International Journal of Informatics and Information Systems ISSN 2579-7069
Vol. 4, No. 1, March 2021, pp. 38-54

Implementing the Expected Goal (xG) Model to Predict Scores in Soccer Matches

Izzatul Umami [1,*], Deden Hardan Gutama [2], Heliza Rahmania Hatta [3]

[1] Darul Ulum University, Indonesia
[2] Alma Ata University, Indonesia
[3] Mulawarman University, Indonesia
[1] izzatul.umami@undar.ac.id*; [2] hardan@almaata.ac.id; [3] heliza_rahmania@yahoo.com
* corresponding author

(Received February 7, 2021 Revised February 21, 2021 Accepted February 27, 2021, Available online March 1, 2021)

Abstract

Football is a sport that has the most fans in the world. What makes sebak patterns so popular are their uncertain and unpredictable results. There are many factors that affect the outcome of a football match, including strategy, skill, and even luck. Therefore, guessing the results of a soccer match is an interesting problem. All shots are grouped into sections on the playing field and theoretical goal scores are applied to each area. The factors analyzed are: distance of shot from goal and angle of shot in relation to goal. When calculating xG, it is recommended that the distance and angle of the shot are important. The combination of the two xG factors is better calculated than each variable only. In addition, this xG check has been able to relatively accurately identify the mid-table teams that score and concede goals.

Keywords: xG model, Soccer match, Distance, Predict scores.

1. Introduction

In an article published by Bloomberg in 2018, 4 out of 10 people stated that they are soccer fans [1]. This makes football the most popular sport in the world. Uncertainty is the very nature of football [2] which makes this one of the factors why football is so included. As The New York Times said in an article entitled Soccer, a Beautiful Game by Chance, there are many components that affect the final outcome of a match, such as strategy, skill, and luck. These factors make the outcome of each match unique and unpredictable. However, from every football match data can be obtained that can be used to analyze how the match is going.

With the development of technology, important data related to soccer matches are easier to obtain. These data can be processed and used to make predictions on future matches. One of the fields in Computer Science that is widely used to make predictions based on data is Machine Learning. Machine Learning, in its definition, is a field in Computer Science that can study certain patterns from data sets and make predictions or classifications based on those data sets. The use of Machine Learning in problems like this is very suitable, because of the large amount of data available, soccer is also difficult to predict based on logic, or other explicit reasons [3]. Some examples of popular Machine Learning algorithms today are Artificial Neural Network (ANN) and Support Vector Machine (SVM).

In a previous study [4], SVM and ANN were used to play soccer, but the results obtained by SVM were disappointing, the accuracy obtained was only 53.3%, while ANN was impressively capable of producing an accuracy above 80%. Reflecting on the results of this study, the method to be used in this research is Expected Goals (xG). Expected Goals (xG) are statistical calculations regarding the quality of an opportunity. This metric can be used as a measure of the performance of a team. In matches, team performance is often not directly proportional to the final result. Here xG is able to provide additional stories about the team's performance in these matches.

2. Literature Review

Since all of these factors have been analyzed separately in various studies (as mentioned above), they have not been analyzed together in terms of what or which factor gives the best indication of scoring the most goals by a player. This is what the Expected Objectives (xG) Method attempts to calculate. In its simplest form, this method involves calculating a team's chances of scoring and conceding goals. The definitions for calculating xG have been interpreted similarly among blogs and websites. The units of measure are 0 (no purpose) to 1 (target) [5]. Through the research carried out on this topic, this appears to be the first academic study to be undertaken in professional football. Other sports such as Ice Hockey have used xG to evaluate players as seen from Macdonald's [6] presentation on the topic at the MIT Sloan Conference in 2012.

Why is the xG method useful in football? This model has proven to have practical applications in the world of football, especially at club level in Denmark. FC Midtjylland football club won their first Danish league title using this method of recruiting players [7]. Although for obvious reasons they keep their models and calculations secret, it does show that there has been success in providing such statistical methods to coaches and clubs to improve the club's future prospects. Elsewhere, Cardoso [8] states that xG is a good indicator of future team performance.

However, the model has also received criticism on a number of levels. Because of the many factors that need to be included in calculating xG, there is no model that accurately measures the most significant factor of xG [9]. Bertin critically questions ExG. He noted that "a study that looked only at the distance of the shot from the goal resulted in a very high probability rating (0.997) and therefore no other factor to take into account." Since then it has been checked that there is not always only one factor that provides an accurate evaluation of team / individual xG [10].

Defensive Pressure (Dp) has proven to be another area that has received critical attention in studies on xG [11] and to what extent it reduces the xG value of a shot [12]. The lack of defensive position of the player according to the specific action of the game (crosses, through passes and through passes) is described by Cardoso [8]. The different views in calculating Dp and xG were successfully examined [13-15].

A third criticism of the model is explained by Van roy [16] and Kim [17]. According to both, the model does not take into account the appearance of famous clubs during matches and explains why they are better than others [18]. Obviously this can easily be explained in the sense that they have better (though not always efficient) players and more resources in the expense of beating their league competitors. In this case a further criticism is mentioned by Caley. The model according to past study does not indicate which player took the shot, which can be seen as a negative factor. Lauer [19] explains that through this model, FC Midtjylland was able to identify the Finnish midfielder TimSparv as their latest signing.

Although criticism is not without reason, the xG method is still seen as a valuable tool in predicting which players and teams will score and the likelihood of conceding 20. This is noted by Lauer [19] especially if a model is good and builds confidence, or if the player / team outperforms the model. The essence of the xG method is prediction, if not noted by Liu [20] as "historical trends". Fitriana [21] has discussed a number of other factors that can influence xG, while others have described it as 'infinite'. Therefore, this model is still under construction. As noted above, the need for a more central approach with one model being considered and tested is required (such as the FC Midtjylland model used) compared to multiple models [22].

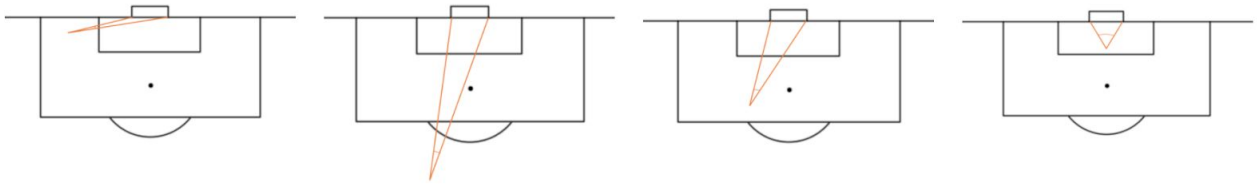
3. Method

Before we build our xG model, we need to consider the type of data we use. Especially in this study we need a large collection of kick data but more importantly we need data to describe the types of shots that result in goals. We can conclude that the most important factors we need are the distance from the goal when the kick is taken, the angle in relation to the goal, and which part of the body the shot was taken.

Football data is usually divided into two forms: event data and tracking data (or score and match analysis). Event data records all events on the ball and where on the field they occurred (such as kicks, passes, tackles, dribbles), while tracking data records the position of players and the ball throughout play at regular intervals. The event data we'll be using today comes from Wyscout. It includes all events from all matches in the top 5 domestic leagues in Europe from the 2019/2020 season. While some of the findings in this section may seem fundamental to those with a broad

understanding of football, we have always believed that it is important to test our assumptions, as they can sometimes be misleading.

Before we explore our newly created dataset, we need to do some data cleanup. It's normal, especially with such large data sets, that there may be some values entered incorrectly, some missing values or just situations we didn't anticipate. For example, we have to check to see why we're getting an error in arccos. There are a number of points answered by clearing the data, namely we know that the error results in nan values & the nan values generated by the error are 6.



	Goal	x	y	playerid	teamid	matchid	header	Y	X	Center_dis	Distance	Angle Radians	Angle Degrees
417224	1	0	49	4131	698	2565801	0	33.32	0.0	1.0	0.68	NaN	NaN
365140	1	0	52	224971	2445	2516954	0	35.36	0.0	2.0	1.36	NaN	NaN
499325	1	0	57	206314	3161	2576251	0	38.76	0.0	7.0	4.76	NaN	NaN

Fig. 1. NaN Values on Features

There seems to be a few goals scored from the touch line which requires us to rethink how we construct our corner attributes. Since there are only 3 occurrences of such occurrences and since they are usually rare, unintended events, we will remove them from our model. This is mainly to keep things simple. After removing it as you can see in the figure below it looks like we have a few unnecessary columns which hold dummy variables when we calculate distances and angles.

