13TH ANNUAL
MIT SLOAN
SPORTS ANALYTICS CONFERENCE
MARCH 1 - 2, 2019 | BOSTON, MA
presented by ESPN

A Data Driven Goalkeeper Evaluation Framework

Derrick Yam
Data Scientist
StatsBomb

Paper Track: Other Sports (Soccer)
Paper ID: 13556

1. Introduction

Professional soccer clubs will rarely pay for quality goalkeeping despite the substantial responsibilities of the position. There are only 2 goalkeepers in the top 50 most expensive transfers worldwide (Most Valuable Transfers, 2018). An explanation for this phenomenon is that clubs struggle to quantitatively evaluate goalkeepers and may believe the position has more parity than any other position on the field[1]. One reason goalkeepers are so difficult to evaluate is their contribution within a game and even a season is largely dependent on their team's defensive system and defensive strengths, as well as the relative strength of their opponents.

Another reason is the scarcity of goalkeeper actions. A goalkeeper in the England Premier League faces only 12 shots a game, 80% of which miss the goal frame completely or were blocked before they reach the goalkeeper. It is not uncommon for a goalkeeper to go an entire game without making one save. If a goalkeeper can go an entire game without preventing one goal, then why should a club spend money on them? Given that 63% of games in the England Premier League last year were decided by one goal or less, that one shot on target that a goalkeeper saves or concedes is the difference between 3 points or 1 and 1 point or none. A team's detriment from poor goalkeeping and a team's success from superior goalkeeping is augmented at small samples.

In this manuscript, we outline a framework to analyze goalkeepers on key responsibilities: Shot Stopping, Cross Collection, Defensive Activities, and Distribution. This framework allows for improved transfer scouting, player development, and tactical analysis. We also construct a matching algorithm to identify similar goalkeepers within specified parameters. Our results quantitatively rank David de Gea as the best shot-stopper in the EPL and Ederson, as the most active goalkeeper with the largest defensive territory, and the top distributor.

2. Data

To date, most readily available data in professional soccer excludes rich information about goalkeepers and therefore, there is a dearth of rigorous, quantitative research with objectives to evaluate goalkeepers (T.A.W., 2018). All data used in the analyses for this manuscript are from StatsBomb[2]. We use all games available from the 2017/2018 seasons in Europe's Big 5 leagues as well as England League One (the third division professional soccer league in England). This data

[1] For more on the historical difficulty in goalkeeper evaluation please see, (T.A.W., 2018).

[2] Free data and data for purchase is available at https://statsbomb.com/data/

includes spatial, time, and other information on all "events" including passes, saves, shots, tackles etc. There is also goalkeeper and field player coordinates within the frame for each shot.

All analyses and models were constructed in R, using packages xgboost, FNN and StatsBombR (R Core Team, 2013) (Chen, 2018) (Beygelzimer, 2018) (Yam, 2018). All visualizations were made using ggplot2 (Wickham, 2018).

3. Shot Stopping

3.1. Goals Saved Above Average (GSAA)

Shot stopping is the single most important responsibility for any goalkeeper. Regardless of a goalkeeper's other strengths, if a goalkeeper cannot make saves, they will not play. Nonetheless, traditional metrics to evaluate shot stopping, like save percentage and shutout counts, are remarkably ineffective. Save percentage is heavily confounded by the types of shots a goalkeeper faces and shutout counts are more representative of a complete defensive system than a goalkeeper's performance (a goalkeeper can record a shutout without making a single save). Using traditional metrics, it is almost impossible to justly evaluate a goalkeeper's shot stopping capabilities. With the limitations of traditional metrics, soccer is desperate for metrics independent of the defensive system in front of a goalkeeper and independent of the quality of shots taken by the opponents.

This necessity has been previously studied with goaltenders in professional hockey (Schuckers, 2011) (Perry, 2015). Some early work on advanced goalkeeper metrics in soccer struck at the heart of our objective in shot stopping analysis, accounting for save difficulty (Trainor, 2014) (11tegen11, 2014). However, these analyses lacked defender information describing each shot and therefore could not to estimate save performance independent of a goalkeeper's defensive system.

We extend this research to professional soccer by estimating the goals saved above average (GSAA) for each goalkeeper.

$$\mathrm{GSAA}_{\mathrm{GK}} = \sum_{i} \mathrm{PSxG}_{\mathrm{i}} - \mathrm{GA}_{\mathrm{i}} * I(GK_i = gk) \tag{1}$$

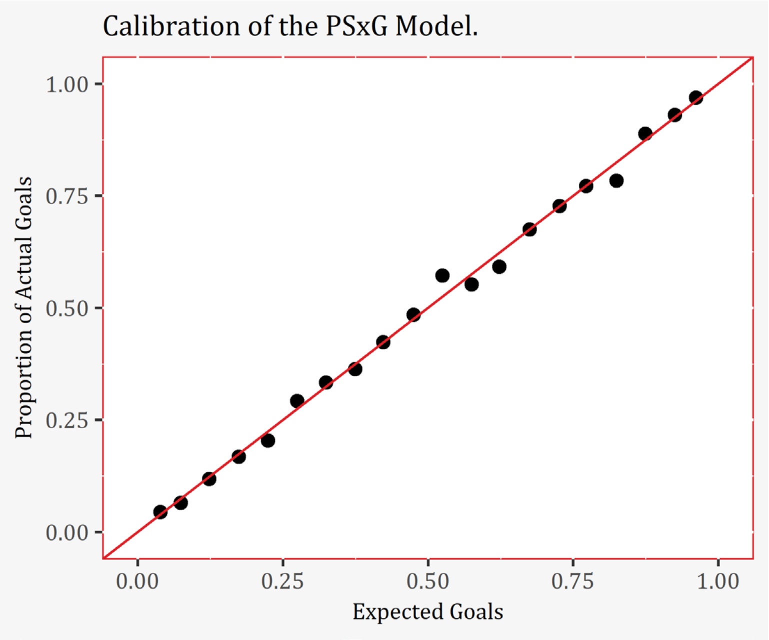
Using the defenders' coordinates on all shots in our data set, we simultaneously adjust for shot quality and defensive system when constructing a Post-Shot Expected Goals model (PSxG). PSxG differs from Pre-Shot Expected Goals (xG) in that it only includes shots that are on target, since all shots that get blocked or miss the goal frame have 0 probability of becoming a goal. Therefore, if we were to include these shots in the sample it would bias the total PSxG that a goalkeeper faces. We aggregate the total PSxG faced by each goalkeeper and then subtract their total goals conceded (GA). At season end, this GSAA is the estimate for how many goals above or below average a goalkeeper saved or conceded respectively (Equation 1) and we standardize the metric by dividing by the total number of shots a goalkeeper faces (SOT) (Equation 2).

