Beyond Expected Goals: A Probabilistic Framework for Shot Occurrences in Soccer

Tianshu Feng, Jonathan Pipping, and Paul Sabin

University of Pennsylvania

August 3, 2025

What Are Expected Goals (xG)?

- Expected Goals (xG) is a metric that estimates the probability that a shot is scored
- Depends on factors like distance from goal, angle to goal, shot type, and player positions
- Estimated by XGBoost models trained on historical shot data
- Often used to measure the quality of a chance
- Aggregated over a match or season to measure team performance

Limitations of xG

- Models are only trained on **observed** shots, inducing significant selection bias
- Skilled attackers who take more shots are over-represented
- Threatening attacks with no recorded shots are omitted
- Aggregating xG across a match double-counts rebound chances

Example 1: No Shot Recorded

Example 2: Multiple Shots Taken

Our Target Metric: xG+

- A more complete picture of goal expectancy
    - Accounts for high-threat attacks with no shots
    - Does not double-count rebound chances
- At each frame, we calculate the probability of a goal:

$$\begin{aligned}\mathbb{P}(\text{goal scored}) &= \mathbb{P}(\text{goal scored} \mid \text{shot taken}) \cdot \mathbb{P}(\text{shot taken}) \\ &= \text{xG} \cdot \text{xShot}\end{aligned}$$

- And then define xG+ for each possession as the probability a goal occurs:

$$\text{xG+} = 1 - (1 - \mathbb{P}(\text{goal scored})^{\text{n\_frames}})$$

Estimating xShot

- `xShot`: the probability that a shot occurs in the next second
- Build a model to estimate `xShot` based on features from tracking data
- Also build our own version of `xG` model using the same features on observed shots

Data Processing

- Remove games where no shots are recorded
- Only keep frames where the ball is in play and a team has clear possession
- Linearly interpolate ball positions to fill in missing frames
- `attack`: Index of the attack the current frame is on (0 if it is not on an attack)
    - Start with the attacking team gaining possession in their attacking third
    - End with the defending team regaining possession or the ball is out of their attacking third
    - Only keep frames with `attack` ¿ 0

Data Processing

- Rotate the coordinates 180° around the center point for frames where the team attacks from right to left to unify the attacking directions and make all x-coordinates positive
- Use a polar coordinate system centered on the goal for the ball
  - r_{ball} and θ_{ball} represent the distance and angle of the ball from the goal
  - Keep the z-coordinate and compute the speed of the ball
- Use a polar coordinate system centered on the ball for each player
  - Choose the 5 closest offense teammates and non-GK defenders to the ball as features
  - Keep goalkeeper positions as a separate feature

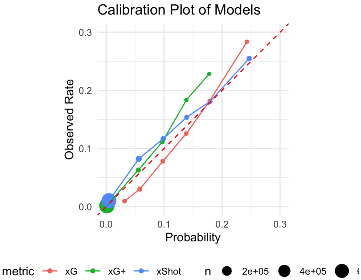
Data Processing

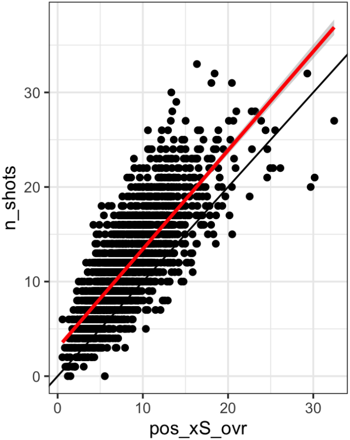
- `openGoal`: Percentage of the goal that is open from the ball's position
  - Simplify every defender as a circle with a radius of 0.75 m
  - Compute the two tangent lines from the ball to every defender in front of the ball their intersection points with the goal line
  - Calculate the length of the open goal as the length of goal not covered by segments formed by the intersection points

Model Specifications

- Trained on all tracking data of 2022-2025 Premier League seasons
- Use a 5-fold cross-validation to evaluate both `xG` and `xShot` XGBoost models
- Choose log loss as the evaluation metric

Results


Calibration Plot of Models
Observed Rate
Probability
0.0
0.1
0.2
0.3
metric
xG
xG+
xShot
n
2e+05
4e+05


n_shots
pos_xS_ovr
0
10
20
30

Cross-Validation Study

- **Objective**: Evaluate xG+ performance using cross-validated Poisson models
- **Dataset**: 3 seasons of match data (114 folds total)
- **Method**: Train on all matchdays except one, predict goals on held-out data
- **Goal**: Determine how well adjusted xG+ explains actual goals scored

Cross-Validation Setup

- Each matchday treated as a fold: 38 matchdays $\times 3 = 114$ folds
- For each fold:
  - Train on all matchdays except one to acquire adjusted metrics
  - Poisson regression on team goals using training data
  - Predict goals scored on held-out matchday test data

Metrics and Aggregation Methods

Metrics:

- **xS**: Probability a player takes a shot in the next second
- **xG**: Probability of a goal (given a shot)
- **xG+**: xS $\times$ xG, probability of scoring in the next second

Aggregation Methods:

1. **Max-per-possession**: Take maximum 1-second prediction in each possession
2. **At-least-one-per-possession**: $1 - \prod(1 - p)$ across possession
3. **Sum-of-shots**: Traditional xG summed over actual shots

Mixed Effects Modeling

Fitted on training data for each fold:

$$\texttt{metric} \sim (1|\texttt{season}) + (1|\texttt{season:team}) + (1|\texttt{season:opp}) + \texttt{home}$$

Extracted effects:

- Team attack (per season)
- Opponent defense (per season)
- Season effect
- Home field advantage

Secondary Poisson Model

Train Poisson regression on adjusted metrics:

$$\texttt{goals} \sim \texttt{home} + \texttt{season} + \texttt{team\_off} + \texttt{opp\_def}$$

Purpose: Assess predictive utility of each adjusted metric on actual goals

Cross-Validation Results

Table: Mean Squared Error (MSE) by Metric and Aggregation Method

Aggregation Method	xG+	xS	xG
At-least-one-per-possession	2.84	2.90	2.94
Max-per-possession	2.84	2.87	2.91
Sum-of-shots			2.90

Cross-Validation Results

Table: Mean Absolute Error (MAE) by Metric and Aggregation Method

Aggregation Method	xG+	xS	xG
At-least-one-per-possession	1.86	1.87	1.89
Max-per-possession	1.86	1.86	1.89
Sum-of-shots			1.87

Conclusions

- **xG+ performs best**: Lowest MSE and MAE across all aggregation methods
- **At-least-one-per-possession** aggregation method shows strongest performance
- **Future work**: Compare to actual shot-based xG, consider time-weighted xS
- **Questions and discussion**

Acknowledgements

- Special thanks to our data providers at PFF FC and the English Premier League.
- All work supported by the Wharton Sports Analytics & Business Initiative (WSABI) Summer Research Lab.