Research Article

Vincent Renner*, Konstantin Görgen, Alexander Woll, Hagen Wäsche and Melanie Schienle

Success factors in national team football: an analysis of the UEFA EURO 2020

https://doi.org/10.1515/jqas-2023-0026
Received March 21, 2023; accepted June 24, 2024;
published online July 22, 2024

Abstract: Identifying success factors in football is of sporting and economic interest. However, research in this field for national teams and their competitions is rare despite the popularity of teams and events. Therefore, we analyze data for the UEFA EURO 2020 and, for comparison purposes, the previous tournament in 2016. To mitigate the challenges of perceived multicollinearity and a small sample size, and to identify the relevant variables, we apply the 'LASSO Cross-fitted Stability-Selection' algorithm. This approach involves iterative splitting of data, with variables chosen via a 'least absolute shrinkage and selection operator' (LASSO) model (Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. B* 58: 267–288) on one half of the observations, while coefficients are estimated on the other half. Subsequently, we inspect the frequency of selection and stability of coefficient estimation for each variable over the repeated samples to identify factors as relevant. By that, we are able to differentiate generally valid success factors such as the market value ratio from on-field variables whose importance is tournament-dependent, e.g. the tackles attempted. As the latter is connected to a team's tactics, we conclude that their observed relevance is correlated to the results of the linked playing style in the specific tournaments. We also show the changing effect of these playing-styles on success across tournaments.

***Corresponding author: Vincent Renner**, Institute of Statistics, Karlsruhe Institute of Technology, Karlsruhe, Germany,
E-mail: vincent.renner@student.kit.edu. https://orcid.org/0009-0005-3359-1841
Konstantin Görgen and Melanie Schienle, Institute of Statistics, Karlsruhe Institute of Technology, Karlsruhe, Germany,
E-mail: konstantin.görgen@kit.edu (K. Görgen),
melanie.schienle@kit.edu (M. Schienle)
Alexander Woll, Institute of Sports and Sports Science, Karlsruhe Institute of Technology, Karlsruhe, Germany, E-mail: alexander.woll@kit.edu
Hagen Wäsche, Department of Sport Science, University of Koblenz, Koblenz, Germany, E-mail: waesche@uni-koblenz.de

Keywords: football; post-lasso double selection; national team competition; stability selection

1 Introduction

The importance of data science and analytics in football has been continuously growing in recent years (see e.g. Collet 2013; dos Reis et al. 2017; Liu et al. 2015; Sarmento et al. 2017). A key goal is to identify the most relevant success factors in order to facilitate player evaluation, spur novel tactical developments, and generally enable more efficient team and player preparations. We show that detecting such relevant drivers in a data driven way using machine-learning and agnostic of ad-hoc assumptions leads to new insights that might not only be beneficial for sporting, but also for the closely intertwined financial success in modern football. The importance of specific influence factors such as ball possession and running distance has not only been discussed in mainstream media, but also in dedicated research, with differing results depending on the study and the respective model assumptions (see e.g. Collet 2013; Lago-Peñas et al. 2011). These disagreements come not only from differences in employed data and methodology, but also from the development of football as a game, with evolving tactics and thus changing success factors over time. Therefore, it is of interest to differentiate between play-style related factors, which might have changed specifically for the most recent tournaments, and generally important factors which remain relatively unchanged over time.

This distinction and the data-driven identification strategy for the respective factors constitutes the main contribution of this paper. We analyze the 2020 UEFA European Football Championship (UEFA EURO 2020) and compare results to the previous event in 2016. In this way, we are able to identify generally important success factors such as the distance covered compared to the opponent, the team's market value as a ratio to the opponent, and the shot accuracy. But we can also determine on-field variables, such as the save rate, the clearances, the dribbles, long passes, tackle rate, and shots from fast-break having a significant positive impact,

with their importance varying between the tournaments. The crosses, errors, and tackles attempted are detected as negative impact factors. To further evaluate this we also inspect the different possible play styles, and see a changing correlation to success for these style buckets in between the two examined tournaments. To ensure the robustness of our results, we consider both a result-based (result as a discrete dependent variable) and a goal-based perspective (result as a goal-difference) to model the outcome of match success. Another important contribution is the proposed use of a cross-fitted stability selection, which combines the 'least absolute shrinkage and selection operator' (LASSO) (see Tibshirani 1996) for identification of important factors with a form of stability selection in order to compensate the relatively small sample size (see Chernozhukov et al. 2018; Görgen and Schienle 2019; Meinshausen and Bühlmann 2010).

We focus on the most recent European big national team tournament – the UEFA EURO 2020 – which took place from the 11th June until the 11th July of 2021, ending with Italy winning the final against England. As a benchmark, we are comparing the EURO 2020 with data from the previous competition, the EURO 2016. This enables us to distinguish between play-style related factors and more generally important ones by comparing which of those are selected for each tournament. Our variables consist of the traditional on-field statistics, such as shots, passes etc., and furthermore contextual variables that represent unique characteristics of the UEFA EURO 2020. These are, for example, the varying home advantages, spectators, and travelling distances, caused by a combination of the tournament being played all across Europe and differing COVID-19 related restrictions. The comparison with the previous tournament also enables us to draw conclusions about the COVID-19 related influences.

With our focus on national team tournaments, club level dominant factors like each team's budget are a secondary concern and general play specific determinants are of higher importance (see e.g. Lepschy et al. 2020, 2021). While research on national team tournaments has so far been limited, those competitions reach an even bigger audience than club football, with the most recent World Cup in 2022 reaching engagement numbers of estimated 5 billion people, more than half of the world's population.[1] Moreover, there is also a major financial interest in national teams, as the described large audience leads to a vast market for sponsorings resulting in big earnings, with e.g. the German national team association (DFB) earning 183 million euros in sponsorships in 2019, 45 % of their total earnings.[2] That exemplifies a need for additional research, especially regarding national teams.

The biggest data challenges of our analysis lie in the high multicollinearity between variables and the small sample size compared to the great number of available variables. To mitigate theses issues, we employ a LASSO-based strategy which allows for selecting important variables. Since the LASSO suffers from selection problems with highly-correlated, high-dimensional data sets such as in our case, we suggest a repeated cross-fitting methodology, which achieves a more stable selection. For this, we combine ideas of stability selection (see Chernozhukov et al. 2018; Meinshausen and Bühlmann 2010) with cross-estimating (see e.g. Wang et al. 2020). Specifically, we repeatedly split the data into two equally large random sub-samples. Variable selection is ensuingly conducted on one subset using the LASSO, whilst the other half of observations is used to estimate the coefficients for these selected variables in a plain OLS model. Subsequently, we inspect the frequency of selection and stability of coefficient estimation for each variable over the repeated samples to identify factors as relevant. This 'LASSO Cross-Fitted Stability Selection' provides a more robust feature selection and coefficient estimation than used in most comparable research.

Existing literature on the sporting side mainly focuses on club level competitions. Lepschy et al. (2020) for instance found multiple factors to be beneficial for success, including the teams market value, the goal efficiency, shots from counter attack, and the home advantage, when analyzing match statistics in the German Bundesliga from 2014/2015 until 2016/2017. On the other hand, Schauberger et al. (2018) examined the Bundesliga for the 2015/2016 season and identified the distance covered as the most important positive factor. Peñas et al. (2010) focuses on the Spanish first division from the 2008/2009 season, with the result of finding the shots on target, the effectiveness among other factors to be most decisive. For national team football, there are, however, only a few studies. Lepschy et al. (2021) for example analyzed the previous two World Cups in 2014 and 2018 and identified the efficiency, the duel success rate, the clearances and the shots from counterattacks to have a positive impact. The 2014 tournament was also analyzed by Liu et al. (2015).

Section 2 describes the model and methodology. The collection and pre-processing of the data is explained in Section 3. In Section 4, we describe and interpret the ensuing results, before Section 5 concludes.

1 FIFA audience report.

2 https://www.dfb.de/fileadmin/_dfbdam/224318-Finanzbericht_DFB_2019_final.pdf.

2 Model and method

The primary purpose of this paper is to identify success factors by determining on- and off-field variables with the largest influence on winning in the UEFA EURO 2020. There is a vast number of data collected for each football match, but only a limited number of matches that are played in each tournament. Our aim is to determine the relevant variables from the large full pool of available potentially influencing factors in a data-driven way. A suitable method for this challenge is the 'least absolute shrinkage and selection operator' (LASSO), as introduced by Tibshirani (1996), which penalizes coefficient size and therefore shrinks the coefficients' estimations, even up to the size of zero, essentially working as a selection device. We combine this with the idea of a cross-fitted stability-selection (see Chernozhukov et al. 2018) to address the issues of influential observations and correlated data, that are pertinent in our data as outlined in Section 3 and cause pure LASSO to produce misleading results. We base the LASSO Cross-Fitted Stability-Selection on a linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + u \tag{1}$$

where X is the vector of all $K = 42$ regressors (see Section 3) with elements x_k $(k = 1, \ldots, K)$. Note that in this set-up, estimated effects can in fact be interpreted as causal if some assumption outlined below is met. Generally from observational data, we can only identify average causal effects, i.e. β_k in (1) can be interpreted as an average effect of variable k over all games while the effect for a specific individual pairing might vary. In particular, in a multivariate linear regression model, β_k marks the marginal effect of a variable x_k on the outcome y given all other variables remain unchanged. Moreover, β_k is also the average causal effect of x_k on y if conditional on all other variables in (1), the error u is independent of x_k (see e.g. Hansen 2022, Section 2.30). In order to ensure that this conditional independence condition is met in our case, we employ the large set of 42 regressors where we take any effort to make it as comprehensive as possible (see Section 3). In this way, we generally aim to capture all potential correlation between x_k and u by the vast amount of the other covariates such that, conditional on the other regressors, there is no remaining dependence between x_k and u. This means that in our case we can interpret a coefficient of model (1) as e.g. an increase of x_{shots} by 1 causes on average over all games an increase of the success variable y measured as the goal difference for that team y_{gb} by $\hat{\beta}_{\text{shots}}$ units. Note that our data set contains for each variable only the final record per match. Thus the only source for a potential collider effect impacting the overall causal interpretation in the model (1) could arise if there was a remaining unobserved common factor of the outcome and a regressor variable conditional on the vast set of other controls. This is very much different from an analysis using within-game information where reversed causality would be a much more prominent concern due to timing effects between regressors and outcome within the game. Please see our detailed discussion in Section 4.1 where we argue that the vast set of explicit controls offers a substantial insurance against missing out on a common factor. Despite the fact that a subset of the other controls already jointly contains quite comprehensive information on game tactics, we construct an additional explicit control strategy for unobserved tactical effects in tight versus less tight games that might influence some regressors and the outcome jointly in very distinct ways. Generally unless otherwise stated, in the sequel all estimated effects are on average and ceteris paribus, i.e. keeping all other variables fixed.

For our dependent success-variable y we use two different notions that give rise to two different approaches: The first one is result-based, which classifies each match by result in the categories win, loss, and draw. The second is goal-based, which uses the goal-difference of the result as the target y. Considering both approaches jointly helps to confirm the validity of the consistent results of the two methods, while differences in results spur further investigation. Therefore, we use the *Result – Goal Difference* as the dependent variable in the goal-based approach and *Result – W/L/D* in the result-based approach, where we adapt our model to a logistic regression by applying the logit-link function.

2.1 Two-step regression with LASSO cross-fitted stability-selection

This Section describes the goal-based approach that is based on a standard linear model where we treat the outcome as continuous variable.

Our methodology builds on the LASSO pre-estimator $\hat{\beta}^L$ in a first model selection step. It is defined as follows:

$$\hat{\beta}^L = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{k=1}^{K} x_{ik} \beta_k \right)^2 + \lambda \sum_{k=1}^{K} |\beta_k| \right\}. \tag{2}$$

Note that the estimate $\hat{\beta}^L$ depends on the choice of the penalty parameter λ which is usually pre-determined in a data driven way. Generally, we select a variable k as relevant if $|\hat{\beta}^L_k|$ deviates sufficiently from zero. When a high number of variables relative to the available sample size must be

included in a model to ensure causal interpretability of effects, their estimated coefficients can become biased with a high variance. By penalizing the cofficients as in Brant (1990) for a pre-set value of λ, this problem is overcome (Hastie et al. 2013, p. 63).

While the LASSO generally offers an excellent way to implement data-driven model selection, in our data we face a situation with a substantial number of high leverage observations (see Section 3.2). When estimating our model with such data, this would lead to highly varying coefficients when dropping one of the high leverage points and thus to varying sets of covariates selected by the LASSO, depending on the subset. As a result, we followed the idea of stability selection (see Chernozhukov et al. 2018; Meinshausen and Bühlmann 2010) and extended it with the concept of cross-fitting or cross-estimating (see 'R-Split' by Wang et al. 2020). The final method is described in Algorithm 1 below in detail.

Algorithm 1. LASSO Cross-Fitted Stability-Selection.

