

Using (Copula) Regression and Machine Learning to Model and Predict Football Results in Major European Leagues

Hendrik van der Wurp, Andreas Groll¹

Department of Statistics, TU Dortmund University, Dortmund, Germany

Abstract *In this manuscript, we compare classical univariate regression approaches with copula models explicitly accounting for the dependency structure as well as with modern machine learning techniques in the context of modelling and predicting of football results in the major European leagues. Particularly, we want to present an extensive data set compiled from publicly available sources containing data and match results from the first men's football divisions from England, France, Germany, Italy, Spain (often referred to as the "big five"), the Netherlands and Turkey. We introduce several modelling approaches to predict upcoming matches and compare their predictive strengths. The gathered data set is presented in detail and made publicly available to motivate further work and modelling ideas.*

Keywords: *Count data regression, Football, Joint modelling, Regularisation, Application.*

1. INTRODUCTION

Generally, international football tournaments such as FIFA World Cups or the big confederation's championships (e.g. UEFA European Championship, CONCACAF Gold Cup, CONMEBOL Copa América) as well as international and national tournaments on the team-level are experiencing an ever increasing standing in terms of popularity and financial relevance. Also, modelling and predicting the results of sport matches and especially football matches has become a quite popular and present topic.

Even though no gold standard approach exists to model football results, a vast selection of methods and model classes has been proposed over the years. On the observed results of scored goals per team, Poisson regression approaches have been commonly used (e.g. by Lee, 1997, or Maher, 1982). These have been extended over the years to include several team-specific covariates in combination with regularisation techniques (e.g. by Groll and Abedieh, 2013 or Groll et al., 2015). The basic Poisson approaches can be extended by including dependency

¹Corresponding author: Hendrik van der Wurp, email: vanderwurp@statistik.tu-dortmund.de

between the numbers of goals scored by competing teams, which Dixon and Coles (1997) investigated early. In particular, the bivariate Poisson approach was then proposed in detail by Karlis and Ntzoufras (2003). A different approach to dependency is the inclusion of copulas, which McHale and Scarf (2007) used to model the number of shots-on-target. Nikoloulopoulos and Karlis (2010) promoted copulas for the application to count data in general. More recently, van der Wurp et al. (2020) and van der Wurp and Groll (2021) extensively applied copulas within the GJRM (Generalised Joint Regression Modelling) framework by Marra and Radice (2019), and added football-specific regularisation into it.

A completely different approach is to dispense with the information of the numbers of goals and to model the nominal/ordinal outcome (win first team, tie, win second team) directly. The usage of ordinal or nominal regression approaches is rather straightforward as well (and e.g. discussed in Hvattum, 2017). Leitner et al. (2010) used national team abilities (depicted by Elo ratings) and bookmakers' odds to directly obtain winning probabilities in a binary (win / loss) setting. This was extended by Tutz and Schauburger (2014) with penalisation approaches for league football data and by Schauburger et al. (2017) to analyse on-field variables such as total running distance per team. A comparison of both score- and result-based approaches has been performed by Egidi and Torelli (2021).

Besides regression approaches, random forests (originally introduced by Breiman, 2001) are a very flexible and frequently used technique in the context of predicting sports results. Random forests were used e.g. by Groll et al. (2019) and Groll et al. (2021) to model FIFA World Cup and European championship data, respectively, and to predict the latest tournament. Also with the tree-based methods, principally both score- and result-based models can be used, see, e.g., Schauburger and Groll (2018).

Bayesian approaches (see, for example, Baio and Blangiardo, 2010) are also promising, but are omitted in this work. It will examine the predictive performance of the mentioned (and some other) approaches via suitable performance measures and will also investigate potential betting results. The probabilities gathered from several online bookmakers will be used as a natural benchmark. While copula regression and the proposed football-specific penalty structures by van der Wurp et al. (2020) and van der Wurp and Groll (2021) will receive special attention, a lot of different modelling approaches and covariate settings will be benchmarked against one another.

The underlying data set was gathered in July 2021 and contains all matches from the respective first men's divisions of England, France, Germany, Italy, Spain








(the “big five”), the Netherlands, and Turkey for ten seasons between 2010 and 2020. Our data set ends just before the start of the COVID-19 pandemic, as these extraordinary circumstances are deemed to be a research topic completely on its own (postponed or completely canceled games, games with less or no fans, etc.). A growing-window approach will be used to assess the approaches’ predictive potential, where the upcoming matchday is predicted using all prior matchdays and seasons.

We present this data set in detail in Section 2 with information about available covariates. Section 3 contains brief descriptions of all used model classes, covariate settings, underlying software packages, and provides an overview about the performance indicators used in our application. The corresponding results are presented and visualised in Section 4, before we conclude in Section 5.

2. DATA

The data set was freely available, gathered from different websites, and published (van der Wurp, 2022). As the analysis of market values by `transfermarkt.com` was started in 2010, we chose the season of 2010/2011 as a starting point and ended in the season of 2019/2020 with the start of the COVID-19 pandemic (see end of Section 2). The sample sizes and more information by country are given in Table 1.

Table 1: Sample sizes per league. The season 2019/2020 was called off for the Ligue 1 and the Eredivisie, while postponed and later completed in the other leagues.

League	matches	league size	teams competed	$\overline{\text{goals}}_{\text{home}}$	$\overline{\text{goals}}_{\text{away}}$
 Premier League	3800	20	36	1.55	1.19
 Ligue 1	3700 (3800)	20	34	1.46	1.07
 Bundesliga	3060	18	28	1.65	1.30
 Serie A	3800	20	34	1.52	1.19
 Primera División	3800	20	33	1.59	1.13
 Eredivisie	2988 (3060)	18	26	1.80	1.34
 Süper Lig	3060	18	34	1.54	1.20

The main information of each match (teams competing, date, day of week, matchday number, and the scored goals is easily available data and was gathered from `kicker.de` in July 2021). Other covariates are:

- **Elo** rating of each team. Calculated and gathered from <http://clubelo.com/> (July 2021; Schiefler, 2015). It ranges from 1223 (FC Dordrecht in 2014) to 2106 (Barcelona in 2012) and can be interpreted via the differences in rating, denoted by $d = \text{Elo}_{\text{home}} - \text{Elo}_{\text{away}}$. The probability for the home team to win is then defined as $\pi = P(\text{HomeWin}) = 1 / \left(10^{\left(\frac{d}{400}\right)} + 1 \right)$ with ties being counted as a half win (Schiefler, 2015). Equal Elo ratings will lead to a probability of 0.5.

After each match, the team's Elo scores are adjusted by $\Delta\text{Elo} = (R - \pi) \cdot 20$ with R corresponding to the results from each team's point of view (1 for a win, 0.5 for a tie and 0 for a loss). The factor of 20 is a weight index chosen by Schiefler (2015). With this scheme, unlikely results like an underdog's win will result in bigger Elo changes.

These (or similar) types of Elo rankings are commonly used in competitive sports. It was originally proposed by Arpad Emmerich Elo (1961) to rank the ability of chess players.

- **Market Value (MV)** of a team. Determined and gathered from transfermarkt.com (July 2021). Given in million euro and ranges from 2.8 (FC Dordrecht in 2014) to 1,300 (Manchester City in 2019/20). The market values of transfermarkt.com are a community project, where each player's market value is discussed and determined by (known or rumoured) transfer fees and the player's standing in his team. The team's value is the simple sum of its current players. The values are updated twice a month to timely include transferred players. The earliest available data is from 2010-11-01, so missing values occur for the first matchdays of the season 2010/11. As the market values are growing over time, we are transforming the raw values to shares of the league's market value, using each matchday's sum as a total market value. Missing values are imputed as averages. With this approach, the dominance of single teams can be modelled over the years without a bias by inflation.
- **Bookmaker Odds** averaged from multiple bookmaker companies. Collected from oddsportal.com (July 2021) and averaged over six different bookmakers in 2010 up to 12 bookmakers in 2019. The odds can be transformed to probabilities by inverting them to $p_j = \frac{1}{\text{odds}_j}, j \in \{1, X, 2\}$. As

these do not sum up to 1 (due to bookmakers' margins²), we adjust these by $\tilde{p}_j = \frac{p_j}{p_1 + p_X + p_2}$ with p_1 and p_2 corresponding to wins of the first or second named team and p_X to a tie. With this, we implicitly assume an evenly distributed margin across these outcomes. An alternative, more complex normalisation approach, which is optimal against insider trading, was proposed by Shin (1991).

