Bayesian Hierarchical Approach for Position and Player Adjusted Expected Goals in Football (Soccer)

Alexander Scholtes

11/09/2023

School of Mathematics,
Cardiff University

A dissertation submitted in partial fulfilment of the requirements for **MSc Data Science and Analytics** by taught programme, supervised by Oktay Karakuş.

Executive Summary

The aim of the project was to use Bayesian methods to investigate if there were position or player effects on the probability of a shot resulting in a goal in football (soccer). The rationale is that if two players were taking exactly the same shot with the same conditions and the only difference being one player is a striker and the other is a defender then one could assume that the striker has a greater chance of scoring than the defender on average. Similarly, if one player was Lionel Messi, one of the greatest footballers of all time, and the other was a player from the English 5th tier, one would also assume Lionel Messi has a greater chance of scoring most chances. However, almost all modern xG models do not adjust xG values for player position or player's themselves and would assign both chances the same probability.

To address this, Bayesian methods were used because they offer some crucial benefits when calculating conditional probabilities as is being suggested with group adjustments. Firstly, they provide a measure of uncertainty when giving results by outputting a distribution as opposed to a single point estimate, which could be very important when making footballing decisions. Moreover, Bayesian hierarchical models are very good at dealing with data groups with very few observations. This is particularly useful for player-adjusted xG models because some players will have only a few shots in the data, but the hierarchical model will still give reasonable estimates for xG by using information from the other groups to inform the new group's values.

The data that was used in this project was from StatsBomb's publicly available data, selecting only men's competitions in case of any inherent differences in modelling between men's and women's football, and because more men's data was available, and filtering out set-pieces because they do not represent typical chances and this paper is not looking to capture set-piece xG.

A few different models were fitted to achieve the goals of the paper. Firstly, a non-Bayesian xG model was built to provide comparable results to StatsBomb's xG model to confirm that the predictors being used were able to predict xG well compared to an industry leading model. Then, three Bayesian models were fitted: two which included positional effects, and one which included player effects. The first positional effects model used only shooter distance to goal

and shot angle as predictors, while the other two models used all of the predictors that were used in the non-Bayesian model.

The results of the first Bayesian model, with only distance to goal and shot angle, suggested there were positional effects on xG, with strikers and attacking midfielders having positive xG adjustments and defenders having significantly negative adjustments. However, the second model, which added a number of additional predictors, resulted in these effects almost disappearing completely, suggesting that the additional predictors were more important in determining goal probability and the position of the player, once these additional factors were controlled for, was less important. Nevertheless, the final model, which still had all the additional predictors but with player effects, did find evidence that some players were more likely to score than other on average.

The choice of prior distributions can be very important in Bayesian modelling, so the choices made for the aforementioned models were analysed against some other methods and showed that the choices were adequate to yield good results.

Some recommendations for future work include applying Bayesian hierarchical modelling to other concept and metrics in football. For example, applying it to injury risk assessment whereby some players are more injury prone than others, so could have their own personal injury risk different to a general model. Additionally, performing the same analysis on a different dataset, for example a different league, would be useful to see if the findings hold. The player-adjustment model in this paper also only looks at the player effects of six players to reduce sampling times. Perhaps the model can be optimised by choosing narrower prior distributions such that sampling takes less time, allowing for a larger more complex model without infeasibly long sampling times.

Acknowledgements

My supervisor Oktay, for being a great teacher and great guy. My amazing partner Lucy, for her emotional support and motivating me when I was struggling. My parents, for their support, and especially my dad for his more technical and theoretical suggestions.

Contents

	Page Number
List of Figures	5
List of Tables	7
Report	
- Abstract	8
- Introduction	8
- Literature Review	9
- Data	11
- Non-Bayesian modelling	14
- Bayesian Modelling	19
- Discussion and Future Work	39
- Conclusion	41
References	43

List of Figures

Figure	Page Number
Figure 1: Relationship between distance to goal (binned in 10s) and the proportion of goals from shots.	16
Figure 2: Relationship between shot angle (binned in 20s) and the proportion of goals from shots.	17
Figure 3: Distributions of xG predictions for each of the non-Bayesian models fitted, with the distribution of StatsBomb's xG model predictions for comparison.	18
Figure 4: Comparison of StatsBomb unadjusted xG sum with the same values adjusted by position, compared against true total of goals scored.	20
Figure 5: Distributions of xG adjustments for Model 1, where adjustment is hierarchical model prediction minus baseline model prediction. Grouped by general position.	25
Figure 6: Point estimates for xG adjustments for Model 1 plotted against distance to goal, grouped by general position of shooter.	26
Figure 7: Point estimates for xG adjustments for Model 1 plotted against shot angle, grouped by general position of shooter.	27
Figure 8: Heatmap of shot locations for each general position, normalized for each group.	28
Figure 9: Distributions of xG adjustments for Model 2, where adjustment is hierarchical model prediction minus baseline model prediction. Grouped by general position.	30
Figure 10: Comparison of point estimates for xG adjustments against distance to goal and shot angle between Model 1 and Model 2, grouped by general position. Adjustments are hierarchical model prediction minus baseline model prediction.	31
Figure 11: Distributions of xG adjustments for Model 3, where adjustment is hierarchical model prediction minus baseline model prediction. Grouped by player.	33
Figure 12: Selected player shot locations and goals.	34
Figure 13: Comparison of Model 3 baseline predictions, Model 3 hierarchical predictions, and actual goals scored for selected players.	35

Figure 14: Distributions of xG predictions for each of the extended single-level Bayesian model, with different choices of prior distributions.	38
Figure 15: Point estimates of the first 100 predictions for Model 3 with different sets of priors. The priors are listed in Table 3 and Table 6.	39

List of Tables

Table	Page Number
Table 1: Summary statistics for most relevant features for xG in the dataset.	12
Table 2: Results of non-Bayesian models where basic model has only distance to goal, interaction model adds shot angle and an interaction between that and distance, and the extended model adds several additional features. Compared to the StatsBomb xG model.	19
Table 3: Listing which predictors are used in each of the Bayesian models, and what prior distribution is given to the coefficient of the given predictor.	22
Table 4: Mean xG adjustment for each general position from Model 1 versus theoretical adjustment of baseline Model 1 predictions using Bayes' Theorem.	29
Table 5: Selected players for Model 3 and their goal scoring statistics.	32
Table 6: Choices of prior distributions for analysing the impact of prior choice on results. Note, the current priors are in Table 3.	36

Report

Abstract

This paper uses Bayesian methods to investigate the presence of position or player effects when determining the probability of a shot being a goal, a metric called expected goals (xG). Using roughly 60,000 shots from StatsBomb's publicly available data, a non-Bayesian, or frequentist, approach is taken to build a baseline model with comparable results to the model produced by StatsBomb, an industry leader in terms of football analytics and data collection. Thereafter, Bayesian hierarchical logistic regressions are built to determine if there are positional or player level effects on xG using roughly 10,000 shots from the Premier League. The results showed that there were positional effects in a basic model with only distance to goal and shot angle as predictors, with strikers and attacking midfielders being more likely to score and defenders being less likely to score on average, but these effects disappeared once more predictors were added. However, even with the additional predictors there were player level effects as some players had large positive or negative xG adjustments, suggesting that they are more or less likely respectively to score a given chance than others, and this effect can be measured using a Bayesian hierarchical model. This paper also analyses the impact of prior distribution choices on results and finds that the priors used in the aforementioned models do yield good model results, but that they could be refined to improve sampling speeds to feasibly build a more complex, larger model.

Introduction

One of the most common advanced football analytics metrics is the idea of expected goals (xG). It estimates the probability of a given shot resulting in a goal based on a number of features about the shot, such as distance of the shooter to the goal or the body part used by the shooter. However, none of the mainstream xG models take into account any player-specific features when estimating these values. To illustrate this, imagine that you have two players taking exactly the same shot from the same position, with defenders in the same place and everything else being the same but one player is Lionel Messi and the other is a random player from the National League (English 5th tier). Obviously, players who play in the National League are good players, but it is not unreasonable to assume that Lionel Messi would be more likely to score. However, xG metrics would assign the same value for both of these chances.

The objective of this paper is to investigate if there are position or player effects on xG, meaning that certain positions or players have higher or lower goal probabilities for a given chance than others. This will be achieved using a Bayesian hierarchical model, where the hierarchies will be the position of the player or the player themselves. The results of this method will be compared to a more traditional frequentist xG model to initially compare baseline results without any group effects. Then the hierarchical models will be compared to the non-hierarchical Bayesian models, to assess the impact of having hierarchies in the data on the results. If the rationale described above of the two players taking the same shot is true, then it is expected that the xG predictions of the hierarchical models will differ significantly from the non-hierarchical models, supporting the idea that there is a position and / or player effect on the xG of a given shot.

This paper will begin with a review of the relevant literature around this topic, looking at the development of football analytics and xG, as well as any attempts to use Bayesian modelling in football analytics. Thereafter, the data will be introduced and described with any changes made. Next, some non-Bayesian, or frequentist, modelling will be done to create a typical xG model for comparison. Then, the Bayesian modelling methodology and results will be given, before discussing the results and recommending future work, and finally a conclusion.

Literature Review

The use of data in football is often not fully embraced, with many decision makers arguing the sport is too complex for data to be used effectively to improve results and performance (Smith, 2022). However, with its successful use in other sports there was sufficient interest for some clubs, companies, and individuals to pursue using data to derive conclusions and make suggestions in football. With growing demand for data, companies that specialise in sports data collection have grown too, along with their ability to track data. The result is that there is now an enormous amount of football data to use for a number of purposes, such as player/club performance, scouting, and player fitness and injury risk, to name a few (Tippett, 2019).

At the heart of the idea of using data in football was the potential to gain a competitive advantage. As a result, clubs that use data tend to be secretive about their operations and procedures (Tippett, 2019). Despite this, there is plenty of publicly available literature and sources showing how data is used in football. Moreover, sports broadcasters have long used data when giving an overview of a match, such as possession statistics, but these have only recently moved away from simple counts and percentages. The Bundesliga, for example, provides a goal probability value after each goal is scored, giving the chance of that given opportunity resulting in a goal (Aberle, et al., 2020).

Goal probability, also commonly called expected goals (xG), has been a central topic in the development of more advanced statistics using football data (Smith, 2022). Crucially, it moves away from the idea of things that did happen and focuses on things that could have happened. With football being such a complex and chaotic sport, outcomes often do not reflect expectations as matches are often decided by fine margins or decisions out of the players' and coaches' control. Nevertheless, by using expectations, it gives decision makers an idea of the underlying performance of their team and allows them to see if their team is over or underperforming according to expectations (Brechot & Flepp, 2020).

There have been many versions of xG models created since the idea was founded, using a variety of machine learning techniques and data sources. Herold et al. (2019) provide a summary of many applications of machine learning in football, including xG models. The most common methods of estimating xG in their paper are logistic regressions, decisions trees, ensemble methods (e.g. random forest), and neural networks. Lucey et al. (2015) use player and ball tracking data from the 10 seconds leading up to a shot to estimate goal probabilities across an entire season and found that "defender proximity, interaction of surrounding players, speed of play, coupled with shot location play an impact on determining the likelihood of a team scoring a goal". Pardo (2020) uses qualitative data from the popular video game FIFA to account for player-effects on xG using a logistic regression and an XGBoost model. They found that an adjusted model can better predict goals over a season for individual players and teams than an overall xG model. Fairchild et al. (2022) built an xG model again using a logistic regression and used it to estimate MLS teams' offensive efficiency in scoring. They also discuss evaluation metrics for expected goals models and suggest the use of the Brier score to compare predicted probability to ground-truth binary outcomes. Cavus and Biecek (2022) apply a variety of ensemble and boosting methods to calculate xG values and find that a random forest model performs best, even compared to models from other papers using other techniques and data.

The closest study to this paper to date is that of Hewitt and Karakuş (2023), who investigate position and player adjusted xG models. They find evidence of positional adjustments with forwards having a positive adjustment, midfielders having a slightly negative adjustment, and defenders having a large negative adjustment. Moreover, they also find evidence of player effects on xG by fitting their model with only data from Lionel Messi and find a large positive adjustment in this case.

One of the features a lot of these models have in common is their frequentist approach, as opposed to using Bayesian methods. Spearman (2018) uses a Bayesian approach to estimate maximum a posteriori effects of parameters in a model for predicting future scoring of teams in games. Joseph et al. (2006) used Bayesian networks and naïve Bayes learners to predict the results of matches played by Tottenham Hotspur and compared the results to K-nearest neighbour and decision tree models. They reiterate one of the benefits of Bayesian modelling which is comparably accurate predictions in the absence of a large amount of data. Zambom-Ferraresi et al. (2018) use Bayesian methods to analyse team performance in Europe's top leagues to determine which features tend to be most significant in predicting team performance. They find that the most important features include number of assists, number of shots conceded, saves made by goalkeeper, passing accuracy, and number of shots on target.

