Beating the bookie in the football betting market without using extensive match data

Barend Verkerk
Research Paper Msc Business Analytics

June 2018

Abstract

The goal of this research is to build a profitable betting system while only using basic historical match data. Two regression methods are used to predict the outcome of football matches: Poisson regression and ordered logistic regression. The performance of the models is evaluated and compared to the performance of a bookmaker over more than eight seasons in the Dutch Eredivisie. In addition, the best models are chosen to simulate placing bets on matches with a positive expected profit. Different betting strategies are evaluated in this simulation. A random betting system that replicates the betting behavior of the different betting strategies is used to evaluate the significance of the realized profits. The results show that it is possible to beat the bookmaker in most seasons.

Contents

1. Introduction

Association football is ranked among the most popular sports worldwide. From the wealthy suburbs of Amsterdam to the the favelas of Rio de Janeiro, everywhere around the world people enjoy the game of football. The entire Dutch nation colors orange when their national football team participates in the World Cup. No less than 50% of the Dutch population watched the national squad play their World Cup semi-final against Argentina in 2014 [1]. The famous coach Arsène Wenger thinks that the sport is so popular because the results are more unpredictable than in any other sport [2]. The unpredictability of football matches is – among other things – caused by the low average number of goals, which is below 3 per match [3]. The combination of its popularity and the unpredictable nature of the results make football the biggest sport to bet on [4]. Most people place their bets for fun, making decisions based on their own perceptions. However, despite the millions of people placing their recreational bets once in a while, the gambling world is serious business for the bookmakers. The (mostly professional) bettors that can outperform the bookmaker on the long-term are a real threat. Some of these bettors exploit their excellent knowledge about the sport, others use statistical models to predict the expected match outcomes.

A lot of papers have been written about predicting football match outcomes. There are two kind of models that are mainly used. On the one hand there are the goal models. This kind of models predict the number of goals that the home and away team will score in their upcoming match. Knowing the expected number of goals for both teams, one can estimate the likelihood of different match outcomes. Maher [5] found that the number of goals per match can be estimated by the Poisson distribution. He showed that the number of goals of both teams follow independent Poisson distributions. The estimated attacking and defensive parameters of both teams represent the means of the Poisson distributions. This idea is further improved by

Dixon and Coles [6]. One of their main improvements is the allowance for fluctuations in team performance. They state that a team's attacking and defensive parameters are not constant during the season. These fluctuations could be caused by an important player getting injured, or the manager getting sacked. Therefore Dixon and Coles state that the attacking and defensive strength parameters of football teams are time-dependent and thus vary during the season. Several recent studies have opted for a different approach in predicting football match outcomes: the toto-models. With this new approach match results can directly be predicted. This can be done by an ordered probit regression model, as proposed by Koning [7]. The same method is applied by Goddard and Asimakopoulos [8]. They then compared the model's performance to the bookmakers and were able to compete with them on four different seasons in the English Premier League.

Although there has been a lot of interest in predicting match outcomes, less attention has been devoted to betting strategies. One of the few who did investigate different strategies is Langseth [9], who explained and tested five of the most popular betting strategies. The easiest among these strategies is the fixed-bet strategy, where a fixed amount of money is placed on each match with a positive expected return. Another popular strategy is the Kelly Criterion, introduced by Kelly [10]. This betting strategy suggests that the stake should be proportional to the presumed edge and the estimated probability of winning. However, Langseth found that none of the five betting strategies significantly outperformed the other in the English Premier League during the seasons 2011/2012 and 2012/2013.

In a global betting industry that is worth billions, there is a constant need for bookmakers to improve their predictions. Goddard and Asimakopoulos already mentioned the increased efficiency of the bookmakers in 2003. At that time, the available data was limited to things like the final result and the number of shots fired by both teams. In recent years the available data has exponentially increased. Companies like Opta provide a detailed description of all match actions (and the corresponding coordinates) in over 1,000 leagues worldwide. Bookmakers have access to these detailed data and therefore it is expected that their efficiency has increased even more since the paper of Goddard and Asimakopoulos in 2003.

The aim of this paper is to investigate whether it is still possible to beat the bookmakers in the football betting market without using extensive match data. The data contains basic match information over more than eight sea-

sons in the Dutch Eredivisie. Multiple statistical models – almost solely based on historical match results – will be performing against a bookmaker. A number of betting strategies will be evaluated.

This paper is organized as follows. Section 2 provides a description of the data that is used in this research. Next, the methodology is treated in Section 3. Thereafter, the results are presented in Section 4. Finally the conclusions about this research are drawn in Section 5.

2. Data

2.1 Data collection

2.1.1 Basic match data

The data that is used in this paper originates from three different sources. The data mainly contains publicly available historical match data gathered from football-data.co.uk. This website provides historical results and betting odds for many different European football competitions. In this paper data from the last 13 seasons in the Dutch Eredivisie is used. The complete dataset consists of 3,832 rows and each row describes one match. The potentially useful variables that can directly be extracted from the dataset are shown in Table 2.1 on the next page.

Variable	Description
date	the date at which the match is played
season	the season at which the match is played
home	the team that plays at home
away	the team that plays away from home
B365H	the odds offered by bookmaker Bet365 for a home team win
B365D	the odds offered by bookmaker Bet365 for the draw to occur
B365A	the odds offered by bookmaker Bet365 for an away team win
fthg	the full time number of goals scored by the home team
ftag	the full time number of goals scored by the away team
ftr	the full time result (H=home, D=draw, A=away)

Table 2.1: Variable description of the football-data.co.uk dataset.

