ISyE 6420 – Final Project

Spring 2023

Garrett Ramos

Projecting Goal Contributions in MLS with Bayesian Regression

Introduction

The hardest thing to do in the sport of soccer is score goals. Scoring goals is a matter of not only having a player talented enough to put the ball in the back of the net when given the opportunity, but also having a player talented enough to provide those opportunities, by way of assisting. Clubs around the globe spend millions of dollars each and every year scouting players looking for those with the ability to influence matches, by assisting or scoring goals. The goal of this work, in particular, is to try and accurately predict how many goals and assists we can expect a prospective player to score per 90 minutes of match time given certain predictive factors.

Our specific case study of choice will be that of Major League Soccer, the top flight of professional soccer in the United States. We will then turn our attention to a more isolated case, that of New York City Football Club (NYCFC) of MLS, in order to see if we can reasonably project how many goals and assists each member of NYCFC contributes on a per 90 minute basis.

Data Selection

All of the data used for this report was accessed through football-reference. Furthermore, all of the data used is in the form of per 90 minutes, rather than total. With a special interest in predicting goals and assists, the response variable is defined as ‘G+A’, goals plus assists per 90 minutes. A number of factors were considered as possible predictors for G+A, and eventually these factors were limited to just 15. The preliminary 15 variables chosen were determined based on background knowledge under the assumption that these variables were strong indicators of creativity, decisiveness, and ability. For example, if a player shoots the ball a lot, there is a reasonably strong chance the player scores a decent amount of goals. Furthermore, if a player completes a lot of passes into the penalty area, there is a strong chance that player will get assists.

The preliminary variables for prediction are outlined in the table below:

Variable	***Description***
Sh	*Shots*
SoT	*Shots on target*
KP	*Key Passes*
Final 3rd	*Passes into final 3rd*
PPA	*Progressive passes into penalty area*
CrsPA	*Crosses into penalty area*
PrgP	*Progressive Passes*
SCA	*Shot creating actions*
Att 3rd	*Touches in attacking third*
Att Pen	*Touches in penalty area*
Live	*Open play touches*
Succ	*Successful take-on*
Fin 3rd Car	*Carries into final third*
CPA	*Carries into the penalty area*
PrgR	*Progressive passes received*

Figure 1 – Variable names and their description.

These stats were accumulated for every currently active player in MLS. The original dataset consisted of 634 players. This dataset was eventually restricted to players who have contributed at least 1 goal or assist in the 2023 MLS season. This left 228 players in the dataset prior to model building.

Model Building

With the response variable defined as G+A per 90 minutes, and the predictor variables defined above, the regression model can be written as follows:

$$\begin{aligned} GA = \beta_0 &+ \beta_1 * X_{Sh} + \beta_2 * X_{SoT} + \beta_3 * X_{KP} + \beta_4 * X_{Final3rd} + \beta_5 * X_{PPA} + \beta_6 * X_{CrsPA} + \beta_7 \\ &* X_{PrgP} + \beta_8 * X_{SCA} + \beta_9 * X_{Att3rd} + \beta_{10} * X_{AttPen} + \beta_{11} * X_{Live} + \beta_{12} * X_{Succ} \\ &+ \beta_{13} * X_{Fin3rdCar} + \beta_{14} * X_{CPA} + \beta_{15} * X_{PrgR} \end{aligned}$$

Using the Bayesian approach, definitions are required for the prior distributions. The intercept, denoted by β_0 is represented by a normal distribution as a non-informative prior, with a mean of 0, and a standard deviation of 100. Additionally, each predictor i is denoted by β_i, with a mean and standard deviation equal to the league average, and standard deviation for each respective predictive metric. For example, Shots per 90 minutes, written as 'Sh', denoted by β_1, is defined as a normal distribution with a mean of 1.79 (league average), and a standard deviation of 1.34

(league standard deviation). Lastly, the standard deviation, denoted by σ, is defined as a Half-Cauchy distribution with a beta value of 1.

The model was fit using 10,000 samples, where the posterior means for the model parameters resulted as follows:

β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}
-0.15	-0.2	0.79	0.13	-0.03	-0.03	0.09	0.05	0.07	-0.05	0.08	0.003	-0.004	0.23	-0.17	0.06

Figure 2 – Posterior means for model parameters.

Generally speaking, this model provided satisfactory results. Over the course of a soccer match, thousands of small actions take place with only a miniscule number of those actions resulting in a goal or assist. That being said, the predictions provided by the model were quite accurate, given the nature of the problem.

