Predicting the outcome of soccer matches in order to make money with betting

Master Thesis

Business Analytics and Quantitative Marketing

Author:
Tim van der Zaan (362006)

Supervisor:
Dr. Andreas Alfons

Econometrics & Operations Research
Erasmus School of Economics
Erasmus University Rotterdam

February 3, 2017

Abstract

For several decades soccer is reckoned among the most beloved sports on the globe. This manifests itself by continuous attention from millions of people from all over the world. The money volume involved in soccer is exceptional and expands consistently. Simultaneously, bookmakers' profit acquired from soccer betting has increased over the recent past. The urge on the construction of statistical betting strategies that systematically generate profit can potentially be accomplished since nowadays more soccer-related data is publicly accessible. This paper presents various betting strategies that exploit the predicted likelihood of the outcome possibilities assigned by match outcome prediction models. In this paper is shown that the extension of an existing betting strategy generates an average out-of-sample profit share of 0.26 per bet euro.

Contents

1 Introduction

Along with the start of the soccer season 2015-16, the bookmaker William Hill set a fixed-odd equal to 5,000 for Leicester City to win the 'Premier League' (England's highest division of soccer). Joe Crilly, Press-Officer of William Hill, reacted on Leicester City leading the 'Premier League' with only twelve matches left to play; "If they hold on to claim the title, it would be the biggest upset in sports history.", quoted in [27].

When Leicester City unexpectedly became the champions of the 2015-16 'Premier League' season it demonstrated once more that the real-life outcome of sport events can be surprising. From the occurrence of this unforeseen denouement one can deduce that the outcome of soccer matches is not always easy to predict for both bettors and bookmakers. According to Dobson and Goddard in [7] the popularity of soccer, which is currently the most watched and participated sport worldwide, was rising across the globe in the last years of the 20^{th} century. Historically the sport has been popular in especially Europe and South-America, however [21] emphasizes the rising popularity of soccer in the USA over the last decades. Furthermore, the recent expenses of Chinese soccer clubs indicate an (expected) growth in Asian club soccer. With the growth in popularity, the growth in online gambling on soccer matches has grown even more rapidly over the last years. The possibility to bet online and on mobile phones contributed largely in this enormous growth according to Hing [18]. Finnigan and Nordsted even state in [10] that betting on soccer provide online bookmakers the largest turnover compared to other sports. Preston and Smith [36] assert that bettors, which are present among all welfare and racial classes in modern society [11], are fond of making decisions depending on their own believes. It is also stated that bettors might think that they personally posses the knowledge and cleverness to select winning bets. Besides, various bettors believe their fate is well blessed. In [36] is asserted that bettors even find pleasure in betting whilst losing money. Sauer endorses this in [35] by asserting that gamblers mainly bet for pleasure and not in the first place to earn money.

"I get nervous before the games because soccer is not mathematics, it's completely unpredictable. That's what makes it so popular." - Arsène Wenger, manager of 'Premier League' club Arsenal, quoted in [29]. Just like Arsène Wengere indicates here, prediction regarding the outcome of soccer matches (and competitions) is associated with lots of uncertainty. The outcome of soccer matches is partly dependent on measurable factors, nevertheless a large share cannot be explained by measurable factors. In papers of Pollard and Pollard as well as Sosik and Vergin, in [32] and [40], respectively, it is proposed that the home team experiences an advantage over the away team. In [5], Clarke and Normand demonstrate that the distance between the stadium of the home and away team is an influential factor. Asimakopoulos and Goddard [12] incorporated these, and other, soccer-related variables in an ordered probit prediction model in an

attempt to accurately forecast the outcome of soccer matches. The predictions from this model are able to compete with the bookmakers' forecasts. In [37], prediction results from various forecasting models are obtained and these assigned match outcome probabilities are used as input for several betting strategies. The obtained profit varies over the different betting strategies. Several of these strategies generate positive profit.