- **Promoted** status of a team. Indicates for each team, whether it has been promoted to the division immediately before the current season. This is used to include the “rookie status”.
- **Titleholder** from last season. Indicates for each team whether it is the league's current titleholder.
- **CupTitleholder** from last season. Indicates for each team whether it is the titleholder of the national cup (DFB-Pokal in Germany, FA CUP in England, Copa del Rey in Spain, Coppa Italia, Coupe de France, KNVB Cup in the Netherlands, Turkish Cup).
- **FormGoals3** is the number of goals scored by the corresponding team i in its last three matches. Easily calculated for matchdays 4 and later. For earlier matchdays the last seasons average of all teams \bar{g} is used.

- matchday 1: $\text{FormGoals3} = \bar{g}$
- matchday 2: $\text{FormGoals3} = \frac{1}{3}g_{\text{team } i, \text{ matchday1}} + \frac{2}{3}\bar{g}$
- matchday 3: $\text{FormGoals3} = \frac{1}{3}g_{\text{team } i, \text{ matchday1}} + \frac{1}{3}g_{\text{team } i, \text{ matchday2}} + \frac{1}{3}\bar{g}$

In rare cases, when a result is missing in the last 3 matches, the average of the remaining 2 matches is used. Instead of 3, the last 5 (or 7, 10, ...) matches could be used. We settled on 3 to capture the most recent form of the teams, which in football can often change quite spontaneously.

Note that, of course, principally many more potential covariates could be collected and added to the data, such as e.g. the teams' *average ages* or the coaches' *job tenure*, or even so-called *hybrid* variables that are derived themselves by statistical model as done in Groll et al. (2019) and Groll et al. (2021). However, we


²The bookmakers' margins can be seen as the fee the bookmakers take for offering their bets. As a simplified example, fair betting odds for a (fair) coin toss would be 2. The offered odds need to be lower than that, maybe 1.9, so the bookmaker is running profits in the long run. For more details, see also the Betting Results paragraph in Section 3.4

abstain to do so here, as we want to present rather standard approaches that can be applied more or less directly by interested practitioners. For this purpose, we have restricted the set of potential covariates to a selection which we deem to be both highly informative and quite directly available.

MISSING DATA AND ABNORMALITIES

As noted above, no market values were available before 2010-11-01. This affects 676 matches in total from all included leagues. The website transfermarkt.com also does not provide data for teams that were dissolved or left professional and semi-professional divisions. This results in missing market values in the following cases:

- Athletic Club Arlésien in the Ligue 1  was dissolved in 2016 and has missing market values in its only season of 2010/11.
- Thonon Évian F.C. in the Ligue 1  was relegated multiple times and left professional and semi-professional football, currently switching between France's 5th and 6th division. This leads to missing values in the four seasons of 2011/12, 2012/13, 2013/14, and 2014/15.
- ACN Siena 1904 in the Serie A  was dissolved in 2014 and has missing market values in the seasons of 2011/12 and 2012/13. Although the team was re-established multiple times, it was never able to reach the higher divisions.
- AC Cesena in the Serie A  was dissolved in 2018 and has missing market values in the seasons of 2010/11, 2011/12, and 2014/15.
- Kayseri Erciyesspor in the Süper Lig  was dissolved in 2018 and has missing market values in the seasons of 2013/14 and 2014/15.
- Orduspor in the Süper Lig  was dissolved in 2019 and has missing market values in the seasons of 2011/12 and 2012/13.
- Mersin İdman Yurdu in the Süper Lig  was dissolved in 2019 and has missing market values in the seasons of 2011/12, 2012/13, 2014/15, and 2015/16.
- Bucaspor in the Süper Lig  was dissolved in 2020 and has missing market values in its only season of 2010/11.

- Gaziantepspor in the Süper Lig  was dissolved in 2020 and has missing market values in the seven seasons between 2010/11 and 2016/17.








In total, 2236 market values are missing, of which 1352 correspond to matches before 2010-11-01 and 884 to the teams mentioned above (after 2010-11-01).

For the bookmakers' odds a total of 346 entries is missing, belonging to 118 matches. In total 1706 matches include missing data, of which 676 are from the start of the season 2010/11. The other 1030 matches are spread throughout the leagues and seasons. Apart from these missing values of single covariates, due to the COVID-19 pandemic full matchdays were missing or performed under different circumstances.

THE PANDEMIC

As noted before, we will omit games played during the COVID-19 pandemic. The dates on which each league was influenced is given in Table 2. As the leagues were handling the situation differently, e.g. in Ligue 1 the season was postponed and later cancelled while the Süper Lig had matches behind closed stadium doors and later postponed the season, we exclude all matches later than the given dates, which were those of the earliest decisions regarding each league. As single matches (e.g., matches in the Eredivisie in February) have been postponed due to different reasons and should have taken place later, those matches before that cut-off point are missing. The corresponding final sample sizes per league are found in Table 2 as well.

Table 2: Start dates of matches under the COVID-19 pandemic influence. Date corresponds to the first decision, not the final one.

League	decision	date	included matches	with missings
 Premier League	postponed	2020-03-13	3696	128
 Ligue 1	cancelled	2020-03-13	3687	325
 Bundesliga	postponed	2020-03-16	2966	116
 Serie A	postponed	2020-03-09	3668	296
 Primera División	postponed	2020-03-12	3674	119
 Eredivisie	cancelled	2020-03-12	2973	118
 Süper Lig	postponed	2020-03-12	2963	597

Given the ever changing situation and decisions, we exclude all matches starting from 2020-03-01. As the remaining missing data points are rather few

compared to the full data set, we will not use any methods for data imputation and instead omit matches whenever a variable is used that is missing.

3. MODELS AND EVALUATION MEASURES

For all models the general notation includes the number of goals scored per team (y_1, y_2) and a covariate or design matrix \mathbf{X} , respectively, containing for each match a set of k different covariates as a single row $\mathbf{x}_i = (1, x_1, \dots, x_k)^\top$. The first column with entries of 1 corresponds to an intercept, which is included depending on the model.

3.1. MODELLING THE NUMBER OF GOALS

All fitting procedures and evaluations were performed within R (R Core Team, 2020).

Most models will be used with two different model equation sets. First, each team's goals are modelled with the team's covariates, indicated by H and A for home and away teams, respectively, in the following pseudo model formulae:

$$\begin{aligned} y_H &\sim \text{elo}_H + \text{MV}_H + \tilde{p}_1 + \text{FormGoals3}_H + \text{Promoted}_H + \text{Title}_H + \text{CupTitle}_H, \\ y_A &\sim \text{elo}_A + \text{MV}_A + \tilde{p}_2 + \text{FormGoals3}_A + \text{Promoted}_A + \text{Title}_A + \text{CupTitle}_A. \end{aligned} \quad (1)$$

And for a second, more complex type of approaches, each team's goals are modelled by the covariates of both teams, including information about the opponents strength.

$$\begin{aligned} y_H &\sim \text{elo}_H + \text{elo}_A + \text{MV}_H + \text{MV}_A + \tilde{p}_1 + \tilde{p}_2 + \text{FormGoals3}_H + \text{FormGoals3}_A + \\ &\quad \text{Promoted}_H + \text{Promoted}_A + \text{Title}_H + \text{Title}_A + \text{CupTitle}_H + \text{CupTitle}_A \\ y_A &\sim \text{elo}_A + \text{elo}_H + \text{MV}_A + \text{MV}_H + \tilde{p}_2 + \tilde{p}_1 + \text{FormGoals3}_A + \text{FormGoals3}_H + \\ &\quad \text{Promoted}_A + \text{Promoted}_H + \text{Title}_A + \text{Title}_H + \text{CupTitle}_A + \text{CupTitle}_H \end{aligned} \quad (2)$$

POISSON REGRESSION

Poisson regression is typically performed via a generalised linear model (GLM) with an exponential link function and often used to model count data. The two margins are typically treated independently (conditional on the covariate information), so no dependency apart from the covariate level is included. For a general overview of these models, see, e.g., Groll and Schaubberger (2019).

