Expected Goals Model and XAI ⚽

Mustafa Cavus

MI DATALAB

Content

Introduction to xG Model

- Definition
- Calculation
- Usage

Motivation

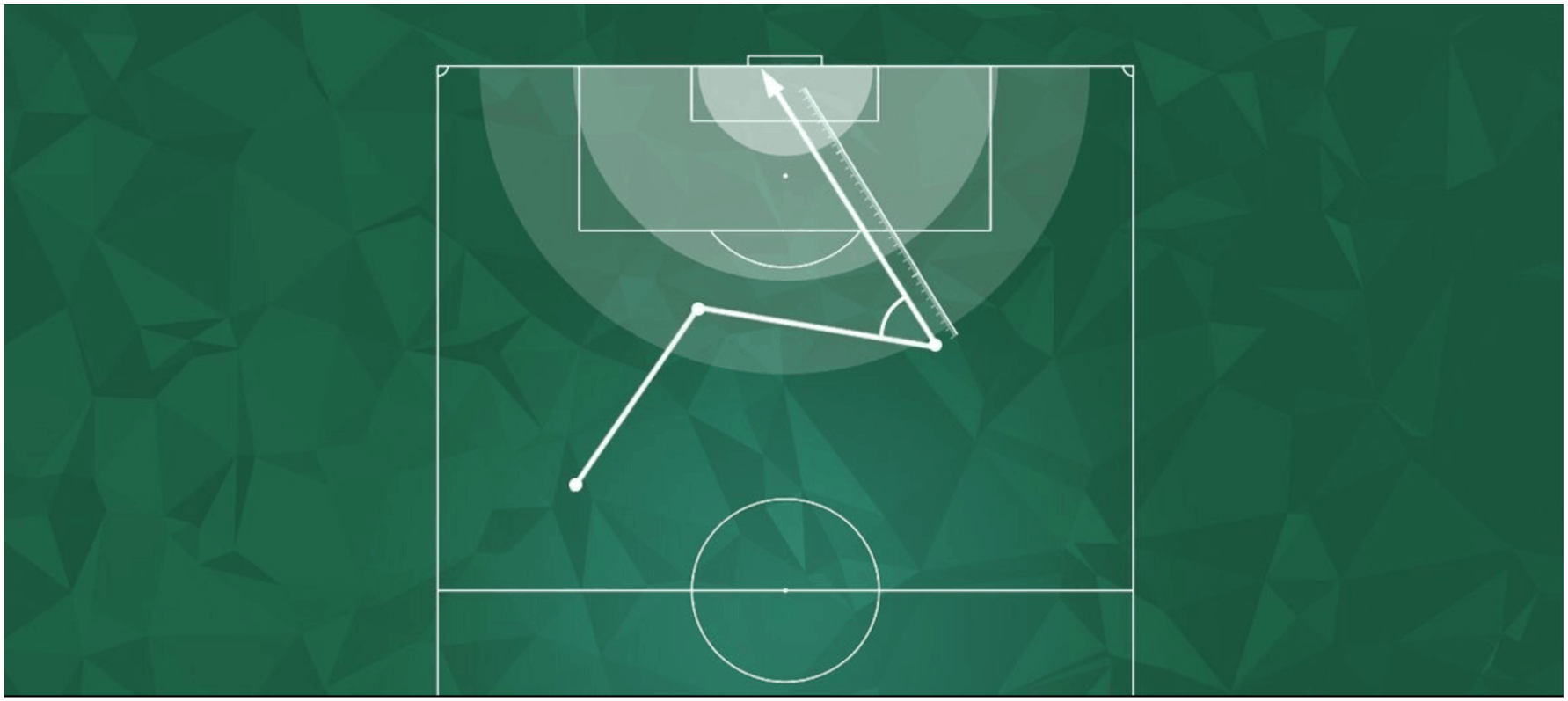
- Dataset
- Model
- Explanations

Open problems

Potential further studies

Expected Goals (xG) model

Expected goals (xG) is proposed to quantify the probability of a shot resulting in a score by **Green (2012).**



Predictors

1. Distance to goal
2. Angle to goal
3. Game setting
4. Part of the body
5. Last action
6. …

Calculation of the xG metric

There are two steps:

1. The individual scoring probabilities of multiple shots are calculated.

2. These probabilities are summed over for a player, or a team to derive the cumulative chance value.

For example, if a team had three shots in a match with probabilities 0.40, 0.10, 0.01. The team has generated chances worth 0.51 expected goals (Brechot and Fleep, 2020).

