Analysing Finishing Skill in Football

Edward St John
City, University of London
Data Science MSc: Principles of Data Science
London, UK
edward.st-john@city.ac.uk

***Abstract*—The aim of this paper is to investigate whether a better finishing skill model could be made by splitting up shot data into long- and short-range shots, and predicting future performance based solely on the over/under performance of those subsets of shots. Seven seasons' worth of data is used underlying this analysis, with predictions performed on unseen 2021 data to provide an unbiased test. This study finds that short-range shots actually provide a better estimation of future performance when compared to a general model (predicting based on all past shots from a player) and a long-range model (predicting solely on long shot ability), rejecting our hypothesis of a long-range model providing a more accurate prediction of future goals scored based off Expected Goals.**

Keywords—xG, finishing, football, expected goals, performance, Understat

I. Introduction

The use of Expected Goals (xG – the probability of a particular shot resulting in a goal on a scale from 0 to 1, explained in more detail in Section II B opposite) in football analytics has increased dramatically over the past few years, whether it be in assessing a player's performance for football scouts or in professional football betting algorithms. However, one aspect which has always been controversial in the use of xG is how to determine finishing skill based on this in order to predict future goals.

Usually, finishing skill can be assessed by seeing how much a player outperforms their xG value (with a constant added into the formula for smoothing to account for small sample sizes – more on this in Section III), as a 1:1 ratio for goals and xG implies an average level of finishing skill. This paper explores whether the model for finishing skill can be improved by predicting future goals by how much players outperform xG on long shots, which are by definition much harder shots (with lower individual xG values, i.e. lower probability of resulting in a goal), which could provide a more reasonable estimation for finishing skill compared to overperformance on all shots. We also explore whether the opposite could be true, with short-range finishing skill being a better indicator of a player's finishing ability.

II. Analytical Questions and Data Materials

A. Analytical Questions

The overriding aim of this paper is to investigate if a better model for estimating finishing skill (and therefore estimating future goal returns) can be made by splitting up data to differentiate long and short shots, predicting future goal returns based on a player's over/under performance of those particular shots. Below are some of the questions this paper aims to answer:

1. Can a more predictive finishing skill model be made utilising long / shot shot performance as opposed to over/underperformance on all shots by a player?
2. Does the optimal smoothing constant change based on the model (general vs long shot vs short shot)?

Question 1 is particularly interesting to investigate as finishing skill often isn't taken into account with predicting future goals for a particular player, but more how much xG their team as a whole is expected to create and how much of a share of that xG the player has.

B. Data

The main (and essential) aspect of our data used is Expected Goals (xG). xG provides a description of the quality of the chances a player / team has had, roughly translating to the number of goals you would expect that particular player / team to score should you simulate the match thousands of times with the same xG values. The value of xG is determined by many factors [1], such as where the shot was taken, what body part was used to hit the shot, how many defenders were in front of the ball (and their positioning) and where the goalkeeper is. These are all determined from hundreds of thousands of previous shots, meaning we can estimate the probability of a shot with all its characteristics based on similar shots in the past and their conversion rate. Note that different xG models have different amounts of sophistication, so may not take into account all the variables listed above. The xG provider Understat is used in this paper due to their data being easily accessible and easy to extract, with most other providers needing a membership to obtain their statistics.

The data used in this paper has many features for us to utilise. In particular, the shot location variable (with x and y coordinates) allows the extraction of distance to goal (converting the axis from 0 – 1 to the actual distance of specific points on the pitch) which is essential for us to define shots as long range or short range, and therefore needed to answer our first research question defined in Part A. Another feature included in the data is the 'situation' variable, allowing us to remove penalties from the equation as we just want to analyse finishing skill regular shots rather than penalty-taking skill which may not reflect a player's true finishing ability. Both of these factors contribute to make this data very suitable for the questions this paper attempts to answer.

