Evaluating Player Performances in Football: A Debiased Machine Learning Approach

Robert Bajons
Institute for Statistics and Mathematics, Vienna University of Economics and Business
Vienna, Austria
robert.bajons@wu.ac.at

ABSTRACT

In this work, a novel framework for the evaluation of individual football (soccer) players using event stream data is introduced. Applying a debiased machine learning approach (DML), we estimate the contribution of players to a possession sequence, i.e. a sequence of consecutive on-ball events stopped either by the opponent team gaining the possession or by an action of the referee. The estimates are then used to derive a metric to rate players, which is able to account for team strengths and game context. To show the potential of our novel rating approach we compare it to existing ratings by measuring the quality of match outcome forecasts generated when the ratings are used as predictor variables.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning algorithms**; **Model development and analysis**.

KEYWORDS

Regression, Double/Debiased Machine Learning, Ratings, Sports Analytics

ACM Reference Format:
Robert Bajons. 2023. Evaluating Player Performances in Football: A Debiased Machine Learning Approach. In *2023 6th International Conference on Mathematics and Statistics (ICoMS 2023), July 14–16, 2023, Leipzig, Germany.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3613347.3613368

1 INTRODUCTION

Evaluating player performances in professional football is a challenging task that is becoming more and more relevant for the professional football industry. On the one hand this problem is highly relevant for media, advertising and betting companies involved in the sport in order to increase fan engagement, provide novel insights ([16]) and shed light on highly debated topics such as the FIFA player of the year award ([14]) as well as to estimate winning odds and analyze betting strategies ([18], [17]). On the other hand sports teams increasingly face decisions where they have to scout and recruit new players in order to remain compatible, while at the same time being compliant with budget restrictions ([22], [19]). The traditional approach for decision-making in football, where humans qualitatively assess situations and performances within a game, is more frequently reinforced by the use of statistical methods. However football suffers from its low-scoring nature, which compared to the other sports makes it difficult to accurately measure factors that influence the game as well as to evaluate players ([26], [5], [1], [6]).

To assess player performance, typically some kind of rating system is used by which players can be compared and ranked. In football, the most traditional approach is to use discrete count-based statistics, such as completed passes, shots on target, assists, goals, etc. to measure the strength of players. However, doing so disregards the context in which actions are performed and thus makes such statistics unreliable and of moderate usefulness ([26]). Fostered by the rapid advances in collecting and creating data, more data-driven and granular methods have recently emerged in order to create novel rating systems. Hvattum and Gelade (2021) [10] divide the existing approaches into two main categories: bottom-up and top-down rating systems. The former assess player performances by assigning values to each action performed and then aggregate them for each player over the course of a relevant period (i.e. a match or a season). Top-down ratings on the other hand evaluate players by breaking down the whole team performance and distributing credit onto players involved.

A popular new metric developed, which slowly gets acknowledged among football experts, is known as expected goals (xG), which uses statistical models to assign a probability of scoring a goal to each shot made during a match (cf. [23], [1]). While xG is one of the early bottom-up approaches to measure player performance, it only accounts for shots, ignoring other important actions such as passes or crosses. A large body of research thus focuses on assessing the quality of other actions. Specifically passes have found a lot of attention as they are the most common actions in soccer. Szczepański and McHale (2016) [26] and Håland et al. (2019,2020) [12], [11] measure passing ability by deriving generalized additive models for different aspects of passes such as success, difficulty, risk and potential. Other approaches focus on the location on the pitch and ball movement ([8]) or a mix of spatio-temporal data and human classification to model values of passes ([3]). Ultimately bottom-up rating systems have been developed that do not consider specific parts of player contribution but rather aim to asses overall performance. Pappalardo et al. (2019a) ([20]) divide players into roles and create a feature vector based on all actions performed to derive their PlayeRank rating system. Similarly Decroos et al. (2019) [5] present VAEP, a bottom-up rating system that evaluates players by measuring a broad set of actions based on their likelihood to change the scoreboard.

