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 Does It Mean to "Deserve" a Win?

- **Match Analysis:** After a match, commentators often remark that a team "didn't deserve the result." But what is this judgment based on?
- **Traditional Measure:** Expected Goals (xG), which estimates the quality of shots *taken*. But is that the right metric?
- **Our Question:** Which is more notable – getting a shot off, or making it?

Video Example

What Are Expected Goals (xG)?

- **Expected Goals (xG):** 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
- **Applications:**
  - Estimating the quality of a shot
  - Measuring of team performance over a match or season
  - Residualizing provides a measure of shooter skill

Limitations of xG

- **Selection Bias:** Models are only trained on *observed* shots
  - Skilled attackers take more shots and are more successful on them
  - Any chance without a recorded shot isn't in the training data
- **Aggregation Issues:** An incomplete measure of team performance
  - Threatening attacks with no recorded shots are omitted
  - Rebounded chances are double-counted

Example 1: No Recorded Shot

- **Match:** Manchester City v Real Madrid (Feb 19, 2025)
- **Question:** Which chance was more likely to produce a goal?

Example 2: > 1 xG on a Possession

- **Match:** Orlando City vs. Philadelphia Union (February 22, 2025)
- **Sequence:** Multiple shots in one possession totaling 1.63 xG

Example 2: > 1 xG on a Possession

Time	Player	Shot Outcome (xG)
78:01	Brekalo	Shot Blocked (0.05)
78:02	Muriel	Shot Post (0.52)
78:04	Pasalic	Shot Post (0.68)
78:05	Pasalic	Shot Goal (0.38)

Table: Sequence of shots leading to a goal, totaling 1.63 xG.

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:

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

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

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

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

Data Overview

- **Source:** Gradient Sports (formerly 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

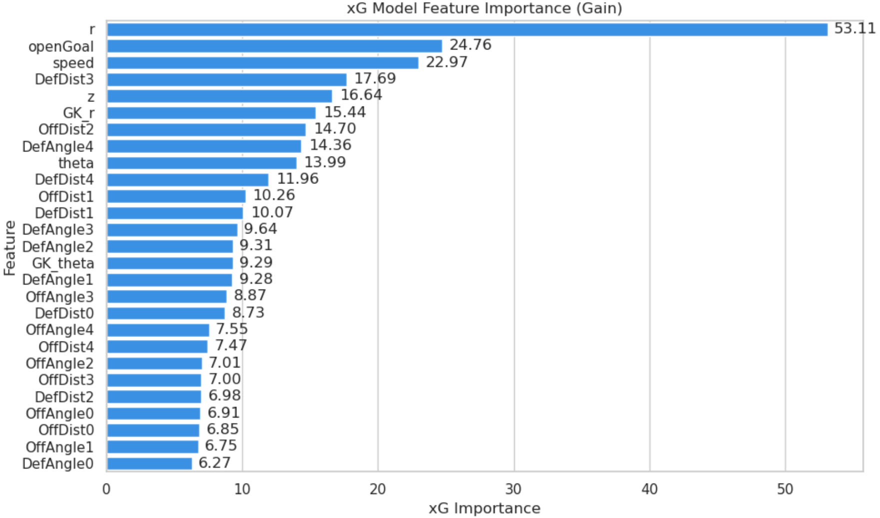
Feature Engineering

- **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`

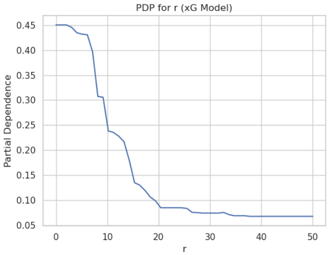
Modeling

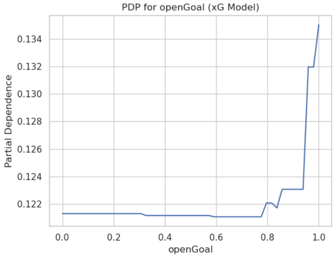
- **Training Data:** All data from the 2022-2025 Premier League seasons
- **Features:** Ball features, player features, and goal openness
- **Models:**
  - **xG:** 5-fold cross-validated XGBoost model estimating the probability that a shot is scored
  - **xShot:** 5-fold cross-validated XGBoost model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

