ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Bachelor Thesis Business Analytics & Quantitative Marketing

Predicting Results of Football Matches: Ordered Random Forest versus Bivariate Poisson Regression

Author

Bart H.P. de Zeeuw (434299)

Supervisor

MSc Nienke F.S. Dijkstra

Second Assessor

Prof. dr. Patrick J.F. Groenen

July 7, 2019

Abstract

Betting on football matches is a huge industry which exists already for a long time. Therefore, predicting sports matches has been the topic of a lot of researches. This paper attempts to replicate and extent the paper of Goller et al. (2018), where a new method to predict football matches is used. Various variables describing the teams in the 1. Bundesliga are used to predict the final league table of the 1. Bundesliga. The main focus of this paper is the comparison between the predictive power of the Ordered Random Forest Model (ORFM) of Goller et al. (2018) and the Bivariate Poisson Regression Model (BPRM). The predictive performance of both models is assessed by different performance measures, for example a hypothetical return on investment (ROI). The BPRM slightly outperforms the ORFM in terms of the performance measures used in this paper. Therefore, I conclude that the BPRM has a slightly higher predictive power than the ORFM. Hence, I would recommend the BPRM over the ORFM for predicting football matches.

Contents

1 Introduction

Betting on football (i.e. soccer) matches is not a new phenomenon. However, the sports betting industry has grown rapidly over the last decades. Due to the growth of the internet and mobile devices, the betting market has become more accessible than it used to be. According to Darren Small[1], director of integrity at Sportradar, a betting and sports data analyst, the industry is worth in between 700 billion and 1 trillion dollars a year, where 70% of the total worth comes from football related activity. Consequently, a proper prediction strategy for football matches is important, for both the firms in the betting industry as for the people who bet on those games.

Besides the importance for the betting industry, football clubs may gain from realistic estimates of their place in the league table as well. For example, having a good estimation of the league table can help the board of the club to get more adequate sponsor deals. The higher your club finishes, the more it can ask from its sponsors. Furthermore, the club's board can make a better judgment of the manager's performance in that particular season.

Machine learning has proved to be a useful method in all sorts of prediction problems. Although, there is little research conducted to examine the performance of these methods on the estimation of probabilities of ordered outcomes. In addition, machine learning is not used frequently when predicting the outcomes of football games. Therefore, the first part of this paper consists of a replication of the paper of Goller et al. (2018). Here, an extension of the classical random forest estimator will be tested on its predictive performance. Goller et al. (2018) use the so-called Ordered Random Forest Model (ORFM), developed by Lechner and Okasa (2018), which takes the natural ordering into account that exists in outcomes of football matches. Moreover, classical random forest estimation has an outstanding predictive performance in predicting probabilities of ordered outcomes, according to Fernández-Delgado et al. (2014). The estimates of these probabilities are then used to construct a prediction of the league table at the end of the season. This league table will be used to examine the predictive performance of this estimation strategy. In the second part of the paper, I will try to further enhance the paper by Goller et al. (2018) by comparing the ORFM with the Bivariate Poisson Regression Model (BPRM) of Karlis and Ntzoufras et al. (2005). Hereby, the performance of the ORFM can be seen in a broader perspective, than when it is not compared with another method.

This research can be used by betting companies when a particular prediction method shows to perform better in predictive power than the prediction method that a betting company uses. Furthermore, as stated before, using the best prediction method, could be useful for fans and in particular the board of a football club. Because of this, I believe that this research could be of use for both the betting and the football industry.

This paper is structured in the following manner. First, there is a short overview of papers that discuss this topic. Hereafter, the sources of the data are presented and the data set itself is discussed. Furthermore, adjustments in the data are being accounted for. The fourth Section contains the methodology of both models. Thereupon, the results of both models are presented. Finally, the limitations of this paper are discussed in Section 7, as well as suggestions for further research.

[1] See article: https://www.bbc.com/sport/football/24354124.

2 Literature Review

There have been numerous studies about predicting football matches and the best methods to do so. For example, Leitner et al. (2018) examined the forecast performance of methods build on ability ratings or bookmakers odds. They showed that the model where the bookmakers odds were aggregated performed best. Besides that, Min et al. (2008) proposed the use of Bayesian inference, rule-based reasoning and the so-called in-game time-series approach, rather than machine learning techniques. This technique allows for different kind of tactics in a game and in-game changes in tides between the two teams. This model outperforms two other so-called historic predictors that they built. Crowder et al. (2002) took another approach. They examined the independent Poisson model with dynamic attack and defense qualities to obtain probabilities for home win, draw or away win and compared their model to the model of Dixon and Coles (1997), which also has its roots in the independent Poisson model. Crowder showed that both of the approaches have similar predictive abilities. Furthermore, Baio and Blangiardo (2010) propose the so-called Bayesian hierarchical model. They have shown that, emperically, there is a certain correlation between the goals scored by the two competing teams. However, Baio and Blangiardo (2010) showed that the Bayesian hierarchical model is not inferior to the bivariate Poisson model, despite the fact that the Bayesian hierarchical model does not explicitly takes the correlation between the amount of goals scored by the two teams into account. In this paper we extend the bivariate Poisson model into a bivariate Poisson regression model.

As stated before, machine learning techniques are widely used to handle predictive challenges. For example, Groll et al. (2018) tested the predictive performance of several estimation methods, namely a Poisson regression models, random forests and ranking methods. In contrast to Baio and Blangiardo, they showed that a combination of the random forest and the ranking methods gives the best model in terms of predictive performance. In addition, Baboota and Kaur (2019) tested machine learning methods, such as the Gaussian naive Bayes model, a support vector machine model, the random forest method and finally gradient boosting. They concluded that out of these four models, gradient boosting had the best performance. However, gradient boosting was not capable of beating the bookmaker's predictions. Finally, Joseph et al. (2006) compared an expert constructed Bayesian network with a naive Bayesian network, a Bayesian network learned form statistical relationships in the date, a k-nearest neighbour implementation and a decision tree. They showed that generally speaking, the expert Bayesian network has the biggest predictive power of all five techniques.

As stated before, there exist numerous ways to predict the outcomes of football matches. The goal of this paper is to replicate the ORFM of Goller et al. (2018) and compare this method with the BPRM of Karlis and Ntzoufras (2005). This will put the findings of Goller et al. (2018) in a broader perspective.

3 Data

This research uses data that comes from various sources. Unfortunately, I was not able to find all the data that Goller et al. (2018) used. However, I think that the core variables from their paper are

incorporated in this research. The data consists of different kind of variables, which can be put into some categories. These categories are now briefly discussed.

First of all, the match results from all the 1. Bundesliga games are gathered from www.datahub.io. This contains all results of the matches played from the beginning of the season 2008/09 until the 33-rd match day of the season 2018/19.

Furthermore, the quality of a particular team can be approximated using certain variables. Examples of these *team characteristics* are: market value of the club, TV revenues received and average age of the team. These variables are obtained from www.transfermarkt.com. The TV revenues of each club are obtained from www.fernsehgelder.com from 2012/13 until 2018/19.

The location of the match can be of a huge influence on the game's result. Therefore, *location related* variables are obtained as well. The travel times between the two stadiums of the two competing teams are computed from www.google.com/maps. Here, I deviate from the paper of Goller et al. (2018). They use public transportation time as their approximation of the travel time between two cities. However, all 1. Bundesliga clubs have their own bus, so I think the travel time of a car is more adequate to use. Moreover, the same source is used to calculate the distance between two cities. Finally, the maximum capacity of each stadium is obtained from Wikipedia.

Another factor that could be of influence in the result of a match is the schedule that is used in a season. Moreover, international football, such as the Champions League and matches for a country, may affect the performance of a club. This information is collected from www.github.com.

As in the paper of Goller et al. (2018) the regional economic situation is retrieved from www.regionalstatistik.de. This contains the GDP and the unemployment rate in the cities of the teams.

Finally, the betting odds used by five of the world's biggest betting companies are obtained from www.football-data.co.uk. The odds of the following bookmakers are used: Bet365, Bwin, Interwetten, Ladbrocker and William Hill. It contains the odds of a home win, a draw or an away win. These odds are used to benchmark the predictions made by the models that are being tested in this paper against the predictions used by the bookmakers. The bookmakers' odds are not used in the estimation procedure.