ICoMS 2023, July 14–16, 2023, Leipzig, Germany

ACM ISBN 979-8-4007-0018-7/23/07.
https://doi.org/10.1145/3613347.3613368

A popular class of top-down ratings are (adjusted) plus-minus (PM) ratings. Originating in sports such as basketball ([25]) and ice hockey ([15], [7]) adjusted PM ratings use regularized regression techniques to assess the contribution of players to team performances, e.g. the goal difference of two teams in a specific segment of a game. Since the work of Sæbø and Hvattum (2015) [24] the idea has been developed in the context of soccer. Hvattum (2019) [9] provides a comprehensive review of adjusted PM ratings in several sports including football. Lately [14] advanced the plus-minus metric by combining it with bottom-up approaches such as xG. A different top-down approach is taken by Wolf et al. (2020) [27], who propose an Elo rating approach that ranks players based on the differences in actual and expected scores.

This paper is focused on the evaluation of players by using event stream data, i.e. granular data that provides information on each action performed during a match. We aim to contribute to the existing literature of player rating models by developing a framework, which is situated at the intersection of existing bottom-up and top-down models. We consider small segments of matches that we term *possessions* and assess player performance by measuring the contribution of players to these possessions. Thus the model could be classified as top-down approach, however, we take into account granular information on actions within the possession sequence. To measure a player's contribution a method from causal inference theory termed double/debiased machine learning (DML, cf. [4]) is employed, which provides means to get unbiased estimates of contribution effects. These estimates are then used to derive a player rating, which accounts for quantity and quality of contributions to the game. The proposed metric is thus able to rate players while accounting for the teammates and team strengths as well as context of the game, i.e. the circumstances in which players contributed to the game. To test the validity of the metric a similar approach as in Hvattum and Gelade (2021) [10] is used. We compare the ability to predict match outcomes using our metric with existing approaches such as VAEP ratings and ELO team ratings. To do so we use two state-of-the-art models for match outcome predictions, a bivariate Poisson model ([13]) and an ordinal logistic regression model ([2]).

2 METHODOLOGY

In this section, we discuss details about the derivation of a player ranking. The idea is to use specific regression models, from which a meaningful strength parameter for each player can be extracted. First, we establish our statistical units, namely possession sequences and explain how to set up regression variables and outcome variable in our framework. Next, we present a methodology on how to estimate the relevant strength parameters using a debiased machine learning approach (DML). Finally, the DML methodology is used in the context of our football data in order to derive a meaningful parameter of playing strength for each player in our dataset.

2.1 Data

The raw data used for this project is a collection of a new data format called event stream data or soccer logs. The data was collected and provided by Wyscout, a football data company located in Europe, and made available via figshare, an online open access data repository. In summary, such event stream data describes all events happening during a football game. Any action or event is annotated by important attributes such as a timestamp, location, action type, player involved, and more. Figure 1, provides a snippet of the event stream data from a match of FC Barcelona. More details about the nature of the data as well as the specific dataset used can be found in a description by Pappalardo et al. (2019b) [21].