REGULARISED POISSON REGRESSION

To achieve some form of sparsity, penalisation techniques such as the LASSO (Tibshirani, 1996) can be used. In this setting, the fitting procedure is able to shrink coefficients or to set them completely to zero. As is typical for LASSO, the penalty strength (commonly denoted as λ) is determined via a cross validation approach, which is e.g. implemented in the `cv.glmnet` function from the `glmnet` R package (Friedman et al., 2010). The LASSO penalisation was used in the context of football, e.g. by Groll and Abedieh (2013) and Groll et al. (2015).

COPULA REGRESSION

Copula regression applies dependency between (in this case) Poisson marginal regressions. The GJM framework and R implementation by Marra and Radice (2019) is used, which was proposed to the application of football in van der Wurp et al. (2020). Detailed insights into the methodology can be found there and in the references therein. As the authors found the F (Frank) and FGM (Farlie-Gumbel-Morgenstern) copulae to be good choices for the application of FIFA World Cups, we concentrate on these dependency structures.

REGULARISED COPULA REGRESSION

Moreover, van der Wurp et al. (2020) proposed a penalty to ensure equal coefficient estimates for the same covariates of both competing teams. Corresponding covariates in this case are e.g. elo_H and elo_A in Equations (1) or (2). The way a team's elo rating is influencing the goals scored by the team should be the same regardless of whether the team is first- or second-named, i.e. home or away team. It is important to note that for the models from Equation (2), elo_H in the first margin and elo_A in the second margin are **not** coinciding, but yielding the same interpretation in different margins. To clarify, they are not the same covariate, but are treated as identical in the penalisation scheme. The covariates' order in Equation (2) highlights this. However, it can be argued that their coefficients should coincide.

A second LASSO-type penalty proposed by van der Wurp and Groll (2021) introduces sparsity to the framework. We will use the two penalties both individually and combined to find the best approach. A fixed grid length of 100 is used for optimising the LASSO-penalty strength. Note that varying the construction of the grid (density or location) would yield slightly different results.

RANDOM FORESTS

Multiple implementations of random forests exist in R. Groll et al. (2019) found the `cforest` from the `party` package by Hothorn et al. (2006) to be the best for the application of FIFA World Cups. Also, in the UEFA European Championship 2020 the `cforest` again yielded very promising results (Groll et al., 2021). We will follow these findings and use this implementation as a representative for random forests. For the general methodology about random forests see Breiman (2001), and Breiman et al. (1984) for the idea of classification and regression trees (CARTs) behind random forests.

EXTREME GRADIENT BOOSTING

Instead of parallel ensemble methods like the random forest approach from above, one can also consider sequential ensembles such as *boosting*, a technique which stems from the machine learning community (Freund and Schapire, 1996) and was later adapted to estimate predictors for statistical models (Friedman, 2001; Friedman et al., 2000). Friedman (2001) introduced the idea of gradient tree boosting, with decision trees as learners. These are repeatedly fitted on the residuals of the previous fitting step and, hence, combined to a sequential ensemble. This technique was then further improved by Chen and Guestrin (2016) via introducing additional regularisation in the objective function. The regularisation terms make the single trees weak learners to avoid overfitting. In a certain boosting iteration, the next tree is additively incorporated into the ensemble after multiplication with a rather small learning rate, which makes the learners even weaker. The method is called *extreme gradient boosting* (XGBoost), and is known in the machine learning community for its high predictive power³ The R package `xgboost` by Chen et al. (2021) contains the implementation of the algorithm.

For a brief summary of the methodology and an exemplary application to football, see e.g. Groll et al. (2021). Finally, note that an important aspect is that XGBoost involves several tuning parameters, such as e.g. the learning rate, the optimal number of boosting steps and several penalty parameters. For this purpose, we specified simple, discrete parameter grids and used multivariate 10-fold cross validation to determine optimal values for three key tuning parameters (namely the learning rate `eta`, the convergence criterion for splits `gamma`, and the max number of boosting iterations `nrounds`) on the training data (prior to

³It lately has been very successful in several prestigious machine learning prediction competitions, such as those launched by Kaggle (<https://www.kaggle.com>).

2014/2015). This is performed for each league individually and on the full training data set. The tuned parameters are kept constant after this.

3.2. MODELLING THE ORDINAL/NOMINAL OUTCOME

Beside modelling the number of goals per match (y_1, y_2) one can also model the three-way outcomes directly, which could be seen as a natural alternative as we are using multiple quality-of-prediction measures on this dimension and betting on these outcomes is rather popular. Hence, we will also model the match results `winHome` (with $y_1 > y_2$), `draw` (with $y_1 = y_2$) and `winGuest` (with $y_1 < y_2$) and from now on will use the common short notation of bookmakers, i.e. $1/X/2$, for these three outcomes. As a draw is clearly positioned between the other two outcomes, ordinal approaches are deemed more suitable than nominal ones, as they can exploit this information. We use the `polr` function of the `MASS` package by Venables and Ripley (2002), which fits a cumulative proportional-odds logit model.

REMARKS

Model approaches from Equations (1) and (2) are used in comparison whenever possible. This includes (regularised) Poisson regression, all copula models, random forests and the XGBoost. The ordinal approach is modelling the one-dimensional outcome $1/X/2$, where all covariates from (one of the two parts from) Equation (2) are used.

For all models predicting independently both scores, the Skellam distribution as a difference between two Poisson distributed variables is used to calculate probabilities for the three-way outcomes. This affects Poisson regression, random forests and XGBoost.

3.3. PREDICTION APPROACH

To simulate a realistic prediction situation, we use all prior matches of a given league to predict the following matchday. For this, we declared the first 5 seasons from (2010 up to 2015) as “burn-in” training data. So, starting from the season of 2015/2016, this training data is used to predict the next matchday. Afterwards, the predicted matchday is added to the training data, continuing throughout all remaining matchdays and seasons.

For the global model, which does not differentiate between the leagues, we use the date instead of the matchday, as the latter is not consecutive anymore. Although this leads to smaller steps (dates vs. matchdays) and slightly changing

sizes of the test data in our prediction approach, we deem the differences to the league-specific approach to be negligible

The quality or goodness of the obtained predictions is observed on multiple levels and calculated with measures from the following Section 3.4.

3.4. GOODNESS OF PREDICTION MEASURES

This section will introduce measures of prediction quality. With these, we cover all interesting response levels, i.e. *goals*, *three-way outcomes*, and *betting results*. It should be noted that not all measures are applicable to all models. The ordinal model for example does not provide estimated goals, so no error measures on this level can be obtained.

RPS

The ranked probability score (RPS) observes the three-way outcomes. It takes the ordinal structure of *win*, *draw* and *loss* into account and is defined in this context as

$$\text{RPS}_i = \frac{1}{2} \sum_{r=1}^2 \left(\sum_{l=1}^r \hat{\pi}_{il} - \delta_{il} \right)^2$$

for each match i (see, e.g. Schauburger and Groll, 2018, for another application, and Gneiting and Raftery, 2007, for the original proposition). Here, $\hat{\pi}_{il}$ are the estimated probabilities for the respective three-way outcomes l and δ_{il} is the Kronecker's delta, containing the observed outcome. In general, the RPS is an error term on the probability-level and is to be minimised. Alternatively, the (multi-category extension of the) Brier score (Brier, 1950) could be used on the three-way outcomes. But as it does not account for the ordinal structure, we use the RPS instead.

MULTINOMIAL LIKELIHOOD

The multinomial likelihood (LH), which also operates on the probability-level, is defined as

$$\text{LH}_i = \hat{\pi}_{i1}^{\delta_{i1}} \hat{\pi}_{i2}^{\delta_{i2}} \hat{\pi}_{i3}^{\delta_{i3}},$$

which is essentially the predicted probability for the observed outcome (van der Wurp et al., 2020), and therefore is to be maximised.

CLASSIFICATION RATE

The classification rate (CR) is maybe the simplest measure. Out of the three-way outcome, we classify the outcome with the highest predicted probability as the estimated outcome. For a single game i , this can be written via

$$CR_i = \mathbb{1} \left(\delta_i = \arg \max_{l \in \{1,2,3\}} (\hat{\pi}_{il}) \right).$$

The global classification rate is then averaged over all matches and is to be maximised.