3.1 Data Adjustments

The TV revenues of the football teams are obtained from www.fernsehergelder.com, as stated before. However, only the data from 2012/13 until 2018/19 were available on this website. There were no alternative data sources for this variable for the seasons 2008/09 until 2011/12. This variable is used in a lot of other variables as well. For example the variable *Market value - TV revenues*, obviously needs the variable TV revenues. Missing this variable would, thus, mean that almost half of all observations had to be discarded. This would have led to a major loss of information, which is needed to construct appropriate models. Therefore, the missing TV revenues are estimated by regressing on the available TV revenues. These estimates are then used in the data set.

Another adjustment that is made in the data set is about the previous season variables. As every year there are teams that get promoted and get relegated, some values of these variables are

unreliable to use. For example, if Hannover 96 plays in the 2. Bundesliga in season 2010/2011 and gets promoted to the 1. Bundesliga. Then the values, in season 2011/12, of for example *PS goals*, which reflects the total amount of goals scored the previous season, will be relatively high, because Hannover 96 was one of the top teams in the second division and consequently scored a lot of goals. This does not adequately reflect the amount of goals they would have scored when they would have been playing in the 1. Bundesliga. Therefore, the teams that got promoted, get the average value of the three lowest teams on the table of that previous season.

4 Methodology

The prediction of football games is a non-standard predictive task. This is the case, because of two reasons. Firstly, the outcome results of football matches are constructed in a goal difference or reflect whether the match ended in a win for the home or the away team, or as a draw. Because of this structure of the outcomes, a standard linear model can be expected to perform poorly. There are several reasons why one could suspect that a standard linear model would not be the most appropriate model in this case. Firstly, a standard linear model does not take into account the ordering of the data. The outcome of football matches clearly has a certain ordering, namely win, draw or lose. Besides that, a standard linear model assumes a continuous dependent variable, while the dependent variable in this paper only can take three, discrete values, corresponding to a win, draw or lose. The second reason why the prediction of football games is a non-standard predictive task is that it is in the nature of football that every game is heavily influenced by unobservable factors, such as luck or possible mistakes of the referees. Because of this uncertainty, it is needed to use prediction methods that fully utilize the available information. This can be used to expose the importance of uncertainty in the outcome of a match. The random forest algorithm is more flexible than a standard linear model, hence, one can expect it would perform better than a standard linear model. Therefore an ordered random forest algorithm is introduced in Section 4.1.

In Section 4.2 another prediction method is described, namely the BPRM. Furthermore, the estimation procedure for this method is explicitly discussed in Section 4.2.1. Finally, in Section 4.3 the construction of the league tables, using the results of the prediction methods, is explained, for both models.

4.1 Random Forests

Developed by Breiman (2001), random forests are used in all kinds of predictions problems. The algorithm of random forest works with a large amount of decorrelated, randomly built, decision trees. But why would one work with a lot of trees and not one?

The problem of using only one single regression tree is that it tend to have a high variance. This is due to the path-dependent structure of a decision tree. These regression trees (Breiman, 2017) choose a certain covariate space, which minimizes the sum of squares, at each split of the decision tree. The final prediction for input $\mathbf{x}$ is then obtained by averaging the amount of observations that end in the same end-node $L(\mathbf{x})$, which is called a leaf. If a decision tree has a lot of splits, it will

have a low bias on the one hand. On the other hand, it will have a high variance, because of the path-dependent structure of the tree. The so-called bagging resolves this problem by taking a lot of relatively small trees and averaging on all those low bias trees. To define this formally, we take the same notation as used in Breiman (1996), where a learning set $\mathcal{L} = \{(y_m, \mathbf{x}), m = 1, 2, 3\}$ is defined, and where y_m denotes the outcome of the match. In other words, y_m expresses whether a particular team wins, ties or loses a particular game, for given $\mathbf{x}$. Here $\mathbf{x}$ denotes the particular covariates used. Next, we assume that we have a predictor $\phi(\mathbf{x}, \mathcal{L})$, which predicts y if we put $\mathbf{x}$ in $\phi(\mathbf{x}, \mathcal{L})$. Suppose, we have a sequence of learning sets $\{\mathcal{L}_k\}$ that all consist of Q independent observations with the same distribution as $\mathcal{L}$. The goal is to get a better predictor than $\phi(\mathbf{x}, \mathcal{L})$, using only the sequence of predictors $\phi(\mathbf{x}, \mathcal{L}_k)$. In our case, as we have numerical outcomes y, we average $\phi(\mathbf{x}, \mathcal{L}_k)$ over k.

Next to bagging, random forest decorrelates the trees to achieve even a higher variance reduction (Hastie et al., 2009). This is done by taking only a random subset of the covariates at each split point in the tree. If we combine the bagging and the above, we obtain the algorithm that lies behind random forests. To summarize, the algorithm picks a bootstrapped sample b of size Q and constructs a regression tree $T_b(\mathbf{x})$ while choosing randomly p covariates out of the total of K covariates at each split point, where $p < K$. This is done until the minimum leaf size is reached. Hereafter, the final random forest estimate $RF^B(\mathbf{x})$ can be obtained. $RF^B(\mathbf{x})$ is defined as follows

$$RF^B(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} T_b(\mathbf{x}),$$

with $T_b(\mathbf{x})$ defined as

$$T_b(\mathbf{x}) = \frac{1}{|\{i : \mathbf{x}_i \in L(\mathbf{x})\}|} \sum_{\{i:\mathbf{x}_i \in L(\mathbf{x})\}} y_i.$$

Although, Random Forest have proved to be a method with a high predictive power, it does not take into account any potential ordered structure in the data. The outcomes of football matches do have an ordered outcome. Therefore, Goller et al. (2018) use the ORFM.

4.1.1 Ordered Random Forest Model

The ORFM is an extension of the conventional random forest. It does take into account the ordered structure within the data. Herewith, the potential loss of valuable information is prevented. The ORFM used in this paper, is developed by Lechner and Okasa (2018). This model explicitly incorporates the ordered structure of the outcomes. Because the basis of this model is a random forest algorithm, the model can deal with high-dimensional covariate spaces and provide some basic econometric output, for example marginal effects and outcome probabilities. This model can, thus, be seen as an alternative for the more traditional econometric models that take the ordered structure into account, such as the ordered probit model or ordered logit model. For the reader who is more interested in the ORFM, I would like to refer to Lechner and Okasa (2018). Their paper provides a thorough discussion of the estimator, the inference procedure and a simulation study. In the next paragraph the ORFM is explained in a more formal manner.

First, consider the ordered outcome variable $\Phi_i \in \{1, ..., M\}$ with m ordered categories. In this paper, $M = 3$, because the only possible outcomes can be a home win, a draw or an away win.

Given sample size N, with $n = 1, ..., N$, estimating the conditional ordered outcome probabilities evaluated at x, $P[\Phi_i = m|X_i = x]$, is based on an estimation of the cumulative probabilities given by binary indicators $Y_{m,i} = \underline{1}(Y_i \leqslant m)$ for $m = 1, ..., M-1$. Hereafter, a regression random forest is estimated for all the $M-1$ binary indicators, which gives the predictions $\hat{Y}_{m,i} = \hat{P}(Y_{m,i} = 1|X_i = x)$. Because the fact that the cumulative probabilities have to sum up to one, $\hat{Y}_{M,i}$ has to be equal to 1. After obtaining the cumulative probabilities, the probabilities of all M categories, for all N games are subsequently computed. For the first outcome category, the probability is defined as $\hat{P}^{tot}_{1,i} = \hat{Y}_{1,i}$. This probability is taken directly from the random forest estimation as in the case of the binary outcome. Here, the estimated conditional mean is a valid estimation of the probability of the first category. For the outcome probabilities $m = 2, ..., M$ the probabilities are computed as follows: $\hat{P}^{tot}_{m,i} = \hat{Y}_{m,i} - \hat{Y}_{m-1,i}$. The cumulative probabilities' nature is used for these categories, because it is possible to isolate the probability of the m-th category by subtracting the estimated probability of the preceding category. In the case that a estimated probability is negative, these probabilities are set to zero, i.e. $\hat{P}^{tot}_{m,i} = 0$ if $\hat{P}^{tot}_{m,i} < 0$. To ensure that all predictions sum up to one, the final step is to normalize all the probabilities. This is done in the following way: $\hat{P}_{m,i} = \frac{\hat{P}^{tot}_{m,i}}{\sum_{m=1}^{M} \hat{P}^{tot}_{m,i}}$. Here $\hat{P}_{m,i}$ denote the conditional ordered outcome probabilities, i.e. $\hat{P}^{tot}_{m,i} = \hat{P}[Y_i = m|X_i = x]$.

