

Group 4

Maxwell Gerhart, Rohan Sampath, Jaikhush Thakkar, and Ryan Kaye

Our goal for the project was to create a model to calculate the probability of a shot resulting in a goal. In soccer this metric is commonly known as expected goals (xG for short). We plan on creating three models: a logistic regression model, Generalized Additive Model (GAM), and random forest model. Once we select the best model, we will put this model to use by calculating the expected points of an entire league season. Calculating expected points requires running simulations based on the shot xG values in order to calculate the proportion of wins, losses, and draws for each team.

$$\text{Expected Points} = 3 \times \text{Win Proportion} + 1 \times \text{Draw Proportion} + 0 \times \text{Loss Proportion}$$

This lets us determine what teams performed as expected and which did better or worse. The reason for the variance could be caused by many factors, but over time teams should regress towards their expected value.

The data is open to the public through the StatsBombR package in R. It includes detailed information about different competitions and the matches and events inside each competition. For our purposes we will filter down the event level data down to only shots. There are many columns within the event dataframe that don't apply to shots so we can remove those.

Logistic

For the initial approach we began with a Logistic Regression model which is an extremely commonly used model for prediction tasks. We wanted to create a baseline model that would help us predict the expected goals scored in a total match by focusing on calculating the probability of a shot resulting in a goal. This is a great match because Logistic Regression works with binary factors, and in our case each shot would be either a Goal or Not a goal depending on the outcome of the shot.

In this case the way that we estimate with the model is:

$$\begin{aligned}
\text{logit}(P(\text{goal} = 1)) = & \beta_0 + \beta_1 \text{ distance} + \beta_2 \text{ angle} \\
& + \beta_3 \text{ LeftFoot} + \beta_4 \text{ OtherBodyPart} + \beta_5 \text{ RightFoot} \\
& + \beta_6 \text{ DivingHeader} + \beta_7 \text{ HalfVolley} + \beta_8 \text{ Lob} + \beta_9 \text{ NormalTechnique} \\
& + \beta_{10} \text{ OverheadKick} + \beta_{11} \text{ Volley} \\
& + \beta_{12} \text{ FromCounter} + \beta_{13} \text{ FromFreeKick} + \beta_{14} \text{ FromGoalKick} \\
& + \beta_{15} \text{ FromKeeper} + \beta_{16} \text{ FromKickOff} + \beta_{17} \text{ FromThrowIn} \\
& + \beta_{18} \text{ OtherPlayPattern} + \beta_{19} \text{ RegularPlay}.
\end{aligned}$$

This is the model we have seen many times in class, and follows the basic commonly used pattern for logistic regression tasks. The incorporation of these variables works to help us to estimate and understand how each characteristic affects the probability of a shot resulting in a goal. For example the numerical variables such as the location variables work to tell us the coordinates of where on the field the shot is being taken from. Naturally, a shot attempt right in front of the goal is far more likely to result in a goal compared to a shot taken from the midfield line. Similarly the angle in which the shot was taken has a major influence on the prediction we decide on. When looking at the categorical predictors such as body parts or under pressure they can provide additional context for the model. We fit and train the model model by using the historical data from the Bundesliga, Serie A, Ligue 1, and La Liga.

Area under the curve: 0.7964

GAM

For our experimental modeling approach, we used a Generalized Additive Model (GAM) to estimate the probability that a shot results in a goal. GAMs extend logistic regression by allowing the model to learn smooth nonlinear relationships between predictors and a response variable. Unlike standard linear or logistic regression, where each variable contributes linearly, a GAM fits flexible spline functions to variable shots such as shot distance and shooting angle. This is useful for soccer analytics because the relationship between shot placement and scoring probability is well known to be nonlinear.