One area of Bayesian modelling which is also often not considered in football analytics is using multi-level, or hierarchical, models. Tureen and Olthof (2022) construct a multi-level model for player-adjusted expected goals, but do not use a Bayesian approach to do so. Still, they use their model to calculate estimated player impact values on xG. On the other hand, Baio and Blangiardo (2010) construct a Bayesian hierarchical model but use it to predict match results as opposed to xG directly and group their data by team as opposed to by player. Still the use of Bayesian hierarchical modelling in football is a relatively unexplored area.

Data

Data Extraction

The data used for this project is all freely available event data from StatsBomb, obtained using their python package *StatsBombPy*. From their database, only men's competitions were used because it could be that there is a difference in given goal probabilities in men's and women's football, and we have more data from men's competitions. Then, all open-play shots were extracted with all relevant information for each shot. Set-pieces were excluded because again goal probabilities could vary for set-pieces, and we are not interested in modelling this effect.

The resulting data has around 60,000 shots from a variety of competitions and years, with 42 columns of information for each shot. Table 1 gives some summary statistics for the most relevant variables in the data.

Table 1: Summary statistics for most relevant features for xG in the dataset.

Variable	Description	Summary (2.d.p)
distance_to_goal	Distance between the shooter and the middle of the goal line, normalized to StatsBomb pitch size of 120x80 due to varying pitch sizes.	N: 63,309 Mean: 18.96 SD: 8.58 Min: 0.63 Max: 88.83
shot_angle	By creating a triangle between the shooter and the two goalposts, the shot angle is the angle that is by the shooter.	N: 63,309 Mean: 25.39 SD: 15.60 Min: 0.66 Max: 168.61
gk_distance_to_goal	Same as *distance_to_goal* but for the goalkeeper instead of the shooter.	N: 63,309 Mean: 3.56 SD: 2.61 Min: 0.00 Max: 118.00
players_in_shot_triangle	The number of players in the shot triangle, created by the shooter and the two goalposts.	N(0): 1,786 N(1): 30,262 N(2): 19,481 N(3): 6,802 N(4): 2,918 N(5): 1,217 N(6): 513 N(7): 211 N(8): 77 N(9): 29 N(10): 11 N(11): 2
opponents_in_radius	The number of opposition players in a 1m radius of the shooter.	N(0): 55,536 N(1): 7,030 N(2): 662 N(3): 71 N(4): 10
shot_body_part	The body part used by the shooter to hit the ball.	N(Preferred Foot): 30,738 N(Other Foot): 11,733 N(Head): 10,647 N(Other): 191
shot_first_time	Whether the shot was a first-time shot, meaning the shooter took no additional touches of the ball before shooting.	N(True): 20,946 N(False): 42,363
gk_in_shot_triangle	Whether the goalkeeper was in the shot triangle created by the shooter and the two goalposts when the shot was taken.	N(True): 60,570 N(False): 2,739

shot_one_on_one	Whether the shooter was one-on-one with the goalkeeper when shooting.	N(True): 3,546 N(False): 59,763
shot_open_goal	Whether the shooter was shooting at an open goal.	N(True): 736 N(False): 62,573
shot_technique	The technique the shooter used.	N(Normal): 47,854 N(Half Volley): 9,371 N(Volley): 4,483 N(Lob): 688 N(Overhead Kick): 385 N(Diving Header): 284 N(Backheel): 244
under_pressure	Whether the shooter was under pressure when shooting.	N(True): 16,149 N(False): 47,160
goal	Whether the shot resulted in a goal.	N(True): 6,559 N(False): 56,750
shot_statsbomb_xg	StatsBomb's own estimated xG value for each shot.	N: 63,309 Mean: 0.10 SD: 0.13 Min: 0.00 Max: 1.00
general_position	The general position of the shooter (striker, attacking midfielder, other midfielder, or defender).	N(ST): 17,073 N(AM): 20,065 N(M): 15,858 N(D): 10,313
player	The name of the shooter.	N(Messi): 1,907 N(Luis Suárez): 596 N(Iniesta): 374 N(Neymar): 352 N(Thierry Henry): 325 … N(Vaclav Migas): 1 N(Roberto Tricella): 1

Feature Transformation

As well as the information given in the columns already in the data, there are several features which were not included by StatsBomb which could be useful for predicting goal probability. Many sources cite distance to goal and shot angle to be two of the most important predictors of goal probability.

The data has the location of the shot, and the StatsBomb data specification (StatsBomb, 2019) provides information about the coordinates of the goal posts. Distance to goal is therefore calculated as the Euclidean distance from the shot to the centre of the goal.

For shot angle, the cosine rule is used to calculate the angle from the shooter to the two goalposts. To calculate this reliably, any shots that are taken from the same x-coordinate as the goal-line are excluded, since this would create a straight line instead of a triangle with no shot angle.

Next, the freeze-frame feature in the data is utilised, which provides information about all other players than the shooter when the shot is taken, including crucially their location on the pitch. From this information, several features are added: goalkeeper distance to goal, whether the goalkeeper is present in the shot triangle formed by the shot and two goalposts, the number of players present in the shot triangle, and the number of opponents within a 1m radius of the shooter. Opponents are used for the 1m radius as opposed to all players because the only time a non-opponent in the radius of a shooter would impact goal probability is when they are in the shot triangle, which is already being accounted for. Otherwise, only opponents will try to put pressure or tackle the shooter outside of the shot triangle.

Then, each players mode general position is added. These positions are grouped into: strikers, attacking midfielders, non-attacking midfielders, and defenders, and we expect goal probability to fall on average for each group respectively. Wingers are included as attacking midfielders, while wide-midfielders and central midfielders are non-attacking midfielders. Wing backs are classed as defenders, and so are the few attempts by goalkeepers. Otherwise, all positions to classes are self-explanatory.

Finally, we modify body-part used by changing “left foot” and “right foot” to “preferred foot” and “other foot” according to the player’s apparent preference. To assign these, we go back through the open data and instead look at the passes for each player, we then assign whichever foot the player made the most passes with as their preferred foot because they often have more time to choose which foot to pass with and will then tend to go safely with their preferred foot, which is less feasible when taking a shot as players tend to be under more pressure and have less time.

Non-Bayesian Modelling

Methodology

First, a frequentist approach was used to try to build an expected goals model which would perform adequately against true outcomes and against the StatsBomb xG values included in the data. While the StatsBomb xG model predictions do not represent a true outcome, it is an industry-leading prediction which still gives indication if the predictions from this model are good.

Several logistic regression models will be fitted and compared using the mean-squared error, mean-absolute error, R-squared coefficient, and the Brier score. The first three will compare model predictions to the StatsBomb xG values, while the Brier score will compare predicted probabilities to the actual outcome of a shot (goal or not), meaning it can also be calculated for the StatsBomb predictions for further model comparison.

Any numeric predictors will be scaled to be between 0 and 1 based on their minimum and maximum values, while categorical features will be encoded using one-hot encoding. Moreover, a random training-test split of 70:30 is used.

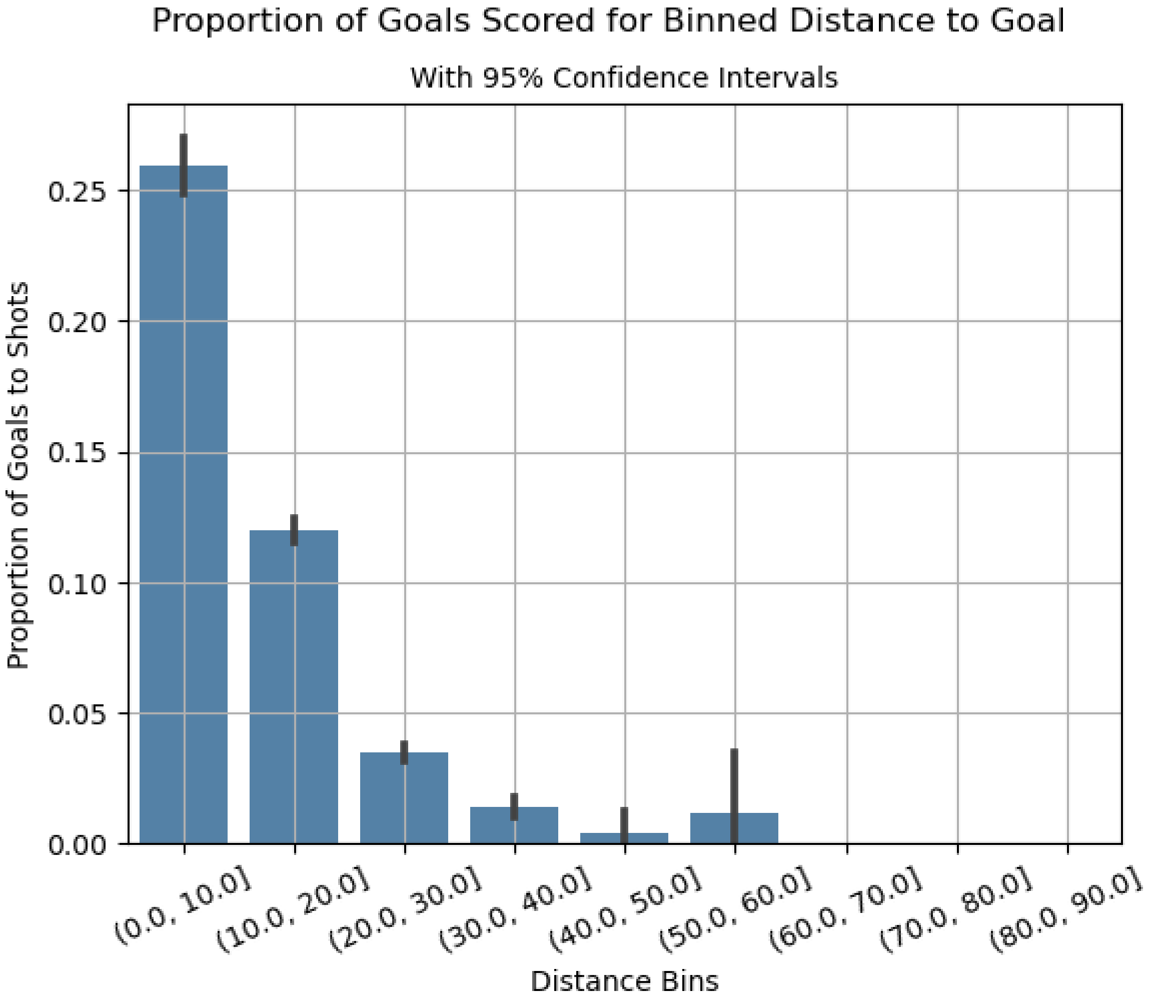
Results

The first model will be a basic model with only the distance between the shooter and the goal as a predictor, as this is widely considered one of the most important factors in determining goal probability. The model is formulated as:

$$logit(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 * distance\ to\ goal_i$$

Logically, it is reasonable to assume that as distance to goal increases, the probability of scoring would decrease on average, meaning β_1 would be negative. By grouping distance into bins of size 10 and then calculating the proportion of shots resulting in goals for each bin, see Figure 1 below, the predicted relationship between distance and goal probability is supported.

Figure 1: Relationship between distance to goal (binned in 10s) and the proportion of goals from shots.