$$\mathrm{GSAA\%}_{\mathrm{GK}} = GSAA_{GK} / SOT_{GK} \tag{2}$$

We estimate PSxG using an extreme gradient boosting model (xgboost). Our objective function in xgboost is logistic regression predicting the binary outcome, 1 if Goal and 0 if Saved, using a set of predictor variables defining the shot characteristics and defensive shape. We cannot include any information on the goalkeeper's location, because goalkeepers who are typically positioned better than others will consistently have lower PSxG and their $GSAA\%_{GK}$ will be deflated, despite the low $PSxG_{GK}$ value being a result of their quality positioning. Some of the most important variables in the model are: the defensive density behind ball, defined as the sum of the inverse distances for each defender between the shot and the goal, the distance from the shot to the center of the goal frame, the estimated location the ball is expected to cross the goal line given the shot's trajectory and the attacking speed from the start of the possession defined as the distance from where the attacking team gained possession of the ball to the shot location divided by the time between those two points.

Xgboost, as a typical tree-based learning method, controls for collinearity between predictors well, and detects variable interactions through the optional parameter of randomly sampling predictor variables in each tree. We set additional xgboost parameters to build an exact tree rather than an approximate tree, the maximum depth of each tree is 6 predictors, we sample 30% of the predictors to be used in each tree, and we set a learnings rate (eta) of 0.03. We use 10-fold cross validation to estimate the accuracy of the model on testing data and avoid over-fitting.

Figure 1*: Calibration of the post shot expected goals (PSxG) model. The closer points are to the line where expected goals = actual goals the better the calibration of the model.*



Since we are using the GSAA% to compare goalkeepers, it is essential that our PSxG model is not only accurate, but well calibrated. If there is a systematic flaw in the calibration of the model, our estimates for GSAA% will be biased and untrustworthy. The calibration is how accurate the model's predicted probabilities are to the true probabilities. This is important because if the calibration was off and low probability shots were consistently estimated too high, or high probability shots were consistently estimated too low, when we are aggregating PSxG across an entire season, goalkeepers who face a lot of low probability shots will have their GSAA% inflated while goalkeepers who face a lot of high probability shots will have their GSAA% deflated. In Figure 1, we confirm the calibration of our model and predict PSxG for each on target shot in our sample and calculate GSAA_{GK} and $\text{GSAA\%}_{\text{GK}}$ for all goalkeepers. The results for all first-string goalkeepers in the EPL is presented in Table 1.

Table 1: *First string goalkeeper for each team in the EPL during the 2017/2018 season.*

Goalkeeper	Team	GSAA%	GA	PSxG	SOT	SV%	xSV%	GSAA
David de Gea	Manchester United	9.2%	24	38	142	82.4%	73.2%	13.0
Nick Pope	Burnley	7.6%	29	41.5	150	80.1%	72.6%	11.5
Lukasz Fabianski	Swansea City	4.5%	49	57.6	190	74.2%	69.7%	8.6
Mat Ryan	Brighton and Hove Albion	4.2%	47	55.4	176	72.9%	68.7%	7.4
Jack Butland	Stoke City	2.7%	54	59.6	206	73.8%	71.1%	5.6
Asmir Begovic	Bournemouth	1.3%	58	60.2	173	66.5%	65.2%	2.2
Wayne Hennessey	Crystal Palace	0.8%	38	39.1	129	70.5%	69.7%	1.1
Fraser Forster	Southampton	0.0%	28	29.0	98	70.4%	70.4%	0.0
Petr Cech	Arsenal	-0.2%	42	41.7	124	66.4%	66.6%	-0.3
Hugo Lloris	Tottenham Hotspur	-0.4%	32	32.5	124	73.4%	73.8%	-0.5
Robert Elliot	Newcastle United	-1.3%	19	18.2	63	70.3%	71.6%	-0.8
Thibaut Courtois	Chelsea	-1.8%	30	28.0	109	72.5%	74.3%	-2.0
Jonas Lössl	Huddersfield Town	-2.2%	52	48.7	151	65.8%	68.0%	-3.3
Jordan Pickford	Everton	-2.8%	55	50.1	177	68.9%	71.7%	-5.0
Heurelho Gomes	Watford	-4.2%	35	30.9	97	64.3%	68.5%	-4.1
Kasper Schmeichel	Leicester City	-4.4%	44	38.8	142	68.3%	72.7%	-6.2
Ederson	Manchester City	-5.0%	25	21.1	79	68.4%	73.3%	-3.9
Ben Foster	West Bromwich Albion	-5.6%	52	44.5	152	65.4%	70.9%	-8.5
Simon Mignolet	Liverpool	-7.0%	21	18.1	57	61.4%	68.4%	-4.0
Joe Hart	West Ham United	-7.1%	33	27.7	87	61.4%	68.5%	-6.3

From Table 1, we see that the top performing goalkeeper in the EPL last season was David de Gea, who we estimate saved 13 goals more than the league average goalkeeper. That is a massive defensive contribution from one player and an estimate presumably independent of the shot quality and the defensive system played by Manchester United.

Analysts estimate that each goal in the EPL is worth roughly one point in the league table (Sears, 2015). Manchester United placed second in the EPL last season and under this estimate, if Manchester United replaced David de Gea with a league average goalkeeper, they would have finished 5th or 6th with 68 points and out of the qualification for the Champions League. Manchester United profited $40.35 million directly from UEFA for their participation in the 2017/2018 UEFA Champions League (statista). Conversely, if a goalkeeper like Ben Foster who conceded more than 8 goals below average were replaced by a league average goalkeeper, West Bromwich Albion would have finished 15th in the EPL avoiding not only relegation, but the loss of an estimated 50 million pounds that comes with it (Smith, 2018). We must note that there is a conservation of points within the EPL, i.e. if multiple teams made goalkeeper changes the effects would diminish. [3] Nonetheless,

[3] The exact profit valuation is also outside the scope of this analysis and is presented to show the under-valuation of goalkeepers that professional soccer has been suffering from for years.

