

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

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 for the players who take the shots, and this assumptions does not seem suitable, since the player's skill could influence the success of shot- conversion.

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

The data are collected by tracking players over the course of football matches 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 collect relevant informations of the shots

Data extrapolation



Figure 1: Messi's Goal against Albacete

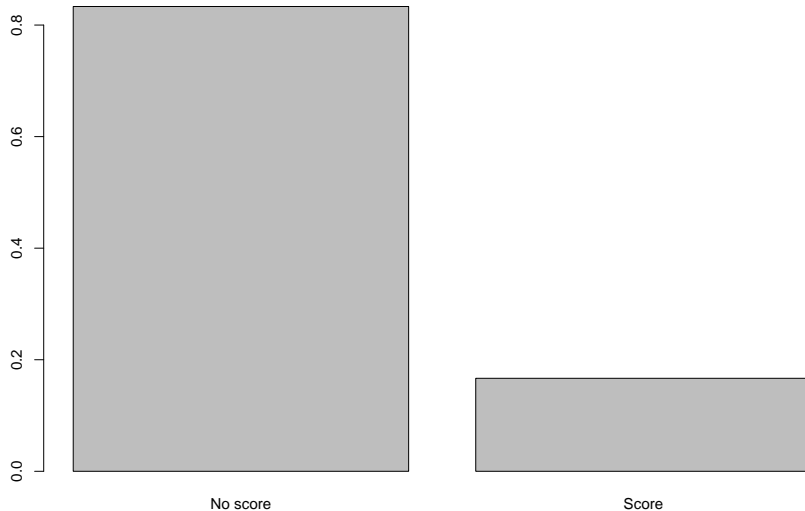
Dataset description

The dataset is composed by 13 variable and 7437 observations that represent shots of different 42 player of Barcelona F.C.

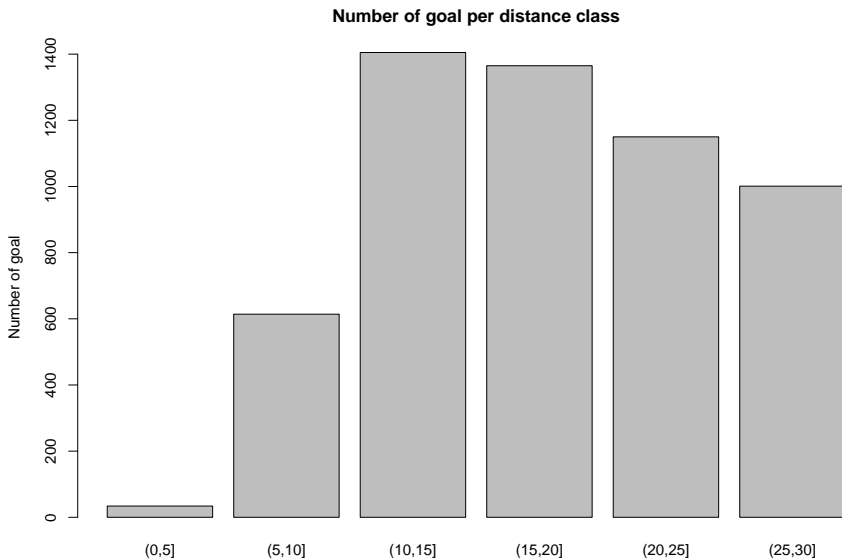
- Three continuous predictors: shot-taker's distance to goal, angle of shot and time.
- Three binary predictors: technique, first touch, inside 18 metres
- Two multi-class predictors: body part and preferred type

Explanatory Analysis

Bar plot of shots results



Explanatory Analysis (2)



Hierarchical assumptions

A reasonable assumption is that player has a unique type of shooting. Therefore, given a player, the shots are inherently correlated to each other and behave independently from the shot of another player. This introduces within-player correlations (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 structure, otherwise it may lead to biased inference.

Example of the hierarchical structure

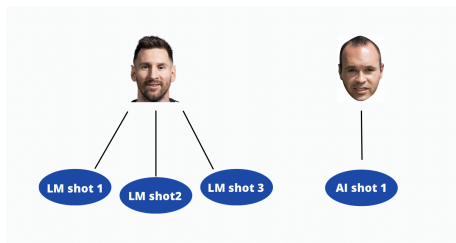


Figure 2: Example of Hierarchical structure

Model specification

For targeting the target variable, two models are chosen.

- GLM (Generalized linear model): do not feed information about the hierarchical structure
- GLMM (Generalized linear mixed model): do feed information about the hierarchical structure

Generalized linear model specifications

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

where:

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

Generalized linear mixed model specifications

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

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
- $\pi_{i,g}$ = odds ratio of score from the i – th shot and the g – th player

Dimensionality of the parameters

Let us define the dimensionality:

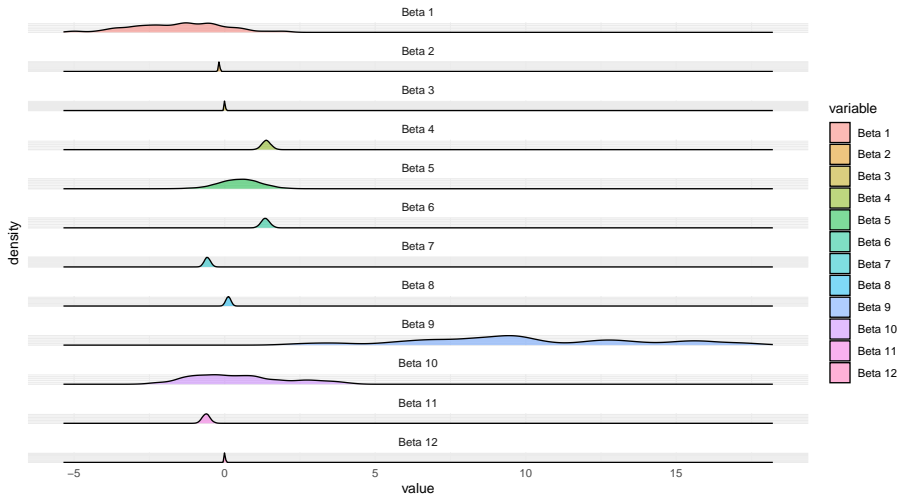
- $X_{g,i} \in (k \times 1)$ where k is the number of predictors.
- $\underline{\beta} \in (k \times 1)$ can be considered as a vector of fixed effect since it does not depend on the player.
- δ_j represent the random effect linked to the g -th player.
- The vector $\underline{\epsilon_g}$ is the vector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

Sampling model assumptions

The $\epsilon_g \perp b_g \forall g = 1 \dots 42$. The b_g are normally distributed with their own variance. The ϵ_g are normally distributed.

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

Poterior Analysis Distribution of the beta coefficient



Betas coefficient

Predictors	Effects
β_1 : Intercept	-1.5087
β_2 : shot_distance	-0.184
β_3 : shot_angle	0.0011
β_4 : bodypartLeft Foot	1.385
β_5 : bodypartOther Bodypart	0.4860
β_6 : bodypartRight Foot	1.3539
β_7 : techniqueVolley	-0.579
β_8 : first_touchTrue	0.1215
β_9 : preferred_typeLeft Foot	10.078
β_{10} : preferred_typeRight Foot	0.523
β_{11} : inside_18True	-0.615
β_{12} : time	0.0021

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 reject the null hypothesis.

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

Discussion

todo

References

todo