Note, that the ORFM makes use of linear combinations from the regression random forest probability estimates. Thus, if the regression random forest meets the conditions for normality and consistency, the ORFM will fulfill these conditions as well. Then, statistical inference can be conducted. Moreover, the ORFM, as described above, needs to estimate $M-1$ random forests in the training set. This seems a rather demanding task, but because the fact that our data only has 3 outcome categories and the existence of fast software implementations, the computation time is not a problem.

4.2 Bivariate Poisson Regression Model

As stated before, this paper compares the ORFM of Goller et al. (2018). with the BPRM of Karlis and Ntzoufras et al. (2005). This model is built to predict the amount of goals that two teams score against each other. Here, consider the random variable G_r for $r = 1, 2, 3$, that follow an independent Poisson distributions with parameters $\lambda_r > 0$. Now, the random variables $H = G_1 + G_3$ and $A = G_2 + G_3$ follow a joint, bivariate Poisson distribution. This gives the following joint probability function

$$P_{H,A}(h,a) = P(H = h, A = a) = \\ \exp(-(\lambda_1 + \lambda_2 + \lambda_3))\frac{\lambda_1^h}{h!}\frac{\lambda_2^a}{a!}\sum_{q=0}^{min(h,a)} \binom{h}{q}\binom{a}{q} q! \left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^q. \tag{1}$$

Here, H and A represent the amount of goals scored by the two teams. The bivariate Poisson distribution allows for dependence between H and A. Marginally, H and A follow a univariate Poisson distribution. More formally, $E[H] = \lambda_1 + \lambda_3$ and $E[A] = \lambda_2 + \lambda_3$. The assumed dependence between H and A is expressed by $\text{Cov}(H, A) = \lambda_3$. Note that if $\lambda_3 = 0$, the two random variables are independent and, thus, this distribution would reduce to the product of two independent Poisson distributions.

All the parameters λ_r, for $r = 1, 2, 3$, in (1) can incorporate covariates by specifying an adequate response function. Specifically, $\lambda_r = (\lambda_{r1}, ..., \lambda_{rn})$ contains all λs for every observation. I follow Groll et al. (2018) and use $\lambda_r = \exp(\eta_r)$, with the linear predictor $\eta_r = \beta_{0r} + c_r^T \beta_r$ and response function $h(\cdot) = \exp(\cdot)$, such that λ_r becomes a non-negative Poisson parameter. Moreover, $c_r = (c_{1r}, ..., c_{pr})^T$ contains all the covariates of predictor r.

For football data, a logical way of modelling the three parameters λ_r, $r = 1, 2, 3$, is to let λ_1 and λ_2 contain the covariate information of the competing teams 1 and 2, respectively. λ_3 contains the covariate information reflecting the match conditions, which are the same for the competing teams, obviously. The model representation becomes

$$\lambda_1 = \exp(\beta_{10} + c_1^T \beta_1), \quad \lambda_2 = \exp(\beta_{20} + c_2^T \beta_2), \tag{2}$$

where c_1 and c_2 contain the covariate information of team 1 and team 2, respectively. Covariance parameter λ_3 generally depends on different covariates and effects and is modelled as follows

$$\lambda_3 = \exp(\alpha_0 + z^T \alpha), \tag{3}$$

with z that could contain certain covariates of c_1 and c_2 or completely new covariates.

Note that, for λ_1 and λ_2, in (2), there is no common intercept. The intuition behind the intercept in this model is that it represents the goals scored, without any further information. I assume there exists a so-called "homFe-advantage" and a "away-disadvantage". In other words, β_{01} represents the home-advantage and β_{02} represents the away-disadvantage, in a particular game.

4.2.1 Estimation

The BPRM does not offer a direct estimate for the parameters λ_1, λ_2 and λ_3. Therefore, a numerical approach is chosen to obtain maximum likelihood estimates for these parameters. There are a lot of different ways to estimate the parameters of the BPRM. For example, Gourieroux et al. (1984) derived a pseudo maximum likelihood estimation method and Kocherlakota and Kocherlakota (2001) estimated the parameters using a Newton-Raphson procedure. In this paper, the Expectation Maximization (EM) algorithm is used to estimate the parameters which are needed in 1. This method is also used in the paper of Karlis and Ntzoufras (2003). For the BPRM, a trivariate reduction derivation of the bivariate Poisson distribution is necessary to construct the EM algorithm. Suppose, for the i-th observation G_{1i}, G_{2i}, G_{3i} correspond to non-observable data. The observable data is represented as $H_i = G_{1i} + G_{3i}$ and $A_i = G_{2i} + G_{3i}$. The estimation would have been simple if the unobservable data was available. The only thing to do was to fit Poisson regression models to G_1, G_2 and G_3. To resolve this problem caused by the unobservables and to be able to construct the EM algorithm, the unobservable data is estimated by their conditional expectations. After this Poisson regression models are fitted to the pseudo-values obtained by the E-step. Let γ denote the vector containing the parameters to be estimated, that is $\gamma = (\beta_1, \beta_2, \alpha)'$. Then, the complete data log-likelihood is given by

$$L(\gamma) = -\sum_{i=1}^{n}\sum_{r=1}^{3} \lambda_{ri} + \sum_{i=1}^{n}\sum_{r=1}^{3} x_{ri} \log(\lambda_{ri}) - \sum_{i=1}^{n}\sum_{r=1}^{3} \log(g_{ri}!),$$

where the λs are given by (2) and (3).

Now, the EM algorithm for the bivariate Poisson model is given by

E-step: Calculate the conditional expected values of G_{3i}, using the current parameter values of k iteration, denoted by $\gamma^{(k)}, \lambda_{1i}^{(k)}, \lambda_{2i}^{(k)}$ and $\lambda_{3i}^{(k)}$, for $i = 1, ..., n$, by

$$s_i = E\big(G_{3i}|H_i, A_i, \gamma^{(k)}\big) = \begin{cases} \lambda_{3i}^{(k)} \frac{P_{H,A}(h_i-1, a_i-1|\lambda_{1i}^{(k)}, \lambda_{2i}^{(k)}, \lambda_{3i}^{(k)})}{P_{H,A}(h_i, a_i|\lambda_{1i}^{(k)}, \lambda_{2i}^{(k)}, \lambda_{3i}^{(k)})} & \text{if } \min(h_i, a_i) > 0 \\ 0 & \text{if } \min(h_i, a_i) = 0 \end{cases}$$

where $P_{H,A}(h, a|\lambda_1, \lambda_2, \lambda_3)$ is given in (1).

M-step: Update the estimates by

$$\begin{aligned} \beta_1^{(k+1)} &= \hat{\beta}(h - s, c_1), \\ \beta_2^{(k+1)} &= \hat{\beta}(a - s, c_2), \\ \alpha^{(k+1)} &= \hat{\alpha}(s, z), \\ \lambda_{ri}^{(k+1)} &= \exp\left(c_{ri}' \hat{\beta}_r^{(k+1)}\right) \quad \text{for} \quad r = 1, 2. \\ \lambda_{3i}^{(k+1)} &= \exp\left(z_i' \hat{\alpha}^{(k+1)}\right) \end{aligned}$$

Here, $s = (s_1, ..., s_n)'$ is the $n \times 1$ vector from the E-step, $\hat{\beta}(h, C)$ are the maximum likelihood estimates of a Poisson model with response the vector h and data matrix $C = (c_1, c_2, z)'$. Each data matrix C_r is a $n \times p_r$ matrix, where p_r is the amount of covariates for $r = 1, 2, 3$.

The **bivpois** package is used in this paper to implement the above described estimation algorithm in R. The package is available from the authors' web page at `http://www.stat-athens.aueb.gr/~jbn/papers/paper14.htm`.

4.3 League Outcomes

In this Section it will be explained how the league tables are constructed. Besides the construction of the league tables, other performance measures will be introduced, which will be used to assess the predictive power of the models.

4.3.1 Ordered Random Forest Model

Once we obtained the probabilities for a win, draw or a loss for all the teams using the ORFM, one can use these probabilities to derive a final league table. I follow the paper of Goller et al. (2018) and use two approaches. Firstly, a logical way of using the estimated probabilities is to compute the expected points (3 for a win, 1 for a draw and 0 for a loss) according to the probabilities of each team in each game. If all these points are summed up, the expected final league table is obtained. Secondly, one may take into account the uncertainty that every game has. To account for this, the points in every game are derived by a random draw of a simulated outcome based on the estimated probabilities. The realization of this random variable determines the amount of points a team gets for a particular game. Again, adding up all the points from all the games will result in a final league

table. This is repeated 10.000 times, so that it becomes feasible to compute the probability to become champions, to get relegated or to reach the play-offs for European football. For example, to compute the probability that Borussia Dortmund will win the league, one divides the total amount of times Borussia Dortmund ended on top of the table by the total amount of simulations. All other places in the table are derived in the same manner.