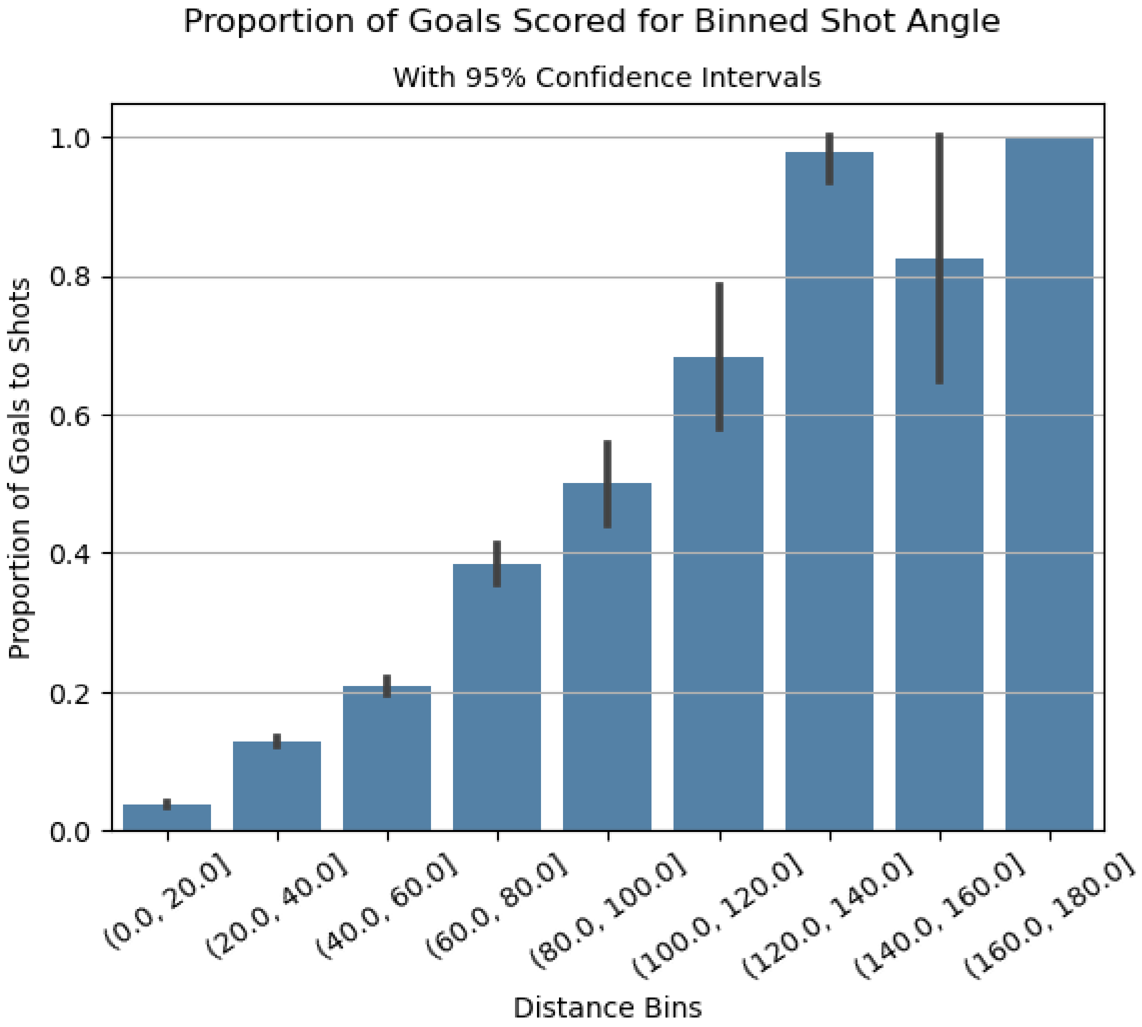


Another feature which is commonly considered to be important for determining goal probability is the angle of the shot, that is the angle at the shooter created by the triangle from the shooter to the two goalposts. Since distance and shot angle are inherently linked, because as distance increases angle decreases regardless of position on the pitch, an interaction term will be needed to account for the combined effect of both features. The final model then, is formulated as:

$$logit(p_i) = \beta_0 + \beta_1 * distance\ to\ goal_i + \beta_2 * shot\ angle_i + \beta_3 * distance * angle_i$$

β_2 will likely be positive to capture the fact that as shot angle decreases the position of a shot will either be further away or from a wider position, both likely resulting in a lower goal probability. Figure 2 (below) shows that this relationship generally holds in the data. β_3 will capture the effect of changing shot angle for a given distance to goal. For a given distance, as shot angle falls this means the shot is going further to either side of the goal while the distance remains the same. Generally, this would result in a decrease in goal probability on average as shots either side of the goal will be harder to score than more central shots for a given distance on average.

Figure 2: Relationship between shot angle (binned in 20s) and the proportion of goals from shots.



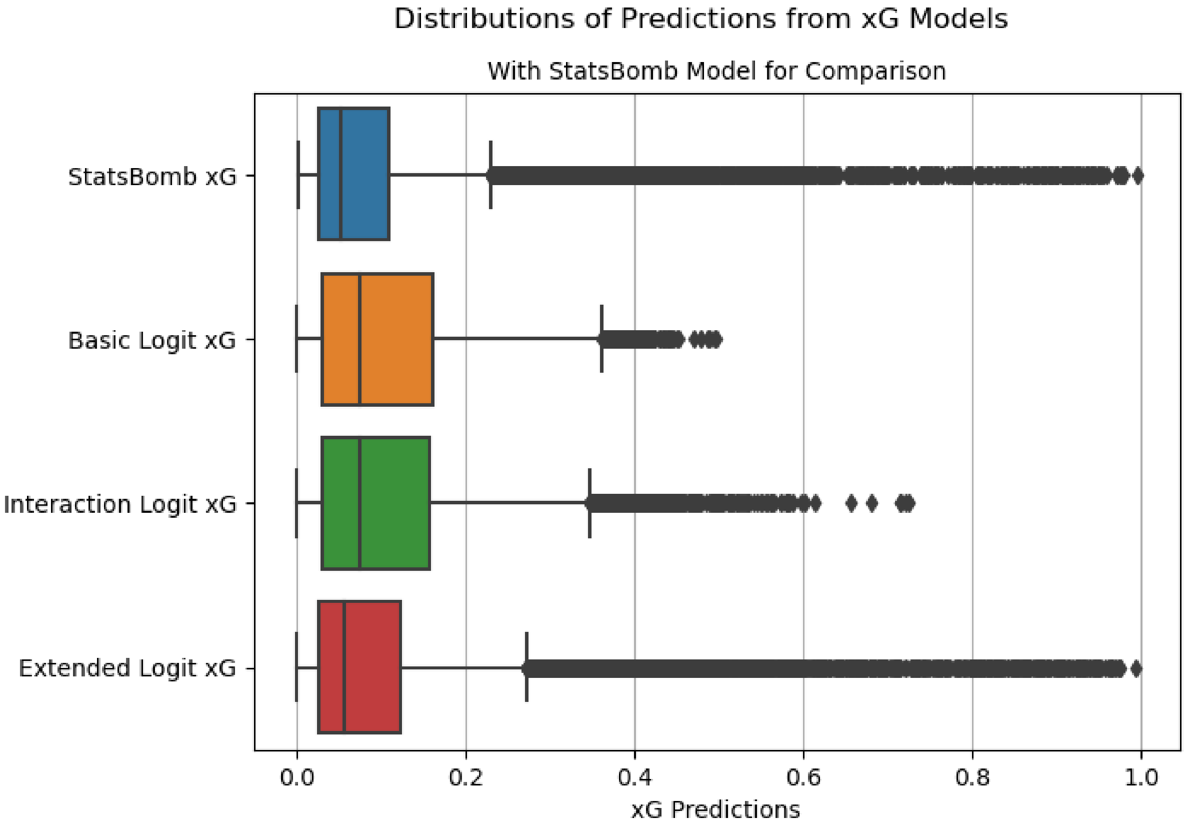
Finally, an extended model is fitted which includes a number of additional features. This model is formulated as:

$$logit(p_i) = \beta_0 + \beta_1 * distance\ to\ goal_i + \beta_2 * shot\ angle_i + \beta_3 * distance * angle_i + \beta_4 * gk\ distance\ to\ goal_i + \beta_5 * players\ in\ shot\ triangle_i + \beta_6 * body\ part_i + \beta_7 * first\ time\ shot_i + \beta_8 * gk\ in\ shot\ triangle_i + \beta_9 * one\ on\ one\ shot_i + \beta_{10} * open\ goal_i + \beta_{11} * technique_i + \beta_{12} * under\ pressure_i$$

Recall that the categorical features are one-hot encoded and therefore these variables and their associated betas are multidimensional vectors.

The distributions of the predicted xG values for each model is shown below in Figure 3, along with the StatsBomb xG values for comparison. As expected, the model with shot angle, distance to goal, and the interaction between the two performs better than the basic model containing only distance to goal by predicting more extreme values above 0.5. Similarly, the extended model performs better than the interaction model by also predicting much more

Figure 3: Distributions of xG predictions for each of the non-Bayesian models fitted, with the distribution of StatsBomb's xG model predictions for comparison.



extreme values and by decreasing the interquartile and whisker ranges. The combination of this is that the extended model has a distribution very close to that of the StatsBomb model.

In terms of the evaluation metrics, Table 2 shows how each of the models performed. The extended model is by far the best of the three models, as is to be expected with having the most information to predict goal probability. It also has a Brier score almost identical to the StatsBomb model, meaning it is performing well compared to an industry-leading xG model with respect to this metric.

Table 2: Results of non-Bayesian models where basic model has only distance to goal, interaction model adds shot angle and an interaction between that and distance, and the extended model adds several additional features. Compared to the StatsBomb xG model.

	Basic Model	Interaction Model	Extended Model	StatsBomb Model
MSE	0.010	0.009	0.003	-
MAE	0.058	0.058	0.029	-
R^2 Score	0.413	0.428	0.826	-
Brier Score	0.086	0.086	0.076	0.075

Bayesian Modelling

Justifying Using Bayesian Methods

To illustrate how using Bayesian methods can be applied to xG models, we start with a simple application of Bayes' Theorem:

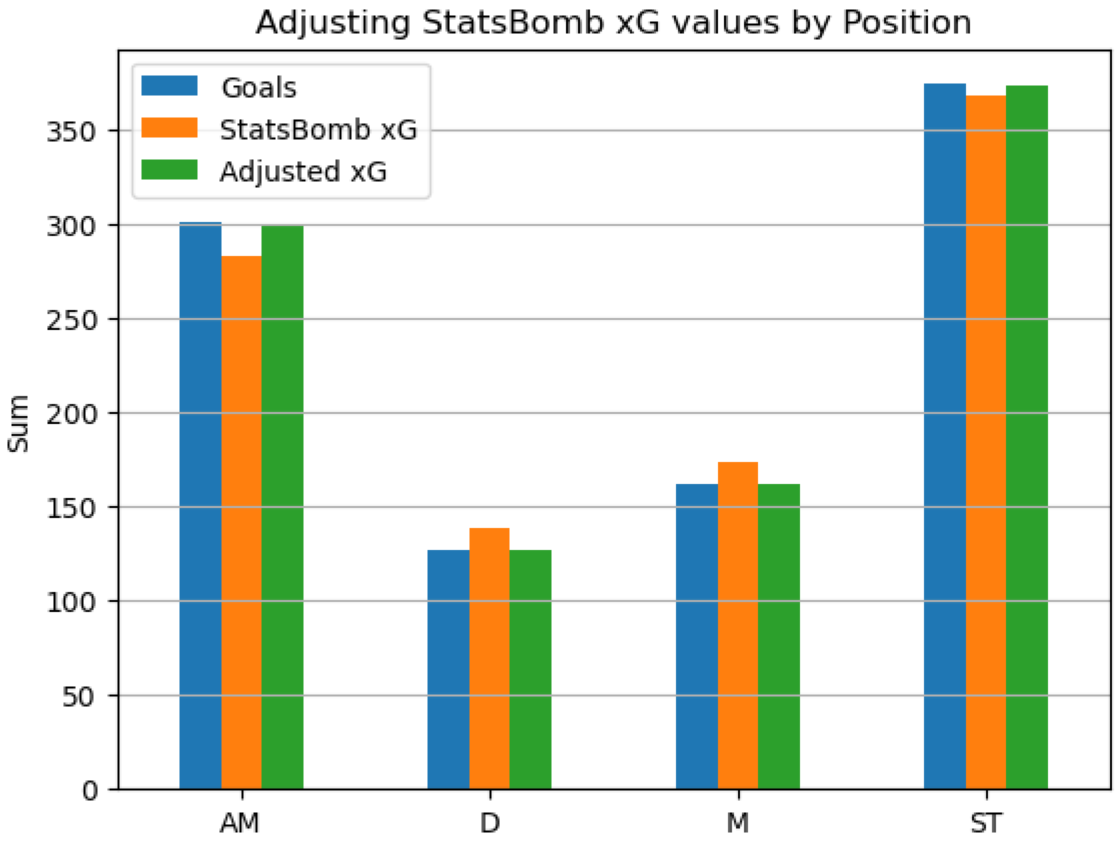
$$P(A|B) = \frac{P(B|A)*P(A)}{P(B)}$$

Imagine two players are attempting to score from exactly the same shot, with the same defensive players' location and shot technique, etc. and the only difference being the player taking the shot. If one player is a striker and the other a central defender, it is not unreasonable to assume the striker has a greater probability of scoring on average than the central defender. We can apply this logic to Bayes' Theorem and adjust xG values based on the position of the shooter:

$$P(goal|position) = \frac{P(position|goal) * P(goal)}{P(position)}$$

To apply this numerically to each position category ST, AM, M, and D, the proportion of goals to shots is used to estimate $P(goal)$, the proportion of each position category in the data is used to estimate $P(position)$, and the proportion of each position category in the data including only goals is used to estimate $P(position|goal)$. This method is applied to each shot in the premier league in the data and summed for each position. The sum of the adjusted xG is then compared to the sum of the StatsBomb xG and actual goals for each position. Figure 4 below shows that the adjusted xG provides more accurate measures of total xG than the unadjusted xG.

Figure 4: Comparison of StatsBomb unadjusted xG sum with the same values adjusted by position, compared against true total of goals scored.



This example is a very basic use of Bayesian techniques, but there are also more advances techniques which can be applied for further benefits. One of the main benefits of using Bayesian modelling over frequentist modelling is that the results of Bayesian models, such as a Bayesian logistic regression as in the next section, is not simply a point estimate but a probability distribution for different values of estimates. This allows the user to see how certain the estimates are, which is not possible using the frequentist approach above.