Formally, the model estimates:

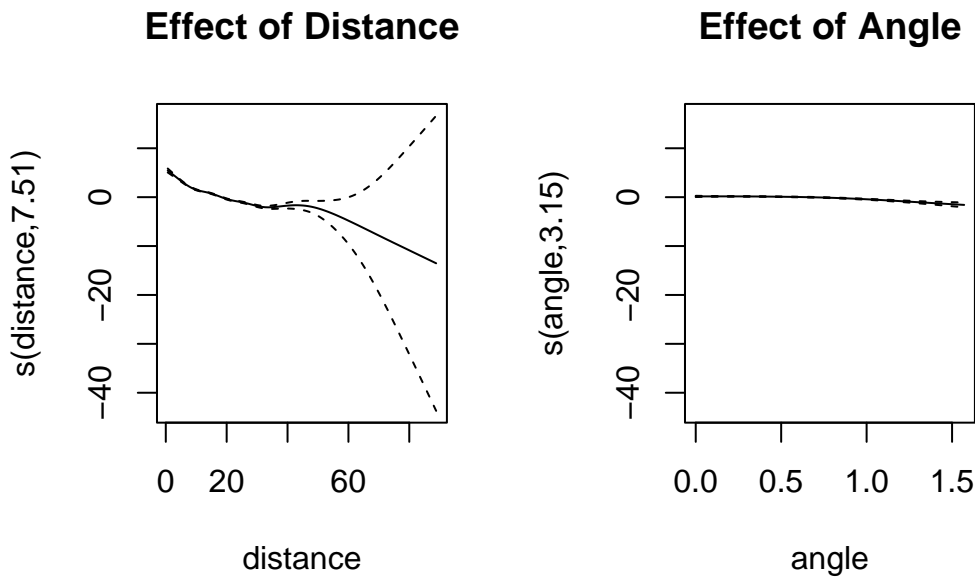
In this formulation, distance and angle are modeled using smooth spline functions $s(\cdot)$, while under pressure, shot body part, shot technique, and play pattern enter the model as standard categorical (parametric) terms through dummy variables.

We chose GAM because:

- Shot distance and angle do not influence scoring in a straight-line manner, and GAM can capture this automatically.
- It balances interpretability and flexibility, making it easy to visualize how each feature affects scoring probability.
- GAM is commonly used in modern soccer analytics as a baseline for xG models.
- It satisfies the project requirement to use at least one method beyond/outside standard class material.

We trained the GAM on shots from four major European leagues in the 2015/16 season and evaluated it on a held-out test dataset using AUC as the primary performance metric.

Area under the curve: 0.7993



The GAM model produced an **AUC of 0.8164** on the test set, which shows strong predictive performance. An AUC above 0.80 suggests that the model successfully distinguishes between shots that become goals and those that do not. This aligns with expectations for simple xG models and confirms that our features, especially distance carry important information about shot quality.

The smooth terms in the GAM reveal clear relationships:

- **Distance** has the strongest effect on scoring probability. The predicted probability declines sharply as the shot is taken farther from goal, which is consistent with established soccer analytics. The GAM curve shows a steep decrease after roughly 20–25 yards and extremely low probabilities beyond 40+ yards. This validates distance as a key driver of expected goals.
- **Angle** exhibits a weaker but still relevant effect. The model shows a slight decline in scoring probability as the shot becomes more narrow in angle. This pattern is expected given our simplified angle formulation; nevertheless, it captures the idea that central shots tend to be more dangerous. More advanced angle features could strengthen this relationship, but for our purposes the effect is reasonable.

Our categorical predictors, **body part, technique, play pattern, and under pressure** were included as linear terms. These variables help adjust the baseline probability upward or downward depending on the context of the shot. For example, headers typically have lower scoring probabilities than shots with the foot, and shots taken under pressure are generally more difficult.