After the final league table is derived, one can examine the predictive performances of this method. This can be done by Spearman's rank coefficient, the root mean squared error and a hypothetical ROI. The latter one is used to compare the methods to the betting odds of the bookmakers.

4.3.2 Bivariate Poisson Regression Model

The bivariate Poisson model, as discussed in Section 4.2, is used to compute the two distributions of the scores of the two teams, for each match. The match result is drawn randomly from the predicted distributions. In other words, the scores are drawn from $G_1 \sim Poisson(\hat{\lambda}_1 + \hat{\lambda}_3)$ and $G_2 \sim Poisson(\hat{\lambda}_2 + \hat{\lambda}_3)$, where $\hat{\lambda}_1$, $\hat{\lambda}_2$ and $\hat{\lambda}_3$ are obtained by the regression model, explained in Section 4.2. If $G_1 > G_2$, the home team has scored more goals than the away team and, thus, the home team wins; if $G_1 = G_2$, it means that both teams scored the same amount of goals, which means that the game ended as a draw; if $G_1 < G_2$ the away team scored more than the home team, resulting in a win for the away team. Again, the winning team receives 3 points, 1 point for a draw and 0 points for a loss. Hereafter, the seasons will be simulated 10.000 times. Based on these simulations, for each team the probability of reaching a certain place at the table will be calculated.

To asses the predictive performance of this model the odds of the bookmakers are used. These odds are used to compute three quantities $\tilde{p}_r = 1/\text{odds}_r, r \in 1, 2, 3$. Here, $\tilde{p}_1$ represents the odds of a home win, $\tilde{p}_2$ the odds of a draw and $\tilde{p}_3$ the odds of a away win. This is normalized with $d := \sum_{r=1}^{3} \tilde{p}_r$, so that the margins of the bookmakers are being account for. The odds can be transformed into probabilities using $\hat{p}_r = \frac{\tilde{p}_r}{d}$. These transformed probabilities serve as an approximation for the bookmakers' probabilities for a home win, an away win or a draw. This approximation is based on the assumption that the margins of the bookmakers follow a discrete uniform distribution on the three possible match results. The true match outcomes $\omega_g \in \{1, 2, 3\}, g = 1, ..., G$, with G the total amount of games, are used to compute $\bar{p}_{three-way} := \frac{1}{G} \sum_{g=1}^{G} \hat{p}_{1g}^{\delta_{1\omega_g}} \hat{p}_{2g}^{\delta_{2\omega_g}} \hat{p}_{3g}^{\delta_{3\omega_g}}$, where δ_{rg}, denotes Kronecker's delta. $\bar{p}_{three-way}$ serves as a good performance measure of the model compared to the bookmakers' odds.

5 Results

This Section contains the results of the two methods introduced in Section 4. First, the results of the ORFM are presented. Hereafter, the results of the BPRM are displayed. Finally, the two methods will be compared by the hypothetical ROI and the $\bar{p}_{three-way}$, which are described in Section 4.1.1 and 4.3.2, respectively.

5.1 Ordered Random Forest Model

To asses the predictive power of the Ordered Random Forest model, the season of 2017/18 is predicted. The training data consisted of the data from the season 2008/09 until 2016/17. The training data is used to compute the probabilities for a home win, a draw or an away win in the season 2017/2018. The algorithm used, is described in Section 4.1.1. The expected points have been gathered and used to construct a prediction of the league table for the 1. Bundesliga of the season 2017/18. Table 1 shows the predicted and the actual league table. The same is done for the season 2018/2019, with training data until the end of the season 2017/2018. The results are shown in Table 9 in the Appendix. Note that Table 9 is not the final league table, but the table after match day 33.

Table 1: The predicted and the actual final table of the 1. Bundesliga season 2017/18 estimated with the ORFM

	Rank		Points	
Team	Predicted	Actual	Predicted	Actual
FC Bayern München	1	1	75.3	84
Borussia Dortmund	2	4	62.0	55
RB Leipzig	3	6	55.0	53
Schalke 04	4	2	52.0	63
Bayer 04 Leverkusen	5	5	51.3	55
Borussia M'gladbach	6	9	51.3	47
TSG Hoffenheim	7	3	48.4	55
VfL Wolfsburg	8	16	46.2	33
Hertha BSC Berlin	9	9	45.7	43
1. FC Köln	10	18	45.0	22
Werder Bremen	11	11	41.9	42
SC Freiburg	12	15	41.9	36
1. FSV Mainz 05	13	14	41.3	36
FC Augsburg	14	12	41.1	41
Eintracht Frankfurt	15	8	40.9	49
Hannover 96	16	13	40.8	39
VfB Stuttgart	17	7	40.2	51
Hamburger SV	18	17	39.4	31

As football results depend for a huge amount on luck and other unpredictable factors, there are surprises in the league table, in a positive and a negative way. For example, in Table 1 it can be seen that Eintracht Frankfurt was predicted to end at the fifteenth place. However, they had a strong season and ended at the eighth place. At the same time, VfL Wolfsburg performed relatively poor, as they had a predicted rank of 8, but a realized rank of 16. These differences between the predicted and actual league table are not really surprising, because it is very difficult to capture all the in-season developments of a team.

To see whether the ORFM did a good job in predicting the final league table, the predictions of Table 1 are compared with other predictions of the final ranking and points. Twelve alternative strategies to predict the final league table and points are retrieved from www.bstat.de. Here, different forecasts of experts or algorithms are collected before each season.

To formally compare the different prediction methods, the Spearman's rank correlation coefficient and the root mean squared error (RMSE) are considered. The Spearman's rank correlation coefficient

is computed using the actual table and the predicted table. This coefficient represents the amount of correlation between the actual and predicted final league table. Thus, the higher the coefficient, the closer the prediction is to the actual table. The RMSE is used to asses the accuracy of the predicted points. In Table 2 the results of the ORFM and the twelve other predictors are shown.

Table 2: Comparison of different predictions for the final league table of the 1. Bundesliga season 2017/18

	Rank Correlation	RSME
Ordered Random Forest	0.65	8.7
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	-

In terms of Spearman's rank correlation, three other predictions showed to perform better than the ORFM. These are the predictions of the experts of the newspapers General-Anzeiger and Spiegel Online. Besides those, the algorithmic prediction of Goalimpact showed to have a higher Spearman's rank correlation. On the other hand, in terms of predicting the amount of points that each team won, the ORFM has the smallest RSME of all predictors. This means that the ORFM performed the best in predicting the amount of points won.

The ORFM calculates for every game a certain probability for a home win, a draw or an away win. These probabilities are compared with the betting odds of five major bookmakers. To asses the performance of the ORFM, different betting strategies are used and a hypothetical ROI is calculated for each strategy. The first strategy that is considered, is the so-called *proportional strategy*. In this strategy, 1 euro is bet on each game. This euro is then split according to the estimated probabilities of the ORFM for each outcome. To clarify, if the estimated probabilities are 50 % for a home win and 25% for a draw and an away win, 50 cents are bet on a home win and 25 cents are bet both on a draw and an away win. To check whether we would win or lose money, assume that the home team wins and the betting odds are 1.9, 2.0 or 2.1. In the first case, we would lose money, because we spent €1,- and get 1.9*€0.5 = €0.95. For the second betting odds, we would not win nor lose money, because 2*€0.5 = €1,-. Finally, for the third case, we would win money, because 2.1*€0.5 = €1.05. The probability that is implied by the betting odds of 2.1 is $\frac{1}{2.1} = 47.6\%$. As our estimated probability was 50%, a ROI is achieved of $\frac{1.05-1}{1} = 5\%$. In the first three columns of Table 3 the ROIs of this strategy for different time frames is shown. In the first column the ROIs for season 2017/18 is shown, the second column displays the ROIs for the season 2018/19 until the 33rd match day, while the third column contains the ROIs if we combine both seasons.