The main assumption in utilising this data is assuming that all the football pitches are the same size for the shot location conversion (to calculate distance from goal in order to classify a long shot) – this is not the case with each pitch being slightly different dimensions, but this paper uses the Premier League (PL) recommended pitch size (which also happens to be the average size for most of the pitches in this dataset) of 105 metres by 68 metres [2].

III. Analysis

A. Data Preparation

1) Extraction of Data: To obtain data from Understat a script was used from GitHub to scrape the required data, having adjusted the code to only include the Premier League (the original code scraped data for Europe's top five Leagues). From the scraper, the raw aquired data was in the form of separate csv's for each team per season – this was then aggregated into one large csv containing all shot data for every player & team for the past seven seasons.

2) Data Analysis & Cleaning: A quick look at the structure of our data shows us we have over 66k rows (i.e. 66k shots to analyse) with 18 different features. After importing the data as a csv (rather than keeping the dataframe loaded from the scraping) all columns were interpreted correctly, for example the season column was recognised as an integer, etc.

3) Identifying Missing Values and Outliers: All columns were checked for missing values, with only the assist variable (listing who assisted the goal) having missing values which is reasonable as not all goal attempts have someone who assisted it. In terms of zero-values, looking at our summary statistics for each column xG is the only column which is unreasonable for there to be a minimum of zero as that should be impossible (i.e. no shot should have a zero probability of going in the goal). On further investigation, it was found that these were all attributed to own goals which makes sense as they don't count towards someone's goal tally (otherwise a clumsy defender could end up as top goalscorer for a club!) so an xG value of 0 is assigned. We removed all of these values to just have a dataset containing actual shots from players. The summary statistics also show there are no xG values above 1 or below 0 (both of which would be impossible as that would imply a probability > 100% or below 0%), also confirmed by the strip plot below in Figure 1 of all seven seasons of data (also showing roughly the same distribution of shots each year with a lower amount of high probability shots).

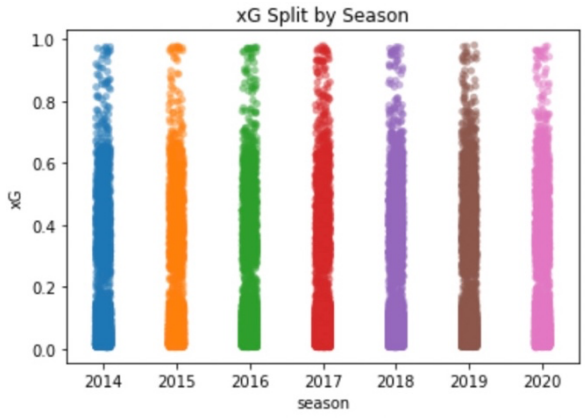

Figure 1

B. Data Derivation

1) Extracting Important Variables: Due to our analysis just focusing on finishing skill, there are a few variables we can remove to leave us with a more digestable dataset, being the first index column (the dataset came with its own indexes but we already have indexes set in the dataframe), the Understat ID (just used to identify the particular event) and the date (as we already have the season the shot takes place in and the respective game).

2) Adding Long Shot Variable: In order to define what is a long shot and what isn't, the distance to goal had to be measured. Python originally interpreted the position colmn as an object, so the *x* and *y* value coordinates had to be split up in order for them to be seen as separate values and converted to numerical features. These coordinates were based on how far the shot was away from 0,0 being the bottom left of a pitch (seen below), so they had to be converted to how far away the shot is to the opponent goal (located at 1,0.5) by calculating the Hypotenuse [3] to obtain the actual distance to goal from the coordinates. The long shot variable could then be created by defining a long shot as anything further than the penalty arc which is 22 yards away from the goal. This was calculated as we knew the coordinates for a penalty (12 yards from goal), so a scale factor could then be applied to measure the penalty arc distance. A goal variable showing if the result of the shot was a goal or not was also added as this was needed for the contruction of the models. Figure 2 provides a rough visualisation of the long shot definition to help understand where it is defined, however the figure is not to exact specifications due to limitations with support for our chosen xG provider Understat with translating coordinates to the pitch from Python library 'mplsoccer'.

