Modelling Expected Goals (xG) Using Bayesian Inference

52086

A report submitted to the Department of Statistics of the London School of Economics and Political Science

April 23, 2023

Abstract

In this report, we explore the notion of expected goals (xG) in association football, and how to deal with the uncertainty that arises in its calculation. Equipped with the Wyscout Soccer Match Event Dataset from Pappalardo et al. (2019*b*) and Pappalardo et al. (2019*a*), we will fit our own xG models with varying levels of complexity, in an attempt to infer the most important factors in determining xG, and reduce uncertainties within these. Our initial models are based on positional data of the shot, and later in the report, we will apply Bayesian techniques to account for the individual abilities of the shot taker.

Contents

Abstract **ii**

1 Introduction **1**
1.1 The Wyscout Dataset 1
1.2 Problem Setting 1

2 Methodology **2**
2.1 Data Collection 2
2.2 Data Augmentation 2
2.3 Baseline Methods 3
2.3.1 Logistic Regression 3
2.4 Bayesian Methods 3
2.4.1 Bayesian Single-Level Logistic Regression 4
2.4.2 Bayesian Multi-Level Logistic Regression 4

3 Results **6**
3.1 Single-Level Models 6
3.2 Multi-Level Model 6
3.3 Diagnostics 7

4 Conclusion **8**
4.1 Findings 8
4.2 Limitations 8
4.3 Further Work 8

Bibliography **9**

Chapter 1

Introduction

1.1 The Wyscout Dataset

The data set contains event data from almost 2000 matches played within 5 countries in the 2017/2018 football season. In total, there are over 3,000,000 events described. An event constitutes any time a footballing action is taken, such as a pass or a shot. For our purposes, we are only interested in the event of a shot. Within each event entry, there are various further characteristics, usually in the form of a tag. We translate the tags using a glossary provided by Wyscout. These tags tell us whether a shot has been blocked, what part of the body the action was taken with and so on.

1.2 Problem Setting

An increasingly popular metric in football is Expected Goals, or xG. The xG metric aims to take into account various spatial factors and produce a probability estimate of a given shot being a goal. In order to construct a model of this kind, there must be plentiful shot data we can fit a model to, however, there is little of this kind in the public domain. Herein lies the novelty of our approach, we plan to apply Bayesian methods to account for this lack of shot data. By incorporating prior beliefs into our model, we can account for uncertainty within the shot distances, angles, and even player abilities.

Chapter 2

Methodology

2.1 Data Collection

As mentioned earlier, we first want to extract only the shot data from the dataset. From here, we extract meaningful information about the shots from the tagging system Wyscout has implemented, we refer the reader to Wyscout (2023) for a comprehensive description of what the tags refer to. The tags of interest in our case are the body part the shot was taken with, whether the shot was blocked, and if the shot came at the end of a counterattack.

Finally, we sort the data by what players recorded the most shots, and keep the top 50 of these, discarding the others. This is to primarily make the computation time for the Bayesian models reasonable. It also acts as a balancing mechanism in the data. For example, there are some players with >100 shots and some with only 1.

2.2 Data Augmentation

The data set provides spatial information, in the form of coordinates. Despite this, there is some augmentation necessary to make the data (and therefore models) more interpretable. For each event, we consider the coordinates of the start of the shot. The coordinates themselves describe the percentage of the pitch covered by the attacking team. The x coordinate refers to the percentage of the pitch towards the opposition goal, and the y is how far right of the goal it is, towards the touch-line.

For our purposes, it is necessary to transform these coordinates, so they describe metres. We then apply simple geometry to calculate the distance from the goal, and the angle to the goal. Upon completion of this, we observed some negative angles within our data. Since the volume of these was comparatively small to the rest of the data, we cut these from the analysis.

2.3 Baseline Methods

In this section, we describe the models we consider to be our baseline. This achieves two goals: firstly, for comparison to the more complicated models later on; secondly, a more subtle point is that we can perform exploratory analysis of these models, for example by extracting the most important variables to consider.