Methodology

All the models we have fitted up until now are pooled models because they all assume there is no grouping in the data. We have one overall model with a set of coefficients, and we apply this model to each shot to obtain xG estimates for each shot in our test data.

At the other end of the spectrum are fully unpooled models, which would split the data by the potential grouping variable, such as general position, and fit four different models on each subset of the data. The test data would then be split by general position and each model applied to each subset of the test data to obtain grouped xG estimates. This would work in this case since we have lots of training data for each general position group, but in other cases there may not be such an abundance of data for each group. For example, if the data was split by the player taking the shot, there may be players with only a few shots, yielding unreliable estimates for that player's effect on xG.

The solution lies between the two extremes and is known as hierarchical or multilevel modelling. The idea of such a model is to add another variable to the model to capture the potential effects of the groups in the data, but also having the parameters of the coefficient for that variable in the model be determined by a probability distribution, such that the parameters for the variable can be different for the different groups in the data. The result is a single model which will provide multiple posterior distributions for each group in the data.

This paper will consider position and player as the grouping variables in these models, with the former using the general position variable and the latter using the player name to group the data. Moreover, for the position-adjusted model there will be a basic model with only distance to goal, shot angle, and the interaction between them as predictors as well as an extended model with the same predictors as used in the non-Bayesian model. For the player-adjusted model, only the extended model will be fitted. For each of these models, there will be a baseline (single-level) model which does not account for the group-level effects as well as the hierarchical model. This will allow for direct comparison to see the extent of the group-level adjustments in each model.

Definitions: Y_i = binary outcome of shot i (1 = goal, 0 = no goal)

p_i = probability of shot i resulting in a goal

$X_{n,i}$ = the value of predictor n for shot i

Likelihood: $Y_i \sim Bernoulli(p_i)$

Baseline Model: $logit(p_i) = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_n X_{n,i}$

Hierarchical Model: $logit(p_i) = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_n X_{n,i} + \beta_{n+1} X_{n+1,i}$

Where $n + 1$ is the index of the grouping variable in the data.

Table 3: Listing which predictors are used in each of the Bayesian models, and what prior distribution is given to the coefficient of the given predictor.

	Model			
Predictor	1	2	3	Prior
Intercept	✓	✓	✓	$Normal(\mu = 0, \sigma = 5)$
Distance to goal	✓	✓	✓	$SkewNormal(\mu = -1, \sigma = 5, \alpha = -1)$
Shot angle	✓	✓	✓	$SkewNormal(\mu = 1, \sigma = 5, \alpha = 1)$
Distance $*$ angle	✓	✓	✓	$Normal(\mu = 0, \sigma = 5)$
Goalkeeper distance to goal		✓	✓	$Normal(\mu = 0, \sigma = 5)$
Players in shot triangle		✓	✓	$SkewNormal\left(\begin{matrix}\mu = 0, \sigma = 5, \\ \alpha = \{5, \dots, -5\}\end{matrix}\right)$ Where α is determined by the value of this feature (0 players = 5, 1 player = 4, etc.)
Opponents in 1m radius		✓	✓	$SkewNormal\left(\begin{matrix}\mu = 0, \sigma = 5, \\ \alpha = \{1, \dots, -2\}\end{matrix}\right)$ Where α is again determined by the value of this feature (0 players = 1, 1 player = 0, etc.)
Body part		✓	✓	$Normal(\mu = 0, \sigma = 5)$

First-time shot		✓	✓	$Normal(\mu = 0, \sigma = 5)$
Goalkeeper in shot triangle		✓	✓	$SkewNormal(\mu = 0, \sigma = 5, \alpha = -2)$
One-on-one shot		✓	✓	$SkewNormal(\mu = 0, \sigma = 5, \alpha = 2)$
Open goal shot		✓	✓	$SkewNormal(\mu = 0, \sigma = 5, \alpha = 4)$
Shot technique		✓	✓	$Normal(\mu = 0, \sigma = 5)$
Shooter under pressure		✓	✓	$SkewNormal(\mu = 0, \sigma = 5, \alpha = -2)$
General position adjustment	✓	✓		$SkewNormal\begin{pmatrix} \mu = 0, \sigma = \sigma, \\ \alpha = \{2,1,0,-2\} \end{pmatrix}$ $\sigma \sim HalfNormal(\sigma = 5)$ Where α is in the order $\{ST, AM, M, D\}$
Player adjustment			✓	$SkewNormal\begin{pmatrix} \mu = 0, \sigma = \sigma, \\ \alpha = 2 \vee 0 \end{pmatrix}$ $\sigma \sim HalfNormal(\sigma = 5)$ Where α is assigned depending on prior beliefs about a player (good finisher = 2, not good finisher = 0)

Table 3 lists the predictors in each of the models and the prior distributions used for them. Any variables for which there is no logical reason for a coefficient to be positive or negative have been given a normal distribution as a prior, while those for which an effect direction can be logically predicted have been given a skewed normal distribution prior because this favours values in that direction while not completely discounting values in the opposite direction in case the prediction is wrong. The justifications for these skews are given below:

Distance to goal ($\mu = -1$ and $\alpha = -1$): On average it is more difficult to score the further from goal you are shooting because the goalkeeper has more time to react to the shot.

Shot angle ($\mu = 1$ and $\alpha = 1$): This is because increasing shot angle will either mean getting closer to the goal or more central to the goal, both of which mean it is easier to score on average.

Players in shot triangle ($\alpha = \{5, \dots, -5\}$): As more players are present in the shot triangle it becomes harder for a player to shoot in such a way that their shot won't hit one of the players and still find the goal.

Opponents in 1m radius ($\alpha = \{1, \dots, -2\}$): Similar to above, as more opponents are in the shooter's radius there will be more players trying to put the shooter off and blocking the shot so it will be harder to score.

Goalkeeper in shot triangle ($\alpha = -2$): When the goalkeeper is in the shot triangle, they are better placed to save a shot than if they are not in the shot triangle.

One-on-one shot ($\alpha = 2$): If a player is one-on-one with the goalkeeper, then they only have to beat the goalkeeper with their shot and not try to avoid or focus on any other players so it should be easier to score.

Open goal shot ($\alpha = 4$): Similar to above, but now there is no goalkeeper so the shooter only has to focus on getting the shot on target so should be very likely to score.

Shooter under pressure ($\alpha = -2$): If a player is under pressure they will have less time to focus on shooting well and on target and so will be less likely to score on average.

General position adjustment ($\alpha = \{ST{:}\,2,\ AM{:}\,1,\ M{:}\,0,\ D{:}\,-2\}$): The values of alpha capture the belief that strikers will be the most prolific finishers, followed by attacking midfielders, then other midfielders, and finally a drop for defenders.

Player adjustment ($\alpha = 2 \vee 0$): If a player is expected to be a good finisher based on their name and reputation they are assigned 2, and otherwise 0 for alpha.

$\sigma = 5$ was the standard value for the priors to allow for enough variation that priors will not be too narrow if the prior knowledge they are based on turns out to be false, but also not so large that convergence takes many rounds of sampling.

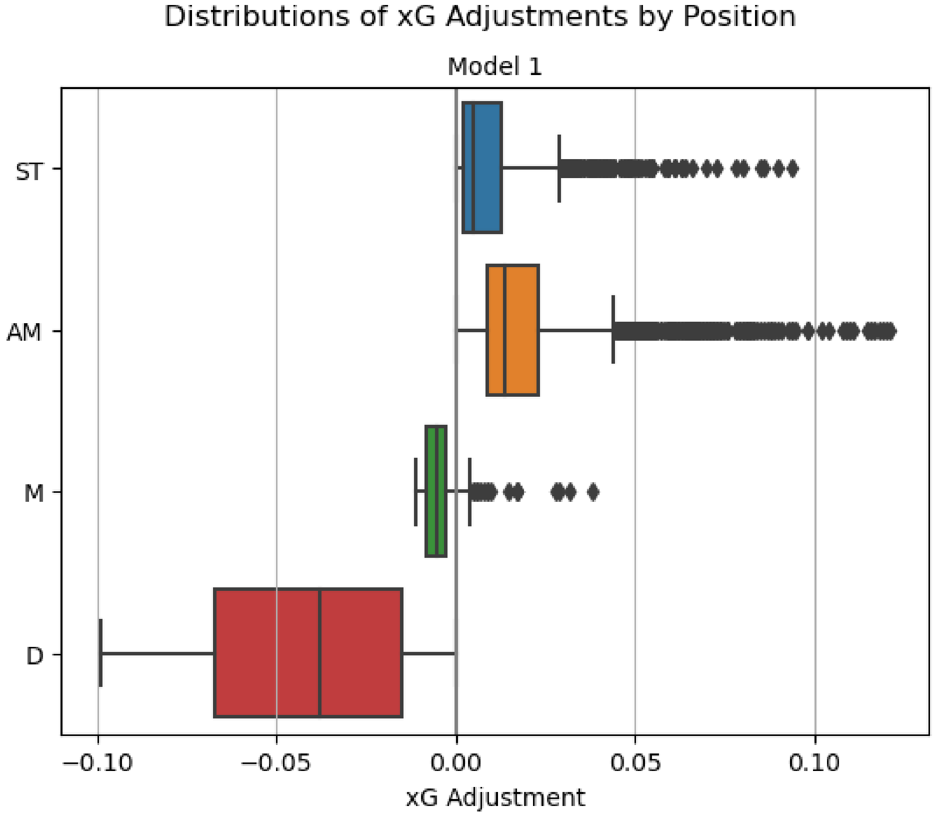
Markov chain Monte Carlo is used to produce posterior distributions in the model with 100 draws as well as 100 additional draws for tuning. This will be done 4 times to produce 4 chains resulting in 800 total samples for each model. Moreover, to reduce sampling times the Bayesian modelling was performed only on Premier League data, which amounted to roughly 10,000 shots from the data. A single league was chosen instead of simply sampling 10,000 shots from the data because it gives a good range of players and matchups of teams, whereas a large chunk of the overall data was just Barcelona matches due to the data which StatsBomb have made publicly available which could have skewed the results.

Results

Model 1 used distance to goal, shot angle, and the interaction between the two as predictors. The differences in xG predictions between the single-level and hierarchical model were taken for each shot in the data to determine the impact of the grouping variable, general position, on xG.

Figure 5 shows the distributions of the xG adjustment after accounting for general position in the hierarchical model for each position category. The prior beliefs about the impact of general

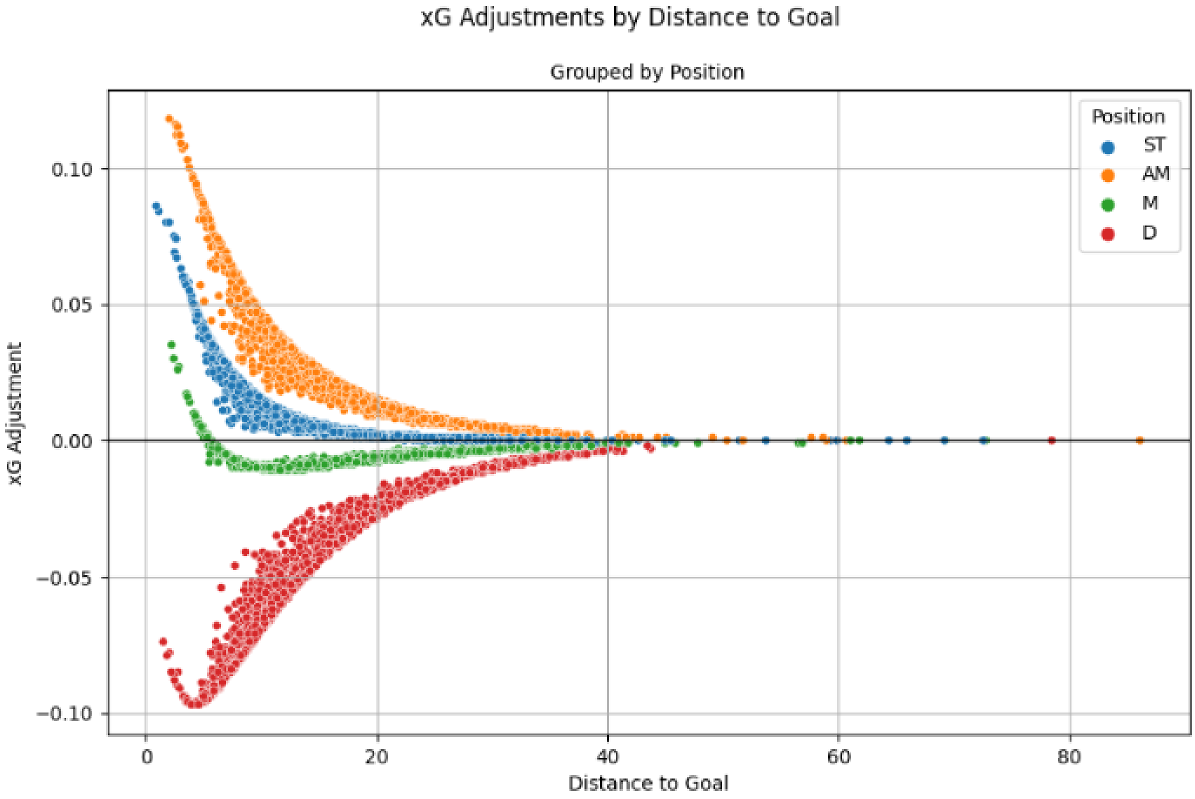
Figure 5: Distributions of xG adjustments for Model 1, where adjustment is hierarchical model prediction minus baseline model prediction. Grouped by general position.



position are mostly supported by these results. Defenders have a large number of negative xG adjustments compared to the baseline model's xG predictions, with some adjustments going as low as 0.1 lower. Non-attacking midfielders have smaller xG adjustments, both positive and negative, as was expected. Both strikers and attacking midfielders have mostly positive xG adjustments as predicted, but attacking midfielders seem to have larger positive adjustments on average than strikers which was not expected. This could be reflective of the fact that strikers tend to be in high xG scoring positions when shooting more frequently than attacking midfielders so their xG values are already high enough without positional adjustment. Whereas attacking midfielders take shots from around the area or either side of the goal more often which tend to be lower xG chances, but they still score them because they are above average at attacking and scoring.

