Universität St.Gallen

Predicting Match Outcomes in Football by an Ordered Forest Estimator

Daniel Goller, Michael C. Knaus, Michael Lechner, Gabriel Okasa

October 2018 Discussion Paper no. 2018-11

School of Economics and Political Science,
Department of Economics

University of St.Gallen

Editor: Vanessa Pischulti
University of St.Gallen
School of Economics and Political Science
Department of Economics
Müller-Friedberg-Strasse 6/8
CH-9000 St.Gallen
Phone +41 71 224 23 07
Email seps@unisg.ch

Publisher: School of Economics and Political Science
Department of Economics
University of St.Gallen
Müller-Friedberg-Strasse 6/8
CH-9000 St.Gallen
Phone +41 71 224 23 07

Electronic Publication: http://www.seps.unisg.ch

Predicting Match Outcomes in Football by an Ordered Forest Estimator [1]

Daniel Goller, Michael C. Knaus[2], Michael Lechner[3], Gabriel Okasa

Author's address:

Daniel Goller
Swiss Institute for Empirical Economic Research (SEW)
University of St.Gallen
Varnbüelstrasse 14
CH-9000 St.Gallen
Email daniel.goller@unisg.ch

Michael C. Knaus
Swiss Institute for Empirical Economic Research (SEW)
University of St.Gallen
Varnbüelstrasse 14
CH-9000 St.Gallen
Email michael.knaus@unisg.ch
Website mcknaus.github.io

Michael Lechner
Swiss Institute for Empirical Economic Research (SEW)
University of St.Gallen
Varnbüelstrasse 14
CH-9000 St.Gallen
Email michael.lechner@unisg.ch
Website www.michael-lechner.eu

Gabriel Okasa
Swiss Institute for Empirical Economic Research (SEW)
University of St.Gallen
Varnbüelstrasse 14
CH-9000 St.Gallen
Email gabriel.okasa@unisg.ch

---

[1] We thank Alex Krumer for his invaluable contributions in the earlier stages of the Soccer Analytics project.

[2] Michael C. Knaus is also affiliated with IZA, Bonn.

[3] Michael Lechner is also affiliated with CEPR and PSI, London, CESIfo, Munich, IAB, Nuremberg, and IZA, Bonn.

Abstract

We predict the probabilities for a draw, a home win, and an away win, for the games of the German Football Bundesliga (BL1) with a new machine-learning estimator using the (large) information available up to that date. We use these individual predictions in order to simulate a league table for every game day until the end of the season. This combination of a (stochastic) simulation approach with machine learning allows us to come up with statements about the likelihood that a particular team is reaching specific places in the final league table (i.e. champion, relegation, etc.). The machine-learning algorithm used, builds on a recent development of an Ordered Random Forest. This estimator generalises common estimators like ordered probit or ordered logit maximum likelihood and is able to recover essentially the same output as the standard estimators, such as the probabilities of the alternative conditional on covariates. The approach is already in use and results for the current season can be found at www.sew.unisg.ch/soccer_analytics.

Keywords

Prediction, Machine Learning, Random Forest, Soccer, Bundesliga

JEL Classification

Z29, C53

1 Introduction

Predicting the outcome of football (i.e. soccer) games based on past information is a non-standard predictive task because of the nature of the game outcome, as well as because of the importance of uncertainty (luck and unobservables). The game outcome consists of the scores of the two teams that are usually either collapsed into a goal-difference or further aggregated to reflect whether the game ended as a win for the home or away team, or as a draw. From a statistical perspective, such outcomes have bounded support and, thus, standard linear modelling can be expected to perform poorly. The large amount of uncertainty in the game outcomes due to just luck or due to game or team specific unobservables (e.g. hidden injuries of players, etc.) makes it imperative to use prediction methods that fully exploit the potential of the available information, as well as to uncover the uncertainty of a match outcome. The latter is also relevant when interest is not only in single games, but also in a league table at the end of the season. Obviously, such league tables should capture the uncertainty for the single games accumulated over a season to be useful guides on what to expect.

Recently, machine-learning methods have shown their power in all sorts of prediction problems,[1] in particular in situations where the relation of the variables capturing the information used to predict with the target of the prediction, i.e. here the outcome of the game, is non-linear. However, so far there has been only little development in gearing these methods explicitly towards the estimation of the probabilities of ordered outcomes, such as score differences, points, or just wins, draws, and losses. Lechner and Okasa (2018) propose to adapt classical random forest estimation, which is known to have excellent predictive performance (e.g. Biau and Scornet (2016), Fernández-Delgado, Cernadas, Barro, and Amorim (2014)) to

[1] For a statistical treatment of many of these methods see Hastie, Tibshirani and Friedman (2009) and James, Witten, Hastie, and Tibshirani (2013).

the problem of predicting probabilities of ordered categorical outcomes, such as the win-draw-loss problem of a football game. In this paper, we use their approach to predict game outcomes of the German Fussball Bundesliga (BL1) based on more than 10 years of data on game outcomes as well as extensive information about teams, their players, and their environment. These predictions are then used to obtain the final season rankings in a way that reflects and shows the magnitude of the inherent uncertainty of football games.