2.3.1 Logistic Regression

For our baseline models, we implement a selection of standard logistic regression models. We hypothesised that distance and angle to goal would be the most important predictors, so we started with just distance as a predictor, and then a combination of distance and angle to goal. Finally, we implement a model based on all of our predictors. By doing this, we can evaluate the p-value for each predictor, and determine what variables have the greatest influence on the model. A lower p-value describes a more important variable. Outlined in 2.1 are the summary statistics for our final model, we see that as expected, distance and angle are among the most influential, along with the body part indicator. These results will inform our variable selection later on.

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.827369   0.313812  -5.823 5.78e-09 ***
is_CA1            0.488807   0.150439   3.249  0.00116 **
body_partleft     1.030124   0.146097   7.051 1.78e-12 ***
body_partright    0.787840   0.171598   4.591 4.41e-06 ***
dist             -0.091253   0.012458  -7.325 2.39e-13 ***
angles            0.026604   0.004658   5.712 1.12e-08 ***
preferred_foot_b1 0.209729   0.126546   1.657  0.09745 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 2.1: Summary statistics for our final logistic regression model

2.4 Bayesian Methods

We next incorporate Bayesian methods into our analysis. As mentioned earlier, in order to make a comprehensive xG model, we necessitate large quantities of data. Our data set provides enough data for our purposes but from a small time frame and few players. Bayesian inference provides us with a way of counteracting this. By incorporating prior beliefs into our model, we inject extra information into our domain, which will in theory go some way to reducing variance in our models.

A key line of investigation in this report is the effect of player ability on xG. For

this, it will be suitable to incorporate another layer to our model, one to account for the differences in groups, which in our case are players. By again using Bayesian methods, we can make up for the limited data for each player.

2.4.1 Bayesian Single-Level Logistic Regression

At first, we only consider priors at the first level. Due to our findings in 2.3.1, we only consider uncertainty in distance and angle. We feel this is justified since there is certainly some variation in each players shooting capability, which will show its effect in the distance and angle data. The variation in the body part is likely to be minimal however since most shots are taken with the player's feet.

A key step in the Bayesian process is assigning priors to our models. Since both distance and angle have large scale and range, it will be suitable to assign a *flat* prior with a large variance. As such, we choose to assign a normal prior to both distance and angle. Furthermore, as shown in Figure 2.2, both of the parameters are somewhat normally shaped, excluding outliers, so this further speaks to the usage of normal priors.

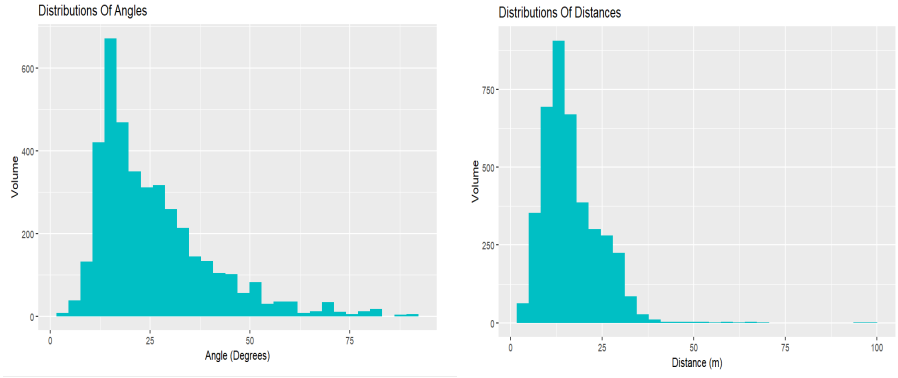

Figure 2.2: Left: Distribution of Angles. Right: Distribution of Distances.

2.4.2 Bayesian Multi-Level Logistic Regression

Finally, we implement a multi-level model, that aims to accurately describe the variation between player abilities. This is primarily motivated by the clear difference in the conversion rate of some players, highlighted in Figure 2.3. If we model our data as a single population, we lose information about differences between players in the data. However on the other end, if we modelled each group independently, we would observe far higher variance in our predictions. A Bayesian hierarchical model strikes a balance between this. In our analysis, we consider only the top 50 players in the data set, so we can assume these players are of a higher quality which drags their abilities to some common distribution. We call this the *hyper-prior*. This

can inform the model about the quality of the players but still allow for differences between them, in the prior.

Due to the computational complexity of considering many predictors, for our multi-level model, we only consider the distance predictor, along with the grouping mechanism. A further avenue of study would be incorporating more predictors into our multi-level model. We then assign the same prior to distance as in 2.4.1, and beneath that, we assign a normal hyper-prior to the scale of the distance prior, to account for the variance in player ability. We then assign a simpler exponential distribution to the variance to prevent problems with over-fit.