	Goal	x	y	playerid	teamid	matchid	header	Y	X	Center_dis	Distance	Angle Radians	Angle Degrees
213	1	6	57	256992	3799	2500686	0	38.76	6.30	7.0	7.896050	0.755576	43.291300
302	0	17	42	334552	3772	2500686	1	28.56	17.85	8.0	18.660549	0.372069	21.317963
498	1	4	43	26389	3772	2500686	0	29.24	4.20	7.0	6.348039	0.851948	48.813019
577	0	16	21	276920	3772	2500686	0	14.28	16.80	29.0	25.905953	0.184838	10.590449
629	0	27	51	366760	3799	2500686	0	34.68	28.35	1.0	28.358154	0.256637	14.704224
...	...	...	...	...	...	...	...	...	...	...	...	...	...
642945	0	28	45	8561	1633	2500098	0	30.60	29.40	5.0	29.595946	0.244517	14.009788
643023	1	14	33	41174	1633	2500098	0	22.44	14.70	17.0	18.700898	0.309646	17.741433
643051	0	12	62	7879	1623	2500098	0	42.16	12.60	12.0	15.011516	0.410444	23.516712
643055	0	8	38	145692	1623	2500098	0	25.84	8.40	12.0	11.710918	0.461143	26.421528
643149	0	14	50	8005	1633	2500098	1	34.00	14.70	0.0	14.700000	0.488036	27.962409

Fig. 2. Dataset after NaN Features deleted

3.1. Data Visualization

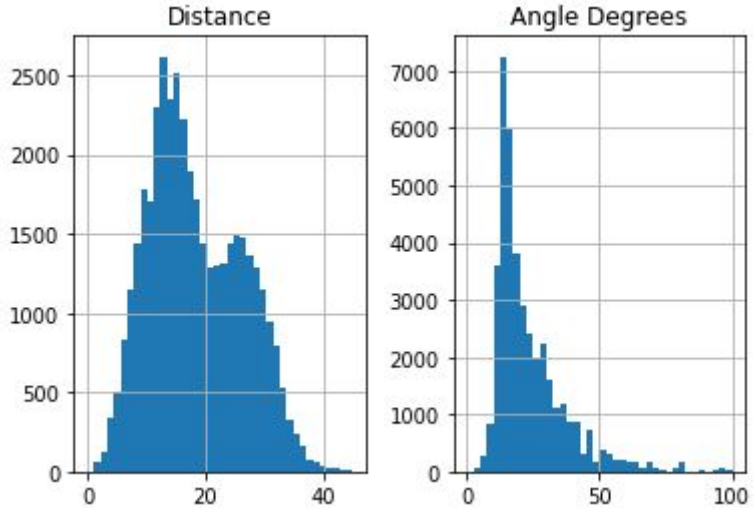


Fig. 3. Distributions of Shots by Distance and Angle

After doing basic visualization such as dividing the kick plot based on side-by-side distance and angle, there are a number of conclusions we can draw. The distribution shows that:

1. The majority of kicks occur between 10 and 20 meters.
2. Shots taken within 6 meters are quite rare compared to shots taken outside 10 meters
3. Surprisingly, there is a trough in the local area which is taken between 18m to 25m husk
4. As expected, players rarely take kicks from less than 5 degrees, corresponding to a center shot close to the goal

The angle distribution corresponds to the distance distribution because shots taken up close (larger angles) are much more difficult to produce. With just a simple distribution chart, we can conclude that it is quite difficult to produce a close and central shot on goal. While we now know how kicks are distributed by distance and angle, we haven't discussed how kicks that lead to goals differ from those that don't. after that we use the seaborn library to check the distribution of kicks by result (goal or no goal) and also the seaborn library to plot the violins and extract the required columns from our dataframe.

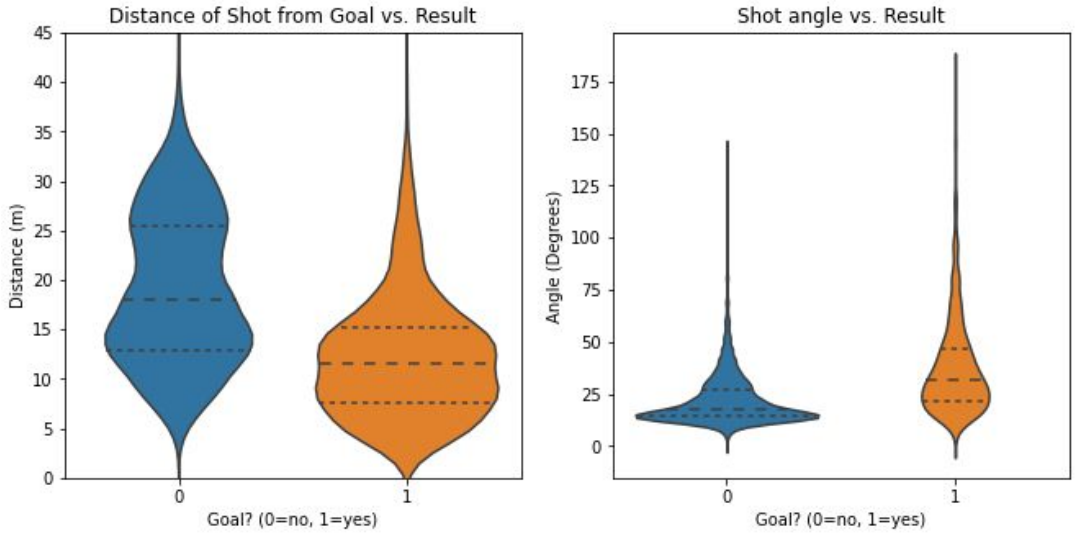


Fig. 4. Shot Distance Comparison Between Goal vs Result & Shot Angle vs Result

The violin plot above plays a similar role to the box and whiskers plot but also provides use with approximate kernel distribution of the data (which is basically a smooth distribution). In dividing the data by kick result, we can see that,

on average, the kicks that result in a goal are taken from a closer distance to the goal than the kicks which are not scored. The average goal-scoring shot is about 12 meters compared to 18 meters for those who don't make the goal stand out. Likewise, goals are usually scored from an angle of 20 degrees to around 50 degrees. So while it is difficult to produce shot opportunities that are close to goal, the plot of the violin shows that near and central opportunities tend to produce goals.

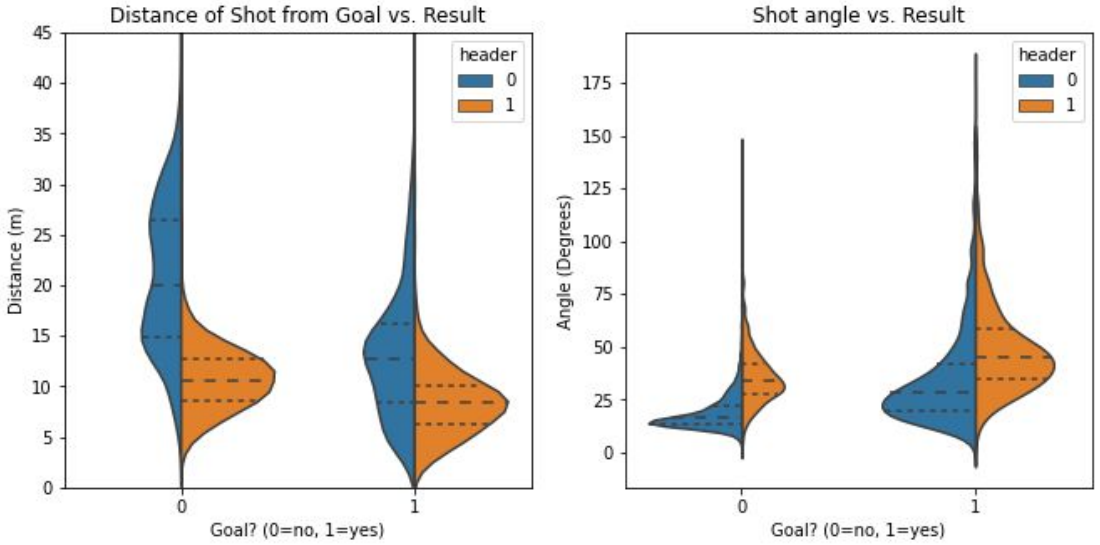


Fig. 5. Shot Distance Comparison

The parameter divides the plot by categorical data, in this case the header. As expected, it is usually taken in an 18 yard (16.5m) box. Interestingly, the means and distribution of results don't differ much, so that's something we should consider. We've gained some great insights through some basic distribution plots, but we can take it a step further. We can better visualize how these variables affect the outcome by plotting possible shots in the field. Namely, we want to divide the pitch into bins, count the number of shots taken inside each bin and then use a color gradient to visualize how the odds differ from one bin to another.

