# Expected Goals in Football

David Dyer - 1786617

## 1 Abstract

In this project, we will investigate how statistics can be used in sport. We will do this by looking at the concept of expected goals in football, explaining what it is, developing a couple of simple models and seeing the uses of more elaborate models.

## 2 Introduction

Often in football, we hear comments such as "he should have scored there" and "it looked harder to miss" said by pundits, commentators, and fans, but are these statements true? Expected goals ($xG$) is a statistical model that measures the quality of shots taken and gives the likelihood of a given shot resulting in a goal. The total expected goals for a team in a game is the sum of each team's shot likelihoods. Football fans often base their views of which team should win depending on the quality of scoring chances that they create; expected goals is a way of quantifying these scoring chances.


Figure 1: Google Trends of expected goals

Figure 1 is a chart showing Google search interest relative to the highest point on the chart for the term 'expected goals' in the UK since 2009 with 100 being the peak popularity (Google Trends, 2020). The term was barely searched before August 2017 when it peaked and then remained at a much higher volume of searches than before the maximum. The sudden rise is because of higher exposure to the football community since the start of the 2017/18 season.

When this season began in August 2017, Match of the Day and Match of the Day 2 (BBC TV programmes showing highlights from all Premier League games on a Saturday and Sunday respectively) started showing each team's expected goals figure from each game. These two programmes

have a combined audience of more than seven million each weekend (BBC Sport, 2018). The expected goals figure being used is from Opta's (a British sports analytics company) expected goals model. There are many expected goals models out there from a wide range of sports websites, each including different variables and based on a different data set of shots. Opta has analysed over 300,000 shots to come up with the likelihood of different shots resulting in a goal.

# 3 Mathematics Behind Expected Goals

To be able to able to produce an expected goals ($xG$) model, some statistical knowledge is needed. First, it is necessary to know what the mathematical understanding of the word 'expected' is.

**Definition 1.** *The **expected value** of a random variable is the average value of several independent realisations of the random variable.*

In the case of expected goals, the random variable is a given shot. The next step is to look at the possible outcomes and the probability distribution of these outcomes.

**Definition 2.** *A **probability distribution** is a mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment.*

In football, either a given shot can go in the goal giving the team an additional score of 1, or it can not go in the goal which would not affect the scoreline. Because of this binary nature, we treat a shot as a Bernoulli random variable.

**Definition 3.** *A **Bernoulli random variable** is the discrete probability distribution of a random variable which takes the value 1 with probability $p$ and the value 0 with probability $q = 1 - p$.*

For a Bernoulli random variable (x) the expectation is

$$E[x] = 1 \times p + 0 \times (1 - p) = p. \tag{1}$$

## 3.1 Simplest Expected Goals Model

The most basic way to determine the probability of a shot resulting in a goal is by looking at shots to goals ratios. For example, looking at the Premier League (the highest division in England) over the past eight seasons, the number of shots and goals per season (Footstats, 2020) can be obtained.

From Table 1, we can obtain the mean number of goals per shot, which is 0.1072. This number means a shot results in a goal on average 10.72% of the time. A goal is worth one score wherever it is scored on the pitch and is the only way to add to your score. We treat a shot as a Bernoulli random variable, and we are assuming that every shot has the same chance of resulting in a goal. Hence, the expectation of a shot resulting in a goal is just the probability of a goal.

Table 1: Shots and goals per Premier League season

| **Season** | **18/19** | **17/18** | **16/17** | **15/16** | **14/15** | **13/14** | **12/13** | **11/12** |
|---|---|---|---|---|---|---|---|---|
| **Goals** | 1072 | 1018 | 1064 | 1026 | 975 | 1052 | 1063 | 1066 |
| **Shots** | 9608 | 9287 | 9689 | 9759 | 9846 | 10214 | 9552 | 9868 |
| **Goals to Shots (4 d.p)** | 0.1116 | 0.1096 | 0.1098 | 0.1051 | 0.0990 | 0.1030 | 0.1113 | 0.1080 |

As the expectation of a goal is 0.1072, the following formula for the total expected goals based on the total number of shots (S) for a team is

$$xG = 0.1072 \times S. \tag{2}$$

The obvious main drawback to this model is that it assumes that every shot has the same likelihood of ending up in a goal which is not the case. The likelihood of a shot resulting in a goal depends on several factors. Factors include distance and angle to goal, and whether the shot was a header or kicked. As the model does not consider any of these factors, it is an inferior basic model as the whole concept of expected goals is to measure the quality of shots taken.