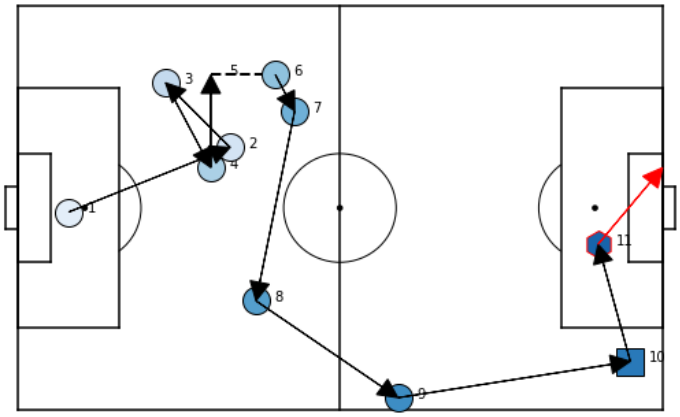
The raw event dataset is then used as a basis in order to create a dataset of possessions. A possession is defined as a sequence of consecutive on-ball events, which ends either by the opponent team gaining the possession or by an action of the referee, i.e. a foul, the ball going out of the pitch or the end of a period. From the possession sequences a number of relevant characteristics are extracted that are deemed important or influential for the probability of scoring a goal from the possession. The features contain spatial (start/end location of possession, total distance covered, goal distance, goal angle, etc.), temporal (game time, speed of possession, etc.) and discrete (freekick/corner indicator, number of actions, score differential, etc.) information. As the main goal is the valuation of players, it is necessary to extract the information on which player was involved in a possession sequence. Extracting information on offensive players involved in a possession is straightforward from the data, however, information on the opposing players would also be desirable. Since event stream data only provides information on the on-ball actions, it is difficult to extract meaningful opponent player information. One way to incorporate such information is to simply record all players on the field for a possession. Thus the regression matrix X_{mod} in this work contains $N_f + N_{inv} + N_{of}$ columns and N_p rows. N_f represents the number of contextual features. Each of the involvement columns N_{inv} is a binary variable for a specific player, that is 1 whenever the player was involved in the possession and 0 otherwise. Each of the N_{of} columns indicates whether a specific player was on the field for the possession team (indicated by 1) or on the field for the defensive team (indicated by -1)[1]. The number of rows N_p depends on the number of possessions analyzed. As we are interested in deriving a player rating it makes sense to only consider valuable possessions, where we define a possession as valuable if it ends in the last third of the pitch (i.e. sufficiently close to the opponents goal).

Finally, a relevant response variable is extracted, which allows us to analyze the effect of players on possessions. Since goals are the most important part of the game of football, a goal indicator is taken as outcome variable. As football is a low scoring game, one drawback is that this outcome variable is quite imbalanced. While there are other possibilities for outcome variables, such as ordinal type variables or hybrid approaches, it is out of the scope of this paper to consider such cases.

2.2 Deriving a Player Rating

In order to infer a player rating from the possession data a naive idea would be to simply fit a (generalized) linear model and to interpret the resulting model coefficients β. However, as many studies point out ([7], [24], [9], [14]), such an approach is not expected to work in the high-dimensional and sparse setup at hand, due to classical

[1]Typically $N_{inv} = N_{of}$ corresponds to the number of players in the data set. In this work however a ranking is only derived for players that have a sufficient amount of involvements in possessions, thus $N_{inv} < N_{of}$.



	time	actiontype	player	team
1	26m2s	pass	M. ter Stegen	FC Barcelona
2	26m13s	pass	Sergio Busquets	FC Barcelona
3	26m16s	pass	Y. Mina	FC Barcelona
4	26m20s	pass	Sergio Busquets	FC Barcelona
5	26m23s	dribble	Y. Mina	FC Barcelona
6	26m26s	pass	Y. Mina	FC Barcelona
7	26m28s	pass	Iniesta	FC Barcelona
8	26m32s	pass	Piqué	FC Barcelona
9	26m34s	pass	Nélson Semedo	FC Barcelona
10	26m36s	cross	O. Dembélé	FC Barcelona
11	26m37s	shot	L. Suárez	FC Barcelona

Figure 1: A possession from a match from FC Barcelona as provided by event stream data.

issues such as multicollinearity, overfitting and the high imbalance in outcome and independent variables. A remedy used by previous studies is to employ regularization on the coefficient estimates. While such a regularized approach might be appealing and also provide reasonable results if set up properly, it however lacks nice properties, especially clear interpretability of the coefficients, which is important for practitioners.

