Modelling Soccer Outcomes

Duncan Ofosu

12/13/2021

Abstract

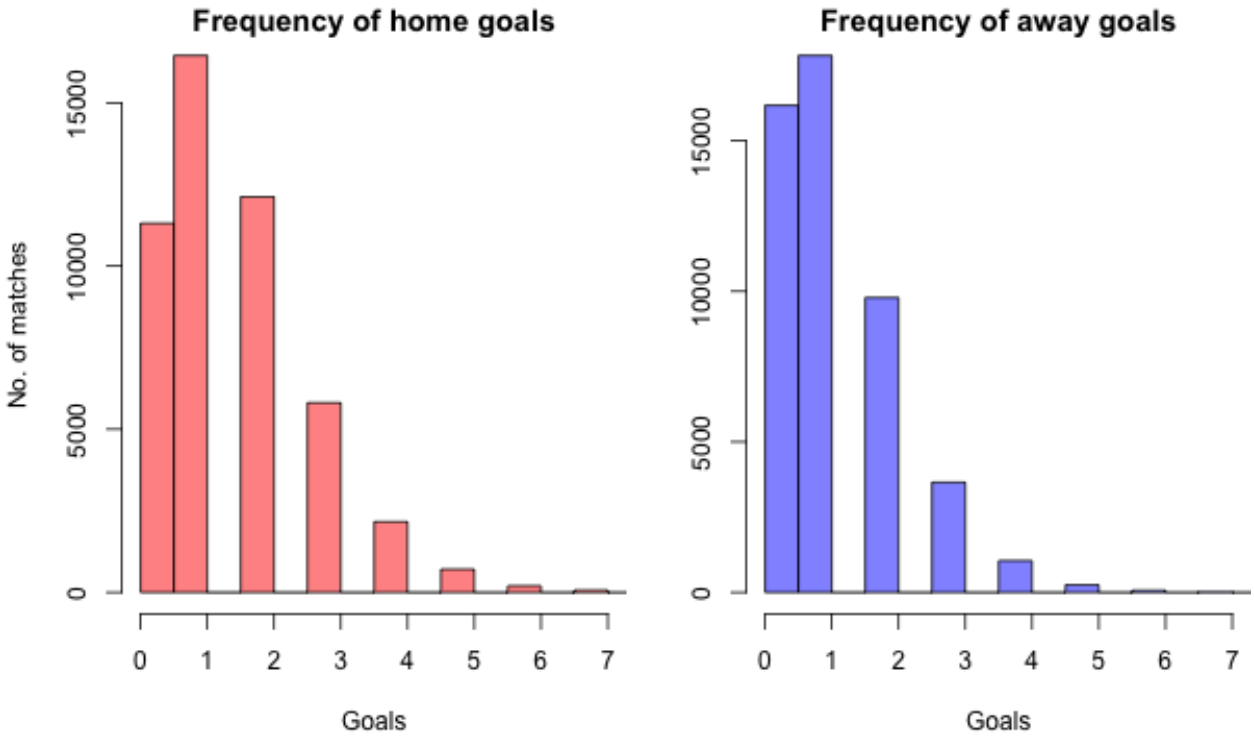
This article looks at the a basic approach to modelling the outcome of soccer matches in the English Premier League using the Poisson regression. The offensive and defensive abilities of the teams are extracted based on their rate of scoring or conceding goals. Probabilities for the outcome on each soccer match is then assigned.