Motivation

- **Low-scoring nature of games.**

- Supporting a new mind-set for performance evaluation and decision-making of football clubs.

The football club FC Midtjylland won their first Danish league title using this method for the recruitment of players (de Hoog, 2015).

Variants of xG metric

Some variations are also used:

- **xGA**: expected goals against
- **npxG**: non-penalty expected goals
- **npxGA**: non-penalty expected goals against
- **xGChain**: total expected goals of every possession the player was involved in.
- **xGBuildup**: total expected goals of every possession the player was involved in without key passes and shots

Evaluation of xG metric

1. Performance-based usage

actual goals < xG means under-performance of the team/player

offensive ratio = actual goals scored / xG

actual goals > xG means over-performance of the team/player

defensive ratio = actual goals conceded / xG allowed

2. Ranking-based usage

created xG - allowed xG

Literature of football analytics

There are three parts of the football analytic literature:

- Scientific papers
- **Blogs**
- Sport data company bulletins

Dataset

Data Source: understat.com
by {worldfootballR} package

- 22 variables
- 395k observations
- 5 leagues
- 10 seasons

```
Rows: 395,824
Columns: 22
$ X.1             <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,…
$ league          <chr> "Ligue_1", "Ligue_1", "Ligue_1", "Ligue_1", "Ligue…
$ id              <int> 375456, 375457, 375458, 375459, 375461, 375453, 37…
$ minute          <int> 36, 38, 42, 46, 61, 24, 27, 34, 51, 71, 86, 3, 5, …
$ result          <chr> "SavedShot", "MissedShots", "MissedShots", "Missed…
$ X               <dbl> 0.714, 0.600, 0.712, 0.974, 0.690, 0.793, 0.803, 0…
$ Y               <dbl> 0.324, 0.282, 0.528, 0.536, 0.564, 0.380, 0.727, 0…
$ xG              <dbl> 0.015786497, 0.008143110, 0.016758630, 0.547621429…
$ player          <chr> "Toma Basic", "Otávio", "Hwang Ui-Jo", "Hwang Ui-J…
$ h_a             <chr> "h", "h", "h", "h", "h", "a", "a", "a", "a", "a", …
$ player_id       <int> 6902, 6060, 7746, 7746, 6060, 6018, 7169, 6018, 10…
$ situation       <chr> "OpenPlay", "OpenPlay", "OpenPlay", "FromCorner", …
$ season          <int> 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 20…
$ shotType        <chr> "LeftFoot", "RightFoot", "RightFoot", "Head", "Rig…
$ match_id        <int> 13977, 13977, 13977, 13977, 13977, 13977, 13977, 1…
$ home_team       <chr> "Bordeaux", "Bordeaux", "Bordeaux", "Bordeaux", "B…
$ away_team       <chr> "Nantes", "Nantes", "Nantes", "Nantes", "Nantes", …
$ home_goals      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ away_goals      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1,…
$ date            <chr> "2020-08-21 17:00:00", "2020-08-21 17:00:00", "202…
$ player_assisted <chr> "Nicolas de Preville", NA, "Otávio", "Nicolas de P…
$ lastAction      <chr> "Pass", "BallRecovery", "Pass", "Cross", "Pass", "…
```

Zivkovic, J. (2022) worldfootballR: Functions to Extract and Clean World Football (Soccer) Data, R package version 0.4.9, https://CRAN.R-project.org/package=worldfootballR