Step 1:
for $i = 1$ **to** $C = 1{,}000$ **do**
Divide the N observations into two random subsamples Y_1, X_1, and Y_2, X_2 where Y_1 is
an n_1-vector and X_1 is an $(n_1 \times k)$ regressor matrix; Y_2 is an n_2-vector and X_2 is an $(n_2 \times k)$ matrix. We set $n_1 = n_2 = 0.5 \cdot N$;
Compute LASSO-model on subsample 1 and select s_i variables;
Estimate unrestricted OLS model on subsample 2 with the s_i variables selected before;
end
Step 2:
for $k = 1$ **to** K **do**
$\bar{\hat{\beta}}_k = \frac{\sum_{c=1}^{C} \hat{\beta}_{kc}}{\sum_{c=1}^{C} \mathbb{1}\{|\hat{\beta}^L_{kc}|>0\}}$;
$\hat{\pi}_k = \frac{1}{C}\sum_{c=1}^{C} \mathbb{1}\left\{|\hat{\beta}^L_{kc}| > 0\right\}$;
$\hat{\Delta}_k = \hat{\pi}_k - \hat{\pi}_{k+1}$;
$\sigma(\hat{\beta}_k)$;
$\hat{\upsilon}_k = \sigma(\hat{\beta}_k)/\bar{\hat{\beta}}_k$;
$\gamma_k = \frac{\sum_{c=1}^{C} \mathbb{1}_{\hat{\beta}^L_{kc} \times \bar{\hat{\beta}}_k > 0}}{\sum_{c=1}^{C} \mathbb{1}\{|\hat{\beta}^L_{kc}|>0\}}$;
end
Result: Table with $\hat{\pi}_k$, $\bar{\hat{\beta}}_k$, $\hat{\Delta}_k$, $\sigma(\hat{\beta}_k)$, $\hat{\upsilon}_k$, γ_k for all variables
Step 3:
Find ideal selection threshold for the final model using the measures obtained in step 2.

The main idea is, that we divide our data set into two equally large random sub-samples for every iteration. We then estimate a LASSO model on our first sub-sample and track which covariates have estimated LASSO coefficients $\hat{\beta}^L_k$ substantially different from zero and are thus selected. Instead of using the estimates obtained in this first step, we now compute a new unrestricted model, using only the selected covariates in the second sub-sample and obtaining estimates $\hat{\beta}_k$. We redo this $C = 1{,}000$ times and in every step c of the algorithm, the penalty parameter λ of the LASSO Equation (2) is obtained by minimizing the 10-fold cross-validated mean squared error (*MSE*). For robustness, we only keep the variables with the highest empirical selection frequency across all iterations c in the final model. In particular, we take the share of selections for each variable $\hat{\pi}_k = \frac{1}{C}\sum_{c=1}^{C} \mathbb{1}\left\{|\hat{\beta}^L_{kc}| > 0\right\}$ as primary screening device. In addition, we also consider the difference in the share of selection to the next less selected covariate $\hat{\Delta}_k = \hat{\pi}_k - \hat{\pi}_{k+1}$ (after ordering variables by $\hat{\pi}_k$'s from large to small), the standard deviation $\sigma(\hat{\beta}_k)$, the coefficient of variation $\hat{\upsilon}_k = \frac{\sigma(\hat{\beta}_k)}{\frac{1}{C}\sum_{c=1}^{C}\hat{\beta}^L_{kc}} = \sigma(\hat{\beta}_k)/\bar{\hat{\beta}}_k$ and the 'share of same direction' $\gamma_k = \frac{\sum_{c=1}^{C} \mathbb{1}_{\hat{\beta}^L_{kc} \times \bar{\hat{\beta}}_k > 0}}{\sum_{c=1}^{C} \mathbb{1}\{|\hat{\beta}^L_{kc}|>0\}}$ around potential cut-off points in $\hat{\pi}$ to determine if a variable is included into the final model. In this way, we account not only for the covariates with the most selections but also with the most stable coefficient estimations. As estimation results we generally report the average of coefficients conditional on the variable being selected in the LASSO step $\bar{\beta}_k = \frac{\sum_{c=1}^{C}\hat{\beta}_{kc}}{\sum_{c=1}^{C} \mathbb{1}\{|\hat{\beta}^L_{kc}|>0\}}$ together with the empirical selection frequency $\hat{\pi}_k$. For completeness, we have also calculated the unconditional average coefficients $\bar{\hat{\beta}}_k \cdot \hat{\pi}_k$ across all $C = 1{,}000$ iterations for any finally selected variable k. Following Meinshausen and Bühlmann (2010), we expect that the obtained selection set and marginal effects are more robust to outliers and high correlation of regressors than standard LASSO.

2.2 Two-step classification via logistic regression with LASSO cross-fitted stability-selection

For the result-based approach, the outcome variable can only take three ordered values and thus can no longer be treated as continuous. We therefore base our analysis on a logistic regression for classification. In particular, we argue that the proportional odds assumption is satisfied in order to reduce model complexity. For data-driven model selection in a first step, the stability selection algorithm of Section 2.1 is adapted to fit a proportional odds logistic regression.

Our dependent variable consists of three levels ($J = 3$). We also possess additional information about it, as it is an ordinal categorical variable, since we can sort the three levels ($-1 < 0 < 1/\text{Loss} < \text{Draw} < \text{Win}$) without being able to give relative differences or calculate ratios. This information can be used by performing an ordinal logistic regression model instead of a multinomial logistic regression, which only requires a discrete dependent variable without any order. We use the so-called proportional odds (PO)

model as first introduced by Walker and Duncan (1967) and first being described as PO model by McCullagh (1980). This model is used to obtain the probabilities $\Pr(Y \geq j|X = x)$. For this, the PO model estimates the log-odds of being in or above the category versus being below that category as a linear model:

$$\ln\left(\frac{\Pr(Y \geq j|X = x)}{\Pr(Y < j|X = x)}\right) = \beta_{j0} + \beta_{j1}x_1 + \beta_{j2}x_2 + \cdots + \beta_{jk}x_k = \beta_{j0} + \beta_j^T x \quad (3)$$

In our case we therefore estimate the log-odds of not losing to losing, and of winning to not winning:

$$\ln\left(\frac{\Pr(Y \geq 0|X = x)}{\Pr(Y < 0|X = x)}\right) = \ln\left(\frac{\Pr(Y = 0|X = x) + \Pr(Y = 1|X = x)}{\Pr(Y = -1|X = x)}\right) = \beta_{10} + \beta_1^T x$$

$$\ln\left(\frac{\Pr(Y \geq 1|X = x)}{\Pr(Y < 1|X = x)}\right) = \ln\left(\frac{\Pr(Y = 1|X = x)}{\Pr(Y = -1|X = x) + \Pr(Y = 0|X = x)}\right) = \beta_{20} + \beta_2^T x$$

According to Harrell (2015) regression coefficients in the PO must be independent of the cutoff level j for Y. As a result, we get two different intercepts β_{10} and β_{20}, but only one set of coefficients β, so $\beta_1 = \beta_2$. That means, in our case, that the effects of the regressors are the same for the log-odds of the three cases not losing versus losing and winning versus drawing or losing. For national team tournaments the ordering of the teams in the recording of the match results has no particular meaning as it is not marking which team obtained home advantage, contrary to league competition matches. Due to the structure of the data, the proportional odds assumption seems therefore justified. In addition, the assumption can also be formally assessed using the Brant-test (Brant 1990), where rejecting the null would be evidence for a violation of the assumption.

Using the logit link function we model the cumulative probabilities by re-transformation as Harrell (2015, p. 313):

$$\Pr(Y \geq j|X = x) = \frac{1}{1 + \exp(-(\beta_{j0} + \beta^T x))}$$

With that the probability for each class can be calculated with:

$$\pi_j = \Pr(Y = j|X = x) = \Pr(Y \geq j|X = x) - \Pr(Y \geq j + 1|X = x)$$

The coefficients are estimated by maximum-likelihood estimation. We adapt the LASSO cross-fitted stability-selection (Algorithm 1) to the PO model by fitting a penalized PO model in step 1 on subset X_1. This penalization is similar to the LASSO described in Section 2.1, as it sanctions coefficient size with the L_1-penalization, this time in the Likelihood function $-\frac{1}{N}\ell(b) + \lambda\sum_{k=1}^{K}|\beta_k|$ (see Wurm et al. 2021). We then use the selected features from the LASSO to estimate a regular PO model using only these features on the remaining data X_2.

3 Data

3.1 Sources and empirical stylized facts

We use the $N = 102$ per team observations of results plus game/team characteristics from the 51 games during the EURO 2016 and EURO 2020. In particular, we have gathered a comprehensive data set containing the most important characteristics and variables used in the literature. Firstly, it contains collected on-field statistics, such as *Shots*, *Passes*, and *Dribbles*, including all sub-categories, which are obtained from https://www.whoscored.com/. This data is directly linked to https://www.statsperform.com/opta/, one of the leading providers for sports data, specifically for football. The reliability of their data was found to be very good marked by high kappa values of 0.92 and 0.94 for the correlation between different data collection operators and thus limiting the human error (Liu et al. 2013). Note that all characteristics denote respective values at the end of each completed game, thus we have no access to their evolution during the course of games.[3]

In addition to this data, we have constructed the following variables capturing effects specific to the EURO 2020 in our data. The Save Rate $= \frac{\text{Saves}}{\text{Shots on target conceded}}$ models goalkeeper performance, Shot Accuracy $= \frac{\text{Shots on target}}{\text{Shots}}$ measures the quality of a team's shots. In addition to the home advantage, which is only present in certain matches where the team plays in their own country, the number of spectators allowed varied massively across countries due to different COVID-19 related restrictions. To take this into account, we define the weighted home advantage $wHA =$ Home advantage $\cdot$ Spectators. We also include *Distance covered difference*$_i$ as the difference of *Distance covered* of the observed team compared to the opposing team. The idea

3 The number of initially collected variables was systematically reduced to a remaining number of 40, due to high similarities of the variables mitigating multicollinearity effects but avoiding omitted variable bias.

here is that simply running much is unrewarding, whereas running more than the opponent could lead to a higher chance of winning. This assumption can be supported by the data, as shown in Table 1. The average distance covered is similar in wins and losses but higher in draws. This could be a consequence of draws usually being much more hard-fought games. However, if we look at the *Distance covered difference*, we can see a positive value of 1.823 in wins, and a negative value for losses, which means that losing teams ran 1.823 km less than the winning side. This observation suggests a more significant influence of the difference in distance, which is why we will review it instead of *Distance covered*. The *Market value ratio* follows a similar idea, as it shows the ratio of the *Market value* of the starting formation of the observed team compared to the opponent, with the assumption that a higher *Market value* implies a better or more skilled team. The motivation here is that just a high *Market Value* does not automatically give a team an advantage over its opponent. Only when compared to the opponent, the market value might be indicative of the result, for example if a team's *Market Value* is a lot higher than that of the opposing team. As visible in Figure 1, there is a much higher average *Market Value* in Draws and Wins than Losses. However, the values in Wins seem to be slightly below the ones in Draws. Looking at Figure 2, we can see a much clearer correlation of the log(*Market value ratio*) to the result. Please see Figures 8 and 9 in the Appendix for details. The EURO 2020 featured one major change in sporting regulations, as teams were now allowed to use a total of five substitutions in three substitution windows. Since 1995, teams were only allowed to make three substitutions in a game. This change was justified by the increased physical demands on the players in the previous season due to the shifts in the game calendar triggered by the COVID-19 pandemic. We have therefore included the variable *Substitutions* indicating how many substitutions were made by each team.

An overview of the descriptive statistics for the final set of 42 covariates is provided in Table 5.

Table 1: Arithmetic means of *Distance covered* and *Distance covered difference* grouped by result (win/loss/draw).

Arithmetic mean	Loss	Draw	Win
Distance covered	109.659	116.9	111.482
Distance covered difference	−1.823	0	1.823

3.2 Data challenges and exploration

We structure our data by using match-team observations as we model team performance through splitting every match into two observations, one for each side. This procedure is necessary due to the low number of matches played resulting in a small sample size. But we also have to be aware of an ensuing problem: The outcomes of these pairs are always directly correlated to each other. This leads to the residuals of these pairs to be negatively correlated. If we estimate a post-regression in the goal-based approach for our final selected variables (which will be discussed in Section 4.2), we see a high correlation for the associated residuals with $r = -0.52$ for 2020 and $r = -0.64$ for 2016. We are, however,

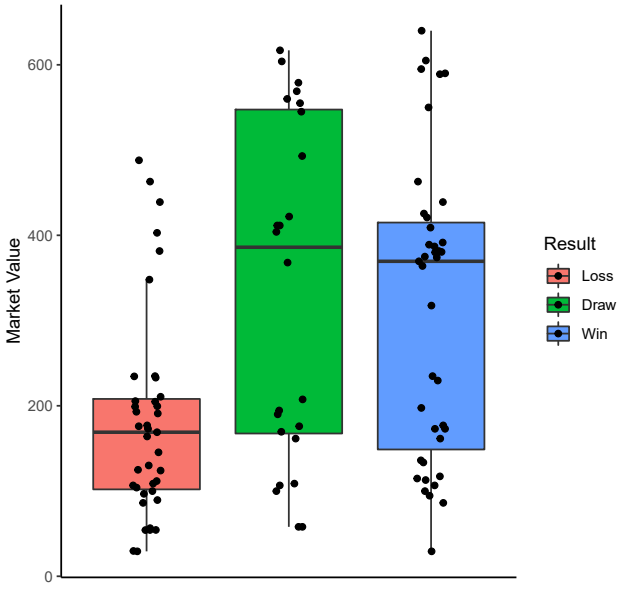

Figure 1: *Market value* by result.

