university of groningen / faculty of science and engineering / mathematics and applied mathematics

Using Poisson regression to model football scores and exploit inaccuracies in the online betting market

Bachelor's Project Mathematics

April 2020

Student: R.S. Bruinsma

First supervisor: dr. M.A. Grzegorczyk

Second assessor: dr. W.P. Krijnen

Abstract

This paper aims to develop a parametric model that indirectly predicts the outcome of football matches by directly predicting the number of goals both teams will make in a match. The model is motivated by a desire to exploit potential inefficiencies in the online betting market. It builds upon existing work on statistical modelling in sports prediction. The theory behind the models used is described, as well as the model selection procedures for selecting the best model. Using historical match data, it finds an optimal prediction method, based on a Poisson regression model, that gives rise to probabilities on match outcomes for assigned matches. These probabilities are compared to the bookmakers odds for the corresponding matches. When the model gives more favourable probabilities, a bet is placed, and it is found that this strategy was profitable employing it on the 2018/2019 Eredivisie season.

Contents

1 Introduction

Betting on football matches has been done since the beginning of the twentieth century and has been growing in popularity ever since. The market size of online gambling in the Netherlands is close to 600 million euros, and sports betting is the main contributor to this with a market size of about 260 million euros. For every football match, one can bet on numerous (and sometimes ridiculous) things, such as which player will receive a yellow card, in which minute a goal will be scored and many many more. Whereas there has been some research done on applying mathematical models to 'beat' the bookmakers in fixed odds match outcome betting (win, lose or draw), considerably less research has been done on the alternative wagers.

The goal of this paper is to develop a generalized linear model to predict the outcome of matches in the 2018/2019 Eredivisie season. We will use this model on 'Draw no Bet' wagers (which team wins if a team wins) to see if we would be able to 'beat' the bookmakers by producing more accurate probabilities on match outcomes than they produced, resulting in a potentially profitable betting strategy.

Chapter 2 reviews the literature discussing the use of mathematical models to predict sports matches, in particular, football matches. There is no literature review about using these models for betting purposes, since this was mostly not done. Chapter 3 then gives the mathematical theory behind the models, after which we discuss how we select the best model in Chapter 4. The match data, as well as the historical odds data that are available and are used are described in Chapter 5. In Chapter 6, we first develop our models, and then see how they would have performed. Thereafter, we also check if we would have been able to make money wagering on 'Draw no Bet' wagers. A conclusion to the paper and suggestions for refinements which could lead to further improvements are given in Chapter 7.

2 Literature Review

A lot of literature has been written about using statistical methods for modelling sports data. However, the majority of this literature is dedicated to the traditionally big American sports: American football, basketball, hockey and baseball. Since the schedules and standings of these sports are more complex than for regular football leagues, most of this research was not aimed at predicting matches, but at ranking teams in a better way than the usual rankings.

An exception to this is by Thompson [34], where the alternative ranking of teams was eventually used for predicting probabilities of upsets between higher and lower ranked teams. Another exception is [4], where NFL teams strengths are measured from historical game results to make predictions on future NFL games, exploiting the paired comparisons models from Thurstone-Mostellor (Thurstone [35], Mosteller [28]) and Bradley-Terry (Bradley and Terry [6]).

But there has also been some research focussing on the worlds biggest sport, football. Through a paper by Jochems in 1958 [20], it was already shown that experts of the game (football journalists) were not able to predict football matches very accurately, and that statistical models may well do better.

This indeed was the case. Through the years, two main methods were developed to predict the outcome of a football match.

The first one is by making a model that directly outputs the probabilities of every match outcome. These models usually use the difference in rating between teams as the predictor. The Bradley-Terry model mentioned before is such a model and was also used to predict football match outcomes after it was extended to accommodate for draws as well (see Davidson [12]). A popular system to rate teams and players in sports is the Elo-rating system, originally developed for rating chess players in 1978 by Arpad Elo [15]. An Elo-rating system that is used for match prediction in football was proposed by Hvattum and Arntzen [19], who used Elo-rating differences as covariates in ordered logit regression models.

Early references to statistical modelling of football data concentrate mainly on the distribution of the number of goals scored in a game. This corresponds to the other main method of predicting the outcome of a football match, namely indirectly predicting the outcome by first predicting how many goals both teams will score and from that calculate the probabilities for each match outcome. A number of different distributions were used to fit on the goals scored in a game, including the Poisson distribution (see [27]) and the negative binomial distribution (see [32]). In this paper we will develop a model of this method, using the papers of the major contributors to this method of prediction.

3 Theory

3.1 Generalized Linear Models

3.1.1 Linear Models

A linear model is used to define how the dependent response variables Y_i depend on p explanatory variables $x_{1i}, \ldots, x_{pi}$, for $i = 1, \ldots, N$ observations from the dataset. In a linear model it is assumed that the response variable Y_i is normally distributed. The linear model looks as follows:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \epsilon_i = \boldsymbol{x_i^T \beta + \epsilon_i}, \text{ for } i = 1, \ldots, N$$

The error terms ϵ_i are assumed to be normally distributed with mean μ equal to 0 and variance σ^2, so that $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. The model is usually written as

$$\boldsymbol{y = X\beta + \epsilon},$$

where

$$\boldsymbol{y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x_1^T} \\ \vdots \\ \boldsymbol{x_N^T} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_N \end{bmatrix}$$

Linear models help understand the data, and are used to predict future behaviour of the data. The least squares estimator and the maximum likelihood estimator are the two most commonly used estimators to estimate the parameter vector $\boldsymbol{\beta}$. For a linear model, these estimators are equal and given by

$$\boldsymbol{\hat{\beta}_{MLE} = \hat{\beta}_{LS} = (X^T X)^{-1} X^T y}$$

The derivation of these estimators can be found in many books on statistics or econometrics. See for example Dobson [14], or Hayashi [16].

3.1.2 Generalized Linear Models

The biggest shortcoming of the linear model is that the response variables Y_i have to be normally distributed. In many cases the response variable is a count variable, or even a binary variable. The generalized linear model extends the linear model in a way such that the response variable is allowed to be any member of the exponential family. A generalized linear model consists of three components:

- The response variables Y_i, coming from the same distribution form (Normal, Exponential, Poisson, etc.). This distribution is a member of the exponential family, in canonical form. That is, the density function of Y_i can be written as
$$f(y; \theta) = e^{yb(\theta) + c(\theta) + d(y)},$$
where b, c and d are known functions.

- The linear predictor η_i, which is a function of the explanatory variables $x_{i1}, \ldots, x_{ip}$, linear in the parameter estimates $\beta_i, \ldots, \beta_p$. Hence,

$$\eta_i = \boldsymbol{x_i^T \beta} \tag{1}$$

- A monotone link function g that links the expected value of the response variable Y_i to the linear predictor η_i:

$$g(E[Y_i]) = \eta_i$$

Generalized linear models were introduced by Nelder and Wedderburn [29] in 1972. In this paper we will use the maximum likelihood estimates to estimate the parameters of the model.

3.2 Poisson Distribution

The Poisson distribution is a non-negative discrete probability distribution. It is named after French mathematician *Siméon Denis Poisson*, who introduced the distribution in 1937 [30]. His work theorized the number of wrongful convictions, focussing on certain random variables that counted the number of discrete occurrences in a given amount of time. However, it was not the first publication of this distribution, since the result had already been given in 1711 by *Abraham de Moivre* [26].

The Poisson distribution expresses the probability of a given number of events occurring in a fixed interval of time, if these events occur with a known and constant average rate, independently of the time since the last event. The rate parameter is the only parameter of the Poisson distribution and it is equal to the average number of events per fixed time-interval.

A discrete random variable X is said to follow the Poisson distribution with rate parameter $\lambda > 0$ if, for $x = 0, 1, 2, \ldots$, the density function of X is given by

$$f(x; \lambda) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{2}$$

The expected value, as well as the variance of this Poisson distributed random variable X are given by

$$E[X] = \text{var}[X] = \lambda$$

Since we can rewrite the density (2) as

$$f(x; \lambda) = e^{x\log(\lambda) - \lambda - \log(x!)}, \tag{3}$$

the Poisson distribution belongs to the exponential family of distributions. This allows it to be used for generalized linear models.

The function that is commonly used to link a response variable Y_i, that is considered to be Poisson distributed with rate parameter λ, to η_i, is the logarithmic link function

$$g(E[Y_i]) = \log(E[Y_i]) = \log(\lambda)$$

This link function keeps the expected value (rate parameter) non-negative, even when the regressors or regression coefficients have negative values.

3.3 Poisson regression

The Poisson regression model is a generalized linear model used to model count data. This regression model is derived from the Poisson distribution by allowing the rate parameter λ to depend on explanatory variables.

Typical data that is used for Poisson regression consists of N independent observations $(Y_i, \boldsymbol{x_i})$, for $i = 1, \ldots, N$. The scalar response variable Y_i is the number of occurrences of the event of interest, and $\boldsymbol{x_i}$ is the vector of linearly independent explanatory variables that are thought to determine Y_i.

A regression model follows by conditioning the distribution of Y_i on a p-dimensional vector $\boldsymbol{x_i}^T = [x_{1i} \ \ldots \ x_{pi}]$, and parameters $\boldsymbol{\beta}$, through a continuous function $\lambda_i(\boldsymbol{x_i}, \boldsymbol{\beta})$, such that $E[y_i|\boldsymbol{x_i}] = \lambda_i(\boldsymbol{x_i}, \boldsymbol{\beta})$. We already stated that the logarithmic link function is most commonly used, so that a regression model is fully defined by the following equations:

$$f(y_i|\boldsymbol{x_i}) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}, \qquad y_i = 0, 1, 2, \ldots$$
$$\lambda_i = e^{\boldsymbol{x_i^T \beta}}$$

Iteratively reweighted least squares can be used to find the maximum likelihood estimates of $\boldsymbol{\beta}$ in Poisson regression.