## 3.2 Advancements of Expected Goals Model

To advance the models of expected goals, we will not assume that every shot has the same probability of resulting in a goal as this is not realistic. One of the main factors that affect if a shot results in a goal is the perpendicular distance between the goal line and the shot location. To be able to improve upon our simple model (equation 2), we will take account of the perpendicular distance between the shot and goal by developing a regression model.

**Definition 4.** ***Regression*** *is a technique for determining the statistical relationship between two or more variables where a change in a dependent variable is associated with, and depends on, a change in one or more independent variables.*

To be able to develop a regression model, we need to gather data of shots. The data needed is the perpendicular distance between the shot and the goal line and if it was a goal. As we are treating a shot as a Bernoulli random variable (it ether results in a goal or not) this means our dependent variable is qualitative. Due to the outcome being of a qualitative nature, this means we are modelling a classification problem as we will have the two classes: goal and no goal. The primary technique for modelling a classification problem is logistic regression (James et al., 2017).

**Definition 5.** ***Logistic regression*** *is a statistical method for analysing a data set in which there are one or more independent variables that determine an outcome. A dichotomous variable (in which there are only two possible outcomes) measures the outcome. It predicts a binary outcome.*

Choosing at random Premier League gameweek 13 ($23^{rd}$ November to $25^{th}$ November) as ten fixtures to collect our shot data. Using WhoScored's chalkboard feature (Who Scored, 2019) for each game, we can view the location of each shot taken.

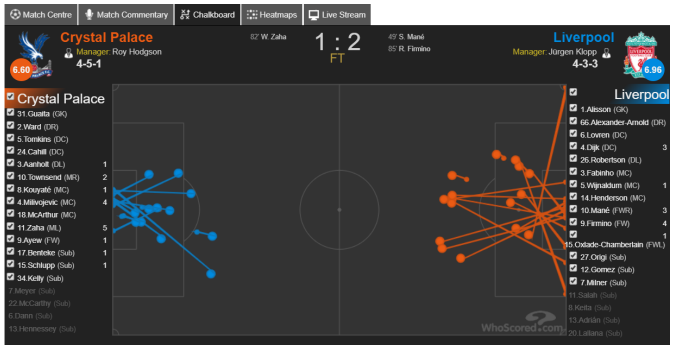


Figure 2: WhoScored's chalkboard for Crystal Palace vs Liverpool

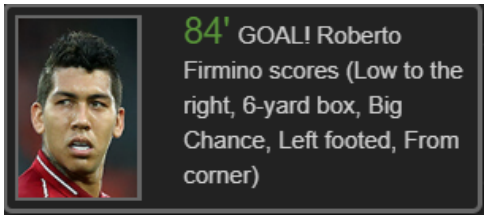


Figure 3: Details of one of Liverpool's shots

The dots on the pitch display the location of the shots and lines represent the path of the ball.

As Figure 3 shows us, clicking on one of the dots shows us more details about each shot, including if the shot resulted in a goal or not. In total there were 238 shots across the gameweek, for logistic regression to work we will give a shot a rating of 1 if it results in a goal and 0 if it does not.

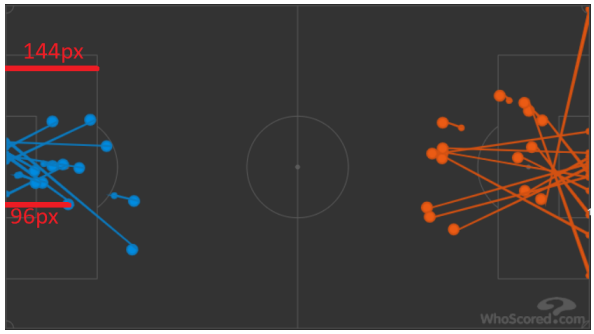


Figure 4: How shot distance is calculated

Although WhoScored's chalkboard shows us the location on the pitch that the player shoots from, it does not tell us the yardage between the shot and goal. Using the program Paint, we can view how many pixels away from the goal line, the edge of the penalty box is. As the edge of the penalty box is always 18 yards from the goal line, we can use this to get our scale. As seen in Figure 4, we can establish that there are 8 pixels in a yard since $144/18 = 8$. We can then use this to scale to get the shot location. For example, the shot highlighted is from 12 yards as $96/8 = 12$. Continuing to do this for Liverpool's other 11 shots generates Table 2. We will assume that this way of calculating the perpendicular distance is accurate throughout the report.