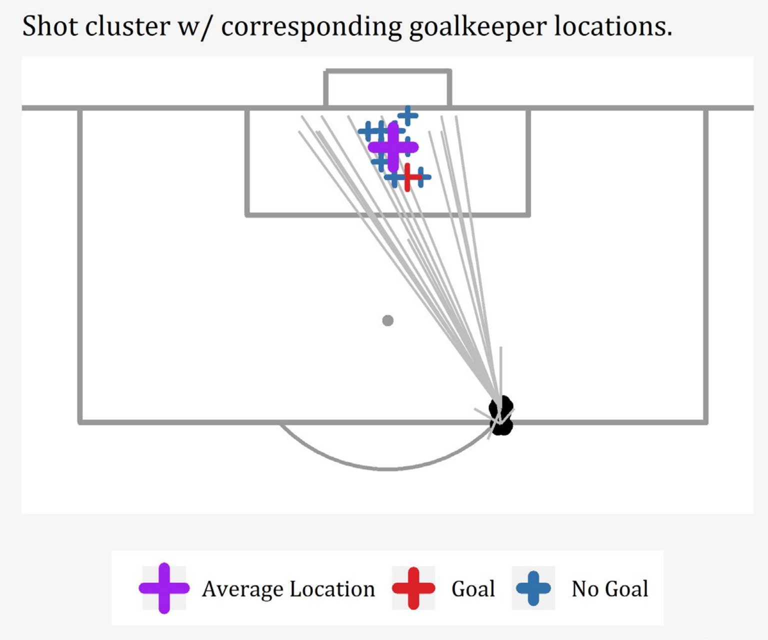
the individual contribution from one player, the goalkeeper, can be huge and if markets were efficient, we would see that reflected in transfer fees.

3.2. Positional Deviations

Positioning is another imperative skill a goalkeeper must learn to maximize their likelihood of making saves. Depending on a goalkeeper's size, reaction speed and personal intuition, goalkeepers' set positions may vary slightly. However, at the professional level, it is safe to say that goalkeepers position themselves correctly on shots against more often than not, and that the optimal position for each shot is similar across all goalkeepers. Working under this assumption, we can then estimate a goalkeeper's positional deviations from average as a proxy for their deviations from optimal position.

Using a K Nearest Neighbor (KNN) algorithm, for each shot we search for the 20 most similar shots based on characteristics that innately influence a goalkeeper's position: x and y location, shot type (free kick or open play) and whether the shot was headed. With this cluster of 20 shots, we then average the goalkeepers' x and y location on the pitch to represent the "optimal position". An example of this is presented through data points in Figure 2.

Figure 2: *Randomly sampled shot and the cluster of most similar shots from the knn algorithm.*

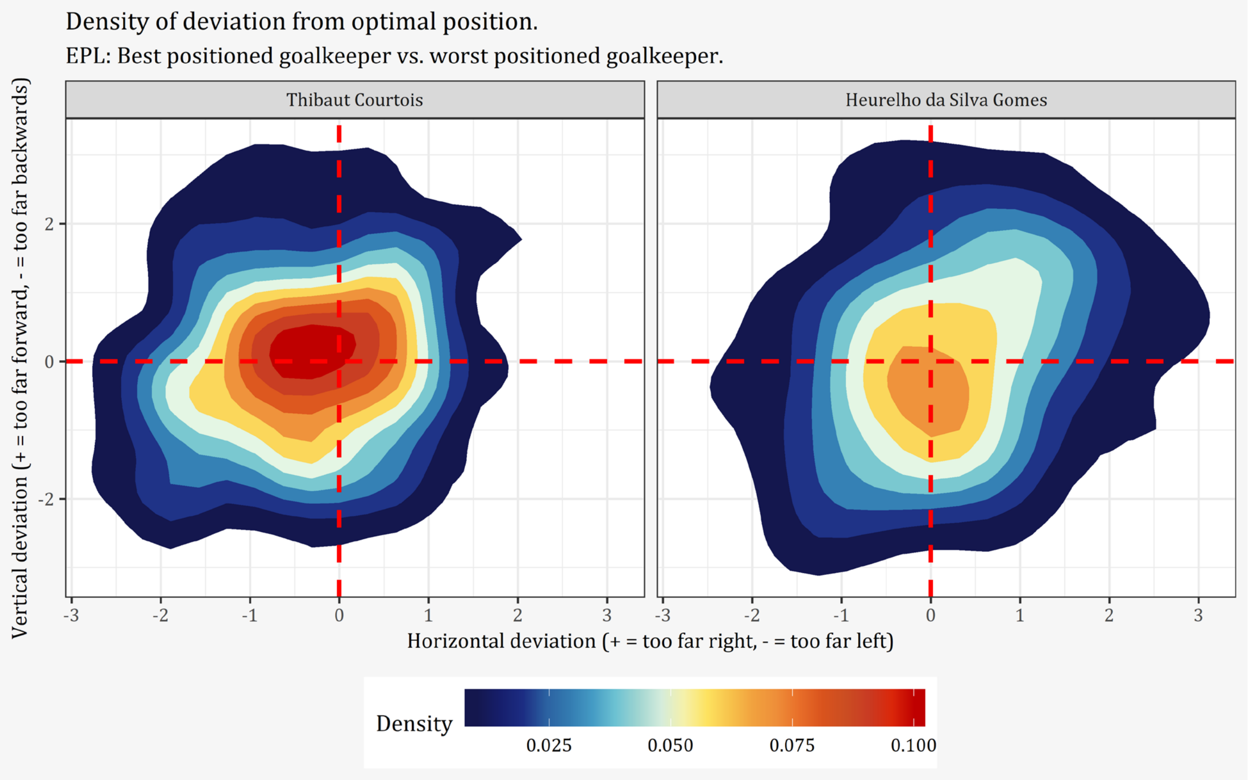


For each shot, we take the difference in horizontal location the goalkeeper is from the average location (xdiff) and the difference in vertical location the goalkeeper is from the average location (ydiff). We then calculate the average deviation for each goalkeeper,

$$PD_{\text{GK}} = \frac{\sum_{i=1}^{n}\sqrt{xdiff_i^2+ydiff_i^2}}{n} * I(GK_i = gk). \tag{3}$$

Using the horizontal and vertical deviations, we can construct densities of goalkeepers' deviations from average to show not only how often a goalkeeper deviates from average, but where they tend to deviate to. An example of the best positioned goalkeeper in the EPL and the worst positioned goalkeeper in the EPL is presented in Figure 3.