3.4 Derivation

The derivation below is based on statistical literature, such as [14], and is adjusted for the Poisson regression case.

Consider $Y_1, \ldots, Y_N$ independent Poisson distributed random variables with rate parameter λ_i, for $i = 1, \ldots, N$. Suppose they satisfy the properties of a generalized linear model. We wish to estimate the parameter vector $\boldsymbol{\beta}$, which is related to the Y_i's through $E[Y_i] = \lambda_i$, and the logarithmic link function $g(\lambda_i) = \log(\lambda_i) = \boldsymbol{x_i^T \beta} = \eta_i$. $\boldsymbol{x_i}$ is a vector $(x_{i1}, \ldots, x_{ip})$ corresponding to the covariate pattern of Y_i.
The log-likelihood for each Y_i is

$$l_i(\lambda_i; y_i) = y_i b(\lambda_i) + c(\lambda_i) + d(y_i), \tag{4}$$

where $b(\cdot), c(\cdot)$ and $d(\cdot)$ are functions arising from the exponential family form of the Poisson distribution, see (3).
Substituting those functions, we can rewrite the log-likelihoods (4) as

$$l_i(\lambda_i; y_i) = y_i log(\lambda_i) - \lambda_i - log(y_i!) \tag{5}$$

The log-likelihood for the entire sample $Y_1, \ldots, Y_N$ is then

$$\begin{aligned} l(\boldsymbol{\lambda}; \boldsymbol{y}) &= \Sigma_{i=1}^N [l_i(\lambda_i; y_i)] \\ &= \Sigma_{i=1}^N [y_i log(\lambda_i)] - \Sigma_{i=1}^N [\lambda_i] - \Sigma_{i=1}^N [log(y_i!)] \end{aligned}$$

To obtain the maximum likelihood estimates for β_j, we need the score function

$$\begin{aligned} U_j &= \frac{\partial l}{\partial \beta_j} \\ &= \Sigma_{i=1}^N [\frac{\partial l_i}{\partial \beta_j}] \\ &= \Sigma_{i=1}^N [\frac{\partial l_i}{\partial \lambda_i} \frac{\partial \lambda_i}{\partial \beta_j}] \end{aligned} \tag{6}$$

using the chain rule.
If we treat both terms of the right hand side of (6) separately, for our Poisson case, we get, using (5), (1) and the chain rule:

$$\frac{\partial l_i}{\partial \lambda_i} = \frac{y_i}{\lambda_i} - 1$$

and

$$\frac{\partial \lambda_i}{\partial \beta_j} = \frac{\partial \lambda_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \lambda_i}{\partial \eta_i} x_{ij}$$

Hence,

$$U_j = \Sigma_{i=1}^N [(\frac{y_i}{\lambda_i} - 1) x_{ij} \frac{\partial \lambda_i}{\partial \eta_i}] \tag{7}$$

Since the expected value of the score function is zero, the variance covariance matrix of U_j has terms

$$\begin{aligned} \mathcal{J}_{jk} &= E[U_j U_k] \\ &= E\{\Sigma_{i=1}^N [(\frac{y_i - \lambda_i}{\lambda_i}) x_{ij} \frac{\partial \lambda_i}{\partial \eta_i}] \Sigma_{l=1}^N [(\frac{y_l - \lambda_l}{\lambda_l}) x_{lk} \frac{\partial \lambda_l}{\partial \eta_l}]\} \\ &= \Sigma_{i=1}^N [\frac{E[(y_i - \lambda_i)^2] x_{ij} x_{ik}}{\lambda_i^2} (\frac{\partial \lambda_i}{\partial \eta_i})^2], \end{aligned} \tag{8}$$

because $E[(y_i - \lambda_i)(y_l - \lambda_l)] = 0$ for $i \neq l$ as the y_i's are independent.
Using that $E[(y_i - \lambda_i)^2] = \text{var}[y_i] = \lambda_i$, we can simplify (8) to

$$\mathcal{J}_{jk} = \Sigma_{i=1}^N [\frac{x_{ij} x_{ik}}{\lambda_i} (\frac{\partial \lambda_i}{\partial \eta_i})^2] \tag{9}$$

The elements $\mathcal{J}_{jk}$ form the information matrix $\mathcal{J}$. This information matrix can be rewritten as

$$\mathcal{J} = \boldsymbol{X^T W X}, \tag{10}$$

where $\boldsymbol{W}$ is the $N \times N$ diagonal matrix with diagonal elements

$$w_{ii} = \frac{1}{\lambda_i}(\frac{\partial \lambda_i}{\partial \eta_i})^2$$

The equation to estimate $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$ at the m^{th} iteration, using the method of scoring (see, for example [14]), then is

$$\boldsymbol{b^{(m)}} = \boldsymbol{b^{(m-1)}} + [\mathcal{J}^{(m-1)}]^{-1}\boldsymbol{U^{(m-1)}}, \tag{11}$$

where $\boldsymbol{b^{(m)}}$ is the m^{th} iteration estimate for $\boldsymbol{\beta}$.
Multiplying both sides of (11) with $\mathcal{J}^{(m-1)}$ gives

$$\mathcal{J}^{(m-1)}\boldsymbol{b^{(m)}} = \mathcal{J}^{(m-1)}\boldsymbol{b^{(m-1)}} + \boldsymbol{U^{(m-1)}} \tag{12}$$

Using equation (7) for the score function U_j and equation (9) for the information, the expression on the right hand side of (12) can be written as a vector with elements

$$\Sigma_{k=1}^{p}\Sigma_{i=1}^{N}[\frac{x_{ij}x_{ik}}{\lambda_i}(\frac{\partial \lambda_i}{\partial \eta_i})^2 b_k^{(m-1)}] + \Sigma_{i=1}^{N}[\frac{y_i - \lambda_i}{\lambda_i}x_{ij}(\frac{\partial \lambda_i}{\partial \eta_i})]$$

evaluated at $\boldsymbol{b^{(m-1)}}$.
Therefore, we can rewrite this right hand side of (12) as

$$\boldsymbol{X^T W z}, \tag{13}$$

with

$$z_i = \Sigma_{k=1}^{p}[x_{ik}b_k^{(m-1)} + (y_i - \lambda_i)(\frac{\partial \eta_i}{\partial \lambda_i})]$$

Hence, using (13) and (10), the estimation equation (12) can be written as

$$\boldsymbol{X^T W X b^{(m)}} = \boldsymbol{X^T W z} \tag{14}$$

This has to be solved iteratively because $\boldsymbol{z}$ and $\boldsymbol{W}$ depend on the estimate, $\boldsymbol{b}$. In this paper we use `R` for fitting generalized linear models. `R` has an efficient algorithm based on (14).

4 Model Selection

4.1 Kolmogorov-Smirnov test

Before we should look at which explanatory variables we can include in our model, we have to convince ourselves that the data is from a Poisson distribution.

The Kolmogorov-Smirnov test [24] is used to decide if a sample comes from a population with a specified distribution. It depends on the empirical distribution function, as well as the specified cumulative distribution function of the sample. For a sample $Y_1, \ldots, Y_N$, the empirical distribution function is defined as:

$$S_N(y) := \text{fraction of sample points less than } y$$

The Kolmogorov-Smirnov test is based on the maximum difference between the empirical distribution function, and the specified cumulative distribution function. The test has several limitations, some of which apply to our case. Namely, the test in its original form does only apply to continuous distributions, and the parameters of the specified distribution need to be known.

Since these are not the case for us, we cannot use the Kolmogorov-Smirnov test. However, adaptations to the Kolmogorov-Smirnov test are made to solve for these problems. To this end, we can use the Kolmogorov-Smirnov test for the Poisson distribution with unknown mean, described in [9].

The rate parameter of the Poisson distribution is estimated by the mean value of the sample, $\bar{y}$. The test still uses the maximal difference between the empirical distribution and the specified distribution as its test statistic. The test statistic is defined as

$$D = \sup_y |F_0(y) - S_N(y)|,$$

where $F(y) = P(Y_i \leq y)$ is the cumulative Poisson distribution function evaluated with $\bar{y}$ as the estimated rate parameter. If D exceeds the critical value found in [9] for the correct sample characteristics[1], we reject the null hypothesis that the sample is Poisson distributed.

4.2 Including Explanatory Variables

Before we include explanatory variables in our model, we will give a strong theoretical basis on why they should be included. This will be done by investigating the data at our disposal.

[1] the value of the mean and the sample size indicate a critical value

4.3 Overdispersion

Overdispersion is a common issue in Poisson regression. For a Poisson random variable Y, we know that $\text{var}[Y] = E[Y]$. Overdispersion in Poisson regression occurs when the variance of the response variable is greater than the expected value of this response variable.

Cameron and Trivedi [8] proposed to test the null hypothesis that $\text{var}[Y] = E[Y] = \mu$ against the alternative hypothesis that $\text{var}[Y] = \mu + \alpha g(\mu)$, where $\alpha > 0$ means overdispersion (and $\alpha < 0$ underdispersion). The function g is some monotone function. The coefficient of α can be estimated by an auxiliary ordinary least squares regression and tested with the corresponding t-statistic, which is asymptotically normal under the null hypothesis that $\alpha = 0$.