Table 2: Liverpool shot data

| **Distance from Goal (Yards)** | **Goal or No Goal** |
|---|---|
| 6 | 1 |
| 6 | 0 |
| 8 | 0 |
| 10 | 0 |
| 10 | 0 |
| 12 | 0 |
| 13 | 0 |
| 16 | 1 |
| 17 | 0 |
| 20 | 0 |
| 20 | 0 |
| 25 | 0 |

As Table 2 shows the shot data collected from Liverpool's shots in the game between Crystal Palace vs Liverpool, we will continue to do this for the other 19 teams from the ten games. After collecting the 238 shots from the gameweek, we can calculate the logistic function. Importing the data into statistical programming software 'r', we can run a few lines of code, and we get the following coefficient table.

Table 3: Coefficient table

| | **Estimate** | **Std. Error** | **z Value** | **Pr($>|z|$ )** |
|---|---|---|---|---|
| **Intercept** | 0.20001 | 0.46322 | 0.432 | 0.666 |
| **Distance** | -0.15280 | 0.03685 | -4.146 | $3.38 \times 10^{-5}$ |

The first thing that Table 3 shows is that distance harms the likelihood because the distance estimate is negative (-0.1528), i.e. the further away from the goal line a shot is taken the less likely it results in a goal. To test the significance of the relationship, we can perform a hypothesis test.

## 3.3 Hypothesis Testing

**Definition 6.** *A statistical* ***hypothesis test*** *is a method of making statistical decisions using experimental data.*

To conduct the hypothesis test, we first assume that there is no relationship between distance away from goal and the probability of a shot resulting goal. The formula for logistic regression is

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 x \tag{3}$$

which can be converted to

$$P = \frac{\exp(\beta_0 - \beta_1 x)}{1 + \exp(\beta_0 - \beta_1 x)}. \tag{4}$$

where $P$ is the probability of the dependent variable occurring, $\beta_0$ and $\beta_1$ are the coefficients and $x$ is the independent variable. If the probability of the dependent variable occurring is not affected by the value of $x$ then $\beta_1$ will be equal to 0. Hence we can formulate the hypothesis test in the following way:

**Hypothesis $H_0$.** $\beta_1 = 0$ *(Null Hypothesis).*

**Hypothesis $H_1$.** $\beta_1 \neq 0$ *(Alternate Hypothesis).*

Now to conduct the hypothesis test, we look at the distance z-value in Table 3. Dividing the distance estimate by the distance standard error calculates the distance z-value. z-values are related to the normal distribution curve (Figure 5). The numbers along the x-axis are the z-values.

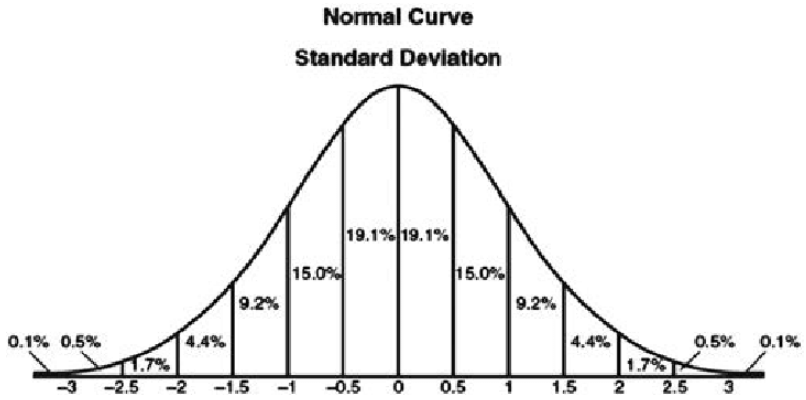


Figure 5: Normal distribution curve

To see if our z-value (which we can call the test statistic) is significant, we need to look at the critical z-value. If the modulus of our test statistic is larger than the critical z-value, then we say the result is significant. The only thing that affects the critical z-value is the level of significance.

**Definition 7.** *The **significance level** is the probability of rejecting the null hypothesis when it is true.*

We can choose our level of significance to be 2%; this means that there is a 2% chance of concluding that a relationship exists when there is no actual relationship. The test is a two-tailed because we are testing that $\beta_1 \neq 0$ rather than if $\beta_1 > 0$ or $\beta_1 < 0$ so it could be bigger or smaller in our test. Due to this, we have to half our level of significance when getting the critical value due to the symmetrical nature of the normal curve, as shown in Figure 5. Hence looking at the z-tables, we get a critical value of 2.326. As $|-4.146| = 4.146 > 2.326$ we can reject $H_0$, as our data suggests that there is a relationship between distance away from the goal line and the probability of a shot ending up in goal.

This is also backed up the Distance's $\Pr(>|z|\ )$ value in Table 3. This value is generated by the statistical software r when conducting a regression test. The actual name for this value is the p-value, it tells the probability of obtaining a z-value at least as extreme as the results observed during the test, assuming that the null hypothesis is correct. In our test, the p-value is $3.38 \times 10^{-5}$; this means there is a $3.38 \times 10^{-5}\%$ chance of obtaining a z-value as extreme as -4.146 if the null hypothesis was true. With the percentage being minute, our results are highly significant, and we are again lead to rejecting $H_0$.

