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

Jonathan Pipping, Tianshu Feng, and Paul Sabin

University of Pennsylvania

August 7, 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
  - Avoids double-counting rebounded chances
- At each frame t, let xG+ be the probability of a goal:

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

- Then define xG+ over a possession with n frames:

$$\text{xG+}_{\text{poss}} = 1 - \prod \left(1 - \mathbb{P}_t \left(\text{goal scored}\right)\right)$$

- Estimating this value requires fitting two models: xG and xShot

Data Overview

- **Source:** Pro Football Focus (PFF) FC video tracking and event data from the 2022-2025 English Premier League
- **Key Features:**
  - Player positions (x, y) at 30 frames per second
  - Ball position (x, y, z) at 30 frames per second
  - Shot events and outcomes
  - Team possession indicators
  - Player and team identifiers

Data Cleaning

- **Filtering:** Keep frames where the ball is in play and a team has clear possession
- **Smooth Ball Tracking:** Linearly interpolate ball positions to fill in missing frames
- **Define Attacking Sequences:**
  - Start: team gains possession in their attacking third
  - End: defending team regains possession or ball exits attacking third
- **Field Standardization:** Flip right-to-left attacks 180° to make all attacks go left-to-right

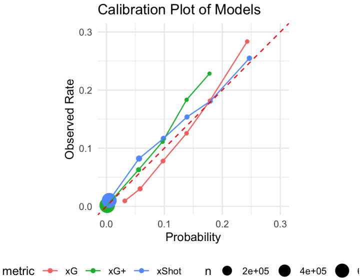
Feature Engineering

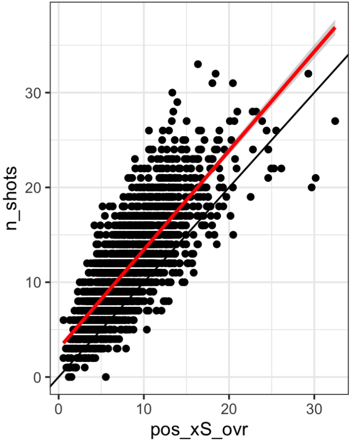
- **Ball Features:**
  - Distance from goal (r_{ball})
  - Angle to goal (θ_{ball})
  - Ball height (z_{ball})
  - Ball speed (v_{ball})
- **Player Features:**
  - Position of 5 closest offensive teammates relative to the ball (r_{off}, θ_{off})
  - Position of 5 closest non-GK defenders relative to the ball (r_{def}, θ_{def})
  - Position of goalkeeper relative to the goal (r_{gk}, θ_{gk})
- **Goal Openness:**
  - Model defenders as circles with diameter 0.75m
  - Draw tangent lines from the ball to defenders and find where they meet the goal line
  - Call the percentage of the goal that's unobstructed `openGoal`

Modeling

- **Training Data:** All data from the 2022-2025 Premier League seasons
- **Features:** Ball features, player features, and goal openness
- **xG**: 5-fold cross-validated XGBoost model estimating the probability that a shot is scored
- **xShot**: 5-fold cross-validated XGBoost model estimating the probability that a shot occurs *in the next second*
- **Evaluation Metric:** Log loss

Results


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


n_shots
0
10
20
30
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 × 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.