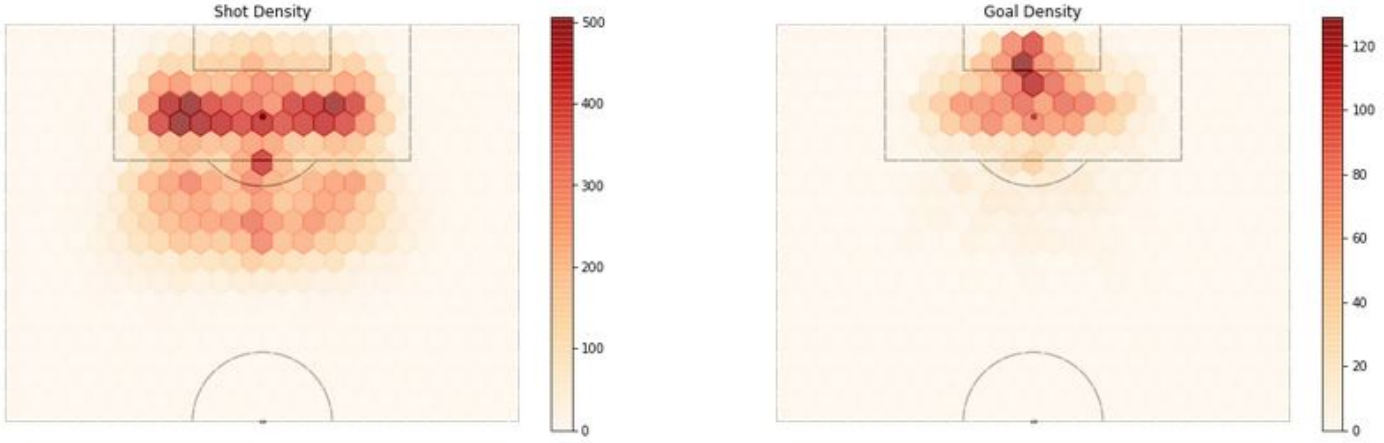


Fig. 6. Shot & Goal Density

As we can see in the figure, it is similar to the violin plot except now that we are plotting all kicks on the left and only aiming on the right, This density plot has a similar function to the violin plot above but gives us a much better visual understanding of the area. which court usually results in the kick. and goals. This is because we can see how distance and angle affect the distribution of shots on the same plot. As we learned with the violin plot:

1. Shots are rarely taken from either side of the box due to the bad angle
2. The majority of kicks are taken around the penalty spot (11m)
3. Goals are usually scored within 11 meters and in very narrow sections

Before we dive deeper, we need to discuss why there is a sharp drop in the number of shots at the edge of the box. This could be for a number of reasons.

1. We have generalized that all soccer fields are 105m x 68m in size when in reality each field has unique dimensions. The trough is probably a product of our generalizations.
2. Another factor may be the way Wyscout records its data. There may be a tendency to record kicks near the penalty box line both inside the box and outside. Therefore, it is possible that the shots that occur on the phone are of the wrong character.
3. Lastly, there may be a psychological effect on the players. Players can resist the temptation to kick while outside the box to dribble into the box in hopes of winning a penalty. Defenders tend to defend more tentatively when opponents are in the box due to costly foul possibilities and it's possible the attacker will want to take advantage of this.

There are other possibilities and it is difficult to draw any conclusions about the troughs but for now we have to accept them. Return to another data visualization. we can plot the probabilities to judge which areas of the court have a high probability of producing a kick on goal and which are not.

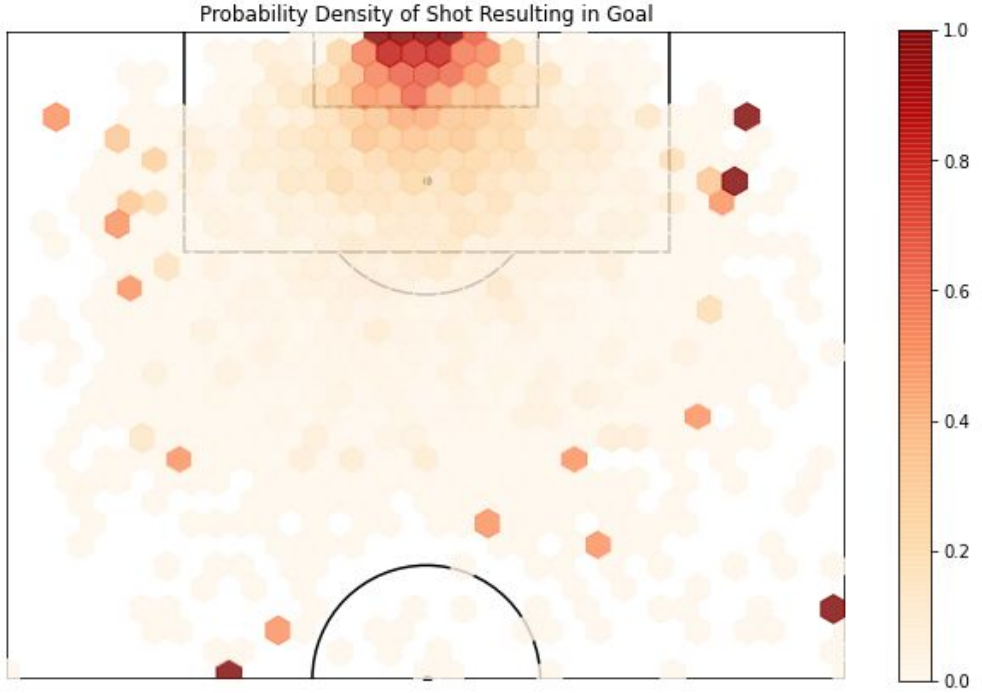


Fig. 7. Probability Density of Shot Resulting in Goal

As expected, the closer you are to shooting from goal, the more likely it is that the shot will score a goal. Note that there are certain outliers for which the probability in the bin is very high. This is because the few shots that are taken from the area lead to a goal. If we said 10 seasons of data, we would see a much more homogeneous possible probability.

One of the most surprising things for those who haven't used xG is that the odds of scoring from over 11 meters are actually lower than we might appreciate when we look at a match in person. It is for this reason that this kind of analysis is important. We tend to overestimate the quality of opportunities, such as shots from outside the box, when in fact there are shots that are quite difficult and inefficient. So the next time your favorite player misses from 11 meters, remember that he only has a 3 out of 10 chance of scoring. then what about the headers? The figure below shows the possible headers probability of producing a goal.

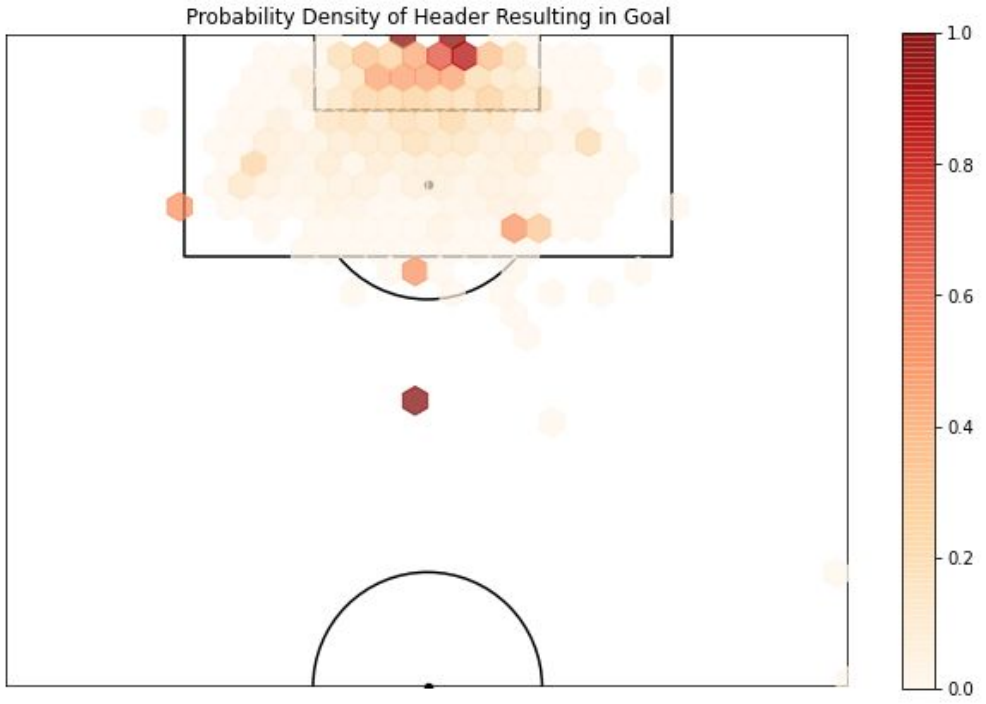


Fig. 8. Probability Density of Header Resulting in Goal

Even though the header shows a trend similar to that of a regular shot, it has a lower probability value overall. This seems to suggest that while the average header is close to the mark, it also represents a much harder chance to clear. As we will see later, this is an important finding and one that influences our interpretation of the opportunity evaluation. and again we will use the seaborn library to check the distribution of shots by result (goal or goalless). we want to create bins to calculate our probability, after that we want to find the mean of the Goal column (our prob density) for each bin and the average distance for each bin.