2.1.2 Club strength

The above data contains the most essential information about all football matches played over a range of 13 seasons, but still lacks a strength indicator per club. Such an indicator is able to estimate the ability of all teams at a certain point in time and can form the basis of a statistical model. In this paper the increasingly popular Elo rating system [11] is used. This system was invented in 1960 as an improved chess rating system, but nowadays it serves as a football rating system as well. For example, the FIFA World Ranking (the official ranking of all men's national football teams in the world) is based on an Elo inspired method since after the World Cup of 2018. The Elo rating system can also be applied to clubteams. Although it is possible to create such a system with the available data from Table 2.1, there exists a publicly available Elo system for clubteams already. Schiefler [12] implemented club rankings for most of the European club competitions. Daily historical rankings can be accessed via an API.

Another strength indicator is the annual budget per club. After all, the richest clubs have the most resources to build up their squad. For all 13 seasons the annual budgets of all clubs are manually gathered from Wikipedia. Both strength indicators add the following 4 variables to each row of the dataset from Table 2.2 on the next page.

Variable	Description
homeElo	the Elo rating of the home team
awayElo	the Elo rating of the away team
homeBudget	the annual budget of the home team
awayBudget	the annual budget of the away team

Table 2.2: Variable description of both club strength indicators.

2.1.3 Artificial turf

There is an increasing number of Dutch Eredivisie clubs playing their home matches on artificial turf (7 in the season 2017/2018). Van Ours [13] showed that clubs in the Dutch Eredivisie playing their home matches on artificial turf have an advantage over clubs that play on grass of approximately 4 points per season. Therefore it could be beneficial to add the information about the home ground surfaces of all clubs in the Dutch Eredivisie. The information about which clubs play their home matches on artificial turf is manually gathered from Wikipedia. This adds the following variables to each row of the dataset (Table 2.3):

Variable	Description
homeTurf	1 if the home team plays their home matches on artificial turf, 0 if they do not
awayTurf	1 if the away team plays their home matches on artificial turf, 0 if they do not

Table 2.3: Variable description of the home surface data.

2.2 Feature engineering

The dataset now contains two long-term measures of club strength. One strength factor is constant throughout the season (annual budget), while the other is updated after each match (Elo rating). Although the Elo ratings are updated weekly, the ratings are also based on matches in the distant

past. Matches played years ago still have influence on the current Elo rating of a club. Therefore it might be useful to create a number of variables that measure recent form only. An appropriate measure can be overall form (i.e. goals scored, goals conceded, points won) over the last $2, 5, ..., 17$ matches that a club played this season. This is not always the best measure; some teams perform significantly better at home than away from home. In that case their probability of winning a home match is relatively high. For scenarios like this we also calculate the home form of the home team and the away form of the away team over the last $2, 5, ..., 11$ matches that each club played this season. However, just counting all the recent form measures is not adequate enough, since this would score teams with a high winning probability the same as the underdog. A solution to this is to multiply the results of all clubs by their predicted probability (bookmaker odds) of winning the match. With this adjustment short-term luckiness gets assigned lower scores than confirmed superiority (i.e. if the favorite wins). Teams are rewarded higher scores when it is expected that they can continue their current form. An example of the overall form measure is shown below:

	FC Gronigen	**PSV**
Estimated probability of winning	0.125	0.688
Goals scored (added to form measure)	3 (0.375)	3 (2.064)
Goals conceded (added to form measure)	3 (2.064)	3 (0.375)
Points (added to form measure)	1 (0.125)	1 (0.688)

As you can see, the points assigned to FC Groningen (the underdog) are less than for PSV, while they actually both scored three goals and won one point. However, chances are that FC Groningen was just lucky this time. On the other hand, if FC Groningen are the favorites in their next match, scores will be assigned accordingly when they win.

The overall form variables that are added to the dataset are shown in Table 2.4 on the next page. For the home and away form measures the added variables are not shown, but they are similar. The only difference is that these measures are calculated for the last $2, 5, ..., 11$ matches per team.

Variable	m	Description
overall_HG$_m$	2, 5, ..., 17	Score for the goals scored by the home team in their last m matches.
overall_AG$_m$	2, 5, ..., 17	Score for the goals scored by the away team in their last m matches.
overall_HA$_m$	2, 5, ..., 17	Score for the goals conceded by the home team in their last m matches.
overall_AA$_m$	2, 5, ..., 17	Score for the goals conceded by the away team in their last m matches.
overall_HP$_m$	2, 5, ..., 17	Score for the points won by the home team in their last m matches.
overall_AP$_m$	2, 5, ..., 17	Score for the points won by the away team in their last m matches.

Table 2.4: Variable description of the overall recent form measures.

3. Methods

When it comes to predicting football matches, two kinds of models are often used: goal models and toto models. Goal models assume that the occurrence of goals can be estimated by certain probability distributions. Match outcomes can indirectly be predicted by the expected goals of both teams. Toto models use a different approach and are able to predict match outcomes directly. In this research two goal models (Poisson regression) and two toto models (logistic regression) will be implemented.

3.1 Generalized Linear Models

Poisson regression and logistic regression are both generalized linear models (GLM). The goal of a GLM is to estimate $\boldsymbol{\beta}$, a parameter vector with intercept β_0. A generalized linear model has a one-to-one link function $g(\mu_i) = \eta_i = \boldsymbol{x_i^T}\boldsymbol{\beta}$, where $\mu_i = E(Y_i)$ for observation Y_i, $i = 1, ..., n$. For Poisson regression this link function equals $\log(\mu_i)$ and for logistic regression this link function equals $\log[\mu_i/(1-\mu_i)]$. An estimate of $\boldsymbol{\beta}$ is the one that maximizes the log-likelihood function:

$$l(\theta) = \sum_{i=1}^{n} \left(\frac{y\theta_i - b(\theta_i)}{\phi/A_i} + c(y, \phi/A_i) \right)$$

3.2 Goal models

3.2.1 Poisson distribution

The first kind of model that will be treated in this research is the goal model. Maher [5] found that the number of goals scored by both teams can be estimated with the Poisson distribution. Some other papers opt for the negative binomial distribution, because of overdispersion. However, Maher stated that a negative binomial distribution may arise from the aggregate of Poisson-distributed scores where each team in the competition has a different mean (i.e. strength). Therefore we will assume in this paper that the occurrence of goals can be approached by a Poisson distribution.