Table 3: Return of investment in percent of different bookmakers in different seasons

	Odds			Odds net of fees		
	2017/18	2018/19	2017/2019	2017/18	2018/19	2017/2019
B365	-5.7	-7.5	-6.7	-0.8	-2.6	-1.7
Bwin	-5.8	-7.4	-6.6	-0.9	-2.7	-1.8
Interwetten	-7.1	-2.4	-4.8	-1.0	-2.4	-1.7
Ladbrockes	-3.3	-7.9	-5.6	-1.1	-2.8	-1.9
William Hill	-5.0	*	*	-1.3	*	*
Value bet:	-13.2	-17.4	-11.3			

For the season 2017/18 the ROIs lie between -3% and -7%. For the next season the results are similar, showing negative ROIs in between -2% and -7%. In the third and sixth column season 2017/18 and 2018/19 are combined, which means that all games of both seasons are used to compute the ROIs. When both seasons are combined, the results lie approximately in the middle of the two former columns for all bookmakers, respectively. The odds for the season 2018/19 for the bookmaker William Hill are not shown, because the data was not correct for this year. This would lead to biased and inadequate ROIs, which would make the results less credible. Therefore, the value, in the column where the two seasons are combined, is also not shown. Table 3 shows that the ORFM does not outperform any bookmaker, for all time frames.

There are two possible explanations why the bookmakers odds outperform the ORFM. First, the bookmakers' odds most likely reflect short-term developments as well, while the ORFM does not have access to this kind of information. Examples of short-time developments can be injuries or other factors that could influence the players and/or the staff. Second, the implied probabilities of the odds of the bookmakers imply a fee for the bookmaker, because the implied probabilities do not count up to one. However, it is unknown how the fees are distributed over the outcomes (for a discussion on this, see e.g. Levitt (2004), Paul and Weinbach (2007), Paul and Weihnbach (2012)). I assume, like in Goller et al. (2018), that the fees are distributed proportionally. This would create odds that are "net of fees" in the following manner. The bookmakers' odds are inverted to get the implied probabilities. Hereafter, these implied probabilities are normalized to make sure that they sum up to one. Finally, the normalized probabilities are inverted again to obtain the "normalized" bookmakers' odds. By doing this, a comparison between the probabilities obtained from the ORFM and from the bookmakers' odds is more fair. These results are displayed in the last three columns of Table 3. There are two remarkable observations to be made when looking at these results. Namely, all the ROIs, for all different time frames, have decreased substantially relative to the corresponding columns where the "conventional" odds were used. For the season 2017/18 all ROIs changed to approximately -1%. Furthermore, the variation between the different bookmakers has reduced, making them very similar. This may indicate that the differences, in the first three columns, are mainly being caused by the different fees that the bookmakers charge.

The second strategy that is being investigated is the so-called *value bet strategy*. In this strategy one will only bet on a certain game when the estimated probabilities are higher than the one implied by the bookmaker's odds. For example, if you have an estimated probability of a home win of 50%. When the betting odds are lower than 2, it would mean that we would lose money in the long-run,

even when we would know that the true probability is 50%. This may happen because of the implicit fee that is being charged or due to a too high probability of a home win (estimated by the bookmaker). It could happen that there are multiple outcomes that have a higher probability than the probabilities implied by the bookmakers or that this is the case for multiple bookmakers. In this case, the best bookmaker-outcome combination is chosen by picking the highest ratio of the ORFM probabilities and the implied probabilities of the bookmakers. The losses using this strategy lie between 17% and 11%. Note that for the values of the *value bet strategy* in the season 2018/19 and the combination of seasons 2017/18 and 2018/19 (under the header 2017/19 in Table 3), the odds of William Hill are not considered. As stated before, the odds of William Hill are not correct for the season 2018/19, therefore the hypothetical ROIs are not reliable and, thus, not presented for the columns that contain this season.

In Section 4.3.1 it is described how the probabilities are derived that a certain team ends at a particular place on the league table. A team can have a goal in the season, for example qualifying for the Champions League (rank 2-4) or to qualify for the Europa League (rank 5-6). The probabilities for certain ranks are therefore aggregated to represent the probability to achieve a certain goal. The probabilities are given in Table 4. The left part of Table 8 represents the probabilities that are computed before the season, so when all teams still have 0 points. The right part of the table contains the probabilities computed after match day 8, so the points won thus far are included in computing the probabilities. I follow Goller et al. (2018) in my chose for match day 8, however every match day could have been chosen to show the change in probabilities within the season, when the materialized results thus far are incorporated. Table 4 shows that if you incorporate the results, the probabilities will change. For example, Hertha BSC had a great start of the season in the season 2018/19, with a shared fifth place with RB Leipzig. Consequently, one can see that the probability that they reach Champions League places (rank 2-4) rose from 16% to 25%. For teams that start the season badly, the probabilities change as well. Schalke 04, for instance, is usually a team that plays for the Champions League spots in the 1. Bundesliga. However, after the eighth match day, Schalke 04 is at the sixteenth place with only 6 points. Because of this, the probability of Schalke 04 to reach the Champions League places has reduced from 17% to only 3%.

Table 4: Probabilities in percentages, obtained with the Ordered Random Forest model, to achieve certain season goals in the 1. Bundesliga season 2018/19

	Before season start						After match day 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	78	20	1				62	37	1			
Dortmund	11	62	14	12			30	64	5	2		
RB Leipzig	3	39	20	36			2	47	25	25		
Leverkusen	2	36	21	39	1		14	23	61	1	1	
Gladbach	3	34	20	40	1	1	5	57	21	18		
Hoffenheim		20	17	56	3	3		12	20	64	2	2
Schalke 04		17	17	58	3	5		3	7	75	6	9
Hertha		16	16	59	4	5	1	25	26	47	1	1
E. Frankfurt		14	14	61	4	6		15	21	60	2	2
Wolfsburg		11	12	62	5	9		4	10	74	5	8
Bremen		7	10	65	7	11		17	23	58	1	1
Stuttgart		6	9	64	8	12		1	3	62	12	23
Hannover		6	9	64	8	14		1	3	64	12	20
Augsburg		3	6	61	10	20		1	5	69	10	15
Mainz		3	5	63	10	19		1	3	70	10	16
Freiburg		2	4	55	12	28		1	2	62	12	23
Düsseldorf		2	3	53	12	30			1	40	14	46
Nürnberg		1	3	49	12	36			1	51	13	35

The percentages below 1 are not reported in this table. Furthermore, the displayed percentages have been rounded to whole numbers.

Note that these probabilities are obtained by aggregating all the points until match day 33, this is been taken as the final league table. This is done, because the last match day was missing in the data set. However, because the algorithm has run 10.000 times and the fact that only one round is missing, the estimated probabilities for each rank are reliable.

5.2 Bivariate Poisson Regression Model

In this Section the results, obtained using the BPRM, will be showed. The first step in the process was the estimation of the coefficients in (2) and (3), so that $\hat{\lambda}_r$ for $r = 1, 2, 3$ was obtained. This is done by the EM algorithm, described in Section 4.2.1. In contrast to the findings of Groll et al. (2018), $\hat{\lambda}_3$ showed to be nonzero and significant. This means that there is a certain covariance between the scores of both teams, although investigation of $\hat{\lambda}_3$ shows that for most observations $\hat{\lambda}_3 \approx 0$. Note that if $\hat{\lambda}_3 = 0$, one would obtain two (conditionally) independent univariate Poisson models. However, the Likelihood-Ratio (LR) test shows that the BPRM and the product of two (conditionally) independent univariate Poisson models differ significantly. More formally, $LR = 2l(\hat{\theta}_1) - 2l(\hat{\theta}_0) = 2(-9945.913 + 9980.576) = 69.326$, where $l(\hat{\theta}_1) = -9945.913$ represents the log-likelihood of the BPRM and $l(\hat{\theta}_0) = -9980.576$ corresponds to the log-likelihood of the product of two (conditionally) independent univariate Poisson models. Asymptotically, the LR test follows a $\chi^2(38) \approx 56$ distribution. The degree of freedoms in this distribution is 38, because there are 37 variables and a intercept used in the estimation of λ_3. Hence, because $LR \approx 70 > 56$, the hypothesis that $\hat{\lambda}_3 = 0$ is rejected. The fact that this paper differs on this point to the paper of Groll et al. (2018) is explainable. Groll et al. (2018) used European championships to test their model. In these

matches the "home-advantage" and "away-disadvantage" is less influential than in regular matches in a national league, because at a international tournament matches are played at a "neutral" ground, except for the country that hosts the tournament. In a regular season the home-advantage and away-disadvantage may be more present, because teams do play at home or away.

The estimated parameters are then used in a univariate Poisson distribution to determine the estimated amount of goals scored in a particular game. More formally, $G_1 \sim Poisson(\hat{\lambda}_1 + \hat{\lambda}_3)$ and $G_2 \sim Poisson(\hat{\lambda}_2 + \hat{\lambda}_3)$, with G_1 and G_2 the amount of goals scored by the home team and the away team, respectively.