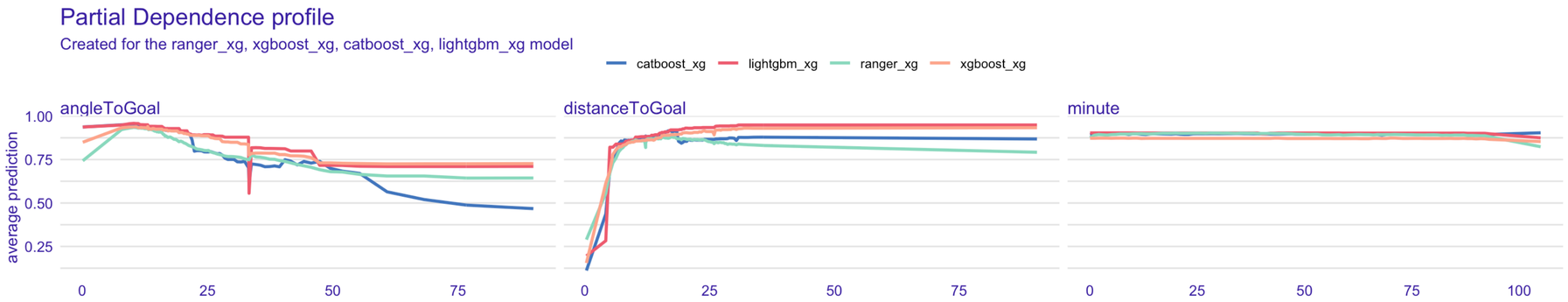
Model

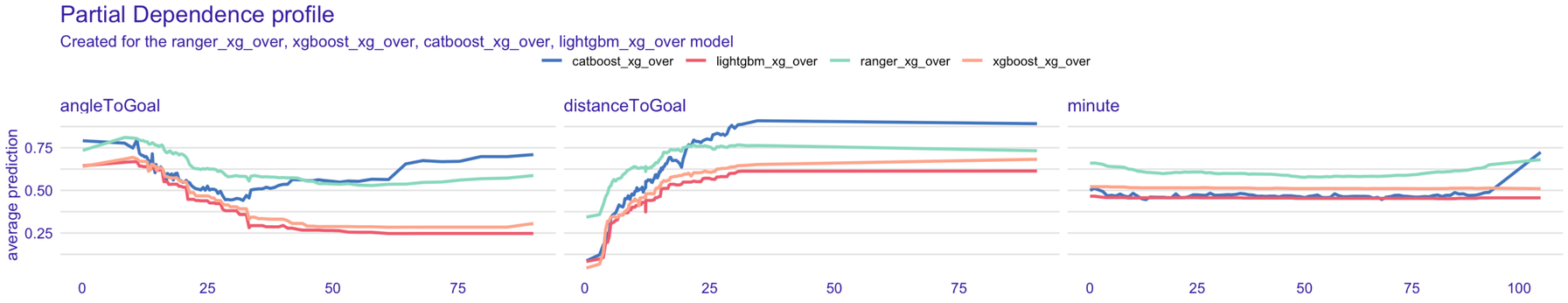
- random forest, catboost, xgboost and lightgbm models are used to train the model by {forester}
- train-test set split is used.