While there are many approaches to predict football games (e.g. Leitner, Zeileis, and Hornik (2010), Nakamura et al. (2018) and references therein), the use of machine learning methods is still rather rare (e.g. for international championships see Groll, Kneib, Mayr, and Schauberger (2018) and Groll, Ley, Schauberger, and Van Eetvelde (2018)). In a recent paper, Baboota and Kaur (2018) used gradient boosting, another machine learning method, to predict the game outcomes of the English Premier League. A major difference of their approach compared to our approach is that their goal was to get the best prediction of a game, i.e. will it be a win, a draw, or a loss, while we are interested in the probabilities of these events occurring. In technical terms, they considered a classification problem, while our problem has the structure of a regression problem. The latter is required if the goal is to use these predictions to predict the final season outcome probabilistically, i.e. to end up with probabilities that a particular team becomes the champion, gets relegated or will play in the Champions League next season.

In the next section, we briefly introduce the machine learning method developed to predict the probabilities of the ordinal outcomes. Sections 3 shows how these predicted probabilities are used to obtain the end-of-season results. Section 4 illustrates the empirical application of the methods to the German Bundesliga 1 and compares the predictions to other publicly available predictions as well as betting odds. Section 5 concludes. Finally, the appendix documents the data used for the application.

2 The Ordered Forest estimator

2.1 Random forests

In the machine learning literature, random forests as developed by Breiman (2001) became a widely used prediction method. The random forest algorithm is based on randomly constructing and combining regression trees. In particular, the trees are combined via bootstrap aggregation, the so-called bagging, with additional randomness within the tree construction. A single regression tree (Breiman, 1984) recursively splits the covariate space into separate regions based on minimizing the sum of squares at each split of the tree. The final prediction for evaluation point x is then the average of the observations falling into the same end-node $L(x)$, the so-called leaf. Regression trees with many splits tend to have low bias, but a rather high variance due to the path-dependent structure. Bagged trees achieve a variance reduction by averaging many low bias trees. However, as opposed to bagged trees, the random forest decorrelates the trees to lower the variance even further (Hastie, et al., 2009). This is achieved within the tree-growing step where at each split point, only a random subset of covariates is considered. More formally, the random forest algorithm draws a bootstrapped sample b of size N and grows a regression tree $T_b(x)$ by choosing m out of p covariates ($m << p$) at random for the split until the minimum leaf size is reached. The final random forest estimate $RF^B(x)$ is the ensemble of B regression trees as

$$RF^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \quad with \quad T_b(x) = \frac{1}{\left|\{i : X_i \in L(x)\}\right|} \sum_{\{i:X_i \in L(x)\}} Y_i$$

where X_i denotes the covariates and Y_i the outcomes. In their recent contributions, Wager and Athey (2018) and Athey, Tibshirani, and Wager (2018) further modify the random forest algorithm and implement the so-called honest splitting rule, which uses different observations for both placing the splits and for estimating the effects. This is important for

statistical inference and contributes to the prediction accuracy as it helps to reduce the bias of the estimates even further.

2.2 Random forest estimation of ordered probability models

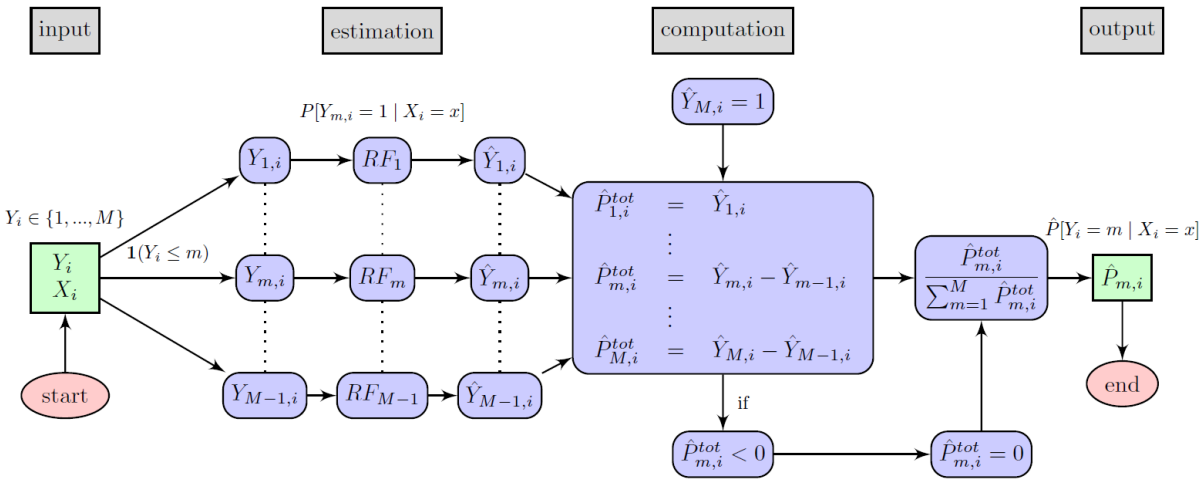
Despite the wide usage of random forests as a prediction tool (Athey, 2018), the major targets are either continuous or discrete outcomes, while the estimation of models involving ordered outcomes, as those of a football game, is not well established.[2] However, similar to the standard econometric ordered probability models (see Wooldridge (2010) for an overview), it is desirable to take the ordering nature into account to prevent loss of valuable information. To do so, Lechner and Okasa (2018) develop an Ordered Forest estimator, which explicitly incorporates the ordering information in the outcomes. Due to the underlying random forest algorithm, the estimator can flexibly deal with high-dimensional covariate spaces, possibly larger than sample size, while still providing the standard econometric output such as ordered outcome probabilities and marginal effects. Moreover, under certain conditions, statistical inference about the estimated effects is feasible as well. Thus, the Ordered Forest estimator can be regarded as a more flexible alternative to traditional econometric models, such as ordered logit or ordered probit. An extensive discussion of the estimator and the inference procedure as well as a simulation study is provided in Lechner and Okasa (2018).