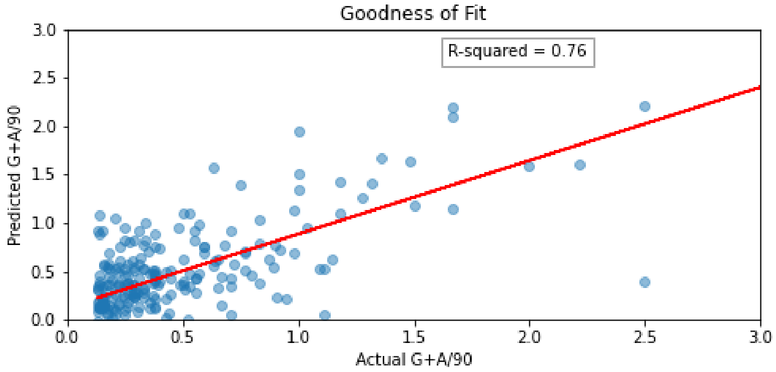


Figure 3 - Goodness of fit plot demonstrating R-squared value of 0.76.

Looking at Figure 3 above, an R-squared value of 0.76 indicates a relatively decent fit to the data. We can see that the predicted values are in the ballpark of the actual values associated with ‘G+A’.

NYCFC Case Study

In order to further evaluate the performance of the model, the squad of New York City Football Club was analyzed, with respect to each player and their goal contributions, both on the pitch and as predicted by the model. That being said, not only are the model’s predictions of interest, but also the industry recognized ‘xG+A’ values are of interest in order to gauge the model performance in comparison to the industry standard. Expected Goals, or xG, quantifies each shot as a probability between 0 and 1, in terms of how likely that shot is to go in. If a shot is taken very close to the goal it might have an xG value of 0.65, which says that the average player would be expected to score a similar chance 65% of the time. Additionally, the pass that led to that shot with an xG value of 0.65 would inherit an Expected Assist, or xA value of 0.65, respectively. These values are accumulated over the course of a 90 minute match in order to reflect a players' xG+A/90. While the collection methodology is quite different to the Bayesian regression model outlined in this report, the xG+A metric represents an intriguing point of comparison with our model's predictions.

Below is a table outlining players of NYCFC's squad, consisting of players who have registered at least 1 goal or assist thus far during the 2023 MLS campaign.

Player	***Actual G+A/90***	***xG+A/90***	***Predicted G+A/90***
Maxime Chanot	*0.14*	*0.13*	*0.18*
Braian Cufre	*0.13*	*0.18*	*0.17*
Mitja Ilenic	*0.47*	*0.14*	*0.31*
Richy Ledezma	*0.67*	*0.73*	*0.65*
Talles Magno	*0.26*	*0.37*	*0.54*
Keaton Parks	*0.13*	*0.22*	*0.43*
Gabriel Pereira	*0.68*	*0.30*	*0.62*
Santiago Rodriguez	*0.36*	*0.40*	*0.37*
Thiago Andrade	*0.67*	*0.27*	*0.59*

Figure 4 – Outfield Players with at least 1 G+A with their actual G+A/90, xG+A/90, and lastly their model predicted G+A/90.

We can see that our model's predictions compare admirably to both the players' actual goal contributions and the industry standard xG+A. In some cases, such as that of Thiago Andrade, Gabriel Pereira, and Mitja Ilenic, our model's predictions provide a much better estimate than the players' respective xG+A value. In another case, that of Talles Magno, we can see that our model's prediction is inflated relative to his actual G+A. Additionally, Talles Magno's xG+A also trumps his actual on field production, which seems to indicate that the underlying data, of both our model and that associated with xG+A, believes that Talles Magno should be scoring or assisting more often than he actually is. This could be for a variety of reasons, not just limited to the model. Players can simply be in poor form, not performing as well as they should, missing simple chances to score goals, or their teammates doing the same, effecting the possibility for assists. Also, the sample size is relatively small in this scenario, with the league campaign only about a quarter of the way through. By the end of the season it would be expected to see most players either progress or regress to the expectations of our model, as well as those of xG+A.