To overcome these issues we consider the following partially linear model

$$\begin{aligned} Y &= D\theta_0 + g_0(X) + \epsilon, \quad \mathbb{E}[\epsilon|D,X] = 0, \\ D &= m_0(X) + v, \quad \mathbb{E}[v|X] = 0. \end{aligned} \tag{1}$$

In this notation Y represents the outcome variable, D is the treatment variable, which is potentially highly correlated with X, X are confounders, θ_0 is the parameter of interest, ϵ and v are stochastic error terms. Note that g_0 and m_0 are (potentially non-linear) nuisance functions, i.e. we need to estimate them, however, this is not our primary interest. In our setup, Y is the outcome variable of whether the possession resulted in a goal or not. The variable D is the involvement variable of a specific player, that we want to estimate, and X is the feature matrix as described in the previous section, such that $X_{mod} = (X, D)$. As Y is a binary variable, the above model specification is clearly only an approximation to the true underlying data-generating process, nevertheless, it allows for nice interpretation. The estimate for θ_0 in this case can be seen as the level shift in the probability of scoring a goal from a possession when the player corresponding to D is involved in the possession.

In order to estimate the parameter θ_0, Chernozhukov et al. (2018) [4] propose a debiased machine learning approach, which provides means for decreasing the regularization and overfitting bias in the estimation procedure. Their approach is based on the idea of partialling out the effect of the confounders X from equation (1), i.e. rewriting it in the form

$$\begin{aligned} W &= v\theta_0 + \epsilon, \quad \mathbb{E}[\epsilon|D,X] = 0, \\ W &= Y - \ell_0(X), \quad \ell_0(X) = \mathbb{E}[Y|X] = m_0(X)\theta_0 + g_0(X), \\ v &= D - m_0(X), \quad m_0(X) = \mathbb{E}[D|X]. \end{aligned} \tag{2}$$

The estimation of the treatment effect θ_0 thus amounts to executing 2 steps:

(1) Estimate m_0 and ℓ_0 using machine learning techniques of choice, i.e. predict D and Y using X.
(2) Estimate θ_0 by regressing the residuals $\hat{W} = Y - \hat{\ell}_0(X)$ onto the residual of $\hat{v} = D - \hat{m}_0(X)$.

Combining these two steps with a suitable cross-validation strategy can be shown to reduce regularization bias as well as overfitting bias, thus providing to some extent better estimates than using a direct approach for estimation of θ_0. The above two-step procedure was used for estimating the effect θ_0, i.e. the procedure is repeated such that for every player i in the dataset an estimate $\hat{\theta}_{0,i}$ is obtained. As mentioned earlier the resulting estimate is nicely interpretable as the level shift in the probability of scoring when player i is involved. However, the raw estimate is still subject to problems that arise from the imbalance in the dataset. The effect for players with little involvement is overestimated, while the effect for players that highly participate in the game is underestimated. This is due to the fact that is it more difficult to maintain a level of effectiveness when a player is a key player and receiving a lot of passes and thus exhibiting a high amount of involvements. Pure strikers for example mostly are only involved at the end of a play and do not have to make risky passes but rather convert the chances they have. Conversely, midfielders and creative players receive the ball far more often during the game. Furthermore, they have to take more risks by passing the ball through defenses or dribbling through a wall of defenders. Thus similarly estimated values of θ_0 might hide the fact, that one player has a lot of involvements in a lot more build-up possessions than the other player and is therefore much more important to the game and the team. In order to account for this

fact and value players with respect to not only quality (respectively effectiveness) but also quantity, a transformation of the $\hat{\theta}_0$ value is proposed, which takes into account the amount of actions a player has. The metric for rating player i termed possession contribution value (PCV) is given as

$$PCV_i = N_i(\hat{\theta}_{0,i} + \bar{p}), \tag{3}$$

where N_i is the number of involvements of players i in possession over the course of the season and $\bar{p}$ is the proportion of goals scored from the possessions. The rationale behind the PCV metric is the following: Considering a general possession, where no details are known about the possession, a good estimate for the probability of scoring is to simply take the proportion of scoring a goal from a possession in our dataset, so $\bar{p}$. Since $\hat{\theta}_{0,i}$ represents the level shift in the probability of scoring a goal from a possession when player i is involved in the possession, then if it is known that player i is involved in this general possession, adding the level shift $\hat{\theta}_{0,i}$ to $\bar{p}$ represents a more accurate estimate of the scoring probability. Multiplying this by the number of involvements results in an estimate for the average contribution to possessions by a player over the course of the season. Of course, a downside of such an adjustment is that players that played fewer matches during a season due to injuries for example, might be undervalued. However as we only consider one season of data in this work and the goal is to rank players over the course of this period, it is justifiable to use this approach.