It is also interesting to look at the xG adjustments across the ranges of the predictors distance to goal and shot angle. Figure 6 shows each of the xG adjustments by distance to goal grouped according to position. The convergence of each position towards an adjustment of 0 reflects the fact that at a certain distance the xG value will be close to zero regardless of any other

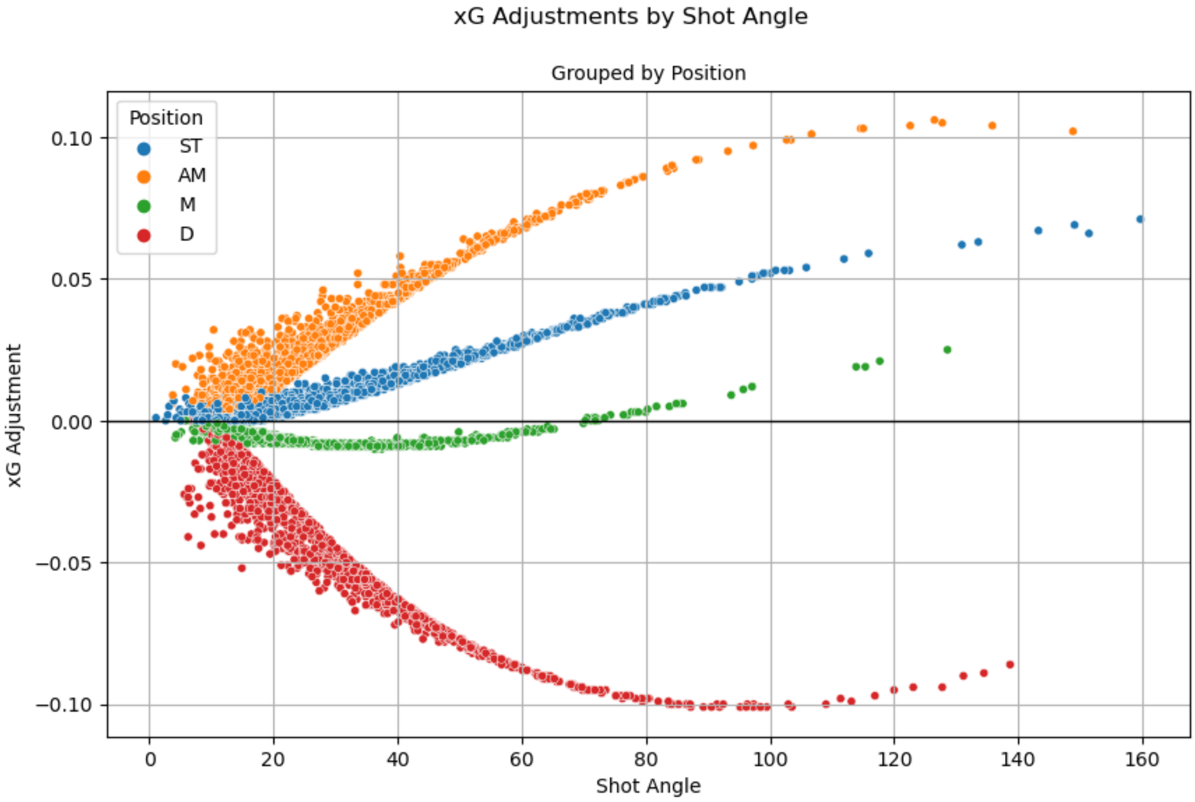
Figure 6: Point estimates for xG adjustments for Model 1 plotted against distance to goal, grouped by general position of shooter.



factors about the shot. Interestingly, there is a slight dip in xG adjustments for defenders at the lowest distance to goal before then increasing to convergence. This could reflect situations where defenders are getting into very good goalscoring positions following set pieces (not directly as these are excluded) or if a team is chasing a game since defenders are typically good in the air, so will be likely to score close range headers in these situations. For the other three positions there is a relatively consistent convergence of xG adjustments, and the same order of adjustments shown in Figure 5 with attacking midfielders having the largest positive adjustments followed by strikers, and other midfielders having close to no adjustment from the baseline model.

A similar analysis can be performed for shot angle, which is shown in Figure 7. Smaller values for shot angle correspond to more difficult chances to score either in the form of being far from goal or from being a tight angle either side of the goal. This is similar to the effect above of convergence where the xG values are so small anyway that features like player's general position have little impact. From there, defenders have gradually larger negative adjustments as shot angle increases, but again this trend reverses after a certain point for very high shot angles which likely corresponds to very close distances to goal. Otherwise, the same results

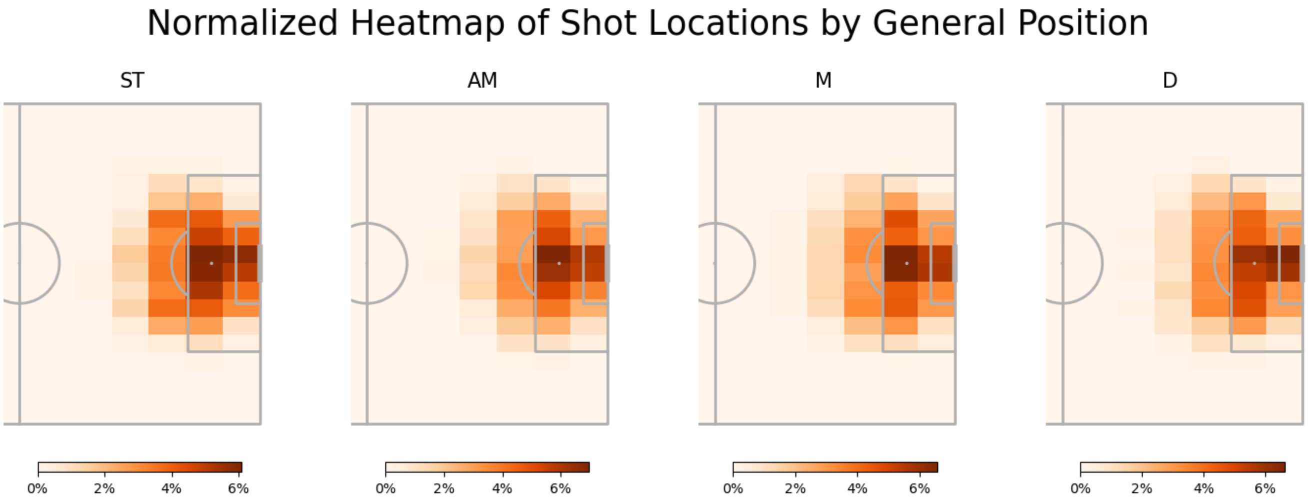
Figure 7: Point estimates for xG adjustments for Model 1 plotted against shot angle, grouped by general position of shooter.



appear again for the other positions with attacking midfielders having the largest positive xG adjustments, followed by strikers and other midfielders respectively. The impact of shot angle for these positions diverges initially until a shot angle of around 60, at which point the effect of an additional degree on shot angle has roughly the same effect on xG adjustment for strikers and all midfielders on average, i.e. the slope of the increase is roughly the same after this point.

Figure 8 shows the shot locations for each general position category and normalizes them for each position. It shows the relatively large number of chances that defenders get right in front of the goal, which could describe the effect reversals seen in Figure 6 and Figure 7. It also shows that attacking midfielders do not tend to take many shots from far away or difficult angles as was theorized when describing the results of Figure 5, and in fact strikers are more likely to have these kinds of shots on average. This finding, paired with the fact that attacking midfielders have larger positive xG adjustments than strikers on average, suggests that attacking midfielders are better on average at converting high xG chances right in front of the goal than strikers.

Figure 8: Heatmap of shot locations for each general position, normalized for each group.



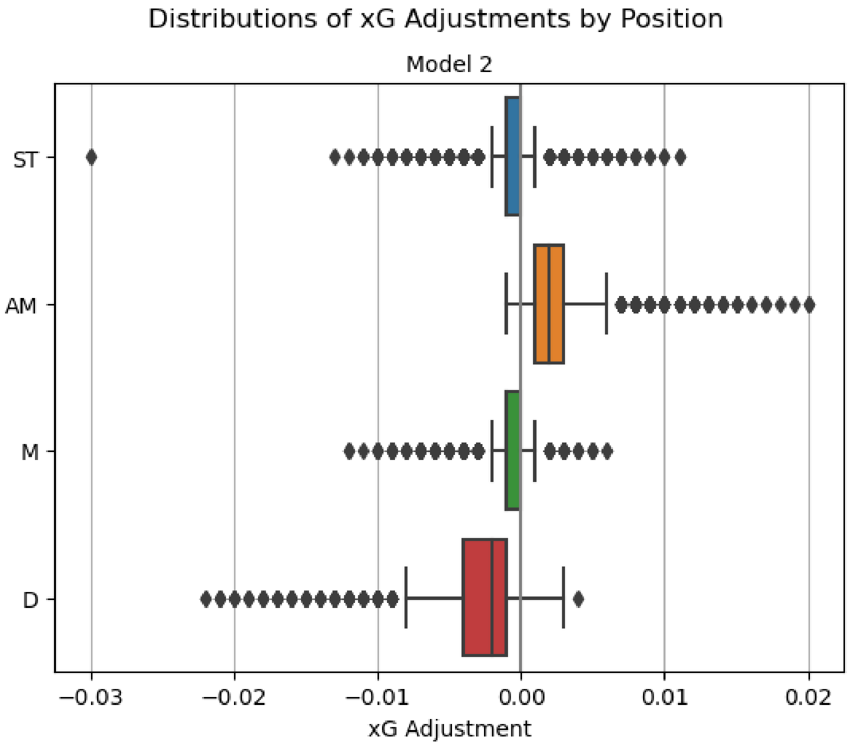
To assess how well the model captures the effects discussed in the justification of Bayesian methods section, whereby non-Bayesian xG values were adjusted by position using Bayes' Theorem, the results of the baseline model can also be adjusted using Bayes' Theorem and compared to the results of the hierarchical model to see if the theoretical adjustment is close to the model adjustment. The results of this process are shown in Table 4. The mean model and theoretical adjustments are very close for each position confirming the model has estimated the positional impact well.

Table 4: Mean xG adjustment for each general position from Model 1 versus theoretical adjustment of baseline Model 1 predictions using Bayes' Theorem.

