Shot analysis in different levels of German football using Expected Goals*

Laurynas Raudonius[1] and Thomas Seidl[2]

[1] ETH Zürich, Rämistrasse 101, 8092 Zürich, Switzerland
lraudonius@student.ethz.ch
[2] VfB Stuttgart 1893 AG, Mercedesstraße 109, 70372 Stuttgart, Germany
T.Seidl@vfb-stuttgart.de

Abstract. Shooting has been one of the most analyzed and researched parts of association football as it directly leads to goals which determine the score of the match. We take a look at it from a previously unseen perspective and analyze if there are differences between four different levels in German football (Bundesliga, Regionalliga, U19 Bundesliga and U17 Bundesliga) in shooting tendencies and efficiency and explore how these change as players get older. To do that we employ statistical analysis and examine the individual weights of Expected Goals models based on logistic regression. We find that players in higher levels tend to be more risky and aim for corners of the goal and are more predictable in terms of their shot origins. A comparison of headers and kicks show that goal likelihood of the latter is much more influenced by whether a shot has happened after a set piece, whereas goal likelihood of headers decreases more steeply with increasing distance from goal. Analysis also reveals that with increasing level goalkeepers tend to be more reliable saving shots at medium height but have a harder time with shots aimed at bottom corners.

Keywords: Football Analytics · Sports Analytics · Performance Analysis · Machine Learning · Expected Goals.

1 Introduction

Johan Cruyff, a Dutch footballer and manager widely regarded as one of the greatest in the history of the sport, once famously said "You have to shoot, otherwise you can't score". And really, just following a very basic deduction, the result of a football match is determined solely by goals scored, and in most cases these are a direct result of shots.
The importance of these particular events occurring in a football match can be proven not only by common sense, but also by historical data, as the number of total shots per match was found to be among the best differentiators between winning and losing teams [7]. Combining this with the shots being relatively basic events, it is no surprise that they are among the most widely researched

* Supported by VfB Stuttgart.

events among academics in the football analytics field [9, 10, 3].
Following this, arguably the most popular regression model in football analytics field called Expected Goals was produced. The model is trained on historical data; it takes shot information such as distance from goal, angle, body part etc. and returns a single number - the likelihood of that shot ending in the back of the net. Expected Goals gained traction in the past decade and is now the core philosophy behind Brentford's successful push for a spot in the Premier League [19]. We will be using our own Expected Goals models and examine their weights to explore differences between different types of shots in four leagues of German football - Bundesliga, Regionalliga, U19 Bundesliga and U17 Bundesliga. As each of the leagues has a higher average player age than the level below (see Figure 1), we will also explore how shooting tendencies and success rates change as players grow older and play in higher levels.

2 Related work

Even though right now there are no particular works that focused on changes in shooting tendencies as footballers get older, there has been some research done on the effects of aging to the psychology of both professional athletes and the general population. Among works on the latter, various studies have found that generally, the tendency to take risks declines with age [17, 5] - one of the proposed causes for that is cognitive aging, but it does not influence the decisions until people are 50+ years old and the difference between, for example, 18 and 35 year olds is not overly noticeable.
A considerable amount of research has been carried out on the effect of age on the physicality and psychology of athletes. A work by Rábano-Muñoz et al. found that there is a large difference between physical demands in under 17 and under 19 year old small-sided football games, with U19 level being even more physically demanding than senior player (aged 20 and over) level [18]. Interestingly, similar results were observed by Benítez-Sillero et al., this time focusing on the psychology of players of different age levels [4]. Players in the U19 level again scored the highest (even compared to older players) in various mental metrics including motivation, attitude control, attention control etc. Somewhat contradictory, Trninić et al. conducted a study that found older athletes in team sports to be more agreeable, conscientious (able to control emotions and impulses) and all around, stable [20]. Staying focused on football, the position of the player also can be taken into account, as players in different positions seem to peak at different ages [24].
As mentioned in the introduction, shots are among the most researched events in football analytics, both by scientists and analytics enthusiasts alike. Most of the studies revolve around Expected Goals, a Machine Learning (mostly regression, but recently Artificial Neural Networks have been used as well) model that produces the likelihood of a shot resulting in a goal based on historical data. The idea was first introduced by Richard Pollard and Charles Reep (widely regarded as the first football analyst) in 1997, their logistic regression model had