If the assumption of equidispersion is not justified, then the negative binomial distribution provides an alternative model with $\text{var}[Y_i] = \phi E[Y_i]$, where $\phi > 1$ is a dispersion parameter that can be estimated. Overdispersion can be a consequence of dependence between observations.

4.4 Mc Fadden's pseudo R^2

R^2 is a statistic that will give information about the goodness of fit of a model. In an ordinary least squares regression, R^2 is the proportion of the observed variation of the dependent variable that can explained by the model. The value of R^2 is thus between 0 and 1, and a higher R^2 means that more of the variance can be explained by the model. If we let y denote the dependent response variable, $\bar{y}$ the mean of the observations y, $\hat{y}$ the fitted value of the prediction and N the number of observations, then R^2 is defined as follows:

$$R^2 = 1 - \frac{\Sigma_{i=1}^{N}(y_i - \hat{y}_i)}{\Sigma_{i=1}^{N}(y_i - \bar{y})}$$

In ordinary least squares the squared error is minimized to estimate the parameter values. Since we use the maximum likelihood estimators instead, and since this maximum likelihood estimators do not necessarily minimize the squared error, we cannot use R^2 as a goodness of fit statistic for our models.

As an alternative to R^2 for generalized linear modelling, we introduce Mc-Fadden's pseudo R^2 [25],

$$R^2_{McF} = 1 - \frac{l(M_{int})}{l(M_{min})},$$

where M_{int} is the model of interest, M_{min} is the model that only includes an intercept and $l(\cdot)$ is the value of the log-likelihood.

Unlike ordinary R^2, R^2_{McF} does not represent the proportion of explained variance, but rather the improvement in model likelihood over a null model.

Considering the size of our dataset, a R^2_{McF} value between 0.15 and 0.32 indicates a good fit[2]. Higher values indicate an excellent fit.

4.5 Penalized model selection criteria

The most common approach on the comparison and selection of statistical models is standard significance testing of nested models. However, such tests potentially have some undesirable properties. They can be very sensitive to small deviations from the null hypothesis when the sample sizes are large, leading to the possibility of rejecting reasonably parsimonious models as having a statistically significant lack of fit. Standard tests also provide little guidance for choosing between non-nested models that cannot be rejected. One common class of an alternative selection method is the penalized model selection criteria.

The two most commonly used penalized model selection criteria are the Bayesian Information Criterion (BIC, Schwarz [33]) and the Akaike Information Criterion (AIC, Akaike [3]).
They are based on the maximum value of the log-likelihood function with an adjustment for the number of parameters that are being estimated and an adjustment for the number of observations. We define them as follows, since `R` follows this definition as well:

$$AIC = -2l_{max} + 2p$$
$$BIC = -2l_{max} + p \cdot ln(N),$$

where p is the number of parameters that are being estimated, N is the number of observations (datapoints) and l_{max} is the maximum value for the log-likelihood.

For predictive modelling, the AIC is the best statistic to test, according to Konishi and Kitagawa [22]. The best model when comparing by the AIC, is the model with the lowest AIC value. We call this value AIC_{min}.

The differnce in AIC between model i and the model with minimal AIC value is then given as

$$\Delta_i = AIC_i - AIC_{min}$$

Table 1 comes from Burnham and Anderson [7]. It shows how big the difference in AIC value, compared to AIC_{min}, of a model can be to still be considered.

[2] see [17]

Δ_i	level of empirical support for model i
0-2	substantial
4-7	considerably less
>10	essentially none

Table 1: *Level of support for model i based on the difference in AIC value.*

5 Data

5.1 Eredivisie

Our research will focus on the prediction of matches in the 2018/2019 season of the Eredivisie. The Eredivisie is the highest division football league in the Netherlands, similar to what the Premier league is in England. In the Eredivisie, eighteen teams participate every season. Every team plays 34 matches in a season: a home and away match against every other team.
At the end of a season, the team at the top wins the title and the team at the bottom will relegate to the second division. Due to a playoff system for relegation (to the second division) and promotion (to the Eredivisie), there will be either one, two or three new teams in the Eredivisie each season. They entered the league via promotion.

Since there is no data available of the second division, this is important for our research, because it means that historical data is not always available for every team. And if it is, the amount of data available could still differ per team.

5.2 Match statistics data

There is an abundance of statistics available for each match of football that has been played. We gathered our match data from the free football data site *football-data.co.uk* [2] and the `engsoccerdata` package [11] from `R`.

For setting up our models, we will use the data of all matches that have been played in the Eredivisie from the 2010/2011 season up until the season we want to predict: the 2018/2019 season. This adds up to a total of 2754 matches that are played over nine consecutive seasons[3].

For every match we filtered out the most important statistics we needed for eventual match prediction. These were the following:

- date on which the match was played
- home team of the match
- away team of the match
- number of goals the home team scored in the match
- number of goals the away team scored in the match

[3]For some figures and results shown later a larger sized dataset has been used. If this is the case, this will be explicitly noted.

Figure[4] 1 shows us the number of matches that all teams from the 2018/2019 Eredivisie season have played in the Eredivisie in the time span of our dataset (from 2010/2011 until 2017/2018). The complete dataset can be found here.
As mentioned before, we can see that a difference between the data available for every team is existing.
For the teams that participated in the 2018/2019 Eredivisie season, there are ten teams that played every possible match they could, meaning that they never relegated from the Eredivisie since the 2010/2011 season.
We also see that there are two teams displayed in Figure 1, Fortuna Sittard and FC Emmen, that have not played in single match in the Eredivisie in the time span of our dataset, meaning that they forced promotion to the Eredivisie in the 2017/2018 season, and played in the second division in the other seasons of our dataset.

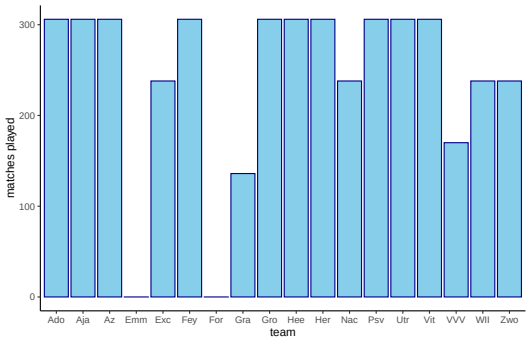

Figure 1: *Total matches played during the time span of our dataset by all teams that participated in the 2018/2019 Eredivisie season.*

5.3 Betting Application

5.3.1 Odds

In this paper, we work with decimal odds, since most online betting sites use these as well. By looking at the odds for every outcome of the match, you can immediately spot the favourite and the underdog. The underdog has the higher odds.

These decimal odds represent the amount one wins for every €1 stake. It represents the total return, and not just the profit. The total potential return on a stake is calculated as

$$\textbf{Total return} = \text{stake} \times \text{decimal odd number}$$

[4]Captions to figures and tables are written in italic so that they are easily distinguishable from regular text

We will bet on 'Draw no Bet' wagers, which means that we can put money on either the home team or the away team. If the match then ends up as a draw, we get our stake back, and we do not lose or win anything with the bet.

5.3.2 "Draw no Bet" data

We want to predict the outcome of all matches of the 2018/2019 Eredivisie season. For this we will use the match statistics data that we gathered. But, thereafter we want to see if our prediction model could have been profitable for 'Draw no Bet' wagering.
For this purpose, we need the 'Draw no Bet' odds that bookmakers had for all of the 2018/2019 Eredivisie matches. These odds can be found on *OddsPortal.com* [1]. The 'Draw no bet' odds found on this website give us the best odds on every match, meaning that it compares odds on the same match for multiple bookmakers[5], and chooses the ones for which the most money is won for each outcome of the match, see Table 2.
We use the highest odds for every match, since this is the most beneficial to us: Betting on the highest odds of a match, gives us the biggest return if we win the bet, making it easier to be profitable.

Bookmaker	odds on home win	odds on away win	odds on draw
A	**1.53**	2.91	no bet
B	1.50	**2.95**	no bet
C	1.51	2.93	no bet

Table 2: *In this example we would use bookmaker A for the odds of a home win and bookmaker B for the odds of an away win.*

[5] bet365, Bethard, bwin, Coolbet, Unibet and Wiliam Hill

6 Model

The outcome of a football match is determined by the number of goals both teams score. The team that scores the most goals wins the game. We are interested in Eredivise league games, which means that if the two teams score the same amount of goals in a match, the match ends up as a draw.
We will predict the outcome of a football match by setting up a model that will assign probabilities to the number of goals each team scores in a match. From this, probabilities on the outcome of the match (win, loss) can be calculated.

In Section 6.1 we will analyze the matchdata that we have gathered for the Eredivisie as described in the previous chapter, as well as checking some assumptions that we will make in setting up and selecting a best possible prediction model, which will be done in Section 6.2. This model will then be used in Section 6.3 to see what would have happened if we had used our model to put money on 'Draw no Bet' wagers for the 2018/2019 Eredivisie season.

6.1 Pre-analysis of data

6.1.1 Poisson fit

Moroney (1956) [27] was the first one who examined if the Poisson distribution was a good fit on the number of goals a team scores in a football match.

From Figure 2 we can see that the Poisson distribution, which is a widely applied standard statistical distribution, does fairly good in explaining how much goals are made by a team during a football match.