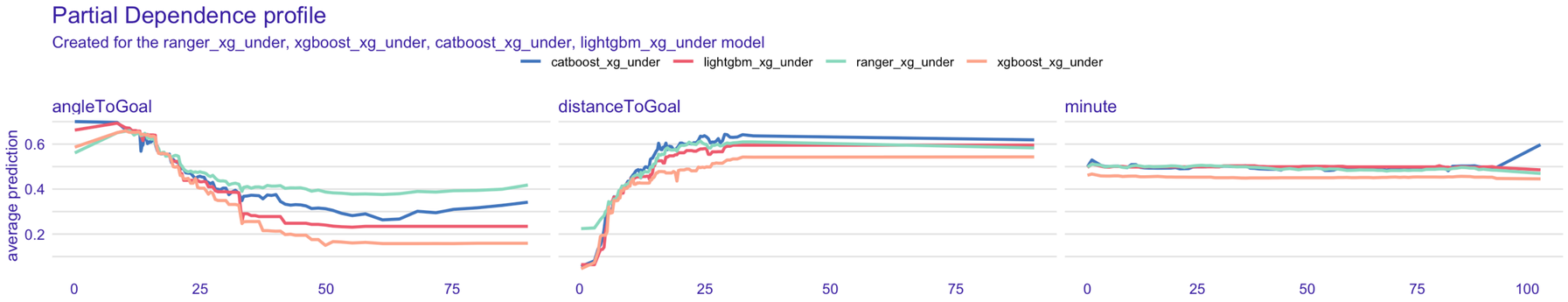
	sampling	recall	precision	F1	accuracy	AUC	MCC	Brier score	log-loss	balanced accuracy
ranger	-	0.9958	0.9251	0.9591	0.9243	0.9762	0.5201	0.0494	0.1717	0.6630
	over	0.9296	0.9569	0.9431	0.9439	0.9875	0.8881	0.0684	0.2643	**0.9439**
	under	0.8864	0.8663	0.8762	0.8750	0.9547	0.7503	0.1049	0.3551	0.8750
catboost	-	0.9913	0.9103	0.9491	0.9052	0.8189	0.3380	0.0755	0.2647	0.5899
	over	0.7653	0.7450	0.7550	0.7517	0.8363	0.5036	0.1657	0.4990	0.7517
	under	0.7539	0.7325	0.7430	0.7398	0.8112	0.4798	0.1726	0.5158	0.7398
lightgbm	-	0.9902	0.9095	0.9481	0.9033	0.8108	0.3198	0.0771	0.2692	0.5854
	over	0.7514	0.7248	0.7379	0.7331	0.8111	0.4665	0.1772	0.5286	0.7331
	under	0.7501	0.7252	0.7374	0.7334	0.8110	0.4672	0.1772	0.5288	0.7334
xgboost	-	0.9917	0.9087	0.9484	0.9037	0.8130	0.3191	0.0766	0.2677	0.5816
	over	0.7512	0.7257	0.7382	0.7336	0.8124	0.4676	0.1766	0.5272	0.7336
	under	0.7564	0.7242	0.7400	0.7347	0.8132	0.4699	0.1762	0.5257	0.7347

Lunardon, N., Menardi, G., and Torelli, N. (2014). ROSE: a package for binary imbalanced learning. *R Journal*, 6:82–92.