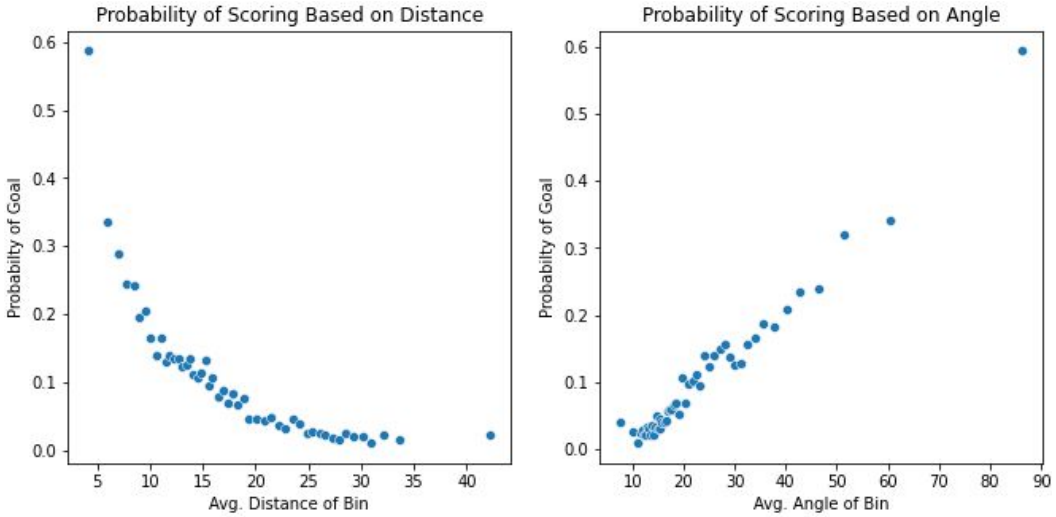


Fig. 9. Probability of Scoring Based on Distance and Angle

The first thing that comes up and it's quite interesting is that as we move away from the goal, the chances of scoring become exponentially more difficult. Now that's great because it greatly reduces the value of shooting from a distance. So why is this happening? Up until now, we had ignored the fact that the angle with the goal was reduced as we moved away from the goal. So we have this kind of 'multiplier factor' for distance. We can hypothesize that this is because as we increased the distance to shoot, not only was the mileage longer but the target was also smaller.

3.2. Modeling Expected Goals

In the previous section, we explored shooting data and trends based on three main variables; distance, angle and categorical variables for identifying lead kicks. We develop an understanding of the distributions and probabilities associated with kicks and goals by representing, modifying and visualizing data. Here, we use this data to develop a model for predicting goals.

	Goal	x	y	playerid	teamid	matchid	header	Y	X	Center_dis	Distance	Angle Radians	Angle Degrees
213	1	6	57	256992	3799	2500686	0	38.76	6.30	7.0	7.896050	0.755576	43.291300
302	0	17	42	334552	3772	2500686	1	28.56	17.85	8.0	18.660549	0.372069	21.317963
498	1	4	43	26389	3772	2500686	0	29.24	4.20	7.0	6.348039	0.851948	48.813019
577	0	16	21	276920	3772	2500686	0	14.28	16.80	29.0	25.905953	0.184838	10.590449
629	0	27	51	366760	3799	2500686	0	34.68	28.35	1.0	28.358154	0.256637	14.704224
...	...	...	...	...	...	...	...	...	...	...	...	...	...
642945	0	28	45	8561	1633	2500098	0	30.60	29.40	5.0	29.595946	0.244517	14.009788
643023	1	14	33	41174	1633	2500098	0	22.44	14.70	17.0	18.700898	0.309646	17.741433
643051	0	12	62	7879	1623	2500098	0	42.16	12.60	12.0	15.011516	0.410444	23.516712
643055	0	8	38	145692	1623	2500098	0	25.84	8.40	12.0	11.710918	0.461143	26.421528
643149	0	14	50	8005	1633	2500098	1	34.00	14.70	0.0	14.700000	0.488036	27.962409

Fig. 10. Datasets after Dropping NaN Values

The first thing to do is to remove the nan values and ensure that our binary data is numeric for further data generation. Since our response variable (kick result) is categorical, we had to apply a classification method to build a predictive model. To introduce this type of approach, let's look at an illustrative example. Assume we plot a random selection of shots from our data and classify them based on the results, taking only distance into account. What we next do is plot specific points to demonstrate what the Heaviside function does to select goals that come in and miss and then plot them with the boundary between them.

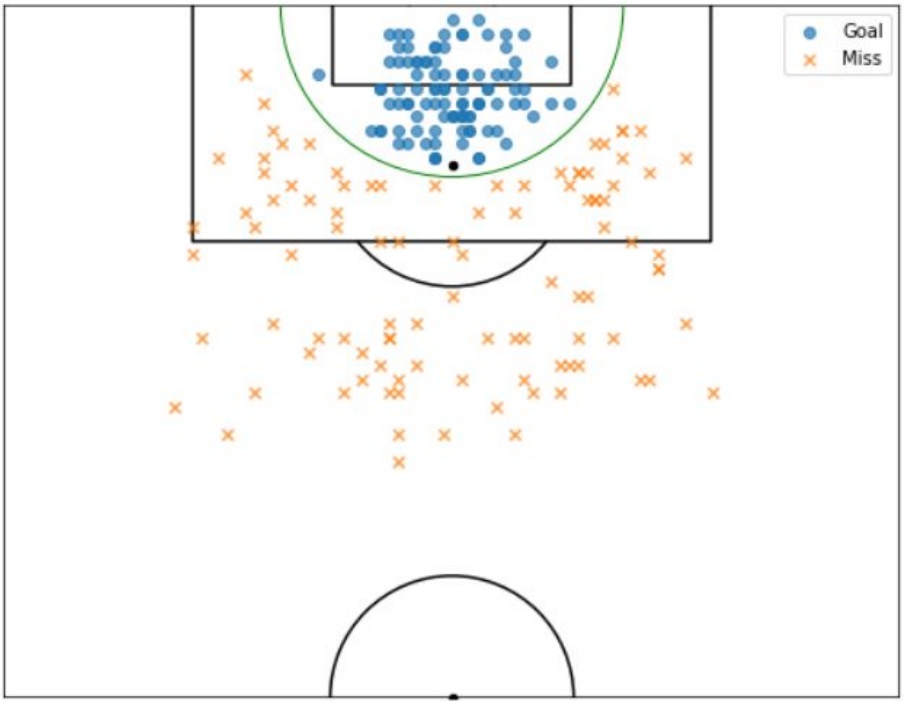


Fig. 11. Relationship between Shots and Distance

The above shows that there is a clear difference between the blue and orange clusters. If we assume that this is the true nature of the relationship between shot and distance, and we want to predict the outcome of future shots based on this data. Because the data is easy to separate, we can draw a boundary between the two clusters. This line will

represent the discriminant function, which we will use to classify each shot as a goal or a miss. In the above case, the discriminant function maps the distance r from the center of the goal.

This discriminant function, depicted in green, is determined by the equation $y = \alpha x + \beta$, where x is the distance from the destination and y is the binary response. Especially for the data above, y = x -12. we can translate this into a classification model by applying the so-called Heaviside function to our discriminant, which will return a value of 0 or 1 because the response variable is based on the y value:

$$H(y) \equiv \begin{cases} 0 & \text{if } y > 0 \\ 1 & \text{if } y \leq 0 \end{cases}$$

The Heaviside function is the simplest hard classifier and classification model, but requires data to be neatly separated. As you can see in the image below, we plotted the heaviside function above the responses to the data which can be separated above.