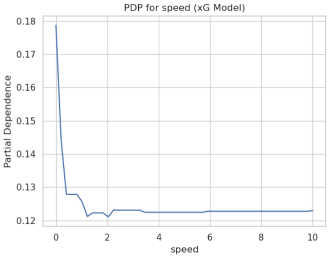
xG Model Results

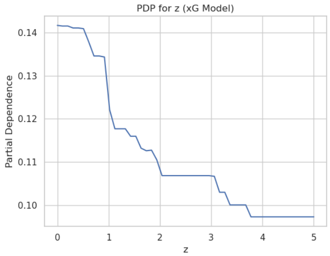


xG Model Results

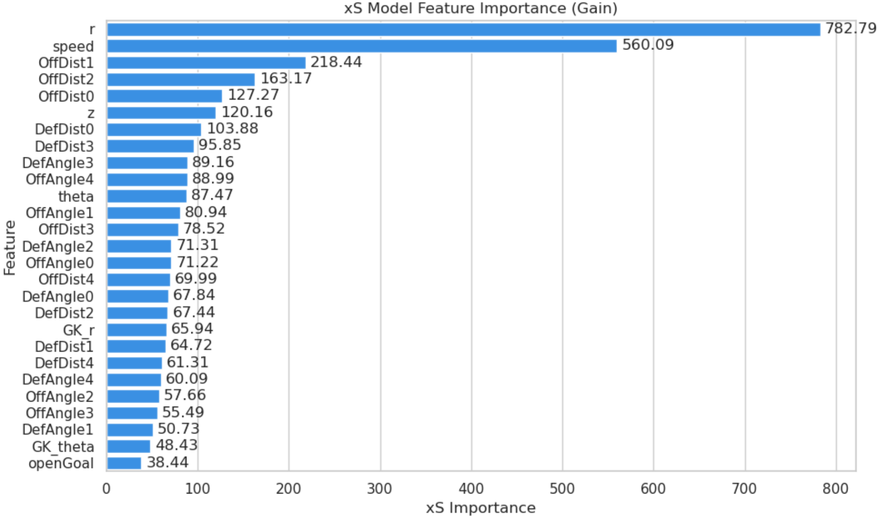




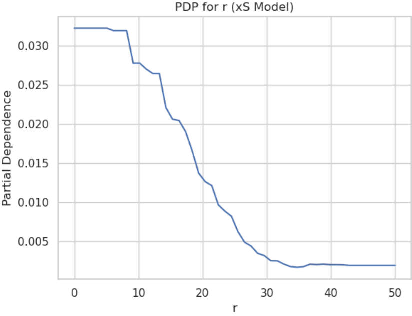


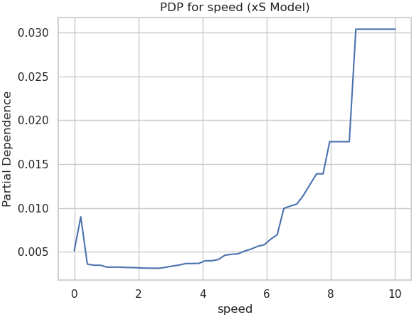


xS Model Results



xS Model Results





Cross-Validation Study

- **Objective:** Evaluate xG+ performance using cross-validated Poisson models
- **Dataset:** 3 seasons of EPL match data
- **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

- **Setup:** Treat each matchday as a fold.

$$38 \text{ matchdays} \times 3 \text{ seasons} = 114 \text{ folds}$$

- **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 that a shot is scored
- **xG+:** xS $\times$ xG, the 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

Player Ability

- **Question:** Does this framework provide insight into individual player skills?
- **Methodology:**
  - Aggregate chances for the closest player to the ball within a possession
  - Sum up per game & season for each of xG, xS, and xG+
  - Compare actual shots & goals to expected metrics
  - Analyze year-over-year performance consistency
- **Key Finding:** Players *do not* consistently over-perform their expected goals (xG)

Player Ability

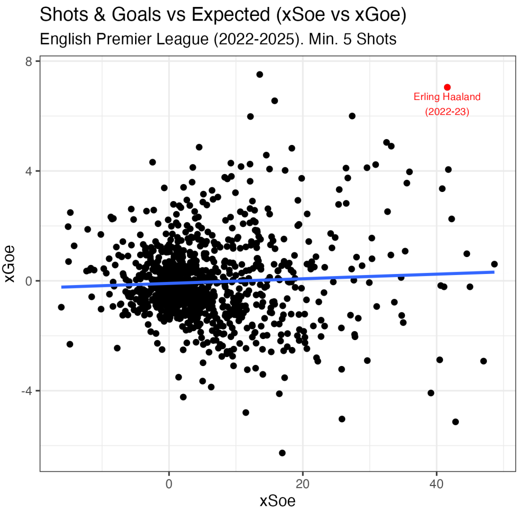
- **Key Finding:** Players *do* consistently over-perform as shot takers!
- **Implication:** Generally, elite goal scorers are the ones converting chances into shots, not converting shots into goals!

Table: Year to Year Correlation (Stability) vs. Expected (Per Game)

xG	xS	xG+
0.12	0.63	0.35

Player Ability

- **Example:** Haaland's record breaking season (36 goals)
- **Key Insight:** Required both higher shots and goals than expected



Top Over-Expectation Performers (Per Game)

xG+ Over Expected

Season	Player	xG+ OE	Goals	Games	Chances
2022–23	Erling Haaland	0.52	36	35	444
2023–24	Erling Haaland	0.44	30	31	513
2024–25	Omar Marmoush	0.41	10	16	293
2024–25	Yoane Wissa	0.35	22	35	429
2024–25	Mohamed Salah	0.33	27	38	1042
2024–25	Chris Wood	0.31	23	36	330
2023–24	Cole Palmer	0.30	19	34	693
2024–25	Alexander Isak	0.30	25	34	558
2023–24	Alexander Isak	0.30	19	28	351
2022–23	Harry Kane	0.29	25	38	589

xG Over Expected

Season	Player	xG OE	Goals	Games	Chances
2024–25	Omar Marmoush	0.41	10	16	293
2024–25	Chris Wood	0.31	23	36	330
2024–25	Michael Keane	0.21	3	10	19
2022–23	Erling Haaland	0.52	36	35	444
2022–23	Roberto Firmino	0.25	10	21	253
2022–23	Martin Ödegaard	0.22	15	37	942
2023–24	Heung-min Son	0.28	20	35	751
2022–23	Matias Viña	0.26	3	10	66
2022–23	Alexander Isak	0.24	11	22	346
2023–24	Taiwo Awoniyi	0.26	7	19	125

Note: Minimum 10 Games with a Chance

Conclusions

- **Best Metric for Team Prediction:** xG+
- **Best Aggregation Method:** At-least-one-per-possession
- **Player Evaluation:** Shots over expected is more predictive of future player performance than goals over xG
- **Future Work:** An instantaneous or time-weighted xShot model
- **Questions and Discussion**

Acknowledgements

- Special thanks to our data providers at Gradient Sports (formerly PFF FC)
- All work supported by the Wharton Sports Analytics & Business Initiative (WSABI)