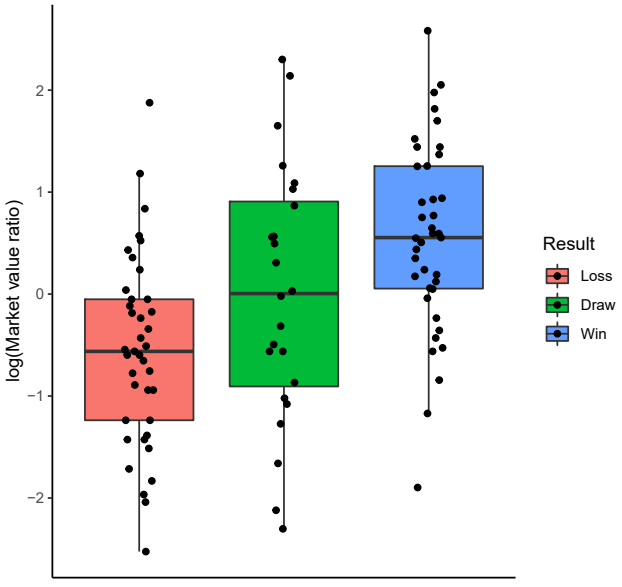

Figure 2: log(*Market value ratio*) by result.

able to alleviate this issue through the applied cross-fitting, as the coefficients are always estimated on a random subsample of the data with size $N/2$. These sub-samples are drawn independent of the pairs, so the correlation of the residuals is reduced.

Inspecting our data we find some observations to be highly influential. This is visualized in Figures 3 and 4, where the 'difference in fits' (DFFITS), proposed by Belsley et al. (1980), is depicted. The DFFITS measures the change in fit when one observation is deleted and by that its influence. A usual threshold to identify the influential observations is $\pi_{\text{DFF}} = 2 \cdot \sqrt{\frac{K}{N}}$, which "accounts both for the sample size [N] and the fact that DFFITS_i increases as [K] does" (Belsley et al. 1980, p. 28). We see 14 influential observations (13.73 %) for either approach (see Figures 3 and 4), which is a significant number of leverage points.

We additionally look at the influence of observations on the coefficient estimates by calculating the DFBETAS (see Belsley et al. 1980). The DFBETAS_{ij} describes the influence of observation i on the estimate of coefficient j. Our cutoff value for this is $\pi_{\text{DFBETAS}} = \frac{2}{\sqrt{N}}$. To compare the amount of influential observations for each covariate, we define $\xi_j = \sum_{i=1}^{N} \mathbb{1}_{\{|\text{DFBETAS}_{i,j}| > \pi_{\text{DFBETAS}}\}}$ (see e.g. Görgen and Schienle 2019), the number of influential observations per covariate. We see a median value of 4 in the result-based approach and 3 in the goal-based approach. These results are somewhat problematic, as they lead to highly different estimations and variable selections for our model depending on the subsample. Moreover, since many of the available variables describe a similar matter and many covariates are subcategories of others, we expect and empirically confirm high multicollinearity of covariates. Both high multicollinearity and influential observations are mitigated by the proposed cross-fitted stability-selection approach described in Section 2.1.

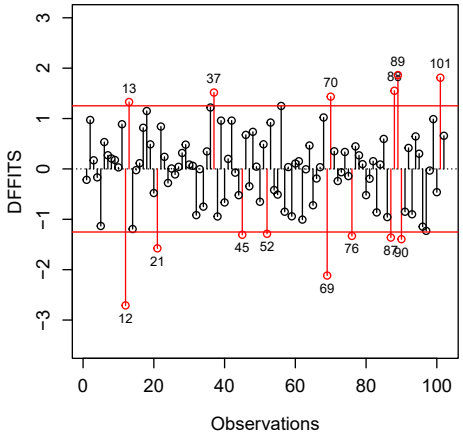

Figure 3: DFFITS in result-based approach.

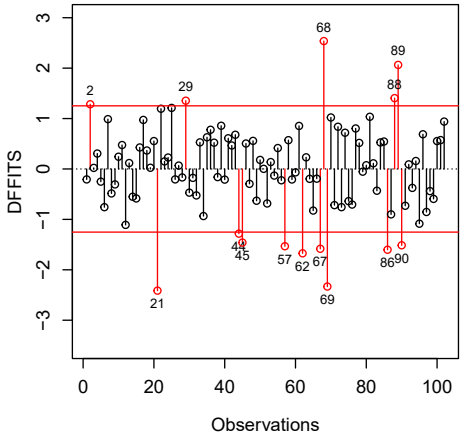

Figure 4: DFFITS in goal-based approach.

4 Empirical results and discussion

4.1 Specific model setup

In assessing team success in tournaments such as the UEFA EURO, we must acknowledge potential differences in team strategies between the group stage and the knockout stages. In the knockout stage, teams often exhibit more aggressive play as draws are not possible, unlike in the group stage. To reflect these potential variations in our model, we incorporate a dummy variable, *Knockout*. This variable is set to 1 for a game in the knockout stage and 0 for a game in the group stage. To evaluate how effects and covariates might vary between these two stages, we further consider the interaction of *Knockout* with all other covariates in our model.

In order to justify the average causal effects interpretation of obtained coefficients, the conditional independence assumption below Equation (1) must be met. Please also see our discussion in Section 2. Note that we only have end of game information in our publicly available data where within-game timing and resulting reversed causality issues play no role. Thus potential collider effects might only arise due to an unobserved common effect driving the result y and a regressor x_k as a source for simultaneous influence of y on x_k. Recall that we use an exceptionally large set of regressors as an insurance against a remaining common unobserved effect after pre-conditioning on the other controls. Let us illustrate this for the on-field variable *Dribbles*. Following the argumentation outlined above, we require that conditional on all other observed characteristics, there is no more remaining dependence of dribbles and results resulting from some common unobserved factor. For

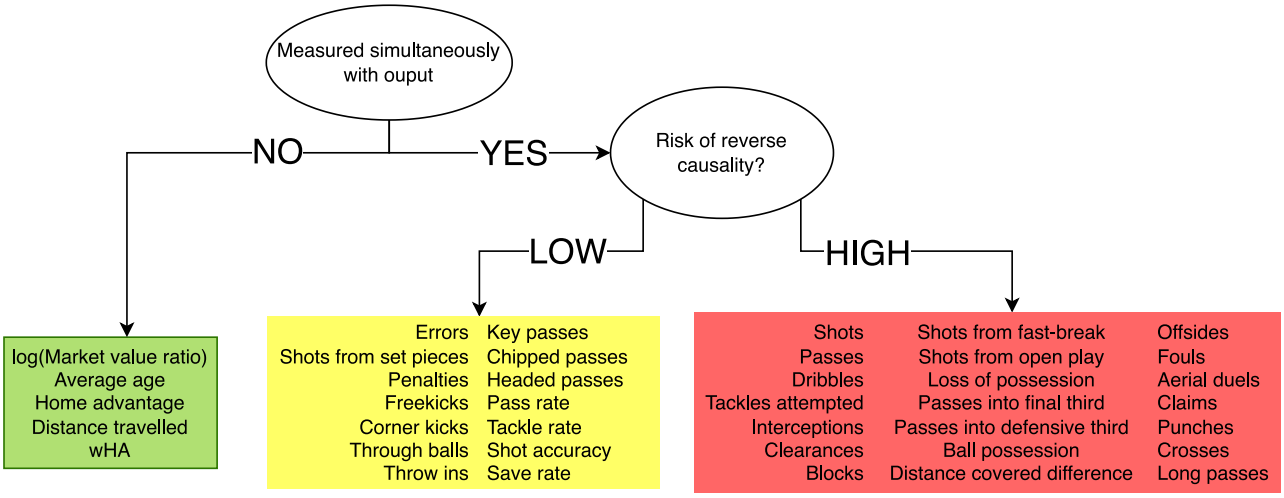

Figure 5: Grouping of variables based on severity of reverse causality issues.

example, one outstanding player could be such a potential unobserved factor that drives dribbles and success simultaneously. Please note, however, that this outstanding player would also majorly impact the market value and all other on-field controls such as three types of pass and shot variables, blocks, tackles etc. that are strongly correlated with the dribbles by the outstanding player impact. Therefore we argue that if we condition on these other variables, the outstanding player information is already fully captured. Hence, conditional on the other controls, the independence assumption for dribbles seems justified and a directed interpretation appears valid. As detailed below the same reasoning applies to most of the other variables in Figure 5 due to the large set of employed controls. Even though the risk of general unobserved common factors seems low, there might still be remaining unobserved factors such e.g. game evolution and related tactical measures that impact both y and x_k even if conditioning on the other regressors. Such unobserved common effects pose a challenge for the causal interpretation of all regressors and their impact is not limited to the subgroup which they directly affect. For instance, a trailing team may adopt a more aggressive style of play, attempting more tackles and covering more ground than the leading team such that game evolution and resulting tactics influence both y and respective variables.

Structuring our investigation for potential collider effects, we have categorized our variables into three groups illustrated in Figure 5 that mark their proneness to unobserved within-game/tactical effects. The first group consists of predictor variables that are measured before the start of the game, such as the log(*Market value ratio*). These variables are marked in green as they pose no issues concerning the challenge previously described. For variables measured concurrently with the output, we further differentiate into two subsets. The first subset consists of variables in the yellow box that are minimally influenced by the game's evolution. This group includes metrics like *Shot accuracy* and *Save rate*, which are mainly driven by general team quality and specific daily ability but less by game evolution. Note that only controlling for variables in the yellow box would actually leave some team ability unobserved that could impact the results outcome. In this case, however, green variables like log(*Market value ratio*), act as observed proxies for the unobserved collider allowing for valid interpretation when conditioning on them. Thus we must control for the green and yellow groups jointly in the model. Only in this case estimated impacts would not suffer from collider problems in their interpretation as average causal effects (see Section 2). But such models would be underspecified missing out on relevant variables from the second (red) subset leaving us with estimates that are biased.

The second (red) subset, however, comprises variables that might be influenced by a distinct game evolution factor that is not covered by the other variables but that also has an impact on the output variable, as for example, *Distance covered difference* and *Tackles attempted* marked in red in Figure 5. Without further measures, this would destroy the conditional independence assumption required for the estimated effects to be average causal effects. As additional more detailed data with information regarding the actual course of a game is not available publicly, we construct a control variable *Close game* in order to recover the conditional independence for all existing regressors when additionally conditioning on *Close game*. Our argument is that

Close game captures the key part of the unobserved game evolution factor that impacts both, response and regressors. We define *Close game* as a dummy variable indicating whether a game was closely contested, coded as '1' if the final goal difference was 1 or lower, and '0' otherwise. By additionally employing interaction variables of *Close game* with the existing regressors, we allow impacts to also appear in marginal effects. In this way, we implicitly split the sample into two parts between which effects of the red variables differ substantially keeping all other variables unchanged. In this way, we argue to explicitly capture the key part of the game evolution effects that impact both regressors and outcome and thus restore the conditional independence assumption conditional on this augmented set of regressors such that derived effects can be interpreted as average causal effects.

To accurately identify crucial success factors in team football, we conduct an analysis using both a goal-based and a result-based approach for EURO 2020 and EURO 2016. Studying different aggregation levels of the outcome variable helps to confirm selection results, and inspecting both competitions enables us to differentiate between longer-term critical factors and short-term ones, which might be influenced by the unique format of EURO 2020. In both approaches, we employ the two-step approaches with the initial LASSO stability selection based model determination as outlined in Section 2.1 and subsampling. In particular for each model, we choose the final set of variables according to the selection proportion $\hat{\pi}_k$ when resampling $C = 1{,}000$ times. The cut-off threshold for the variable selection is set adaptively taking $\hat{\Delta}, \sigma(\hat{\beta}), \hat{v}$, and γ in decreasing importance into account. This will determine our final model and, consequently, the relevant covariates.