ERRORS IN GOALS

On the response-level of the goals scored, one can easily calculate the difference between the number of estimated and observed goals per team. For each match, we calculate the squared and absolute errors via

$$\begin{aligned} SE_i &= (\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2, \\ AE_i &= |\hat{y}_1 - y_1| + |\hat{y}_2 - y_2|. \end{aligned}$$

BETTING RESULTS

Last, as maybe the most popular benchmark measure, we will investigate each model's performance in regard to betting. For this, we use the bookmakers' odds from `oddsportal.com`. It is important to note that these odds are averaged over a selection of bookmakers, so the results are not necessarily the same using a single or even a selection of bookmakers.

To create a betting strategy, we calculate the expected return of a given bet via

$$E[\text{return}_i] = \hat{\pi}_{il} \cdot \text{odds}_{il} - 1. \quad (3)$$

As soon as the expected return is positive, one should take that bet (a threshold value of 0 marks a fair bet). Larger thresholds than zero may be chosen.

If multiple bets for a single match yield a positive expected return, we will simply take the one with the highest expected return, limiting us to a single bet per match. Other approaches, such as a variance-minimising strategy, taking the bet with (a positive expected return and) the highest probability of success are also possible.

We are using a stake of 1 fiscal unit for each bet, indicated by the -1 in the expected return (3). Other strategies are possible as well, e.g. the Kelly criterion (Kelly, 1956), which gives weights and therefore different stake sizes to each bet. The outcome in terms of gains is then calculated via

$$\text{gains}_i = \begin{cases} -1, & \text{if bet failed} \\ \text{odds}_{il} - 1, & \text{if bet was successfull} \end{cases}$$

and summed up over all matches of a given league. Making a profit (i.e. beating the bookmakers) is a very optimistic and challenging objective. Hence, achieving betting losses close to zero with rather simple models is already considered an achievement, especially considering bookmakers' costs and (presumably) taxes.

When transforming bookmakers odds to probabilities (see Section 2), the probabilities do not sum up to 1 because of margins. As bookmakers are offering smaller odds than a fair bet would be, the transformation yields higher probabilities. These sums average to 1.05. The downward outlier (see Figure 1) may be the result of the averaging process from `oddsportal.com` and is not further investigated. The distribution indicates the 5% winning margin (median) the bookmakers are collecting.

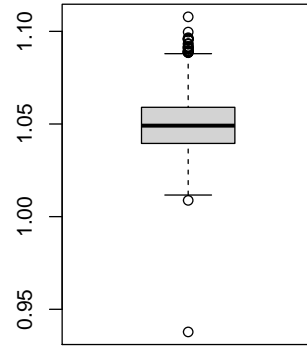









Figure 1: Values of inverted and summed up odds. For fair bets, this would always sum up to exactly 1. The difference can be interpreted as bookmaker margin.

4. RESULTS

For all models and leagues, the resulting measures are averaged throughout all predicted matches. Exemplarily, the results for the simple independent Poisson model from Equation (1) are shown in Table 3. The quality of prediction differs between the national leagues. This is especially visible in the betting results, ranging from a profit of 38.41 stakes (fictional money unit) in the Premier League to a loss of 189.51 stakes in the Süper Lig. Relative to the “invested” stakes this corresponds to a winning rate (i.e. $\text{bet.gains}/\text{bets}$) of 2.35% and a loss of 20.18%,

respectively. The ratio of matches that are bet on (i.e. bets/n) also varies strongly and ranges from 92.36% for the Premier League to 70.50% for the Süper Lig. This should be taken with a grain of salt, as the leagues receive widely different attention in the national and international media and betting markets.

Table 3: Results for the simple independent Poisson model from Equation (1).

	RPS	LH	CR	SE	AE	n	bets	bet.gains
 Premier League	0.191	0.434	0.552	2.601	1.804	1768	1633	38.41
 Bundesliga	0.203	0.418	0.520	2.972	1.921	1414	1189	-117.80
 Primera División	0.191	0.435	0.537	2.583	1.770	1746	1353	-113.76
 Ligue 1	0.199	0.411	0.514	2.523	1.746	1784	1358	-76.70
 Serie A	0.185	0.442	0.577	2.534	1.755	1744	1539	-89.17
 Eredivisie	0.189	0.446	0.583	2.961	1.912	1442	1061	-83.06
 Süper Lig	0.200	0.405	0.527	2.653	1.808	1332	939	-189.51

First, to be able to compare our big selection of models, we average the measures throughout all leagues. We are using a weighted average by sample sizes for the measures of RPS, LH, CR, SE, and AE and a simple sum for the number of matches n , the number of bets and the bet gains. The results for all models can be found in Table 4. Goodness-of-prediction results, exemplarily in terms of RPS and betting returns, for each league can be found in the appendix, Tables 9 and 10.

The RPS is, ever so slightly, improving with the copula models becoming more complex. Both the equal and the LASSO penalty are improving the results. Regarding the average multinomial likelihood the BIC models with lasso penalisation are performing worse than their AIC counterparts. We found no noteworthy differences between models using both marginal covariates in both marginal regressions and their simpler counterparts (see Equation (2) in Section 3.1 compared to Equation (1)). The classification rate CR has little to no variation in any direction. Sadly, no model was able to end with a net gain in betting from thousands of matches and bets. But some models are getting close to break-even. The simple copula models with all available covariates are achieving losses of less than 2.5% of stakes from more than 8600 bets. As discussed and shown above in Section 3.4, the calculated margin of bookmakers can be assumed to be about 5%, as they have expenses to cover. A selection of our models is solidly beating that threshold and might create frowning reactions with bookmaker companies. The equalisation penalty from van der Wurp et al. (2020) is impairing the models with

Table 4: Results for all modelling approaches. Calculated separated by leagues, but then aggregated. Cell colors best (green) to worst (red) for visualisation. See digital version.

Model	Eq	Cop.	regul.	RPS	LH	CR	SE	AE	bets	gainratio
pois	1	-	-	0.1938	0.428	0.545	2.674	1.811	9072	-0.0696
pois	2	-	-	0.1941	0.428	0.542	2.683	1.812	9506	-0.0430
pois	1	-	LASSO	0.1938	0.425	0.544	2.672	1.808	8915	-0.0693
pois	2	-	LASSO	0.1939	0.425	0.543	2.676	1.809	9352	-0.0524
RF	1	-	-	0.1975	0.427	0.536	2.753	1.840	10365	-0.0718
RF	2	-	-	0.1961	0.427	0.539	2.721	1.827	10188	-0.0617
XGboost	1	-	-	0.1967	0.412	0.543	2.732	1.818	10188	-0.0765
XGboost	2	-	-	0.1970	0.412	0.542	2.733	1.819	10312	-0.0609
Cop	1	F	-	0.1937	0.429	0.544	2.676	1.812	7874	-0.0549
Cop	1	FGM	-	0.1937	0.429	0.544	2.676	1.812	7936	-0.0573
Cop	2	F	-	0.1940	0.429	0.542	2.683	1.812	8696	-0.0250
Cop	2	FGM	-	0.1940	0.429	0.542	2.683	1.812	8753	-0.0243
Cop	1	F	equal	0.1937	0.429	0.543	2.674	1.808	7200	-0.0898
Cop	1	FGM	equal	0.1938	0.429	0.543	2.674	1.808	7276	-0.0897
Cop	2	F	equal	0.1938	0.429	0.542	2.675	1.810	8109	-0.0517
Cop	2	FGM	equal	0.1938	0.429	0.542	2.674	1.810	8164	-0.0497
Cop AIC	1	F	LASSO	0.1937	0.428	0.544	2.676	1.810	7578	-0.0670
Cop BIC	1	F	LASSO	0.1939	0.425	0.544	2.681	1.809	8016	-0.0800
Cop AIC	1	FGM	LASSO	0.1937	0.428	0.543	2.676	1.810	7669	-0.0733
Cop BIC	1	FGM	LASSO	0.1940	0.425	0.544	2.682	1.810	8035	-0.0921
Cop AIC	2	F	LASSO	0.1939	0.428	0.543	2.684	1.811	8150	-0.0363
Cop BIC	2	F	LASSO	0.1944	0.423	0.544	2.694	1.814	8315	-0.0715
Cop AIC	2	FGM	LASSO	0.1939	0.428	0.542	2.684	1.812	8259	-0.0408
Cop BIC	2	FGM	LASSO	0.1943	0.423	0.544	2.693	1.813	8456	-0.0717
Cop AIC	1	F	both	0.1937	0.429	0.543	2.679	1.808	6028	-0.0858
Cop BIC	1	F	both	0.1936	0.429	0.543	2.679	1.808	5729	-0.0807
Cop AIC	1	FGM	both	0.1935	0.429	0.543	2.670	1.806	5864	-0.0901
Cop BIC	1	FGM	both	0.1935	0.429	0.543	2.669	1.806	5560	-0.0960
Cop AIC	2	F	both	0.1938	0.428	0.543	2.673	1.808	6505	-0.1004
Cop BIC	2	F	both	0.1938	0.428	0.543	2.675	1.808	5816	-0.0963
Cop AIC	2	FGM	both	0.1936	0.429	0.543	2.670	1.807	6257	-0.0899
Cop BIC	2	FGM	both	0.1935	0.428	0.543	2.673	1.807	5729	-0.1037
ordinal	-	-	-	0.1944	0.430	0.542	-	-	9419	-0.0433

and without LASSO penalisation. The gain in interpretability (see van der Wurp and Groll, 2021 and the aforementioned reference from 2020) comes at a cost of prediction quality.