distance and angle from goal, whether the player touched the ball before shooting, whether the shot was pressured and whether it happened after a set piece as features [14]. A study that followed found distance to goal and to the nearest defender to be the most important variables that help predict whether a shot will be successful [13]. A surprisingly accurate model that predicted Premier League and Bundesliga shot likelihoods only took distance and angle from goal as features [15]. On the other end of the spectrum, football data providers nowadays have highly complex Expected Goals models with numerous features such as goalkeeper position or ball height at the time of impact - in some cases their models are no longer even based on regression [1, 22, 23].

3 Methodology

As mentioned in the introduction, we will first analyze the statistics between the four leagues and then examine the performance of Expected Goals models with different designs.

3.1 Data

The analysis is based on anonymized event data collected by Wyscout and consisted of shots from 2020/2021 and 2021/2022 seasons in four different German football leagues [3], general facts one these can be found in Table 1. While U17 and U19 Bundesligas have constraints on player age, in Regionalliga there are many reserve teams with younger players. Those particular four leagues were selected as they are a somewhat general pathway for players from academies: going from U17s to U19s, to reserves and finally to Bundesliga, with each representing a next level of German football pyramid.

Table 1: Facts of the four leagues.

	Bundesliga	Regionalliga	U19 League	U17 League
Tier	1st	4th	1st for U19s	1st for U17s
ø-Age	25.9	24.7	18.5	16.9
No. of matches	597	949	509	573
No. of shots	13461	19844	11617	12213

Every shot in the dataset has the following attributes [2, 12]:

- X and Y coordinates
- body part qualifier
- whether the shot was after a set piece
- whether the shot was on target
- whether the shot resulted in a goal
- what part of the goal did the shot go in

[3] Some matches were missing from the datasets as they were cancelled or not yet recorded, hence the odd number of matches in leagues.

3.2 Statistical analysis

Across the four leagues, we will analyse general trends, differences in frequency of shot destinations, shot origins and goalkeeper performance with respect to different zones in the goal. Additionally, we will use analysis of variance (ANOVA) to investigate if the leagues differ significantly.

3.3 Expected goals models

Design considerations As mentioned in the literature review, state of the art Expected Goals models are highly complex, have numerous features and sometimes replace regression with neural network as the base model, which in turn no longer allows them to examine individual weights to explore the effect each parameter has on the goal probability. This is all done to achieve highest possible accuracy, also noteworthy is that robust training is made possible by large amounts of historical data. As our ultimate goal isn't perfect accuracy and we have access to a limited amount of data, we have gone a different direction. Our models are quite simplistic as we have sacrificed some of the accuracy to prevent possible overfitting and obtain robust models that could provide some general insights into how the effect of variables changes based on league. Another argument for simple models is the amount of data we have - according to research, for a robust complex model data from at least five seasons might be necessary [16], and sadly we don't have that luxury. Since we want to examine actual weights of the models, we will separate them in two parts - one for foot shots and one for headers. A single model could be tailored to accommodate both types of shots using dummy variables [8], however it would likely make weight examination cluttered.