In Poisson regression it is assumed that the independent response variables Y_i, $i = 1, ..., n$ are Poisson distributed with mean μ_i. In football matches there are two teams that score their goals according to a Poisson distribution. However, the number of goals that both teams score during the match is not entirely independent. In fact, their Pearson correlation coefficient is approximately -0.145. Therefore it could improve the model slightly if we implement a bivariate Poisson model as proposed by Maher [5]. However, in this paper we will assume independence between the goals of both teams because of the small correlation coefficient. Since we assume independence the home

and away goals can be modeled separately. The number of home goals follows the following Poisson distribution: $HG_i \sim \text{Poisson}(\mu_{HG_i})$. In the same way the number of away goals can be estimated by $AG_i \sim \text{Poisson}(\mu_{AG_i})$.

3.2.2 Skellam distribution

Although we are able to calculate the expected goals of both teams for every match in the Dutch Eredivisie, the probabilities of all possible match results are not known yet. This can be accomplished by the Skellam distribution, which is known to be the difference of two Poisson-distributed random variables N_1 and N_2 with expected values μ_1 and μ_2 respectively. The density function of this distribution function is given by the convolution of two Poisson distributions with $k = \mu_1 - \mu_2$:

$$f(k; \mu_1, \mu_2) = e^{-\mu_1+\mu_2} \left(\frac{\mu_1}{\mu_2}\right)^{k/2} I_k(2\sqrt{\mu_1\mu_2})$$

Here $I_k(z)$ is the modified Bessel function of the first kind. The Skellam distribution can be applied to football matches if the difference in expected goals $k = \mu_{HG_i} - \mu_{AG_i}$ is taken. The probabilities of the three different outcomes for this match are then described as follows:

$$P(\text{home}) = P(k > 0), \quad P(\text{draw}) = P(k = 0), \quad P(\text{away}) = P(k < 0)$$

3.3 Toto models

3.3.1 Ordered logistic regression

The next kind of model that will be treated in this research is the toto model, for which logistic regression can be used. Since football matches have three possible outcomes the binary logistic regression method is not applicable. Other regression methods are available to solve this problem. For example, Koning [7] used an ordered probit model for predicting football matches. In

this paper the ordered logistic regression method will be used. This method can be described as follows. It is already mentioned that each football match outcome has three categories. The ordered logistic model uses two threshold values to separate these three categories. This makes the categories binary at each of the thresholds. It is know that for GLM's the link function equals $g(\mu_i) = \eta_i = \boldsymbol{x_i^T}\boldsymbol{\beta}$, with $i = 1, ..., n$. For an ordered logistic regression model with J categories we have to add the thresholds α, which causes the link link function to change as follows: $g(\mu_i) = \text{Logit}[P(Y_i \leq j)] = \alpha_j - \boldsymbol{x_i^T}\boldsymbol{\beta}$, with $i = 1, ..., n$ and $j = 1, ..., J-1$. Since the cumulative probabilities are increasing and $P(Y_i) = 1$ only $J-1$ probabilities have to be modeled. If this is applied to a football with ordered categories away (1), draw (2) and home (3), the estimated log odds of categories 1 and 2 can be calculated as follows:

$$\begin{aligned} \text{Logit}[P(Y_i \leq 1)] &= \alpha_1 - \boldsymbol{x_i^T}\boldsymbol{\beta} \quad \text{(away)} \\ \text{Logit}[P(Y_i \leq 2)] &= \alpha_2 - \boldsymbol{x_i^T}\boldsymbol{\beta} \quad \text{(draw)} \end{aligned}$$

The corresponding probabilities can be derived by taking the inverse logit: $P(Y_i \leq j) = \exp(\alpha_j - \boldsymbol{x_i^T}\boldsymbol{\beta})/[1 + \exp(\alpha_j - \boldsymbol{x_i^T}\boldsymbol{\beta})]$. If this is applied to the above two equations all three probabilities can easily be obtained.

3.4 Train set size

In this paper a fixed-size train set will be used to train a model and predict the next match with this model. The size of the train set will be h matches, so that the train set for some match i contains the range of matches $\{i - h - 1, i - h, ..., i - 2, i - 1\}$. According to Hamadani [14] the explanatory variables of a model change each season in American football. Although this is a different sport, it could be beneficial to test multiple values for h in order to determine the optimal train set size. The used explanatory variables are the same for all seasons. This makes the models perform similar in each season, which allows for a better examination of the development of the bookmaker efficiency. It should be mentioned that the optimal variable coefficients do actually change, since the train set is constantly moving per predicted match. In this way the model still adjusts for season-dependent changes in optimal model parameters.

3.5 Explanatory variables

For both methods (Poisson regression and ordered logistic regression) two kinds of models are introduced: a basic model and an extended model. The basic model is solely based on the estimated team strength parameters (i.e. long-term performance indicators), without looking at more detailed information. This can be modeled by two explanatory variables: the difference in Elo rating and annual budget between the home team and the away team.

Variable	**Description**
eloDiff	homeElo − awayElo
budgetDiff	homeBudget − awayBudget

Table 3.1: Explanatory variables used for the basic models.

The basic model can be used as a basis for the extended model. Although the Elo ratings already account for recent form, matches from seasons ago still influence the current Elo rating of a club. Therefore, the basic model lacks an adjustment for recent form changes over the last series of matches that a team played. The extended model will capture this by adding multiple variables about a team's recent form in the current season. Another variable that will be used in the extended model is the surface on which each team plays its home matches, in order to investigate the influence of artificial turf. Note that there are actually three distinct extended models: two models for Poisson regression (home goals and away goals) and one model for the ordered logistic regression method.

