite a reasonable outcome (which is not captured by the baseline model). As future works, it should be interesting to examine in-depth this topic by a larger sample size, and maybe comparing others classification models (for example, Gompit [3] or Classification Trees) performances. Further developments might also regard the generalization of the adopted methodology to other sports, for example in estimating the probability to make a shoot in basketball.

Acknowledgements. I want to thank Math&Sport s.r.l. for providing the tracking data for the Italian Serie A matches used in the statistical analysis.

References

1. Anzer, G., Bauer, P.: A goal scoring probability model for shots based on synchronized positional and event data in football (soccer). Front. Sports Active Living **3**, 53 (2021)
2. Bishop, C.M., Nasrabadi, N.M.: Pattern Recognition and Machine Learning, vol. 4. Springer, New York (2006)
3. Cameron, A.C., Trivedi, P.K., et al.: Microeconometrics Using Stata, vol. 2. Stata Press, College Station (2010)
4. Cefis, M.: Football analytics: a bibliometric study about the last decade contributions. Electron. J. Appl. Stat. Anal. **15**(1), 232–248 (2022)
5. Cefis, M., Brentari, E.: Formative vs reflective constructs: a CTA-PLS approach on a goalkeepers' performance model. Book of the Short Papers, 51st Scientific Meeting of the Italian Statistical Society, pp. 323–328 (2022)
6. Cefis, M., Carpita, M.: The higher-order PLS-SEM confirmatory approach for composite indicators of football performance quality. Comput. Stat. 1–24 (2022)
7. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
8. Dal Pozzolo, A., Caelen, O., Johnson, R.A., Bontempi, G.: Calibrating probability with undersampling for unbalanced classification. In: 2015 IEEE Symposium Series on Computational Intelligence, pp. 159–166. IEEE (2015)

9. Fairchild, A., Pelechrinis, K., Kokkodis, M.: Spatial analysis of shots in MLS: a model for expected goals and fractal dimensionality. J. Sports Anal. **4**(3), 165–174 (2018)
10. Hair Jr, J.F., Hult, G.T.M., Ringle, C.M., Sarstedt, M., Danks, N.P., Ray, S.: Partial least squares structural equation modeling (PLS-SEM) using R: a workbook (2021)
11. Hair Jr, J.F., Sarstedt, M., Ringle, C.M., Gudergan, S.P.: Advanced Issues in Partial Least Squares Structural Equation Modeling. Sage Publications, Thousand Oaks (2017)
12. Henseler, J., Ringle, C.M., Sarstedt, M.: Testing measurement invariance of composites using partial least squares. Int. Mark. Rev. **33**(3), 405–431 (2016)
13. Hossin, M., Sulaiman, M.N.: A review on evaluation metrics for data classification evaluations. Int. J. Data Mining Knowl. Manag. Process **5**(2), 1–12 (2015)
14. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning, vol. 112. Springer, New York (2013). https://doi.org/10.1007/978-1-4614-7138-7
15. Menardi, G., Torelli, N.: Training and assessing classification rules with imbalanced data. Data Min. Knowl. Disc. **28**(1), 92–122 (2014)
16. Rathke, A.: An examination of expected goals and shot efficiency in soccer. J. Hum. Sport Exerc. **12**(2), 514–529 (2017)
17. Robberechts, P., Davis, J.: How data availability affects the ability to learn good xG models. In: Brefeld, U., Davis, J., Van Haaren, J., Zimmermann, A. (eds.) MLSA 2020. CCIS, vol. 1324, pp. 17–27. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-64912-8_2
18. Ruan, L., Ge, H., Shen, Y., Pu, Z., Zong, S., Cui, Y.: Quantifying the effectiveness of defensive playing styles in the Chinese football super league. Front. Psychol. 1–10 (2022)
19. Umami, I., Gautama, D.H., Hatta, H.R.: implementing the expected goal (xG) model to predict scores in soccer matches. Int. J. Inf. Inf. Syst. **4**(1), 38–54 (2021)