## 3.4 Summary of Advanced Model

This means we can input our estimated coefficients from Table 3 into the logistic equation (4) giving us

$$P = \frac{\exp(0.20001 - 0.15280D)}{1 + \exp(0.20001 - 0.15280D)} \tag{5}$$

where $P$ is the probability of a shot resulting in a goal and D is the distance between the shot and the goal line.

Using r, we can plot the line and show the logistic curve displaying the negative relationship between distance away from the goal line and probability of a shot resulting in a goal as seen in Figure 6.

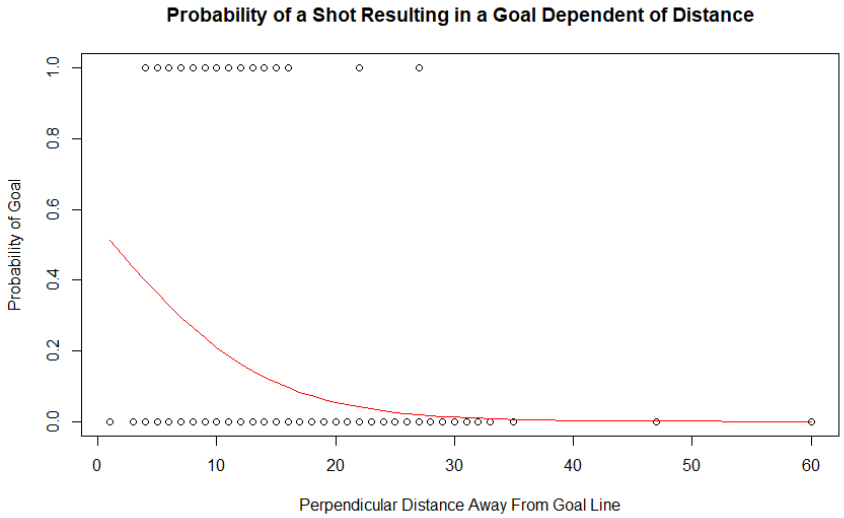


Figure 6: Graph showing relationship between shot distance and probability of a goal

Combining all our information from above we can then say the total expected goals for a team in a game is

$$xG = \sum_{i=1}^{n} \frac{\exp(0.20001 - 0.15280D_i)}{1 + \exp(0.20001 - 0.15280D_i)}. \tag{6}$$

where $D_i$ is the perpendicular distance between the $i^{th}$ shot and the goal line and $n$ is the total number of shots a team makes during a game.

## 3.5 Validation of Advanced Model

To check the suitability and quality of our model, we can perform a method of validation.

There are many different validation techniques that we could use for evaluating our logistic model including, K-fold cross-validation, leave-one-out validation or bootstrap. In this case, we will use K-fold cross-validation.

The idea behind K-fold cross-validation is that we will split the data into $K$ equal-sized (or near equal if not perfectly divisible) parts. We leave out part $k$ (called the test data), fit the model to the other $K-1$ parts (called the training data), and then obtain predictions for the left-out $k^{th}$ part. This process is repeated for $k = 1, ..., K$. We test to see if the actual class ($y_i$) is equal to the predicted class ($\hat{y}_i$).

For this model, we will use 5-fold cross-validation, meaning our sample of 238 shots will be split into five parts. A visual representation can be seen in Figure 7, where each square represents a part of the data, a blue square indicates the test data and a grey square represents the training data for each of the five iterations.

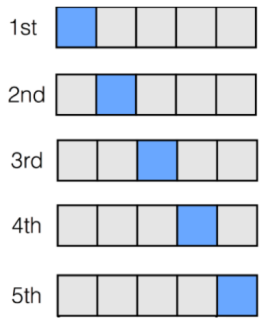


Figure 7: Visual representation of 5-fold cross-validation

To work out the accuracy of our model, we use the cross-validation formula for classification problems which is

$$CV_K = \sum_{k=1}^{K} \frac{n_k}{n} Err_k, \tag{7}$$

where

$$Err_k = \sum_{i \epsilon C_k} \frac{I(y_i \neq \hat{y}_i)}{n_k}. \tag{8}$$

Equation 8 uses the indicator function, this means if $y_i = \hat{y}_i$ (i.e if actual class is the same as the predicted class) then it assigns a value of 1 if not then it assigns a value of 0. Our model has a cross-validation value of 0.8697, meaning that 86.97% of the time our model correctly classifies the test set. To look at the accuracy in more detail, we can produce, using r, a confusion matrix. The confusion matrix shows us which shots from the 5 test tests the model correctly and incorrectly classified.