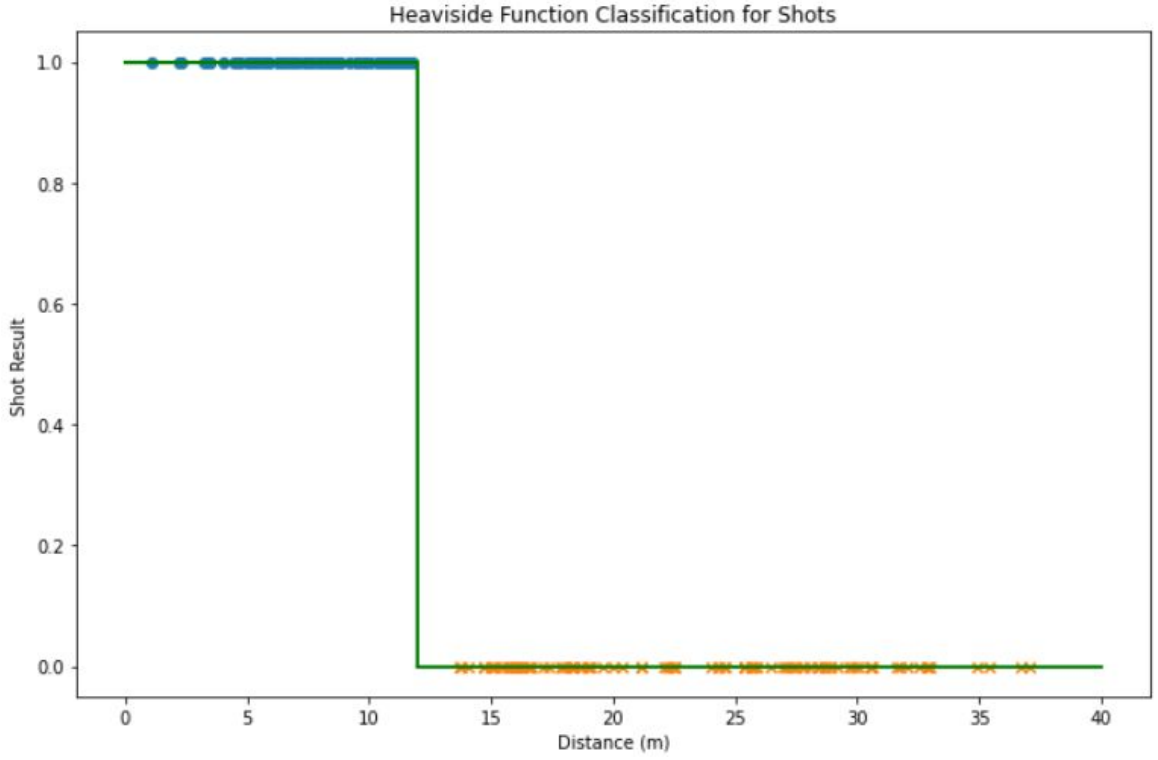


Fig. 12. Heaviside Function Classification for Shots

We'll create a contour plot to see how the ceiling function is classified into pitch. first we have to simulate the above data and we can do that by initializing. In reality, the data describing the shots cannot be neatly separated, making goals difficult to predict with any certainty. Nevertheless, Heaviside's function served as a natural progression into a more advanced classification model.

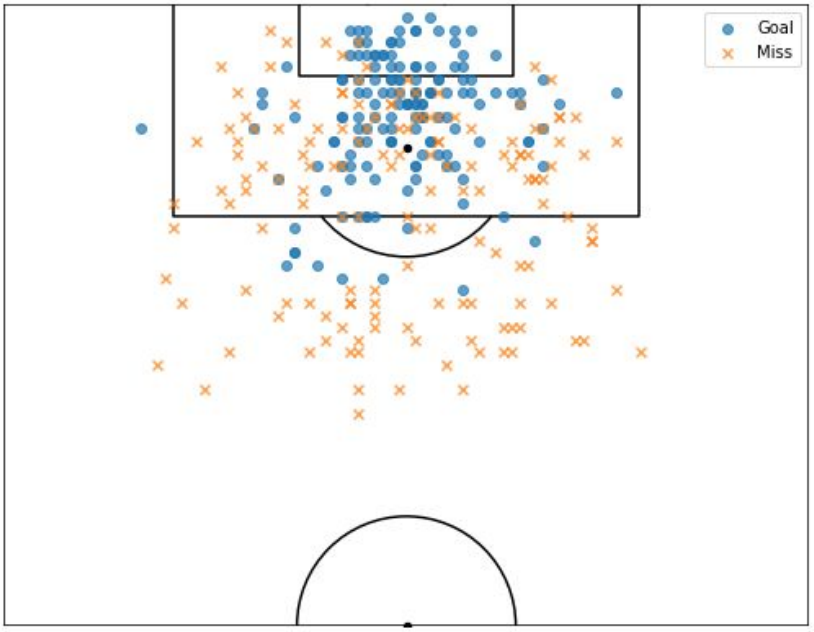


Fig. 13. Relationship between Shots and Distance with Different Characteristic

If we sampled actual data, it might be similar to the plot below. If the data couldn't be neatly disaggregated, we would have to look at the probability model as opposed to assigning zeros and ones. We have seen in our data exploration that there is a clear trend for probability with respect to the distance and angle variables. What functions should we adopt to model this trend? There are many functions that map probabilities and fit inseparable data, but we use the logistic function (also known as the sigmoid function) because of its simplicity.

$$G(y) \equiv \frac{1}{1+e^{-y}} \equiv \frac{1}{1+e^{-(\alpha x_1+\beta)}}$$

Similar to the weight function, the logistic function takes our predictor (distance in this case), but returns a value between zero and one. The logistic function takes any value in the domain $(-\infty, +\infty)$ and returns a value in the range (0.1). Thus, given a value of y, we can interpret $G(y)$ as the conditional probability that the shot is scored, $G(y) \equiv$Pr [label is 1 | y]. As we can see below, we plot a general logistics function for a demonstration.

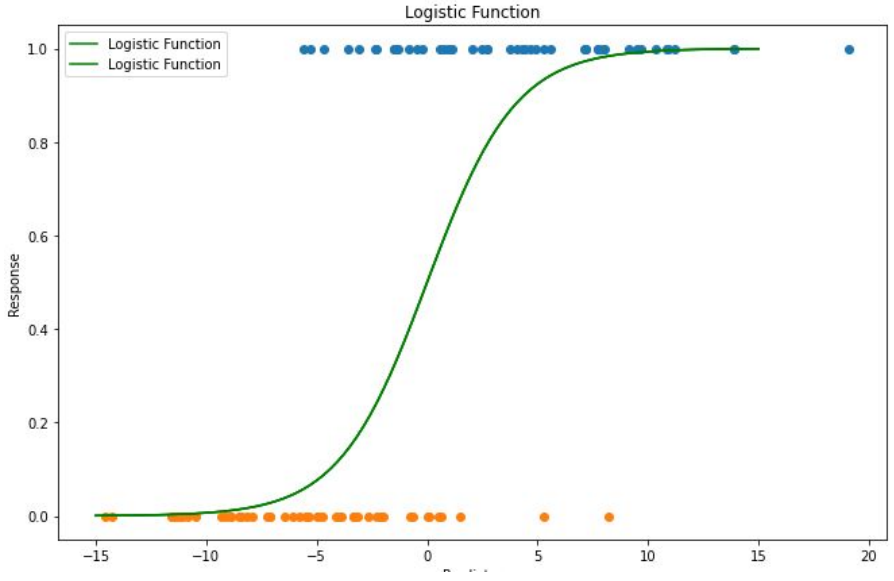


Fig. 14. Logistic Function

The logistic function is an S-shaped curve that changes its slope and path based on its coefficient value. Now the question is, how do we use the logistic function to model our image capture data? for each predictor variable we use, we optimize the corresponding coefficients (α, β, etc.) to best fit the data.

3.3. Expected Goals Model

Our goal is to create a model to describe the existing data as accurately as possible, and ultimately predict future events. Before applying logistic regression to our entire data set, we must divide the data into training and test sets. The training set works on the data where we build the model, and the test set is the data we use to evaluate how well our model is performing. Matching the training data to the logistics function will generate coefficients for our predictors. If we start by adjusting only the distance variable, we have to arrive at the optimal parameters to describe the training data.

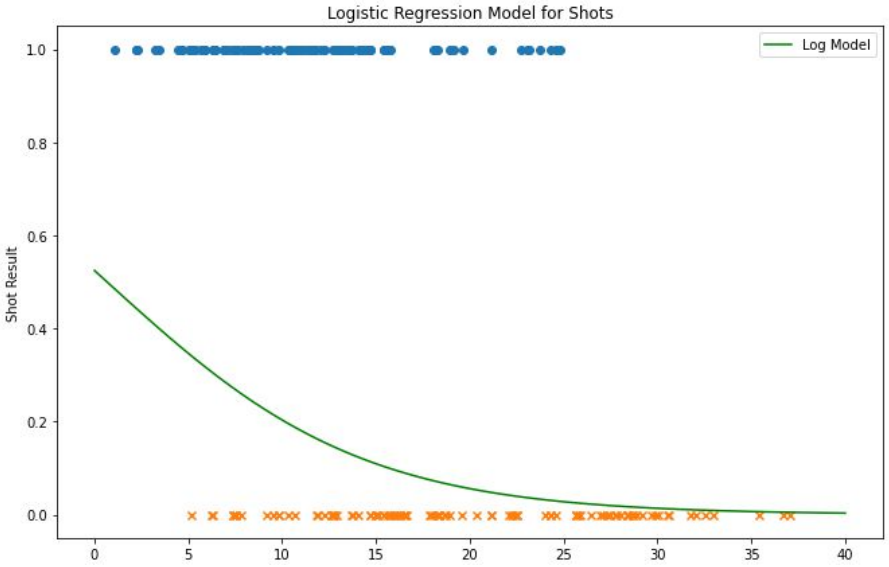


Fig. 15. Logistic Regression Model for Shots

Although the graph above yields several certifications for which our coefficients produce a reasonable match, it is not very clear graphically how well this match is, especially considering this is only a small sample of our training data. Instead of measuring the goodness of conformity numerically, let's examine another graph. As we built in Part I, we were able to collect data based on distance, calculate the ratio of shots to one goal per bin, and then spread the trash plots onto a graph. If we now put the logistics model into a scatter plot, we can see how well the function maps the data.