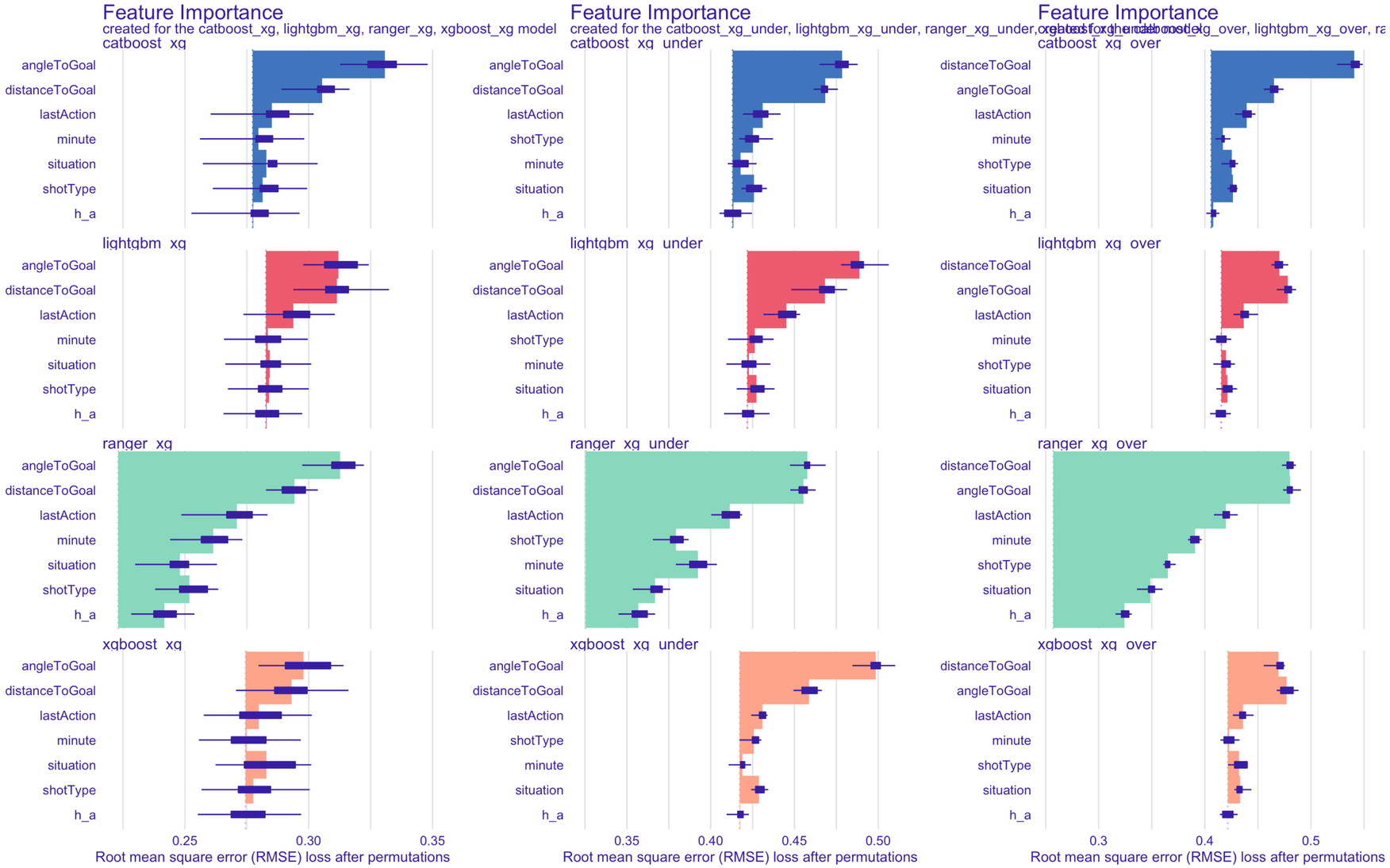
Explanations







Explanations



Open problems

- Proposing validated model by metrics for imbalanced data.
- Using XAI methods for player/team based explanations.
- Expanding the results of the following RQs in **Robberechts and Davis (2020)**:
  - How much data is needed to train an accurate xG models?
  - Does the data go out of date?
  - Are xG models league-specific?
- …

Potential studies

1. Writing a paper about the explanation of the xG models:
    - Promote to use {forester} and {DALEX}
2. Writing a story for Beta and Bit
3. Use-case for teaching

References

Brechot and Fleep (2020) Dealing with randomness in match outcomes: how to rethink performance evaluation in European club football using expected goals, *Journal of Sports Economics*, 21(4), 335-362.

Eggels et al. (2016) Explaining soccer match outcomes with goal scoring opportunities predictive analytics, *CEUR Workshop Proceedings*.

Kharrat et al. (2020) Plus–minus player ratings for soccer, *European Journal of Operation Research*, 283(2), 726-736.

Pardo (2020) Creating a model for expected goals in football using qualitative player information, *Master thesis in BarcelonaTech*.

Partida et al. (2021) Modeling of football match outcomes with expected goal statistic, *Journal of Student Research*, 10(1), 1-10.

Rathke (2017) An examination of expected goals and shot efficiency in soccer, *Journal of Human Sport and Exercise*, 12, 514-529.

Spearman (2018) Beyond expected goals, *MIT Sloan Sports Analytics Conference.*

Sam (2012) Assessing the performance of premier league goalscorer, *OptaPro Blog.*

Tiippana (2020) How accurately does the expected goals model reflect goalscoring and success in football?, *Master thesis in Aalto University*.