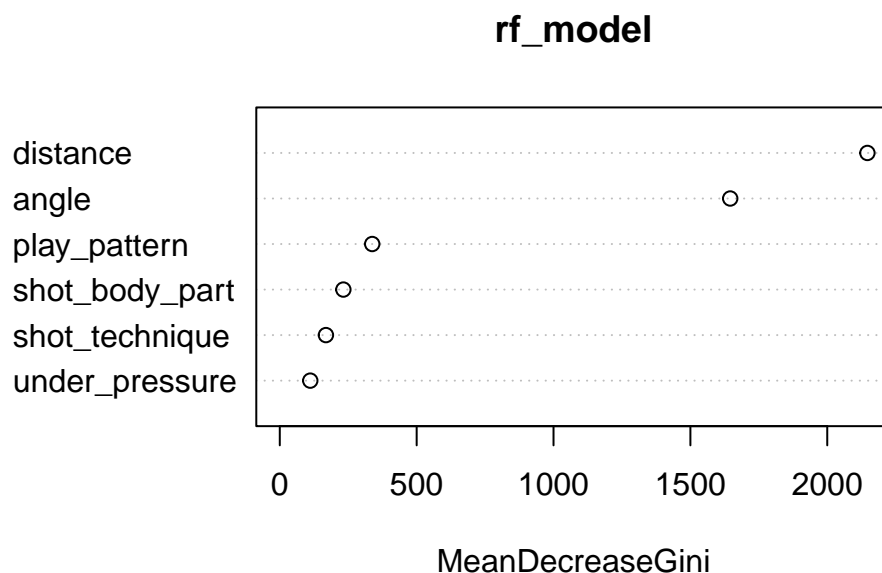
The partial-effect plots generated by the GAM provide interpretable visual summaries of these patterns. The distance plot clearly shows the nonlinear relationship between distance and goal probability, while the angle plot highlights the role angle plays in our simplified model. Together, these results demonstrate that the GAM effectively captures realistic shot-scoring dynamics and forms a strong foundation for comparison with more advanced models like Random Forest or XGBoost.

Random Forest

For the last model here we wanted to employ a Random Forest classifier to help with our prediction goals. We maintain our goals with the task of this ensemble model with the use of many decision trees, and finding the average of our results. This is a good fit for this problem because we have many predictors and variables, this way we prevent overfitting issues and improve accuracy compared to a single decision tree. Similarly to GAM this works to help capture nonlinear patterns and interactions.

We still employ the same predictors as aforementioned, maintaining consistency and reproducibility in our approach. For random forest we also made sure to convert to factors to appropriately run the model. That way we have a predictor for not a goal and goal scored, helping to effectively create our xG. We used 500 trees in our model, and kept the mtry value to 3.

Area under the curve: 0.7528



The Random Forest Model produced an **AUC of 0.7762**. This is a relatively strong predictor, showing the model does well when trying to predict whether or not a goal is scored. Although it is not on the level we achieved with GAM, we can still see that the model can meaningfully differentiate between a good and a poor shot.

Furthermore, when taking a deeper dive into our visualization, which we created above, we can analyze the importance of each predictor. As you would naturally assume, the most important factor was the distance, which is how far away the shot is taken from the location of the goal. Then, not too far behind, was the angle at which the shot was taken towards the goal. This makes sense because a shot taken heavily from one side would allow the goalkeeper to position themselves in a manner where they are essentially covering all the area the ball could travel to based on the angle.

Overall, the use of the Random Forest was successful and allowed us to compute xG values in an accurate and meaningful way. We were able to compute values for each shot that align with historical data, and have a quality accuracy measurement in AUC. Although it was not the best performance, it is reliable and a good tool to have when comparing between models.

Model Selection

League Table

After fitting our xG model, we applied it to all Premier League matches in the 2015–16 season to estimate each team's expected points (xPTS). Expected points reflect the average number of points a team should have earned given the quality of shots they created and conceded, assuming a match is replayed many times using our xG-driven simulation procedure.

For each match, we simulated 10,000 outcomes using the predicted xG for every shot, calculated team-level win/draw/loss probabilities, and translated those into expected points using the standard scoring rule:

Aggregating across all matches yields a full model-based Premier League table, which we compared to the actual results.