The advantage of this method, compared to, for example the ORFM, is that the exact match outcomes is drawn for every match. The season 2018/19 has been simulated 10.000 times. Based on these simulations, the probabilities to reach a certain goal in this season is obtained. Note that, again, Table 5 presents the probabilities of being at that certain place after match day 33, due to the fact that the data of the last match day is not available in the data set.

Table 5: Probabilities in percentages, obtained with the bivariate Poisson regression model, to achieve certain season goals in the 1. Bundesliga season 2018/19

	Before start season						After match day 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	84	16					67	33				
Dortmund	10	82	6	1			28	70	1			
RB Leipzig	5	79	13	4			4	80	14	3		
Leverkusen	1	59	27	13	1			27	41	32		
Gladbach		30	36	34			1	62	26	10		
Hertha		21	10	76	5	5		6	21	72	1	
Schalke 04		8	23	67	1	1		1	8	86	3	2
Wolfsburg		7	22	68	2	1		3	13	82	2	1
E. Frankfurt		5	15	74	4	3		6	21	71	1	1
Hoffenheim		5	15	74	3	3		3	13	81	2	1
Bremen		4	15	76	3	3		10	31	58		
Mainz		1	6	78	7	8			3	83	7	6
Augsburg		1	5	77	7	9			4	84	6	6
Stuttgart		1	4	73	9	12			1	64	14	20
Freiburg			2	63	13	22			1	70	13	16
Hannover			1	51	15	33				44	19	37
Nürnberg				35	16	49			1	38	18	44
Düsseldorf				34	15	50				21	14	65

The percentages below 1 are not reported in this table. Furthermore, the displayed probabilities have been rounded to whole numbers.

Table 5 shows the same structure as Table 4. For example, in 84% of the cases, FC Bayern Munich becomes champions if we predict the season from the first match day. This probability slinks, however, if we incorporate the first 8 match days, to 67%. This is because of the fact that FC Bayern Munich had a relative bad start of the season, while for example Borussia Dortmund had an excellent start of the season, increasing their probability to become champion from 10% to 28%.

As stated in Section 4.3.2 the predictive power of the BPRM is assessed by the quantity $\bar{p}_{three-way}$. This quantity is compared with the predictive power of the bookmakers' odds and also the $\bar{p}_{three-way}$ of the ORFM. In Table 6 it is shown that the BPRM has approximately an equal predictive power

as the bookmakers' odds.

Table 6: $\bar{p}_{three-way}$ of five bookmakers' odds and the bivariate Poisson regression model

	$\bar{p}_{three-way}$		
	2017/2018	2018/19	2017/2019
B365	40.52	42.44	41.47
Bwin	40.51	42.29	41.38
Interwetten	40.38	42.33	41.34
Ladbrockes	40.74	42.31	41.52
William Hill	40.54	*	*
ORFM	36.56	36.96	36.76
BPRM	40.65	42.30	41.46

The numbers reported in this table are in percentages.

Furthermore, it outperforms the ORFM in all time spans. One may conclude from this that the BPRM has a higher predictive power than the ORFM. In addition, the BPRM performs better relative to the bookmakers' odds than the ORFM, which is outperformed at all time spans by every bookmakers' odds. It is worth to note that this is a remarkable result, taking into account that the bookmakers' odds are usually only released a couple of days before a match, hence, containing the most recent information regarding the teams.

To see whether the estimated probabilities obtained from the BPRM could help one with betting, a hypothetical ROI is computed. This ROI is computed using the same betting strategy as was used for the ORFM. This means that for the first three columns of Table 7 the *proportional strategy* has been used. The last three columns of Table 7 report the hypothetical ROIs where first the odds of the bookmakers have been corrected for their implicit fees. Lastly, the last row of Table 7 shows us the ROIs when the *value-bet strategy* is used.

Table 7: Return of investment in percent of different bookmakers in different seasons

	Odds			Odds net of fees		
	2017/18	2018/19	2017/2019	2017/18	2018/19	2017/2019
B365	-5.4	-5.0	-5.2	-0.5	0.1	-0.2
Bwin	-5.7	-4.7	-5.2	-0.7	0.2	-0.3
Interwetten	-6.9	0.9	-3.0	0.3	-2.4	-0.3
Ladbrockes	-3.0	-5.3	-4.1	0.0	-2.8	-0.4
William Hill	-4.6	*	*	-0.9	*	*
Value bet:	-5.4	-16.3	-10.8			

If the resutls of Table 3 and Table 7 are compared, one can see that the ROIs obtained using the BPRM are a bit less negative. This means that this model slightly outperforms the ORFM in terms of the hypothetical ROI. However, the main structure within and relationships between the columns are the same for both tables. In addition, Table 7 shows that in general the BPRM, does not outperform the bookmakers. The same holds for the OFRM.

6 Conclusion

This paper consists of two parts. The first part of the paper is an attempt to replicate the paper of Goller et al. (2018). Here, the ORFM is implemented to obtain predictions of the outcome of sport leagues. This is done in a probabilistic manner. Thanks to the particular algorithm of the ORFM, predictions could be made for every game in the season. As expected, the same conclusions are drawn in this paper, as in the paper of Goller et al. (2018), although the data in both papers was different. The data set which is used, is obtained using different sites. The retrieving of the data appeared to be a rather demanding job. Because of the limited time available for this paper not all the variables used in Goller et al. (2018) have been used. However, I think the most important variables have been used in this paper, making the model strong enough to obtain similar and reliable results.

The ORFM showed to be a decent prediction model. First of all, the predicted final league table showed to be a good estimation compared to other predictions of the final league table made by experts or algorithms. In terms of the Spearman's rank correlation coefficient only three other predictors performed better. Moreover, in terms of the RSME, the ORFM performed best of all predictors. However, the ORFM did not outperform the bookmakers' odds, but the model was close, especially in the case that the implicit fee for the bookmakers was taken into account. It is important to note that the bookmakers use a much more up-to-date information set, compared to the ORFM. This could be an explanation for the fact that the ORFM is outperformed by the bookmakers, in every case.

In the second part of the paper, another method is used to predict the final league table in the 1. Bundesliga. The BPRM showed to predict a similar outcome of the 1. Bundesliga after 33 match days, compared to the prediction made by the ORFM. Using an EM algorithm the estimates for $\hat{\lambda}_r$ for $r = 1, 2, 3$ are obtained. Hereafter, these estimates are used to predict the amount of goals by the two competing teams, resulting in a particular match outcome. It was interesting to see how the two different models differ from each other. An advantage of the BPRM is that in this model also the amount of goals scored and conceded is predicted. This could improve the prediction of the final league table, relative to the prediction made by the ORFM. For example, if two teams have the same amount of points, their ranking is based on the amount of goals scored minus the amount of goals conceded. The ORFM is not capable of estimating the amount of goals scored per game and, thus, cannot decide which teams has to be higher in this kind of situations. Furthermore, two other performance measures were used to compare the two methods. As stated before, the ORFM has been compared to the bookmakers' odds. This is also done for the BPRM, showing slightly better results. However, the BPRM is also not able to outperform the bookmakers, just as the ORFM. Next to this measure, the prediction power of both models is assessed by the quantity $\bar{p}_{three-way}$, which is described in Section 4.3.2. This is also a useful performance measure for a comparison between the predictive power of a model and the bookmakers' odds, but also between the models. In Table 6 it is shown that the BPRM does not outperform nor underperform the bookmakers' odds. This is remarkable, because, again, the bookmakers' odds contain more recent information compared to the BPRM. Besides that, the BPRM does outperform the ORFM in all time frames by approximately 4%. If everything is taken into account, the advantage of obtaining the exact match results, having a

less negative hypothetical ROI for all bookmakers, in all time spans and the higher $\bar{p}_{three-way}$ result in a minor preference for the BPRM over the ORFM.

7 Limitations and Further Research

The models in this paper use a long list of variables that may or may not contribute to the predictive power of the models. Goller et al. (2018) use a data set containing around 300 variables. In this paper, more or less 90 variables are used. Not all variables have been retrieved, because of a limited time span in which this paper had to be written. Although the missing variables, both the OFRM and the BPRM worked quite well. It would be interesting to see how the models would perform with the same data set that is used in Goller et al. (2018).

Furthermore, the models do not limit themselves to only work for football matches. The ORFM works for all sports where you can win, draw or lose, while the BPRM needs sports where the goal is to score more points than the other team. There are several sports that meet these conditions. Therefore, research to further investigate the predictive performance of both models in other sports or in other leagues can be interesting.