Model As logistic regression is a model most widely used to predict the probability of one event taking place (in our case the event is goal), we decided to use it for both models. The particular implementation to obtain the models was R's Generalized Linear Model (*glm* command), it was trained on 80% and tested on the remaining 20% of the shots and did not use regularization as we want to examine individual weights. Following other works on Expected Goals [14, 15, 13], we transformed X and Y coordinates to distance from the center of the goal and angle from the center of the goal in radians. Since in the general statistics shot success in all leagues seemed to be heavily influenced by whether the shot was a result of a set piece, it was also included as a parameter for our models. As we do not really have any more information on the goals, these three numbers will serve as our features, it also leaves us with relatively simple models to examine the weights of. You can find the mathematical definition of our model in Equation 1. Because for both models we have the same features, they are defined exactly the same way - they will just be trained on different data.

$$G(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 * x_{distance} + \beta_2 * x_{angle} + \beta_3 * x_{isSetPiece})}} \tag{1}$$

4 Results

4.1 Statistical analysis

Analysis of variance Using ANOVA we found that the shot distances vary significantly between leagues ($p_{value} < 2 * 10^{-16}$), whereas goal zone efficiencies and frequencies and shot origin frequencies do not seem to be significantly different in different levels ($p_{value} > 0.05$). However, we can still examine individual numbers of these to extract minor patterns.

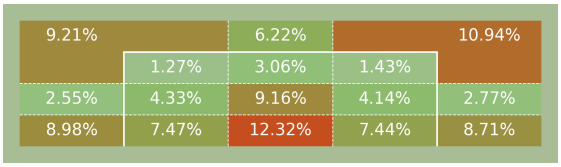
Table 2: General statistics of the four leagues.

Measure	Bundesliga	Regionalliga	U19 League	U17 League
Shots	13461	19844	11617	12213
Goals	1594	2355	1531	1737
Shots per match	22.55	20.91	22.82	21.31
Goals per match	2.67	2.48	3.01	3.03
Goals per shot	0.12	0.12	0.13	0.14
Saved shots, %	25.67%	26.27%	26.40%	27.72%
Off-target shots, %	62.69%	62.08%	60.73%	58.43%
On-target shots saved, %	68.56%	69.06%	66.92%	66.20%
Headers, %	17.44%	15.47%	13.59%	13.98%
Header success rate	0.13	0.13	0.15	0.16
Left foot shots, %	31.73%	31.14%	32.59%	31.97%
Left foot shot success rate	0.11	0.11	0.12	0.14
Right foot shots, %	50.83%	53.39%	53.82%	54.06%
Right foot shot success rate	0.12	0.12	0.14	0.14
Penalty box shots, %	64.12%	61.54%	60.02%	60.27%
Penalty box shot success rate	0.16	0.17	0.19	0.20
Shots from set pieces, %	20.57%	22.31%	20.95%	21.98%
Set pieces success rate	0.10	0.11	0.12	0.14
Average distance, m (shots)	17.21	17.58	18.09	17.89
Average distance, m (goals)	12.02	11.75	12.63	12.57

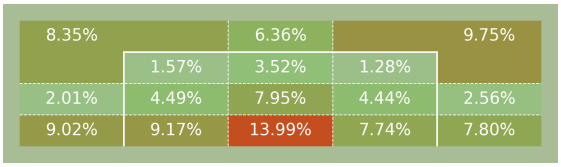
General statistics & Shot destinations Looking through the general statistics that can be found in Table 2, there are both some expected results which we have anticipated and unexpected findings which contradict our prior hypotheses. With Bundesliga being the highest level of the four, we knew it was a physical league (especially compared to U17 and U19 levels), so it having the biggest share of headers should not surprise anyone too much. Similarly to the percentage of shots that are taken within the penalty box - at the highest level of German football we expected the players to take the best quality chances.
On the other hand, there are some findings that, at least initially, raised our eyebrows. For one, with Bundesliga players obviously being of the highest level and, likely, more ambidextrous than younger ones, we expected a larger share of

left foot shots there, when in reality the scores between leagues are very much similar. Another unexpected result was the success rates of every single type of shot decreasing as the players got older, with Bundesliga players being the least efficient. Adding to that, we certainly did not expect them to on average have the most off target shots of the four leagues. Let's explore this counter-intuitive phenomenon further.