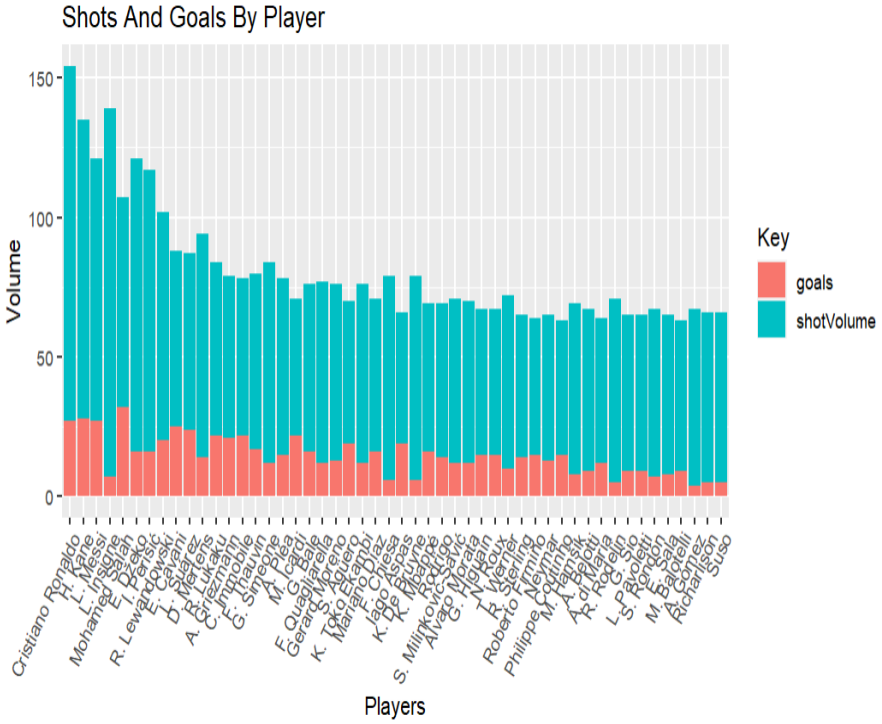

Figure 2.3: Conversion rate of top player in data set

Chapter 3

Results

3.1 Single-Level Models

Ultimately, our single-level models did not offer any significant advantage over our standard logistic regressions. There was a slight improvement in predictive accuracy in the Bayesian models, but it was not significant. Further, our confidence intervals changed marginally, implying that our injection of information was not helpful. We found that ultimately, the model that focused on distance and body part gave us a posterior that best fit the data, as described in Figure 3.1. Here we have conducted a Posterior Predictive Check (PPC). We use the model to make predictions on a new set of data, and then calculate the mean and standard deviation at every iteration of the sampling conducted. The distribution of these statistics is then plotted. This gives us an idea of how well our model fits the data. We observe that although it is not perfect, it leads us to the next section, where we consider a higher level of complexity in our model.

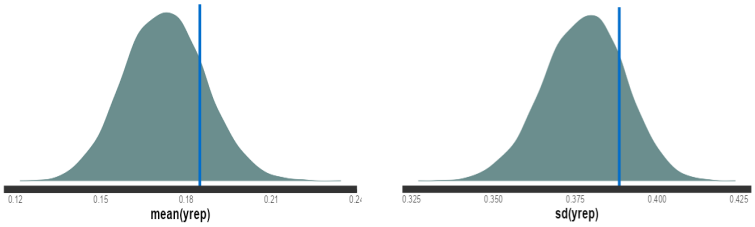

Figure 3.1: Left: Distribution of Means. Right: Distribution of Standard Deviations.

3.2 Multi-Level Model

Our final model aimed to take into account differences in player ability from shooting at a given distance. Although there was again no predictive power increase, we can be sure from Figure 3.2 that the model is indeed taking into account player ability.

The plot shows that for a given shot, Mohamed Salah, who had a high conversion rate in Figure 2.3, has a higher xG on average than the population. We also see in Figure 3.3 that each player's coefficient has a varying mean, indicating the outcome of the shot will be impacted differently by each player.