8 Acknowledgment

I would like to thank MSc Nienke Dijkstra for her help and advice during the whole process of writing this paper.

9 References

Baboota, R, & Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. International Journal of Forecasting, 35(2), 741-755.

Baio, G., & Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football results. Journal of Applied Statistics, 37(2), 253-264.

Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Breiman, L. (2017). Classification and regression trees. Routledge.

Crowder, M., Dixon, M., Ledford, A., & Robinson, M. (2002). Dynamic modelling and prediction of English Football League matches for betting. Journal of the Royal Statistical Society: Series D (The Statistician), 51(2), 157-168.

Dixon, M. J., & Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. Journal of the Royal Statistical Society: Series C (Applied Statistics), 46(2), 265-280.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?. The Journal of Machine Learning Research, 15(1), 3133-3181.

Freund, Y., & Schapire, R. E. (1996, July). Experiments with a new boosting algorithm. In icml (Vol. 96, pp. 148-156).

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics, 28(2), 337-407.

Gourieroux, C., Monfort, A., & Trognon, A. (1984). Pseudo maximum likelihood methods: applications to Poisson models. Econometrica: Journal of the Econometric Society, 701-720.

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.

Goller, D., Knaus, M. C., Lechner, M., & Okasa, G. (2018). Predicting Match Outcomes in Football by an Ordered Forest Estimator (No. 1811). University of St. Gallen, School of Economics and Political Science.

Joseph, A., Fenton, N. E., & Neil, M. (2006). Predicting football results using Bayesian nets and other machine learning techniques. Knowledge-Based Systems, 19(7), 544-553.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction, Springer Series in Statistics.

Hofner, B., Mayr, A., & Schmid, M. (2014). gamboostLSS: An R package for model building and variable selection in the GAMLSS framework. arXiv preprint arXiv:1407.1774.

Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. Journal of the Royal Statistical Society: Series D (The Statistician), 52(3), 381-393.

Karlis, D., & Ntzoufras, I. (2005). Bivariate Poisson and diagonal inflated bivariate Poisson regression models in R. Journal of Statistical Software, 14(10), 1-36. Kocherlakota, S., & Kocherlakota, K. (2001). Regression in the bivariate Poisson distribution.

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, 26(3), 471-481.

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

Levitt, S. D. (2004). Why are gambling markets organised so differently from financial markets?. The Economic Journal, 114(495), 223-246.

Mayr, A., Fenske, N., Hofner, B., Kneib, T., & Schmid, M. (2012). Generalized additive models for location, scale and shape for high dimensional data—a flexible approach based on boosting. Journal of the Royal Statistical Society: Series C (Applied Statistics), 61(3), 403-427.

Mayr, A., Binder, H., Gefeller, O., & Schmid, M. (2014a). The evolution of boosting algorithms. Methods of information in medicine, 53(06), 419-427.

Mayr, A., Binder, H., Gefeller, O., & Schmid, M. (2014b). Extending statistical boosting. Methods of information in medicine, 53(06), 428-435.

Min, B., Kim, J., Choe, C., Eom, H., & McKay, R. B. (2008). A compound framework for sports

results prediction: A football case study. Knowledge-Based Systems, 21(7), 551-562.

Paul, R. J., & Weinbach, A. P. (2012). Does sportsbook. com set pointspreads to maximize profits?. The Journal of Prediction Markets, 1(3), 209-218.

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, 3(3), 137.

Schmid, M., & Hothorn, T. (2008). Boosting additive models using component-wise P-splines. Computational Statistics & Data Analysis, 53(2), 298-311. Schmid, M., Potapov, S., Pfahlberg, A., & Hothorn, T. (2010). Estimation and regularization techniques for regression models with multidimensional prediction functions. Statistics and Computing, 20(2), 139-150.

10 Appendix

10.1 Data

As mentioned in Section 3, variables are used that represent team characteristics, match, schedule, economic and location related information. The data set consists of "home", "away" and "home-away" variables, which numbers concern the home team, the away team or the difference between the home and away team, respectively.

The first variables that are presented are the team characteristics, which are captured in a lot of different variables. For example, the wealth of different clubs is captured in *TV Revenue*, which denotes the amount of TV revenue a certain club received in a particular season, in millions. The variable *market value* denotes the market value of a club. Furthermore, there are several variables that combine the two above variables. Other team characteristics such as the inequality within a team is measured as the ratio of Top 3 (11) most valuable players to the market value of the players ranked 12-14 (12-21). Also, the diversity in a team is represented by several age related variables, such as the mean, minimum, maximum or the standard deviation of the age in a team. Other examples are, the share of left-footed players in a team, etc. Finally, there are three variables that classify the different clubs. *traditional club* indicates teams with a rich history, for example Borussia Dortmund and Bayern Munich. *yo-yo club* indicates a team that gets relegated and/or promoted from/to the 1. Bundesliga. *Other clubs* denote all other clubs that are neither traditional, nor yo-yo clubs.

The next set of variables are variables that contain information about the previous played games, for each team, as well as the difference between the home and away team. For example, the amount of points the home team got the last match day is incorporated in *PG Home*. The share of total possible points the home team has won, is also used in *PG Share Home*.

Schedule related variables capture information such as if the home or the away team is in the Champions League this year. If the match played is two weeks before or after a European game for the home or away team. Finally, there are two dummy variables that indicate if it is a season before the World cup or after, respectively.

Regional economic indicators are incorporated in the covariate set as well. *Unemployment rate* and *log GDP* are the variables that try to reflect the economic situation in the city where a particular team comes from.

Location related factors denote for example the maximum capacity of the home stadium, the distance in km between the two cities of the two competing teams and the travel time between the two cities.

The last set of variables contain variables regarding the results from the previous season. The amount of points won, amount of goals in total and at half time, amount of shots, fouls, yellow cards, red cards, corners and the difference in the amount of points won the previous season between the home and away team, as well as the difference in goals scored the previous season.

Table 8: Descriptive Statistics

Variables	Reference	Unit	Mean (St. dev.)	Update
Team Characteristics				
TV Revenue	Home	EURO (in millions)	26.31(13.25)	Yearly
TV Revenue difference	Home-Away	EURO (in millions)	0.94(8.62)	Yearly
Market value	Home	EURO (in millions)	122.13(124.93)	Yearly
Market value difference	Home-Away	EURO (in millions)	0.08(197.64)	Yearly
Market value / TV Revenue	Home	EURO (in millions)	4.60(3.35)	Yearly
Market value / TV Revenue difference	Home-Away	EURO (in millions)	-0.09(4.78)	Yearly
Market value - TV Revenue	Home	EURO (in millions)	95.82(117.73)	Yearly
Market value - TV Revenues difference	Home-Away	EURO (in millions)	-0.86(162.35)	Yearly
Market value share	Home, Away	Ratio	0.06(0.05)	Yearly
Standardized market value	Home	-	0.05(1.01)	Yearly
Standardized market value difference	Home-Away	-	0.0004(1.47)	Yearly
Average market value	Home	EURO (in millions)	121.18(108.50)	Yearly
Average market value difference	Home-Away	EURO (in millions)	-0.02(157.77)	Yearly
St.dev. market value	Home	EURO (in millions)	52.15(42.94)	Yearly
St.dev. market value difference	Home-Away	EURO (in millions)	0.05(62.13)	Yearly
Ratio of Top 3 to ranked 12-14 players' market value	Home	Ratio	3.60(1.44)	Yearly
Ratio of Top 3 to ranked 12-14 players' market value difference	Home-Away	Ratio	0.002(1.83)	Yearly
Ratio of Top 11 to ranked 12-21 players' market value	Home	Ratio	3.54(1.14)	Yearly
Ratio of Top 11 to ranked 12-21 players' market value difference	Home-Away	Ratio	-0.0007(1.41)	Yearly
Age mean difference	Home-Away	Numerical	-0.001(1.11)	Yearly
Age st. dev. difference	Home-Away	st. dev.	-0.0005(0.74)	Yearly
Age 11 most valuable players difference	Home-Away	Numerical	-0.001(1.54)	Yearly
Age ratio of top 11 to ranked 12-21 difference	Home-Away	Numerical	0(0.11)	Yearly
Age of those above 20 difference	Home-Away	Numerical	-0.006(4.56)	Yearly
Minimum age in the squad difference	Home-Away	Numerical	-0.0002(1.16)	Yearly
Maximum age in the squad difference	Home-Away	Numerical	-0.0002(1.16)	Yearly
Share left footed players difference	Home-Away	Ratio	-0.0001(0.09)	Yearly
Share two footed players difference	Home-Away	Ratio	-0(0.07)	Yearly
Share left footed among 11 most valuable players difference	Home-Away	Ratio	-0.0002(0.17)	Yearly
Share two footed among 11 most valuable players difference	Home-Away	Ratio	0(0.10)	Yearly
Mean height difference	Home-Away	Numerical	0(0.01)	Yearly
St. dev. height difference	Home-Away	Numerical	0(0.01)	Yearly
Mean height top 11 difference	Home-Away	Numerical	0(0.02)	Yearly
St. dev. height top 11 difference	Home-Away	Numerical	0(0.02)	Yearly
Traditional club	Home	Categorical	0.66(0.47)	Once
Traditional club	Away	Categorical	0.66(0.37)	Once
Yo-yo club	Home	Categorical	0.17(0.37)	Once
Yo-yo club	Away	Categorical	0.17(0.37)	Once
Other clubs	Home	Categorical	0.18(0.38)	Once
Other clubs	Away	Categorical	0.17(0.38)	Once
Previous Game (PG) Outcomes				
PG points last match	Home	Numerical	1.14(1.28)	Match
PG points last match	Away	Numerical	1.54(1.33)	Match
PG points share of total	Home	Ratio	0.44(0.20)	Match
PG points share of total	Away	Ratio	0.45(0.21)	Match
Schedule related				
Season ID		Categorical	5.98(3.16)	Yearly
Before European match	Home	Dummy	0.08	Match
Before European match	Away	Dummy	0.08	Match
After European match	Home	Dummy	0.09	Match
After European match	Away	Dummy	0.09	Match
In champions league	Home	Dummy	0.18	Yearly
In champions league	Away	Dummy	0.18	Yearly
Season before World cup		Dummy	0.27	Yearly
Season after World cup		Dummy	0.27	Yearly
Regional Economic Indicators				
Log GDP per capita difference	Home-Away	EURO	0(0.27)	Yearly
Unemployment difference	Home-Away	Percentage	-0.02(4.60)	Yearly
Location Related Variables				
Stadium capacity	Home	Discrete	47990.56(17807)	Match
Distance between cities		Kilometer	371.32(185.41)	Once
Transport time between cities		Minutes	218(102.828)	
Previous Seasons (PS) Outcomes				