2.3 Assessing the Quality of Ratings

When deriving a player rating it is necessary to provide an evaluation framework that is able to objectively quantify whether the rating makes sense. One idea is to analyze the results from the rating process using a domain-specific viewpoint. However, there is no ground truth for a player ranking. Thus relying on opinions from practitioners would not result in an objective evaluation of the metric. Comparing rating results with commonly used key performance indicators such as goal scored, assists or shots provides information on quality to some extent. Nevertheless, the main reason for deriving a more granular player rating system is to observe player strengths beyond classical performance measures which do not account for context.

To overcome the above-mentioned issues, a similar path is taken as in Hvattum and Gelade (2021) [10]. The authors approach is based on the fact that a rating is useful or relevant for football if it can be related to an outcome of interest such as matches won. Thus they derive a validity check which relies on predicting match results using only information from the player rating. To evaluate the player ranking derived in this work, this approach is adapted appropriately. Since only data from the 2017/18 season is used, match outcome data from that season is collected and divided into a training set (all matches up to the end of February 2018) and a test set (matches from March 2018 to the end of the season). In order to predict a match outcome two state-of-the-art models are considered. First, a bivariate Poisson model as discussed by Karlis and Ntzoufras (2003) [13], which models the number of goals by the home and away team jointly and second an ordered logistic regression model as proposed in Arntzen and Hvattum (2021) [2], which uses an ordered categorical variable with three levels as response, namely home win (H), draw (D) and away win. Using the team strength as derived as the average rating from players on that team as a covariate, the models are trained on the training set and the predictive performance of the models is then analyzed on the test set. In order to compare the derived rating to existing ones, four different variants are considered. First, to obtain a baseline comparison, a model is trained without any covariate information (intercept-only model). Second differences in average team VAEP values (cf. [5]) for the home and away team are used as covariates. Third, differences in ELO team ratings before the matches are used as predictor variable (the values are taken from http://clubelo.com/). And lastly the difference in home and away team as per the PCV metric derived is taken as a regressor. To compare the predictive performances of the four models at hand, proper scoring rules as suggested by Arntzen and Hvattum (2021) [2] are used. First, the Brier score

$$BS = \sum_{i=1}^{3}(d_i - p_i)^2, \tag{4}$$

for each match in the test set is computed and second the informational loss is computed, which is given as

$$IL = -\log_2(\sum_{i=1}^{3}(p_i d_i)), \tag{5}$$

where p_i is the model's probability of outcome $i = H, D, A$ and $d_i = 1$, if match ended with outcome i and 0 otherwise.

3 RESULTS

Table 1 displays the top 20 players rated by the presented framework for the 2017/18 season of the 5 big European football leagues. While the list provides interesting insights that might be worthwhile to analyze in detail or from a domain-specific viewpoint, the focus is laid on the validity results presented in Table 2. The predictive performance when using different covariates to measure team strengths is compared. First, it can be observed, that between the two match outcome modeling approaches (bivariate Poisson and ordinal logistic regression), there is no notable difference. Second, the models with covariates for team performance perform similarly, but all of them outperform the intercept-only model. A t-test confirms that these differences are even highly statistical, with p-values far below any usual confidence level (see Table 3). The best performing models, i.e. the models with the lowest predictive loss, are the ones using the ELO team rating. Using the newly derived PCV rating as predictor variable leads to a slightly better score than using VAEP ratings. However, these differences are no longer statistically significant when comparing them via a t-test, (see again Table 3). Furthermore, it has to be mentioned that the ELO ratings use far more data, i.e. more seasons of matches to arrive at their rating (for details we refer to http://clubelo.com/), while for PCV and VAEP only 2017/18 data was accessed.