One possible explanation for that is that the tempo of the game is a lot higher in higher levels and players don't have that much time to prepare an accurate shot. To add to that, we can also take a look at where the shots actually went - for example, let's examine the difference between Bundesliga (Figure 1a) and U17 League (Figure 1b), as the difference in Shot off target % is the largest between these two. There is a substantial difference in what share of the shots are aimed towards the middle, which generally is the least challenging portion of the goal for the goalkeeper. This particular part of goal is more favoured by the young players and even though it does count as a shot on target, the expected return from it is quite small. Players in the 1st tier, on the other hand, chose the sides of the goal more frequently, the difference is most noticeable in the top corners, which unsurprisingly are the most efficient shot destinations if the player is able to guide the ball there, especially with the goalkeepers getting better in higher levels. The effectiveness of such a decision is also backed up by research which suggests that in order to score efficiently, players should aim very close to the inside post [21].



(a) Bundesliga

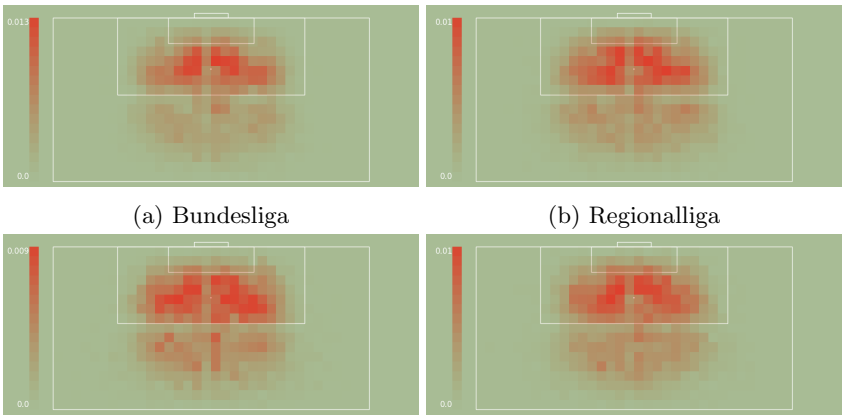


(b) U17 Bundesliga

Fig. 1: Differences in shot destinations between leagues.

Shot origin Now that we've covered shot destinations, let's look into where the shots came from in the four leagues - the frequency heatmaps can be found

in Figure 2[4]. For this, the final third of the pitch was divided in 2x2 meter squares, similarly to a proposition already observed in research on dangerousity in football [11].



(a) Bundesliga (b) Regionalliga

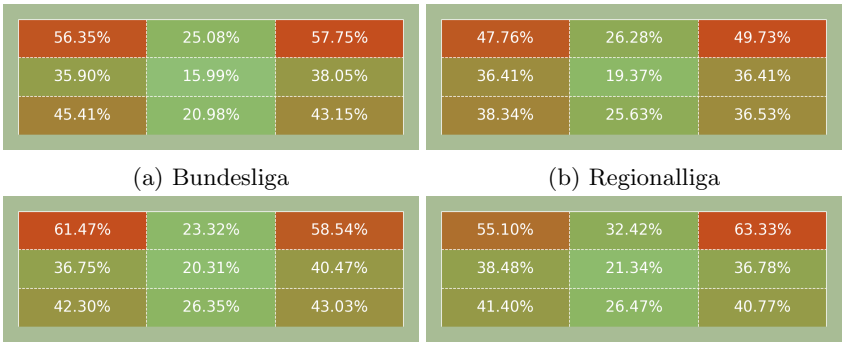
(c) U19 Bundesliga (d) U17 Bundesliga

Fig. 2: Differences in shot locations between leagues. The frequency of the location corresponds to the color of the square, with most frequent locations being colored solid red.

To begin with, it has to be said that players in Bundesliga seem to shoot less from outside the box than players from other leagues. This is backed up by the already discussed general statistics. A consequence of that is that in Bundesliga, the squares that are inside the penalty box have a larger frequency - the maximum observed in Bundesliga was 0.013 whereas in the other leagues that number was around 0.010. This would suggest that Bundesliga players are more predictable, as they choose to shoot from the places in the pitch they know very well, while shots from younger players are more evenly distributed in that respect.
One more phenomenon that can be observed in the graphs is the asymmetry between left and right parts of the pitch, with the right side having more shots in each of the leagues. This, although interesting, can easily be attributed to more players being right footed and being more comfortable with shooting from the right side.