4.2 Results

4.2.1 Goal-based regression

The results for employing the suggested LASSO-procedure from Section 2 for the goal-based approach are presented in Tables 6 and 11 for 2020 and 2016, respectively, sorted by decreasing selection probability. The empirical selection probabilities $\hat{\pi}$ are shown in Figure 6. For completeness, we have also calculated the unconditional average coefficients $\bar{\hat{\beta}}_k \cdot \hat{\pi}_k$ across all $C = 1{,}000$ subsampling iterations of the algorithm for any finally selected variable k, which can be found in Appendix in Table 16. The 1000 LASSO models in step 1 of the cross-fitted stability-selection algorithm are estimated with the help of the cvglmnet function of the glmnet package by Simon et al. (2011).

Variable selection: For the 2020 tournament, the three most highly selected covariates are the *Shot accuracy*, the *Distance covered difference*, and the log(*Market value ratio*), with selection shares π_k between 0.976 and 0.997. Figure 6 marks a significant drop in $\hat{\pi}_k$ after $k = 3$ suggesting a cut-off point according to $\hat{\Delta}$. Though the following 6 variables are still quite stable as indicated by the standard deviation $\sigma(\hat{\beta}_k)$, the variation coefficient $\hat{v}_k$, and the share of coefficients with the same sign $\hat{\gamma}_k$. The following variable *Through balls* at $k = 10$, however, is less stable, indicated by the much higher $\hat{v}_k = 37.144$ and lower $\hat{\gamma}_k = 0.5$. Therefore, we select the variables *Shot accuracy*, *Distance covered difference*, log(*Market value ratio*), *Save rate, Tackles attempted, Errors, Tackle rate, Dribbles*, and *Clearances* × *Close games* into the final model.

For the 2016 installment, we see the similar top three most selected variables. The most selected variable are the *Shots from fastbreak* with a $\pi_k = 0.915$. Interestingly, this

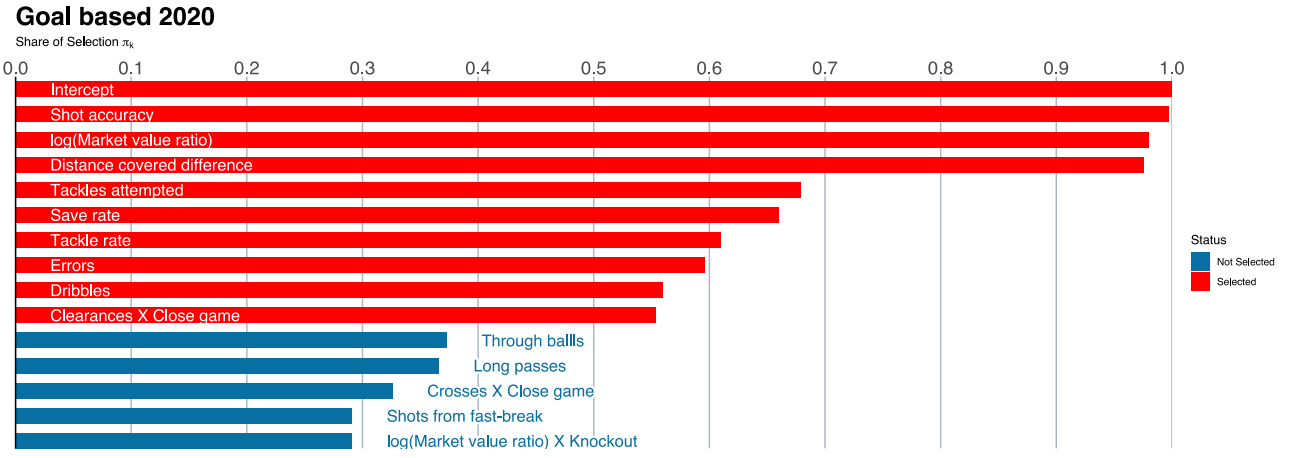

Figure 6: Selection share π_k for the 15 most selected variables in the goal-based model for 2020. Variables selected for final model are red.

variable is not selected in either of the models for 2020. The next two covariates, the *Clearances* and the log(*Market value ratio*) × *Knockout*, seem to be very stable in υ_k and γ_k and therefore are selected. This changes between the *Clearances* and the *Tackle rate*, which has a much higher $\upsilon_k = 19.064$ and lower $\gamma_k = 0.549$. We see this as an indication of less significance, and therefore cut off all subsequent variables with lower selection shares. Note that reported estimates are averages over the 1,000 subsamples aiming for robustness of the effects with respect to prominent outliers in the data. The corresponding results can be found in Tables 7, 9, 12, and 14, which highlight a generally high similarity to our aggregated coefficients.

For the interpretation of the reported subsample average estimates of coefficients in Table 2 we have to note that they do not originate from a single linear model but we still interpret them as marginal effects robust to outliers. Therefore we can say that for e.g. every additional dribble by a team is positively associated with an average increase of 0.02651 in the predicted goal difference of our model for 2020 ceteris paribus.

Comparison to existing literature and interpretation: The observed importance of most of the variables is in line with existing literature. The shot accuracy is also deemed important by Brito de Souza et al. (2019), as well as by Lago-Peñas et al. (2011), Lepschy et al. (2020), Liu et al. (2015), and Peñas et al. (2010), who all examined the shots on goal instead. The effect for the number of *Clearances* is confirmed in Lepschy et al. (2020), and also in the form of a variable measuring successful defensive actions by Carmichael et al. (2000). A negative effect of *Errors* was also found by Lepschy et al. (2020), however, no similar variable was examined in most other research. The importance of the *Tackle rate* was confirmed by Schauberger et al. (2018), and a similar variable in the total number of successful tackles was significant in the research of Lepschy et al. (2021).

This is not the case for the negative effect of the number of *Tackles attempted* that we found, for which Carmichael et al. (2000) and Liu et al. (2015) found a positive impact. It is difficult at first to see a negative implication of the number of tackles, as an additional tackle should not give the team a disadvantage. However, the total number of tackles attempted is no indicator of the quality of defending of a team, as this variable does not measure defensive success but rather how much a team has to defend, different to the *Clearances*, which have a positive impact, as described later. A possible interpretation of this would be that defensive-minded teams seemingly are less victorious, as the number of tackles is directly linked to the amount of time a team needs to defend. As mentioned in Section 4.1, we might also interpret this as the result actually influencing the variable and not vice-versa, as teams that are down becoming more aggressive in defending. However, if this would be the case, the effect would be increased in lopsided games, and the interaction *Tackles attempted* × *Close Game* would be significant, which it is not. The number of *Dribbles* was found to have a negative impact by Liu et al. (2015), in contrast to our positive coefficient. This fits our explanation for the coefficient of the *Tackles attempted*, as more offensive-minded teams will attempt more one-on-one situations, and seemingly were more successful in 2020 than more defensive-minded teams. We can also see for the 2020 tournament that the number of *Clearances* standalone are

Table 2: Comparison of arithmetic means of coefficients and selection shares for goal-based (GB) and result-based (RB) approaches for both tournaments.

Variable	EURO 2020		EURO 2016	
	GB	RB	GB	RB
log(Market value ratio)	0.76525	2.47420	0.46060	1.06464
	(0.980)	(0.992)	(0.893)	(0.885)
Distance covered difference	0.14363	0.40981	0.11802	0.30960
	(0.976)	(0.927)	(0.776)	(0.730)
Shot accuracy	0.03602	0.08515	0.02235	0.04258
	(0.997)	(0.986)	(0.856)	(0.800)
Save rate	0.00668	0.02131	–	–
	(0.660)	(0.719)		
Tackles attempted	−0.03387	−0.11055	–	–
	(0.724)	(0.565)		
Dribbles	0.02651	0.10677	–	–
	(0.560)	(0.627)		
Clearances × Close game	0.05116	0.17047	–	–
	(0.554)	(0.378)		
Errors	0.46555	–	–	–
	(0.596)			
Tackle rate	0.03422	–	–	–
	(0.610)			
Long passes	–	0.10283	–	–
		(0.394)		
Crosses	–	−0.17915	–	–
		(0.468)		
Claims × Close game	–	1.30948	–	–
		(0.418)		
Shots from fast-break	–	–	1.01076	2.41468
			(0.915)	(0.857)
Clearances	–	–	0.03395	0.06162
			(0.551)	(0.447)
log(Market value ratio) × Knockout	–	–	0.33127	–
			(0.670)	

Note: For the selected variables for both tournaments and approaches, the respective average coefficient is depicted. The parenthesized number indicates the share of selection for each variable in the cross-fitted stability-selection algorithm.

not selected as a relevant factor, but only in combination with the *Close game* variable. This can be interpreted as the *Clearances* becoming a positive success factors when a game is close.

The variables *Distance covered difference*, log(*Market value ratio*), and *Save rate* have not been examined before. We see that a higher market value than the opponent leads to an increase in goal difference, which is not a surprising result. A better *Save rate* also leads to an improved predicted goal difference, meaning goal keeper performance is an important factor. Interestingly, running more than the opponent, as displayed in the *Distance covered difference*, is an important positive success factor as well.

4.2.2 Result-based classification

Variable selection: Similarly for the result-based approach,[4] Table 8 displays the full results for the 2020 tournament. We can see the same group of mostly selected and thus relevant variables as before. The following group in terms of selection share consists of the *Save Rate*, *Dribbles*, *Tackles attempted*, *Crosses*, *Claims* × *Close game*, *Long passes*, and *Clearances* × *Close game*, which all have a reduced π_k, but remain very stable in γ_k and υ_k. Again, the variable *Through balls* is quite unstable, which is why we choose our cut-off value there and select all prior variables.

For the 2016 tournament, the results are provided in Table 13. Once again, the group of log(*Market value ratio*), *Distance covered difference*, and *Shot accuracy* shows the highest selection proportions $\hat{\pi}$. The variable *Shots from fast-break* with a $\hat{\pi}_k = 0.857$ also belongs to this group in accordance with the goal-based approach for 2016. Behind these in the $\hat{\pi}$-ranking, we see the *Clearances* with a reduced selection share, but still substantial values of υ_k and γ_k. The latter drop substantially at the next variable, the *Claims* × *Knockout*, leading to our cutoff being set here.

Interpretation of coefficients: The main underlying assumption of the proportional odds model (3) is that the slopes of different levels are equal, leading to the same set of coefficients for all levels. The Brant test is largely not rejected providing no empirical counter-evidence against holding up the proportional odds assumption. Test results can be found in the Appendix in Table 15.

As discussed in Section 4.2.1, recall that the coefficients derived from our algorithm are subsampling averages, not coefficients from a single PO model but robust to outliers. Additionally, it is important to note that we have estimated the log-odds, and thus, the calculated effects are applicable only to these log-odds. To obtain a meaningful interpretation of these coefficients, we apply the exponential function to the arithmetic mean of the coefficient estimates from the algorithm (refer to Table 2). We proceed with all non-logarithmic variables accordingly (refer to Table 3), providing us with the respective odds-ratios (OR).

The marginal effects can be interpreted with respect to the odds of drawing or winning compared to losing, as well as winning compared to drawing or losing (see Section 2.2). According to the main assumption of the model (see Table15 for an empirical check on our data), the marginal effects coincide on both odds, as the proportional odds assumption. Please note that we estimated the log-odds and therefore can only calculate effects on those. Consequently, we apply the exponential function to the coefficients of all our non-logarithmic variables (see Table 3), which gives us the so-called odds-ratios (OR). This leads us to the interpretation that e.g. a one percentage point increase of *Shot accuracy* is associated with an increase in the odds in our model on average by 8.89 % ceteris paribus in 2020, and 4.35 % in 2016. As the variable log(*Market value ratio*) is already log-transformed, a one percent increase in the market value ration increases the odds in our model by 2.47 % in 2020 ceteris paribus and 1.06 % in 2016.

Comparison to existing literature: Variables exclusively significant in the result-based classification are the *Long passes*, *Crosses* and *Claims* × *Long passes*. In addition to the discussion in Section 4.2.1 for the common factors, we complement the comparison with the literature for these result specific variables. The negative coefficient for the number of *Crosses* might come as a surprise, similar to that

Table 3: Odds-ratios (OR) of all non-logarithmic coefficients in the PO model.

Variables	EURO 2020		EURO 2016	
	Coefficient	OR	Coefficient	OR
Shot accuracy	0.08515	1.08888	0.04258	1.04350
Distance covered difference	0.40981	1.50654	0.30960	1.36288
Save rate	0.02131	1.02153	–	–
Dribbles	0.10677	1.11268	–	–
Tackles attempted	−0.11055	0.89534	–	–
Crosses	−0.17915	0.83598	–	–
Claims × Close game	1.30948	3.70425	–	–
Long passes	0.10283	1.10831	–	–
Clearances × Close game	0.17047	1.18587	–	–
Clearances	–	–	0.06162	1.06356
Shots from fast-break	–	–	2.41468	11.18615

4 Here, the 1,000 LASSO-PO-models of the cross-fitted stability-selection algorithm are estimated with the help of the ordinalNet package (Wurm et al. 2021). The ensuing non-penalized model is estimated with the help of the vglm function of the VGAM R-package.

of the *Tackles attempted* before. We could argue, that an unsuccessful cross can lead to highly dangerous counter attacking situations for the opponent. However, it would be more plausible to interpret the total number of crosses as a more general description of how a team is trying to attack the goal. Teams that find more direct and central ways into the penalty area appear to be more successful than teams that to rely on crosses from the wing areas. This negative impact is in fact confirmed in most of the literature with a similar negative coefficient (see Lepschy et al. 2020, 2021; Liu et al. 2015; Peñas et al. 2010). A dedicated research on this topic (Sarkar 2018) found teams that have a high chance of scoring from crosses to be less likely to attempt them, as their opponent adapts to their strength and uses offside traps more frequently, decreasing the number of crosses attempted. The positive impact of the *Long passes* is also contrary effects found in existing research for other tournaments. dos Reis et al. (2017), who analyzed the 2014 World Cup, found these to have a negative impact on the ball possession due to their low success rate, and also not helpful in creating chances, as only 1 % of them lead to a shot on goal. The *Claims* × *Close game* have not been examined elsewhere. The positive coefficient can be interpreted as the *Claims* only being a relevant success factor in close games, similar to the *Clearances* × *Close game*, which are also significant in the result based model for 2020.