Since the recent form measures are sequential (see section 2.2) there is a large dependency between the variables within each recent form measure. An example for one of the recent form measures (the difference in points won between both teams) is shown in Figure 3.1 on the next page. The numbers after *Po_diff* in the figure indicate the number of previous matches that is used for calculating that variable (i.e. *Po_diff11* uses information about the previous 11 matches of both teams). It can be seen that the correlation between many variables is quite high, often even equal to 1. The other recent form measures shown the same pattern.

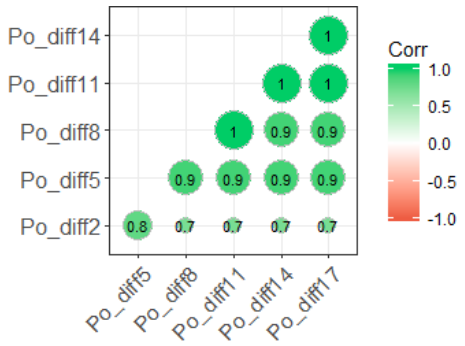

Figure 3.1: Correlogram of all the *Po_diff* variables.

Because of the large dependency between many variables, it is not necessary to try all possible combinations of variables for the extended models. A couple of different variables per recent form measure is sufficient. The method used to determine the best extended models is similar to the stepwise forward selection method. However, since the best combination might not reveal itself using this method, some combinations that have not been tried with the stepwise forward selection method are also manually evaluated. The main restriction for testing a specific combination manually is that the dependence between the variables should be small.

3.6 Model evaluation

3.6.1 Log Likelihood

All models will be evaluated by a number of different evaluation methods. The first one is the Log Likelihood (or Binomial Deviance) statistic. This evaluation metric is used to evaluate the predictions of Chess game outcomes in a Kaggle competition [15] and can easily be applied to football. For each model the average Log Likelihood statistic is calculated per season. The estimated probabilities of all outcomes for a certain match i can be applied to calculate an expected match score $\mathrm{S}_i = P_i(\text{home}) + 0.5 \cdot P_i(\text{draw})$. When we know the true result Y_i, the Log Likelihood statistic can be calculated

as follows: $LL_i = -Y_i \cdot \ln(\mathrm{S}_i) + (1 - Y_i) \cdot \ln(1 - \mathrm{S}_i)$. The true result Y_i is 1 for a home win, 0.5 for a draw, and 0 for an away win. When all matches in a season are evaluated, the average of all scores is calculated to obtain the Log Likelihood statistic for the whole season. This metric heavily penalizes large differences between the predictions and the actual outcomes. That shows a good fit with this research, since we do not want to make big mistakes with our predictions. Betting on matches without having an edge on the bookmaker can turn out to be a costly misjudgment.

3.6.2 Root Mean Squared Error

As similar reasoning applies to the next evaluation method: the Root Mean Squared Error (RMSE). For each model the RMSE is calculated per season. This measure can be seen as the root of the average MSE of the three possible match outcomes. The first step is to calculate the squared error (SE) for each match individually:

$$\mathrm{SE}_i = \begin{cases} \big(P(\text{home}) - 1\big)^2 + P(\text{draw})^2 + P(\text{away})^2 & \text{if result is home win} \\ P(\text{home})^2 + \big(P(\text{draw}) - 1\big)^2 + P(\text{away})^2 & \text{if result is draw} \\ P(\text{home})^2 + P(\text{draw})^2 + \big(P(\text{away}) - 1\big)^2 & \text{if result is away win} \end{cases}$$

When all matches in a season are evaluated, the RMSE can be calculated by the following formula: $\mathrm{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{N} \mathrm{SE}_i}$.

3.7 Betting strategies

In this section two betting strategies will be introduced. Creating a profitable betting system does not only depend on the quality of the models. Different betting strategies can have a positive effect on the profitability of the system. Let us first define the main restriction on betting on a certain match: a bet will only be placed if the expected profit is greater than zero. This holds whenever our predicted probability P is greater than the bookmaker's predicted probability P_{BOOK} for one of the three different match

outcomes. The difference between both predictions for some match i is the expected edge $E_i(\text{prf})$ that we have on the bookmaker. These matches are the only ones considered placing a bet on.

3.7.1 Fixed bet

The first and most obvious strategy is the fixed bet. If the expected profit is greater than a certain threshold level T, with $T \geq E_i(\text{prf})$, a fixed-size bet of size S will be placed.

3.7.2 Kelly Criterion

The second betting strategy is called the Kelly Criterion. This strategy suggests that the percentage of your bankroll that should be bet is equal to $\big(B \cdot P - (1 - P)\big)/B$, where B is equal to the decimal odds -1. For example, take a match for which the probability of the home team winning is 50%. Then the decimal odds are equal to $1/0.5 = 2$ and $B = 1$. The Kelly Criterion increases the stake size S whenever the expected profit increases. The stake size is positively related to the probability of success as well. This betting strategy is preferred if you want to grow your bankroll more quickly. However, this come with more risk of getting bankrupt. A couple of adjustments can be made to make this strategy more safe. For example, one can restrict the maximum percentage of bankroll that will be bet. Another option is to add a threshold T to the expected profit, so that bets only will be placed whenever $T \geq E_i(\text{prf})$.