Consider an ordered outcome variable $Y_i \in \{1, \ldots, M\}$ with ordered categories m. For a sample of size $N\,(i = 1, \ldots, N)$, the estimation of the conditional ordered outcome probabilities evaluated at x, i.e. $P\left[Y_i = m \middle| X_i = x\right]$ is based on an estimation of cumulative probabilities given by binary indicators $Y_{m,i} = \underline{1}(Y_i \leq m)$ for $m = 1, \ldots, M-1$. Then a regression random forest is estimated for all $M-1$ binary indicators, obtaining the predictions $\hat{Y}_{m,i} = \hat{P}(Y_{m,i} = 1 \middle| X_i = x)$.

[2] The few exceptions are the works of Hothorn, Hornik, and Zeileis (2006) and Hornung (2017).

The prediction for the *M*-th category is given as $\hat{Y}_{M,i} = 1$ as the cumulative probabilities must sum up to one. Based on the cumulative probabilities, the probabilities for each respective category m for all i are subsequently computed. The probability of the first outcome category is defined as $\hat{P}_{1,i}^{tot} = \hat{Y}_{1,i}$ and stems directly from the random forest estimation as in the case of a binary outcome; the estimated conditional mean translates to a valid probability estimate. For the following outcome categories $m = 2,\dots M$ the algorithm exploits the nature of cumulative probabilities and as such isolates the probability of the *m*-th category by subtracting the estimated probability of the preceding category as $\hat{P}_{m,i}^{tot} = \hat{Y}_{m,i} - \hat{Y}_{m-1,i}$. If some of the resulting probabilities become negative, these are set to zero, i.e. $\hat{P}_{m,i}^{tot} = 0$ if $\hat{P}_{m,i}^{tot} < 0$.[3] Lastly, it is ensured that the probabilities sum up to one and as such for all outcome categories $m = 1,\dots M$ the probabilities are normalized as $\hat{P}_{m,i} = \dfrac{\hat{P}_{m,i}^{tot}}{\sum_{m=1}^{M} \hat{P}_{m,i}^{tot}}$ where the probabilities $\hat{P}_{m,i}$ correspond to the conditional ordered outcome probabilities, i.e. $\hat{P}_{m,i} = \hat{P}\left[Y_i = m \middle| X_i = x\right]$. A graphical illustration of the algorithm is depicted in Figure 1.

Figure 1: An Illustration of the Ordered Forest Algorithm



[3] In practice, this is a rather rare case, especially when the forests are estimated with honesty.

Notice, that the above approach makes use of linear combinations of probability estimates from the regression random forest. Hence, if a regression random forest fulfils the conditions needed for the consistency and normality, the ordered random forest shares these properties too and thus enables conducting statistical inference. Computationally, the Ordered Forest as described above requires the estimation of $M-1$ regression random forests in the training data. Although this might appear as a rather demanding exercise, given the majority of empirical applications features limited number of outcome categories and the fast software implementations available, this becomes less of an issue.

3 Predict league outcomes – basic methodology

Once the probability of a win, draw or loss for a particular team has been predicted, such predictions can be aggregated to obtain the final league table. It appears natural to compute the expected points per team per game (3 points for a win, 1 for a draw, 0 for a loss) using the estimated probabilities, and add the points over all games leading to expected end-of-season points for all teams. Ordering the teams according to their expected points in the season leads to the final ranking. In order to capture the uncertainty in game outcomes, we also use an alternative approach. Instead of computing the expected points of a game, we randomly draw a simulated outcome based on the predicted probabilities. Depending on the realisation of the random variable, we assign 3, 1, or 0 points to teams, do this for all games, and then add-up the points. This process is repeated many times.[4] Finally, the probability of becoming champion, for example, is computed by counting the number of times a team was the first in the simulations, and dividing this number by the number of simulations. In the same way, all other probabilities of ranking positions of interest can be obtained.

[4] Predictions are dynamic in the sense that the points achieved so far are also part of the covariates. If this variable is not observed (e.g. because we are predicting games for round 34, but so far only 23 rounds have been played), then the unknown points are either substituted by their expectations or their simulated values, depending on the particular method used.

4 An illustration: The German Fussball Bundesliga

4.1 The database

Starting from the 2007/08 season, we collected data on all matches in the German Bundesliga. The season 2018/19 is therefore the 12th year of data in this database, which is continuously updated before every game day. This resulted in 3366 observations before the 2018/19 season's first game day. About 300 variables are used for the predictions and gathered from various sources. In the following, we briefly explain the categories of variables and their sources.

We collect a wide range of player information, teams' composition and club specifics to approximate teams' abilities. Various variables are constructed using information about teams, players and coaches from www.transfermarkt.com.[5] Those are, e.g., market values, as well as height and age structure within the team. Since teams' compositions change before as well as after the first half of the season, this information is updated whenever there are potential changes. The same source is used for the weekly updates of the reported stadium attendance, as well as potential managerial changes. TV Revenues are calculated using the allocation key published by the DFL for the seasons 2006/07 until 2009/10 and taken from www.fernsehgelder.de for the years from 2010/11 onwards.

Other factors influencing the performance of a team may be location related. Thus, data regarding distance and travel time are calculated as shortest routes between the two competing cities from www.google.com/maps. Capacity of the stadiums is taken from www.worldstadiums.com, as well as the respective Wikipedia pages.

[5] We refer the reader to the works of Bryson, Frick, and Simmons (2013) and Franck and Nüesch (2012) for a discussion on the reliability of the information generated by this source, as well as how well this approximates teams' abilities.