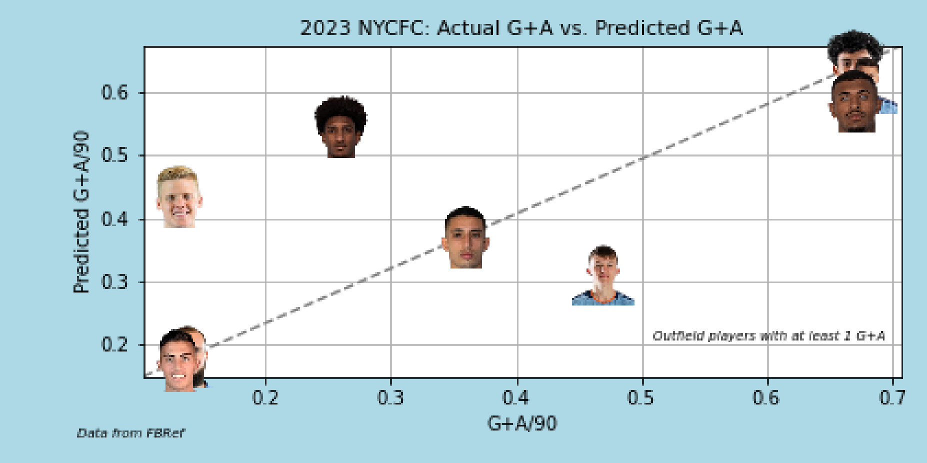


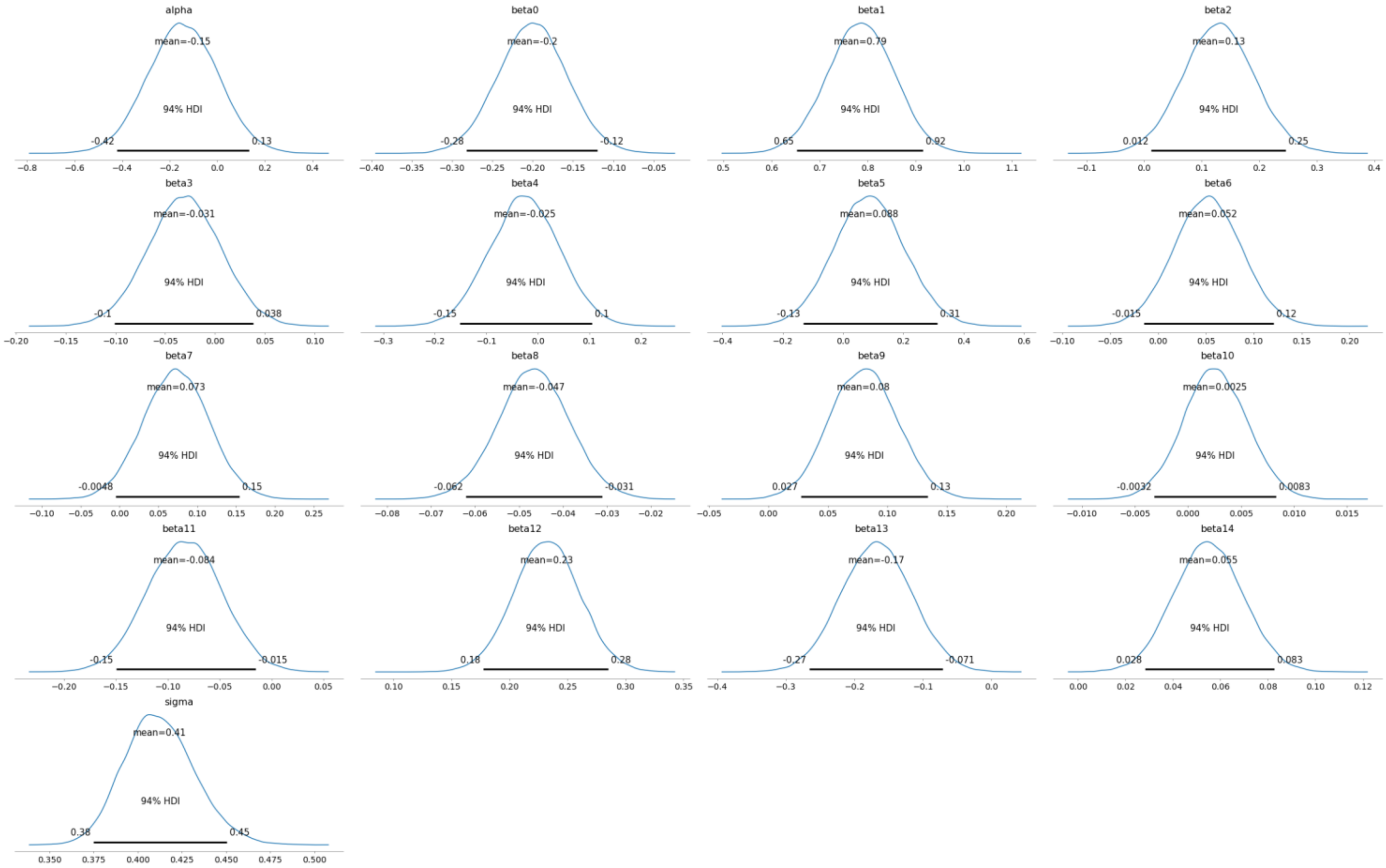
Figure 5 – NYCFC outfield players with the model predictions in respect to actual G+A.

Figure 5, above, demonstrates many of the thoughts echoed above. Many of the NYCFC players are well modeled, with predicted G+A values well in line with their actual on field performance.

References

[1] https://fbref.com/en/comps/22/stats/Major-League-Soccer-Stats

Appendix



Appendix 1 – Posterior densities of model parameters