Although Arsène Wenger asserted that soccer is not mathematics, statistics will be used in this paper to come up with models that predict the outcome of soccer matches. The focus is on soccer in Europe, in particular the matches within the highest national leagues of European countries. Cluster analysis is used to classify the competitions into groups. The resulting clusters assist by the selection of the competitions that will be checked more thoroughly. Because Hamadani found in [16] that for a certain competition the set of explanatory variables varies each season, models are build to forecast the outcomes of soccer matches in several seasons and competitions. More precisely, four different prediction models are constructed for three recent seasons of four different competitions. These forecast models are evaluated by applying several measures on out-of-sample outcome predictions, and compared to the outcome probabilities assigned by bookmakers. Subsequently, several betting strategies will be considered in an attempt to systematically make profit. The probabilities for the outcome possibilities estimated by the prediction models are input for these betting strategies. The link between the obtained profit and the season predictability is studied over the considered seasons and competitions. Three different measures are applied to indicate the season predictability.

The remainder of this paper is structured as follows. In Ch. 2 existing related literature is discussed. The data used during the research is described in Ch. 3. In Ch. 4 the methods used during the research are described into detail. The results of the performed methods are denoted and evaluated in Ch. 5, whereas in Ch. 6 a conclusion is drawn and some suggestions for further research are provided.

2 | Related work

The goal of this research is to construct models that can accurately predict the outcome of soccer matches, based on historic match results and other statistics. At first, cluster analysis will be applied to split the competitions into groups. Existing literature on clustering methods will be discussed in Sect. 2.1. Subsequently, Sect. 2.2 presents an overview of existing literature regarding the prediction of soccer match outcomes. Numerous papers exist that propose varying models in order to predict soccer match outcomes. Sect. 2.3 elaborates on literature concerning strategies specialized in betting on soccer matches.

2.1 | Clustering algorithms

Clustering methods provide the possibility to split a set of data points into several groups based on the similarity of the attributes. Within a cluster the data points contain similar characteristics whereas data points assigned to different clusters contain smaller similarity of attributes. The K-means clustering algorithm, also called Lloyd's method [25], is a widely used clustering method. A requested number of clusters is the only necessary input to perform this method. The amount of clusters, K, is acquired applying the pseudo-F measure [41]. In [33] it is suggested to reduce the dimensionality using principal component analysis (PCA) [30] before clustering is applied in case of either a high-dimensional dataset or large correlation within the dataset. In [24] it is showed that the approach of using PCA before K-means clustering provides good cluster output on the original data.

In this paper a clustering method will be used to determine a selection of European competitions that will be investigated. Because of correlation among the data attributes prior to the K-means cluster analysis PCA will be performed.

2.2 | Match outcome models

All the proposed models in existing literature concerning the prediction of soccer match outcomes can mainly be divided into two kinds of models; the first kind of models forecasts the amount of goals scored per team and so, indirectly, the match outcome is predicted. These models are called *goal-models*. Maher [26] proposes such a *goal-model* that forecasts the amount of goals scored by both the home and away team using poisson distributions. This prediction is based on attacking and defending parameters for both the home and away team, as well as a parameter representing the home town advantage. Plenty of literature is published in which several models are proposed to indirectly forecast soccer match outcomes based on this model of Maher.

More recently several papers are published in which the outcome of a soccer match is predicted directly. These *toto-models* predict whether a soccer match ends in a victory for the home team, away team, or that a draw occurs. Koning [22] proposed to use an ordered probit regression model in order to investigate the competition balance of Dutch soccer. Based on this model of Koning, Graham and Scott [14] proposed a comparable model. They compared the resulting probabilities of their dynamic probit regression model on English leagues, from August 2004 to November 2006, with the outcome probabilities obtained from bookmaker William Hill. Although the predicted probabilities of William Hill did outperform the constructed model, it was not a significant difference. In addition to current season match results all sort of other explanatory variables are included in the prediction models proposed in [12]. The English match outcomes of seasons 1998-99 through 2000-01 are predicted in Asimakopoulos and Goddard [12] applying an ordered probit model as well. The resulting probabilities can compete with the predictions of the bookmakers. A comparison of the betting profit obtained from the probabilities assigned by several prediction models, containing both *toto-models* and *goal-models*, is applied in [37]. Snyder shows in this paper that the betting profit of various strategies is significantly higher for *toto-models* than for *goal-models*.

