ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

Measuring scoring efficiency through goal expectancy estimation

H. Ruiz[1,2], P. J. Lisboa[1], P. J. Neilson[2] and W. Gregson[3]

1- School of Computing and Mathematical Sciences – Department of Mathematics and Statistics – LJMU, Liverpool L3 3AF – UK
2- Performance Lab – Prozone Sports Ltd., Leeds LS11 8BN – UK
3- The Football Exchange – LJMU, Liverpool L3 3AF – UK

Abstract. Association football is characterized by the lowest scoring rate of all major sports. A typical value of less than 3 goals per game makes it difficult to find strong effects on goal scoring. Instead of goals, one can focus on the production of shots, increasing the available sample size. However, the value of shots depends heavily on different factors, and it is important to take this variability into account. In this paper, we use a multilayer perceptron to build a goal expectancy model that estimates the conversion probability of shots, and use it to evaluate the scoring performance of Premier League footballers.

1 Introduction

The open and dynamic nature of football is one of the key features that define the sport. The constant fluidity in the movement of the ball and the players makes it a fascinating sport to watch and analyse, but at the same time makes it challenging to model and evaluate individual and collective behaviours. Other sports like baseball have a clearly structured order of play and much weaker dependencies between the actions of different players, which helps analysts break down games and isolate performances.

Accepting the more intricate dynamics of football, a separate complication is the low prevalence of goals: average shot conversion rates typically lie around 10% in professional competitions, with less than 3 goals per game as the norm. Game outcomes therefore depend on very scarce events compared to the number of other player actions, and finding connections between the two requires sampling a large amount of different games. A possiblity to increase the number of samples when analysing offensive output is to take a step back from goals and consider shots as the target event. Since all goals (except for own goals) come from shots, it makes sense to to look at shot production to assess offensive success. However, using just the number of shots to measure offensive output can be misleading, because the threat that a shot poses to the opposition varies depending on factors like location, shot type, power, goalkeeper position, etc. It is important to weight shots according to how these factors affect their conversion probability so that those with a higher chance of producing a goal are assigned a larger offensive impact.

The concept of goal expectancy [1] was introduced with this principle in mind. By modelling the observed goal conversion of a large sample of these events, a smooth probability function can be derived that approximates the underlying expected number of times that a shot with certain characteristics will lead to a goal. This is interesting from two different perspectives: Firstly, analysing the resulting model may provide insightful information about the relative efficiency of different shooting strategies. Secondly, the production of expected goals may be used as a supplement to the observed number of goals to more accurately reflect offensive throughput, giving rise to a different perspective that alleviates the inherent variability in the shot to goal conversion process.

In this work, a multilayer perceptron is used to model the success probability of shots using real-world data. Then, a visual interpretation of the model is provided using pitch map representations of the resulting probability estimates. Finally, an example application is provided where the shooting efficiency of a selection of professional football players is measured using their expected goal distribution.

2 Methodology

2.1 Goal expectancy model

The principle behind goal expectancy models was introduced to the field of football analysis through the study of goal conversion rates in different regions of the pitch [2,3]. This approach divides the field into a number of areas and calculates the goal expectancy for each of them as the average success rate of observed shots originating from it. A more uniform model can be obtained using the shot data to fit a probability estimator, providing a continuous output with respect to the location of the strike [4].

Following this idea, we model the probability of a shot resulting in a goal given some contextual information as follows:

$$p(g|shot) = p(g|\mathbf{x}) = \phi\big(a(\mathbf{x})\big) = \frac{e^{-a(\mathbf{x})}}{1 + e^{-a(\mathbf{x})}}$$

The input map $a(\mathbf{x})$ is given by a multilayer perceptron (MLP) with sigmoid activation functions,

$$a(\mathbf{x}) = \mathbf{W}^{O}\phi(\mathbf{W}^{H}\mathbf{x} + \mathbf{B}^{H}) + \mathbf{B}^{O}$$

where $\mathbf{x}$ is the input vector, $\mathbf{W}^{H}$ is the hidden layer weight matrix, $\mathbf{B}^{H}$ is the hidden layer bias weight vector, $\mathbf{W}^{O}$ is the output layer weight vector and $\mathbf{B}^{O}$ is the output layer bias weight. The input $\mathbf{x}$ is a vector of contextual variables used to characterise shots, and contains six elements: two continuous variables that represent the pitch coordinates where the shot is taken from, and four binary variables that code the type of shot (open play footed shot, header, free kick shot and penalty shot).