Accounting for schedule related factors constitutes another large part of the database. Information on international association competitions, qualification rounds and friendly matches are collected from www.fifa.com once before the season. The schedules of the European association club competitions are taken from www.uefa.com, as well as www.kicker.de. This is updated for the teams in the European competitions, i.e. the Champions League and the Europa League on a regular basis. The Bundesliga schedule is obtained from www.transfermarkt.com and updated as soon as the exact timing is published.

Information regarding the regional economic situation is collected from www.regionalstatistik.de.[6] Previous seasons' game outcomes are constructed using match specific information from www.football-data.co.uk. Here we collect outcomes of the last weeks' matches before every game day. Additionally we are using the same source to account for previous games' outcomes using information on the last 1-4 matches of each team.

Finally, we obtain pre-match betting odds from seven of the major bookmakers, i.e. Bet365, Bwin, Interwetten, Ladbrocker, Pinnacle, William Hill and BetVictor from www.football-data.co.uk. Odds are collected Friday afternoon for the weekend and Tuesday afternoon for the midweek game days and include the odds for a home win, draw and away win.[7] Those are not used in the estimation but to benchmark the predictions against the bookmakers.

4.2 The 2017/18 and 2018/19 seasons

We started using the Ordered Forest to predict the Bundesliga season 2017/18. The prediction model was estimated with the information in our database before the first match day. This model was then used to predict the expected ranking and the expected points at the end of

[6] In more detail: For the unemployment we take table code 13211-01-03-4, for GDP table code 82111-01-05-4.

[7] For the full set of covariates, as well as some descriptive statistics, the interested reader is referred to Appendix A.

the season as described in Section 3. Table 1 shows the resulting predictions and the comparison to the actual outcomes of the season. Every season produces its positive and negative surprises. For the season 2017/18, VfB Stuttgart performed unexpectedly well with a predicted rank of 16 and a realized rank of 7. In contrast, 1. FC Köln was predicted to finish in the midfield of the table but finished last and was unexpectedly relegated. Such differences between predicted and actual outcomes are common for football predictions because it is unrealistic to predict all developments and dynamics of a whole season.

Table 1: Comparison of the predicted and the actual final table of Bundesliga season 2017/18

	Rank		Points	
Team	Predicted	Actual	Predicted	Actual
FC Bayern München	1	1	74.5	84
Borussia Dortmund	2	4	64.7	55
Bayer 04 Leverkusen	3	5	54.1	55
Borussia Mönchengladbach	4	9	52.4	47
Schalke 04	5	2	52.1	63
RB Leipzig	6	6	50.0	53
TSG Hoffenheim	7	3	47.0	55
VfL Wolfsburg	8	16	46.1	33
1. FC Köln	9	18	45.2	22
Hertha BSC Berlin	10	10	43.4	43
Werder Bremen	11	11	40.5	42
FC Augsburg	12	12	40.0	41
Mainz 05	13	14	39.9	36
Eintracht Frankfurt	14	8	39.9	49
SC Freiburg	15	15	39.7	36
VfB Stuttgart	16	7	38.8	51
Hamburger SV	17	17	38.5	31
Hannover 96	18	13	37.0	39

To assess the predictive performance of the Ordered Forest, we compare the predictions of Table 1 with other predictions that forecast the final ranking and points. To this end, we

access twelve alternative predictions from www.bstat.de that collects different forecasts of experts or algorithms before each season.[8]

We consider Spearman's rank correlation coefficient of the actual table and each prediction as performance measure for the ranks and the root mean squared error (RMSE) to assess the accuracy of the predicted points. Table 2 shows that the Ordered Forest is outperformed by three alternative predictions in terms of rank correlation. The expert predictions of the newspapers General-Anzeiger and Spiegel Online, as well as the algorithmic prediction of goalimpact show larger rank correlations, which indicates that their predictions were closer to the true ranking than the one based on Ordered Forests. However, in terms of accurately predicting the final points, the Ordered Forest performs best showing the smallest RMSE of all available forecasts.

Table 2: Comparison of different predictions for the final table of Bundesliga season 2017/18

	Rank correlation	RMSE
Ordered Forest	0.64	8.9
bundesliga-prognose.de	0.43	16.8
Club Elo	0.62	-
Euro Club Index	0.61	9.0
FiveThirtyEight	0.63	9.9
Fupro.de	0.63	11.5
fussball-manager.com	0.58	16.4
fussballmathe.de	0.63	12.1
General-Anzeiger	0.74	-
Goalimpact	0.71	9.0
kickform.de	0.61	9.4
Spiegel Online	0.75	-
transfermarkt.de	0.60	-

Section 3 described that the predictions of the final table produce probabilities for the outcomes of each match as a by-product. These predictions are updated after each match day to

[8] The results of SEW Soccer Analytics that are reported there were estimated with an old version of the algorithm based on Lasso prediction, while the Ordered Forest was still in the test phase. However, note that all results that are reported here were obtained by using only the information that were available before the season.

incorporate the recent developments. In the following, we compare these predictions to the betting odds of the seven bookmakers in our database.