Table 4: Confusion matrix

| | **Actual No Goal** | **Actual Goal** |
|---|---|---|
| **Predicted No Goal** | 206 | 30 |
| **Predicted Goal** | 1 | 1 |

Logistic regression will classify depending on just which class has the greater probability. In our case, it will only classify a shot as a goal if it has a probability of 0.5 or greater. Although our model has a high accuracy rating, this is mainly due to it correctly predicting a shot that is not a goal correctly, as can be seen by Table 4. Our model only actually predicts two shots will result in a goal, with one of these being correct. This is to be expected as only shots that are less than 1.3 yards away from the goal have a higher than 50% chance of resulting in a goal according to our model.

If a team took three shots with a perpendicular distance away from the goal of 5 yards, the logistic model would classify them all as no goal as it has a probability of 0.3626. However, there is only a 25.89% chance that all three shots do not result in a goal. The 50% cut off is little importance in our model, as we are not using the model to classify whether a shot is a goal or not. We are more interested in the actual probability figure as this is also our expected goal figure for a shot.

We can look to improve our model by seeing if an equation of a different form fits our data better. Instead of using logistic regression of the form as seen in equation 3, we can see if including a quadratic term improves the model so we get

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 x + \beta_2 x^2, \tag{9}$$

which can be converted to

$$P = \frac{\exp(\beta_0 - \beta_1 x + \beta_2 x^2)}{1 + \exp(\beta_0 - \beta_1 x + \beta_2 x^2)}. \tag{10}$$

We can use the same data as we used before to fit a logistic model in the form of equation 10. Using the data, we get the following coefficient table (Table 5).

Table 5: Coefficient table

| | **Estimate** | **Std. Error** | **z Value** | **Pr(>\|z\| )** |
|---|---|---|---|---|
| **Intercept** | -0.197607 | 0.900421 | -0.219 | 0.826 |
| **Distance** | -0.083196 | 0.140408 | -0.593 | 0.553 |
| **Distanceˆ2** | -0.002484 | 0.004930 | -0.504 | 0.614 |

Putting the coefficient estimates from Table 5 into equation 10 we get

$$P = \frac{\exp(-0.197607 - 0.083196D - 0.002484D^2)}{1 + \exp(-0.197607 - 0.083196D - 0.002484D^2)}, \tag{11}$$

where $P$ is the probability of a shot with perpendicular distance $D$ between the shot and goal line resulting in a goal.

However, the p-values for the distance and distance squared show us that this model is unsuitable. Taking a 5% significance level, both values are much greater than 0.05, showing that there is not a significant relationship between either distance and distance squared with the probability of a shot resulting in a goal. Another way we can compare equation 5 with equation 11 is by looking at their Akaike information criterion value (which r gives us after running the regression).

**Definition 8.** *The **Akaike information criterion (AIC)** is an estimator of out-of-sample prediction error and thereby the relative quality of statistical models for a given set of data.*

As the AIC is an estimator of the prediction error, the smaller the number - the better the quality of the model (Hastie et al., 2017). For the model with just distance (equation 5) the AIC is 162.42, whereas the model with distance and distance squared (11) the AIC is 164.14. This is a further backup that including a distance squared term makes our model worse. Hence, the model we produced that most accurately calculates the probability of a shot resulting in a goal is equation 5, which we used to create the expected goals model seen in equation 6. It is also better than the first simple model (equation 2) that we created as it does not assume all shots have the same likelihood of ending up in a goal, however it does not take into account many other affecting factors such as the angle to the goal.

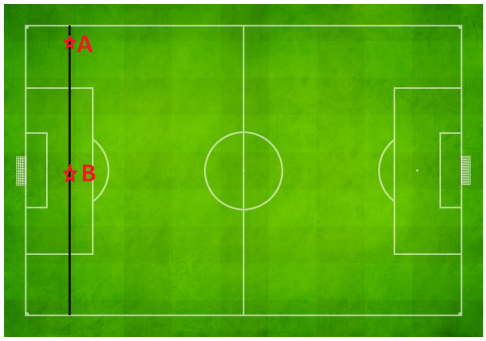


Figure 8: 12 yard line marked on football pitch

For example, Figure 8 shows a line that is 12 yards away from the goal line. Using our model, any shot taken from this distance away has a probability of 0.1633 (4 significant figures) of resulting in a goal and hence has an $xG$ of this value. However, it is much more likely that a shot from point B would be scored than a shot from point A due to it having a smaller euclidean distance to the goal and being a less tight angle.