Figure 3: *Deviations from optimal positioning as calculated from the knn algorithm.*



In Figure 3, we see the best positioned goalkeeper, Thibaut Courtois has a greater density (denoted in dark red) close to the optimal position. On the other hand, Heurelho Gomes has less density at the optimal position and greater density further away from the optimal position. Most notably, Gomes stretches too far forward and too far to his right. Perhaps Gomes is more confident moving and diving to his left and over compensates for that. Gomes ranked 15th in the EPL in GSAA% last season.

4. Defensive Activities

4.1. Claiming crosses and other long passes

As many people know a goalkeeper is responsible for much more than saving shots. One of the most prominent secondary responsibilities for a goalkeeper is claiming crosses and long balls. Claiming a cross, free kick or long ball before it can possibly connect with the intended attacker typically ends an attack and grants the goalkeeper's team possession of the ball. A goalkeeper, the only player allowed to use his hands has the advantage over every player on the pitch so long as he can get to the ball first. Nonetheless despite the obvious advantage, the decision to attempt to claim a ball differs greatly between goalkeepers.

The likelihood a keeper decides to claim a ball depends heavily on the type of pass. Therefore, in order to compare goalkeepers based on their aggression on claimable passes, we must standardize a goalkeeper's claims by the probability an average goalkeeper would attempt to claim those same passes. We construct another probabilistic model to predict the probability a pass is collected by a goalkeeper. We restrict our sample to "claimable" passes, which for simplicity we refer to as "claimables" (for more information on claimables, please see Appendix A1). We define the probability a claimable is collected as expected claims (xC),

$$\begin{gathered}\mathrm{xC} = \Pr(\mathrm{Collected}|\mathrm{X}), \\ where\ X\ is\ the\ set\ of\ all\ pass\ characteristics\ x.\end{gathered} \tag{4}$$

Once we filter for all claimables, we define certain characteristics about each pass as covariates (X) for our xC model. We know the end location of each claimable, however using the end location to predict the probability a claimable is collected would bias our results given that where a pass ends depends on the result of the pass. A pass that is collected will typically end near the goal mouth and a pass that is not collected will continue to travel until it either connects with an attacker or defender or stops moving. Therefore, we use the beginning and end location of all claimables to estimate the trajectory (in point-slope form) of each claimable. We then use this trajectory to calculate where each pass would intersect a radius of 6 yards around the center of the goal frame and 15 yards around the center of the goal frame. These intersection points offer an unbiased proxy for where the goalkeeper would be able to claim the ball. Additional covariates we include in our model include the beginning x and y location of the pass, the pass angle, the pass height, the pass type (Free Kick, Corner, Open Play), the time in the match, and the competition. We train the model using the same parameters as our xG model in Section 3.1 and similarly check the accuracy and calibration which can be found in Appendix A1.

After confirming the validity of our xC model, we apply xC values to all claimables and define two new metrics for a goalkeeper's ability to claim passes. The first is Claimables Collected Above Average (CLCAA) which is very similar to GSAA and is calculates as,

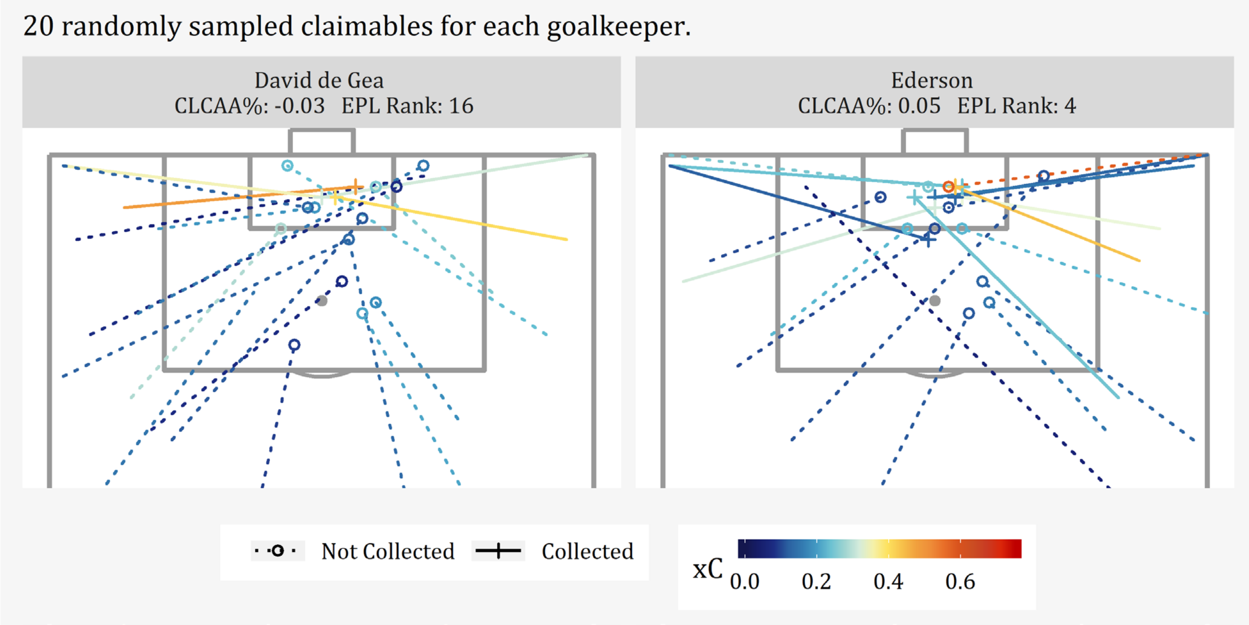
$$\mathrm{CLCAA_{GK}} = \textstyle\sum_i \mathrm{xC_i} - Collected_{\mathrm{i}} * I(GK_i = gk). \tag{5}$$

And the second is a form of CLCAA standardized by the number of claimables each goalkeeper faced (CLCAA%),

$$\mathrm{CLCAA\%_{GK}} = \ \mathrm{CLCAA_{GK}}/\mathrm{Claimables_{GK}}. \tag{6}$$

Since some of the modelling information can be difficult to understand in writing, we present an example of some claimables with their xC values in Figure 4. In Figure 4, we see a stark contrast between Ederson, the 4th best goalkeeper in the EPL (based on CLCAA%), and David de Gea, the 16th best goalkeeper in the EPL. Ederson collects claimables further up the pitch, outside of the center of the six-yard box, and he rarely leaves claimables with a high xC uncollected. De Gea on the other hand, only collects high xC claimables near the center of the six-yard box and leaves some high xC claimables uncollected.