[4] The column to the left of the penalty mark and the row below penalty box being considerably less intense than their neighbours is attributed to limitations in the data as shot positions are represented as integers. The column and row in question have only two possible y-positions associated with it while their neighboring rows and columns have three positions, therefore these particular two have less recorded shots.

Goalkeeper performance The last thing we will cover in this section is performance of the goalkeepers in the different leagues. The most important part of the general statistics table for this is the 'On-target shots saved, %' row. From this we can see that there is a relatively significant disconnect between two leagues with higher average age and two youth leagues, with goalkeepers at the higher level saving a bigger share of shots they face.
We can examine their performance faced with shots heading into different parts of the goal as well in Figure 3. The first very obvious thing is that the top corners are the worst zones for the goalkeeper to concede a shot in, shots that end up in there have 30-40% more chance to end up in a goal than, for example, shots straight to the middle of a goal. This in a way backs up the already discussed tendency of Bundesliga players to aim for the top corners even if the probability of missing the target is large.
Another tendency of the goalkeepers is that the ability to save shots at medium height seems to increase with age quite linearly, with goalkeepers at every level being better there than their counterparts on level below. On the other hand, bottom corners are zones that display an opposite trend - except for goalkeepers in Regionalliga, the ones playing in a higher level seem to have a harder time saving shots to bottom corners than the ones playing at U17 or U19 level.



(a) Bundesliga (b) Regionalliga

(c) U19 Bundesliga (d) U17 Bundesliga

Fig. 3: Differences in success rate for different zones in goal across the leagues. Most efficient locations are colored solid red.

4.2 Expected Goals models

Performance In order to examine individual weights of the trained models, we first must be sure that we have models that are accurate. To check that, we compared the performance of our models on unseen shots against performance of wyscout's Expected Goals model which we treated as the baseline. We have chosen Brier score as our performance measure [6] because in comparison to

AUROC it punishes uncalibrated models and has been used to evaluate Expected Goals models in the past [16].
Surprisingly, our model was even able to slightly outperform wyscout's in some cases and all around had a very similar performance to the baseline. The full scores can be found in Table 3.

Table 3: Brier score comparison between our and wyscout models accross the leagues. Smaller Brier Score suggests better performance.

	Headers		Kicks	
League	Our	Wyscout	Our	Wyscout
Bundesliga	0.086	0.091	0.088	0.086
Regionalliga	0.096	0.098	0.090	0.087
U19 League	0.111	0.111	0.095	0.091
U17 League	0.131	0.128	0.099	0.096

Since our models have very similar performance to the baseline, we can examine individual weights and expect to find meaningful results. These differences also suggest that tailoring a model to a specific league does not seem to make it perform better, which is also backed up by research [16].

General insights A very obvious and quite surprising result was that in all of the 8 models (2 models * 4 leagues), the angle was found to be an insignificant parameter. This, although unexpected, could be explained by the fact that players do not really attempt to shoot from very acute angles that often, so there might not be enough data for the model to deduce that those types of shot rarely go in.
Aside from the angle, every model recognized the intercept, distance and whether the shot was the result of a set piece as significant variables. In all models the distance weights are negative, which means that with increasing distance from goal while holding other variables constant, the probability of the shot resulting in a goal decreases. This is very much expected and quite obvious.
On a more interesting note, all set piece weights are also negative, which means that a shot being a result of a set piece reduces the goal likelihood as well if other variables don't change. Before carrying out the project we could not really anticipate this, but, after examining the general statistics and seeing that shots from set pieces are indeed less likely to end up in the back of the net than other shots, this makes sense.

Headers The trained model weights for headed shots can be found in Table 4.