3.8 Betting evaluation

After all models are evaluated by the methods described in section 3.6, the best regression model is chosen for each of the two regression methods. The betting strategies introduced in section 3.7 will then be applied to both models in order to compete with the bookmaker. In this research one performance metric will be used to evaluate the betting performance of the

models: the return on investment (ROI). Next, a bootstrap method is added to estimate whether the betting performance could have been caused by random luck. This bootstrap method can best be explained by an example: suppose the betting strategy has placed 500 bets out of 5,000 possible matches and the ROI was 2%. Out of these 500 bets, 300 bets were placed on the home team, 150 on the draw and 50 on the away team. The bootstrap method then aims to randomly replicate the same behavior. Therefore this method now randomly picks (without replacement) 500 matches out of the total of 5,000 matches. The same distribution of home team, draw and away team bets is randomly divided over these 500 selected matches. Now the exact distribution of simulated bets is determined, it is easy to calculate the corresponding simulated ROI. Replicate this procedure 10,000 times, so that the probability of luck for the true performance can be estimated.

4. Results

4.1 Extended model parameters

It is already mentioned that the Poisson regression models and the ordered logistic regression models both have a basic and an extended version. The basic version of both regression methods is already described in the methods section. All used variables that are not yet introduced in this paper are shown in Table 4.1 on the next page. The optimal extended models for the Poisson home goals and away goals can be explained by the diagrams of Figures 4.1 and 4.2 on page 17. One can observe that the expected home goals are significantly influenced by the difference in both teams home/away form in the last 5, 8 and 11 matches. The overall recent form does not improve the basic model for the expected home goals. This is in contrary to the extended away goals model, which is slightly improved by the overall recent form of both teams in their last 2 matches. It is also remarkable that the home goals model seems to be influenced more by the recent form of both teams (up

Variable	Description
HoAw_G_mDiff	the difference in goals scored by the home team in their last m home matches and the away team in their last m away matches this seasons
HoAw_A_mDiff	the difference in goals conceded by the home team in their last m home matches and the away team in their last m away matches this seasons
HoAw_P_mDiff	the difference in points won by the home team in their last m home matches and the away team in their last m away matches this seasons
Overall_G_mDiff	the difference in goals scored by the home team in their last m matches and the away team in their last m matches this seasons
Overall_A_mDiff	the difference in goals conceded by the home team in their last m matches and the away team in their last m matches this seasons
Overall_P_mDiff	the difference in points won by the home team in their last m matches and the away team in their last m matches this seasons

Table 4.1: Explanatory variables used for the extended models.

to 11 matches) when compared to the away goals model (up to 2 matches). Another remarkable fact is that the artificial turf variables did not influence the performance of the models at all. However, this could be explained by the paper of Van Ours [13], in which he concluded that the teams playing on artificial turf only win about 4 extra points per season. Therefore the influence of artificial turf per individual match could be negligible.

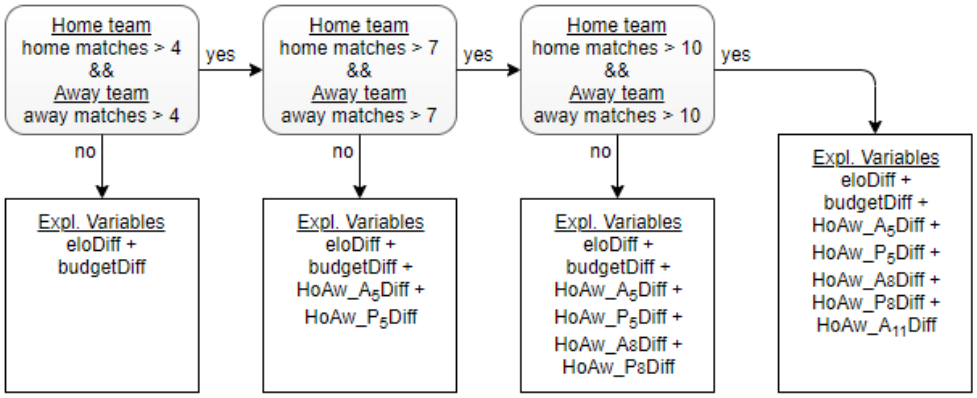

Figure 4.1: Diagram of the extended Poisson home goals model.

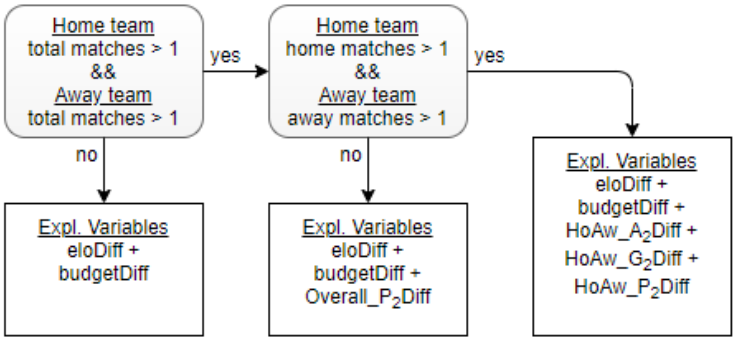

Figure 4.2: Diagram of the extended Poisson away goals model.

The diagrams for the extended version of the ordered logistic regression models is shown in Figure 4.3 on the next page. The difference with the Poisson models is that OLR directly predicts the probabilities of different match outcomes. This can be modeled with one model that predicts the probabilities of all full time results (home win, draw and away win). The most important added variables with respect to the basic model are the differences between both team's won points in their last 8 and 11 matches. Again, the artificial turf variables did not improve the model significantly.

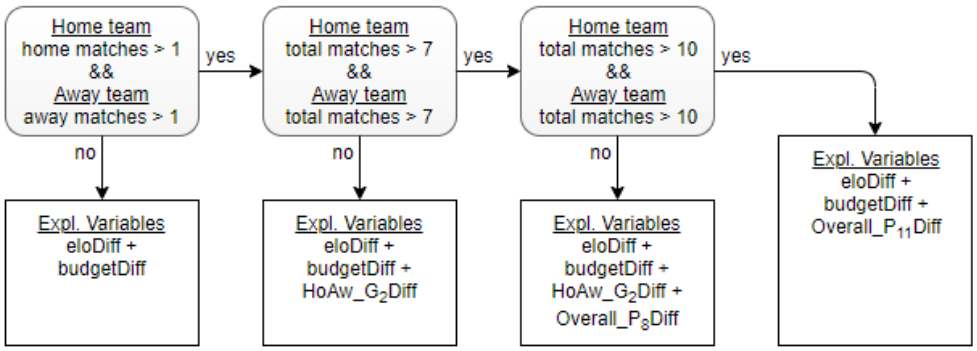

Figure 4.3: Diagram of the extended ordered logistic model.

