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

Tianshu Feng, Jonathan Pipping, and Paul Sabin

University of Pennsylvania

2025-07-30

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
- Has many practical uses
    - A measure of individual shot quality
    - Provides many individual-player metrics
    - A comparison of team performance in a match
    - A measure of team performance over a season

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 additively biases performance estimates
  - Rebounded chances are double-counted
  - Players only receive credit when a shot occurs

Visual Example 1

Video: Player has clear chance but passes instead of shooting

Visual Example 2

Video: Multiple shots from same attacking sequence w/ xG that add over 1

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 t, we calculate xG+ as the probability a goal is scored

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

- Then over a possession with n frames, xG+ is defined as

$$\text{xG+}_{\text{poss}} = 1 - \prod_{t=1}^{n} \left[1 - \mathbb{P}_t\left(\text{goal scored}\right)\right]$$

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

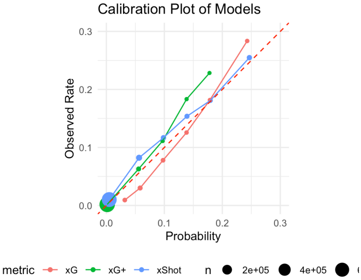
Model Specifications

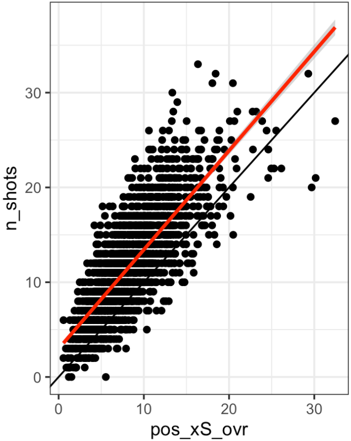
- Trained on all tracking data of 2022-2025 Premier League seasons
- Use a 5-fold cross-validation to evaluate both `xG` and `xShot` XGBoost models
- Choose log loss as the evaluation metric

Results

- Key findings
- Statistical significance
- Practical implications

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
6


n_shots
0
10
20
30
pos_xS_ovr
0
10
20
30

Conclusions

- Summary of main points
- Future work
- Questions and discussion

Thank You

Questions?

Wharton
UNIVERSITY of PENNSYLVANIA