It should be noted that the mentioned measures are operating on the three-way-outcome dimension, while most model fitting procedures are using the likelihood on the number of goals. So errors on goals (SE and AE) might be a fairer measurement with regard to the models' original purpose apart from football modelling. With the exception of BIC models being constantly worse than their AIC counterparts, more sophisticated models in terms of penalisation are achieving better prediction performances. The combined models with equalisation and LASSO penalties are yielding the best results, albeit quite close to the LASSO-penalised Poisson model.

To summarise, it is not possible to declare a clear winning model. Depending on the context and the user's aims and scope, we deem multiple models to be suitable. For pure interpretability very simplistic models such as the ordinal or the simple Poisson model might be favoured. The equalisation approach allows for a better insight into coefficients, as they are cleaned of home- and away-team-specific differences in covariate effects. The best model – if the objective is to beat bookmakers – is, in this case, neither the most complex nor the simplest approach. In the following, we will present selected models in detail to highlight certain advantages and disadvantages.

The results by league are rather interesting, see Tables 9 and 10 in the appendix. Regarding the RPS our predictions for the French Ligue 1 and the Turkish Süper Lig are considerably worse than for the other leagues. The fictional betting returns show a similar pattern for the Süper Lig - matches in this league seem to be harder to predict than those of other leagues. Especially for the English Premier League and the German Bundesliga the models seem to perform quite well. As the investigated leagues receive quite different amounts of international attention, some difference in data quality can be assumed particularly for bookmaker odds and market values, the latter variable originating from a community project.

SELECTED MODELS IN DETAIL

We begin with examining the clear winner model regarding the betting outcome, which is the copula model with all available covariates and no penalisation whatsoever. As the differences between FGM and F copula are negligible we will show examples from both. Some resulting coefficients, exemplarily for the Premier League, can be found in Table 5.

Table 5: Estimated coefficients for the copula FGM model with all covariates and no penalisation, exemplarily for the Premier League

	$\beta^{(1)}$	$SE(\beta^{(1)})$	$\beta^{(2)}$	$SE(\beta^{(2)})$
(Intercept)	-0.9255	0.6616	-0.1517	0.7508
elo Team	-0.0001	0.0004	-0.0006	0.0004
elo Opponent	0.0004	0.0003	0.0007	0.0004
MV Team	0.3625	0.9023	1.2796	1.0520
MV Opponent	1.6485	1.0235	-2.0060	1.1524
p Team	1.6295	0.3602	1.4432	0.3845
p Opponent	-0.3531	0.3756	-0.4858	0.4138
FormGoals3 Team	-0.0205	0.0204	-0.0007	0.0230
FormGoals3 Opponent	-0.0081	0.0217	-0.0227	0.0249
Promoted Team	0.0443	0.0464	-0.1448	0.0549
Promoted Opponent	0.0516	0.0397	-0.0139	0.0453
Title Team	-0.0853	0.0598	-0.0347	0.0668
Title Opponent	-0.0477	0.0780	0.1858	0.0865
Title Cup Team	-0.0303	0.0602	-0.0657	0.0686
Title Cup Opponent	0.0185	0.0792	-0.0470	0.0969

These coefficients (and especially the differences between the two margins) are rather hard to interpret. While each respective team's market value has a positive influence on the team itself, the opponent's market value is behaving quite differently. For home teams, the market value of their opponents has a positive impact and for away teams, the respective market value of their opponents has a negative influence. Due to high levels of multicollinearity, think for example of elo, market value and bookmaker probabilities p , the exact values cannot be taken at face value. But rather big differences between the first and second margin are still hard to justify.

Models with higher value regarding interpretability may be desired, even if they offer a slightly worse performance in specific measures or even overall. The results for the separate leagues (see Table 8 in the appendix) are varying strongly between the leagues and each league's margins. This could be for two reasons: A) The covariates' influence is immensely different in each league and the leagues should therefore be fitted independently. We will discuss this in Section 4.1 in more detail. Or B) a lot of noise and artefacts are included in the models. Therefore, some form of sparsity should be incorporated. We will take a closer look at

other well-performing models from Table 4, i.e. applying the LASSO-type penalty and a second model using both the LASSO and the presented equalisation penalty.

SPARSER MODELS

The (LASSO-) penalised Model via BIC with an F copula from Table 4 has a slightly worse performance regarding betting and no noteworthy changes in the other measures. The resulting coefficients can be found in Table 6.

Table 6: Estimated coefficients for the LASSO-penalised copula F model (left) and with both penalties combined (right) with all covariates, exemplarily for the Premier League. For both models the optimal tuning parameters were selected via BIC.

	$\beta^{(1)}$	$\beta^{(2)}$	$\beta^{(1)}$	$\beta^{(2)}$
(Intercept)	-0.6960	-0.2231	-0.3593	-0.3637
elo Team	0.0005	0.0002		
elo Opponent	0.0001	-0.0002		
MV Team	0.0488	0.3249		
MV Opponent	0.6337	-0.7649		
p Team	0.8067	1.2049	1.6492	1.6498
p Opponent	-0.7372	-0.1131		
FormGoals3 Team		0.0019		
FormGoals3 Opponent		-0.0084		
Promoted Team	0.0190	-0.0535		
Promoted Opponent	0.0153			
Title Team	-0.0702			
Title Opponent		0.0389		
Title Cup Team				
Title Cup Opponent		-0.0081		

With eight coefficients shrunk to zero, the model is slightly sparser and easier to interpret, while maintaining virtually the same quality of prediction. Some oddness remains: Playing against the current titleholder has a positive impact on the away team, but no influence at all on the home team. The opponent's market value even changes its sign completely if a team is playing at home or away. This can be rationalised with strong interdependencies and collinearities or with missed features such as psychological factors and others.

To (partly) tackle this issue, we will take a look at the model with the combined penalties (BIC tuning and F copula again) in Table 4. The results can also

be found in Table 6. With only five coefficients estimated different from zero (including the copula parameter, which, interestingly, was estimated to be virtually zero), the resulting model is extremely sparse and easy to interpret. Here, $\beta^{(1)}$ and $\beta^{(2)}$ are virtually equal, allowing straightforward interpretations. The predicted probabilities by bookmakers p – which can be interpreted as a substitute variable for team strength – are estimated yielding a positive influence on each respective team. Note here that the intercept was only penalised by the equalisation penalty and not by the LASSO-type approach, as is common for the LASSO framework.

4.1. DIFFERENCES BETWEEN LEAGUES

In this section, we investigate whether the leagues are different regarding their assumed underlying model. Instead of comparing or testing the models' coefficients, we compare the quality of prediction in the ever updating models when differentiating between the national leagues and when treating them as one global training data set. Instead of predicting the next matchday (as done before), we are using the dates of matches. This results in 1793 unique dates of which the first 913 are solely used as training data and the other 880 are predicted using all matches before the given date. The results in comparison to Table 4 from before are shown in Table 11. Unsurprisingly, the results are not wildly different. Instead, the results seem to be more homogeneous than before. Especially, the betting results are clearly more consistent between models.