4.2 Train set size

Looking at the optimal train set size for both basic models (Figure 4.4 on the next page), it is obvious that both error measures (RMSE and Log Likelihood) perform best with a train set of two full seasons (612 matches). However, the extended models also use data gathered during the season. Therefore the train set size per season decreases up to 50% for the Poisson home goals model (the HoAw_A_{11}Diff variable). Therefore the fixed-size train set will contain four full seasons (1224 matches). The error measures of both basic models are still close to optimal with this larger train set.

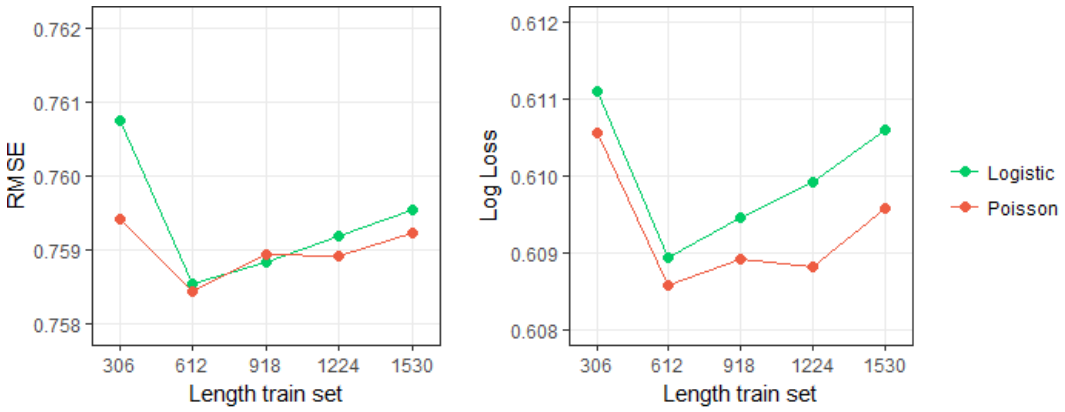

Figure 4.4: Error measures RMSE (left) and Log Likelihood (right) for determining the optimal train set size.

4.3 Model evaluation

All four models are used to predict the probabilities of all match outcomes. The performance per season of the models is evaluation by two metrics: the Log Likelihood statistic and the root mean squared error (RMSE). All results are shown in Tables 4.2 and 4.3 on the next page. Because bookmakers always add a small error margin to their prediction, their probabilities add up to almost 108% on average. Their probabilities had to be normalized in order to fairly compare them to the models created in this paper. Comparing the four models with the bookmaker, it can be seen that all models predictions perform nearly as good as the bookmaker. The extended Poisson model scores the best compared to the other models, although the other three models perform only slightly worse than the extended Poisson model. Based on the errors per season it appears that the bookmaker did not improve its average predictions a lot. The errors fluctuate a lot each season and do not decrease structurally. Moreover, the extended Poisson model performs better than the bookmaker in the two most recent seasons. However, because of the large fluctuations it is premature to conclude anything about the bookmaker efficiency.

	09/10	**10/11**	**11/12**	**12/13**	**13/14**	**14/15**	**15/16**	**16/17**	**avg.**
Basic Poisson	0.711	0.742	0.740	0.770*	0.782	0.760	0.753	0.746	0.750
Extended Poisson	0.713	0.741	0.737	0.771	0.785	0.758	0.750*	0.741*	0.749
Basic OLR	0.709	0.742	0.743	0.771	0.781	0.760	0.753	0.746	0.751
Extended OLR	0.710	0.742	0.739	0.770*	0.782	0.760	0.753	0.745	0.750
Bookmaker	0.707	0.739*	0.735*	0.772	0.778*	0.757*	0.752	0.746	0.748*

Table 4.2: Performance of all models and the bookmaker evaluated by the RMSE. Per season, the smallest error is given an asterisk.

	09/10	**10/11**	**11/12**	**12/13**	**13/14**	**14/15**	**15/16**	**16/17**	**avg.**
Basic Poisson	0.549	0.585	0.590	0.623*	0.632	0.613	0.600	0.589	0.598
Extended Poisson	0.552	0.584	0.587	0.623*	0.636	0.610	0.595*	0.583*	0.596
Basic OLR	0.546	0.585	0.594	0.624	0.632	0.614	0.601	0.588	0.598
Extended OLR	0.547	0.585	0.589	0.624	0.634	0.616	0.600	0.588	0.598
Bookmaker	0.542*	0.583*	0.584*	0.627	0.628*	0.609*	0.600	0.590	0.595*

Table 4.3: Performance of all models and the bookmaker evaluated by the Log Likelihood. Per season, the smallest error is given an asterisk.

4.4 Betting evaluation

In this section the two different betting strategies (fixed bet and Kelly Criterion) will be tested. In this paper 5 different versions (thresholds) per strategy are implemented. Whenever the expected profit $T \geq E_i(\text{prf})$ a bet will be placed. The ROI's of both betting strategies and its different versions are shown in the tables below. Using a bootstrapping method the probability of reaching the same ROI by random betting is estimated. These values are also shown in the tables below. Each row starting with 'Poisson' corresponds to the extended Poisson model, each row starting with 'OLR' corresponds to the extended logistic model. Whenever the average ROI of a certain strategy is significantly better than random betting, an asterisk is given in the table. This occurs whenever the bootstrapping method performed worse than the actual strategy in more than 99% of the simulations.