Figure 4: *Random sample of claimables for two prominent goalkeepers in the EPL with their corresponding xC and whether or not they were collected.*



In Appendix A1, we show the CLCAA%, the CLCAA, and the number of claimables faced for each first-string EPL goalkeeper last season. The goalkeepers who were most aggressive at collecting claimables are those with the highest CLCAA%.

Although collecting claimables does not have an explicit value in the same way that making an unlikely save quite clearly is worth ≈ -1 goals against for your team, they are undeniably important. Collecting a claimable is the only way to surely end an attack, gain possession of the ball and begin a counter attack at your team's own pace. When a goalkeeper decides to stay on his goal line and trust his defenders to clear the ball away, they cannot be certain that the ball will clear out of their defensive zone safely or that an attacker won't win the ball first and earn a shot on goal. In Section 7, we discuss how we may actually estimate the value of collecting claimables.

4.2. Defensive territory

In a simpler analysis, we investigate the distribution of the distance from goal for goalkeeper actions that are not passes, saves or claims. This includes clearances, interceptions, tackles and ball recoveries which we refer to as defensive activities. This shows the presence a goalkeeper has further up the pitch and measures their defensive contribution in a manner more common to field players. We then derive the distance from goal for each defensive activity and rank goalkeepers based on the quantiles of the distance from goal, which we refer to as the goalkeeper's defensive territory (DT_{GK}).

$$\mathrm{DT_{GK}} = \{DT_i\}_{(||0.75*n||)} * I(GK_i = gk)$$
$$Where\ DT\ is\ the\ set\ of\ all\ defensive\ activities \qquad (7)$$

The goalkeeper in the EPL with the greatest defensive territory was Ederson, commonly known as one of the most aggressive goalkeepers in professional soccer. Comparing him to the best shot stopper in the EPL, David de Gea, ranked in the 5 smallest defensive territories of EPL keepers. We find a statistically significant difference between them[4]. On average, Ederson extends his defensive territory 5 yards further from goal than De Gea.

5. Distribution

5.1. Contribution to the attack

Arguably the most rapidly evolving aspect of a goalkeepers' game is ball distribution. In modern soccer, goalkeepers are not only expected to distribute the ball well, but to recognize attacking opportunities and play quickly. We propose a weighted function for measuring a goalkeeper's contribution to the attack based on team's positive outcomes within the 20 seconds ensuing a goalkeeper's pass. Positive outcomes are defined as all free kicks earned in the attacking half, corner kicks earned, shots attempted, and penalties earned. We calculate the total positive outcome contribution (TPO) for each goalkeeper, using intuitively defined weights based on the probability of scoring a goal for each positive outcome listed in Equation 8.

$$\mathrm{TPO_{GK}} = 0.5\sum_f FreeKick_f \ + \ 0.4\sum_c Corner_c \ + \ 1\sum_s Shot_s \ + \ 7.5\sum_p Penalty_p * I\big(GK_{f,c,s,p} = gk\big) \qquad (8)$$

In a more advanced study, we would construct individual models for each team to estimate the weights (or, "value") of each positive outcome. We could then account for goalkeepers who not just contribute to the overall attack but recognize scenarios beneficial to their team's style, where their team can earn a dangerous free kick if they are better at free kicks or take a high-quality shot if they do not convert well on corners and free kicks.

5.2. Change in distribution tendencies under pressure

One of the obvious signs of a player's and especially a goalkeeper's distribution ability is how their tendencies change under pressure. Weaker, less confident, players often struggle under pressure, they may panic and clear the ball away, or fail to complete passes, mistakenly kicking the ball out of bounds or to the other team. Using "Pressure" events in StatsBomb data, we look at changes in a goalkeeper's pass completion percentage and the change in pass length when the goalkeeper is and is not under pressure (description of how pressure is defined is in Appendix A2). Goalkeeper's who are confident under pressure will maintain their distribution tendencies.

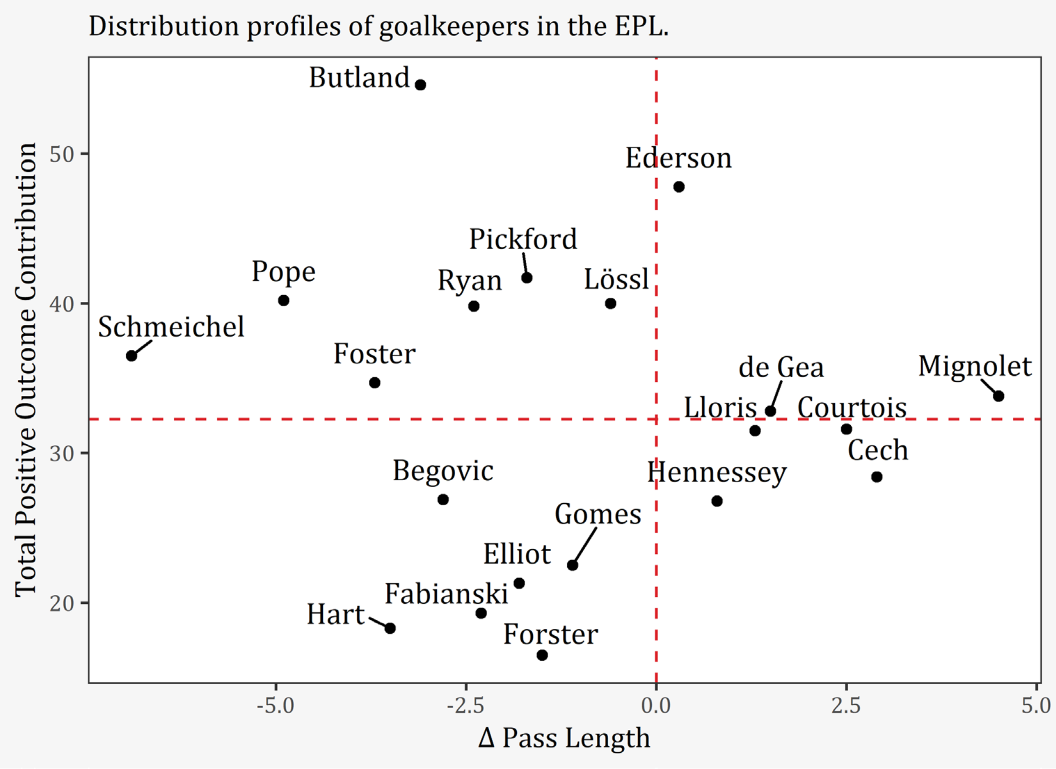
$$\Delta\mathrm{PL_{GK}} = \frac{\sum_p^n \mathrm{PassLength}}{n} - \frac{\sum_i^n \mathrm{PassLength}}{n} * I\big(GK_{i,p} = gk\big), \qquad (9)$$