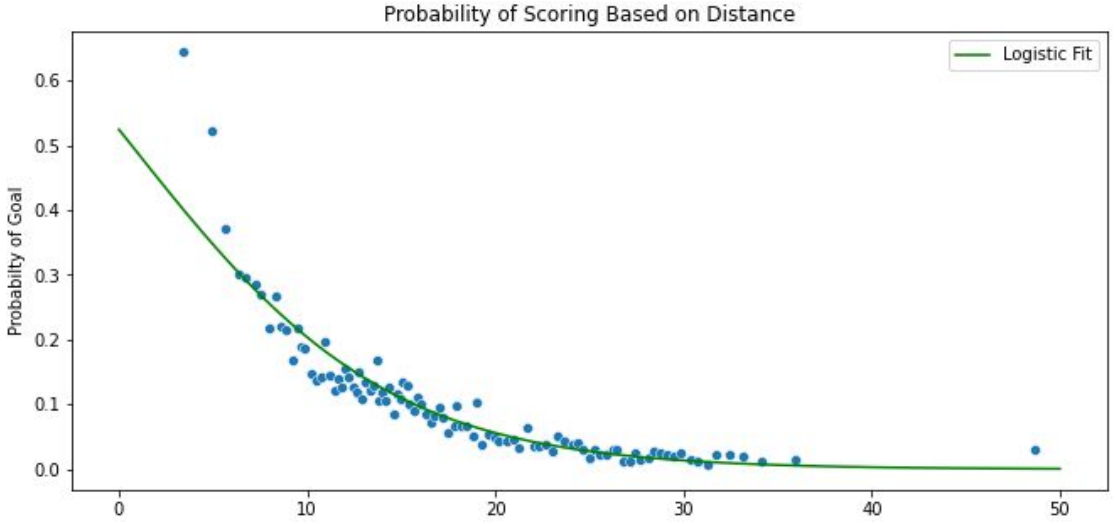


Fig. 16. Probability of Scoring Based on Distance

Please note that we didn't adapt our model to the bullet points in the image above but rather the 32,000 shots in our training set. The purpose of this chart is to measure where our models are performing well and which are not. Plotting 32,000 points on a graph isn't great for visualization, so we decided to plot a sample representation of the population. we can see from figure 7 that the model predicts data well for values greater than 6 meters, but the model underestimates the probability of targeting from a closer distance. This is the kind of advantage the graphic approach will offer. We can try to better predict the shots that are closer to the goal by adding the squared term to the logistics function. The next thing is we have to use the pipeline function from sklearn to the mesh of the polynomial function with logistic regression.

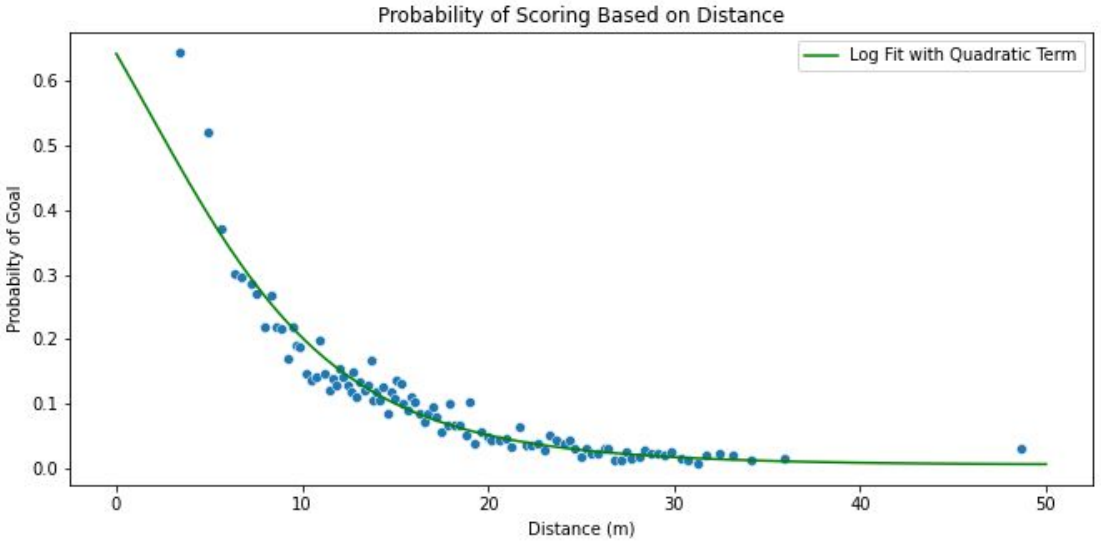


Fig. 17. Probability of Scoring Based on Distance with Quadratic Term

We find it to be an improvement. If we add the squared term to the distance variable we do a much better job at predicting shots that are close to goal. Numerical evaluations do not offer us this luxury. This kind of analysis is more of an art than a science, so it's important to play around with the different options. Of course this is just an ad hoc method of evaluating our model, for the purpose of graphically representing how the model compares to the data. We'll be studying a more concrete method for assessing the accuracy of a model shortly, but let's first add an angle predictor to the mix. As you can see in the figure below, this represents the simplest expected goal model. If we go back to part one and compare the two-dimensional density plot with the xG model, we see that our two-parameter model fits the probability distribution to a reasonable extent.

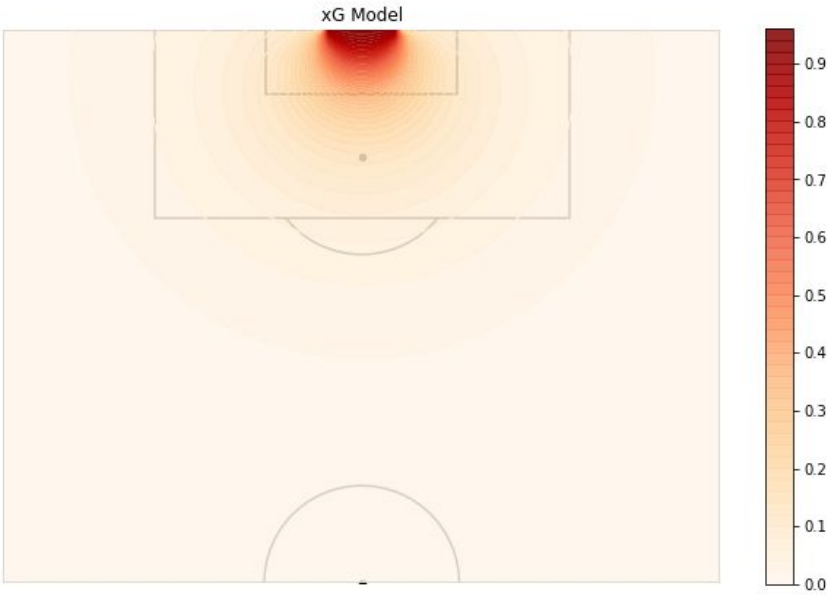


Fig. 18. xG Model

Now for a closer look at the contour plot in the image below, This is where we need to invoke football knowledge. While the probability values in the center position resemble the density plots from Part I, they are too large for the small corner positions near the goal line. This is an obvious drawback with this simplified model. However, it represents an opportunity to experiment with a variety of possibilities and variables. we can choose to add polynomial and interaction terms or add other variables at the same time.