2.2 Distribution of the number of goals

To model the expected number of goals scored from a given amount of shots, we consider all shots as a sequence of independent Bernoulli trials with success probability $p(g|\mathbf{x}_i)$, where $\mathbf{x}_i$ represents the contextual information of the i-th shot. Since these probabilities will, in general, be different from one shot to another, the trials are non-identically distributed and the number of successes follows a Poisson binomial distribution.

Existing work used Poisson [5,6] and negative binomial [7] distributions to model the number of goals scored by teams assuming independent distribution means calculated from each team's previous performances. We could not find any precedent in the scientific literature, however, for the use of individual shot expectancy estimates to model the distribution of goals. Only in the non-academic domain we found this approach: in [4,8] the authors use a Monte Carlo simulation to estimate the values of the compound probability mass function.

We propose a more accurate and computationally efficient calculation using polynomial multiplication to generate all possible goal/miss combinations, weighted by their corresponding probability. If we use p_i to denote the scoring probability of shot i, we can express the probability of the success and failure of a shot as the coefficients of a first order polynomial, $(1 - p_i) + p_i a$. To obtain the probability of a number of goals in a sequence of independent shots, we multiply all the corresponding polynomials:

$$\sum_{k=0}^{n} p(X = k)\, a^k = \prod_{i=1}^{n} \big((1 - p_i) + p_i a\big)$$

where *X* represents the number of goals scored in *n* attempts. Calculating the complete pmf requires n multiplications by a degree 1 polynomial, which takes $O(n^2)$ computation time using standard algorithms.

3 Results

The dataset used in this study contains all the 10318 shots taken during the 2013/14 English Premier League, extracted from Prozone Matchviewer event data [9]. The shots are divided into 8087 open play footed shots, 1678 headers, 466 free kicks and 87 penalty shots. The training stage of the probability estimator was carried out using half of the dataset for training and the other half for validation, with samples selected randomly from each of the four subgroups to ensure a balanced representation of all shot types. The MLP used had 6 hidden nodes, and the value of its weights was calculated using the backpropagation of the log-likelihood errors of the outputs.

3.1 Visual interpretation of the model

Given the input space of the study, an intuitive way to represent the resulting model is to generate a heatmap of the estimated shot probabilities for a grid of plausible shot locations. Figure 1 shows the result, highlighting the smoothing effect of the probability function for the first three shot types (penalties have been omitted here, as they are restricted to a unique pitch location).

Unsurprisingly, the plots capture the decreasing magnitude of the conversion expectancy of free kicks, open play shots and headers, for a given position. An interesting non-linear characteristic of the model is its asymmetry with respect to the horizontal axis (the axis perpendicular to the goal line). This indicates a slight increase in the success rate of shots coming from the right (from the goalkeeper's point of view), possibly due to a higher prevalence of right footed players, for whom those areas provide a more natural shooting position.