The estimated coefficients for a selected copula regression model can be found in Table 7. As interpretability is limited with wildly different marginal coefficients, the equalisation penalty is applied again and the resulting coefficients are compared. The resulting model contains four covariates for each margin. The bookmakers' p was consistently chosen in both settings. Interestingly, the estimated copula parameter θ was again estimated to be virtually zero in terms of Kendall's τ (0.0297 and < 0.0001 in absolute value, respectively for the models from Table 7), indicating no correlation structure whatsoever.

5. CONCLUSIONS

In this work, we presented an extensive data set of football matches in European leagues and the application of different modelling approaches to it. Comparing methodologies, we found regularised copula regression approaches to yield good results. The very flexible machine learning techniques of Random Forests and XGBoost are very sensible to tuning - their rather mediocre results in this

Table 7: Estimated coefficients for the LASSO-penalised copula F model (left) and with both penalties combined (right) with all covariates for all leagues combined in comparison to Table 6. For both models the optimal tuning parameters were selected via BIC.

	$\beta^{(1)}$	$\beta^{(2)}$	$\beta^{(1)}$	$\beta^{(2)}$
(Intercept)	-0.5769	-0.1050	-0.1143	-0.1104
elo Team	-0.0005	-0.0003	-0.0002	-0.0002
elo Opponent	0.0003	0.0000		
MV Team	0.7206	0.4752	0.7013	0.6972
MV Opponent	-0.3280			
p Team	2.2575	1.9893	1.7061	1.7082
p Opponent	0.6245	0.1176		
FormGoals3 Team	0.0204	0.0094		
FormGoals3 Opponent	0.0339	0.0060	0.0178	0.0169
Promoted Team	-0.0130	-0.0579		
Promoted Opponent	-0.0221	-0.0168		
Title Team		-0.0385		
Title Opponent	-0.0437	-0.1077		
Title Cup Team	0.0558	-0.0123		
Title Cup Opponent	0.0285	-0.0165		

application can almost certainly be improved via extensive tuning. The (copula-)regression approaches yield models that are both easy to interpret and to use. However, the gain compared to simple approaches such as standard independent Poisson modelling is rather small.

We found a set of covariates that are more important than others. Unsurprisingly, especially the bookmakers' probabilities (converted from odds) are deemed to be full of information and can be a solid predictor on their own. Differences between the investigated seven European leagues were found considering relevant covariates. The common ground was found to be the previously mentioned bookmakers' odds. The influence of other coefficients varies greatly in different countries in both strength and sign. As these can be interpreted as correction factors onto the immense importance of bookmakers' odds, the variation can be caused by the leagues themselves or different prediction strategies by the bookmakers.

Principally, one reason for all regarded modelling approaches yielding rather similar results could be that they all base on the highly informative bookmakers' odds, as described above. Hence, the specific type of modelling (linear vs. non-

linear, interactions, dependence structure, etc.) here seems to play a minor role. We believe that extending the regarded set of covariates by additional features which cover new types of information, such as e.g. the “hybrid” features regarded in Groll et al. (2021, 2019) for the modelling of national team tournaments could on the one hand side increase the overall predictive performance of the models, on the other hand manifest more distinctive results across model classes. Unfortunately, the calculation of these features is rather extensive and went beyond the scope of this work. Besides, as mentioned above, the machine learning approaches are subject to complex tuning. Hence, they typically need a large training data set to utilize their full potential.

The data indicates that bookmakers are calculating with a betting margin of about 5%. While some models were able to beat this margin, we can not claim to have beaten the bookmakers, as other models ran significant losses. There are obvious limitations due to the available data. Our data set was completely compiled from publicly available sources and from a fixed point in time. Bookmakers are able to shift existing odds depending on betting behaviour of customers or depending on external events, such as a core player getting injured before a match. A public-data driven approach such as this cannot be that flexible.

While this work is focussed around national leagues, all models can principally be applied to different tournaments as well, such as FIFA World Cups, UEFA European Championships, or the UEFA Champions League and comparable tournaments on the club-level on other continents. However, some additional aspects need to be considered. For one the existing sample sizes are considerably smaller, causing issues for complex machine learning approaches. Also each tournament’s specific structure (how groups are built in group stages, tournament schedule, potential extra time and penalty shoot-outs etc.) needs to be taken into account. See, for example, thoughts by Egidi and Torelli (2021), van der Wurp et al. (2020), and van der Wurp and Groll (2021).

All in all, our aim was mainly to create an interesting data set and motivate different statistical and machine learning modelling approaches to it, rather than finding the actual/virtual/definite best prediction approach on the regarded data, e.g. in terms of betting profits. The manuscript shall give an overview of their general predictive potential in this field of application as well as other aspects such as interpretability, which might also be relevant for the practitioner. The underlying data set is publicly available in an R package `EUfootball` (van der Wurp, 2022). The reader is both invited to create their own modelling ideas for the underlying football data and to apply the here presented approaches to other

fields and applications. Also, we hope that this work inspires other researches to use and extend our data set, and to build upon and further improve the modelling strategies presented here.

References

- Baio, G. and Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football results. In *Journal of Applied Statistics*, 37 (2): 253–264.
- Breiman, L. (2001). Random forests. In *Machine Learning*, 45: 5–32.
- Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, J.C. (1984). *Classification and Regression Trees*. Wadsworth, Monterey, CA.
- Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. In *Monthly Weather Review*, 78: 1–3.
- Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.
- Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., and Li, Y. (2021). *xgboost: Extreme Gradient Boosting*. URL <https://CRAN.R-project.org/package=xgboost>. R package version 1.3.2.1.
- Dixon, M.J. and Coles, S.G. (1997). Modelling association football scores and inefficiencies in the football betting market. In *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46 (2): 265–280.
- Egidi, L. and Torelli, N. (2021). Comparing goal-based and result-based approaches in modelling football outcomes. In *Social Indicators Research*, 156 (2): 801–813.
- Elo, A.E. (1961). New uscf rating system. In *Chess Life*, 16: 160–161.
- Freund, Y. and Schapire, R.E. (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning*, 148–156. Morgan Kaufmann, San Francisco, CA.
- Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. In *Annals of Statistics*, 29: 337–407.









- Friedman, J.H., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. In *Annals of Statistics*, 28: 337–407.
- Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. In *Journal of Statistical Software*, 33 (1): 1.
- Gneiting, T. and Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. In *Journal of the American Statistical Association*, 102 (477): 359–378.
- Groll, A. and Abedieh, J. (2013). Spain retains its title and sets a new record - generalized linear mixed models on European football championships. In *Journal of Quantitative Analysis in Sports*, 9 (1): 51–66.
- Groll, A., Hvattum, L.M., Ley, C., Popp, F., Schauburger, G., Van Eetvelde, H., and Zeileis, A. (2021). Hybrid machine learning forecasts for the uefa euro 2020. In *arXiv preprint arXiv:2106.05799*.
- Groll, A., Ley, C., Schauburger, G., and Van Eetvelde, H. (2019). A hybrid random forest to predict soccer matches in international tournaments. In *Journal of Quantitative Analysis in Sports*, 15: 271–287.
- Groll, A. and Schauburger, G. (2019). Prediction of soccer matches. In *Wiley StatsRef: Statistics Reference Online*, 1–7.
- Groll, A., Schauburger, G., and Tutz, G. (2015). Prediction of major international soccer tournaments based on team-specific regularized Poisson regression: an application to the FIFA World Cup 2014. In *Journal of Quantitative Analysis in Sports*, 11 (2): 97–115.
- Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., and van der Laan, M.J. (2006). Survival ensembles. In *Biostatistics*, 7: 355–373.
- Hvattum, L.M. (2017). Ordinal versus nominal regression models and the problem of correctly predicting draws in soccer. In *International Journal of Computer Science in Sport*, 16 (1): 50–64.
- Karlis, D. and Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. In *The Statistician*, 52: 381–393.