During this research the accessible data combined with the results found in [37] made us conclude to concentrate on the direct modelling of soccer match outcomes. Four different *toto-models* will be constructed incorporating current season match results, previous season match results and various other explanatory variables.

2.3 | Soccer betting

The amount of papers that elaborate upon strategies to bet on soccer matches is limited. A 'naive' betting strategy is proposed in [38]. The strategy systematically bets on the outcome that historically occurred most frequently, a home win. For the soccer seasons 1999–00 through 2001-02 of the German highest national league the obtained betting profit was not significantly different than the generated profit obtained from several more sophisticated betting strategies. In [14], Graham and Scott studied betting odds of bookmaker William Hill for the English 'Premier League' seasons 2001-02 and 2002-03. They found that the assigned odds of William Hill for home wins and draws are better than for away wins. Betting on a home win, draw or away win for all matches generated a profit share per bet euro of -0.111, -0.103 and -0.160, respectively.

Langseth suggests in [23] to solely bet on betting options with an expected positive profit. Equal betting on these matches results in positive profit for the English 'Premier League' season 2011-12. Depending the weights on the assigned probability, as that the expected profit is equal for all betting options, results in a similar profit profile. Another betting strategy in which the expected value of the betting budget, expressed in

logarithm, is maximized, called Kelly's betting strategy [20], is performed as well. In [34] a betting strategy is introduced where the betting weights depend on the expected profit and the variance of this profit. These two strategies gave approximately similar profit as the aforementioned strategies for the seasons 2011-12 and 2012-13 in the English 'Premier League'. Not one of these four betting strategies outperforms the others systematically based on the models build by Langseth in [23]. Snyder [37] proposes twelve betting strategies, of which multiple are build upon the betting strategy of Kelly. Ten of these twelve betting strategies generate a positive profit share per bet euro for multiple seasons and competitions.

In this paper extensions on several existing betting strategies are applied on the results of the constructed models. In addition, the 'naive' prediction model is used as a baseline indication. The profit share per bet euro obtained by the betting strategies will be studied for three recent seasons of four different European competitions. Furthermore, a potential relation between the obtained profit share and the season predictability for the considered seasons is investigated

3 | Data

In this paper models are provided that predict the outcome of soccer matches in several European national leagues. To be able to build such models it is essential to have data regarding matches played in these leagues over the past years. From the sport database of Infostrada data is obtained for the highest national leagues of 51 European countries. This dataset covers club and score information regarding matches of the seasons 1986-87 through 2015-16. Besides the data covering the matches in the national leagues, data regarding European matches and national cup matches is included in this dataset. Unfortunately, the data for several competitions contains a lot of missing values for certain attributes, e.g. the number of spectators per match. Nevertheless, for sixteen national leagues the dataset is complete. Two of these competitions are 'summer' competitions, the remaining fourteen competitions are 'winter' competition. Summer competitions start in the spring and end in the autumn as that no matches are played in the winter period. On the other hand, winter competitions start in the autumn and end before the summer. The data from Infostrada has been the input for a tool of Hypercube Business Innovation B.V. This tool provides an up-to-date strength indication for all clubs in the highest division of the 51 European countries, called the European Club Index (ECI)[1]. This model is built based on the popular chess rating system called the ELO-rating, introduced in 1978 by Arpad Elo in [8]. The ECI value of each club depends on the results of historic matches and is updated after either a national or international round of matches for club teams. Clubs' ECI strength indication estimated by this tool is available from the start of the winter competition season 2007-08 through 2015-16. Besides, a dataset containing betting odds for matches of European competitions is obtained from *www.oddsportal.com.* These match odds represent for all three outcome possibilities the average of odds assigned by multiple global bookmakers, such as 'Bet365', 'Unibet' and 'Bwin'. The online retrieved match odds cover the matches played in the highest divisions of European competitions since the start of the winter season 2007-08.

