Quantifying the effects of individual player on goal scoring through a Bayesian

Barba Paolo

July 3, 2023

Introduction (1)

Nowdays, football clubs have started to analyze advanced metric for player performance analysis that can lead in

- player scouting
- decision making of player's contract.

Introduction (2)

While comparing teams or player, the expected goal metric provides a good statistic in the tactical match analysis.

The xG model is a probabilistic model that assign score between 0 and 1 from any observed shot in a match. The model has development using event-level football data from StatsBomb's data.

Usually, xG models do not account fot the players who take the shots, and this assumptions does not seem suit- able, since the player's skill could influence the succes of shot- conversion.

Dataset

The data are collected by tracking players over the course of football matchs and recording their actions such shots, passes or others. For the xG model the data were taken from the shot event. These data, collected from StatsBomb are spatially manipulated in order to colecct relevan informations of the shots

Slide with Plot

#TODO mettere la immagine del plot

Aim

The aim of the analysis is to understand:

- The variables that can influence the shots result
- If the player' skill influence the shot results

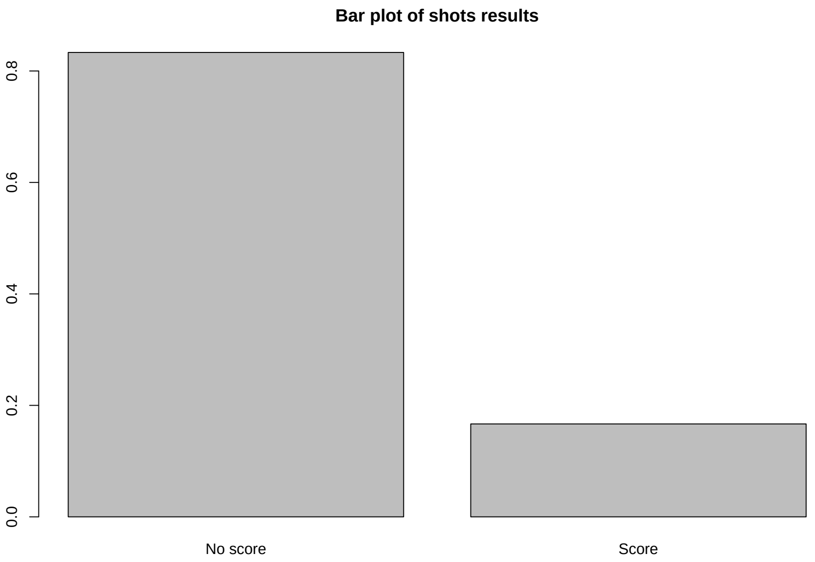
Before going any further on the analysi, it is need to describe the variables that we are going to use for conducting this analysis

Dataset description

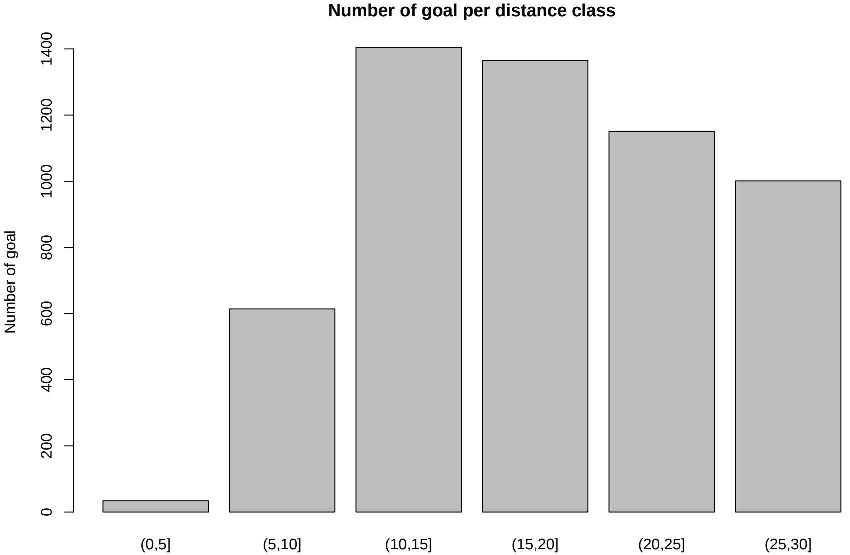
The dataset is composed by ... variable and ... observations that reppresent shots of different 42 player of Barcelona F.C.

```
##     X statsbomb_xg  goal time                           player
## 1  1  0.072952810 False    5 Lionel Andrés Messi Cuccittini
## 2  2  0.015879327 False    6                    Thierry Henry
## 3  3  0.040939737 False   15 Lionel Andrés Messi Cuccittini
## 4  4  0.101285940 False   16             Gnégnéri Yaya Touré
## 5  6  0.007017402 False   22              Daniel Alves da Silva
## 6 10  0.008805739 False   32                    Thierry Henry
##   shot_distance inside_18 shot_angle   bodypart technique f
## 1      17.14060      True     145.94 Right Foot    Ground
## 2      31.16168     False      59.32 Right Foot    Ground
## 3      17.69802      True     137.75  Left Foot    Ground
## 4      19.59719     False      80.01  Left Foot    Ground
## 5      41.74686     False      72.43 Right Foot    Ground
## 6      20.82402     False      16.17  Left Foot    Ground
```