	T	09/10	10/11	11/12	12/13	13/14	14/15	15/16	16/17	avg.
Poisson	0%	-0.182	-0.002	0.063	0.116	-0.053	0.042	-0.097	0.003	-0.027
OLR	0%	-0.130	0.082	0.039	0.154	-0.019	0.181	-0.151	0.097	0.015*
Poisson	2.5%	-0.278	-0.053	0.225	0.145	-0.041	-0.033	-0.079	0.221	0.007
OLR	2.5%	-0.190	0.350	0.313	0.157	0.086	0.075	-0.055	0.133	0.100*
Poisson	5%	-0.268	0.203	-0.037	0.223	-0.040	-0.096	0.100	0.157	0.021
OLR	5%	-0.107	0.311	0.546	0.352	0.138	0.004	0.189	-0.067	0.172*
Poisson	7.5%	0.132	0.300	0.245	0.611	0.146	-0.020	0.277	-0.047	0.185*
OLR	7.5%	0.504	-0.070	-0.073	0.751	-0.005	-0.109	0.096	-0.170	0.122
Poisson	10%	0.250	-0.271	0.641	1.921	-0.055	-0.408	0.277	0.330	0.238*
OLR	10%	-1.000	0.000	0.600	1.338	0.071	-0.583	0.112	0.890	0.072

Table 4.4: Return on investment (ROI) of both models by using the fixed bet strategy for different values of threshold T.

	T	09/10	10/11	11/12	12/13	13/14	14/15	15/16	16/17	avg.
Poisson	0%	-0.147	0.032	0.155	0.125	-0.074	-0.029	0.083	0.152	0.032*
OLR	0%	-0.090	0.093	0.171	0.143	-0.004	-0.032	0.042	0.097	0.050*
Poisson	2.5%	-0.167	0.035	0.193	0.165	-0.055	-0.058	0.080	0.217	0.046*
OLR	2.5%	-0.080	0.100	0.298	0.144	0.020	-0.055	0.073	0.098	0.076*
Poisson	5%	-0.184	0.155	0.102	0.220	-0.035	-0.089	0.181	0.159	0.063*
OLR	5%	-0.049	0.096	0.449	0.253	0.082	-0.091	0.195	0.013	0.126*
Poisson	7.5%	0.105	0.174	0.322	0.583	0.104	-0.112	0.329	0.093	0.186*
OLR	7.5%	0.341	-0.087	0.224	0.506	0.067	-0.164	0.119	0.049	0.114
Poisson	10%	0.248	-0.329	0.662	1.854	-0.025	-0.419	0.336	0.334	0.212
OLR	10%	-1.000	0.000	0.707	0.866	0.094	-0.498	0.134	0.767	0.063

Table 4.5: Return on investment (ROI) of both models by using the Kelly Criterion for different values of threshold T.

It can be seen from Table 4.4 and 4.5 above that almost all strategies have a positive average ROI over all years. Many of them perform significantly better than random as well. The best strategies are profitable in almost all seasons. It appears that the ROI increases with a more conservative (increasing) threshold T. However, since less matches meet the requirements of an increasing T the variance between the ROI per season significantly increases as well. This causes the average profit for the Kelly Criterion with $T \geq 10\%$ to be not significantly better than random. When picking the best strategy for both the extended Poisson model and the extended logistic model it is therefore beneficial to take the number of bets (and the variance) into ac-

count. Two of the most promising strategies apply the Kelly Criterion: the Poisson model with $T \geq 7.5\%$ and the logistic model with $T \geq 5\%$. A visualization of the performance of both models is shown in Figure 4.5 below. The average stake size for both models is a little over €6,-. The number of bets placed per season is not constant, which is shown by the width of each interval (season) in the plots of Figure 4.5. A larger range between the vertical dotted lines indicates more matches have been placed in that particular season. From both the left and the right plot it can be seen that the profits seem to stagnate a little since the season $2014-2015$. This could indicate an increased efficiency of the bookmaker. However, both models (in particular the logistic model on the right) are still profitable.

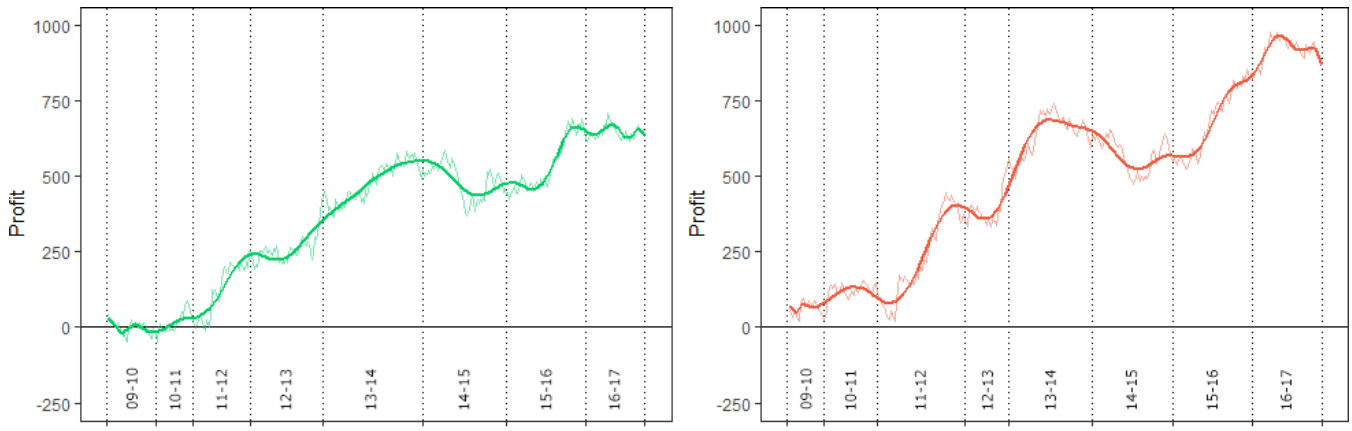

Figure 4.5: Growth of the betting profit over all seasons for the best Poisson model (left) and the best logistic model (right). Each season is indicated above the x-axis. All seasons are divided by a vertical dotted line.