Taking a look in the particular values in distance weights, there seems to be a disconnect between the first three leagues and the U17 Bundesliga. In U17 level the goal likelihood seems to be somewhat less connected to distance from goal than elsewhere. Looking at the set piece weights, the magnitude of those has to

Table 4: Weights of the header model across the leagues.

League	Intercept	Distance weight	Set piece weight
Bundesliga	0.60	-0.26	-0.27
Regionalliga	0.50	-0.26	-0.13
U19 League	0.78	-0.26	-0.22
U17 League	0.65	-0.22	-0.39

Table 5: Weights of the foot shot model across the leagues.

League	Intercept	Distance weight	Set piece weight
Bundesliga	0.81	-0.17	-0.68
Regionalliga	0.96	-0.18	-0.53
U19 League	0.98	-0.17	-0.59
U17 League	0.81	-0.16	-0.46

be taken into account - set pieces seem to have three times as much influence on the Expected Goals from headers in the U17 League than in Regionalliga.

Kicked shots The trained model weights for foot shots can be found in Table 5.

Examining the intercept values, it seems that Regionalliga and U19 Bundesliga players are in general better at converting shots with their feet than their counterparts in Bundesliga and U17 Bundesliga. Similar to headers, the U17 Bundesliga shots seem to be the least affected by increasing distance from goal, whereas in Regionalliga goal likelihood decreases most rapidly with an increase in distance.

Comparison between headers and kicks Comparing the values between the header and foot shot models, we can first observe that distance from goal seems to have a bigger negative effect to headers than to kicks. This is very much expected and intuitive, as headers tend to be scored from closer range, moving the player away from the goal while holding other variables constant then decreases the goal likelihood more drastically. A more interesting phenomenon can be examined in the intercept and set piece weights. Even though from general statistics we found that headers are a more efficient shot type than kicks in all leagues, foot shot models from all leagues having larger intercepts seem to contradict that. However, set piece weights cannot be forgotten, and these are also larger in magnitude than the header shots, but negative. If we combine the two pieces of information, it no longer contradicts the general statistics, but rather tells us that whether a shot happened after a set piece has a 3 times larger negative effect on the goal likelihood to kicks than it does to headers.
After analysis of variance (ANOVA) we found that header ($p_{value} = 4.24 * 10^{-12}$) and kicked shot ($p_{value} < 2 * 10^{-16}$) Expected Goal values from our models

varied significantly between the four leagues - the distributions can be observed in Appendix A.

5 Conclusions

In this work we analysed and presented differences between four football leagues in Germany (Bundesliga, Regionalliga, U19 Bundesliga and U17 Bundesliga) in terms of shooting tendencies and efficiency. We found players in higher levels to be more risky and aim for top corners even though the possibility of missing the target is higher there. They are also more predictable in terms of where they shoot from. Goalkeepers seem to get better at saving shots at body level but have a harder time with shots to bottom corners as they age. We also found out that distance from goal distributions are significantly different between the four leagues.
In addition to statistical analysis, we designed, trained and tested Expected Goals models to examine their weights. They achieved very similar performances to wyscout's Expected Goals model.