We evaluate the performance by using different betting strategies and calculating the hypothetical returns on investment (ROI) for each. First, we consider a *proportional strategy.* This means that we bet one Euro on each match and split it according to the predicted probabilities of the Ordered Forest for each outcome. E.g., if the predicted probability of a home win is 50%, we bet 50% of our hypothetical money on a home win. To see how we would earn or lose money, consider the cases where the home team actually wins and where the betting odds are 1.9, 2.0, or 2.1. In the first case we lose money because we spent 1€ and receive 1.9*0.5€ = 0.95€. Accordingly, we break even for the second case and earn if the odds were 2.1. In the latter case, the probability that is implied by the betting odd is 1/2.1 = 47.6%. This means our predicted probability of the actual outcome was higher and we thus realize a ROI of (1.05-1)/1 = 5%. The first three columns of Table 3 show the ROI of this strategy for the season 2017/18, the season 2018/19 until the seventh match day, and both season combined.

Table 3: Return of investment in percent of different betting strategies in different seasons

	Odds			Odds net of fees		
	2017/18	2018/19	2017/19	2017/18	2018/19	2017/19
B365	-5.7	0.2	-4.6	-0.8	5.5	0.4
Bwin	-5.9	-0.1	-4.8	-1.0	5.0	0.2
Interwetten	-5.9	-1.1	-5.0	-0.6	4.2	0.3
Ladbrockes	-7.1	-0.7	-5.9	-1.0	5.7	0.3
Pinnacle	-3.3	2.9	-2.1	-1.1	5.6	0.2
William Hill	-6.3	-1.0	-5.3	-0.6	4.9	0.5
BetVictor	-4.9	1.3	-3.7	-1.2	5.2	0.1
Value bet:	-6.9	34.5	1.0	-	-	-

Notes: Results are based on the probabilities obtained by the Ordered Forests and betting odds provided by www.football-data.co.uk.

The results show negative ROIs between -3% and -7% for the season 2017/18.[9] However, after the first eight matches of the season 2018/19, some ROIs are positive showing up to 2.9% for Pinnacle.

There are at least two explanations for the mostly negative results. First, the Ordered Forest has no access to short-term developments like injuries of players or other player related information. However, this information is most likely reflected in the betting odds. Second, the betting odds include implicit fees as the implied probabilities of the three outcomes sum up to more than 100%. To correct for this and to get a 'fair' comparison of our probabilities and the probabilities used by the bookmakers, we would need access to their probabilities. However, it is not clear how they distribute their fees over the outcomes (see for discussions e.g., Levitt (2004), Paul and Weinbach (2007), Paul and Weinbach (2008)). For simplicity, we assume that the fees are proportionally distributed over the different outcomes and create odds that are 'net of fees' in the following way. We invert the odds of the three outcomes to get the implied probabilities, we normalize those to sum to 100%, and invert the normalized probabilities again to obtain the net odds. The results in the last three columns of Table 3 using these net odds show two things. First, the losses in season 2017/18 are dramatically reduced to about -1% and the returns after the first eight matches of season 2018/19 would be clearly positive around 5%. Second, the variation of the returns across different bookmakers is much smaller and the hypothetical returns are very similar. This implies that the differences using the unadjusted odds are mostly driven by different fees charged by the bookmakers.

This correction for fees is rather ad-hoc. Thus, we implement a second betting strategy that bets only on those outcomes where our estimated probabilities exceed those that are implied by the bookmaker's odds. E.g., consider again a predicted probability of a home win of 50%. Even if we knew that this is the true value, we would lose money in the long run if the betting

[9] Also Baboota and Kaur (2018) find that their method is slightly outperformed by the bookmakers odds.

odds are below 2.[10] This could happen because the bookmakers wrongly expect a higher probability of a home win or because of the implicit fee they charge. The so-called *value bet* strategy therefore only bets on outcomes for which the predicted probabilities are larger than the probabilities implied by the odds. If this happens for several bookmakers or outcomes, we bet on the bookmaker-outcome combination with the highest ratio of our predicted probability and the bookmaker's implicit probability. The last row of Table 3 shows the ROIs of such a strategy where we bet 1€ if our predicted probabilities exceed the bookmakers implicit probabilities. The -6.9% return suggests that the prediction model was not competitive in season 2017/18. However, for the first eight matches of season 2018/19, this strategy would have created a positive return of 34.5%. Note that this is not the result of few lucky bets but we bet on 71 out of 72 matches.

Section 3 describes how we use the Ordered Forest to obtain probabilities for each rank. Our data base comprises no betting odds for specific ranks or competing forecasts to validate these probabilities. However, we illustrate the information obtained and how the predictions evolve during the season in Table 4. It aggregates the probabilities for each rank for different aspirations, like champion (rank 1), qualification for the champions league (rank 2-4), etc., in the current season 2018/19. The left part of the table shows the probabilities before the season and the right part the most recent updates after matchday 8.[11] Table 4 shows how taking into account the materialized results so far change the probabilities in the course of the season. For example, the probability that Borussia Dortmund wins the league increased from 6% before the season to 21% because they are currently 4 points ahead of Bayern München. In the opposite direction, VfB Stuttgart started out as a promising candidate to qualify for the Champions League (16%) or the Europa League (15%). However, the bad season start decreased these

[10] It means that we would earn 50% of the time less than double of our bet, which leads to an expected loss.

[11] The most recent results can be found on www.sew.unisg.ch/soccer_analytics.

chances substantially and instead more than doubled the probabilities to finish on the relegation ranks.