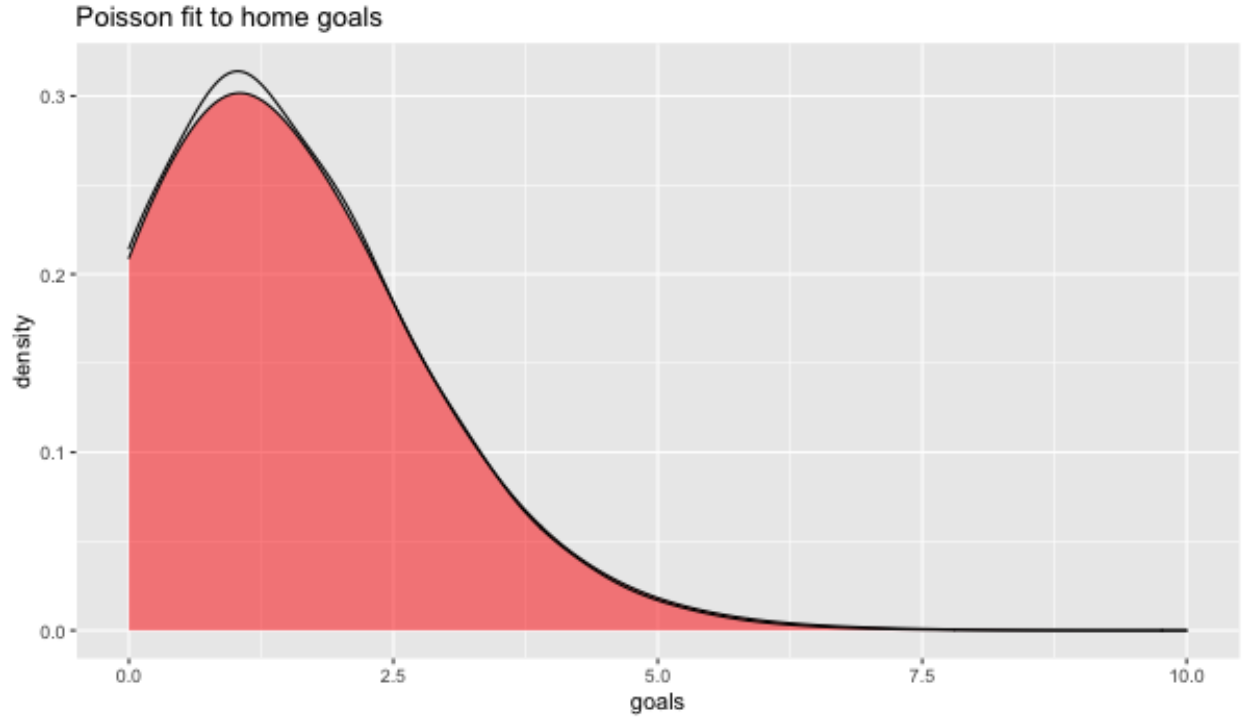
Introduction

With the ease to information and the increased global audience's willingness to consume sporting activities, there has been a surge in betting across most sporting disciplines. Th English Premier league is one of the most popular sporting activities in the world and many consumers of the league are increasingly wanting to bet on the outcome of matches. It leads to the simple question. Can we use Mathematics to model these outcomes and what are the chances of success?

Problem Formulation

Soccer, unlike other sports is a low scoring sport. Similarly, there are so many random acts that could influence the outcome of a match. A poor refereeing decision, an injury to a key player, dressing room issues or a lucky deflection on a shot. Despite all these randomness and unpredictability, the key principles usually remain the same. A good team probably scores more and concedes less. Thus, the ability to assess the offensive and defensive ability of a team will be good baseline approach to solving this problem. To win a football match, you need to score more goals than your opponent. The underlying assumption is goals can be scored at any time and is independent of time. Goals are also discrete events and thus a discrete probability distribution will be a good baseline approach[1]. A plot of goals scored in the Premier league over 12 year league seasons is shown in Figure 1. As many have pointed in numerous literatures [2], the Poisson distribution is a good baseline approach. Figure 2 shows a statistical fit to home goals scored on a given Poisson distribution rate (lambda). It fits home goals scored reasonably well and similar results would be attained for away goals scored. Now that we have a baseline Probability distribution to work with, we need to determine the appropriate covariates needed to build this model. There are two main soccer modelling decisions: predicting the probability of the outcome of a game (home/away/draw) or predicting the exact scoreline and subsequently forecasting an entire season of soccer matches. For this work, we will predict the probabilities of the outcome of a match.





Modelling

To predict the probabilities of an outcome on a game, you need to know the probability of a team scoring a certain amount of goals in a game. For a team to score a goal, it depends mainly on the offensive abilities of the team and the defensive abilities of the opponent. Given teams playing at home win more matches in comparison to their away games, it's reasonable to assign a home advantage weighting as well [3]. Thus, if Team X plays against Team Y at home, we can the number of goals Team X cab score and similarly the

number of goals Team Y can also score. We establish the following relationship:

$$P(X = x) = \frac{\lambda_1^x e^{-x}}{x!} \tag{1}$$

$$P(Y = y) = \frac{\lambda_2^y e^{-y}}{y!} \tag{2}$$

where :

$P(X = x)$ is the probability that Team X will score x number of goals against Team Y and $P(Y = y)$ is the probability that Team Y will score y number of goals against Team X

To find the parameters λ_1 and λ_2 We run a poisson regression to generate the parameters for each team and each game :

$$log(\lambda_1) = (attackX) + (defenceY) + (homeadvantage)$$
$$log(\lambda_2) = (attackY) + (defenceX)$$