Premier League 2015-16 Season

Actual vs Expected Performance

	Team	W	D	L	GF	GA	PTS	xG	xGA	xPTS
1	Leicester City	23	12	3	68	36	81	70.7	52.9	65.1
2	Arsenal	20	11	7	65	36	71	73.3	36.7	75.3
3	Tottenham Hotspur	19	13	6	69	35	70	63.5	44.8	64.0
4	Manchester City	19	9	10	71	41	66	69.3	35.6	72.8
5	Manchester United	19	9	10	49	35	66	46.0	43.4	54.1
6	Southampton	18	9	11	59	41	63	57.7	45.2	60.7
7	West Ham United	16	14	8	65	51	62	61.5	59.1	53.5
8	Liverpool	16	12	10	63	50	60	60.1	41.9	63.5
9	Stoke City	14	9	15	41	55	51	40.8	56.8	43.3
10	Chelsea	12	14	12	59	53	50	59.0	49.3	59.0
11	Everton	11	14	13	59	55	47	54.3	60.3	50.3
12	Swansea City	12	11	15	42	52	47	44.2	56.6	43.4
13	Watford	12	9	17	40	50	45	46.1	55.9	47.2
14	West Bromwich Albion	10	13	15	34	48	43	43.8	54.1	44.5
15	AFC Bournemouth	11	9	18	45	67	42	45.4	50.8	50.1
16	Crystal Palace	11	9	18	39	51	42	48.5	57.6	46.2
17	Sunderland	9	12	17	48	62	39	46.1	65.5	39.9
18	Newcastle United	9	10	19	44	65	37	43.4	59.7	43.1
19	Norwich City	9	7	22	39	67	34	43.5	61.3	41.7
20	Aston Villa	3	8	27	27	76	17	32.4	62.0	32.4

Key Insights

- Teams that overperformed their xPTS gained more points than our model predicted based on shot quality alone. These differences may be attributed to finishing skill, goalkeeper performance, tactical game states, or random variation.
- Teams that underperformed their xPTS likely created enough high-quality chances but finished poorly or conceded goals at an unsustainably high rate.

For example, teams like Leicester City (who won the league in 2015–16) typically show notable overperformance relative to xPTS due to a combination of elite finishing, efficient defending, and strong tactical coherence. Conversely, teams that performed below expectation may have been unlucky or lacked clinical finishing.

Our final table includes each team’s actual points (PTS), expected points (xPTS), expected goals for (xG), and expected goals against (xGA). This allows us to diagnose where teams deviated from expectation whether through attack, defense, or both.

Interpretation

- xG and xGA summarize underlying performance, focusing on process rather than outcomes.
- xPTS contextualizes these performances in terms of league standings, showing what the table would look like if results depended only on the quality of chances.

This approach is widely used in modern soccer analytics and helps identify teams whose results are driven by variance rather than sustainable underlying play. Our results generally align with trendlines observed in professional models such as Understat’s xPTS tables, demonstrating that our modeling and simulation framework captures meaningful elements of team performance.

Shot Map

To complement our modeling results, we created detailed shot maps for two Premier League clubs from the 2015–16 season: Liverpool and Leicester City. Each plot visualizes every shot taken by the team across the entire league campaign and is sized and shaded according to the predicted xG from our model. These visualizations help us understand how teams generated their chances, where those chances occurred on the pitch, and whether our model assigns reasonable probabilities to different shooting situations.