4 CONCLUSION

In this work, we present a novel methodology for rating football players. The presented approach provides a semi-top-down rating using event stream data and a debiased machine learning approach.

Table 1: Top 20 players based on PCV metric.

	Player	Role	$\hat{\theta}_0$	Inv	PCV	G+A	G+A per 90
1	L. Messi	FW	0.014	1,005	36.060	46	1.38
2	K. De Bruyne	MF	0.007	1,179	33.606	24	0.70
3	Malcom	FW	0.013	783	27.775	19	0.60
4	Luis Alberto	FW,MF	0.011	836	27.236	25	0.84
5	L. Sané	FW	0.021	630	27.185	25	0.93
6	F. Thauvin	FW	0.011	816	27.105	33	1.01
7	T. Kroos	MF	0.007	904	26.296	12	0.48
8	L. Suárez	FW	0.019	630	25.988	37	1.15
9	R. Sterling	FW	0.017	656	25.633	29	1.01
10	C. Eriksen	MF,FW	0.002	1,051	25.565	20	0.56
11	Mohamed Salah	FW	0.019	623	25.389	42	1.30
12	Suso	FW,MF	0.009	807	25.233	12	0.38
13	I. Perišić	FW	0.010	761	24.543	18	0.49
14	K. Walker	DF	0.011	751	24.420	6	0.19
15	Cristiano Ronaldo	FW	0.025	519	24.317	31	1.22
16	Neymar	FW	0.013	693	24.053	32	1.61
17	C. Immobile	FW	0.030	460	23.927	35	1.17
18	L. Insigne	FW	0.001	1,037	23.798	19	0.55
19	Son Heung-Min	FW	0.019	557	23.041	18	0.70
20	D. Payet	MF,FW	0.008	753	22.957	20	0.77

Table 2: Predictive performance as measured by Brier score (BS) and informational loss (IL) of the models with different covariates used. Lower values indicate better predictive performance.

	Bivariate Poisson		Ordinal logistic	
Covariates	BS	IL	BS	IL
Intercept Only	0.638	1.526	0.638	1.526
PCV	0.575	1.400	0.575	1.397
VAEP	0.578	1.410	0.578	1.405
ELO	0.574	1.395	0.573	1.392

Table 3: Restults of individual t-tests comparing each of the 4 models with respect to Brier score (BS) and informational loss (IL) differences. Result only shown for bivariate Poisson model.

Scoring	Group 1	Group 2	Estimate	t	p value
BS	Intercept Only	PCV	0.063	4.474	<0.001
BS	Intercept Only	VAEP	0.06	4.412	<0.001
BS	Intercept Only	ELO	0.065	4.409	<0.001
BS	PCV	VAEP	-0.003	-0.499	0.618
BS	PCV	ELO	0.002	0.218	0.828
BS	VAEP	ELO	0.005	0.563	0.574
IL	Intercept Only	PCV	0.126	4.22	<0.001
IL	Intercept Only	VAEP	0.116	3.814	<0.001
IL	Intercept Only	ELO	0.131	4.246	<0.001
IL	PCV	VAEP	-0.01	-0.789	0.43
IL	PCV	ELO	0.005	0.293	0.77
IL	VAEP	ELO	0.015	0.849	0.397

Using this strategy the rating is able to account for team strength, circumstances of play as well as quantity and quality of players contributions to the game.

The novel framework was applied to data from the 2017/18 season of 5 European leagues, the English Premier League, the French Ligue 1, the German Bundesliga, the Italian Seria A and the Spanish LaLiga. A domain-specific validity check shows that the results from the novel rating system are promising, being able to compete with existing approaches for evaluating football players.