Table 4: Probabilities to achieve different season goals in Bundesliga season 2018/19

	Before season start						After matchday 8					
Season goals	1	2-4	5-6	7-15	16	17-18	1	2-4	5-6	7-15	16	17-18
FC Bayern	79	18	2				71	28				
Dortmund	6	47	19	27			21	70	6	3		
Leipzig	5	49	17	28			4	56	22	19		
Leverkusen	4	44	19	32				19	22	55	1	2
Gladbach	2	27	20	45	2	4	3	60	21	16		
Schalke	2	20	18	53	3	4		2	7	73	7	10
Hoffenheim		20	13	55	5	7		9	18	66	3	4
Stuttgart		16	15	57	4	8			6	66	11	17
Hertha		11	11	61	7	11		18	24	55	1	1
Bremen		8	10	62	7	13		20	27	51		
Wolfsburg		8	10	63	7	11		2	7	72	7	12
Augsburg		7	8	62	8	15		3	8	73	7	10
E. Frankfurt		5	8	65	7	15		10	17	67	2	3
Mainz		6	5	64	9	16		1	5	72	9	13
Hannover		5	8	60	10	18			3	59	11	27
Freiburg		4	7	58	10	21			3	64	12	20
Düsseldorf		2	4	55	12	28				36	15	48
Nürnberg		3	5	54	9	29			2	52	13	32

Notes: Probabilities in percent. Those below one percent are not shown.

5 Conclusion

In this paper, we presented a machine learning based algorithm to predict the season outcome of sport leagues in a probabilistic fashion. As a by-product, we also obtain predictions of particular games in any round of interest. This approach is applied since season 2017/18 to predict the match and season outcomes of the German Bundesliga 1. The fact that the target of the prediction problem is to estimate the league table at the end of the season limits the number of variables that can be updated during the season as such variables would require their own prediction models. Despite this, when comparing our game predictions to the ones of the betting firms, which use a much more up-to-date information set, they are surprisingly close and we can form even strategies that outperform them for the current season.

In the future, it will be interesting to apply the suggested methods to other leagues and other sports.

6 Bibliography

Athey, S., 2018. The Impact of Machine Learning on Economics. In: *The Economics of Artificial Intelligence: An Agenda (forthcoming).* s.l.:University of Chicago Press.

Athey, S., Tibshirani, J. & Wager, S., 2018. Generalized Random Forests. *Annals of Statistics,* Issue forthcoming, pp. 1-49.

Baboota, R. & Kaur, H., 2018. Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting.*

Biau, G. & Scornet, E., 2016. A random forest guided tour. *Test.*

Breiman, L., 1984. *Classification and regression trees.* s.l.:s.n.

Breiman, L., 2001. Random Forests. *Machine Learning,* 45(1), pp. 5-32.

Bryson, A., Frick, B. & Simmons, R., 2013. The Returns to Scarce Talent: Footedness and Player Remuneration in European Soccer. *Journal of Sports Economics.*

Fernández-Delgado, M. et al., 2014. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?. *Journal of Machine Learning Research.*

Franck, E. & Nüesch, S., 2012. Talent and/or popularity: What does it take to be a superstar?. *Economic Inquiry.*

Groll, A., Kneib, T., Mayr, A. & Schauberger, G., 2018. On the dependency of soccer scores - A sparse bivariate Poisson model for the UEFA European football championship 2016. *Journal of Quantitative Analysis in Sports.*

Groll, A., Ley, C., Schauberger, G. & Van Eetvelde, H., 2018. Prediction of the FIFA World Cup 2018-A random forest approach with an emphasis on estimated team ability parameters. *arXiv preprint arXiv:1806.03208.*

Hastie, T., Tibshirani, R. & Friedman, J., 2009. The Elements of Statistical Learning. *Elements,* Volume 1.

Hornung, R., 2017. *Ordinal Forests,* Munich: s.n.

Hothorn, T., Hornik, K. & Zeileis, A., 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics,* 15(3), pp. 651-674.

James, G., Witten, D., Hastie, T. & Tibshirani, R., 2013. *An Introduction to Statistical Learning.* s.l.:s.n.

Lechner, M. & Okasa, G., 2018. Random Forest Estimation of the Econometric Ordered Choice Model. *Unpublished Manuscript.*

Leitner, C., Zeileis, A. & Hornik, K., 2010. Forecasting sports tournaments by ratings of (prob)abilities: A comparison for the EURO 2008. *International Journal of Forecasting.*

Levitt, S. D., 2004. Why are gambling markets organised so differently from financial markets?. *Economic Journal.*

Nakamura, L. R. et al., 2018. A new continuous distribution on the unit interval applied to modelling the points ratio of football teams. *Journal of Applied Statistics*.

Paul, R. J. & Weinbach, A. P., 2008. Price setting in the nba gambling market: Tests of the levitt model of sportsbook behavior. *International Journal of Sport Finance.*

Paul, R. & Weinbach, A. P., 2007. Does Sportsbook.com set pointspreads to maximize profits? Tests of the Levitt model of Sportsbook behavior. *The Journal of Prediction Markets.*

Wager, S. & Athey, S., 2018. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*.

Wooldridge, J. M., 2010. *Econometric Analysis of Cross Section and Panel Data.* s.l.:s.n.

A Data appendix

As mentioned in chapter 4.1 we collected players and team characteristics, as well as match, schedule, location related and regional economic information.

The first set of variables are previous seasons' outcomes. We collected the average numbers of shots, shots on target, fouls committed, corners, yellow and red cards, final points, attendance as well as share of capacity of each team in the previous season. The resulting variables are marked as "home" if the numbers concern the home team, "away" for the away team and "home - away" if the numbers are the difference between the home and the away team.

The database contains several previous game outcomes. We constructed variables with mean points in the last one to four matches, for each team, as well as the difference between the teams. As an example, *PG points last 3 matches | home* captures the mean points, which the home team earned during the last three matches. Further, there are variables, e.g. *PG points share of total | away*, with the share of all potential points gained by the away team from the start of the season.