[4] Significance tested using a Wilcoxon rank sum test to control for the skewness in the distance from goal.

$$for\ all\ unpressured\ passes, n_i, and\ all\ pressured\ passes, n_p.$$

Generally, goalkeepers will kick the ball long when under pressure. A conservative tactic that limits the risk of an immediate attack but reduces the probability of retaining possession drastically.

Figure 5: *Scatterplot of the distribution statistics defined in Section 5. For all first-string goalkeepers in the EPL during the 2017/2018 season.*



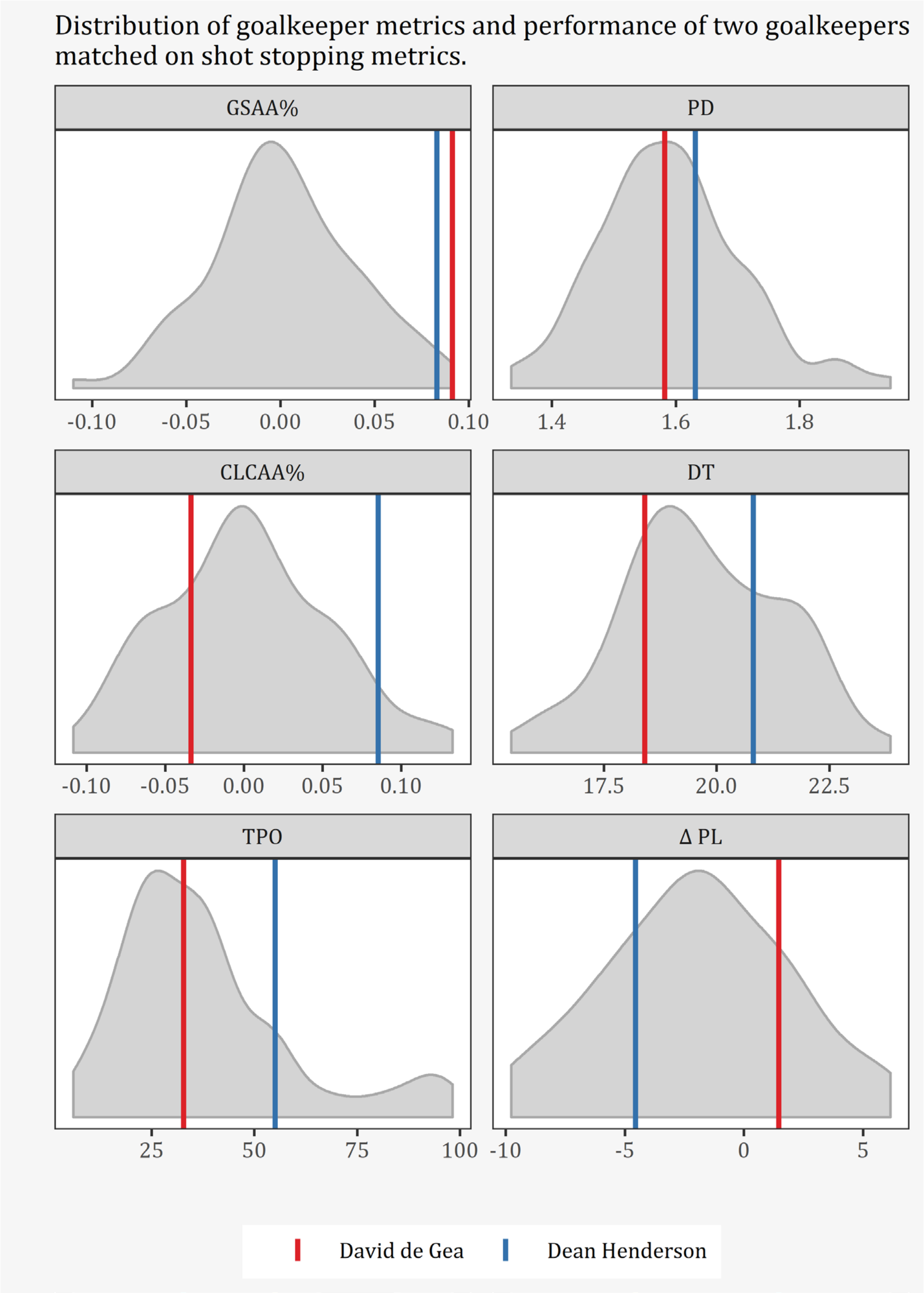
In Figure 5, we look at goalkeepers' ability to contribute to the attack and play under pressure. The average TPO for the EPL is represented in the dashed horizontal line and a change in pass length of 0 when under pressure is represented in the dashed vertical line. Goalkeepers that we intuitively regard as strong distributors will contribute to the attack and maintain their distribution tendencies when under pressure. These goalkeepers are high on the y axis and very close to the dashed vertical line where $\Delta \text{PL} = 0$. The best distributor in the EPL last season was Ederson. Although Ederson did not exceed expectations in shot stopping, he had the largest defensive territory, collected the fourth most claimables and appears to be the best distributor in the EPL. It is becoming clear why, at the time, Manchester City acquired him for the world record transfer fee for goalkeepers in July of 2017 (Most Valuable Transfers, 2018).