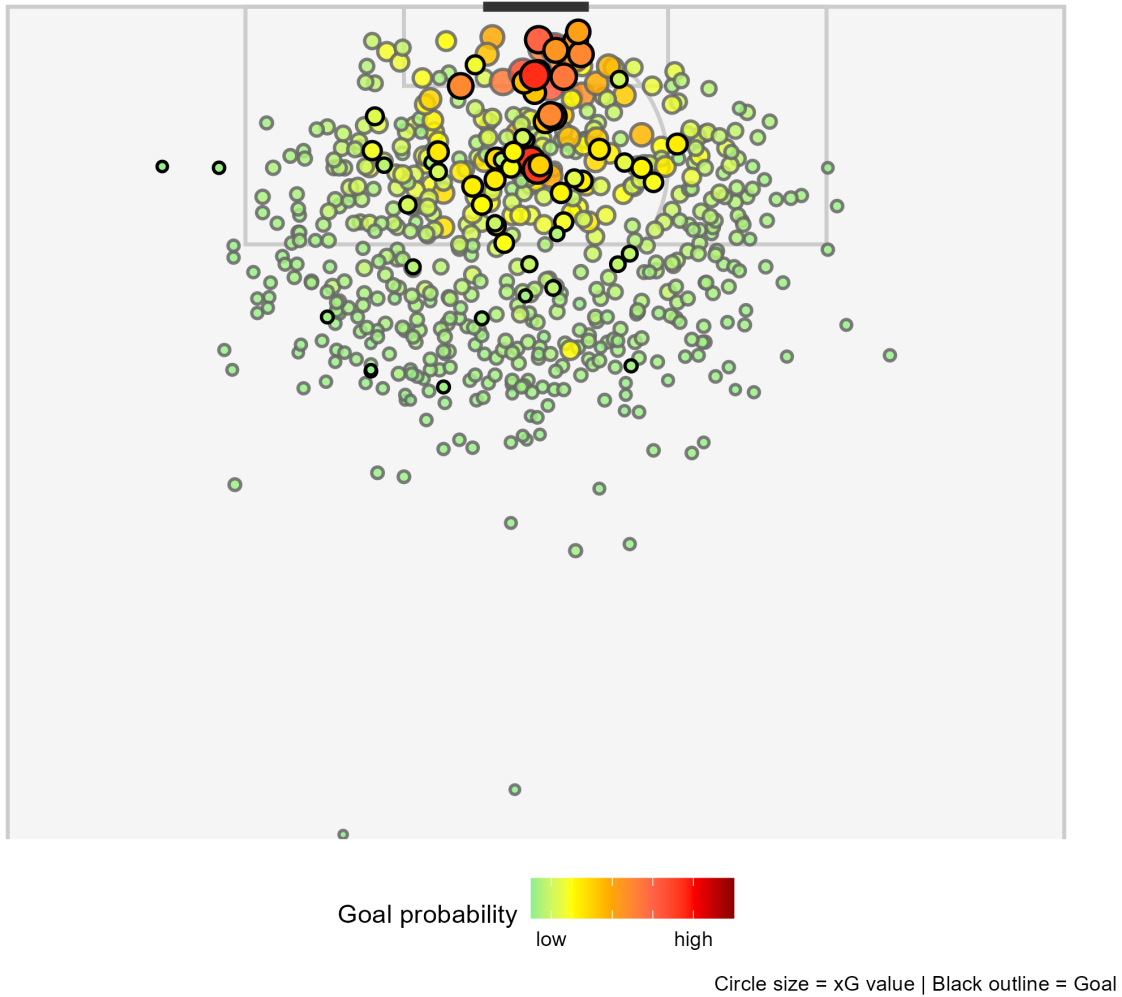
We used a standardized pitch template and colored the points by team identity (red for Liverpool, blue for Leicester City). Larger circles represent higher-quality chances, while lighter/smaller circles correspond to low-probability shots. Because the shot maps reflect the predictions from our logistic regression xG model, they also serve as a qualitative diagnostic tool for model behavior.

Liverpool showed a high volume of shots clustered inside the box, especially near the central channels. Their chance profile reflects a team generating sustained pressure and short-range shooting opportunities, which aligns well with their historical playing style during the 2015–16 season. The model appropriately assigns higher xG values to these closer-range shots, which appear as larger red circles.