While the results of the work are interesting, we plan to extend the study in future work in two ways. The first goal would be to perform a reliability or robustness analysis for the derived metric. Using more seasons of data as well as different leagues, it would be interesting to see how robust our rating is when analyzing players over the course of a season or whether there are adjustment to be done for leagues that are considered weaker. However, such an analysis is reliant on having access to a sufficient amount of event stream data. Second, we plan to extend the work and consider classical regularized regression techniques, similar to the plus-minus approaches, in order to compare our methodology to a broader range of player rating systems.

REFERENCES

[1] Gabriel Anzer and Pascal Bauer. 2021. A Goal Scoring Probability Model for Shots Based on Synchronized Positional and Event Data in Football (Soccer). *Frontiers in Sports and Active Living* 3 (2021), 53. https://doi.org/10.3389/fspor.2021.624475

[2] Halvard Arntzen and Lars Magnus Hvattum. 2021. Predicting match outcomes in association football using team ratings and player ratings. *Statistical Modelling* 21, 5 (2021), 449–470. https://doi.org/10.1177/1471082X20929881 arXiv:https://doi.org/10.1177/1471082X20929881

[3] Sanjay Chawla, Joël Estephan, Joachim Gudmundsson, and Michael Horton. 2017. Classification of Passes in Football Matches Using Spatiotemporal Data. *ACM Trans. Spatial Algorithms Syst.* 3, 2, Article 6 (aug 2017), 30 pages. https://doi.org/10.1145/3105576

[4] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21, 1 (01 2018), C1–C68. https://doi.org/10.1111/ectj.12097 arXiv:https://academic.oup.com/ectj/article-pdf/21/1/C1/27684918/ectj00c1.pdf

[5] Tom Decroos, Lotte Bransen, Jan Van Haaren, and Jesse Davis. 2019. Actions Speak Louder than Goals: Valuing Player Actions in Soccer. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 1851–1861. https://doi.org/10.1145/3292500.3330758

[6] Javier Fernández, Luke Bornn, and Daniel Cervone. 2021. A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. *Machine Learning* 110, 6 (01 Jun 2021), 1389–1427. https://doi.org/10.1007/s10994-021-05989-6

[7] Robert B. Gramacy, Shane T. Jensen, and Matt Taddy. 2013. Estimating player contribution in hockey with regularized logistic regression. *Journal of Quantitative Analysis in Sports* 9, 1 (2013), 97–111. https://doi.org/doi:10.1515/jqas-2012-0001

[8] L. Gyarmati and R. Stanojevic. 2016. QPass: a merit-based evaluation of soccer passes. *Proceedings of the KDD-16 Workshop on Large-Scale Sports Analytics* (2016), 1–4. http://www.large-scale-sports-analytics.org/Large-Scale-Sports-Analytics/Submissions_files/paperID09.pdf

[9] Lars Magnus Hvattum. 2019. A comprehensive review of plus-minus ratings for evaluating individual players in team sports. *International Journal of Computer Science in Sport* 18, 1 (2019), 1–23. https://doi.org/doi:10.2478/ijcss-2019-0001

[10] Lars Magnus Hvattum and Garry A. Gelade. 2021. Comparing bottom-up and top-down ratings for individual soccer players. *International Journal of Computer Science in Sport* 20, 1 (2021), 23–42. https://doi.org/doi:10.2478/ijcss-2021-0002

[11] Else Marie Håland, Astrid Salte Wiig, Lars Magnus Hvattum, and Magnus Stålhane. 2020. Evaluating the effectiveness of different network flow motifs in association football. *Journal of Quantitative Analysis in Sports* 16, 4 (2020), 311–323. https://doi.org/doi:10.1515/jqas-2019-0097

[12] Else Marie Håland, Astrid Salte Wiig, Magnus Stålhane, and Lars Magnus Hvattum. 2019. Evaluating passing ability in association football. *IMA Journal of Management Mathematics* 31, 1 (04 2019), 91–116. https://doi.org/10.1093/imaman/dpz004 arXiv:https://academic.oup.com/imaman/article-pdf/31/1/91/34157145/dpz004.pdf