4. Results and Model Evaluation

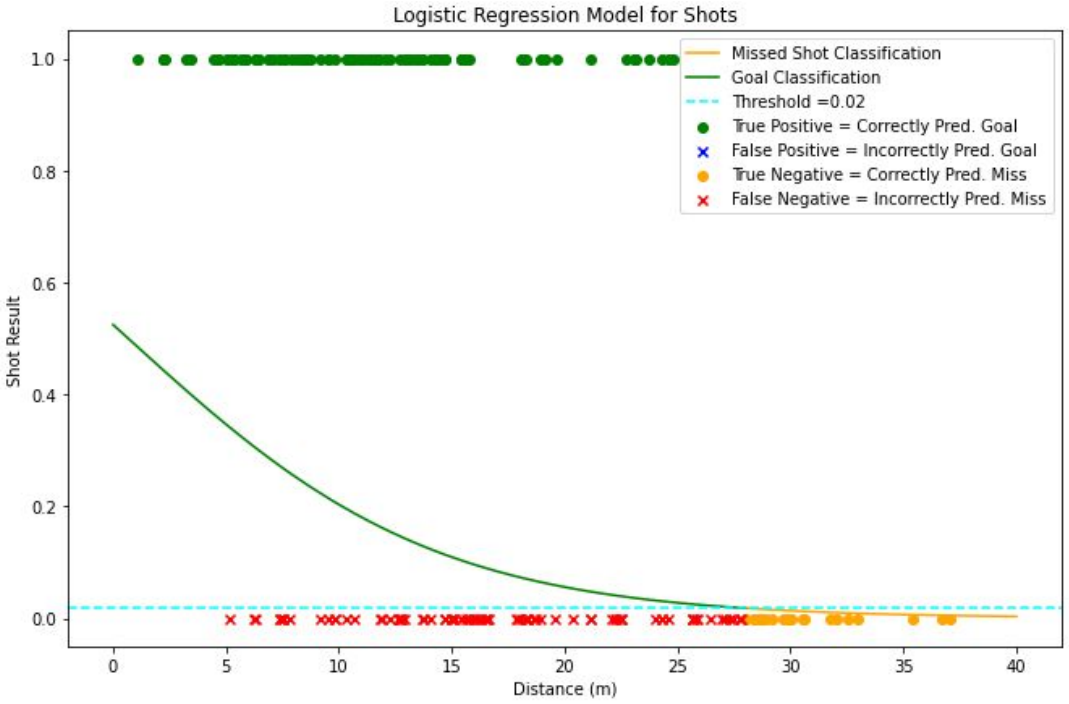


Fig. 19. Logistic Regression Model for Shots

After we have finished building the model, we want to assess the accuracy of our model, we have to test how well it can predict future events. But this raises other concerns. How do we classify the shots based on the expected new target model? In contrast to the Heaviside function, which provides a strict classification, a logistic regression model returns the likelihood of shots resulting in goals. To make a classification, we have to define a threshold. This threshold essentially divides the logistical function, setting goals for which the model falls above the threshold and misses below it. For example, With a threshold of 0.3, the logistic model yields four possible classifications, which can be summarized in a confusion matrix.

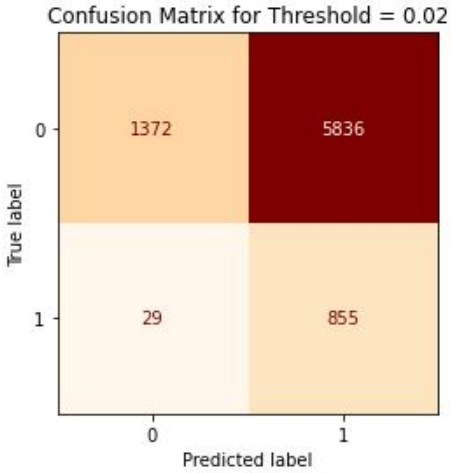


Fig. 20. Confusion Matrix for Threshold 0.2

From the above confusion matrix, we can collect which classifications the model provide promising predictions for and which don't. For the example above, the model does a great job predicting errors but a poor job of predicting objectives, and this is understandable if we examine where the threshold intersects the model. We can determine the model's ability to predict the objectives correctly with a matrix called sensitivity and the model's ability to correctly predict errors is given by specificity.

$$Sensitivity = \frac{TruePositives}{TruePositives + FalseNegatives}$$

$$Specificity = \frac{TrueNegatives}{FalsePositives + TrueNegatives}$$

For the threshold values we checked above, the model will generate the following confusion matrix across the test data, the sensitivity itself, the model's ability to correctly identify the shot that resulted in the goal. On the other hand, specificity is the model's ability to pinpoint non-goal shots. the model yields sensitivity = 0.9671945701357466 & specificity = 0.19034406215316316. Now that we can predict goals and miss on a more balanced level.

If we use a logistic regression model to identify if a patient has cancer, then we will adopt a high threshold, to ignore high specificity in favor of high sensitivity. We prefer to give false positives over false negatives. We don't want someone to walk away thinking they passed the cancer screening when in fact they have cancer. When we try to model goal probability, we don't have such a preference. In fact, trying to predict goals straight away is not very useful to start with. As we will see in the next section, the power of the expected goal model does not lie in making predictions for one shot. So what is the purpose of investigating threshold and confusion matrices? we can use it to compare different models using something known as a Receiver Operator Characteristic (ROC) graph.

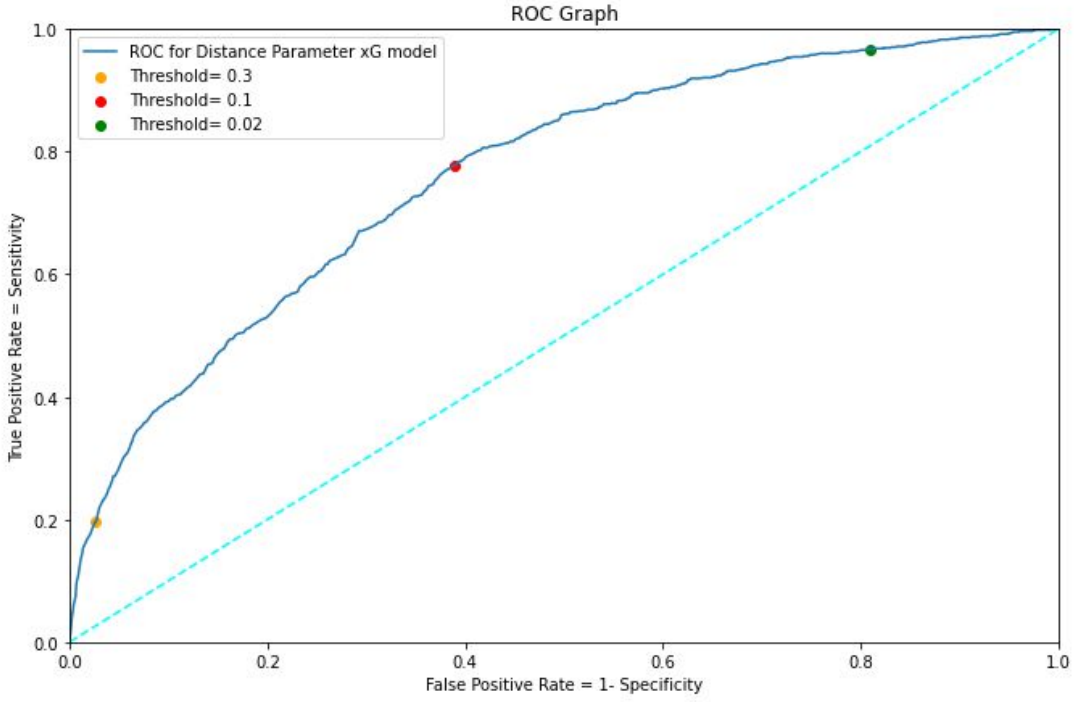


Fig. 21. ROC Graph for Distance Parameter xG Model

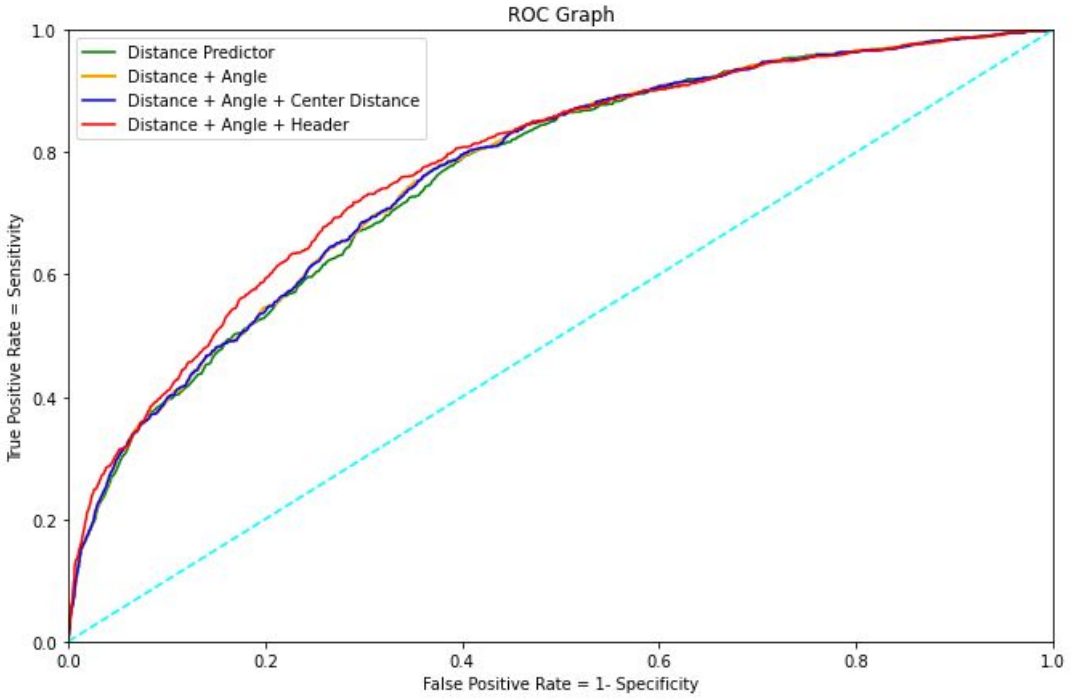


Fig. 22. ROC Model for Distance Predictor

Using a small step size, the ROC curve plots the model's ability to correctly predict objectives against its ability to correctly predict misses for different threshold values. As you move up the y-axis, the model better predicts the goal, and as we move left along the x-axis, the model predicts misses better. It basically maps out the trade-offs between predicting goals and predicting mistakes. The dotted line represents a model that has no predictive power and is basically useless because for every correct classification the model also predicts an incorrect classification. Therefore, the further our ROC curve is from the 45 degree line, the better overall job it is in classifying the test data. Another way to look at it is that the larger the area under the curve, the better our model will describe the test data. This is useful to us because we can use it to compare different models and see if there is a substantial advantage to adding more variables to our model.