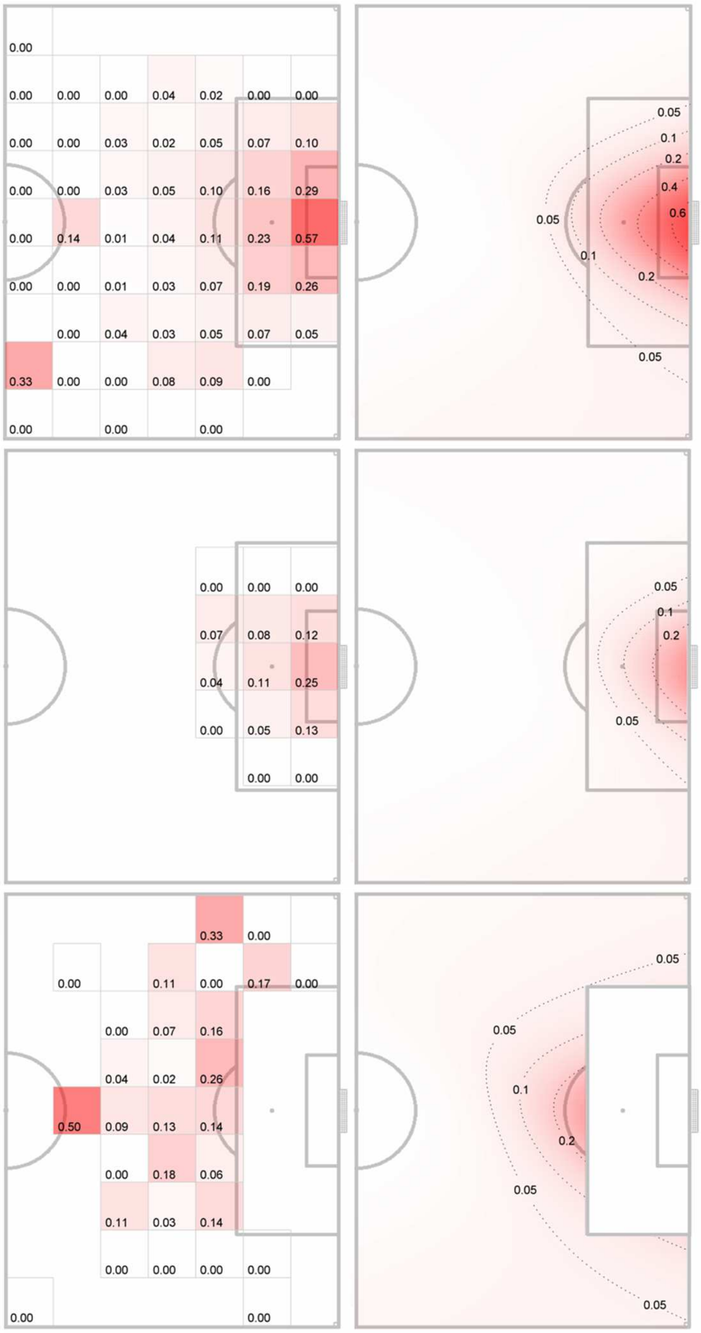

Fig. 1: Observed (left) and estimated (right) shot conversion rates for open-play footed shots (top), headers (middle) and free kicks (bottom). Dotted contour lines connect areas of the pitch with a goal expectancy equal to the corresponding label.

3.2 Measuring scoring efficiency

Working with the expected value of shots adds valuable information to the analysis of their observed outcomes. For instance, we can assess the scoring performance of players with respect to the aggregated expected value of the chances they had [8]. Figure 2 shows the probability distribution of the number of goals expected to be scored by Sergio Agüero given the shoots that he took during the 2013/14 season.

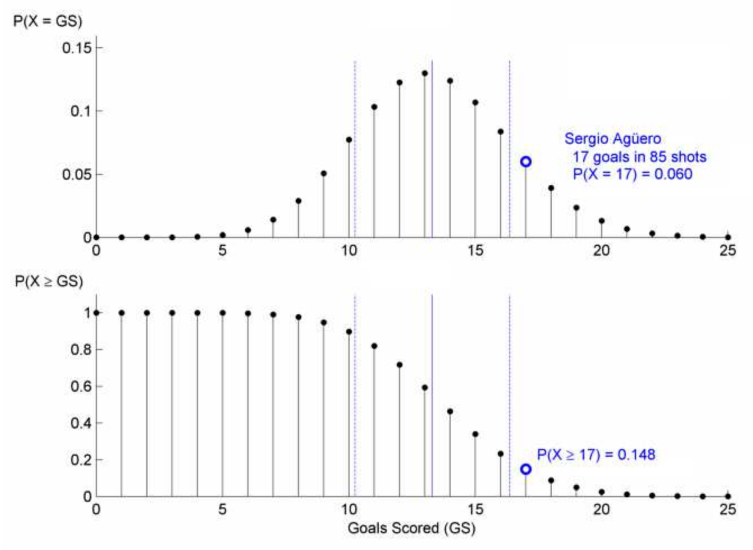

Fig. 2: Probability mass function (top) and cumulative distribution (bottom) of the number of goals expected from Agüero's 2013/14 shots, calculated using the Poisson binomial distribution discussed in Section 2.2. The vertical lines indicate the distribution mean, 13.3, and ±1 standard deviations from it ($\sigma = 3.1$).