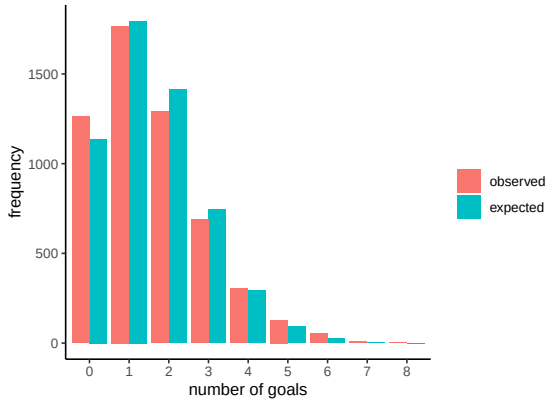

Figure 2: *Observed frequency a certain number of goals was scored by a team in a match (red) versus the expected frequency this amount of goals was scored by a team in a match from the Poisson distribution with the mean of the number of goals scored per match as rate parameter (blue).*

Performing the Kolmogorov-Smirnov test for the Poisson distribution with unknown mean discussed earlier, we get that (see Table 3)

$$D = \sup_y |F_0(y) - S_N(y)| = 0.023$$

This D value does not exceed the critical value we found from [9][6], and therefore we cannot reject that the data is indeed Poisson distributed.

Number of Goals y	Observed Frequencies	Empirical $S_n(y)$ Distribution $S_N(y)$	Theoretical Distribution $F_0(y)$	$\lvert F_0(y) - S_N(y)\rvert$
0	1265	0.230	0.207	$0.023 = D^*$
1	1767	0.550	0.532	0.018
2	1290	0.785	0.789	0.005
3	687	0.909	0.924	0.015
4	304	0.965	0.978	0.013
5	129	0.988	0.994	0.006
6	52	0.997	0.999	0.001
7	10	0.999	1.000	0.000
8	3	1.000	1.000	0.000
9	0	1.000	1.000	0.000
10	1	1	1.000	0.000

Table 3: *Kolmogorov-Smirnov test values for our dataset. Every value is rounded to three decimal places. *Highest value for the last column, this value is compared to the critical value for the test.*

Overdispersion is not a problem in our dataset. Performing the test as described in Section 4.3, we get the estimate $\alpha = -0.026$, with a p-value of 0.915 so that we cannot reject the null hypothesis that $\alpha = 0$ and our data is equidispersed.

6.1.2 Generalized Linear Model: Explanatory Variables

Because the Poisson distribution is part of the exponential family, we can fit a generalized linear model to the number of goals a team scores in a match. When using such a generalized linear model, the goal is to estimate the parameter vector $\boldsymbol{\beta}$ in

$$g(E[Y_i]) = \boldsymbol{x_i^T}\boldsymbol{\beta},$$

where $\boldsymbol{x_i^T}$ denote the values of the explanatory variables corresponding to observation i. $\boldsymbol{\beta}$ will then determine our estimate for $E[Y_i]$, the expected number of goals that is scored.

[6]Using that the number of observations is equal to 5508, and the mean value of the data is equal to 1.577

When building our generalized linear model, it is necessary to know what explanatory variables we need to be able to predict the number of goals that teams score in a match.

For people who watch the game of football regularly, it will not come as a surprise that the quality of different teams are not generally the same, and so different teams will likely to be expected to score different amount of goals in a match. However, there was a time that the question whether football was a game of chance or skill was a difficult one to answer.

Benjamin and Reep(1968) [31] even thought they had proven that football was more a game of chance than skill by looking at how passes effected the number of goals scored by a team. But eventually, Hill(1971) [18] proved that football is a game dominated by skill by performing significance tests on the predictions of experts on league outcomes.

To make it explicit that the difference in quality between teams indeed exists, we can look at Figure 3 which shows the average number of goals scored and conceded per match over the time span of our dataset for every team that participated in the 2018/2019 Eredivisie.

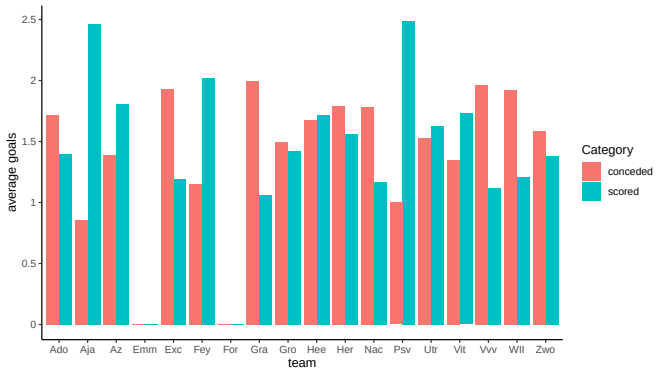

Figure 3: *Average number of goals scored (blue) and conceded (red) per match for all teams that participated in the 2018/2019 Eredivisie , over the time span of our dataset. FC Emmen and Fortuna Sittard have not played a match in the Eredivisie in our dataset, so they have not scored or conceded any goals.*

From this figure we can see that teams are not equally good. There are teams that are clearly better than most other teams, they scored more goals than that they conceded (for example Ajax, Psv and to a lesser degree Vitesse), and there are also teams that are worse than most other teams (for example de Graafschap and Ado den Haag). We can also see that, as noted before, Fortuna Sittard and FC Emmen have not played any matches in the Eredivisie in the time span of our dataset our dataset (average of goals scored and conceded is both equal to zero).

Another important factor in sports is home advantage. In particular, it is also significant in football. Courneya and Carron [10] give a summary of all the work done on home advantage. They make the point that future research should not be directed to the existence of home advantage, since it is proven enough times that this indeed exists.

Without rigorously proving, but for the sake of clarity that home advantage indeed exists, also in the Dutch Eredivisie, Figure 4 is added. We can see that there were more wins by a home team than by an away team in every season from 1993/1994 until 2017/2018. During these twenty-five consecutive seasons there were 3660 wins by a home team and 2181 wins by an away team.

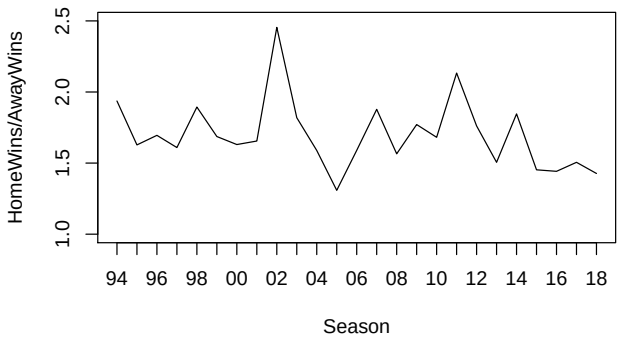

Figure 4: *The ratio of home wins to away wins for every Eredivisie season from the 1993/1994 (denoted 94) season until the 2017/2018 (denoted 18) season.*

From the analysis above, we conclude that we have a theoretical basis to include both the attacking and defensive strength of teams in our model, as well as a home effect.

6.2 Selection of generalized linear model

6.2.1 Generalized Linear Model

In the paper by Maher, *Modelling association football scores* [23], an independent Poisson model for scores was adopted using attacking strength and defensive weakness of teams, including a home effect. In particular, if team i is playing a match against team j, and the observed score is (x_{ij}, y_{ij}), it is assumed that X_{ij} is Poisson distributed with parameter $\alpha_i\beta_j$, that Y_{ij} is Poisson

distributed with parameter $\gamma_j\delta_i$, and that X_{ij} and Y_{ij} are independent.
Then, we can think of α_i as the attacking strength of team i when playing at home, β_j as the defensive weakness of team j when playing away from home, γ_j the offensive strength of team j away from home and δ_i the defensive weakness of team i playing at home.

This means that the model has 4 parameters for every team included in the dataset. To get a unique set of parameters, we impose the sum-to-zero constraints

$$\Sigma_i\alpha_i = \Sigma_i\beta_i = \Sigma_i\gamma_i = \Sigma_i\delta_i = 0,$$

where the sum ranges over all teams included in the dataset.

In its generalized linear model form, using the logarithmic link function, the model for the score of the match looks like

$$\begin{cases} \log(homegoals) = & \mu_1 + \alpha_{Aja}home_{Aja} + \cdots + \alpha_{Zwo}home_{Zwo} + \beta_{Aja}away_{Aja} + \cdots + \beta_{Zwo}away_{Zwo} \\ \log(awaygoals) = & \mu_1 + \gamma_{Aja}away_{Aja} + \cdots + \gamma_{Zwo}away_{Zwo} + \delta_{Aja}home_{Aja} + \cdots + \delta_{Zwo}home_{Zwo}, \end{cases} \tag{15}$$

where *homegoals* and *awaygoals* denote the expected number of goals the home team and the away team will score, respectively. μ_1 denotes the intercept, $home_i$ is the indicator variable indicating whether team i plays at home or not. $away_i$ is defined similarly, so

$$home_i = \begin{cases} 1, \text{ if team } i \text{ plays at home} \\ 0, \text{ otherwise, so if team } i \text{ plays away from home or is not included in the match} \end{cases}$$

$$away_i = \begin{cases} 1, \text{ if team } i \text{ plays away from home} \\ 0, \text{ otherwise, so if team } i \text{ plays at home or is not included in the match} \end{cases}$$

Using that

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_{Aja} \\ \vdots \\ \alpha_{Zwo} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_{Aja} \\ \vdots \\ \beta_{Zwo} \end{bmatrix}, \boldsymbol{\gamma} = \begin{bmatrix} \gamma_{Aja} \\ \vdots \\ \gamma_{Zwo} \end{bmatrix}, \boldsymbol{\delta} = \begin{bmatrix} \delta_{Aja} \\ \vdots \\ \delta_{Zwo} \end{bmatrix},$$

$$\boldsymbol{home} = \begin{bmatrix} home_{Aja} \\ \vdots \\ home_{Zwo}, \end{bmatrix}, \boldsymbol{away} = \begin{bmatrix} away_{Aja} \\ \vdots \\ away_{Zwo}, \end{bmatrix}$$

we can rewrite (15) more readable as

$$\textbf{Model 1}: \begin{cases} \log(homegoals) = \mu_1 + \boldsymbol{home^T\alpha} + \boldsymbol{away^T\beta} \\ \log(awaygoals) = \mu_1 + \boldsymbol{away^T\gamma} + \boldsymbol{home^T\delta} \end{cases}$$

The question arises whether all these parameters are necessary for an adequate description of the match scores. We have noted that there must be real differences between quality of teams, but are these indeed in both attack and defense, and is it really necessary to have different parameters for the qualities of teams at home and away?