Now we have something concrete to do when assessing our model. If we take a closer look, the model with distance and angle as input variables produces the same area under the curve (AUC) as the same model but with the "distance to center" parameter added. So, contrary to some of the assumptions we made earlier, the "distance to center" predictor didn't add much to the performance of our model and we had to exclude it. While the "distance to center" parameter shows a slight change to the model in the area close to the goal line, this change doesn't mean much as only a few shots were taken from that position. Therefore the AUC tells us that adding the center to the distance variable is somewhat useless. Note that we could easily use the p-value to learn whether the center-to-distance parameter is useful for our model.

Another advantage of the ROC curve is that if we choose to use other classification techniques, such as SVM, random forest, neural network, etc., we can compare and compare the performance of the model with the one we have built here. We think this is an exaggeration for the problem at hand, but it is an avenue for further exploration. While this alternative model would produce similar results, logistic regression provided us with a fairly simple and digestible method for describing the snap result, whereas this other method required a deeper and more sophisticated understanding of machine learning techniques.

5. Conclusion

This study has demonstrated the value and reliability xG has in professional football. The distance and angle variables are considered together to have a greater impact on calculating xG than distance as the variable alone. There may not be a direct practical application of this method available yet, but it can be incorporated into practice

(attacking and defending) to aid players' understanding and needs about the game. For example, how the attacking player takes a particular shot and how the defender must be positioned to defend this shot. Goals, as we discussed earlier, are largely random. we need to remember that each shot is unique, made up of hundreds of different variables that we've tried to model using only three. For that we hope to predict the destination with certainty or ending it. Some of this is due to our inability to model all of these variables. Providing that although distance and angle give us a good understanding of the likelihood of a goal-scoring shot, we do not take into account which one is the goalkeeper, if a shot is taken with a weak foot or strong leg, a high shot at a certain point of contact, the state of play, the home advantage , if there are many objects between the goal and the shot, etc. These are just a handful of the measured variables that have an effect. There are also variables that are not easily measured: field conditions (such as on a rough court or carpet), role play, morale and confidence, and other soft factors. Even after ignoring other variables, there is also some degree of intrinsic randomness to image capture. We have tried to model a very complex situation with a simple 3 parameter model; we shouldn't expect high R-squared values. Despite this, xG is far from useless and has in fact been revolutionary in the way we were before games. However, a caution to note is that these methods have not been examined in more detail regarding the different types of shots, passes, through balls and set pieces.

References

[1] A. Rathke, “An examination of expected goals and shot efficiency in soccer,” *J. Hum. Sport Exerc.*, vol. 12, no. Proc2, 2017, doi: 10.14198/jhse.2017.12.proc2.05.

[2] P. Lucey, A. Bialkowski, M. Monfort, P. Carr, and I. Matthews, “‘Quality vs Quantity’: Improved Shot Prediction in Soccer using Strategic Features from Spatiotemporal Data,” *Proc. 8th Annu. MIT Sloan Sport. Anal. Conf.*, pp. 1–9, 2014, [Online]. Available: http://www.sloansportsconference.com/?p=15790.

[3] O. Hubáček, G. Šourek, and F. Železný, “Deep learning from spatial relations for soccer pass prediction,” *CEUR Workshop Proc.*, vol. 2284, pp. 162–169, 2018.

[4] L. Combining, A. Segmentationshape, A. For, R. Fruit, and D. M. W. Hannan, “a M Achine V Ision a Lgorithm Combining Adaptive,” *2017 IEEE Int. Conf. Power, Control. Signals Instrum. Eng.*, vol. XI, no. 2001, pp. 1–17, 2009.

[5] H. M. Zawbaa, N. El-Bendary, A. E. Hassanien, and T. hoon Kim, “Event detection based approach for soccer video summarization using machine learning,” *Int. J. Multimed. Ubiquitous Eng.*, vol. 7, no. 2, pp. 63–80, 2012.

[6] B. M. Faria, L. P. Reis, N. Lau, and G. Castillo, “Machine Learning algorithms applied to the classification of robotic soccer formations and opponent teams,” *2010 IEEE Conf. Cybern. Intell. Syst. CIS 2010*, pp. 344–349, 2010, doi: 10.1109/ICCIS.2010.5518540.

[7] X. X. Xu *et al.*, “Research on the optimal design of soccer robot based on the mechanical analysis,” *Adv. Mater. Res.*, vol. 1049–1050, no. Mmebc, pp. 1033–1037, 2014, doi: 10.4028/www.scientific.net/AMR.1049-1050.1033.

[8] F. da S. L. Cardoso, S. González-Víllora, J. Guilherme, and I. Teoldo, “Young Soccer Players With Higher Tactical Knowledge Display Lower Cognitive Effort,” *Percept. Mot. Skills*, vol. 126, no. 3, pp. 499–514, 2019, doi: 10.1177/0031512519826437.

[9] D. Bao, H. Wang, B. Fang, and L. Li, “HfutEngine3D Soccer Simulation Team Description 2008,” *Sci. Technol.*, 2008.

[10] J. Carlos Núñez and B. Dagnino, “Exploring the application of soccer mathematical models to game generation on a simulated environment,” pp. 1–10, 2020, [Online]. Available: www.metrica-sports.com.

[11] K. Pelechrinis and W. Winston, “A Skellam regression model for quantifying positional value in soccer,” *J. Quant. Anal. Sport.*, pp. 1–31, 2021, doi: 10.1515/jqas-2019-0122.

[12] H. Igarashi, K. Nakamura, and S. Ishihara, "Learning of soccer player agents using a policy gradient method: Coordination between kicker and receiver during free kicks," *Proc. Int. Jt. Conf. Neural Networks*, no. 2, pp. 46–52, 2008, doi: 10.1109/IJCNN.2008.4633765.

[13] B. Larrousse, "Improving decision making for shots," *StatsBomb Innov. Footb. Conf. 2019*, pp. 1–15, 2019.

[14] H. I. Ayman, "DEVELOPING A COOPERATIVE BEHAVIOR FOR MULTI AGENTS SYSTEM," no. July, 2007.

[15] J. Bu, S. Lao, and L. Bai, "Automatic line mark recognition and its application in camera calibration in soccer video," *Proc. - IEEE Int. Conf. Multimed. Expo*, 2011, doi: 10.1109/ICME.2011.6012137.

[16] M. Van Roy, P. Robberechts, T. Decroos, and J. Davis, "Valuing On-the-Ball Actions in Soccer: A Critical Comparison of xT and VAEP," 2019, [Online]. Available: www.aaai.org.

[17] J. H. Kim, "An Analysis of Comparison on Performances in Soccer Attacking-Third," *Korean J. Sport Sci.*, vol. 24, no. 4, pp. 653–661, 2013, doi: 10.24985/kjss.2013.24.4.653.

[18] J. Fernández, L. Bornn, and D. Cervone, "A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions," *arXiv*, 2020.

[19] M. Lauer, S. Lange, and M. Riedmiller, "Calculating the perfect match: An efficient and accurate approach for robot self-localization," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4020 LNAI, no. c, pp. 142–153, 2006, doi: 10.1007/11780519_13.

[20] Y. Liu, Q. Huang, Q. Ye, and W. Gao, "A new method to calculate the camera focusing area and player position on playfield in soccer video," *Vis. Commun. Image Process. 2005*, vol. 5960, p. 59604H, 2005, doi: 10.1117/12.632721.

[21] A. N. Fitriana, K. Mutijarsa, and W. Adiprawita, "Color-based segmentation and feature detection for ball and goal post on mobile soccer robot game field," *2016 Int. Conf. Inf. Technol. Syst. Innov. ICITSI 2016 - Proc.*, pp. 1–4, 2017, doi: 10.1109/ICITSI.2016.7858232.

[22] C. Lago-Peñas, J. Lago-Ballesteros, and E. Rey, "Differences in performance indicators between winning and losing teams in the UEFA Champions League," *J. Hum. Kinet.*, vol. 27, no. 1, pp. 135–146, 2011, doi: 10.2478/v10078-011-0011-3.