Position	Mean Model Adjustment	Mean Theoretical Adjustment
ST	0.009	0.010
AM	0.019	0.020
M	-0.006	-0.005
D	-0.042	-0.044

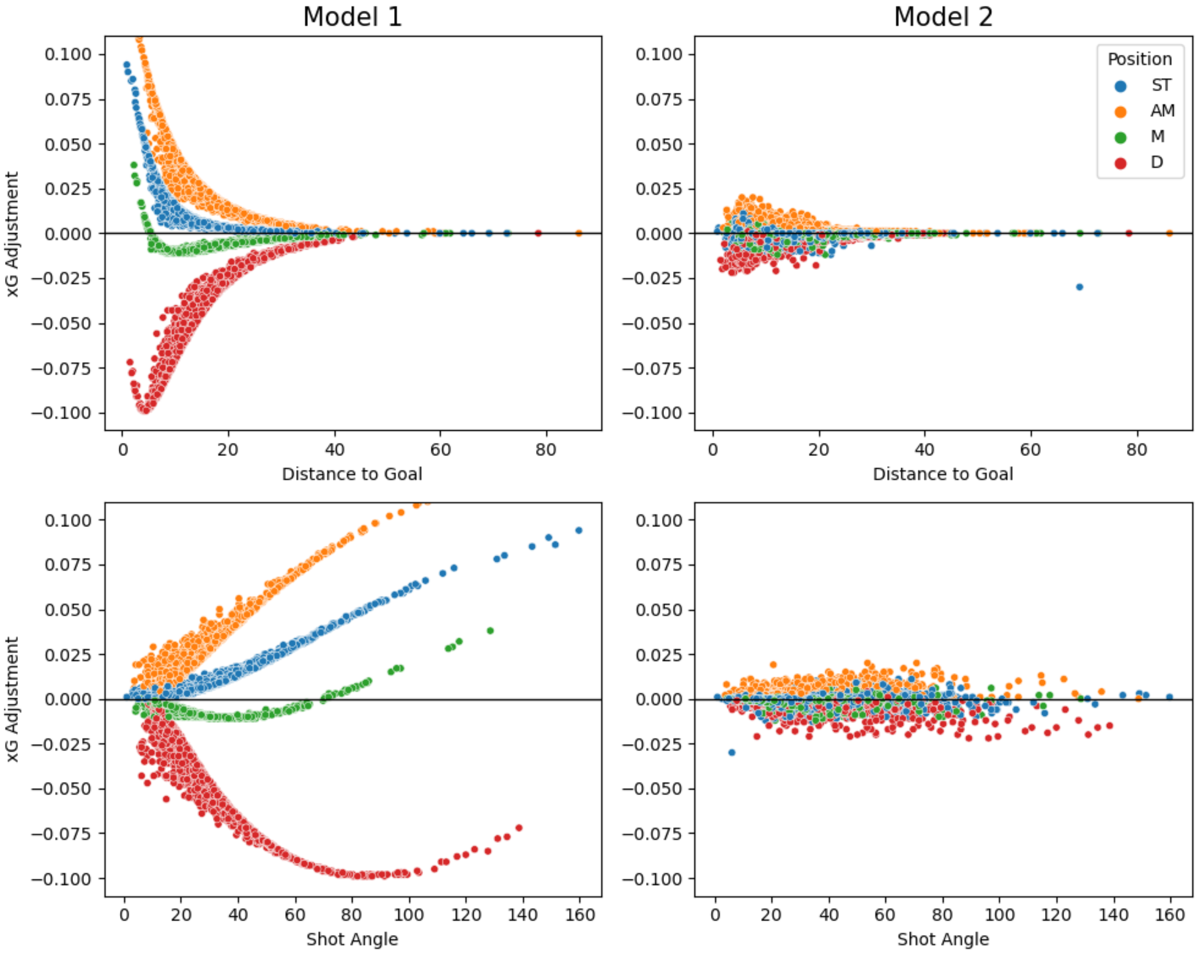
Model 2 added numerous additional predictors to improve the baseline predictions of xG before positional effects are considered. Now looking at the distributions of the xG adjustments by position in Figure 9, the values are much smaller than for model 1. There are now very few adjustments greater than 0.01 away from the baseline xG values. This suggests that the additional predictors are able to predict the position well and that xG advantages come less from the position of the player and more from the play situation that players are in when shooting. For example, attackers will tend to have better scoring chances on average which contributes to the positive xG adjustment they received in the basic model, but by controlling for a lot of the features that define those better scoring chances (e.g. one-on-ones) the effect of the position is diminished.

Figure 9: Distributions of xG adjustments for Model 2, where adjustment is hierarchical model prediction minus baseline model prediction. Grouped by general position.



This point is emphasised if the same plots for distance to goal and shot angle that were shown above are also shown with the predictions from the extended model, shown in Figure 10. Where before there was a clear distinction in effects of distance to goal and shot angle to xG adjustment for each position, there is no no clear difference between positions. The convergence towards 0 for distance is still present but for shot angle there is no significant relationship between shot angle and xG adjustment. The likely explanation for the diminishing effect of position on xG adjustment is that each position category contains a wide variety of players of different abilities and with different roles. Some strikers are primarily playmakers instead of a traditional poacher so will perhaps not be picked based on their goal scoring ability, while some central midfielders (not in the attacking midfielder category) will be very good at scoring from just outside of the box compared to other midfielders or more attacking players. This is the rationale behind model 3, which groups the data according to the player taking the shot as opposed to the player's general position. Fitting the model with

Figure 10: Comparison of point estimates for xG adjustments against distance to goal and shot angle between Model 1 and Model 2, grouped by general position. Adjustments are hierarchical model prediction minus baseline model prediction.



all of the players in the data as a group would take far too long, so instead a few players will be selected, and all other players will be classed as "other". To select the players, it makes sense to have some players for which either a positive or negative xG adjustment is expected. To do so, the data was filtered to only include players with at least 50 shots before calculating their conversion rates. The selected players and their statistics are shown in Table 5. Pirès, Agüero, and Vardy were all selected because they are prolific goal scorers, and it is expected that they will have positive xG adjustments. Pirès especially has a brilliant conversion rate for the shots that are included in this data. On the other hand, Coutinho and Barkley both have below average conversion rates so may have slight negative xG adjustments, while Shelvey converted none of his 51 shots in the data so it is likely he will have a more significant negative xG adjustment.

Table 5: Selected players for Model 3 and their goal scoring statistics.

Player	Shots	Goals	Conversion Rate
Robert Pirès	56	14	25%
Sergio Agüero	112	20	17.9%
Jamie Vardy	111	19	17.1%
Phillippe Coutinho	105	8	7.6%
Ross Barkley	82	6	7.3%
Jonjo Shelvey	51	0	0%

As was outlined in the methodology, the effects of each of the players will have a prior distribution defined by a skewed normal distribution with either $\alpha = 2$ or $\alpha = 0$ depending on if prior beliefs suggest that the player is a good goal scorer. The data that will be used for model fitting should not directly be used to inform priors, so instead qualitative beliefs are used. Pirès, Agüero, Vardy, and Coutinho are all generally considered quality attacking players, so they are assigned $\alpha = 2$. Barkley and Shelvey are not typically associated with being first-class attackers and have other qualities that better define their game, so they are assigned $\alpha = 0$.

Figure 11 shows the distributions of xG adjustments for each player as well as the "other" players group. The most noticeable result is the huge positive xG adjustments for Robert Pirès, with some adjustments being as high as 0.3 above the baseline xG, even with all the additional predictors which eliminated the group effects in Model 2. He also has a very wide spread of adjustments, with some adjustments being close to 0. Such a wide spread is likely indicative of Pirès having lots of different types of shots, some which were high xG chances anyway so didn't need much adjustment, while others were more difficult to score but Pirès scored them relatively consistently hence the large positive adjustments. Agüero also has only positive xG adjustments, but smaller than those of Pirès on average and with a smaller spread. Interestingly, Vardy and Coutinho have only very small positive xG adjustments not much greater than those of Barkley. Shelvey does have the expected result of large negative xG adjustments based on his conversion rate in the data. Finally, the "other" group has very little xG adjustment centralised around 0. This is expected since this group contains many different players so there is no group-effect to capture here.

Figure 11: Distributions of xG adjustments for Model 3, where adjustment is hierarchical model prediction minus baseline model prediction. Grouped by player.

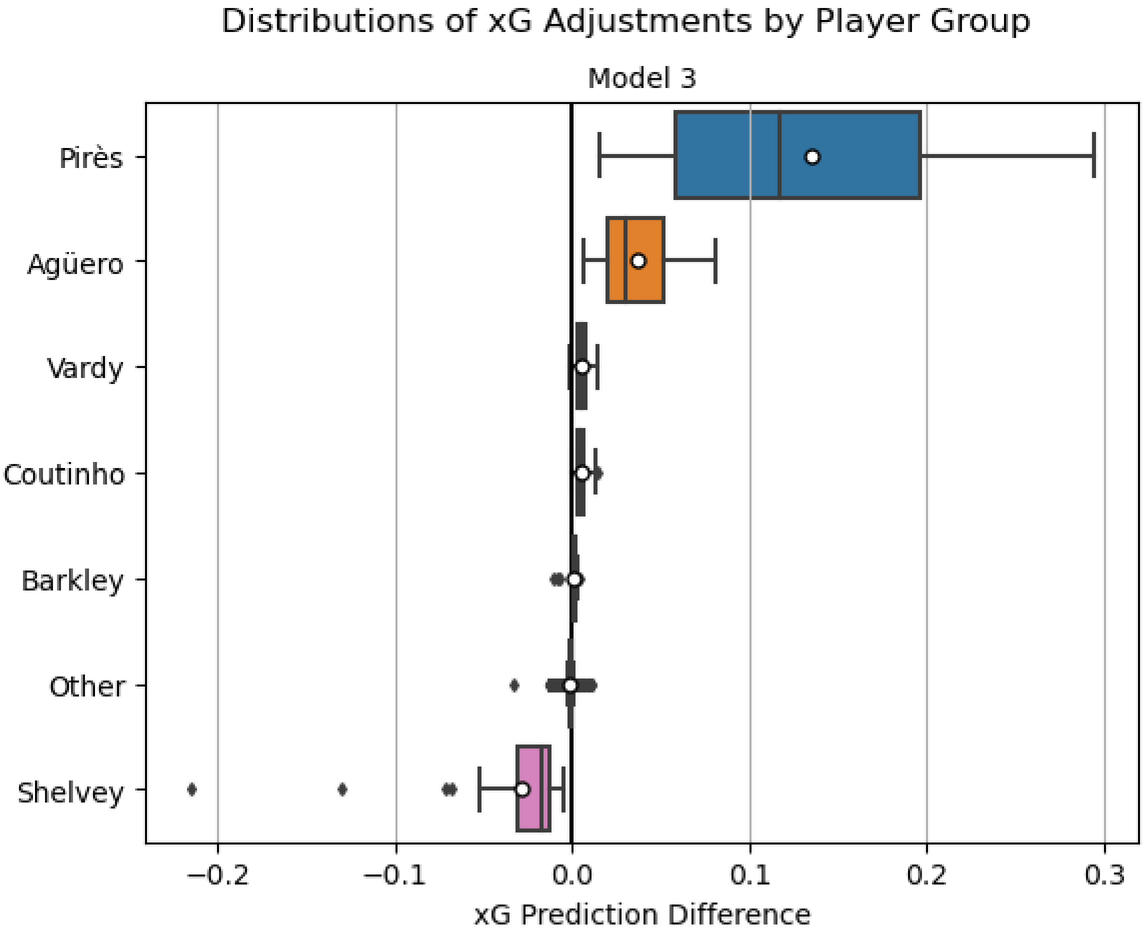


Figure 12 shows the shot locations of the selected players and the outcomes of the shots, and it provides potential rationales for the results shown in Figure 11. Starting with Pirès, who had the biggest positive xG adjustments, the first thing that stands out is how few shots he took and how many of them resulted in goals. This alone would suggest Pirès is a great goalscorer and should get positive xG adjustments based on this data. Moreover, he scored some goals from very difficult positions, e.g. both corners of the box and outside the area, while not having many shots from these locations. This implies that even for hard to score chances Pirès is a good goal scorer which will further increase his xG adjustments. Agüero and Vardy's shot maps look very similar, with lots of shots and goals from both inside and just outside of the area, but the model results had significantly larger positive xG adjustments for Agüero than for Vardy. A possible explanation for this is that the types of shots Vardy was having were high xG chances anyway, e.g. one-on-one shots, while Agüero's shots were harder to score based on defender positioning or other features but he still scored them, hence the larger adjustments. Vardy's xG adjustments according to Figure 11 were also quite similar to those of Coutinho and Barkley but the latter's shot maps look less impressive than Vardy's on face value with fewer goals scored for both from similar shots. The fact that all three have small xG adjustments is indicative of their goal scoring from the shots they took being in line with the xG values from the baseline model considering all the features about a shot. In short, most of the shots they take do not require player-adjustment and are accurately predicted without the player-effects. Finally, Shelvey's shot map obviously has no goals scored from a variety of different positions. For difficult to score shots such as those outside the area or from tight angles either side of the goal there would likely to small adjustments because these are hard to score anyway, but the easier chances centrally in the box which were not converted are likely the ones which had the largest negative adjustments due to Shelvey's poor conversion rates from these positions in this data.

Figure 12: Selected player shot locations and goals.