Leicester City, meanwhile, display a more compact but efficient shot map. Many of their high-xG chances occur slightly deeper but still in central, high-value areas. This aligns with their counterattacking style during their title-winning season, often creating clear chances with fewer overall shots. Their shot map shows fewer circles overall, but a larger proportion of them are medium-to-high xG opportunities.

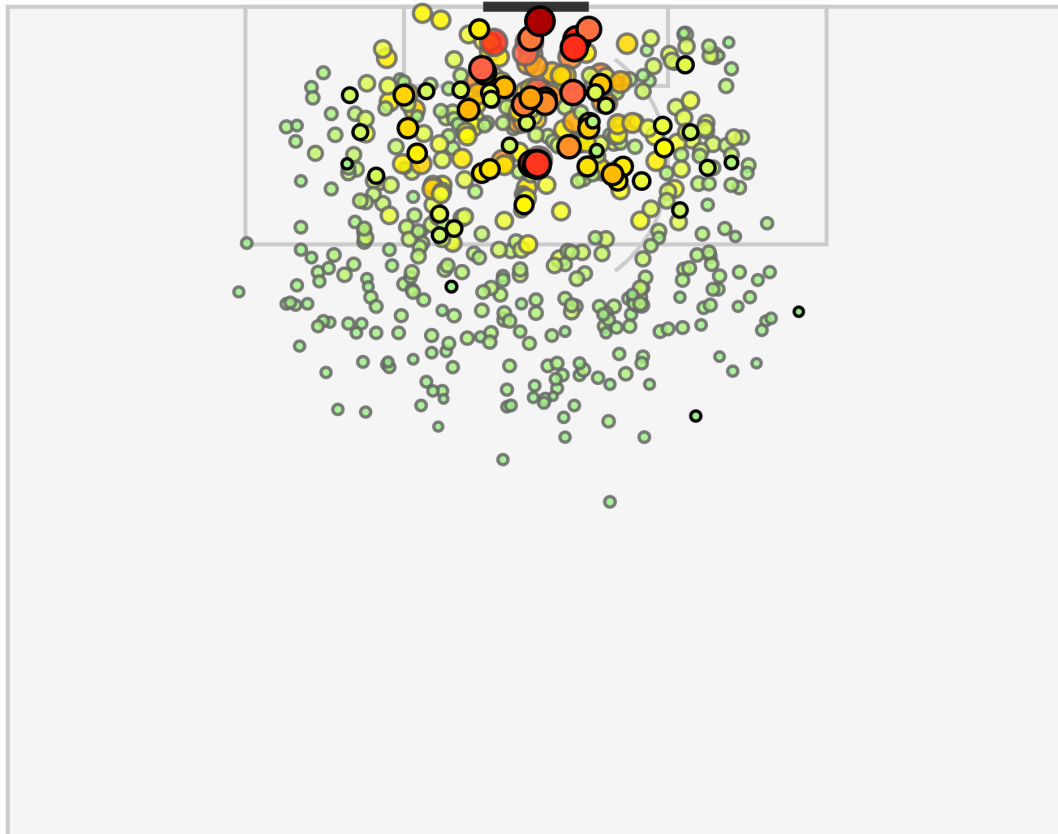
Liverpool – 2015-16 Season

Expected Goals: 60.08 (62 goals / 635 shots)



Leicester City – 2015-16 Season

Expected Goals: 70.66 (67 goals / 525 shots)



Goal probability 
low high

Circle size = xG value | Black outline = Goal

These visual patterns validate two important insights:

1. The xG model produces reasonable spatial patterns, giving higher probabilities to close-range, high-angle shots.
2. The two teams' tactical styles emerge naturally from the visualizations heavy-volume creation for Liverpool vs. selective, high-quality chances for Leicester.

Overall, the shot maps serve as a visual confirmation that our model behaves consistently with real soccer strategy, and they illustrate meaningful differences in chance creation between teams.

Code Appendix