- Kelly, J.L. (1956). A new interpretation of information rate. In *Bell System Technical Journal*, 35 (4): 917–926. doi:10.1002/j.1538-7305.1956.tb03809.x. URL <http://dx.doi.org/10.1002/j.1538-7305.1956.tb03809.x>.
- Lee, A.J. (1997). Modeling scores in the Premier League: is Manchester United really the best? In *Chance*, 10: 15–19.
- Leitner, C., Zeileis, A., and Hornik, K. (2010). Forecasting sports tournaments by ratings of (prob)abilities: A comparison for the EURO 2008. In *International Journal of Forecasting*, 26 (3): 471–481.
- Maher, M.J. (1982). Modelling association football scores. In *Statistica Neerlandica*, 36: 109–118.
- Marra, G. and Radice, R. (2019). *GJRM: generalised joint regression modelling*. R package version 0.2.
- McHale, I. and Scarf, P. (2007). Modelling soccer matches using bivariate discrete distributions with general dependence structure. In *Statistica Neerlandica*, 61 (4): 432–445. doi:10.1111/j.1467-9574.2007.00368.x. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9574.2007.00368.x>.
- Nikoloulopoulos, A.K. and Karlis, D. (2010). Regression in a copula model for bivariate count data. In *Journal of Applied Statistics*, 37: 1555–1568.
- R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Schauberger, G. and Groll, A. (2018). Predicting matches in international football tournaments with random forests. In *Statistical Modelling*, 18 (5–6): 1–23.
- Schauberger, G., Groll, A., and Tutz, G. (2017). Analysis of the importance of on-field covariates in the German Bundesliga. In *Journal of Applied Statistics*, 45 (9): 1561–1578.
- Schiefler, L. (2015). *Football Club Elo Ratings*. <http://clubelo.com/> [Accessed: July 2021].
- Shin, H.S. (1991). Optimal betting odds against insider traders. In *The Economic Journal*, 101 (408): 1179–1185.

- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. In *Journal of the Royal Statistical Society*, B 58: 267–288.
- Tutz, G. and Schauburger, G. (2014). Extended ordered paired comparison models with application to football data from german bundesliga. In *AStA Advances in Statistical Analysis*, 99 (2): 209–227. doi:10.1007/s10182-014-0237-1. URL <http://dx.doi.org/10.1007/s10182-014-0237-1>.
- van der Wurp, H. (2022). *EUfootball: Football Match Data of European Leagues*. URL <https://CRAN.R-project.org/package=EUfootball>. R package version 0.0.1.
- van der Wurp, H. and Groll, A. (2021). Introducing lasso-type penalisation to generalised joint regression modelling for count data. In *AStA Advances in Statistical Analysis*. URL <https://doi.org/10.1007/s10182-021-00425-5>.
- van der Wurp, H., Groll, A., Kneib, T., Marra, G., and Radice, R. (2020). Generalised joint regression for count data: a penalty extension for competitive settings. In *Statistics and Computing*, 30 (5): 1419–1432.
- Venables, W.N. and Ripley, B.D. (2002). *Modern Applied Statistics with S*. Springer, New York, 4th edn.

APPENDIX

Table 8: Estimated coefficient for the copula FGM model with all covariates and no penalisation for all leagues; left columns: home team; right columns: away team

								
(Intercept)	-0.99	0.25	-0.48	-0.49	-1.73	-1.24	-0.15	-0.18
elo Team	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
elo Opponent	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.00
MV Team	0.43	1.52	0.81	-0.18	-0.26	-0.39	0.63	1.10
MV Opponent	1.70	-1.73	-1.07	-0.38	-1.56	0.87	0.19	-0.46
p Team	1.55	1.49	1.87	2.27	1.98	2.12	2.67	1.65
p Opponent	-0.38	-0.42	0.59	0.27	-0.24	-0.08	0.79	0.07
FormGoals3 Team	-0.02	0.01	0.03	0.01	0.01	0.00	0.03	-0.01
FormGoals3 Opponent	-0.01	-0.02	0.02	-0.03	0.04	0.02	0.04	-0.00
Promoted Team	0.04	-0.17	-0.06	-0.11	0.05	-0.01	-0.01	0.04
Promoted Opponent	0.06	-0.03	-0.06	-0.00	0.06	0.11	-0.04	-0.01
Title Team	-0.07	-0.03	-0.06	0.01	-0.01	-0.12	0.05	-0.07
Title Opponent	-0.05	0.17	-0.25	-0.16	-0.18	-0.53	-0.02	0.10
Title Cup Team	-0.01	-0.06	0.10	0.03	0.12	-0.05	0.05	0.13
Title Cup Opponent	0.01	-0.07	0.18	-0.04	0.12	0.09	0.08	0.07







						
(Intercept)	-1.57	-1.12	-1.37	1.31	-0.59	-0.99
elo Team	-0.00	0.00	-0.00	-0.00	-0.00	-0.00
elo Opponent	0.00	-0.00	0.00	-0.00	0.00	0.00
MV Team	-0.38	0.48	0.50	0.81	1.25	-0.33
MV Opponent	-0.84	-0.19	-1.57	0.33	-1.17	0.35
p Team	2.99	2.60	2.28	0.22	2.45	3.37
p Opponent	1.43	0.89	0.97	-1.40	1.11	1.33
FormGoals3 Team	0.03	0.02	-0.01	-0.00	0.03	0.00
FormGoals3 Opponent	0.04	-0.04	0.05	0.03	-0.01	0.02
Promoted Team	-0.01	-0.01	-0.03	-0.17	-0.01	-0.02
Promoted Opponent	-0.00	-0.02	0.01	-0.04	-0.09	-0.06
Title Team	0.11	-0.05	-0.10	0.06	0.08	-0.01
Title Opponent	0.15	-0.12	-0.06	-0.27	0.07	-0.20
Title Cup Team	0.09	-0.12	0.07	-0.04	-0.02	-0.01
Title Cup Opponent	-0.16	0.08	0.02	-0.14	0.05	-0.06

Table 9: RPS (ranked probability score) results for all models and leagues.
Cell colors best (green) to worst (red) for visualisation. See digital version.








Model	Eq	Cop	regul.							
pois	1	-	-	0.191	0.191	0.185	0.191	0.199	0.189	0.200
pois	2	-	-	0.192	0.192	0.185	0.191	0.199	0.189	0.202
pois	1	-	LASSO	0.192	0.192	0.186	0.191	0.199	0.189	0.201
pois	2	-	LASSO	0.192	0.192	0.185	0.191	0.199	0.189	0.201
RF	1	-	-	0.195	0.195	0.189	0.194	0.202	0.192	0.207
RF	2	-	-	0.194	0.194	0.188	0.192	0.202	0.191	0.203
XGboost	1	-	-	0.196	0.196	0.189	0.194	0.202	0.191	0.203
XGboost	2	-	-	0.196	0.196	0.190	0.194	0.202	0.190	0.204
Cop	1	F	-	0.191	0.191	0.185	0.191	0.199	0.189	0.200
Cop	1	FGM	-	0.191	0.191	0.185	0.191	0.199	0.189	0.201
Cop	2	F	-	0.192	0.192	0.185	0.191	0.199	0.190	0.202
Cop	2	FGM	-	0.192	0.192	0.185	0.191	0.199	0.190	0.202
Cop	1	F	equal	0.191	0.191	0.185	0.191	0.199	0.189	0.200
Cop	1	FGM	equal	0.191	0.191	0.185	0.191	0.199	0.189	0.200
Cop	2	F	equal	0.192	0.192	0.184	0.191	0.199	0.189	0.201
Cop	2	FGM	equal	0.192	0.192	0.184	0.191	0.199	0.189	0.201
Cop AIC	1	F	LASSO	0.191	0.191	0.185	0.191	0.199	0.189	0.200
Cop BIC	1	F	LASSO	0.191	0.191	0.185	0.192	0.199	0.190	0.200
Cop AIC	1	FGM	LASSO	0.191	0.191	0.185	0.191	0.199	0.189	0.200
Cop BIC	1	FGM	LASSO	0.192	0.192	0.185	0.192	0.199	0.190	0.200
Cop AIC	2	F	LASSO	0.191	0.191	0.185	0.191	0.199	0.190	0.201
Cop BIC	2	F	LASSO	0.192	0.192	0.187	0.192	0.199	0.190	0.201
Cop AIC	2	FGM	LASSO	0.192	0.192	0.185	0.191	0.199	0.190	0.201
Cop BIC	2	FGM	LASSO	0.192	0.192	0.187	0.192	0.199	0.190	0.201
Cop AIC	1	F	both	0.191	0.191	0.185	0.192	0.199	0.189	0.200
Cop BIC	1	F	both	0.191	0.191	0.185	0.192	0.199	0.189	0.200
Cop AIC	1	FGM	both	0.191	0.191	0.185	0.191	0.199	0.189	0.200
Cop BIC	1	FGM	both	0.191	0.191	0.185	0.191	0.198	0.189	0.200
Cop AIC	2	F	both	0.191	0.191	0.185	0.192	0.199	0.190	0.200
Cop BIC	2	F	both	0.192	0.192	0.185	0.192	0.199	0.190	0.200
Cop AIC	2	FGM	both	0.191	0.191	0.185	0.191	0.199	0.190	0.200
Cop BIC	2	FGM	both	0.191	0.191	0.185	0.191	0.199	0.189	0.200
ordinal		-	-	0.192	0.192	0.185	0.192	0.200	0.190	0.201

Table 10: Fictional betting results (in gain ratio, gains per betted unit of currency) for all models and leagues. Cell colors best (green) to worst (red) for visualisation. See digital version.