References

1. StatsBomb xG, https://statsbomb.com/articles/soccer/statsbomb-release-expected-goals-with-shot-impact-height/
2. Wyscout shot definintions & parameters, https://dataglossary.wyscout.com/shot/
3. Armatas, V., Yiannakos, A., Papadopoulou, S., Skoufas, D.: Evaluation of goals scored in top ranking soccer matches: Greek Superleague 2006-07. Serbian Journal of Sports Sciences **3**(1), 39–43 (2009)
4. Benítez-Sillero, J.d.D., Martínez-Aranda, L.M., Sanz-Matesanz, M., Domínguez-Escribano, M.: Determining Factors of Psychological Performance and Differences among Age Categories in Youth Football Players. Sustainability **13**(14), 7713 (Jul 2021)
5. Bonsang, E., Dohmen, T.: Risk attitude and cognitive aging. Journal of Economic Behavior & Organization **112**, 112–126 (Apr 2015)
6. Brier, G.: Verification of Forecast Expressed in Terms of Probability. Monthly Weather Review **78**(1), 1–3 (1950)
7. Castellano, J., Casamichana, D., Lago, C.: The Use of Match Statistics that Discriminate Between Successful and Unsuccessful Soccer Teams. Journal of Human Kinetics **31**(2012), 137–147 (Mar 2012)
8. Draper, N.R., Smith, H.: Dummy Variables. In: Applied Regression Analysis. Wiley Series in Probability and Statistics, Wiley, 1 edn. (Apr 1998)
9. Lago-Peñas, C., Lago-Ballesteros, J., Dellal, A., Gómez, M.: Game-related statistics that discriminated winning, drawing and losing teams from the Spanish soccer league. Journal of Sports Science and Medicine **9**(2010), 288–293 (2010)
10. Lago-Peñas, C., Lago-Ballesteros, J., Rey, E.: Differences in performance indicators between winning and losing teams in the UEFA Champions League. Journal of Human Kinetics **27**(2011), 135–146 (Mar 2011)
11. Link, D., Lang, S., Seidenschwarz, P.: Real Time Quantification of Dangerousity in Football Using Spatiotemporal Tracking Data. PLOS ONE **11**(12), e0168768 (Dec 2016)

12. Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., Giannotti, F.: A public data set of spatio-temporal match events in soccer competitions. Scientific Data **6**(1), 236 (Dec 2019)
13. Pollard, R., Ensum, J., Taylor, S.: Estimating the probability of a shot resulting in a goal: the effects of distance, angle and space. International Journal of Soccer and Science **2**(1), 15 (2004)
14. Pollard, R., Reep, C.: Measuring the effectiveness of playing strategies at soccer. Journal of the Royal Statistical Society: Series D (The Statistician) **46**(4), 541–550 (Dec 1997)
15. Rathke, A.: An examination of expected goals and shot efficiency in soccer. Journal of Human Sport and Exercise **12**(Proc2) (2017)
16. Robberechts, P., Davis, J.: How Data Availability Affects the Ability to Learn Good xG Models. In: Brefeld, U., Davis, J., Van Haaren, J., Zimmermann, A. (eds.) Machine Learning and Data Mining for Sports Analytics, vol. 1324, pp. 17–27. Springer International Publishing, Cham (2020)
17. Rolison, J.J., Hanoch, Y., Wood, S., Liu, P.J.: Risk-Taking Differences Across the Adult Life Span: A Question of Age and Domain. The Journals of Gerontology: Series B **69**(6), 870–880 (Nov 2014)
18. Rábano-Muñoz, A., Asian-Clemente, J., Sáez de Villarreal, E., Nayler, J., Requena, B.: Age-Related Differences in the Physical and Physiological Demands during Small-Sided Games with Floaters. Sports **7**(4), 79 (Apr 2019)
19. Tippett, J.: The Expected Goals Philosophy: A Game-Changing Way of Analysing Football. Independently Published (2019)
20. Trninić, V., Trninić, M., Penezić, Z.: Personality differences between the players regarding the type of sport and age. Acta Kinesiologica **10**(2), 69–74 (2016)
21. Vars, F.E.: Missing Well: Optimal Targeting of Soccer Shots. CHANCE **22**(4), 21–28 (Sep 2009)
22. Witmore, J.: The Analyst - What are expected Goals? (Mar 2019), https://theanalyst.com/eu/2021/07/what-are-expected-goals-xg/
23. Witmore, J.: Evolving Expected Goals (xG) (Mar 2022), https://theanalyst.com/eu/2022/03/evolving-expected-goals-xg/
24. Worville, T.: What age do players in different positions peak? (Nov 2021), https://theathletic.com/2935360/2021/11/15/what-age-do-players-in-different-positions-peak/

A Box plots of significantly different distributions

Here you can find the box plots for distributions we've encountered that, according to ANOVA, significantly differ ($p_{value} < 0.05$) between the four leagues.