Location related factors are captured with a variable that accounts for the capacity of the home stadium, as well as the shortest distance in km between the two home cities of the competing teams and the travelling time of the shortest route in minutes.

To capture potential schedule related effects we created a set of indicators for the season[12], day of the week[13] or round. *Weekend home advantage* indicates if the home team is playing home on Friday, Saturday or Sunday.

[12] Season id is constructed as: 07=2007/2008,….,17=2017/2018.

[13] Weekday is constructed as: 1= Monday, 2= Tuesday, …, 7= Sunday.

If there was an international game day with friendly or qualification matches of the national teams the dummy variable *After International Break* is equal to 1 and 0 otherwise. For matches, which were not held according to the schedule the variable *Delayed match* is capturing this. *Short week* is indicating a week in which there is a midweek match day additional to the usual weekend matches, while *Weekend after midweek round* is separately indicating those weekends.

Those seasons after a European or World Championship are marked by the variable *World cup/European championship season*, with two separate dummy variables for the two months before and after those events. The season in which there is the African cup is indicated by *African cup season*, as well as the specific months in which the African cup took place by *African cup months*. Further, *already champion | home - away* and *already relegated | home - away* are 0 if none (or both) of the two teams are already champion / relegated, 1 if the home team and -1 if the away team is. Teams are "already champion" or "already relegated" if there is no theoretical chance left that the outcome of the season would be different.

The variables *before (after) European match* indicate if the home/away team played the days before, respectively has to play the days after this match another match in an European competition. *Round * begin/mid/end* each indicate the beginning, mid and end of the season, with the respective rounds in brackets.

Those teams which promoted last year from the second division are for example marked with the *promoted | away* variable. The variables *market value* carry information on the market values of the teams, as well as values standardized by season. *TV Revenue* is the national revenue from sales of the broadcasting rights.

Further, there are variables constructed from information on the team composition. Those are the (normalized) Herfindahl-Hirschman Index, (d) HHI, which is defined as the sum of squares of the shares of the market value of each player within the team. Also there are variables

capturing the within team inequality measured as ratio of Top 3 (11) most valuable players to the market value of those ranked 12 – 14 (12-21). *New Coach* is defined as 1 if the team got a new coach after the start of the season.

Variables capturing diversity of a team are represented in several age related variables, like minimum, maximum or the standard deviation of age, the share of left or two footed players, as well as variables regarding the height of the players in the squad or the eleven most valuable players.

Moreover, *traditional club* is a selection of clubs with a history, like Borussia Dortmund or VfB Stuttgart, *yo-yo club* is a club that was often relegated to the second division and/or promoted to the first division, and *other clubs* are clubs which are neither traditional, nor yo-yo clubs.

Finally, two regional economic variables capture the economic situation in form of the log of GDP per capita in the city of the team, as well as the unemployment.

Table A.1: Descriptive Statistics

Variables	Reference	Unit	Mean (Std. dev.)	Update
Previous seasons (PS)' outcomes				
PS shots	home	numerical	14.70 (2.31)	yearly
PS shots difference	home–away	numerical	2.71 (2.90)	yearly
PS shots on target	home	numerical	5.65 (1.50)	yearly
PS shots on target difference	home–away	numerical	1.10 (1.46)	yearly
PS fouls	home	numerical	15.66 (2.11)	yearly
PS fouls difference	home–away	numerical	-1.23 (2.48)	yearly
PS corners	home	numerical	5.63 (1.08)	yearly
PS corners difference	home–away	numerical	1.23 (1.30)	yearly
PS yellow cards	home	numerical	1.62 (0.36)	yearly
PS yellow cards difference	home–away	numerical	-0.34 (0.50)	yearly
PS red cards	home	numerical	0.07 (0.06)	yearly
PS red cards difference	home–away	numerical	-0.03 (0.10)	yearly
PS points	home	numerical	1.70 (0.43)	yearly
PS points difference	home–away	numerical	0.50 (0.60)	yearly
PS attendance	home	numerical	43837 (15424)	yearly
PS attendance difference	home–away	numerical	1336 (15499)	yearly
PS share of capacity	home	percentage	0.92 (0.08)	yearly
PS share of capacity difference	home–away	percentage	0.001 (0.08)	yearly
Location related				
Public transport time between cities		minutes	197.58 (91.76)	once
Distance between cities		kilometer	373.37 (185.98)	once
Home Stadium capacity		discrete	46813 (17550)	match
Previous game (PG) outcomes				
PG points last match	home	numerical	1.18 (1.26)	match
PG points last 2 matches	home	numerical	1.35 (0.92)	match
PG points last 3 matches	home	numerical	1.32 (0.76)	match
PG points last 4 matches	home	numerical	1.36 (0.69)	match
PG points share of total	home	percentage	0.45 (0.19)	match
PG points last match	away	numerical	1.58 (1.31)	match
PG points last 2 matches	away	numerical	1.40 (0.93)	match
PG points last 3 matches	away	numerical	1.43 (0.79)	match
PG points last 4 matches	away	numerical	1.39 (0.69)	match
PG points share of total	away	percentage	0.46 (0.19)	match
PG points last match difference	home–away	numerical	-0.40 (1.83)	match
PG points last 2 matches difference	home–away	numerical	-0.05 (1.30)	match
PG points last 3 matches difference	home–away	numerical	-0.11 (1.08)	match
PG points last 4 matches difference	home–away	numerical	-0.04 (0.96)	match
PG points share of total difference	home–away	percentage	-0.01 (0.27)	match
Schedule related				
Season id		categorical	12.00 (3.16)	yearly
Weekday		categorical	5.91 (1.00)	match
Weekend home advantage		dummy	0.47	match
Round		categorical	17.50 (9.81)	yearly
After international break		dummy	0.10	match
Delayed match		dummy	0.002	match
Short week		dummy	0.10	match
Weekend after midweek round		dummy	0.05	match
World cup/European championship season		dummy	0.45	yearly
African cup season		dummy	0.55	yearly
African cup months		dummy	0.09	yearly