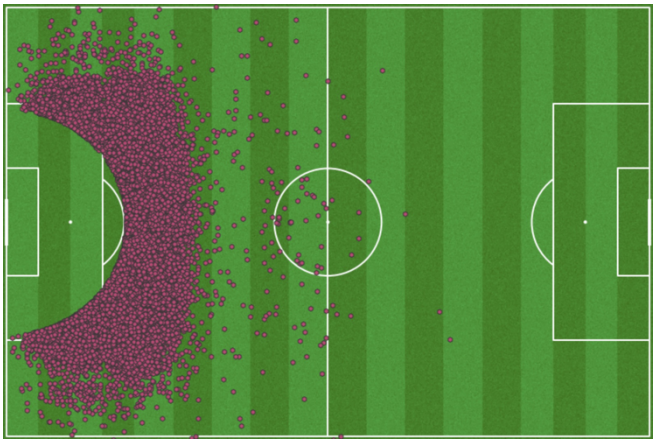

Figure 2

C. Construction of Models

1) General Model

A general model for determining finishing skill would be the following (where G = Goals and xG = Expected Goals):

$$\frac{G + c}{xG + c} \quad (1)$$

In the above, a constant is used to alleviate the discrepancies in using small data samples where variance has too much of an effect, e.g. if a player scored 3 goals from a total of 0.3 xG, you can't possibly expect them to carry on this goalscoring form and elite finishing ability for too long before they start regressing towards the mean, with an overperformance (likely) increasing confidence in the player, leading them to take more shots and therefore bring down their G:xG ratio. This constant used therefore can't be too small (as then we just have the same problem as before, with a small sample with high variance predicting another sample leading to a high error) or too large as all players would then show an average finishing ability, as the higher the constant the more the fraction in (1) above tends towards 1. An optimal value of c minimising the Mean Square Error (MSE) [4] for three different models were obtained:

1. General Model containing all shots

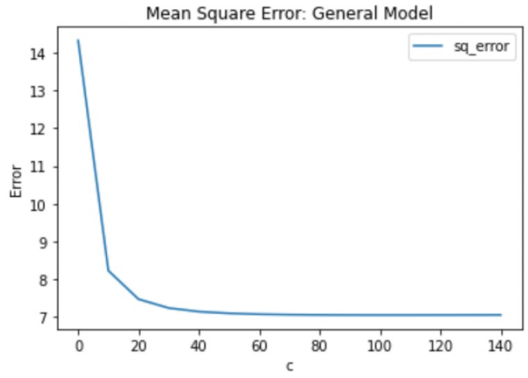

Figure 3

1. Long shot model only using G and xG for long shots (further out than the penalty arc, i.e. 22 yards from goal)

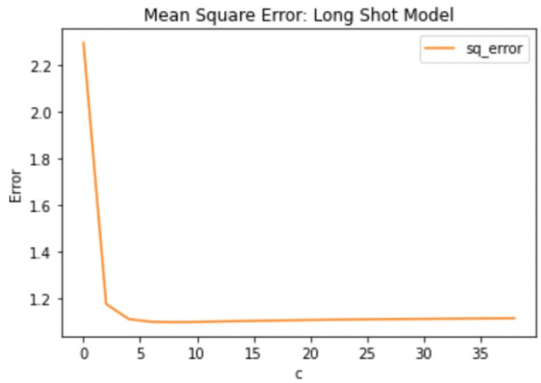

Figure 4

3. Short shot model only using G and xG for shots within 22 yards

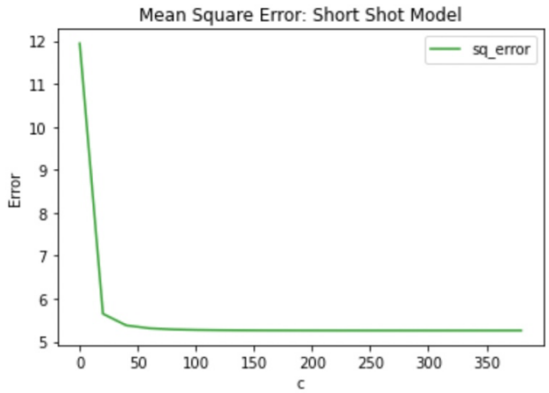

Figure 5

D. *Validation of Results*

After obtaining all the optimal values of c (the constant) for all three models based on shot data from 2014 to 2020, they were then tested on unseen 2021 data to see which model performed best with the lowest mean squared error of predictions for all players. This allows a fair validation for the models using data that was not utilised in the creation of the model, as this would have altered the smoothing constant and therefore potentially distorted our results.

IV. Findings, Reflections and Future Work

A. *Findings & Reflections*

1) Findings: Rather surprisingly, our analysis has found that out of the three models tested (general, long shot and short shot) the long shot model performs the worst with the highest MSE out of the three. The short shot model is the surprise victor here, providing the lowest MSE and therefore what seems to be the best performer when it comes to trying to predict future goalscoring rate based off Expected Goals.

These findings could be utilised within the football analytics domain, whether used for adjusting xG figures to help describe how a game panned out for game reviews, for scouting potential new signings by utilising short range finishing ability to predict future returns if signed for their club, or potentially in betting systems to adjust the likelihood of a player scoring a goal with the implied probability increased if they are an above average finisher, or decreased if they are below average.

2) Reflections: Our data ended up being very suitable for answering our analytical questions (especially being able to extract position data to define longshots), and while though xG as a a statistic isn't absolutely perfect for predicting future goals with off the ball actions not taken into account [5], it seems it is the best indicator we have at our disposal compared to other statistics. It was surprising to see the short shot model perform best in our testing, however looking at the bigger picture it does make some sense given that the majority of players' shots are likely to be short range, and therefore should have the most weight when assessing finishing ability.

3) Limitations: In using our data spanning across seven different seasons, there were a number of assumptions we had to make regarding the homogeneity of the data throughout the years. Firstly, we are assuming no position changes for players which would impact their xG creation, as this could skew the data towards a certain year or streak of over/under performance of xG and therefore affect the smoothing constant for the models providing a slightly less reflective prediction than originally thought. In addition to this, we are assuming there are no changes in set-piece duties, e.g. free kicks as that would more relate to a player's ability to take free kicks rather than their usual shot finishing ability (which is what this paper tries to assess). There is also the possibility of a small misclassification of long shots due to different pitch sizes in the Premier League (if you remember from above we took the average / recommended size), meaning something that we classify as a long shot might actually be less than our defined distance in real terms due to the pitch size difference.

B. Future Work

Further work could be done in order to improve the model and add some complexity to account for different variables faced. One example would be adding penalty-taking skill into the model, allowing us to include penalty goals (and xG) into the equation for a more rounded assessment on a player's finishing ability. Goalkeeper (GK) ability could also be accounted for in future improvement to the model by adjusting xG based on the ability of the GK to save goals compared to the average keeper, i.e. Expected Goals Conceded (xGC) over / under performance. This would hopefully create a more accurate model as with smaller sample sizes the mean GK quality may not be close to the average so this would account for that.

Word Count	
Abstract	123
Introduction	223
Analytical Questions & Data Materials	563
Analysis	1017
Findings, Reflections & Future Work	581
TOTAL	**2,507**

V. BIBLIOGRAPHY

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

[2] The Football Association Premier League Limited, 2021. *Premier League Handbook Season 2021/22*. In: s.l.:s.n., p. 159.

[3] Sierpinski, W., 2003. *Pythagorean traingles.* 9 ed. s.l.:Courier Corporation.

[4] Schluchter, M.D., 2005. Mean square error. *Encyclopedia of Biostatistics*, *5*.

[5] Spearman, W., 2018, February. Beyond expected goals. In *Proceedings of the 12th MIT sloan sports analytics conference* (pp. 1-17).