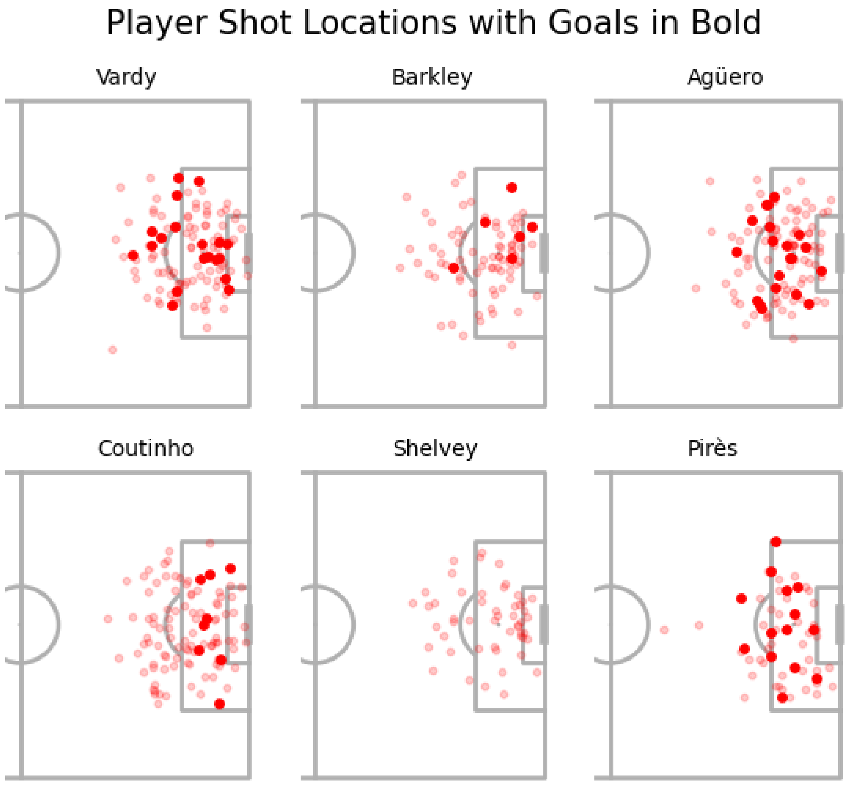
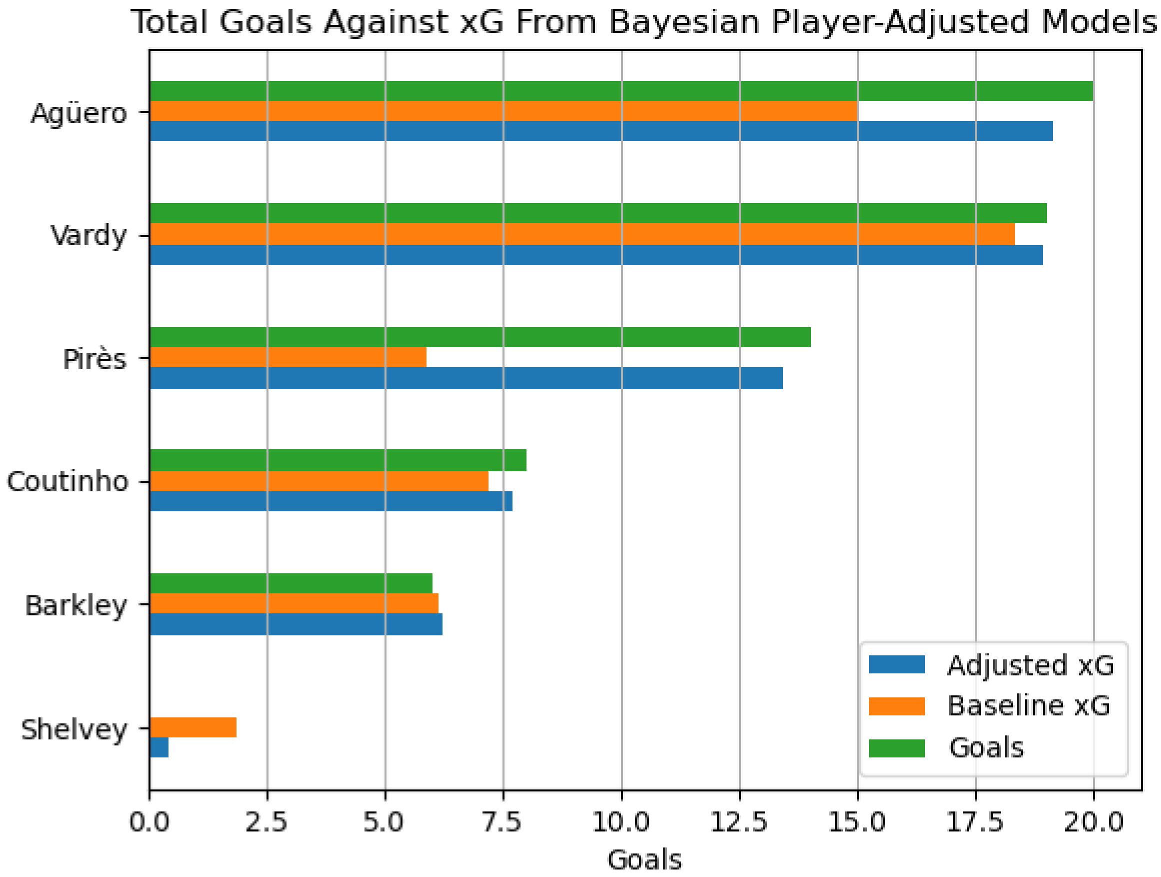


Figure 13 shows the totals of goals scored, baseline xG from the single-level model, and adjusted xG from the player-corrected model for each of the selected players for this analysis. It shows that the player-corrected model gives closer estimates for total goals scored for all the players apart from Barkley, where the difference is very small. Pirès and Agüero, whose actual goals scored is far higher than their baseline xG meaning they are prolific at scoring hard to score chances, have adjusted xG total much closer to their actual goals scored total. On the other hand, Shelvey's adjusted xG total is much closer to the zero goals he scored but crucially is not zero. This reiterates the benefit of using Bayesian methods over frequentist methods because in a frequentist mindset Shelvey's xG should be 0 because he has not scored any of the 51 shots he took in this data, or the model will be unable to provide an estimate due to there being no variation in the dependent variable goal. This is not realistic as clearly being a top-level professional footballer he still has a chance to score, which is accounted for in the Bayesian model and adjusted appropriately using the hierarchical model.

Figure 13: Comparison of Model 3 baseline predictions, Model 3 hierarchical predictions, and actual goals scored for selected players.



Analysing Choice of Prior Distributions

The choice of prior distributions can be crucial for Bayesian modelling for a few different reasons. Firstly, if you use very wide prior distributions that consider a large number of values likely, then it may take many samples for the model to find the true values of the variables in the model which will take a long time. On the other hand, if prior distributions are too narrow and the true value is not likely to appear from the distribution, then the results of the model could be incorrect.

To assess whether the choice of prior distributions here was appropriate, the extended baseline model (single-level) will be refitted with different priors and the predictions will be compared. The sets of priors to be used will include: the current priors, all wide uniform priors, and a set of deliberately bad priors. Additionally, the resulting predictions from these models will be compared to the extended non-Bayesian model's predictions of the same data to provide a baseline for good predictions.

Table 6 shows the prior distributions for each of the predictors in the model. For the uniform priors, each variable has been given a wide uniform distribution to allow for a whole host of values to occur. On the other hand, the bar priors have been given very narrow distributions by using a small value for σ. Moreover, some of the skews in the distributions have been flipped such that the prior belief about the effect is the reverse of what was actually used.

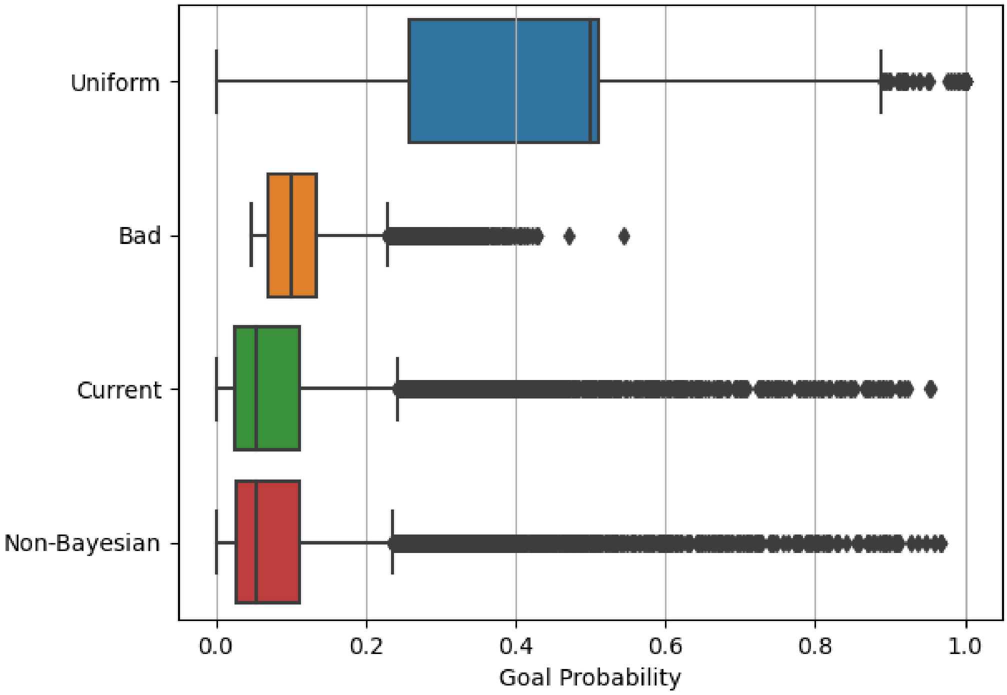
Table 6: Choices of prior distributions for analysing the impact of prior choice on results. Note, the current priors are in Table 3.

Predictor	Uniform Priors	Bad Priors
Intercept	$Unif(a = -100, b = 100)$	$SkewNormal(\begin{matrix}\mu = 0, \sigma = 0.25, \\ \alpha = 2\end{matrix})$
Distance to goal	$Unif(a = -100, b = 100)$	$SkewNormal(\begin{matrix}\mu = 0, \sigma = 0.25, \\ \alpha = 2\end{matrix})$
Shot angle	$Unif(a = -100, b = 100)$	$Normal(\mu = 0, \sigma = 0.25)$
Distance * angle	$Unif(a = -100, b = 100)$	$SkewNormal(\begin{matrix}\mu = 0, \sigma = 0.25, \\ \alpha = -2\end{matrix})$
Goalkeeper distance to goal	$Unif(a = -100, b = 100)$	$Normal(\mu = 0, \sigma = 0.25)$

Players in shot triangle	$Unif(a = -100, b = 100)$	$SkewNormal(\mu = 0, \sigma = 0.25, \alpha = \{-5, \dots ,5\})$
Opponents in 1m radius	$Unif(a = -100, b = 100)$	$SkewNormal(\mu = 0, \sigma = 0.25, \alpha = \{1, \dots , -2\})$
Body part	$Unif(a = -100, b = 100)$	$Normal(\mu = 0, \sigma = 0.25)$
First-time shot	$Unif(a = -100, b = 100)$	$Normal(\mu = 0, \sigma = 0.25)$
Goalkeeper in shot triangle	$Unif(a = -100, b = 100)$	$SkewNormal(\mu = 0, \sigma = 0.25, \alpha = 2)$
One-on-one shot	$Unif(a = -100, b = 100)$	$SkewNormal(\mu = 0, \sigma = 0.25, \alpha = -2)$
Open goal shot	$Unif(a = -100, b = 100)$	$SkewNormal(\mu = 0, \sigma = 0.25, \alpha = -4)$
Shot technique	$Unif(a = -100, b = 100)$	$Normal(\mu = 0, \sigma = 0.25)$
Shooter under pressure	$Unif(a = -100, b = 100)$	$SkewNormal(\mu = 0, \sigma = 0.25, \alpha = 2)$