4.3 Robustness checks

A potential drawback with our algorithm and consequent variable selection is the risk of overlooking significant variables if they rank below insignificant variables in terms of selection share ($\hat{\pi}_k$). To mitigate this, we conduct a sanity check using the Bayesian Information Criterion (BIC) model evaluation criteria (see Schwarz 1978). Accordingly, we add the variables to a new post model in the order of their selection share and calculate the BIC for each model.

If we observe an improved BIC when adding variables that we did not include in our final model, it suggests we might have missed important variables. Figure 7 illustrates the resulting plot for each of the four models, with the *x*-axis representing the variables in order of their selection share. In general, the addition of further variables does

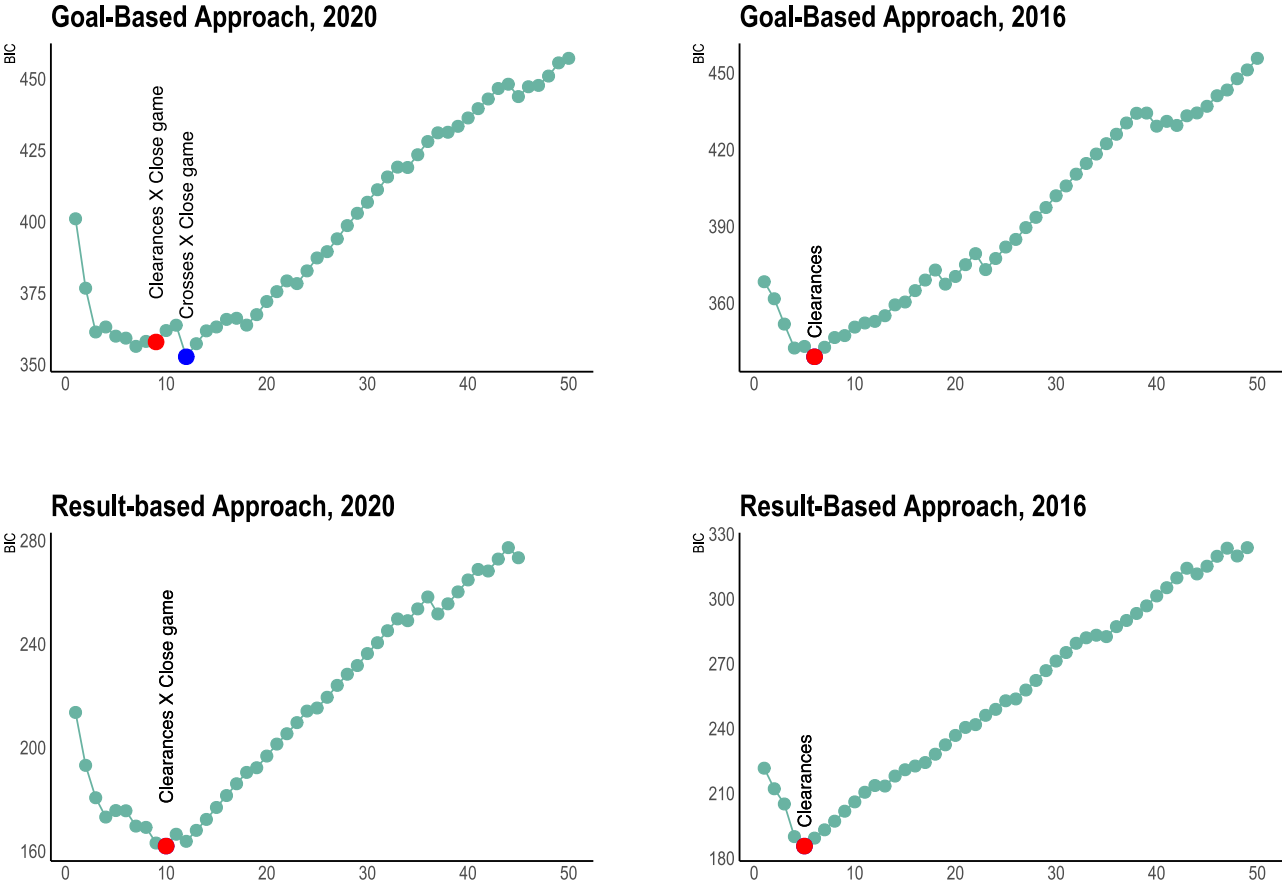

Figure 7: BIC for a post model when variables are added in the order of Algorithm 1. The blue dot is the minimum value. The red dot is the last selected variable (these are the same variables if only a red dot is visible).

not enhance model performance in terms of BIC. The lone exception is the Goal-Based Approach for 2020. Here, including all variables up to the *Crosses* × *Close game* interaction would yield a better BIC than stopping at the *Clearances* × *Close game*, as we have done.

However, this approach would necessitate including the *Through balls* and *Long passes* variables, with *Through balls* being particularly insignificant, as indicated by its extraordinarily high v_k of 37.144 and γ_k of 0.5. Therefore, we conclude that halting at the *Clearances* × *Close game* is indeed the optimal choice since the BIC does not significantly increase. In summary, these results do not suggest that we have omitted any potentially important variables through our model selection method.

4.4 Discussion

The selected variables can be categorized into three different groups based on their selection frequency. For the first group, we see a group of highly selected variables with also stable coefficients across both examined tournaments and approaches in the log(*Market value ratio*), *Distance covered difference*, and *Shot accuracy*. Their coefficients mostly lie in the same range when comparing these two tournaments for the same approach, with the biggest anomaly being the log(*Market value ratio*) in the result based approach ranging from 1.157 to 2.369. The following two groups consist of variables only selected for either of the two tournaments. The *Shots from fastbreak* have to be emphasized here, as they were not significant for the EURO 2020, but seem hugely relevant for the EURO 2016, with the selection share being between $\pi_k = 0.967$ (highest of all) and $\pi_k = 0.958$, while also showing extremely high coefficients. That is especially interesting, considering the average number of *Shots from fast break* had increased by 88.1 % for the EURO 2020 (see Table 10). Therefore it seems like counterattacking play was more common in 2020, but nonetheless less successful.

The three identified groups also provide insights for one of our central original research questions: the distinction of success factors that are generally important from those whose importance varies between tournaments comprising mostly on-field variables. The variables from the first group seem to be commonly significant off-field success factors in national team football, whereas the other groups apparently contain on-field factors that might vary between tournaments. This variation between tournaments might be caused by the different teams' ever-adapting play styles, and successful tactics being countered with adapted new play styles, as well as general tactical trends in football. Indeed, the selections with our methodology support that proposition. All of the selected covariates in the first group are only mildly, if at all, associated with the style of play. In contrast, most other success factors are linked to how a team approaches the game, especially the *Tackles attempted*, *Dribbles*, *Long passes*, *Crosses*, and *Shots from fast-break*. This hypothesis is also supported by the tournaments-specific characteristics. In the EURO 2020, the average goals scored per match was at 2.78, the highest value since 1976, resulting in offensive-minded teams being more successful, as well as defensive-minded teams performing worse. This can be seen from our results, for example in the negative effect of *Tackles attempted* or the positive effect of offensive actions as well as possession linked variables such as the *Dribbles*. The EURO 2016, on the other hand, was not only subjectively perceived as a defensive-minded tournament,[5] but actually had the second-lowest goal average since 1980 with 2.12. Our estimated model reflects this with the *Tackles attempted* having no significant influence as well as in the high importance of fast-breaks represented by the significance of the *Shots from fast-break*, which are often the main source of goal scoring chances for defensive-minded, deep-defending teams. Table 4 shows changes in the impact of different playing styles in 2016 and 2020. In particular, it confirms and highlights that the counter-attacking style was more successful in 2016 compared to 2020, with the effect being reversed for the possession-oriented playing style that contains e.g. the *Dribbles*. It is important to note here that these play-style related differences might have been caused by the mentioned increase of available substitutions from three to five, enabling teams to use more run-heavy play-styles. The variable *Substitutions* itself however was not significant, which is not surprising as both teams had the same amount of subs available.

As we have seen in the previous sections, there are different variables being selected in the goal-based and result-based approach. For the 2020 tournament the *Errors* and *Tackle rate* are only selected in the goal-based approach, whereas the *Long passes*, *Crosses* and *Claims* × *Close game*

Table 4: Impact of different playing styles across tournaments.

Play-style groups	Accuracy 2016	Accuracy 2020	Difference
Counter-attacking	0.53	0.42	−0.11
High-pressing	0.46	0.23	−0.23
Set-piece oriented	0.46	0.54	0.08
Possession-oriented	0.41	0.57	0.16
Defensive	0.37	0.40	0.03

We display the cross-validated prediction accuracy for results from play-style sorted groups of all variables using a Linear Discriminant Analysis (LDA) classifier. The assignment of the variables into the groups of different playing style buckets is visualized in Figure 10.

5 https://www.theguardian.com/football/blog/2016/jul/11/euro-2016-fairytales-wales-iceland-defence.

are only present in the result-based approach. A possible explanation would be that the goal-based approach is better in modeling factors directly connected with the scoring of goals, whereas in the result-based approach, the features describing the style of play of a team are more relevant. That would strengthen our point made before that the negative sign of the marginal effect of *Crosses* is related to the way a team creates its chances. We can see another argument for this in the variable *Errors* only being selected in the goal-based approach, as every error made is directly related to a goal or at least a shot on goal, but does not describe the way a team plays or creates its chances. This could also explain the appearance of the *Tackle rate* in only the goal-based approach, as a higher rate of successful tackles leads to more situations in which the opposing defense is left unsorted, resulting in more goal scoring opportunities. One of the secondary goals of this paper was to evaluate the unique circumstances of the EURO 2020. Some of those were non-COVID-19 related, such as the tournament being played all over the continent instead of as usually in one or two countries, which we modeled in the variables *Home advantage* and *Travel distance*. Others were associated to the pandemic, mainly since the number of spectators allowed varied massively between the different countries. This was contained in our model through the feature *wHA*. Interestingly, in neither of the models was one of these variables deemed significant, implying that there was no significant influence of these circumstances.

5 Conclusions

This paper aims to identify the important success factors in national team football by analyzing the UEFA EURO 2020.

We achieved our results by applying a new method in the 'cross-fitted stability-selection', which helped us to overcome both the sparsity of our data, as well as the issue of multicollinearity. Through the LASSO estimation, we were able to select the relevant variables, and the repeated subsampling enabled us to evaluate the significance of the covariates mainly based on their share of selection, as well as other metrics such as the variance of their coefficient, and the similarity in the direction of the effects.

Comparing the results for the two most recent European Championships in 2020 and 2016 allowed us to differentiate a group of generally valid, tournament independent significant variables in primarily the log(*Market value ratio*), the difference in running distance *Distance covered difference*, and the *Shot accuracy*, from factors that only played a role in specific tournaments. These factors are mainly dependent on a team's style of play, especially the number of *Tackles attempted*, *Dribbles*, *Long passes*, *Crosses*, and *Shots from fast-break*. This could be attributed to changing football tactics trends and the varying success of differing playing styles. We can also show that the success of certain playing styles changes across the examined competitions as displayed in Table 4. It further reflects that successful teams are often countered by adapted tactics, leading to again shifting success related to the specific playing styles. Therefore, those variables do not necessarily represent a success factor but rather show that the linked playing style was more winning in the examined tournament. In contrast, the first group of variables seems important regardless of playing style and tournament. Another secondary goal was to evaluate the influence of the exceptional circumstances of the COVID-19 pandemic. These were mainly represented in the weighted home advantage *wHA*, which was not found to be significant in either of the observed models. Hence, our results do not show a pandemic-related influence of the varying number of spectators. Another unique circumstance of the EURO 2020, the tournament being hosted in a multitude of cities across Europe, did not show significance, as the corresponding variable *Travel distance* was not relevant in any of the models.