Opta's model is more advanced as they take account of more factors. The factors they include are the type of assist, body-part that hit the ball last; if it is considered a big chance; and the angle and distance to the goal (Opta, 2017). As well as this, their training sets are far bigger than the one we have used to create model 6. The data they used to fit their model was based on shots from 3 Premier League seasons (1,140 games) whereas our model is based on shots from just 10 games leading to their model being more accurate at assigning $xG$ values to shots.

# 4 Analysing Team Performances

Unlike sports like basketball or cricket, football is a low scoring game. For example, in the top flight of English football, a team has, on average, scored 1.485 goals per game (Worldfootball.net, 2020). The number of goals a team scores in a game follows the Poisson distribution (a distribution used to model discrete events). As the mean and variance are equal in the Poisson distribution, this means the standard deviation is equal to the square root of 1.485, which is 1.219. Standard

deviation is a measure variation, as the standard deviation is only slightly smaller than the mean; this shows that luck and randomness play a massive role in football and so make it a harder game to predict. Expected goals models can be used to reduce the role of luck to determine the underlying performance levels.

For example, a team may be creating lots of chances but not scoring, which could be due to several reasons such as the opposition goalkeeper playing very well or their strikers not being in good form. Hence, their actual goal figures would be less than their expected goals figure; in this case, we say the team is under-performing their expected goals figure. In the long term, we would expect this not to continue and that their actual goal figure will be very similar to their expected. Likewise, if a team is over-performing their expected goals figure (scoring more goals than their expected goals figure), we would expect the difference between their actual goals and expected goals to reduce.

For example, Table 6 shows the 2018/19 Bundesliga (Germany's top division) top 5 table after 12 games.

Table 6: Bundesliga top 5 table after 12 games

| **Position** | **Team** | **Points** | **Goals** | $xG$ | **Conceded** | $xGA$ |
|---|---|---|---|---|---|---|
| 1 | Dortmund | 30 | 35 | 22.84 | 13 | 15.16 |
| 2 | Gladback | 26 | 30 | 23.43 | 14 | 14.59 |
| 3 | Frankfurt | 23 | 29 | 22.55 | 10 | 16.32 |
| 4 | Leipzig | 22 | 22 | 24.55 | 10 | 13.76 |
| 5 | Bayern | 21 | 23 | 24.65 | 17 | 11.24 |

Table 6 shows that Bayern Munich was in $5^{th}$ place in the league having scored fewer and conceded more than the four teams above them. However, at this point, their expected goals figure was higher than the four teams above them and their expected goals conceded was less than any of the others. This data would suggest Bayern Munich were extremely unlucky up to this point and that teams like Dortmund were very fortunate having exceeded their expected goals by 12.16.

Table 7: Bundesliga top 5 final league table

| **Position** | **Team** | **Points** |
|---|---|---|
| 1 | Bayern | 78 |
| 2 | Dortmund | 76 |
| 3 | Leipzig | 66 |
| 4 | Leverkusen | 58 |
| 5 | Gladbach | 55 |

Table 7 shows the final 2018/19 Bundesliga league table (after 34 games). As was expected by looking at the data from Table 6, Bayern Munich moved up the table and ended up finishing top. Also, the teams that were over-performing their data after 12 games reverted towards their expected levels of goals and goals conceded, which caused them to finish lower down in the table.

# 5 Analysing Individual Performances

One of the factors that expected goals models do not take account of is the player that is taking the shot. The reason for each player not having their own $xG$ model is because it is tough to put a numeric value on the quality of a player. Hence, taking all other factors to be constant a shot from a Premier League striker will be given the same $xG$ figure as a lower league defender. Nevertheless, how much of a problem is this?

Cristiano Ronaldo and Robert Lewandowski are often considered two of the best strikers in the world. They were ranked best and $7^{th}$ best male footballers by the Guardian in 2016 (The Guardian, 2016). Due to this, we would expect more of their shots to result in a goal than the average footballer. However, Ronaldo has underperformed his expected goals in each of his last four seasons, and Lewandowski has in two of his previous four including massively in his most recent season as shown in Table 8 and Table 9 respectively (Understat, no date).