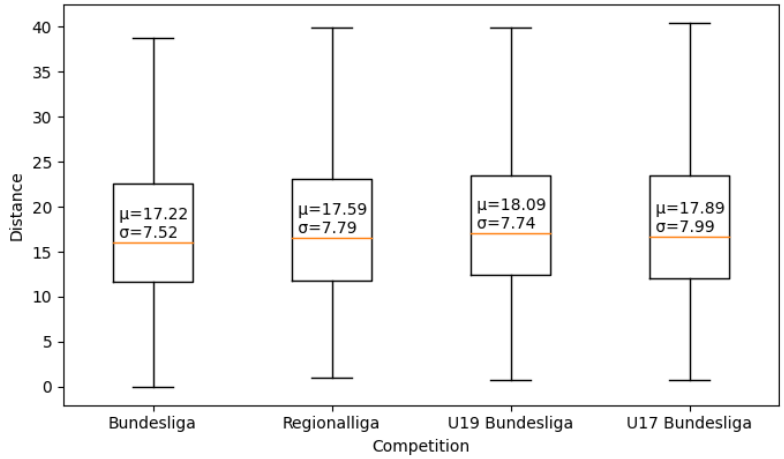


Fig. 4: Distance to goal distributions in different leagues

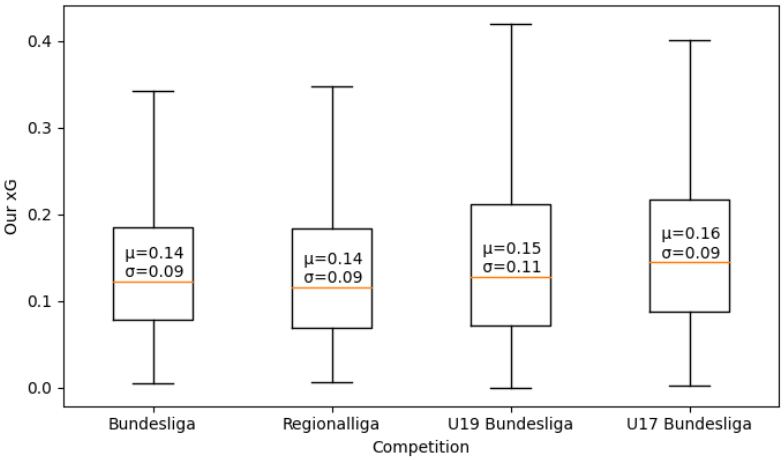


Fig. 5: Our xg score distributions for headers in different leagues

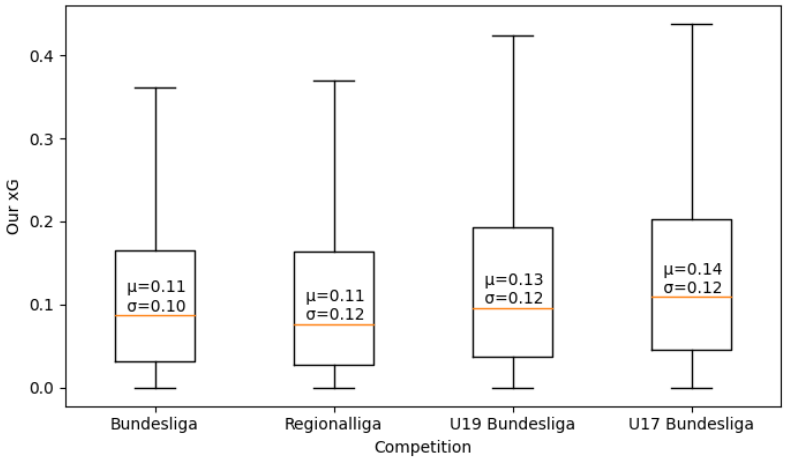


Fig. 6: Our xg score distributions for kicks in different leagues