The absence of the *Home advantage* in our models for the EURO 2020 is another important finding, as this tournament for the first time ever saw many different teams host matches. This enabled us to better examine the effect of the home advantage itself, as the strength of the effect in previous tournaments directly depended on the strength of the host nation itself.

For future research, it would be interesting to employ variables that are better suited to describe the tactical components, for example, the time of pressing, the average height of the pressing line, or the defensive compactness. More covariates representing contextual variables, such as, for example, the recent form of a team in terms of their last results, might also lead to more concise models. These variables are unfortunately not publicly available. Caused by the increased awareness for analytics in football, as we described in Section 1, there is a great number of advanced statistical variables developed explicitly to measure footballing success in a better way, such as the prominent 'Packing' statistic,[6] which measures the number of opponents outplayed. These variables are specifically put together for professional use and were unfortunately not available for our academic purpose. However, it is debatable how much value those features would add to our conducted research on top of the employed comprehensive set of

6 https://www.tz.de/sport/fussball/em-2016-netz-spott-packing-bedeuten-neuen-daten-zr-6483530.html.

observable variables. The advanced variables are designed explicitly as important factors but are essentially based on multiple input from many of the regressors that we cover explicitly and can already handle due to the LASSO-based approach. Therefore expect their additional information to be limited rather deluting the impact of their underlying original inputs in a multiple regression setting.

Research ethics: Not applicable.
Author contributions: The authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: The authors state no conflict of interest.
Research funding: None declared.
Data availability: The raw data can be obtained on request from the corresponding author.

Appendix

Table 5: Descriptive statistics for remaining covariates.

Covariate	Arithmetic mean	SD	Coef. of variation	Lower bound 95 % CI	Upper bound 95 % CI
Shots	12.36	5.93	0.48	11.20	13.53
Passes	504.47	156.68	0.31	473.69	535.25
Dribbles	15.78	6.36	0.40	14.54	17.03
Tackles attempted	22.78	8.41	0.37	21.13	24.44
Interceptions	10.03	4.55	0.45	9.13	10.92
Clearances	16.68	8.78	0.53	14.95	18.40
Blocks	11.46	4.30	0.38	10.62	12.30
Offsides	1.74	1.50	0.86	1.44	2.03
Fouls	11.68	3.97	0.34	10.90	12.46
Aerial.duels	15.10	7.02	0.46	13.72	16.48
Loss.of.possession	23.53	6.75	0.29	22.20	24.86
Errors	0.26	0.51	1.91	0.16	0.36
Claims	0.42	0.68	1.62	0.29	0.56
Punches	0.43	0.79	1.83	0.28	0.59
Shots from open play	8.56	4.78	0.56	7.62	9.50
Shots from fastbreak	0.46	0.69	1.49	0.33	0.59
Shots from set pieces	3.00	2.12	0.71	2.58	3.42
Penalties	0.17	0.40	2.40	0.09	0.24
Crosses	15.99	8.36	0.52	14.35	17.63
Freekicks	11.49	4.14	0.36	10.68	12.30
Corners	4.52	2.90	0.64	3.95	5.09
Through balls	2.10	1.82	0.87	1.74	2.46
Throw ins	19.19	5.99	0.31	18.01	20.36
Key passes	9.34	5.16	0.55	8.33	10.36
Long passes	53.31	11.18	0.21	51.12	55.51
Chipped passes	52.08	13.64	0.26	49.40	54.76
Headed passes	30.00	10.73	0.36	27.89	32.11
Passes into defensive third	107.57	33.25	0.31	101.04	114.10
Passes into final third	157.43	65.23	0.41	144.62	170.24
Average age	28.06	1.23	0.04	27.81	28.30
Market value	268.38	174.83	0.65	234.04	302.72
Ball possession	50.00	13.01	0.26	47.45	52.55
Pass rate	82.46	7.38	0.09	81.01	83.91
Tackle rate	50.00	6.07	0.12	48.81	51.19
Distance covered	112.06	14.12	0.13	109.29	114.83
Home advantage	0.00	0.73	–	−0.14	0.14
Travel distance	988.74	988.98	1.00	794.48	1,182.99
wHA	0.00	21184.84	–	−4,161.10	4,161.10
Shot accuracy	33.78	19.01	0.56	30.05	37.51
Save rate	65.32	32.58	0.50	58.92	71.72
Market value ratio	1.81	2.25	1.24	1.37	2.25
Distance covered difference	0.00	4.93	–	−0.97	0.97
Substitutions	4.45	1.05	0.24	4.25	4.66

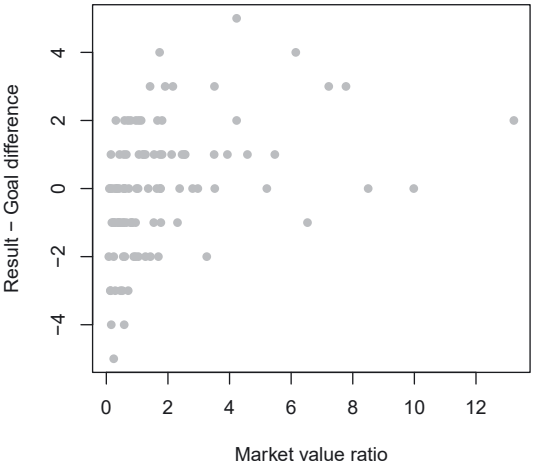

Figure 8: *Result – Goal difference* by *Market value ratio*.

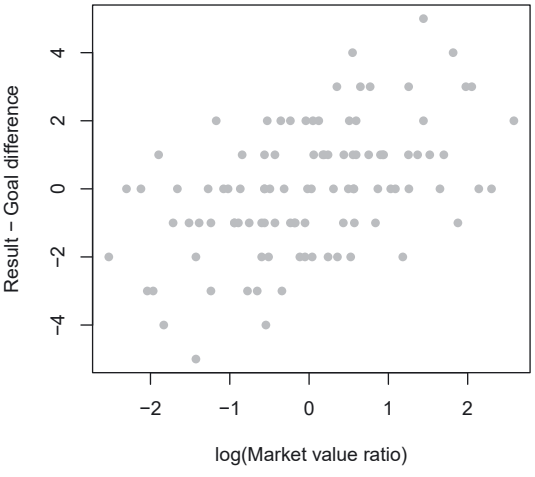

Figure 9: *Result – Goal difference* by log(*Market value ratio*).

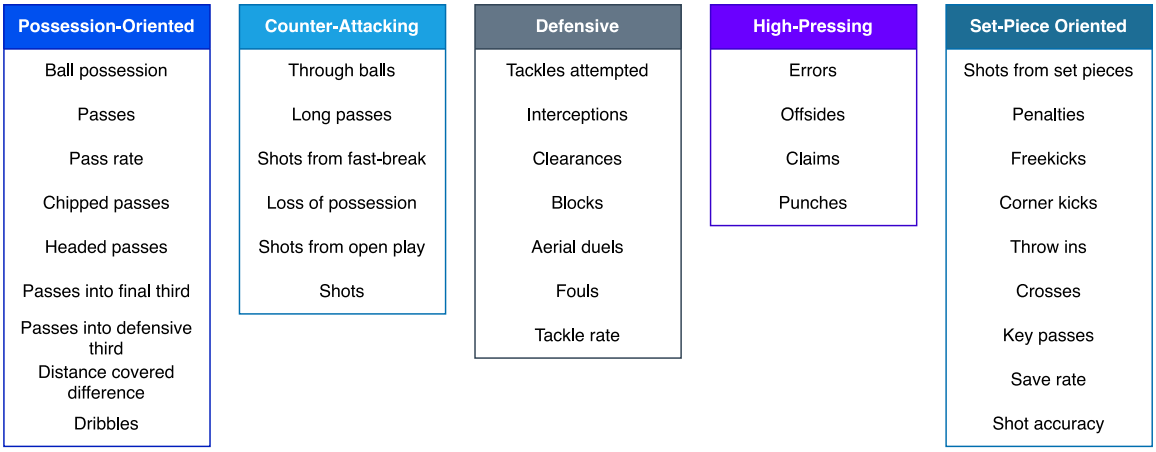

Figure 10: Assignment of all variables to playing style buckets.

Table 6: Full results of cross-fitted stability-selection for the goal-based approach for 2020.

Variable	π_k	Δ_k	$\bar{\beta}_k$	$\sigma(\beta_k)$	v_k	γ_k	AIC
Intercept	1.000	0.000	−3.78674	4.311	−1.138	0.919	0.00
Shot accuracy	0.997	0.003	0.03602	0.011	0.312	0.997	400.98
log(Market value ratio)	0.980	0.017	0.76525	0.248	0.323	0.994	376.67
Distance covered difference	0.976	0.004	0.14363	0.044	0.306	0.998	361.34
Tackles attempted	0.679	0.297	−0.03960	0.033	−0.821	0.932	363.09
Save rate	0.660	0.019	0.00668	0.005	0.771	0.932	359.88
Tackle rate	0.610	0.050	0.03422	0.032	0.944	0.884	359.18
Errors	0.596	0.014	−0.46555	0.395	−0.849	0.894	356.31
Dribbles	0.560	0.036	0.02651	0.041	1.560	0.784	358.04
Clearances × Close game	0.554	0.006	0.05116	0.039	0.769	0.964	357.85
Through balls	0.373	0.181	0.00360	0.134	37.144	0.499	361.84
Long passes	0.366	0.007	0.01670	0.022	1.345	0.806	363.72
Crosses × Close game	0.326	0.040	−0.02799	0.045	−1.599	0.801	352.66
Shots from fast-break	0.291	0.000	−0.09402	0.534	−5.680	0.656	357.24
log(Market value ratio) × Knockout	0.291	0.035	−0.26407	0.837	−3.168	0.698	361.74
Shots from set pieces	0.273	0.018	0.04782	0.163	3.412	0.630	363.10
Average age	0.270	0.003	0.10380	0.173	1.670	0.774	365.75
Offsides	0.252	0.018	−0.04779	0.181	−3.796	0.575	366.14
Ball possession	0.227	0.025	0.01203	0.055	4.611	0.612	363.75
Crosses	0.225	0.002	−0.01426	0.056	−3.910	0.640	367.42
Throw ins	0.208	0.017	−0.00642	0.052	−8.144	0.591	372.00
Distance travelled × Knockout	0.199	0.009	−0.00007	0.001	−8.093	0.533	375.47
Penalties	0.194	0.000	−0.14147	0.647	−4.573	0.629	379.23
Passes into final third × Close game	0.194	0.005	−0.00580	0.008	−1.436	0.799	378.29
Loss of possession	0.190	0.004	0.01747	0.049	2.781	0.674	382.74
Freekicks	0.185	0.005	0.00691	0.071	10.289	0.503	387.27
Shots	0.181	0.004	0.08210	0.155	1.883	0.796	389.47
Shots from open play	0.180	0.001	0.00165	0.161	97.877	0.522	393.97
Distance travelled	0.177	0.003	0.00004	0.000	6.777	0.576	398.59
Fouls	0.173	0.004	0.01710	0.074	4.332	0.613	402.90
Punches	0.171	0.000	0.03362	0.569	16.931	0.573	406.77
Substitutions	0.171	0.000	−0.01792	0.298	−16.651	0.538	411.15
Claims × Close game	0.171	0.002	−0.02942	0.504	−17.136	0.550	415.63
Passes	0.167	0.000	0.00077	0.004	5.749	0.503	419.06
Headed passes	0.167	0.004	−0.00687	0.035	−5.141	0.563	418.94
Claims × Knockout	0.165	0.002	0.18180	0.922	5.074	0.697	423.40
Errors × Knockout	0.162	0.003	−0.04842	2.810	−58.044	0.469	428.02

Table 6: (continued)

Variable	π_k	Δ_k	$\bar{\beta}_k$	$\sigma(\beta_k)$	v_k	γ_k	AIC
Blocks	0.155	0.007	0.01057	0.070	6.613	0.581	431.08
Save rate × Knockout	0.151	0.000	0.00098	0.015	15.695	0.411	431.28
Distance covered difference × Knockout	0.151	0.004	−0.06177	0.120	−1.939	0.709	433.36
Shots from open play × Close game	0.149	0.002	−0.02827	0.128	−4.524	0.658	436.28
Home advantage × Knockout	0.146	0.003	−0.01373	8.895	−647.767	0.397	439.54
Interceptions	0.140	0.000	0.00564	0.071	12.563	0.493	442.93
Aerial duels	0.140	0.006	−0.01381	0.050	−3.647	0.621	446.59
Home advantage	0.138	0.002	−0.35928	0.665	−1.850	0.841	448.09
wHA	0.134	0.004	−0.00000	0.000	−32.356	0.485	443.76
Corner kicks	0.128	0.006	−0.08417	0.104	−1.231	0.805	447.21
Passes into defensive third	0.118	0.010	0.00238	0.012	5.043	0.686	447.68
Penalties × Knockout	0.102	0.016	−0.83550	2.408	−2.882	0.725	450.95
Chipped passes	0.101	0.001	0.00519	0.031	5.910	0.634	455.57
Pass rate	0.087	0.014	0.02840	0.077	2.718	0.678	457.16
Clearances	0.082	0.005	−0.01702	0.059	−3.487	0.646	457.24
Punches × Knockout	0.078	0.000	0.10535	1.714	16.267	0.423	461.76
Punches × Close game	0.078	0.004	−0.14460	0.693	−4.790	0.603	463.82
Claims	0.076	0.002	0.02791	0.716	25.673	0.592	468.39
Shots from fast-break × Close game	0.075	0.001	−0.08879	1.055	−11.886	0.573	468.01
Through ballls × Knockout	0.074	0.001	−0.21973	0.670	−3.048	0.635	466.80
Corner kicks × Knockout	0.072	0.002	0.08054	0.327	4.060	0.667	470.26
Shot accuracy × Knockout	0.063	0.009	0.00301	0.051	16.775	0.365	473.55
Shots from fast-break × Knockout	0.057	0.000	0.17978	1.196	6.654	0.561	477.36
Dribbles × Close game	0.057	0.006	−0.00335	0.104	−31.127	0.456	481.89
Passes into defensive third × Knockout	0.052	0.005	−0.00210	0.023	−10.864	0.346	486.41
			Variables with $\pi_k < 0.05$ not shown				

Table 7: Post-regression with HC3 standard errors for final model in goal-based approach for 2020.

