An innovative xG Model for football analytics

Mattia Cefis & Maurizio Carpita

Department of Economics and Management,

Contrada S. Chiara, 50 – 25122, Brescia. - University of Brescia (IT)

Abstract

In the field of football analytics, we want to improve (in terms of prediction performance) one of the emerging tool: the expected goal (xG) model. With this final goal, we merged match event data with some players' performance composite indicators obtained using a Partial Least Squares - Structural Equation Model (PLS-SEM). Using a sample of match tracking data relying to season 2019/2020 of the Italian Serie A, composed by 660 shots and 25 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 indices, compared with a benchmark.

KEYWORDS: EXPECTED GOAL, PLS-SEM, LOGISTIC REGRESSION, IMBALANCED SAMPLE.

Introduction

The expected goal (xG) models are more and more used in the football world as proxy for measuring players' finalization performance and teams' offensive production during a match (Fairchild et al., 2018). The main lack is that currently xG models are based just on event data and do not take in consideration the sample imbalanced, since the target (i.e. the "goal") is a rare event (Rathke, 2017). The aim of this study is to merge data from different sources (e.g. Understat - understat.com- for event data, Math&Sport - mathandsport.com- for tracking data and Sofifa for the players' performance indicators) for improving the xG in terms of model sensitivity and performance (Robberechts & Davis, 2020). The initial dataset was composed by a sample of 660 shots and 30 features for each-one, relying the season 2019/2020 of the Italian Serie A.

Methods

As preliminary step, six covariates with problematic collinearity problems were removed. Take in mind that some covariates refer to different composite latent traits of players' performance and have been previously estimated by a PLS-SEM (Carpita et al., 2021 ; Cefis & Carpita, 2021). A logistic regression model was applied on different samples scenarios, by splitting randomly the dataset in training and test set (75%-25%), using different machine learning sample-balanced techniques (Menardi & Torelli, 2014; Chawla et al., 2002): oversampling, undersampling, SMOTE and ROSE. Mean results after 1000 replications are summarized in Tab. 1. The benchmark adopted was the xG model by Understat (understat.com) and the software used for the analysis is R (version 4.1.3, r-project.org).

Typical classification metrics have been used to assess the models performance (Hossin & Sulaiman, 2015).

Results

Table 1. Models performance comparison between the 4 balanced sample techniques, the imbalalnced approach and the benchmark using logistic regression (1= goal, 0= no goal): mean scores after 1000 resampling.

Indices	Oversam.	Undersam.	ROSE	SMOTE	Imbal.	Understat
Accuracy	0.75	0.67	0.65	0.78	0.90	0.91
Sensitivity	0.69	0.72	0.75	0.62	0.15	0.14
Specificity	0.76	0.66	0.64	0.79	0.98	0.96
F1	0.34	0.29	0.28	0.34	0.21	0.19
AUC	0.80	0.76	0.77	0.79	0.80	0.72

Discussion

In Tab. 1 we can see how the four balanced-approaches outperform in terms of sensisitvity and F1 index the imbalanced and the benchmark (i.e. understat). Integrating tracking and players' performance data seems to improve also the AUC index despite the benchmark. It could be interesting to in-depth this analysis with other seasons and leagues or applying other classification algorithms.

Conclusion

The main result of this study suggests us that including new features in the xG model could improve it in terms of goal-detection (i.e. sensitivity) but also for the global model performance (AUC), helping in a more accurate way football insiders in players' and teams' evaluation. In addition we want to thank the BDSport Lab (bodai.unibs.it/bdsports) for the financial support.

References

Carpita, M., Ciavolino, E., & Pasca, P. (2021). Players' role-based performance composite indicators of soccer teams: A statistical perspective. Social Indicators Research, 156(2), 815-830.

Cefis, M., & Carpita, M. (2021). Football analytics: a Higher-Order PLS-SEM approach to evaluate players' performance. *Book of Short Papers SIS 2021*, 508-513.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

Fairchild, A., Pelechrinis, K., & Kokkodis, M. (2018). Spatial analysis of shots in MLS: a model for expected goals and fractal dimensionality. *Journal of Sports Analytics*, *4*(3), 165-174.

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process, 5(2), 1.

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. Data mining and knowledge discovery, 28(1), 92-122.

Rathke, A. (2017). An examination of expected goals and shot efficiency in soccer. *Journal of Human Sport and Exercise, 12*(2), 514-529.

Robberechts, P., & Davis, J. (2020). How data availability affects the ability to learn good xG models. In *International Workshop on Machine Learning and Data Mining for Sports Analytics*, 17-27. Springer, Cham.