6. Matching Method

As we have seen, the style of goalkeeping differs greatly, and the value one goalkeeper has to one specific organization can differ substantially from their value to another organization. Because of this, it would be unjust to rank goalkeepers on an overall value when one soccer club needs a world class shot stopper and the other needs a keeper sweeper to clean up through balls and win everything in the air. Therefore, we propose a flexible matching algorithm to detect matches within specified parameters. Using the metrics defined in this manuscript (GSAA%, PD, CLCAA%, DT, TPO, ΔPL) users can specify closeness parameters for each metric as well as specify the competitions to search for the most similar goalkeepers.

As an example, we find the 5 most similar goalkeepers to David de Gea in the England League One. Through this report we have profiled David de Gea as an exceptional shot stopper, timid in defensive actions and average distributor. We match only on shot stopping metrics, to detect the most similar shot stoppers (De Gea's greatest asset), while allowing potential matches to deviate in their aggressiveness (De Gea's greatest weakness) and distribution. In Figure 6, we see that League One goalkeeper, Dean Henderson, matched closest to De Gea.

Figure 6: *Distribution of the goalkeeper metrics defined in this manuscript. Results of matching League One goalkeepers to David de Gea's shot stopping metrics shown through the vertical lines.*



Henderson ranks very similar to De Gea in terms of GSAA% and PD. Unlike De Gea, Henderson is more aggressive than the average goalkeeper, and may be an improvement to De Gea if Manchester United needed a goalkeeper to play further off his goal line. Coincidentally, Dean Henderson is on loan from Manchester United and they may see him as a potential replacement for David de Gea, but at the very least they should recognize his transfer value. Although Dean Henderson far exceeded expectations in League One, at this time we do not know how closely these metrics translate across leagues with varying talent levels.

7. Discussion

To date, this is the most comprehensive framework for evaluating an underappreciated position in soccer, the goalkeeper. As the first goalkeeper evaluation framework, we do not have a testing set available to validate our metrics or compare them to predetermined expectations. However, with some knowledge of professional soccer, and consulting with analysts who work in professional soccer, our methods have passed the "eye-test". We see the most well renowned shot stopper in the Premier League, David de Gea, with the greatest GSAA. The most aggressive goalkeeper and the best distributor, Ederson, with high CLCAA% and the largest Defensive Territory changing his behavior minutely under pressure while contributing immensely to the attack.

It would be unjust to present these results without noting possible limitations. Overall, the use of only one year's worth of data restricts our ability to make longitudinal inferences about goalkeepers and discover the most important metrics for goalkeepers at different points in their career. Also, it is not possible to test the repeatability of shot stopping performance with only one season of data. Some argue that goalkeeper's saves are largely random and their performance in one season has little correlation with their performance in the following season (11tegen11, 2014). However, we investigated this briefly by looking at shot stopping trends at various points throughout the 2017/2018 season and at this time see no reason to believe shot stopping is due more to luck than skill.

Our positional deviations analysis works under an assumption that the average goalkeeper position is most optimal, but that may not be the case. When developing these methods, we first began by extrapolating an xG model to find the goalkeeper position that minimized the xG for each individual shot. Unfortunately, under that framework we were extrapolating too far outside of the domain of this sample size and at times produced illogical results. We conclude that, for now, this is the best proposed methodology to investigate positioning in the given sample.

Lastly, we have little knowledge of the actual value of collecting crosses and distributing the ball well. Therefore, we can say one goalkeeper is more aggressive and better at recognizing the attack than another, but quantitatively, how does that compare to the goals he saved above average? The xC model extends easily into a causal inference framework to quantify the alluding *value* of collecting claimables. Defining the decision to come and collect a cross as a treatment, we can match treated cases (where a goalkeeper came to collect) to control cases (where a goalkeeper did not come to collect) based on similar xC. If we define the potential outcome as the xG against each goalkeeper or xG for a goalkeeper's team in the ~45 seconds following a claimable, then we can take the average difference in xG for treatment and control cohorts while controlling for differing attacking and defensive systems. This would finally give us a value of collecting claimables relative to GSAA.

The models trained for this analysis and the metrics derived can easily be extended to construct age curves and career trajectories. This will help clubs immensely, as they can recognize talented goalkeepers early in development and signal younger goalkeepers from weaker leagues that can be acquired for a cheaper price. We will also be able to calculate more precise valuation of older goalkeepers by better understanding, on average, when goalkeepers are expected to reach their prime and how long they can play at it.

Evaluating players in soccer is difficult, evaluating goalkeepers is even harder. With that said; scouting, recruitment and player development is the most important part of any professional sports organization, it is also the best way to gain a competitive advantage. This framework outlines a straightforward, easy to understand, and fairly simple to implement methodology to profile goalkeepers and shrink scouting lists to the goalkeepers that best fit your system. The data driven evaluation framework is the most comprehensive and objective evaluation system at this time and will improve player valuation in an industry struggling to efficiently assess the most alienating position in professional soccer.

References

[1] 11tegen11. (2014, 02 03). *Never judge a goal keeper by his saves*. Retrieved from 11tegen11: http://11tegen11.net/2014/02/03/never-judge-a-goal-keeper-by-his-saves/

[2] Beygelzimer, A. et. al. (2018, 12 10). Fast Nearest Neighbor Search Algorithms and Applications. Retrieved from https://cran.r-project.org/web/packages/FNN/FNN.pdf

[3] Chen, T. et. al. (2018, 07 09). xgboost: Extreme Gradient Boosting. Retrieved from https://cran.r-project.org/web/packages/xgboost/xgboost.pdf

[4] FC rSTATS. (2018, 10 16). *Uniqueness Passing Model.* Retrieved from https://github.com/FCrSTATS/Statsbomb_WorldCupData/blob/master/Uniqueness%20Passing%20Model.pdf

[5] *Most Valuable Transfers*. (2018, 12 06). Retrieved from transfermkt: https://www.transfermarkt.com/transfers/wertvollstetransfers/statistik/top/plus/0/galerie/0?land_id=&ausrichtung=Torwart&spielerposition_id=&w_s=&art=nf

[6] Perry, E. (2015). *Goalie Stats*. Retrieved from Corsica 2.0: http://corsica.hockey/

[7] R Core Team. (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/.