	Estimate	Std. error	*t* value	Pr(|*t*|)
(Intercept)	−3.7769	1.2098	−3.1219	0.0024
Shot accuracy	0.0369	0.0066	5.5725	0.0000
log(Market value ratio)	0.5501	0.2122	2.5921	0.0111
Distance covered difference	0.1453	0.0281	5.1778	0.0000
Save rate	0.0080	0.0044	1.8116	0.0733
Tackles attempted	−0.0363	0.0181	−2.0074	0.0476
Errors	−0.4754	0.2684	−1.7708	0.0799
Tackle rate	0.0433	0.0222	1.9540	0.0537
Dribbles	0.0348	0.0277	1.2548	0.2127
Clearances × Close game	0.0235	0.0126	1.8734	0.0642

Table 8: Full results of cross-fitted stability-selection for the proportional odds model for 2020.

Variable	π_k	Δ_k	$\bar{\beta}_k$	$\sigma(\beta_k)$	v_k	γ_k	AIC
Intercept1	0.998	0.000	−3.07532	10.394	−3.380	0.781	0.00
Intercept2	0.998	0.000	−5.77452	10.569	−1.830	0.896	0.00
log(Market value ratio)	0.992	0.006	2.47420	1.644	0.665	1.000	213.56
Shot accuracy	0.986	0.006	0.08515	0.053	0.621	0.992	193.12
Distance covered difference	0.927	0.059	0.40981	0.277	0.676	0.999	180.60
Save rate	0.719	0.208	0.02131	0.026	1.214	0.887	173.13
Dribbles	0.627	0.092	0.10677	0.139	1.306	0.863	175.64
Tackles attempted	0.565	0.062	−0.11055	0.145	−1.308	0.903	175.56
Crosses	0.468	0.097	−0.17915	0.185	−1.030	0.942	169.64
Claims × Close game	0.418	0.050	1.30948	2.295	1.752	0.911	169.15
Long passes	0.394	0.024	0.10283	0.115	1.121	0.929	163.09
Clearances × Close game	0.378	0.016	0.17047	0.198	1.161	0.974	161.92
Through balls	0.362	0.016	−0.08139	0.396	−4.860	0.610	166.49
Crosses × Close game	0.309	0.053	−0.14415	0.183	−1.271	0.848	163.76
Punches × Knockout	0.245	0.064	−0.79240	3.881	−4.898	0.735	167.99
Shots from fast-break	0.240	0.005	−0.15183	1.992	−13.119	0.500	172.23
Punches × Close game	0.212	0.028	0.21140	2.167	10.251	0.618	176.85
Clearances	0.181	0.031	0.08789	0.167	1.898	0.773	181.45
Distance covered difference × Knockout	0.172	0.009	−0.24756	0.424	−1.712	0.808	186.00
Home advantage × Knockout	0.161	0.011	23.63529	101.626	4.300	0.634	190.39
Tackle rate	0.154	0.007	−0.00013	0.125	−999.715	0.552	192.19
Chipped passes	0.143	0.011	0.01765	0.096	5.453	0.587	196.70
Errors	0.135	0.008	0.47919	1.911	3.988	0.637	201.33
Blocks	0.131	0.000	−0.04766	0.246	−5.172	0.603	205.33
Loss of possession	0.131	0.004	0.01795	0.130	7.248	0.603	209.63
Passes into defensive third	0.122	0.009	0.01291	0.027	2.093	0.746	214.08
Shots from open play	0.121	0.001	0.02550	0.328	12.870	0.488	215.26
Substitutions	0.117	0.004	0.21304	0.869	4.079	0.684	219.42
Claims × Knockout	0.113	0.004	0.30693	4.684	15.261	0.558	224.01
Shots from fast-break × Close game	0.098	0.015	0.57214	3.746	6.546	0.582	228.32
Fouls	0.095	0.000	0.08968	0.272	3.030	0.684	231.68
Home advantage	0.095	0.003	−1.09318	1.648	−1.507	0.832	236.30
Punches	0.090	0.005	−0.53029	2.379	−4.487	0.578	240.51
Average age	0.086	0.004	−0.22045	0.917	−4.159	0.628	245.13
Throw ins	0.080	0.000	−0.10398	0.232	−2.231	0.675	249.70
Offsides × Knockout	0.080	0.006	−0.26058	1.974	−7.575	0.575	249.01
Penalties	0.079	0.001	−0.20333	2.198	−10.812	0.582	253.57
Penalties × Knockout	0.075	0.004	2.46726	6.770	2.744	0.640	258.17
Interceptions × Close game	0.073	0.002	−0.02254	0.289	−12.808	0.425	251.62
Freekicks	0.069	0.000	−0.04046	0.235	−5.808	0.536	255.53
Distance travelled	0.069	0.004	0.00007	0.002	24.506	0.435	260.14
Headed passes	0.067	0.002	0.03381	0.129	3.808	0.701	264.76
Key passes	0.066	0.001	−0.00102	0.228	−223.286	0.530	268.81
Passes	0.062	0.004	−0.00075	0.014	−18.049	0.613	268.20
Loss of possession × Close game	0.059	0.003	0.00714	0.154	21.566	0.475	272.82
Offsides	0.058	0.001	−0.06927	0.973	−14.045	0.448	277.23
Corner kicks	0.053	0.005	−0.02538	0.619	−24.366	0.698	273.32
Claims	0.052	0.001	−1.49310	5.016	−3.359	0.404	
Variables with $\pi_k < 0.05$ not shown							

Table 9: Post-regression for final model in result-based approach for 2020.

	Estimate	Std. error	*z* value	Pr(|*z*|)
Intercept1	−4.1604	1.9179	−2.169	0.0301
Intercept2	−6.9968	2.0613	−3.394	0.0007
log(Market value ratio)	1.9972	0.4039	4.945	0.0000
Shot accuracy	0.0645	0.0183	3.525	0.0004
Distance covered difference	0.3357	0.0702	4.782	0.0000
Save rate	0.0133	0.0081	1.630	0.1030
Dribbles	0.1216	0.0499	2.438	0.0148
Tackles attempted	−0.1126	0.0367	−3.071	0.0021
Crosses	−0.1504	0.0462	−3.256	0.0011
Claims × Close game	0.7372	0.4852	1.519	0.1287
Long passes	0.0853	0.0290	2.941	0.0033
Clearances × Close game	0.0684	0.0292	2.344	0.0191

Table 10: Comparison of means of all variables for EURO 2020 and 2016. *p*-values for two-sided Mann–Whitney-*U*-test.

Variable	Mean UEFA EURO 2020	Mean UEFA EURO 2016	Change (in %)	*p*-Value
Shots	12.363	13.422	−7.9	0.182
Passes	504.471	453.049	11.3	0.018
Dribbles	15.784	15.716	0.4	0.703
Tackles attempted	22.784	23.765	−4.1	0.284
Interceptions	10.029	13.510	−25.8	0.000
Clearances	16.676	23.431	−28.8	0.000
Blocks	11.461	13.480	−15.0	0.005
Offsides	1.735	1.882	−7.8	0.617
Fouls	11.676	12.598	−7.3	0.063
Aerial duels	15.098	17.333	−12.9	0.013
Loss of possession	23.529	21.755	8.2	0.097
Errors	0.265	0.275	−3.6	0.784
Claims	0.422	0.863	−51.1	0.001
Punches	0.431	0.529	−18.5	0.298
Shots from open play	8.559	9.412	−9.1	0.182
Shots from fast-break	0.461	0.245	88.0	0.004
Shots from set pieces	3.000	3.588	−16.4	0.075
Penalties	0.167	0.118	41.7	0.405
Crosses	15.990	20.343	−21.4	0.002
Freekicks	11.490	12.127	−5.2	0.146
Corner kicks	4.520	5.284	−14.5	0.108
Through ballls	2.098	1.412	48.6	0.001
Throw ins	19.186	22.451	−14.5	0.001
Key passes	9.343	10.147	−7.9	0.182
Long passes	53.314	63.157	−15.6	0.000
Chipped passes	52.078	47.245	10.2	0.016
Headed passes	30.000	39.147	−23.4	0.000
Passes into defensive third	107.569	82.500	30.4	0.000
Passes into final third	157.431	163.784	−3.9	0.365
Average age	28.057	28.257	−0.7	0.275
Ball possession	50.000	50.000	0.0	1.000
Pass rate	82.461	78.657	4.8	0.001
Tackle rate	50.000	50.000	0.0	1.000
Home advantage	0.000	0.000	0.0	1.000
Distance travelled	988.737	418.526	136.2	0.003
wHA	0.000	–	–	–
Shot accuracy	33.779	31.571	7.0	0.695
Save rate	65.324	68.832	−5.1	0.161
log(Market value ratio)	0.001	−0.001	−264.4	0.995
Distance covered difference	0.000	0.000	0.0	1.000
Substitution	4.451	2.873	55.0	0.000

Table 11: Full results of cross-fitted stability-selection for the goal-based approach for 2016 *wHA* was cut out due to the almost perfect correlation to *Home advantage* (as only France had home games, with almost the same amount of spectators every time).

Variable	π_k	Δ_k	$\bar{\beta}_k$	$\sigma(\beta_k)$	v_k	γ_k	AIC
Intercept	0.982	0.000	−1.83156	3.352	−1.830	0.852	0.00
Shots from fast-break	0.915	0.067	1.01076	0.620	0.613	0.986	368.31
log(Market value ratio)	0.893	0.022	0.46060	0.180	0.391	0.991	361.65
Shot accuracy	0.856	0.037	0.02235	0.011	0.487	0.991	351.75
Distance covered difference	0.766	0.090	0.11802	0.052	0.442	0.983	342.39
log(Market value ratio) × Knockout	0.670	0.096	0.33127	0.556	1.679	0.830	343.02
Clearances	0.551	0.119	0.03395	0.028	0.811	0.946	338.97
Tackle rate	0.415	0.136	0.00223	0.042	19.064	0.549	342.69
Claims × Knockout	0.411	0.004	0.29415	1.035	3.519	0.774	346.49
Save rate	0.405	0.006	0.00354	0.006	1.596	0.770	347.25
Distance covered difference × Knockout	0.386	0.019	0.03927	0.284	7.230	0.676	350.57
Crosses	0.355	0.031	−0.02955	0.050	−1.680	0.823	352.18
Key passes	0.344	0.011	0.03009	0.081	2.681	0.602	352.81
Home advantage	0.299	0.045	0.12635	0.719	5.693	0.602	355.02
Claims	0.292	0.007	−0.03126	0.671	−21.477	0.575	359.27
Fouls	0.273	0.019	0.04512	0.080	1.768	0.773	360.38
Punches × Knockout	0.271	0.002	0.14025	1.206	8.603	0.531	364.84
Penalties × Knockout	0.259	0.012	0.04147	1.898	45.778	0.556	368.97
Home advantage × Knockout	0.249	0.010	0.29546	1.203	4.073	0.699	372.89
Shots from fast-break × Close game	0.237	0.012	−1.36613	1.117	−0.817	0.903	367.37
Errors	0.232	0.005	−0.03037	0.451	−14.844	0.483	370.37
Tackles attempted	0.226	0.006	0.01142	0.030	2.632	0.726	374.99
Dribbles	0.224	0.002	0.00531	0.074	13.859	0.460	379.30
Offsides × Knockout	0.219	0.005	−0.29467	0.539	−1.828	0.804	373.06
Shots from set pieces	0.206	0.013	−0.00856	0.097	−11.349	0.549	377.39
Errors × Knockout	0.195	0.011	−0.01993	1.546	−77.598	0.626	381.88
Blocks	0.174	0.021	−0.00541	0.058	−10.735	0.586	384.86
Aerial duels	0.166	0.008	−0.02553	0.091	−3.580	0.669	389.48
Distance travelled	0.163	0.003	−0.00000	0.001	−272.261	0.479	393.49
Through ballls	0.151	0.012	−0.09747	0.191	−1.957	0.781	397.33
Substitutions	0.145	0.006	0.05553	0.882	15.891	0.497	401.92
Average age	0.142	0.003	0.06778	0.211	3.107	0.634	405.77
Passes into defensive third	0.141	0.001	0.00085	0.010	11.912	0.596	410.37
Interceptions	0.136	0.005	−0.00838	0.054	−6.447	0.581	414.55
Punches	0.135	0.001	−0.08358	0.500	−5.976	0.563	418.23
Throw ins	0.127	0.008	0.00670	0.053	7.904	0.559	422.29
Freekicks	0.125	0.002	0.00242	0.083	34.379	0.560	425.98
Offsides	0.119	0.006	0.02265	0.205	9.037	0.588	430.38
Shots	0.114	0.005	0.00130	0.075	57.644	0.482	434.17
Loss of possession × Close game	0.113	0.001	0.01724	0.065	3.770	0.584	434.25
Shots from fast-break × Knockout	0.112	0.001	−0.80522	1.174	−1.458	0.875	429.13
Loss of possession	0.111	0.001	0.00260	0.080	30.771	0.577	431.02
Through ballls × Knockout	0.109	0.002	−0.05042	0.441	−8.753	0.532	429.43
Headed passes	0.107	0.002	−0.00874	0.025	−2.911	0.598	433.19
			Variables with $\pi_k < 0.1$ not shown				

Table 12: Post-regression with HC3 standard errors for final model in goal-based approach for 2016.