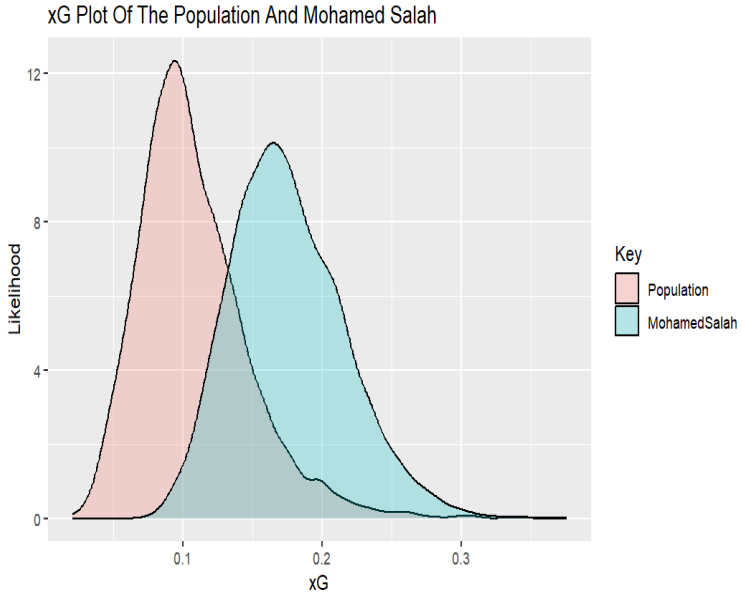

Figure 3.2: xG for given shot for salah, and the population

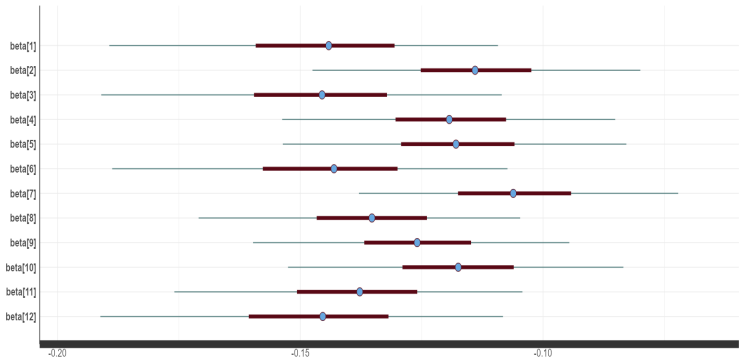

Figure 3.3: Coefficient variation between players, each *beta* represents a player.

3.3 Diagnostics

For a comprehensive review of the diagnostic elements of our models, we refer the reader to the associated R notebook.

Chapter 4

Conclusion

4.1 Findings

Our results show that Bayesian modelling goes some way to addressing our empirical problem, that is, formulating an xG model on limited data. As a purely predictive model, we would still recommend using a simple logistic regression or a method of that family. However, we have shown applying a multi-level model allows for differences in player abilities, even on the limited data we have available. Furthermore, while our Bayesian models did not offer much improvement, they still maintained a high level of predictive power, and this coupled with its player-level inferences, means that the use of Bayes is certainly justified.

4.2 Limitations

There are a few obvious limitations to our analysis. First, we have not achieved significantly higher predictive power in determining the outcome of a shot. This could be remedied by more data, and perhaps constructing a more complicated multi-level model. This leads to our second limit: our model is very simplified when considering the inherent unpredictability of football. We focus on a few, easily observable statistics, but the reality is many key determinants in football aren't observable. We could still further complicate our models, but due to the nature of the game, there must be some tolerated level of inaccuracy.

4.3 Further Work

As mentioned earlier, a potential line of study would be factoring in other players' positions on the field. Of course, this will drastically increase the complexity and computational intensity of our models, but it is a key area of uncertainty that could be considered.

Bibliography

Pappalardo et al. (2019*a*). Playerank: Data-driven performance evaluation and player ranking in soccer via a machine learning approach. *ACM Transactions on Intelligent Systems and Technologies (TIST) 10* 5.

Pappalardo et al. (2019*b*). A public data set of spatio-temporal match events in soccer competitions. *Nature Scientific Data* 6–236.

Wyscout (2023). Wyscout data glossary. *https://dataglossary.wyscout.com/* .