Table 8: Cristiano Ronaldo's expected goals

| **Season** | **Appearances** | **Goals** | $xG$ | **Difference** |
|---|---|---|---|---|
| 2018/19 | 31 | 21 | 23.32 | +2.32 |
| 2017/18 | 27 | 26 | 27.00 | +1.00 |
| 2016/17 | 29 | 25 | 25.41 | +0.41 |
| 2015/16 | 36 | 35 | 35.59 | +0.59 |

Table 9: Robert Lewandowski's expected goals

| **Season** | **Appearances** | **Goals** | $xG$ | **Difference** |
|---|---|---|---|---|
| 2018/19 | 33 | 22 | 33.14 | +11.14 |
| 2017/18 | 30 | 29 | 27.90 | -1.10 |
| 2016/17 | 33 | 30 | 30.10 | +0.10 |
| 2015/16 | 32 | 30 | 28.46 | -1.54 |

From this, we can conclude that the reason that these strikers are highly rated is not that they are scoring more often from a given shot than the average player. As if this was the case, the difference between their $xG$ and goals would be consistently negative. Instead, we can look at how their expected goals figure compares with the rest of the league in that season. Lewandowski has played for Bayern Munich in the Bundesliga (Germany's top division) for the past four seasons. Whereas, Ronaldo played for Juventus in Serie A (Italy's top division) in the 2018/19 season and for Real Madrid in La Liga (Spain's top division) for the previous three seasons.

Table 10: Cristiano Ronaldo's expected goals comparison

| **Season** | **Largest $xG$ Player** | **Largest $xG$** | **Ronaldo $xG$** | **Rank** | **Players** |
|---|---|---|---|---|---|
| 2018/19 | Quagliarella | 23.93 | 23.32 | $2^{nd}$ | 543 |
| 2017/18 | Messi | 28.95 | 27.00 | $2^{nd}$ | 557 |
| 2016/17 | Messi | 26.89 | 25.41 | $2^{nd}$ | 541 |
| 2015/16 | Suárez | 32.12 | 35.59 | $2^{nd}$ | 539 |

Table 11: Robert Lewandowski's expected goals comparison

| **Season** | **Largest $xG$ Player** | **Largest $xG$** | **Lewandowski $xG$** | **Rank** | **Players** |
|---|---|---|---|---|---|
| 2018/19 | Lewandowski | 33.14 | 33.14 | $1^{st}$ | 468 |
| 2017/18 | Lewandowski | 27.90 | 27.90 | $1^{st}$ | 471 |
| 2016/17 | Aubameyang | 31.14 | 30.10 | $2^{nd}$ | 469 |
| 2015/16 | Lewandowski | 28.46 | 28.46 | $1^{st}$ | 476 |

From Table 11 we can see that Lewandowski has the highest expected goals figure in his league in three of the last four seasons and is $2^{nd}$ in the other out of over 450 players each season. Ronaldo is $2^{nd}$ in each of the last four seasons in the league. From these tables, we can see that both players consistently have one of the highest $xG$ in their respective leagues. The only players to have higher $xG$ figures than either Ronaldo or Lewandowski are Fabio Quagliarella, Luis Suárez, Lionel Messi and Pierre-Emerick Aubameyang who are also known for being elite footballers.

We can conclude from these tables that we value players more highly that have larger $xG$ figures than them over-performing their $xG$ (scoring more than their $xG$ figure). Having a high $xG$ means that these strikers are taking shots that are quite likely in resulting in a goal. Hence, if a striker has a higher $xG$ figure, it means they are getting in the best position possible to maximise the chance of their shot resulting in a goal.

# 6 Application of Expected Goals

The idea of expected goals was first mentioned in April 2012 when an analyst called Sam Green posted a blog discussing the importance of shot quality rather than quantity (Green, 2012). Other amateur analysts started developing their own expected goals models taking inspiration from similar models that already existed in ice hockey. Companies such as Smartodds (a betting consultancy who collect data, analyse it and then sell it to professional gamblers) started to use these new models. High-rolling clients of these companies look for the teams that have been performing well but have been suffering bad results as they believe these sides have been unlucky and their form will regress to the mean. The professional gamblers ended up making millions in the football betting markets as many betting firms had not yet incorporated these new findings when setting their odds and were just looking at their recent results, not performances.

Professional football clubs are increasingly using the Expected goals metric. One of the trailblazers were Championship ($2^{nd}$ highest division in England) side Brentford due to their owner Matthew Benham also being the owner of Smartodds. Benham not only sells the data to pro gamblers but also uses it to help Brentford in many ways. For example, he uses it to scout players before signing them. Since 2015, Brentford's net transfer spend (transfer revenue - transfer expenditure) is £63.8 million (Transfermarkt, 2020) and are reaping the rewards of many bargain signings. Due to this, many clubs have since implemented $xG$ into their scouting methods (Tippett, 2019). Brentford also uses a 'table of justice' using $xG$ data as they believe 'the league table lies, even a long way into the season' (Burt, 2016) due to the role of luck in a low scoring sport. The hierarchy at the club believes that the 'table of justice' helps them assess how they are genuinely performing compared to their opponents more so than the actual league table.