	Estimate	Std. error	t value	Pr($\|t\|$)
(Intercept)	−1.6987	0.4202	−4.0424	0.0001
Shots from fast-break	0.6302	0.3410	1.8478	0.0677
log(Market value ratio)	0.4618	0.0940	4.9114	0.0000
Shot accuracy	0.0220	0.0074	2.9516	0.0040
Distance covered difference	0.1226	0.0362	3.3852	0.0010
log(Market value ratio) × Knockout	0.2934	0.2062	1.4228	0.1581
Clearances	0.0363	0.0146	2.4848	0.0147

Table 13: Full results of cross-fitted stability-selection for proportional-odds model for 2016 *wHA* was cut out due to the almost perfect correlation to *Home advantage* (as only France had home games, with almost the same amount of spectators every time).

Variable	π_k	Δ_k	$\bar{\beta}_k$	$\sigma(\beta_k)$	v_k	γ_k	AIC
Intercept1	0.995	0.000	−1.11107	5.909	−5.319	0.659	0.00
Intercept2	0.995	0.000	−3.04457	5.948	−1.954	0.899	0.00
log(Market value ratio)	0.885	0.110	1.06464	0.478	0.449	0.999	221.63
Shots from fast-break	0.857	0.028	2.41468	2.171	0.899	0.993	212.11
Shot accuracy	0.800	0.057	0.04258	0.024	0.574	0.988	205.07
Distance covered difference	0.730	0.070	0.30960	0.124	0.401	0.999	189.97
Clearances	0.447	0.283	0.06162	0.052	0.847	0.926	185.58
Claims × Knockout	0.352	0.095	0.38655	0.910	2.355	0.778	189.28
Home advantage	0.289	0.063	0.93299	3.112	3.336	0.682	193.11
Distance covered difference × Knockout	0.287	0.002	0.04660	0.227	4.862	0.592	197.10
Claims	0.268	0.019	−0.11567	0.491	−4.247	0.586	201.73
Shots from set pieces	0.201	0.067	−0.02531	0.170	−6.719	0.582	206.04
Key passes	0.194	0.007	−0.03071	0.138	−4.494	0.660	210.50
Save rate	0.192	0.002	0.00337	0.010	3.114	0.625	213.61
Crosses	0.163	0.029	−0.04480	0.072	−1.608	0.828	213.36
Punches × Knockout	0.160	0.003	−1.32656	4.571	−3.446	0.750	217.98
log(Market value ratio) × Knockout	0.134	0.026	−0.14766	0.734	−4.971	0.657	220.91
Fouls	0.120	0.000	0.03536	0.138	3.897	0.625	222.63
Penalties	0.120	0.014	−0.00925	1.799	−194.476	0.433	224.25
Offsides × Close game	0.115	0.005	0.06065	0.257	4.234	0.513	228.10
Errors	0.109	0.006	0.20787	0.961	4.624	0.679	232.45
Average age	0.100	0.009	0.17423	0.433	2.483	0.760	236.79
Home advantage × Knockout	0.097	0.003	−0.95604	7.001	−7.323	0.505	240.50
Offsides × Knockout	0.090	0.007	−0.13208	1.308	−9.900	0.678	241.79
Tackles attempted	0.087	0.003	0.04868	0.075	1.540	0.851	246.10
Through ballls × Knockout	0.072	0.015	0.09456	0.504	5.328	0.556	248.85
Passes into defensive third	0.070	0.002	0.00319	0.018	5.523	0.686	252.82
Headed passes	0.068	0.002	−0.01509	0.050	−3.345	0.618	253.67
Crosses × Knockout	0.067	0.001	−0.03140	0.162	−5.146	0.507	257.79
Shots	0.066	0.001	−0.04090	0.148	−3.621	0.606	262.23
Blocks	0.063	0.003	−0.04455	0.126	−2.829	0.714	266.78
Dribbles	0.059	0.004	−0.04131	0.068	−1.634	0.847	271.19
Freekicks	0.055	0.004	0.03358	0.199	5.931	0.655	275.15
Offsides	0.053	0.002	−0.22551	0.403	−1.788	0.887	279.43
Punches	0.052	0.001	0.21469	0.866	4.033	0.788	281.96
Variables with $\pi_k < 0.05$ not shown							

Table 14: Post-regression for final model in result-based approach for 2016.

	Estimate	Std. error	*z* value	Pr(|*z*|)
Intercept1	−2.2825	0.7990	−2.857	0.0043
Intercept2	−4.2924	0.8978	−4.781	0.0000
log(Market value ratio)	1.0507	0.2076	5.060	0.0000
Shots from fast-break	1.6253	0.6559	2.478	0.0132
Shot accuracy	0.0367	0.0149	2.461	0.0139
Distance covered difference	0.2839	0.0701	4.051	0.0000
Clearances	0.0777	0.0258	3.007	0.0026

Table 15: Results of the Brant-test for the proportional odds assumption.

Test for	EURO 2020				EURO 2016			
	υ_k	df	*p*-value	decision	υ_k	df	*p*-value	decision
Omnibus	12.29	10	0.266	H_0	1.41	5	0.923	H_0
log(Market value ratio)	0.69	1	0.405	H_0	0.1	1	0.756	H_0
Shot accuracy	0.17	1	0.681	H_0	0.78	1	0.378	H_0
Distance covered difference	0.12	1	0.734	H_0	0.01	1	0.938	H_0
Save rate	0.42	1	0.517	H_0	–	–	–	–
Dribbles	0.19	1	0.665	H_0	–	–	–	–
Tackles attempted	0.03	1	0.869	H_0	–	–	–	–
Crosses	0.11	1	0.742	H_0	–	–	–	–
Claims × Close game	1.07	1	0.301	H_0	–	–	–	–
Long passes	0.37	1	0.545	H_0	–	–	–	–
Clearances × Close game	4.31	1	0.038	H_1	–	–	–	–
Shots from fast-break	–	–	–	–	0.19	1	0.667	H_0
Clearances	–	–	–	–	0.26	1	0.613	H_0

Table 16: Average coefficients obtained for selected variables when using all 1,000 iterations.

Variable	EURO 2020		EURO 2016	
	GB	RB	GB	RB
log(Market value ratio)	0.74994	2.45441	0.41132	0.94221
Distance covered difference	0.14018	0.37990	0.09040	0.22601
Shot accuracy	0.03591	0.08396	0.01913	0.03406
Save rate	0.00441	0.01532	–	–
Tackles attempted	−0.02689	−0.06246	–	–
Dribbles	0.01485	0.06694	–	–
Clearances × Close game	0.02834	0.06444	–	–
Errors	−0.27747	–	–	–
Tackle rate	0.02087	–	–	–
Long passes	–	0.04052	–	–
Crosses	–	−0.08384	–	–
Claims × Close game	–	0.54736	–	–
Shots from fast-break	–	–	0.92485	2.06938
Clearances	–	–	0.01871	0.02754
log(Market value ratio) × Knockout	–	–	0.22195	–

References

Belsley, D., Kuh, E., and Welsch, R. (1980). *Regression diagnostics: identifying influential data and sources of collinearity*. Wiley Series in Probability and Statistics, Wiley.

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics* 46: 1171—1178.

Brito de Souza, D., López-Del Campo, R., Blanco-Pita, H., Resta, R., and Del Coso, J. (2019). An extensive comparative analysis of successful and unsuccessful football teams in LaLiga. *Front. Psychol.* 10: 2566.

Carmichael, F., Thomas, D., and Ward, R. (2000). Team performance: the case of English premiership football. *Manag. Decis. Econ.* 21: 31—45. .

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* 21: C1—C68.

Collet, C. (2013). The possession game? A comparative analysis of ball retention and team success in European and international football, 2007—2010. *J. Sports Sci.* 31: 123—136.

dos Reis, M.A.M., Vasconcellos, F., and de Almeida, M.B. (2017). Analysis of the effectiveness of long distance passes in 2014 Brazil FIFA world cup. *Braz. J. Kinanthropometry Hum. Perform.* 19: 676—685.

Görgen, K. and Schienle, M. (2019). *How have German university tuition fees affected enrollment rates: robust model selection and design-based inference in high-dimensions*. arXiv working paper, https://doi.org/10.48550/arXiv.1909.08299.

Hansen, B. (2022). *Econometrics*. Princeton University Press, Princeton, NJ.

Harrell, F.E. (2015). *Ordinal logistic regression*. Springer International Publishing, Cham, pp. 311—325.

Hastie, T., Tibshirani, R., and Friedman, J. (2013). *The Elements of statistical learning: data mining, inference, and prediction*. Springer New York, New York.

Lago-Peñas, C., Lago-Ballesteros, J., and Rey, E. (2011). Differences in performance indicators between winning and losing teams in the UEFA champions league. *J. Hum. Kinet.* 27: 135—146.

Lepschy, H., Wäsche, H., and Woll, A. (2020). Success factors in football: an analysis of the German bundesliga. *Int. J. Perform. Anal. Sport* 20: 150—164.

Lepschy, H., Woll, A., and Wäsche, H. (2021). Success factors in the FIFA 2018 world cup in Russia and FIFA 2014 world cup in Brazil. *Front. Psychol.* 12: 525.

Liu, H., Hopkins, W., Gómez, A.M., and Molinuevo, S.J. (2013). Inter-operator reliability of live football match statistics from OPTA Sportsdata. *Int. J. Perform. Anal. Sport* 13: 803—821.

Liu, H., Ruano, M., Lago-Peñas, C., and Sampaio, J. (2015). Match statistics related to winning in the group stage of 2014 Brazil FIFA world cup. *J. Sports Sci.* 33: 1205—1213.

McCullagh, P. (1980). Regression models for ordinal data. *J. Roy. Stat. Soc. B* 42: 109—142.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *J. Roy. Stat. Soc. B Stat. Methodol.* 72: 417—473.

Peñas, C., Lago Ballesteros, J., Dellal, A., and Gómez López, M. (2010). Game-related statistics that discriminated winning, drawing and losing teams from the Spanish soccer league. *J. Sports Sci. Med.* 9: 288—293.

Sarkar, S. (2018). Paradox of crosses in association football (soccer) — a game-theoretic explanation. *J. Quant. Anal. Sports* 14: 25—36.

Sarmento, H., Figueiredo, A., Peñas, C., Milanović, Z., Barbosa, A., Tadeu, P., and Bradley, P. (2017). The influence of tactical and situational variables on offensive sequences during elite football matches. *J. Strength Condit Res.* 32: 1.

Schauberger, G., Groll, A., and Tutz, G. (2018). Analysis of the importance of on-field covariates in the German Bundesliga. *J. Appl. Stat.* 45: 1561—1578.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6: 461—464.

Simon, N., Friedman, J.H., Hastie, T., and Tibshirani, R. (2011). Regularization paths for Cox's proportional Hazards model via coordinate descent. *J. Stat. Software* 39: 1—13.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. B* 58: 267—288.

Walker, S.H. and Duncan, D.B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika* 54: 167—179.

Wang, J., He, X., and Xu, G. (2020). Debiased inference on treatment effect in a high-dimensional model. *J. Am. Stat. Assoc.* 115: 442—454.

Wurm, M.J., Rathouz, P.J., and Hanlon, B.M. (2021). Regularized ordinal regression and the ordinalNet R package. *J. Stat. Softw.* 99: 1—42.