[13] Dimitris Karlis and Ioannis Ntzoufras. 2003. Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)* 52, 3 (2003), 381–393. https://doi.org/10.1111/1467-9884.00366 arXiv:https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9884.00366

[14] Tarak Kharrat, Ian G. McHale, and Javier López Peña. 2020. Plus–minus player ratings for soccer. *European Journal of Operational Research* 283, 2 (2020), 726–736. https://doi.org/10.1016/j.ejor.2019.11.026

[15] Brian Macdonald. 2012. Adjusted Plus-Minus for NHL Players using Ridge Regression with Goals, Shots, Fenwick, and Corsi. *Journal of Quantitative Analysis in Sports* 8 (2012).

[16] Ian G. McHale, Philip A. Scarf, and David E. Folker. 2012. On the Development of a Soccer Player Performance Rating System for the English Premier League. *Interfaces* 42, 4 (2012), 339–351. https://doi.org/10.1287/inte.1110.0589 arXiv:https://doi.org/10.1287/inte.1110.0589

[17] Rouven Michels, Marius Ötting, and Roland Langrock. 2022. Bettors' reaction to match dynamics – Evidence from in-game betting. https://doi.org/10.48550/ARXIV.2202.10085

[18] Marius Ötting, Roland Langrock, and Antonello Maruotti. 2021. A copula-based multivariate hidden Markov model for modelling momentum in football. *AStA Advances in Statistical Analysis* (19 Mar 2021). https://doi.org/10.1007/s10182-021-00395-8

[19] G. Pantuso and L. M. Hvattum. 2021. Maximizing performance with an eye on the finances: a chance-constrained model for football transfer market decisions. *TOP* 29, 2 (01 Jul 2021), 583–611. https://doi.org/10.1007/s11750-020-00584-9

[20] Luca Pappalardo, Paolo Cintia, Paolo Ferragina, Emanuele Massucco, Dino Pedreschi, and Fosca Giannotti. 2019. PlayeRank: Data-Driven Performance Evaluation and Player Ranking in Soccer via a Machine Learning Approach. *ACM Trans. Intell. Syst. Technol.* 10, 5, Article 59 (sep 2019), 27 pages. https://doi.org/10.1145/3343172

[21] Luca Pappalardo, Paolo Cintia, Alessio Rossi, Emanuele Massucco, Paolo Ferragina, Dino Pedreschi, and Fosca Giannotti. 2019. A public data set of spatio-temporal match events in soccer competitions. *Scientific Data* 6, 1 (28 Oct 2019), 236. https://doi.org/10.1038/s41597-019-0247-7

[22] Vineet M. Payyappalli and Jun Zhuang. 2019. A data-driven integer programming model for soccer clubs' decision making on player transfers. *Environment Systems and Decisions* 39, 4 (01 Dec 2019), 466–481. https://doi.org/10.1007/s10669-019-09721-7

[23] Pieter Robberechts and Jesse Davis. 2020. How Data Availability Affects the Ability to Learn Good xG Models. In *Machine Learning and Data Mining for Sports Analytics*, Ulf Brefeld, Jesse Davis, Jan Van Haaren, and Albrecht Zimmermann (Eds.). Springer International Publishing, Cham, 17–27.

[24] Olav Drivenes Sæbø and Lars Magnus Hvattum. 2015. Evaluating the efficiency of the association football transfer market using regression based player ratings. In *Norsk Informatikkonferanse*.

[25] Joseph Sill. 2010. Improved NBA Adjusted +/- Using Regularization and Out-of-Sample Testing. MIT Sloan Sports Analytics Conference.

[26] Lukasz Szczepański and Ian McHale. 2016. Beyond completion rate: evaluating the passing ability of footballers. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 179, 2 (2016), 513–533. http://www.jstor.org/stable/43965554

[27] Stephan Wolf, Maximilian Schmitt, and Björn Schuller. 2020. A football player rating system. *Journal of Sports Analytics* 6 (2020), 243–257. https://doi.org/10.3233/JSA-200411 4.