Peer-Review

Singh, Shriyansh. 2025. “Improving Expected Goals (xG) Models: Methods, Results, and Novel Extensions.” *Journal of High School Science* 9 (3): 475–85. https://doi.org/10.64336/001c.144180.

Good work. I enjoyed reading the manuscript and I think it contributes meaningfully to the existing corpus of knowledge. I have some concerns that I list below.

1.Your sparse description does not do justice to the considerable amount of work that you have done. I would like you to expand your reporting of methods such that your model can be reproduced by anyone reading your paper. This includes making your database public and providing links to a reputable repository such as GitHub. Please also provide access to the .csv file with the freezeframes and the python algorithm. This will significantly increase adoption and enable your work to be reproduced or improved upon. Please also provide a link to your stream app.

2.I did not understand what the difference is between “absolute goal diff” and “goal difference” in Figure 2. Please explain. Also explain what the difference is between “distance” and “gk distance” ? Also provide explaination to interpret this chart. For example, the larger the gk distance, the more the probability of a goal? On the other hand, the smaller the ‘distance’ the more the probability of a goal? Include a table with what these features represent as well as their *quantitative* contribution to the outcome .

3.Contrary to your claims, the “is home” and “n prev passes” do not seem to contribute much to the shot outcome (their SHAP values are less than absolute 1); or am I not interpreting this correctly? That is why a quantitative contribution of each factor to shot outcome will be helpful (see point 2).

4.SHAP assumes feature independence; i.e. no collinearity. Did you perform a multicollinearity analysis - such as a heat map for instance - among your different features? For example, intuitively, it would seem to me that the larger the goal difference, the lesser would be the n-previous-passes because the losing team would have a sense of urgency. Similarly, the larger the goal difference, the larger would be the distance to goal at which shots are attempted; again due to the urgency of the situation. In this case, are you not essentially incorporating redundant variables in the model? Please describe and explain in the manuscript.

5.Again, can you explain Figure 1 in more detail? Why am I seeing probabilities and frequencies when the outcome is binary (0 or 1) ?

In summary, the paper is solid but is rough around the edges. Much more description of procedure, detail and discussion will make this exceptional. Please therefore address my comments. I look forward to the improved next iteration.

Good work. I enjoyed reading the manuscript and I think it contributes meaningfully to the existing corpus of knowledge. I have some concerns that I list below.

1. **Your sparse description does not do justice to the considerable amount of work that you have done. I would like you to expand your reporting of methods such that your model can be reproduced by anyone reading your paper. This includes making your database public and providing links to a reputable repository such as GitHub. Please also provide access to the .csv file with the freezeframes and the python algorithm. This will significantly increase adoption and enable your work to be reproduced or improved upon. Please also provide a link to your stream app.**

    a. My Github repo has been linked in the references section of my research paper as requested. Code and files including pre-processed raw data and processed freeze frames have also been attached, all in a separate downloadable zip file accessible in the ‘Manuscript Details’ section, subsection ‘Files’. This includes the .csv files with the freezeframes and python algorithm as well. The Streamlit app has been linked in the references section of my research paper as well. Github repo also contains instructions in a readme file explaining how to clone and reproduce the project from scratch. Public Statsbomb Dataset which was used for training and testing the model has also been attached in the references section.

2. **I did not understand what the difference is between "absolute goal diff" and "goal difference" in Figure 2. Please explain. Also explain what the difference is between "distance" and "gk distance" ? Also provide explaination to interpret this chart. For example, the larger the gk distance, the more the probability of a goal? On the other hand, the smaller the 'distance' the more the probability of a goal? Include a table with what these features represent as well as their *quantitative* contribution to the outcome.**
    a. goal_difference is a signed variable that indicates whether the shooting team is leading or trailing at the time of the shot (for example, +1 means the shooter's team leads by one goal; −1 means they trail by one). It captures directional match-state effects (for instance, teams that are trailing may shoot from different locations or take riskier chances, like you said, in a desperate attempt to claw themselves back into the game). abs_goal_diff is the absolute value of that quantity (i.e., abs(goal_difference)), and reflects the magnitude of the score gap irrespective of which team is ahead. It captures how "balanced" or "blowout" the match is (large absolute differences often correlate with different tactical behaviours, for instance defensive players such as CDMs or even CBs being taken off for more attack-minded options, and players being instructed to focus less on lengthy build-ups from the back, reducing n_prev_passes, and being more direct, vertically stretching the opponent in the process, which results in more through balls and often even crosses if the opposition sits in a deep-lying low block to see out their lead).

       Likewise, distance is the Euclidean distance from the shot location to the goal centre (StatsBomb coordinate system, units of pitch coordinates, expressed in metres, are used). Shorter distance generally makes a goal more likely because the target is simply physically closer.

       On the other hand, gk_distance is the Euclidean distance from the shooter to the goalkeeper (the goalkeeper freeze-frame location closest to the shot). It captures how exposed / out of position the goalkeeper is for that particular shot.

       The SHAP summary plot in Figure 2 is a global interpretability visualisation. Each row is a feature and each point is one shot; the x-axis shows the SHAP value (impact on the model output in log-odds space). Points to the right (positive SHAP) increase the predicted probability of a goal; points to the left (negative SHAP) decrease it. Colour encodes the feature value (red = high value, blue = low value). The width of the violin indicates density of observations at each SHAP value (wider = more shots).