5. Conclusions

This research aims to answer the question whether it is still possible to beat the bookmaker without using extensive match data. Bookmakers have a constant need for improvement and have access to detailed match data that is not freely available to the public. Therefore it is expected that it is becoming more and more difficult to create a profitable betting system. The data used for this research contains basic match information (i.e. number of goals scored, full time result) over the last 13 seasons in the Dutch Eredivisie.

There are two main movements in predicting football matches: goal models and toto models. In this paper two goal models (Poisson distribution) and two toto models (ordered logistic regression) have been implemented. Two of the four models solely depend on the long-term strength indicator of both opponents, while the other two models also take into account more variables (i.e. recent form). All models have been evaluated by two error measures: the Log Likelihood measure and the Root Mean Squared Error (RMSE). Next, the best goal model and the best toto model are chosen to compete against the bookmaker over a range of 8 seasons. The betting performance of both models is evaluated for multiple betting strategies: the fixed bet strategy and the Kelly Criterion. Multiple threshold levels have been evaluated for both betting strategies.

The results show that both the goal models and the toto models perform nearly as good as the bookmaker. It should be noted that none of the models were able outperform the bookmaker on both error measures. It is remarkable that the bookmaker does not seem to improve on their predictions over the last 8 seasons. However, because of the large fluctuations in error measures per season it might be premature to conclude anything about the bookmaker efficiency. The actual betting performances of the best models show that it is still possible to compete against the bookmaker. Both evalu-

ated betting strategies show a positive average return on investment (ROI) over all seasons for almost all threshold levels. The best models even have an average ROI of over 15%, with a negative ROI in only 2 seasons.

Although the results look promising, it should be noted that all results are obtained by simulating historical matches. Therefore there is always the chance that a certain level of overfitting occurs. Even though this paper intended to keep the models relatively simple, the obtained results do not guarantee a similar performance in the future. Not in the least because the bookmaker is always seeking for improvement of their models. It is reasonable to expect that in the future more sophisticated data is needed to compete with the bookie. Further improvement on this paper would therefore be to implement more detailed data in the models. Another improvement would be to add information about the players. When an important player is injured or suspended, a team is expected to perform worse. It could also be beneficial to keep track of the number of days rest between the matches of a team. According to exercise physiologist Raymond Verheijen[16], a team's performance decreases significantly when they have had only 2 or 3 days of rest. This might improve the model on matches containing teams that have played in European club competitions during the week and have to play in the national competition only a couple of days later.

We can conclude that it is still possible to compete with the bookmaker in the Dutch Eredivisie. This paper can form as a basis to create a profitable betting system. With the above mentioned improvements the results might even get better, which is expected to be necessary since the bookmakers will also aim to improve their models continuously.

Bibliography

[1] Remie, M. (2014): *De halve finale van gisteren is het best bekeken tv-programma ooit*. From: https://www.nrc.nl/nieuws/2014/07/10/halve-finale-best-bekeken-tv-programma-ooit-a1423807.

[2] Sheen, T. (2015): *Arsène Wenger: 'Of course I get nervous... football is not mathematics,' says Arsenal boss*. From: https://www.independent.co.uk/sport/football/premier-league/arsene-wenger-of-course-i-get-nervous-football-is-not-mathematics-says-arsenal-boss-10149481.html.

[3] *Soccer leagues ordered by number of goals*. From: https://www.soccervista.com/soccer_leagues_ordered_by_number_of_goals.php.

[4] Keogh, F., Rose, G. (2013): *Football betting – the global gambling industry worth billions*. From: https://www.bbc.com/sport/football/24354124.

[5] Maher, M.J. (1982): *Modelling association football scores*. In: Statistica Neerlandica, 36: 109-118.

[6] Dixon, M.J., Coles, S.G. (1997): *Modelling association football scores and inefficiencies in the football betting market*. In: Applied Statistics, 46: 265-280.

[7] Koning, R.H. (2000): *Balance in competition in Dutch soccer*. In: Journal of the Royal Statistical Society, Series D: The Statistician, 49: 419-431.

[8] Goddard, J., Asimakopoulos, I. (2003): *Modelling football match results and the efficiency of fixed-odds betting*. Working Paper, Department of Economics, Swansea University.

[9] Langseth, H. (2014): *Beating the bookie: a look at statistical models for prediction of football matches*. Paper presented at the SCAI.

[10] Kelly, J.L. (1956): *A new interpretation of information rate*. In: IRE Transactions on Information Theory, 2: 185-189.

[11] *Elo rating system*. From: https://en.wikipedia.org/wiki/Elo_rating_system.

[12] Schiefer, L.: *European Football Club Elo Ratings*. From: https://www.clubelo.com.

[13] Van Ours, J.C. (2017): *Artificial pitches and unfair home advantage in professional football*. In: Centre for Economic Policy Research Discussion Paper 12341.

[14] Hamadani, B. (2005): *Predicting the outcome of NFL games using machine learning*. From: http://cs229.stanford.edu/proj2006/BabakHamadani-PredictingNFLGames.pdf.

[15] Kaggle: *Deloitte/FIDE Chess Rating Challenge*. From: https://www.kaggle.com/c/ChessRatings2.

[16] *Door Europese verplichtingen krijg je competitievervalsing*. From: https://www.voetbalzone.nl/doc.asp?uid=172830.