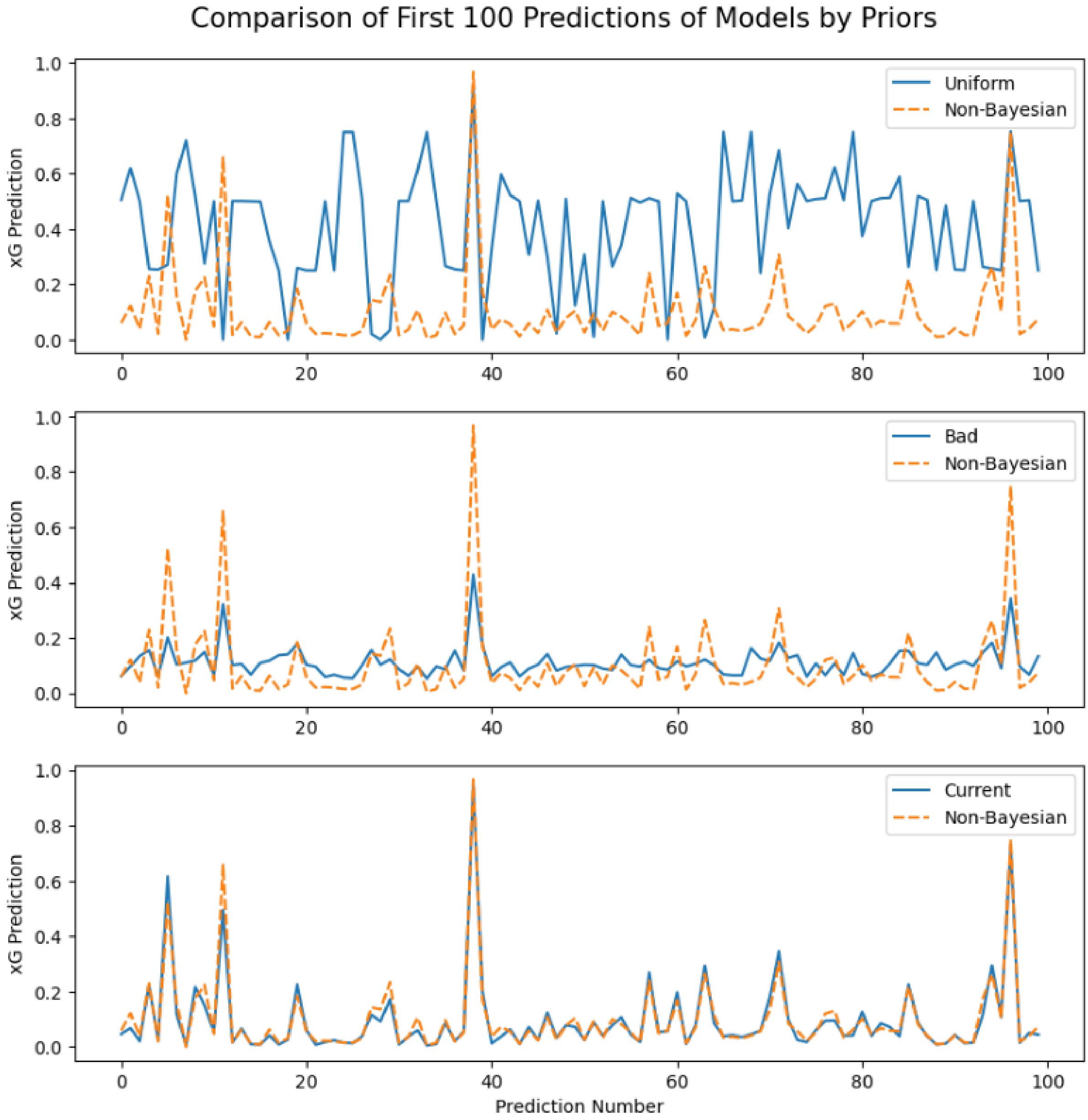
Figure 14 shows the distributions of the predictions of each of the models. The model fitted with the uniform prior distributions performs very poorly compared to the non-Bayesian baseline model, with a very large spread of predictions. On the other hand, the bad prior choices provide too narrow a spread, with almost no xG predictions above 0.5. However, the bad priors do perform better than the uniform priors in terms of getting the average close to the non-Bayesian predictions and having a similar sized interquartile range. This is likely due to the number of samples used to estimate the parameters. There were not enough samples for the model with uniform priors to settle on a good value for each of the parameters in the model so the model predictions were poor. Given enough samples, this model would indeed converge to give good results, but this could take far longer to run and the exact number of samples needed is unknown. The model with the current prior distributions performs the best by far, and almost mirrors the non-Bayesian model’s predictions in terms of the distribution.

Figure 14: Distributions of xG predictions for each of the extended single-level Bayesian model, with different choices of prior distributions.



It is also insightful to see how the models predicted given values in the data to solidify the results of Figure 14. Figure 15 shows the predictions of each of the Bayesian models for the first 100 observations compared to the non-Bayesian predictions. The first plot, with the uniform model predictions, shows the poor predictive power of this model with very few predictions being close to the non-Bayesian predictions, and a large volatility in the predictions. Once again the model with the bad priors gives the opposite of the uniform model, with very little volatility of predictions, but still with better accuracy than the first model. Finally, the model with the current priors fits very closely to the non-Bayesian predictions, solidifying the point that the current priors are sufficient for the model to provide good xG predictions.

Figure 15: Point estimates of the first 100 predictions for Model 3 with different sets of priors. The priors are listed in Table 3 and Table 6.



Discussion and Future Work

This paper has developed multiple Bayesian models to assess the impact of player's position and individual player effects on xG predictions. A basic xG model with only distance to goal, shot angle, and the interaction between them suggested that there were positional effects on xG, with strikers and attacking midfielders having positive xG adjustments, midfielders having minimal adjustments, and defenders having significantly negative xG adjustments on average. However, adding more predictors to the models reduced the positional effects to the point where the effects were almost insignificant, implying player position had very little impact on xG when more factors about shots are considered. Next, player effects were considered by using the same extended model as for the second positional-effects model, but grouping the data based on the player shooting instead of the shooter's position. Six players were used to illustrate how the model worked, and the results showed that there were player effects on xG even when controlling for a number of factors related to the shot. These effects were also present in both directions, most notably very positive for Robert Pirés and negative for Jonjo Shelvey.

The evidence that there are player-specific effects for determining goal probability could be useful for football scouting and player selection. If you calculated the adjusted xG values for many different players and compared their adjusted values to their non-adjusted values, as we have done here, you can identify which players are truly good at finishing hard chances and which players simply tend to be in the right place at the right time. Consider the findings for Jamie Vardy in this analysis, whereby his total adjusted xG is not that different to his baseline xG (Figure 13). This implies that based on the quality of chances Jamie Vardy is having he is scoring them at a fairly average rate, while Sergio Agüero is scoring more consistently from more difficult positions due to his larger adjusted xG. Obviously, this doesn't mean that Agüero is a better player or attacker than Vardy, just that for chances that tend to have lower xG values, this data suggests that Agüero is better at converting those chances than Vardy on average.

It is worth noting that there may also be team style effects being captured here. Continuing with the comparison of Vardy and Agüero, Vardy plays for a Leicester team that plays with high tempo and with direct attacking play, likely resulting in shooting situations where the ball is played behind the defence and there are fewer defenders to block or interfere with a shot, or simply a one-on-one with the goalkeeper, meaning Vardy is getting lots of high xG chances anyway. On the other hand, Agüero was playing for one of the best teams in the league meaning opponents tend to play conservatively and don't give Manchester City much space, resulting in more difficult chances with multiple players in the shot triangle among other features.

It would be beneficial to perform the same analysis on a different dataset, perhaps on a different league, to see if there are still significant player effects and a lack of positional effects when multiple features are considered. Moreover, being able to fit the model with more than just seven player groups (six players and "other") would be beneficial to see if the expected results hold for other players too. Perhaps the sampling times can be reduced by reducing σ for the prior distributions so that convergence is quicker, allowing for a more complex model without infeasible waiting times.

The results of the Bayesian modelling were performed with the backdrop of non-Bayesian, or frequentist, modelling. The main benefit of Bayesian modelling, especially for the player-corrected case, is its ability to capture uncertainty due to the output of the models being posterior distributions and not directly point estimates. Additionally, its ability to deal with data groups with few observations through hierarchical modelling is very useful for dealing with younger players who have not played many matches. In addition to scouting and player selection, Bayesian hierarchical modelling could also be applied to other metrics and areas in football. For example, injury risk assessment of players using data is performed within many large football clubs, where there could be an element of grouping in the data as some players are more injury prone than others overall. Using hierarchical modelling one could potentially more accurately determine injury risk for individual players based on their injury history.

Conclusion

This paper aimed to investigate if there are position or player effects on xG values, implying different positions or players have certain xG adjustments for given chances. The expectations were that better attacking players (strikers, attacking midfielders, or good attacking players) would have positive xG adjustments while players who are not generally lauded for their attacking output (defenders, or defensive players) would have negative xG adjustments.

This was investigated using three Bayesian models: Model 1 which used distance to goal, shot angle, and an interactions term between these as predictors as well as including positional effects, Model 2 which included the same positional effects but added a host of other predictors about each shot, and Model 3 which used all the predictors from Model 2 but changed from positional effects to individual player effects. Model 1 provided evidence of positional effects on xG, with strikers and attacking midfielders having positive xG adjustments, other midfielders having minimal adjustments, and defenders having large negative adjustments. However, Model 2 found that once additional features about shots are added the positional effects are minimal, implying that the variation in xG comes more from the type of chance itself than the position of the player taking the chance. It was theorized that the positional categories contained too much variation in actual player quality and styles, providing the rationale for a player adjusted model. Model 3 showed that even after accounting for all the predictors in Model 2 there were still significant player effects in both positive and negative directions, which were in line with the shot conversion rates for the given players.

The results of this paper have shown the benefit of using Bayesian modelling over non-Bayesian, or frequentist, modelling. Bayesian modelling not only improves estimates for groups with few observations in the form of hierarchical modelling, but also gives distributions instead of point estimates as outputs, meaning the uncertainty of each estimate can be seen. Regardless, Bayesian modelling is rarely used in the literature of football analytics. Future work should look to apply Bayesian methods to pre-existing football analytics concepts, such as injury prevention and other metrics than xG, for its clear benefits in terms of uncertainty in a game which is defined by chaos and complexity.

References

Aberle, M., Figdor, L., Mongrand, L. & Janetzke, M., 2020. *The tech behind the Bundesliga Match Facts xGoals: How machine learning is driving data-driven insights in soccer.* [Online]
Available at: https://aws.amazon.com/blogs/machine-learning/the-tech-behind-the-bundesliga-match-facts-xgoals-how-machine-learning-is-driving-data-driven-insights-in-soccer/
[Accessed 17 August 2023].

Baio, G. & Blangiardo, M., 2010. Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics,* 37(2), pp. 253-264.

Brechot, M. & Flepp, R., 2020. Dealing With Randomness in Match Outcomes: How to Rethink Performance Evaluation in European Club Football Using Expected Goals. *Journal of Sports Economics,* 21(4), pp. 335-362.

Cavus, M. & Biecek, P., 2022. *Explainable expected goal models for performance analysis in football analytics.* Shenzhen, IEEE.

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,* Volume 4, pp. 165-174.

Herberger, T. A. & Litke, C., 2021. *The Impact of Big Data and Sports Analytics on Professional Football: A Systematic Literature Review.* s.l., Springer.

Herold, M. et al., 2019. Machine learning in men's professional football: Current applications and future directions for improving attacking play. *International Journal of Sports Science & Coaching,* 14(6), pp. 798-817.

Hewitt, J. H. & Karakuş, O., 2023. *[2301.13052] A Machine Learning Approach for Player and Position Adjusted Expected Goals in Football (Soccer).* [Online]
Available at: https://arxiv.org/abs/2301.13052
[Accessed 14 August 2023].

Joseph, A., Fenton, N. & Neil, M., 2006. Predicting football results using Bayesian nets and other machine learning techniques. *Knowledge-Based Systems,* 19(7), pp. 544-553.

Lucey, P. et al., 2015. *"Quality vs Quantity": Improved Shot Prediction in Soccer using Strategic Features from Spatiotemporal Data.* Boston, MIT Sloan.

Madrero Pardo, P., 2020. *Creating a model for expected Goals in football using qualitative player information.* [Online]

Available at: https://upcommons.upc.edu/handle/2117/328922
[Accessed 14 August 2023].

Mead, J., O'Hare, A. & McMenemy, P., 2023. Expected goals in football: Improving model performance and demonstrating value. *PLoS ONE,* 18(4).

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

Raudonius, L. & Seidl, T., 2023. *Shot Analysis in Different Levels of German Football Using Expected Goals.* Grenoble, Springer.

Robberechts, P. & Davis, J., 2020. *How Data Availability Affects the Ability to Learn Good xG Models.* Ghent, Springer.

Smith, R., 2022. *Expected Goals: The story of how data conquered football and changed the game forever.* 2nd ed. London: Mudlark.

Spearman, W., 2018. *Beyond Expected Goals.* Boston, MIT Sloan.

StatsBomb, 2019. *StatsBomb Data Specification v1.1.* [Online]
Available at: https://github.com/statsbomb/open-data/blob/master/doc/StatsBomb%20Open%20Data%20Specification%20v1.1.pdf
[Accessed 08 September 2023].

Tippett, J., 2019. *The Expected Goals Philosophy: A Game-Changing Way of Analysing Football.* UK: James Tippett.

Tureen, T. & Olthof, S., 2020. *Tahmeed Tureen, Sigrid Olthof - StatsBomb 2022.* [Online]
Available at: https://researchonline.ljmu.ac.uk/id/eprint/17636/1/Tahmeed-Tureen-and-Sigrid-Olthof-%E2%80%93-Estimated-Player-Impact-EPI-Quantifying-The-Effects-Of-Individual-Players-On-Football-Actions-.pdf
[Accessed 14 August 2023].

Umami, I., Hardan Gutama, D. & Rahmania Hatta, H., 2021. Implementing the Expected Goal (xG) Model to Predict Scores in Soccer Matches. *International Journal of Informatics and Information Systems,* 4(1), pp. 38-54.

Zambrom-Ferraresi, F., Rios, V. & Lera-López, F., 2018. Determinants of sport performance in European football: What can we learn from the data?. *Decision Support Systems,* Volume 114, pp. 18-28.