# 7 Limitations of Expected Goals

Due to the infancy of expected goals models, there are still problems with them. As discussed in the report, shot location is one of the main factors that affect how likely it is that a shot results in a goal. However, how accurate is this shot location? Opta's expected goals models are the ones used by most football clubs and media outlets. Despite this, an employee pinpointing on a digital pitch map the location of each shot is how they determine their shot location. The shot location is likely to be inaccurate in many instances, leading to incorrect goal likelihoods. As technology continues to advance, GPS tracking devices on players would mean shot locations will become far more accurate.

Another glaring weakness of expected goals models is that they do not take into account the exact position of opposition players. Because of this, a player shooting into an open net has the same $xG$ figure as a shot through a crowd of players (holding all other factors constant). It would be much easier for a footballer to score if there are fewer opposition players between him and the net.

The final main drawback of the models is that expected goals data do not account for dangerous attacks that do not result in a shot. For example, a dangerous ball across the face of goal which narrowly misses the outstretched foot of a player would have no affect on the $xG$ as the player did not connect with the ball. In instances like this, the team has created an opportunity to score, but because no actual shot took place, there will be no record in the data. On top of this, $xG$ models do not recognise own goals despite the fact the Premier League has averaged 36 own goals per season since the 2012/13 season (EPL Review, 2020).

# 8 Conclusion

In this report, we have investigated how statistical techniques can be used in football to enhance the football experience for everyone who appreciates the merits of the method. For an average football fan, it is now possible to see a numerical value on the quality of chances that a team has created. For professional gamblers, they now have more extensive data aiding their ability to outsmart the bookmakers. Moreover, for football clubs, they can use the metric in order to improve their recruitment of players which will benefit them in terms of results and possibly financially.

This report has shown that expected goals models can be produced by anyone using data found off the internet and can be made more advanced and accurate by large corporations. As technology advances, the metric will also become more advanced. In the future, The Football Association may decide to place a computer chip into a football. This change would improve the accuracy of the shot location and allow analysts to incorporate additional factors into their $xG$ models. Factors could include the speed, acceleration and the curvature of the ball. They could also place computer chips into player's boots which would allow models to include the position of opposition players. The more advanced and accurate the $xG$ models become increases the chance of more interest. As exposure increases, the creation of more novel ways of interpreting and advancing the model will occur, leading to an exponentially positive cycle.

# References


BBC Sport (2018), 'Match of the Day Viewing Figures', Available at: https://www.bbc.co.uk/sport/football/42873526. (Last accessed on 2 April 2020).

Burt, J. (2016), 'Brentford FC - The club thinking outside the box'.

EPL Review (2020), 'EPL Own Goal Statistics History', Available at: https://www.eplreview.com/statistics-owngoal.htm. (Last accessed on 2 April 2020).

Footstats (2020), 'Shots and goals per Premier League season', Available at: http://www.footstats.co.uk/index.cfm?task=league_shots. (Last accessed on 2 April 2020).

Google Trends (2020), 'Google Trends of Expected Goals in UK', Available at: https://trends.google.com/trends/explore?date=2009-01-01%202019-11-25&geo=GB&q=expected%20goals. (Last accessed on 2 April 2020).

Green, S. (2012), 'Assessing the performance of Premier League goalscorers', Available at: https://www.optasportspro.com/news-analysis/assessing-the-performance-of-premier-league-goalscorers/. (Last accessed on 2 April 2020).

Hastie, T., Friedman, J. and Tisbshirani, R. (2017), *The Elements of Statistical Learning: data mining, inference, and prediction*, Springer.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2017), *An introduction to statistical learning: with applications in R*, Springer.

Opta (2017), 'Advanced metrics: Expected Goals', Available at: https://www.optasports.com/news/advanced-metrics-expected-goals/. (Last accessed on 2 April 2020).

The Guardian (2016), 'The 100 best footballers in the world 2016'.

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

Transfermarkt (2020), 'Championship Transfers', Available at: https://www.transfermarkt.co.uk/championship/transfers/wettbewerb/GB2. (Last accessed on 2 April 2020).

Understat (no date), 'xG stats for teams and players from the TOP European leagues ', Available at: https://understat.com/. (Last accessed on 2 April 2020).

Who Scored (2019), 'WhoScored Chalkboard', Available at: https://www.whoscored.com/Matches/1376023/Live/England-Premier-League-2019-2020-Crystal-Palace-Liverpool. (Last accessed on 2 April 2020).

Worldfootball.net (2020), 'Goals per season', Available at: https://www.worldfootball.net/stats/eng-premier-league/1/. (Last accessed on 2 April 2020).