Variables	Reference	Unit	Mean (St. dev.)	Update
PS goals	Home	Numerical	49.12(14.33)	Yearly
PS goals	Away	Numerical	49.11(14.35)	Yearly
PS points	Home	Numerical	46.95(13.91)	Yearly
PS points	Away	Numerical	46.95(13.93)	Yearly
PS goals at halftime	Home	Numerical	21.59(6.84)	Yearly
PS goals at halftime	Away	Numerical	21.59(6.85)	Yearly
PS total shots	Home	Numerical	444.33(68.57)	Yearly
PS total shots	Away	Numerical	444.26(68.64)	Yearly
PS shots on target	Home	Numerical	160.19(32.39)	Yearly
PS shots on target	Away	Numerical	160.18(32.42)	Yearly
PS total fouls	Home	Numerical	538.14(71.26)	Yearly
PS total fouls	Away	Numerical	538.13(71.30)	Yearly
PS total corners	Home	Numerical	166.76(28.84)	Yearly
PS total corners	Away	Numerical	166.75(28.87)	Yearly
PS total yellow cards	Home	Numerical	61.10(10.38)	Yearly
PS total yellow cards	Away	Numerical	61.08(10.39)	Yearly
PS total red cards	Home	Numerical	3.00(1.77)	Yearly
PS total red cards	Away	Numerical	3.00(1.77)	Yearly
PS points difference	Home-Away	Numerical	0.004(20.25)	
PS goals difference	Home-Away	Numerical	0.01(20.66)	

The standard deviation is shown in parentheses. This is not reported for the dummy variables. "Home": the home team; "Away": the away team; "Home-Away": the value of the home team minus the value of the away team. Update category *match* updates the corresponding variables before every new match day.

10.2 Season 2018/19

Table 9 shows the predicted league table at match day 33 in the 1. Bundesliga in the season 2018/19. The reason why this league table is estimated and not the final league table, is because the data of the last match day was missing in the data set.

Table 9: The predicted and the actual table of the 1. Bundesliga season 2018/19, until match day 33, estimated with the Ordered Random Forest model

	Rank		Points	
Team	Predicted	Actual	Predicted	Actual
FC Bayern München	1	1	71.0	75
Borussia Dortmund	2	2	59.1	73
RB Leipzig	3	3	52.2	66
Bayer 04 Leverkusen	4	5	52.2	55
Borussia M'gladbach	5	4	51.8	55
TSG Hoffenheim	6	8	47.8	51
Hertha BSC Berlin	7	10	46.7	43
Schalke 04	8	15	46.7	32
Eintracht Frankfurt	9	6	45.9	54
VfL Wolfsburg	10	7	44.1	52
VfB Stuttgart	11	16	42.7	27
Werder Bremen	12	9	42.3	50
Hannover 96	13	17	41.9	21
FC Augsburg	14	14	40.1	32
1. FSV Mainz 05	15	12	39.7	40
SC Freiburg	16	13	38.2	33
1. FC Nuremberg	17	18	21.5	19
Fortuna Düsseldorf	18	11	20.4	41

10.3 Programming Code

The programming code used in this paper is gathered in the zip file called "Programs_434229_Thesis". In this Section all the different programming codes are briefly described.

10.3.1 MATLAB Code

- betting_odds1718.m

  This program is used to compute the hypothetical ROIs using the proportional strategy for season 2017/18 (see Table 3 and Table 7).

- betting_odds1718_2.m

  This program is used to compute the hypothetical ROIs using the value-bet strategy for season 2017/18 (see Table 3 and Table 7).

- betting_odds1719.m

  This program is used to compute the hypothetical ROIs using the proportional strategy where season 2017/18 and season 2018/19 are combined (see Table 3 and Table 7).

- betting_odds1719_2.m

  This program is used to compute the hypothetical ROIs using the value-bet strategy where season 2017/18 and season 2018/19 are combined (see Table 3 and Table 7).

- betting_odds1819.m

  This program is used to compute the hypothetical ROIs using the proportional strategy for season 2018/19 (see Table 3 and Table 7).

- betting_odds1819_2.m

  This program is used to compute the hypothetical ROIs using the value-bet strategy for season 2018/19 (see Table 3 and Table 7).

10.3.2 R Code

- Kronecker_ORF_1718.R

  This program is used to compute the $\hat{p}_{three-way}$ in Table 6 for season 2017/18 (see Table 6).

- Kronecker_ORF_1719.R

  This program is used to compute the $\hat{p}_{three-way}$ in Table 6 where season 2017/18 and season 2018/19 are combined (see Table 6).

- Kronecker_ORF_1819.R

This program is used to compute the $\hat{p}_{three-way}$ in Table 6 for season 2018/19 (see Table 6).

- Poisson.R

  This program is used to compute the probabilities to reach a particular goal in season 2018/19, using the BPRM (see Table 5).

- Prob_8_Matchday.R

  This program is used to compute the probabilities to reach a particular goal in season 2018/19 considering the points already materialized on match day 8, using the ORFM (see Table 4).

- Prob_Before_Season.R

  This program is used to compute the probabilities to reach a particular goal in season 2018/19, computed before the beginning of the season, using the ORFM (see Table 4).

- RandomForest17_18.R

  This program is used to predict the final league table for season 2017/18, using the ORFM (see Table 1).

- RandomForest18_19.R

  This program is used to predict the final league table for season 2018/19, using the ORFM (see Table 9).

- Results_Poisson_171.R

  This program is used to compute the $\hat{p}_{three-way}$ for the bookmakers and the BPRM, where the season 2017/18 and season 2018/19 are combined (see Table 6).

- Results_Poisson_1718.R

  This program is used to compute the $\hat{p}_{three-way}$ for the bookmakers and the BPRM, for the season 2017/18 (see Table 6).

- Results_Poisson_1819.R

  This program is used to compute the $\hat{p}_{three-way}$ for the bookmakers and the BPRM, for the season 2017/18 (see Table 6).

- RF.R

  This program is not used for the content in this paper. It contains a self-written code for the random forest algorithm. The code contains a function called SingleTree1, which is captured in Single regression tree2.R.

- Single regression tree2.R

  This program is not used for the content of this paper. It contains a self-written code that constructs a single decision tree, which is the basis for the random forest code in RF.R.