Explanatory Analysis



Explanatory Analysis (2)



TODO Titolo

A reasonable assumption is that player has a unique type of shooting. Therefore, given a player, the shots are inerehntly correlate to each each other and behave indipendently form the shot of another player. This introduces within-player correla- tions (event-level data is longitudinal data). As a result, every shot is essentially grouped or nested under a player. Therefore, models need to be fed information about the hierarchical struc- ture, otherwise it may lead to biased infe

Example of the hierarchical structure

TODO mettere l'immagine..

Model specification

For targeting the target variable, two models are choosen.

- GLM do not fed information about the hierarchical structure
- GLMM do fed information about the hierarchical structure

Generalized linear model specifications

$$g(\pi_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

where the g is the logit link function and the π_i is the odds for a goal $\frac{Pr(Y=1)}{Pr(Y=0)}$.

Generalized linear mixed model specifications

In order to fed the information about the players structure a Linear mixed effect model only with a random intercept it is choosen.

Let us define the equation of the model and some basic notion:

$$g(\pi_{i,j}) = \beta_0 + \beta_1 x_{i,j,1} + \beta_2 x_{i,j,2} \cdots \beta_p x_{i,j,p} + \delta_j$$

- $g = 1, 2, \dots 42$ number of players
- $i = 1 \dots n_g$ Number of shots done by the $g - th$ player
- $\pi_i, g =$ probability of score from the $i - th$ shot and the $g - th$ player
- $\pi_g =$ odds ratio of scoring for the $g - th$ player. $\in (n_g \times 1)$

Dimensionality of the parameters

Let us define the dimensionality for the model:

- $X_{g,i} \in (k \times 1)$ where k is the number of predictors
- $\underline{\beta} \in (k \times 1)$ can be considered as a vector of fixed effect since it does not depend on the player.
- δj represent the random effect linked to the g-th player
- The vector $\underline{\epsilon_g}$ is the vector of error terms.

Sampling model assumptions

The ϵ_g indipendent $b_g \forall g = 1 \cdots 42$ The b_g are noramlly distribuited with their own variance The ϵ_g are noramally distribuited

Conditionally on the group effect, the probability of scoring can be considered independent since we are deleting their common factors.

Remarks

- Interpretation of the coefficients
- Uninformative priors
- Significant level

Beta coefficient

todo

Poterior Analysis Distribution of the beta coefficient

todo

Convergency of the beta variable

todo ## Player impact on goal

todo ## Model comparison via DIC

$$\begin{cases} H_0 : \sigma_{b_0} = 0 \\ H_1 : \sigma_{b_0} > 0 \end{cases}$$

Since the DIC the GLM is lower than the DIC of the GLMM , it is possible to claim that data do not show enough evidence in order to rejcet the null hypothesis.

So we can conlcude that the assumptions of the hierarchical structure does not fit with this data

Discussion

todo

References

todo