6.2.2 Alternative models

Consideration of such questions leads to a possible hierarchy of models that can be tested against each other. As most general model, we have **Model 1**, as described above, in which all four types of parameters are allowed to take different values for the different teams. **Model 2** is the model in which only the attacking parameters (called $\boldsymbol{\alpha}, \boldsymbol{\gamma}$) are allowed to take different values for the different teams. Likewise, **Model 3** only includes the defensive weakness parameters (called $\boldsymbol{\beta}, \boldsymbol{\delta}$) for the different teams.

Model 4 then is the model that takes the home advantage to be constant, whereas it was different for every team in our first models, where we included the home advantage in the attack and defense parameters. In this case, there is a constant home advantage parameter, and an attacking strength and defensive weakness parameter per different team, called ζ_i and ω_i respectively, for team i. The last two models, **Model 5** and **Model 6** are the models that depend only on a home advantage parameter and an attacking strength or defensive weakness parameter per team, respectively.

$$\textbf{Model 2}: \begin{cases} \log(homegoals) = & \mu_2 + \boldsymbol{home}^{\boldsymbol{T}}\boldsymbol{\alpha} \\ \log(awaygoals) = & \mu_2 + \boldsymbol{away}^{\boldsymbol{T}}\boldsymbol{\gamma}, \end{cases}$$

$$\textbf{Model 3}: \begin{cases} \log(homegoals) = & \mu_3 + \boldsymbol{away}^{\boldsymbol{T}}\boldsymbol{\beta} \\ \log(awaygoals) = & \mu_3 + \boldsymbol{home}^{\boldsymbol{T}}\boldsymbol{\delta}, \end{cases}$$

$$\textbf{Model 4}: \log(goals) = \mu_4 + h_4 atHome + \boldsymbol{att.team}^{\boldsymbol{T}}\boldsymbol{\zeta} + \boldsymbol{def.team}^{\boldsymbol{T}}\boldsymbol{\omega},$$

$$\textbf{Model 5}: \log(goals) = \mu_5 + h_5 atHome + \boldsymbol{att.team}^{\boldsymbol{T}}\boldsymbol{\zeta},$$

$$\textbf{Model 6}: \log(goals) = \mu_6 + h_6 atHome + \boldsymbol{def.team}^{\boldsymbol{T}}\boldsymbol{\omega}$$

To get a unique set of parameters, the following sum-to-zero constraints are used:

$$\begin{aligned} &\text{Model 2}: \Sigma_i \alpha_i = \Sigma_i \gamma_i = 0 \\ &\text{Model 3}: \Sigma_i \beta_i = \Sigma_i \delta_i = 0 \\ &\text{Model 4}: \Sigma_i \zeta_i = \Sigma_i \omega_i = 0 \\ &\text{Model 5}: \Sigma_i \zeta_i = 0 \\ &\text{Model 6}: \Sigma_i \omega_i = 0 \end{aligned}$$

Starting from Model 4, we no longer differentiate in home advantage per team, so we can use a single home effect parameter, and use the same model for the number of goals the home team makes, as the number of goals the away team makes.
atHome denotes an indicator variable indicating whether the attacking team plays at home or not. $att.team_i$, $(def.team_i)$ is defined as an indicator variable indicating whether team i is the attacking team (defending team) or not:

$$\boldsymbol{att.team} = \begin{bmatrix} att.team_{Aja} \\ \vdots \\ att.team_{Zwo}, \end{bmatrix}, \boldsymbol{def.team} = \begin{bmatrix} def.team_{Aja} \\ \vdots \\ def.team_{Zwo}, \end{bmatrix}$$

	AIC	R^2_{McF}	log-likelihood
Model 1	9238.0	.154	-8360.296
Model 2	9394.8	.103	-8542.283
Model 3	9431.2	.078	-8632.938
Model 4	9172.1	.150	-8374.949
Model 5	9309.5	.103	-8544.536
Model 6	9399.6	.077	-8634.951

Table 4: *AIC, Mc Fadden's pseudo R^2 and the value of the log-likelihood for our generalized linear models.*

Comparing the models with each other, as in Table 4, we conclude that Model 4 is our best model. Its R^2_{McF} value indicates a good fit, but we pick Model 4 based on the AIC value that is by far the lowest of the different models.

6.2.3 Dynamic Model

We now have a model that fits good on the dataset. However, our goal is to predict future matches. A limitation to the model we have right now is that every match in the dataset is weighted equally.

A teams' performance is likely to be more closely related to recently played matches than to earlier played matches. This should be incorporated into our model. In principle, this behaviour can be modelled by formalizing a stochastic development of the model parameters. However, taking the dimensionality of the model into account, and since we shall always estimate the parameters at fixed time points (right before a matchday), rather then forecasting ahead, we take a more simplistic approach here.

Dixon and Coles [13] proposed to weigh matches $k = 1, \ldots, 2754$ according to the function

$$\phi(t - t_k) = exp(-\xi(t - t_k)), \qquad \xi > 0,$$

where t is the day that our match of interest is played, and t_k is the day that match k was played: $(t - t_k)$ is thus the number of days between the match we want to estimate and match k. All previous results, downweighted exponentially according to ξ, are included in the inference at day t. The static model we had (Model 4), arises as the case $\xi = 0$, and increasingly large values of ξ give relatively more weight to more recently played matches (see Figure 5).

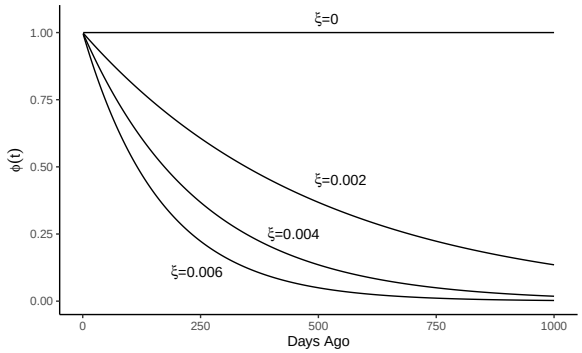

Figure 5: *The weight $\phi(t)$ that is given to a match, where t denotes the number of days ago this match was played, displayed for some different values of our rate parameter ξ.*

We need to choose an optimal ξ. We want this ξ to be such that the overall predictive capabilities, in terms of match outcomes of 2018/2019 Eredivisie matches, of our model is maximized. We will pick this ξ based on which value performed best on the 2017/2018 Eredivisie season.
First note that the probability that match k ends up as a home win is estimated by our model as

$$p_k^H = \Sigma_{l,m \in A_H} P(X_k = l, Y_k = m),$$

where $A_H = \{(l, m) | l > m\}$ and the score probabilities are determined from maximizing the likelihood of the dynamic model at the day of match k. p_k^A is similarly defined as the estimated probability that match k ends up as an away win. We then want our ξ to maximize

$$S(\xi) = \Sigma_k (\rho_k^H \log(p_k^H) + \rho_k^A \log(p_k^A)), \tag{16}$$

over all matches k of the 2017/2018 Eredivise season, with $\rho_k^H = 1$ if match k was won by the home team and zero otherwise, while $\rho_k^A = 1$ if match k was won by the away team and zero otherwise. A plot of $S(\xi)$ against ξ is given in Figure 6. The subsequent results are given using $\xi = 0.0017$, the value that maximizes $S(\xi)$.

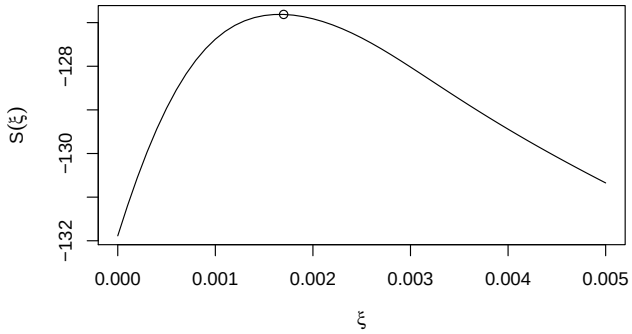

Figure 6: (ξ) *against* ξ *for the 2017/2018 Eredivisie season. The maximum occurs at* $\xi = 0.0017$*, rounded to 4 decimal places.*

Before looking at how well our model performed compared to the bookmakers to see if we could have made money betting on 'Draw no Bet' wagers, we will look at how accurate our model was at predicting the goals scored per match per team, the goal margin by which matches were won and the actual winning team of the matches.

6.2.4 Model performance

Throughout the remainder of this paper, we will call our final model the dynamic model, and the model which does not weigh matches the static model (corresponding to the case $\xi = 0$).

We will first look at how the dynamic model performed. Since we cannot give a correct insight on how good the promoted teams are in the beginning of the season, we use the models on all of the 2018/2019 Eredivisie matches except for the first two matches played by the promoted teams. This means we predict 300 matches instead of 306.