[8] Schuckers, M. E. (2011). "DIGR: A Defense Independent Rating of NHL Goaltenders using Spatially Smoothed Save Percentage Maps". *MIT Sloan Sports Analytics Conference.* Boston, MA. Retrieved from https://pdfs.semanticscholar.org/5ae2/74cb131011799573cb14041ab8b5d570ae07.pdf

[9] Sears, B. (2015, 11 06). *PREMIER LEAGUE: A POINT PER GOAL, BUT KEEPING THEM OUT WORTH MORE*. Retrieved from sportingintelligence: http://www.sportingintelligence.com/2015/11/06/premier-league-a-point-per-goal-but-keeping-them-out-is-worth-more-061101/

[10] Smith, P. (2018, 05 08). *Premier League: The cost of relegation: what is the financial impact of dropping out of the Premier League?* Retrieved from Sky Sports: https://www.skysports.com/football/news/11661/11358620/the-cost-of-relegation-what-is-the-financial-impact-of-dropping-out-of-the-premier-league

[11] statista. (n.d.). *UEFA Champions League revenue distribution to clubs in the 2017/18 season (in million euros)*. Retrieved 12 06, 2018, from https://www.statista.com/statistics/247156/uefa-champions-league-revenue-distribution-to-clubs/

[12] T.A.W. (2018, 02 09). Why football's goalkeepers are cheap and unheralded. *The Economist*. Retrieved from https://www.economist.com/game-theory/2018/02/09/why-footballs-goalkeepers-are-cheap-and-unheralded

[13] Trainor, C. (2014, 10 21). Retrieved from StatsBomb: https://statsbomb.com/2014/10/goalkeepers-how-repeatable-are-shot-saving-performances/

[14] Wickham, H. et. al. (2018). Create Elegant Data Visualisations Using the Grammar of Graphics. *ggplot2*. Retrieved from https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf

[15] Yam, D. (2018, 05 30). StatsBombR. Retrieved from https://github.com/statsbomb/StatsBombR

Appendices

Appendix A1: Claimables

We define claimables as all passes that are labelled in StatsBomb data as a "cross" plus all high passes, whose paths intersect the box at some point. Typical soccer jargon usually refers to goalkeepers "collecting crosses" for simplicity. However, a large proportion of the balls goalkeepers collect are actually direct, long balls that travel little in the horizontal direction and are not referred to as a "cross" and not tagged in StatsBomb data as a "cross".

Using this definition of claimables, we train our xC model described in Section 4.1. The calibration of our model is shown in Figure 6. The results for goalkeepers in the 2017/2018 EPL season are presented in Table 2.

Figure 7: *Calibration of the expected claims model. No systematic bias in the calibration of the xC model.*

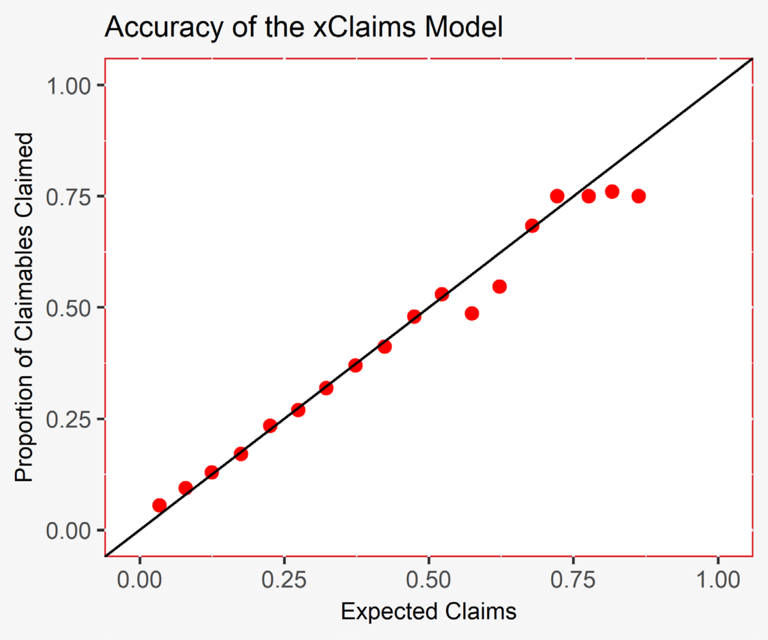


Table 2: *Claimable results from the 2017/2018 EPL season, all first-string goalkeepers are shown*

Goalkeeper	Team	xC	Collected	Claimables	Collected%	xCollected%	CLCAA%
Petr Cech	Arsenal	48.49	77	215	0.36	0.23	0.13
Jonas Lössl	Huddersfield Town	50.17	65	218	0.30	0.23	0.07
Nick Pope	Burnley FC	69.05	84	278	0.30	0.25	0.05
Ederson	Manchester City FC	34.16	42	148	0.28	0.23	0.05
Heurelho Gomes	Watford FC	40.35	50	183	0.27	0.22	0.05
Wayne Hennessey	Crystal Palace FC	44.24	52	196	0.27	0.23	0.04
Kasper Schmeichel	Leicester City FC	55.64	62	244	0.25	0.23	0.03
Joe Hart	West Ham United FC	28.32	30	114	0.26	0.25	0.01
Asmir Begovic	AFC Bournemouth	61.09	65	281	0.23	0.22	0.01
Alex McCarthy	Southampton FC	37.09	39	158	0.25	0.23	0.01
Rob Elliot	Newcastle United FC	34.79	35	144	0.24	0.24	0.00
Thibaut Courtois	Chelsea FC	56.91	55	228	0.24	0.25	-0.01
Lukasz Fabianski	Swansea City FC	64.51	62	282	0.22	0.23	-0.01
Jordan Pickford	Everton FC	69.46	66	315	0.21	0.22	-0.01
Hugo Lloris	Tottenham Hotspur FC	44.57	41	186	0.22	0.24	-0.02
David de Gea	Manchester United FC	46.03	39	209	0.19	0.22	-0.03
Ben Foster	West Bromwich Albion FC	65.35	48	300	0.16	0.22	-0.06
Mathew Ryan	Brighton & Hove Albion	72.41	53	313	0.17	0.23	-0.06
Jack Butland	Stoke City FC	70.71	52	300	0.17	0.24	-0.06
Adrián	West Ham United FC	37.36	24	156	0.15	0.24	-0.09

Appendix A2: Pressures

Pressures are collected in StatsBomb Data whenever a defender enters a certain radius around the player with the ball. Over 200 pressure events are recorded per team per match. Knowing the defensive pressure in a given situation adds more context to event data and helps us better understand player decisions and outcomes. For more on pressures visit, https://statsbomb.com/data/.