Model	Eq	Cop	regul.							
pois	1	-	-	0.024	0.024	-0.058	-0.084	-0.056	-0.078	-0.202
pois	2	-	-	0.017	0.017	0.003	-0.032	-0.077	-0.052	-0.138
pois	1	-	LASSO	0.017	0.017	-0.097	-0.088	-0.075	-0.135	-0.093
pois	2	-	LASSO	0.015	0.015	-0.019	-0.067	-0.048	-0.105	-0.130
RF	1	-	-	-0.004	-0.004	-0.046	-0.087	-0.046	-0.111	-0.190
RF	2	-	-	-0.039	-0.039	-0.116	0.001	-0.069	-0.077	-0.097
XGboost	1	-	-	-0.023	-0.023	-0.136	-0.067	-0.062	-0.102	-0.125
XGboost	2	-	-	0.001	0.001	-0.104	-0.067	-0.050	-0.092	-0.092
Cop	1	F	-	0.068	0.068	-0.049	-0.073	-0.043	-0.078	-0.215
Cop	1	FGM	-	0.054	0.054	-0.034	-0.070	-0.044	-0.081	-0.236
Cop	2	F	-	0.035	0.035	-0.001	0.027	-0.037	-0.062	-0.134
Cop	2	FGM	-	0.037	0.037	0.012	0.012	-0.040	-0.062	-0.127
Cop	1	F	equal	-0.016	-0.016	-0.083	-0.098	-0.086	-0.141	-0.212
Cop	1	FGM	equal	-0.025	-0.025	-0.082	-0.102	-0.077	-0.142	-0.198
Cop	2	F	equal	-0.003	-0.003	0.009	-0.071	-0.060	-0.072	-0.147
Cop	2	FGM	equal	-0.007	-0.007	0.023	-0.068	-0.064	-0.070	-0.150
Cop AIC	1	F	LASSO	0.037	0.037	-0.080	-0.068	-0.066	-0.113	-0.168
Cop BIC	1	F	LASSO	-0.007	-0.007	-0.059	-0.140	-0.045	-0.119	-0.153
Cop AIC	1	FGM	LASSO	0.017	0.017	-0.091	-0.088	-0.055	-0.111	-0.168
Cop BIC	1	FGM	LASSO	-0.006	-0.006	-0.099	-0.143	-0.097	-0.125	-0.141
Cop AIC	2	F	LASSO	0.035	0.035	0.047	-0.014	-0.091	-0.065	-0.197
Cop BIC	2	F	LASSO	-0.021	-0.021	-0.121	-0.070	-0.039	-0.115	-0.112
Cop AIC	2	FGM	LASSO	0.042	0.042	0.019	-0.025	-0.078	-0.061	-0.208
Cop BIC	2	FGM	LASSO	-0.024	-0.024	-0.136	-0.070	-0.047	-0.100	-0.123
Cop AIC	1	F	both	0.022	0.022	-0.045	-0.086	-0.161	-0.167	-0.340
Cop BIC	1	F	both	0.032	0.032	-0.045	-0.100	-0.125	-0.180	-0.337
Cop AIC	1	FGM	both	0.024	0.024	-0.070	-0.119	-0.146	-0.156	-0.357
Cop BIC	1	FGM	both	0.018	0.018	-0.070	-0.120	-0.191	-0.152	-0.357
Cop AIC	2	F	both	-0.022	-0.022	0.005	-0.154	-0.178	-0.155	-0.242
Cop BIC	2	F	both	0.025	0.025	-0.013	-0.149	-0.122	-0.233	-0.293
Cop AIC	2	FGM	both	-0.003	-0.003	-0.031	-0.136	-0.135	-0.134	-0.310
Cop BIC	2	FGM	both	0.011	0.011	-0.103	-0.125	-0.235	-0.135	-0.349
ordinal	-	-	-	0.024	0.024	0.006	-0.050	-0.083	-0.017	-0.152

Table 11: Results for all modelling approaches. Calculated by combining all leagues to one large data set. Cell colors best (green) to worst (red) for visualisation. See digital version.

Model	Eq	Copula	regul.	RPS	LH	CR	SE	AE	bets	gainratio
pois	1	-	-	0.1935	0.429	0.544	2.668	1.810	8316	-0.0734
pois	2	-	-	0.1936	0.428	0.544	2.666	1.808	9017	-0.0468
pois	1	-	LASSO	0.1935	0.428	0.544	2.667	1.809	8207	-0.0790
pois	2	-	LASSO	0.1935	0.428	0.544	2.664	1.808	8833	-0.0415
RF	1	-	-	0.1969	0.427	0.536	2.736	1.835	10318	-0.0612
RF	2	-	-	0.1953	0.427	0.539	2.694	1.822	9982	-0.0503
XGboost	1	-	-	0.1943	0.424	0.544	2.680	1.812	9691	-0.0690
XGboost	2	-	-	0.1944	0.424	0.543	2.674	1.811	9855	-0.0729
Cop	1	F	-	0.1933	0.430	0.545	2.668	1.810	6456	-0.0659
Cop	1	FGM	-	0.1932	0.430	0.545	2.668	1.810	6509	-0.0663
Cop	2	F	-	0.1934	0.429	0.544	2.666	1.808	7622	-0.0432
Cop	2	FGM	-	0.1934	0.429	0.544	2.666	1.808	7675	-0.0456
Cop	1	F	equal	0.1935	0.429	0.543	2.670	1.808	6004	-0.0639
Cop	1	FGM	equal	0.1935	0.429	0.543	2.670	1.808	6087	-0.0635
Cop	2	F	equal	0.1935	0.429	0.544	2.668	1.808	7551	-0.0565
Cop	2	FGM	equal	0.1935	0.429	0.544	2.668	1.808	7605	-0.0513
Cop AIC	1	F	LASSO	0.1933	0.430	0.545	2.668	1.810	6315	-0.0675
Cop BIC	1	F	LASSO	0.1934	0.429	0.545	2.669	1.809	6286	-0.0677
Cop AIC	1	FGM	LASSO	0.1933	0.430	0.545	2.668	1.810	6385	-0.0688
Cop BIC	1	FGM	LASSO	0.1934	0.429	0.545	2.669	1.809	6374	-0.0683
Cop AIC	2	F	LASSO	0.1935	0.429	0.544	2.667	1.808	7527	-0.0490
Cop BIC	2	F	LASSO	0.1934	0.429	0.544	2.669	1.810	7461	-0.0510
Cop AIC	2	FGM	LASSO	0.1935	0.429	0.544	2.665	1.807	7565	-0.0521
Cop BIC	2	FGM	LASSO	0.1935	0.429	0.544	2.667	1.809	7551	-0.0623
Cop AIC	1	F	both	0.1939	0.429	0.543	2.693	1.812	6255	-0.0687
Cop BIC	1	F	both	0.1939	0.429	0.543	2.692	1.812	6050	-0.0651
Cop AIC	1	FGM	both	0.1934	0.429	0.544	2.670	1.807	5773	-0.0660
Cop BIC	1	FGM	both	0.1934	0.429	0.544	2.670	1.807	5565	-0.0617
Cop AIC	2	F	both	0.1948	0.428	0.543	2.721	1.815	7103	-0.0859
Cop BIC	2	F	both	0.1947	0.428	0.544	2.722	1.816	5775	-0.0774
Cop AIC	2	FGM	both	0.1935	0.429	0.544	2.668	1.806	6786	-0.0930
Cop BIC	2	FGM	both	0.1934	0.429	0.545	2.670	1.806	5321	-0.0626
ordinal	-	-	-	0.1936	0.430	0.544	-	-	7937	-0.0506