Figure 7 shows us how accurate the dynamic model was at predicting the goals a team scores in a match during the Eredivisie 2018/2019 season. It consists of 600 observations (two for every match played). All observations that are inside the two lines correspond to a 'correctly predicted' number of goals by a team in a match in the sense that for these observations the closest integer to the expected number of goals found from the dynamic model was the number of goals that was actually scored.
It is clear that the dynamic model almost never gives the highest probability to a team scoring 0 goals in a match, whereas we observed that this happened a

lot of times. It also shows that the expected number of goals a team scores in a match, provided by the dynamic model, is not more than 4 goals (except for one time). However, we can see that it happened quite a few times that a team actually managed to score 4 or more goals in a match. Lastly, we can see that when a team scored 1,2 or 3 goals, that the dynamic model was 'correct' a lot of times.

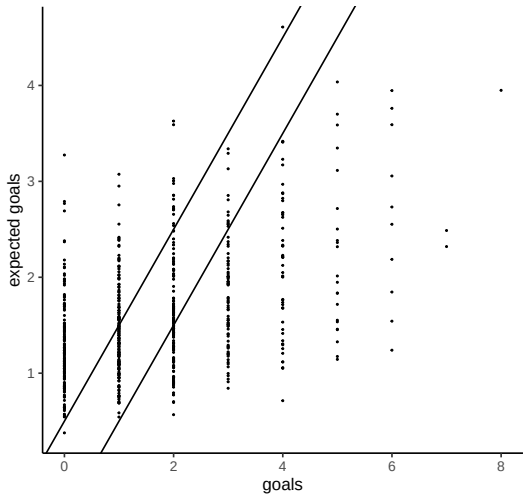

Figure 7: *The actual goals scored in a match by one team plotted against the expected goals scored in that match by that team. The plot shows all 600 observations corresponding to the 2018/2019 Eredivisie. Points inside the lines correspond to observations for which the absolute difference between the expected goals and observed goals is less than a half.*

Table 5 then gives us the exact number of times the dynamic model gave the highest probability to the correct number of goals a team scored.

		frequency	correctly predicted
goals	0	144	1
	1	165	99
	2	130	56
	3	79	12
	≥ 4	82	0
total		600	168

Table 5: *Results on how well our model predicted the amount of goals scored by a team in a match. The frequency column denotes the number of times a certain amount of goals was scored in the 2018/2019 Eredivisie and the last column tells us how many of these times our model gave the highest probability to this amount of goals.*

Figure 8 shows us how accurate the dynamic model was in predicting the goal margin as seen from the home team for every match in the Eredivisie 2018/2019 season. We can see a clear positive correlation between the expected goal margin and the observed goal margin.

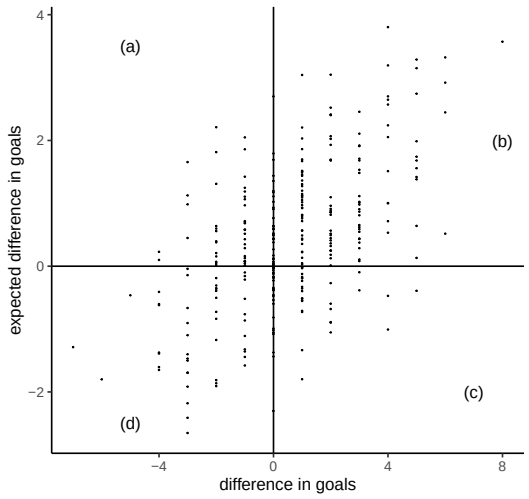

Figure 8: *The expected goal margin of a match plotted against the observed goal margin as seen from the home team. The plot shows all 300 matches corresponding to the 2018/2019 Eredivisie. It is divided into 4 areas: (a) Dynamic model expected a home win, but the away team won. (b) Dynamic model expected a home win, and the home team won. (c) Dynamic model expected an away win, but the home team won. (d) Dynamic model expected an away win, and the away team won.*

Table 6 then shows us how many times the dynamic model was 'correct' in the sense that it gave the highest probability to the observed goal margin between two teams in a match. Since the model almost never gives the highest probability to a team scoring more than 3 goals, it won't give the highest probability to a big goal margin as well.

		frequency	correctly predicted
goal margin	$\geq$3	58	1
	2	34	7
	1	60	23
	0	58	29
	-1	35	7
	-2	25	3
	$\leq$-3	30	1
total		300	71

Table 6: *Results on how well the dynamic model predicted the correct goal margin as seen from the home team for matches in the 2018/2019 Eredivisie. The frequency column denotes how often a certain goal margin was observed, whereas the last column tells us how often the dynamic model gave this goal margin the highest probability of happening for these matches.*

The scatter plot of Figure 8 is divided into four different areas, so that we can see how well the dynamic model performed as predictor of the outcome of a match. Observations in area (b) and (d) correspond to matches where the expected goal margin actually pointed to the right winning team of a match. It can be seen from this that the dynamic model pointed to the correct winning team for most of the matches, since there are far more points in the areas (b) and (d) combined, than there are in (a) and (c) combined.

Since the model gives the expected number of goals both teams will score in a match as a rate parameter of the poisson distribution, we can calculate the probability the dynamic model gives to both teams winning the match by comparing those values.

Surprises happen in football and so we do not necessarily want the model to point to the correct winning team every match, we want it to give correct probabilities on how likely it is that a team wins a match. Table 7 gives us a better view on how well the dynamic model actually did in predicting match outcomes. Since the goal of the paper was to make money on 'Draw no Bet' wagers, we do not care about draws. We want to know which team would win, should one of the two win. The first column of the table states the probability range that the dynamic model gives to the team that is most likely to win that match, according to the model. The table shows us that for matches in which the model gave the favourite a probability of winning between 50% and 60%, the favourite only won in 47,9% of the matches. Optimally, this percentage would be between 50% and 60% as that would mean that the model did well in giving probabilities to which team wins a match. For the other ranges of percentages the model actually did this.

model probability for favourite	number of matches*	matches won by favourite	percentage of matches won by favourite
50% - 60%	48	23	47.9%
60% - 70%	58	36	62.1%
70% - 80%	46	34	73.9%
80% - 90%	56	47	83.9%
90% - 100%	34	32	94.1%
total	242	172	71.1%

Table 7: *The performance of the dynamic model displayed as follows: The first column shows the probability the model gave the favourite of the match to win, the second column tells how many matches of the 2018/2019 Eredivisie fell into this probability range, the third column tells us how many of these were actually won by the favourite and the last column denotes the corresponding percentage of matches won by the favourite. *Excluding draws.*

We did the same for the static model in Table 8. We see that the static model also did well, since the percentage of matches won by the favourite is in the correct percentage range in four of the five rows. It only did not as we wanted in the 60% - 70% range.

model probability for favourite	number of matches*	matches won by favourite	percentage of matches won by favourite
50% - 60%	62	36	58.1%
60% - 70%	47	27	57.4%
70% - 80%	53	40	75.5%
80% - 90%	53	45	83,9%
90% - 100%	27	25	94,1%
total	242	173	71.5%

Table 8: *The performance of the static model ($\xi = 0$) displayed as follows: The first column shows the probability the static model gave the favourite of the match to win, the second column tells how many matches of the 2018/2019 Eredivisie fell into this probability range, the third column tells us how many of these were actually won by the favourite and the last column denotes the corresponding percentage of matches won by the favourite. *Excluding draws.*

6.3 Betting Application

6.3.1 Betting strategy

By comparing the estimated result probabilities from our models with the bookmakers' probabilities for the Eredivisie season 2018/2019 matches, we can determine on which matches we should bet, assuming that our model is more accurate than the bookmakers. We can thereafter calculate if we would have

made money, or if we would have lost money betting on those matches.

To calculate the probabilities the bookmakers have on each result for a "Draw no Bet" wager, we need to use the odds they putted on the match. If we denote by $Odds^H$ and $Odds^A$ the odds corresponding to a home team win and away team win respectively, these transform to probabilities p^H and p^A for a home team win and away team win as follows:

$$p^H = \frac{\frac{1}{Odds^H}}{(\frac{1}{Odds^H} + \frac{1}{Odds^A})}$$
$$p^A = \frac{\frac{1}{Odds^A}}{(\frac{1}{Odds^H} + \frac{1}{Odds^A})}$$

The probabilities of bookmakers usually add up to a sum that is more than 1. This is standard in betting markets. This surplus of probability is called 'the bookmakers' take', which is equal to their expected profit if the bookmakers are accurate in their probability specifications. We rescale the probabilities of bookmakers so that they add up to one for every match. We define b_k^H as the scaled bookmakers' probability of a home win in match k and b_k^A as the scaled bookmakers' probability of an away win in match k. Similarly, we define $\hat{p}_k^H$ and $\hat{p}_k^A$ as the corresponding probabilities our model estimates for match k.

For every match played in the Eredivisie season 2018/2019, Figure 9 shows the comparisons of the probability estimates: each dot corresponds to one match [7]. Overall, there is a lot of agreement between the probabilities, but the variability of these plots indicates the potential for positive profit, if our model is a better predictor than the bookmakers.

If our model produces probabilities without any error, then the expected profit from a €1 stake bet on a home win is

$$E(\text{profit}) = €(\frac{\hat{p}_k^H}{b_k^H} - 1) \tag{17}$$

We will obtain a positive profit if our model probabilities are sufficiently more accurate than the scaled bookmakers' probabilities.
A natural betting strategy arising from (17) for match k is to bet on a home win if

$$\frac{\hat{p}_k^H}{b_k^H} > r \tag{18}$$

and on an away win if

$$\frac{\hat{p}_k^A}{b_k^A} > r, \tag{19}$$

[7] Excluded are the first two matches played by promoted teams: they had not played any matches before in our dataset, so we only included them starting from matchday 3.