Results

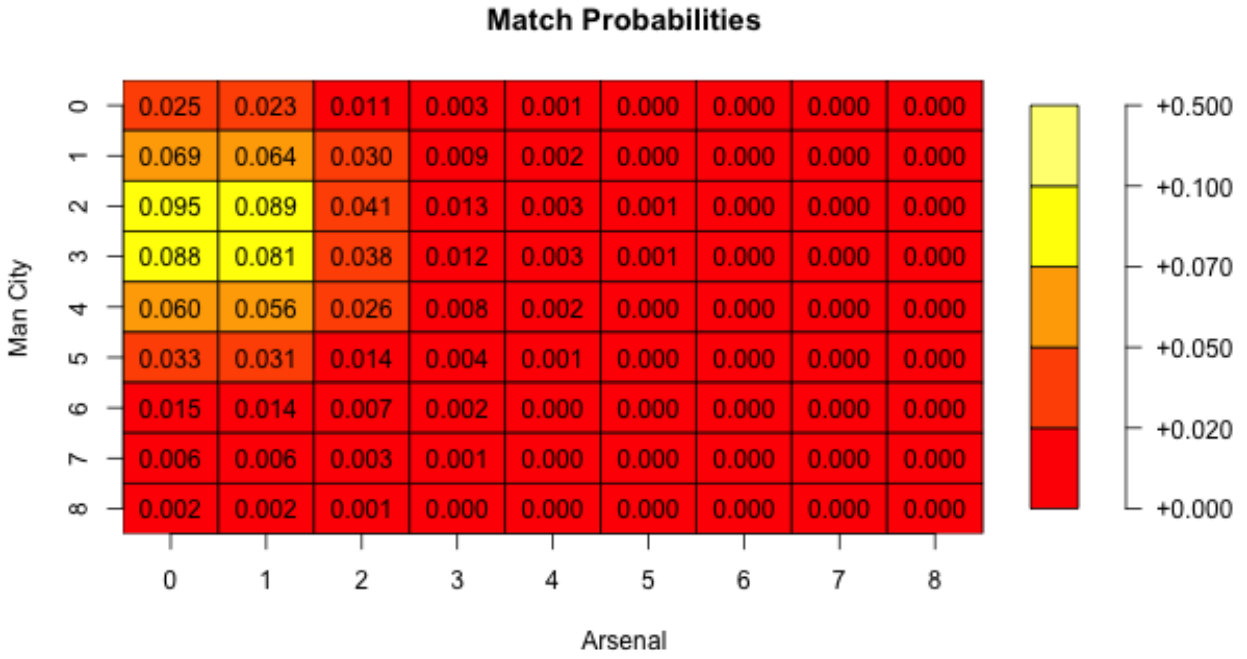
Once we have found the parameters and using the assumption that goals scored are independent, we can find the joint probability distribution as follows :

$$P(X = x, Y = y) = P(X = x)P(Y = y) \tag{3}$$

Thus, for a given match, to find the probability that Team X scores 2 goals and Team Y scoers 1 goal is as follows:

$$P(X = 2, Y = 1) = P(X = 2)P(Y = 1) \tag{4}$$

For our case study from the data, we will use the Man City vs Arsenal game. The diagonal entries in the matrix shows the probabilities that the game ended in a draw, the upper triangular matrix shos the entries where the away team won, whiles the lower triangular matrix shows the entries where the home team won



Conclusion

This is a very good baseline model. The obvious question to ask? Can we start betting and actually win monies from this? Unfortunately, NO! Well, there is a possibility but not that big. For starters, the model runs under the assumption that the rate of goals scored (λ) is invariant of time. That is not entirely accurate and others have extended that framework to using bivariate Poisson distribution[4]. Again, what will be an appropriate time weighting on matches? Matches played a long time shouldn't carry as much weighting or importance as matches recently played. Other approaches to modelling prediction of soccer has been explored[5]

References

1. Groll A, Schauberger G, Tutz G. 2015 Prediction of major international soccer tournaments based on team-specific regularized poisson regression: An application to the FIFA world cup 2014. *Journal of Quantitative Analysis in Sports* **11**, 97–115.

2. Dixon MJ, Coles SG. 1997 Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **46**, 265–280.

3. Forrest D, Goddard J, Simmons R. 2005 Odds-setters as forecasters: The case of english football. *International journal of forecasting* **21**, 551–564.

4. McHale I, Scarf P. 2007 Modelling soccer matches using bivariate discrete distributions with general dependence structure. *Statistica Neerlandica* **61**, 432–445.

5. Owramipur F, Eskandarian P, Mozneb FS. 2013 Football result prediction with bayesian network in spanish league-barcelona team. *International Journal of Computer Theory and Engineering* **5**, 812.