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

Tianshu Feng, Jonathan Pipping, and Paul Sabin

University of Pennsylvania

July 29, 2025

Wharton
University of Pennsylvania

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

Wharton
University of Pennsylvania

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

Wharton
University of Pennsylvania

Visual Example 1

Video: Player has clear chance but passes instead of shooting

Wharton
University of Pennsylvania

Visual Example 2

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

Wharton
University of Pennsylvania

Methods

- Statistical approach
- Data processing steps
- Model specifications

Wharton
University of Pennsylvania

Statistical Approach

- `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
- Estimate the probability of goal as $P(goal) = P(shot) \cdot P(goal|shot)$

Wharton
University of Pennsylvania

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`

Wharton
University of Pennsylvania

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

Wharton
University of Pennsylvania

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

Wharton
University of Pennsylvania

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

Wharton
University of Pennsylvania

Results

- Key findings
- Statistical significance
- Practical implications

Wharton
University of Pennsylvania

Conclusions

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

Wharton
University of Pennsylvania

Thank You

Questions?

Wharton
University of Pennsylvania