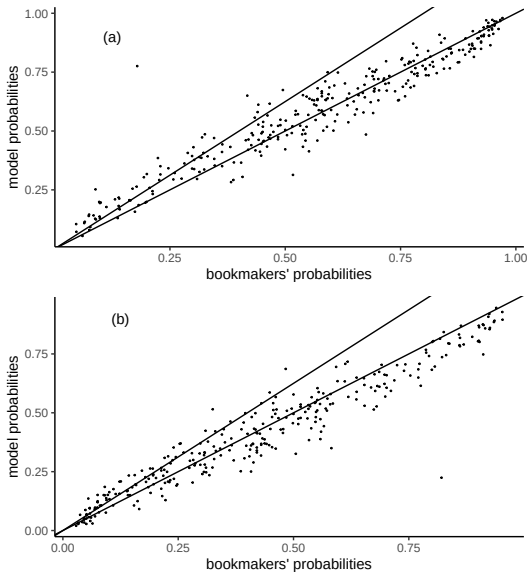

Figure 9: *Model probability estimates for a home win (a) and away win (b) plotted against bookmakers' probabilities for the 2018/2019 Eredivisie matches. Also, the lines* $r = \frac{\hat{p}}{b}$ *are added for* $r = 1$ *and* $r = 1.25$.

where r is a predetermined value greater than 1.

The higher we take r, the more difference there has to be in our models' probabilities and the bookmakers' probabilities for a bet to be placed. In Figure 9 the line $\hat{p} = rb$ is plotted for $r = 1$ and $r = 1.25$. After choosing a value for r, points above the line $\hat{p} = rb$ correspond to matches we should place a bet on.

6.3.2 Results

We can determine if our betting strategy would have been profitable by calculating the profit we would have obtained (using the results of the matches to see if a bet was won or not) for different values of r in (18) and (19). The green bars in Figure 10 show us the total profit we would have obtained using different values of r for our dynamic model. The maximum total profit is obtained for $r = 1.5$, where we stake €1 on 27 wagers to get a total profit of €34. For $r = 1.15$ we would get a similar value for the total profit (€33,15), staking €1 on 129 wagers. We can also see that we would have made a positive profit, no matter what value of r take, so that our dynamic model would have been profitable if it were used on the 2018/2019 Eredivisie season.
The red bars in Figure 10 correspond to the total profit we would have obtained

using the same betting strategy with the static model. For most values of r this model gives us a lower profit than the dynamic model, and for $r = 1$ this model even gives us a negative profit. However, for $r = 1.35$, the static model gives us a total profit of €33,61 staking €1 on 88 wagers, which is very close to the maximal profit we got from our dynamic model.

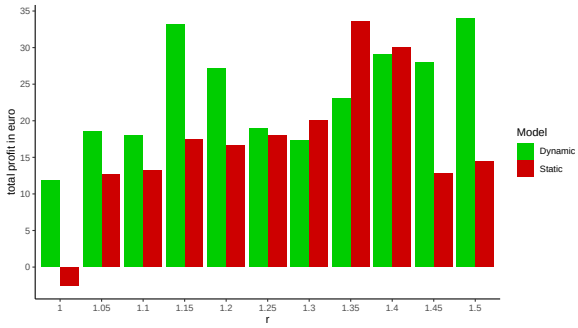

Figure 10: *Total profit for employing particular values of r, using the dynamic (green) and static model (red). A stake of €1 was used for every bet.*

Hence, we might believe that including the weighting of matches in our model, did not make an impact on our betting results. However, Figure 11 shows us the average profit per bet for different values of r for both the dynamic and the static model. The average profit per bet mostly increases by increasing r, but also lowers the numbers of wagers that are placed. We can see that the profit per bet for our dynamic model is bigger than that of the static model for every value of r that is used. We can also see that if we use $r = 1.5$ for the dynamic model, we have an astonishing return of €2,26 for every €1 that is staked.

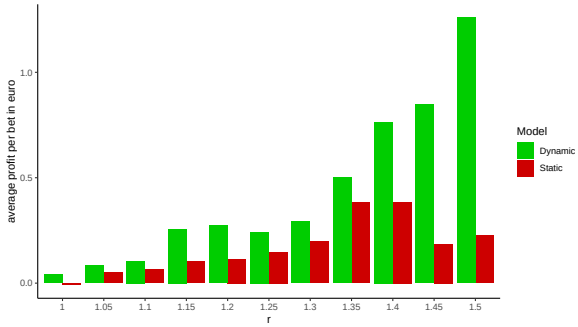

Figure 11: *average profit per bet for employing particular values of r, using the dynamic (green) and static model (red). A stake of €1 was used for every bet.*

Table 9 gives the characteristics for wagers placed in different ranges of $\frac{\hat{p}}{b}$ using the dynamic model. Interestingly, we lose money in most ranges of $\frac{\hat{p}}{b}$, but we make enough in the last columns to have a positive profit for every value of r. We can also see that most of the matches are actually predicted fairly close to the bookmakers' prediction, seeing that over 200 matches fall into the first two rows of the table.

range of $\frac{\hat{p}}{b}$	total wagers	total return	total profit	average profit per bet
$1 < \frac{\hat{p}}{b} < 1.1$	128	€121,83	-€6,17	-4.8%
$1.1 < \frac{\hat{p}}{b} < 1.2$	73	€63,90	-€9,10	-12.5%
$1.2 < \frac{\hat{p}}{b} < 1.3$	40	€49,81	+€9,81	$+24.5\%$
$1.3 < \frac{\hat{p}}{b} < 1.4$	21	€9,25	- €11,75	-56.0%
$1.4 < \frac{\hat{p}}{b} < 1.5$	11	€6,06	-€4,94	-44.9%
$\frac{\hat{p}}{b} > 1.5$	27	€61,-	+€34,-	$+125.9\%$

Table 9: *characteristics of the wagers we placed on matches of the Eredivisie 2018/2019 season using the dynamic model, subdivided by range of $\frac{\hat{p}}{b}$. A stake of €1 was used for every bet.*

We gathered the same characteristics for the static model in Table 10. We can see that there is only a positive return on the ranges $1.4 < \frac{\hat{p}}{b} < 1.5$ and $\frac{\hat{p}}{b} > 1.5$. For smaller values of $\frac{\hat{p}}{b}$, money is lost. Also, we see that there are bigger differences between the bookmakers' probabilities and the model probabilities now, as compared to when we used the dynamic model, because there are less matches placed in the first two rows than in Table 9.

range of $\frac{\hat{p}}{b}$	total wagers	total return	total profit	average profit per bet
$1 < \frac{\hat{p}}{b} < 1.1$	100	€84,23	-€15,77	-15.8%
$1.1 < \frac{\hat{p}}{b} < 1.2$	56	€52,66	-€3,34	-6.0%
$1.2 < \frac{\hat{p}}{b} < 1.3$	43	€39,57	-€3,43	-8.0%
$1.3 < \frac{\hat{p}}{b} < 1.4$	23	€12,98	- €10,02	-43.57%
$1.4 < \frac{\hat{p}}{b} < 1.5$	14	€29,63	+€15,63	$+111.6\%$
$\frac{\hat{p}}{b} > 1.5$	64	€78,40	+€14,40	$+22.5\%$

Table 10: *characteristics of the wagers we placed on matches of the Eredivisie 2018/2019 season subdivided by range of $\frac{\hat{p}}{b}$, using the static model. A stake of €1 was used for every bet.*

Table 11 shows us some more details when we are using the dynamic model to bet on matches for certain values of r. An interesting fact from this table is that the average odds of the wagers we place are quite large, which means that a lot of wagers that we place are on the underdog of the match to win, especially

for high values of r. Also interesting is that although the average odds of bets placed when using $r = 1.5$ are a lot bigger than the average odds when using $r = 1.3$, we win relatively more wagers when using $r = 1.5$.

r	total wagers*	number of of winning wagers (accuracy in %)	average odds of a winning wager	average odds of all wagers
1	242	103(42.6%)	2.46	4.56
1.1	143	41(28.7%)	3.93	6.02
1.2	88	24(27.3%)	4.80	7.55
1.3	52	10(19.2%)	6.93	8.88
1.4	37	9(24.3%)	7.34	9.41
1.5	26	7(21.9%)	8.57	10.67

Table 11: *More characteristics of the wagers placed using the dynamic model for particular values of r, including the total bets placed, the accuracy and information on the odds of the wagers. *Excluding the wagers on matches that ended up as a draw.*

If we want to take a better look at when in the season we make the most money, we can divide the season up into four periods: period 1 ranges from matchday 1 up until matchday 9, period 2 ranges from matchday 10 up until matchday 17, period 3 ranges from matchday 18 up until matchday 26 and period 4 ranges from matchday 27 up until matchday 34.
Figure 12 now shows us the profits we made on every period using different values of r, employing the dynamic model. We see immediately that a lot of money was lost during the first period. A reason for this could be that our model does not take the summer transfer market into account, in which teams can strengthen or weaken their teams significantly by buying and selling players. Once our model is through the first period, money gets made. While the profits for $r = 1$ and $r = 1.15$ behave similarly (with $r = 1.15$ getting higher profits), we lose a lot less money in the first period using $r = 1.5$.