It is therefore possible to express the actual number of goals scored by Agüero as a deviation from the expected value, in this case of +1.2 standard deviations. Presented with the same chances, an average player in that competition would be expected to match or outperform Agüero 14.8% of the time.

The results in Table 1 compare the top 10 scorers of the competition in terms of observed and expected shooting efficiency. The first three players displayed a proficient scoring skill, with their p-values indicating that their chance conversion accuracy was significantly higher than the average estimates. It is interesting to see how Yaya Touré's 67 shots gave rise to 12.6 expected goals, while Sturridge's 10.6 came from 93 attempts. This is due, at least partly, to Touré having taken (and converted) 6 penalty shots, which exemplifies the importance of looking beyond simple counts of shots and goals.

4 Conclusions

Modelling the expected goal value of shots gives an additional level of detail to the analysis of offensive and defensive performance in football. This paper proposes a goal expectancy model based on a neural network estimator followed by an efficient calculation of the expected goal distribution given a sample of shots and their characteristics. The model provides insightful outcomes using very simple contextual information. Further work could easily improve the conversion estimates by using more detailed information such as shot power and placement, and location of the players between the shooter and the goal.

Rank	Player	Goals scored	Shots taken	Expected goals	Standard deviation	$p(X \geq GS)$
1	Luis Suárez	31	184	16.9	3.7	≈ 0
2	Daniel Sturridge	22*	93	10.6	2.8	≈ 0
3	Yaya Touré	20	67	12.6	2.5	0.005
4	Sergio Agüero	17	85	13.3	3.1	0.148
	Wayne Rooney	17	101	13.5	3.0	0.159
	Wilfried Bony	17*	106	15.3	3.1	0.346
7	Edin Džeko	16	103	13.6	3.3	0.269
	Olivier Giroud	16	113	16.7	3.5	0.629
9	Romelu Lukaku	15	102	11.2	3.0	0.132
	Jay Rodríguez	15	103	10.6	3.0	0.099

Table 1: 2013/14 Premier League top scorer table, with added goal expectancy information. The competition's goal panel deducted one goal from these figures, classing them as own goals. Given the borderline nature of these two cases, we decided to include them in the analysis as goals.

References

[1] C. Trainor and C. Chappas, Goal expectation and efficiency, 06 August 2013. Retrieved from http://www.statsbomb.com/2013/08/goal-expectation-and-efficiency/.

[2] P. Riley, Shot Position Average Model – SPAM, 29 December 2012. Retrieved from http://www.differentgame.wordpress.com/2012/12/29/shot-position-average-model-spam/.

[3] Anonymous, Where do the best shots come from?, 7 June 2013. Retrieved from http://11tegen11.net/2013/06/07/where-do-the-best-shots-come-from/.

[4] M. Taylor, Using an expected goal model to re-evaluate results, 18 March 2014. Retrieved from http://www.bettingexpert.com/blog/using-an-expected-goal-model-to-reevaluate-results/.

[5] M. J. Maher, Modelling association football scores, *Statistica Neerlandica*, 36:109:118, Wiley, 1982.

[6] M. J. Dixon and S. G. Coles, Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2):265:280, Wiley, 1997.

[7] C. Reep, R. Pollard and B. Benjamin, Skill and chance in ball games, *Journal of the Royal Statistical Society: Series A (General)*, 623-629, JSTOR, 1971.

[8] P. Riley, The Spanish Inquisition – Roberto Soldado, 23 March 2014. Retrieved from http://statsbomb.com/2014/03/the-spanish-inquisition-roberto-soldado/

[9] P. Bradley, P. O'Donoghue, B. Wooster and P. Tordoff, The reliability of ProZone MatchViewer: a video-based technical performance analysis system, *International Journal of Performance Analysis in Sport*, *7*(3):117-129, University of Wales Institute, 2007.