       For example, for distance, red points (high distance) lie more to the left, meaning higher distances decrease predicted goal probability. For gk_distance, red points generally lie to the right, meaning larger goalkeeper separation increases predicted goal probability. For defensive features like defenders_in_5m or angular_pressure, higher values (red) are concentrated on the left, indicating greater defensive pressure reduces a shot's xG. Angle behaves as expected: larger angles (more of the goal in view) shift SHAP rightwards, increasing xG. As requested, a new Table 3 has been added which reports mean absolute SHAP values and each feature's share of the model's average explanatory magnitude (how much it contributes towards influencing shot outcome). All this was computed using code available in the original google colaboratory notebooks found in the github repository of the project which is as mentioned in point 1, now linked in the references section of the research paper.

3. **Contrary to your claims, the "is home" and "n prev passes" do not seem to contribute much to the shot outcome (their SHAP values are less than absolute 1); or am I not interpreting this correctly? That is why a quantitative contribution of each factor to shot outcome will be helpful (see point 2).**
    a. Thank you for this helpful request for clarification. We have revised the manuscript to (1) clarify the difference between goal_difference (signed) and abs_goal_diff (absolute magnitude), and between distance (shot → goal centre) and gk_distance (shot → goalkeeper), and (2) temper our language about the contribution of is_home and n_prev_passes. The SHAP summary now accompanies a new supplementary table (Table 2) which addresses your various comments regarding clarifying the ambiguity associated with some of the variables mentioned in the 2nd point, and another table (Table 3) that reports mean absolute SHAP (importance) and mean signed SHAP (direction) for every feature so readers can see both magnitude and sign quantitatively for each features contribution towards xG. Numerically, distance, angle, angular_pressure, and gk_distance account for the largest share of the model's average explanatory power, while is_home and n_prev_passes have substantially smaller mean absolute SHAP values (i.e. lower quantitative contribution), as you pointed out. We have added a short explanatory caption to Figure 2 describing how to read SHAP units (additive log-odds) and the practical interpretation in probability space. As requested in point 2, a new Table 3 has also been added which reports mean absolute SHAP values and each feature's share of the model's average explanatory magnitude (how much it contributes towards influencing shot outcome).
4. **SHAP assumes feature independence; i.e. no collinearity. Did you perform a multicollinearity analysis - such as a heat map for instance - among your different features? For example, intuitively, it would seem to me that the larger the goal difference, the lesser would be the n-previous-passes because the losing team would have a sense of urgency. Similarly, the larger the goal difference, the larger would be the distance to goal at which shots are attempted; again due to the urgency of the situation. In this case, are you not essentially incorporating redundant variables in the model? Please describe and explain in the manuscript.**
    a. We thank the reviewer for highlighting this important point. We have now conducted a full multicollinearity analysis. Pairwise correlations confirmed substantial overlap among shot geometry variables (e.g., distance ↔ goalkeeper distance, $r = 0.83$; distance ↔ angle, $r = –0.74$). VIF analysis further identified distance (18.6) and goalkeeper distance (14.2) as multicollinear, while all other predictors remained below conventional thresholds.
    To evaluate practical impact, we computed permutation importance, which showed that each correlated feature remained individually important (AUC drops of 0.032–0.103). We also tested redundancy by residualising n_prev_passes on goal_difference; model AUC and SHAP rankings were unchanged (ΔAUC ≈ +0.0005). This suggests that, although collinearity exists, it does not destabilise model inference.
    We have added a new subsection ("Multicollinearity Analysis") to the Methods/Feature Engineering section of the manuscript describing these diagnostics, their results, and the implications. We also acknowledge this issue as a limitation and outline decorrelation strategies as future directions.
5. **Again, can you explain Figure 1 in more detail? Why am I seeing probabilities and frequencies when the outcome is binary (0 or 1) ?**
    a. The model predicts a probability for each shot (e.g. xG = 0.23). The calibration plot aggregates shots into bins by their predicted probability and compares the mean

predicted probability within a bin (x-axis) to the *observed frequency* of goals in that bin (y-axis). Observed frequency is simply the proportion of shots that were goals (count of 1s divided by total shots) inside the bin. Thus, although each individual outcome is binary, the average across many binary outcomes becomes a frequency (a fraction between 0 and 1) which can be compared directly to the predicted probability. If the model is well calibrated, mean predicted probability ≈ observed frequency (points lie on the diagonal).

Interpretation of Figure 1. The blue points (xG-NextGen) lie close to the dashed perfect-calibration line across most bins, indicating that predicted probabilities match observed goal rates well on average. Small deviations in some bins (particularly at the high-probability end) are expected because these bins typically contain fewer shots: empirical frequencies there have higher sampling uncertainty. To quantify this uncertainty we report Brier score and log loss as quantitative calibration metrics in Supplementary Table 1.

6. **The n-previous passes include crosses and through passes?**
   a. Yes. But in this model the through balls and crosses are restricted to describing the assist type only (not necessarily whether or not it results in a goal, which is when the final pass is actually said to go down as an assist), which means it would only be used when describing the final nth prev pass (n=1) or last pass before the shot was attempted. Assist type is assessed as a separate variable by the model compared to n_prev_passes altogether, as crosses are generally attempted by a header or aerial attempt of some kind which usually results in a much lower quality shot as opposed to one struck using one's foot off the ground, due to the greater control the shooter has over the ball in this case. Also, through balls tend to play the shooter in-behind the opposition backline which effectively means angular defensive pressure is often minimal in such situations (one of the leading SHAP drivers), and creates one-on-ones with the opposition goalkeeper which are historically much more frequently put away as opposed to a cross, meaning a higher xG.

In summary, the paper is solid but is rough around the edges. Much more description of procedure, detail and discussion will make this exceptional. Please therefore address my comments. I look forward to the improved next iteration.

Thank you for addressing my comments. Accepted.