The results we get from using our model on 'Draw no Bet' wagers of the 2018/2019 Eredivisie season are promising. However, we could have been lucky to have picked a season in which making money was easy. Therefore, Table 12 shows the result our dynamic model gave for $r = 1$, (g), against some other basic betting strategies that we could have used on the matches of the 2018/2019 Eredivisie season. While we would still have a positive profit for two of these strategies, our model gives the best total return of all the strategies. Note that using $r = 1.5$ would triple the profit, so that we conclude this section by saying that our model was indeed highly profitable when using it on 'Draw no Bet' wagers on the 2018/2019 Eredivisie season.

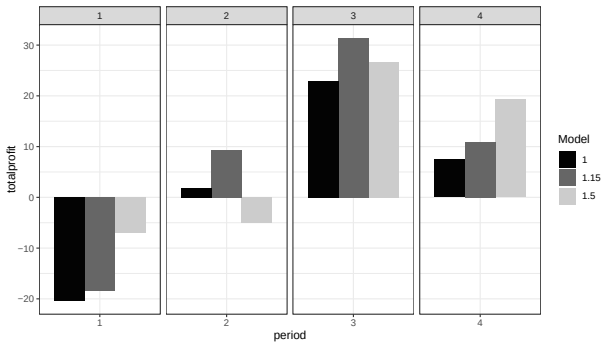

Figure 12: *profit of all four periods of the 2018/2019 Eredivisie for different values of r using the dynamic model.*

betting strategy	total return	total profit
(a)	€306,77	+ €6,77
(b)	€284,57	- €15,43
(c)	€292,29	- €7,71
(d)	€305,06	+ €5,06
(e)	€285,03	- €14,97
(f)	€297,47	- €2,53
(g)	€311,85	+ €11,85

Table 12: *Total return and profit if a €1 bet was placed on every match (300) of the 2018/2019 Eredivisie as follows: (a) hometeam always wins, (b) awayteam always wins, (c) favourite (lowest odds) always wins, (d) underdog (highest odds) always wins, (e) team that ranked higher last season always wins, (f) static model ($\xi = 0$) with $r = 1$ and (g) dynamic model ($\xi = 0.0017$) with $r = 1$.*

7 Discussion

The goal of this paper was to develop a model that would predict the outcome of matches in the 2018/2019 Eredivisie season in such a way that we would make profit by betting money on 'Draw no Bet' wagers. As we have shown in the preceding chapter, we managed to succeed: For a range of values for r in (18) and (19), a positive return was found, if the model we created was used on the 2018/2019 Eredivisie.

7.1 Reliability of results

Although the results are thus satisfying on a first look, we should address certain limitations of the model.
First of all, we saw in Chapter 7 that our total profit would have been maximized by using $r = 1.25$ in our betting strategy. We could not have said this if the matches had not been played yet. However, as we have seen in Figure 10, we would have made profit for every value of r between 1 and 1.5.
Also, throughout this paper we have modelled the scores of the home team and the away team in the same match independently from each other. Maher [23] stated in his paper that there might be a correlation between the number of goals scored by both teams, and later the existence of this correlation has been proven as well (see Karlis and Ntzoufras (2000) [21]).
We could have accounted for this correlation by modelling the match score as a bivariate Poisson distribution, with the number of goals the home team scores as one variate, and the number of goals the away team scores as the other. However, the distribution of the difference of the two variates in a bivariate Poisson distribution, which is called a Skellam distribution, is invariant to the correlation coefficient. Therefore, the outcome of a match does not depend on the correlation between the number of goals scored by both teams. By that reasoning we have not used the bivariate Poisson distribution in our paper: We were only interested in the outcome of a match and not in the actual final score, since we wanted to use our model for 'Draw no Bet' wagering.
Finally, we should also note that the results are based on just one Eredivisie season, using the model on 300 matches. We can see from Table 9 that the results vary a lot for different ranges of $\frac{\hat{p}}{b}$, so the model should be tested on more seasons to say it is indeed a reliable profitable model.

7.2 Future improvements

The simplicity of the model presented is appealing. However, modifications to the model might help getting even better results.
In Chapter 6, we eventually abandoned the idea that home advantage is different for different teams, and we went straight to a constant home advantage parameter that is the same for every team. In the 2018/2019 Eredivisie season,

six of the eighteen participating teams played their home matches on an artificial pitch. It is shown by Barnett and Hilditch [5] that the home advantage of a team playing their home matches on artificial turf is relatively bigger than the home advantage of a team that plays their matches on normal grass. Their research was only based on English football results, so it will not necessarily improve the model, but a different home advantage parameter for teams playing their home matches on an artificial pitch could improve the model.

Another way to improve the model would be to add additional covariates. Although our model does include the form of a team by weighting the matches by date, we do not model the 'matchday situations' of teams. By 'matchday situations' we mean for example the injuries of starting players and suspensions due to an overload of cards.
Also, it should be looked at how one could implement transfers in the model. When a team buys a very good new player, they will usually be better than before. When they sell a very good player, they will usually be worse than before. Our model does not account for this at the moment.

Lastly, we used a very basic betting strategy, in which we stake €1 on every bet we place. There might be more complex betting strategies, with stake adjusted for probabilities that could help optimize the profit. However, this will only theoretically improve the model. In practice, if an individual is betting non-integer stakes on online betting sites, and wins more than he loses, he will be removed from the betting sites.

References

[1] Historical 'draw no bet' odds on eredivisie matches. `https://www.oddsportal.com/soccer/netherlands/eredivisie/results/`. Accessed: 2019-08-31.

[2] Historical football results and betting odds data. `https://www.football-data.co.uk/data.php`. Accessed: 2019-08-29.

[3] Htrotugu Akaike. Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, 60(2):255–265, 1973.

[4] Jim Albert, Mark E Glickman, Tim B Swartz, and Ruud H Koning. *Handbook of Statistical Methods and Analyses in Sports*. CRC Press, 2017.

[5] V Barnett and S Hilditch. The effect of an artificial pitch surface on home team performance in football (soccer). *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 156(1):39–50, 1993.

[6] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[7] Kenneth P Burnham and David R Anderson. Model selection and multimodel inference: A practical information-theoretic approach. 2002.

[8] A Colin Cameron and Pravin K Trivedi. Regression-based tests for overdispersion in the poisson model. *Journal of econometrics*, 46(3):347–364, 1990.

[9] DB Campbell and CA Oprian. On the kolmogorov-smirnov test for the poisson distribution with unknown mean. *Biometrical Journal*, 21(1):17–24, 1979.

[10] Kerry S Courneya and Albert V Carron. The home advantage in sport competitions: a literature review. *Journal of Sport & Exercise Psychology*, 14(1), 1992.

[11] James Curley. *engsoccerdata: English and European Soccer Results 1871-2016*, 2016. R version 0.1.5.

[12] Roger R Davidson. On extending the bradley-terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329):317–328, 1970.

[13] Mark J Dixon and Stuart G Coles. 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, 1997.

[14] Annette J Dobson and Adrian G Barnett. *An introduction to generalized linear models*. Chapman and Hall/CRC, 2008.

[15] Arpad E Elo. *The rating of chessplayers, past and present*. Arco Pub., 1978.

[16] Fumio Hayashi. Econometrics. 2000. *Princeton University Press. Section*, 1:60–69, 2000.

[17] Giselmar AJ Hemmert, Laura M Schons, Jan Wieseke, and Heiko Schimmelpfennig. Log-likelihood-based pseudo-r 2 in logistic regression: Deriving sample-sensitive benchmarks. *Sociological Methods & Research*, 47(3):507–531, 2018.

[18] ID Hill. Association football and statistical inference. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 23(2):203–208, 1974.

[19] Lars Magnus Hvattum and Halvard Arntzen. Using elo ratings for match result prediction in association football. *International Journal of forecasting*, 26(3):460–470, 2010.

[20] Dirk Bernard Jochems. Voetbalprognoses. *Statistica Neerlandica*, 12(1):17–31, 1958.

[21] Dimitris Karlis and Ioannis Ntzoufras. On modelling soccer data. *Student*, 3(4):229–244, 2000.

[22] Sadanori Konishi and Genshiro Kitagawa. *Information criteria and statistical modeling*. Springer Science & Business Media, 2008.

[23] Michael J Maher. Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118, 1982.

[24] Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[25] Daniel McFadden et al. Conditional logit analysis of qualitative choice behavior. 1973.

[26] Abraham De Moivre. De mensura sortis, seu, de probabilitate eventuum in ludis a casu fortuito pendentibus. *Philosophical Transactions of the Royal Society of London*, 27(329):213–264, 1711.

[27] Michael Joseph Moroney. Facts from figures. 1956.

[28] Frederick Mosteller. Remarks on the method of paired comparisons: Ii. the effect of an aberrant standard deviation when equal standard deviations and equal correlations are assumed. *Psychometrika*, 16(2):203–206, 1951.

[29] John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

[30] Siméon Denis Poisson. *Recherches sur la probabilité des jugements en matière criminelle et en matière civile precédées des règles générales du calcul des probabilités par sd poisson*. Bachelier, 1837.

[31] Charles Reep and Bernard Benjamin. Skill and chance in association football. *Journal of the Royal Statistical Society. Series A (General)*, 131(4):581–585, 1968.

[32] Charles Reep, Richard Pollard, and Bernard Benjamin. Skill and chance in ball games. *Journal of the Royal Statistical Society: Series A (General)*, 134(4):623–629, 1971.

[33] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[34] Mark Thompson. On any given sunday: Fair competitor orderings with maximum likelihood methods. *Journal of the American Statistical Association*, 70(351a):536–541, 1975.

[35] Louis L Thurstone. A law of comparative judgment. *Psychological review*, 34(4):273, 1927.

A R code

The complete `R` script that was used for this paper can be found here.