Table A.1: continued

Variables	Reference	Unit	Mean	Update
Post World cup/European championship		dummy	0.08	yearly
Pre World cup/European championship		dummy	0.11	yearly
Already champion difference	home–away	categorical	0.001 (0.10)	match
Already relegated difference	home–away	categorical	-0.001 (0.07)	match
Before European match	home	dummy	0.10	match
Before European match	away	dummy	0.10	match
After European match	home	dummy	0.10	match
After European match	away	dummy	0.10	match
Round * begin (matches 1-11)		categorical	1.94 (3.33)	yearly
Round * mid (matches 12-22)		categorical	5.50 (8.16)	yearly
Round * end (matches 23-34)		categorical	10.06 (13.78)	yearly
Team characteristics				
Promoted	home	dummy	0.13	yearly
Promoted	away	dummy	0.13	yearly
Promoted difference	home–away	categorical	0 (0.49)	yearly
Market value	home	EURO	112329501 (102768312)	regular
Market value difference	home–away	EURO	22825 (144964414)	regular
Standardized market value	home	-	0 (1.00)	regular
Standardized market value difference	home–away	-	0.0002 (1.45)	regular
Market value share	home, away	ratio	1.60 (1.94)	regular
TV Revenue	home	EURO	24216389 (9618865)	yearly
TV Revenue difference	home–away	EURO	0 (7764625)	yearly
Market value / TV revenue	home	EURO	4.44 (2.82)	regular
Market value / TV revenue difference	home–away	EURO	0.001 (3.99)	regular
Market value - TV revenue	home	EURO	88113112 (97807758)	regular
Market value - TV revenue difference	home–away	EURO	22825 (139769918)	regular
HHI	home	ratio	0.06 (0.01)	regular
HHI difference	home–away	ratio	-0.000001 (0.02)	regular
dHHI	home	ratio	0.03 (0.01)	regular
dHHI difference	home–away	ratio	0.000004 (0.01)	regular
Average market value	home	EURO	3757742 (3783521)	regular
Std. dev. market value	home	std. dev.	3607418 (3635022)	regular
Ratio of Top 3 to ranked 12 – 14 players' market value	home	ratio	3.07 (0.94)	regular
Ratio of Top 11 to ranked 12 – 21 players' market value	home	ratio	2.99 (0.85)	regular
Average market value difference	home–away	EURO	-49.62 (5301092)	regular
Std. dev. market value difference	home–away		1843 (5126814)	regular
Ratio of Top 3 to ranked 12 – 14 players' market value difference	home–away	ratio	0.0001 (1.34)	regular
Ratio of Top 11 to ranked 12 – 21 players' market value difference	home–away	ratio	0.001 (1.21)	regular
New coach	home	dummy	0.18	match
New coach difference	home–away	categorical	0.0003 (0.52)	match
Age mean difference	home–away	numerical	-0.001 (1.21)	regular
Age std. dev. Difference	home–away	std. dev.	0.0002 (0.72)	regular
Age 11 most valuable players difference	home–away	numerical	0.001 (1.48)	regular
Age ratio of top 11 to ranked 12 – 21 difference	home–away	numerical	0.0002 (0.09)	regular

Table A1: continued

Variables	Reference	Unit	Mean	Update
Age of those above 20 difference	home–away	numerical	0.0003 (1.12)	regular
Minimum age in the squad difference	home–away	categorical	-0.001 (1.20)	regular
Maximum age in the squad difference	home–away	categorical	-0.01 (2.89)	regular
Share left footed players difference	home–away	percentage	0 (0.08)	regular
Share two footed players difference	home–away	percentage	0 (0.08)	regular
Share left footed players among 11 most valuable players difference	home–away	percentage	0 (0.16)	regular
Share two footed players among 11 most valuable players difference	home–away	percentage	0 (0.12)	regular
Mean height difference	home–away	numerical	0 (1.51)	regular
Std. dev. height difference	home–away	std. dev.	0 (1.03)	regular
Mean height top 11 difference	home–away	numerical	0 (2.32)	regular
Std. dev. height top 11 difference	home–away	std. dev.	0 (1.92)	regular
Traditional club	home	categorical	11.69 (13.45)	once
Yo-yo club	home	categorical	6.45 (9.38)	once
Other clubs	home	categorical	1.27 (4.75)	once
Traditional club	away	categorical	11.69 (13.45)	once
Yo-yo club	away	categorical	6.45 (9.38)	once
Other clubs	away	categorical	1.27 (4.75)	once
Regional Economic Indicators				
Log GDP per capita difference	home–away	EURO	0 (0.25)	yearly
Unemployment difference	home–away	percentage	0 (5.15)	yearly

Notes: The standard deviation is reported in parentheses and not reported for dummy variables.

"home": the home team, "away": the away team. Updates which are indicated as *regular* are updated at least three times each season, i.e. before the season starts and after the transfer window closed in summer and winter, but as soon as there are major changes. Update category *match* points to updates in this variable before each new match day.