[1] http://www.euroclubindex.com/asp/Ranking.asp

4 | Method

In order to answer the main research questions several econometric methods will be applied in this paper. In this chapter, these methods are introduced and discussed. In Sect. 4.1, cluster analysis is introduced for the selection of a number of European national competitions. Sect. 4.2 introduces four different ordered probit models to predict the outcome of soccer matches. Multiple evaluation measures to compare the results of the models are discussed in Sect. 4.3, whereas Sect. 4.4 introduces several betting strategies. Sect. 4.5 elaborates on measures that indicate the predictability of a certain competition.

4.1 | Clustering of the competitions

In Ch. 3 is stated that the dataset used during this research contains soccer-related information about the national league of sixteen European countries. These leagues differ in composition, strength and other characteristics. Based on season 2015-16, fifteen different league-specific variables are considered to reflect the characteristics of these competitions. The competition characteristics vary from the average stadium capacity to the mean age of the players that have played during the season. A cluster analysis will be applied to split these competitions into several groups. The competitions within a cluster contain comparable characteristics whereas competitions in different clusters vary relatively more in characteristics.

Because a number of league characteristics seem to be correlated the results of 'normal' clustering algorithms would be affected and the reliability of corresponding results can be questioned. To overcome this problem, [1] suggests to perform PCA before the cluster analysis is applied. A normalized version of the initial dataset is used. This mathematical procedure reduces the dimensionality by finding a number of new uncorrelated underlying components that are a composition of the original (possibly) correlated variables. This is achieved by maximizing the variance explained by these new components. Components with the largest eigenvalues contain the most useful information about the initial data. Decision on the number of components is an important step in PCA. A popular criterion, suggested in [42], is to use the components with eigenvalues larger than one. This corresponds to using components explaining at least one Q^{th} of the total explained variance, where Q gives the number of variables.

Subsequently, a K-means clustering algorithm will be applied on the set of orthogonal components. The number of clusters, K, is the input for this clustering algorithm. The pseudo-F measure [41] is seen as the rule of thumb to decide on the number of clusters. Using the K-means clustering method the starting points, cluster means, are randomly assigned at first. Thereafter, the method will edit these cluster means until the optimal ones are obtained. Each of the data points is assigned to the closest cluster mean.

4.2 | Proposed models

In this section four models will be introduced to predict match outcomes. Each of these proposed models can be categorized as an ordered probit model [15]. An introduction of the ordered probit models will be given in Subsect. 4.2.1. The first constructed model, based on solely the ECI values of the competing clubs, is described in this subsection as well. Subsect. 4.2.2 elaborates upon an ordered probit model where the parameters are estimated based on data of the past three seasons. The third and fourth model, incorporating several explanatory attributes, are introduced in Subsect. 4.2.3.

4.2.1 Ordered Probit prediction model

The prediction of soccer match outcomes in this paper is performed using ordered probit models. These models generate probability estimations for a home win, draw and away win for each match. The statistics behind the ordered probit model will be given as the first model will be introduced. This first model is referred to as the basic ECI model in the remainder of the paper. The ordered probit model incorporates an unobserved latent variables. The latent variable is dependent on explanatory variables concerning the competing clubs of match i, given by x_i. This gives the following unobserved latent variable $y_i^* = \beta x_i + \epsilon_i$, where $\epsilon_i \sim N(0, 1)$. As ϵ_i is standard normally distributed and y_i^* is linearly related to ϵ_i, y_i^* is also normally distributed under the assumption of fixed predictors.

Now, imagine a match between two clubs with the exact same characteristics and similar historic results, playing on neutral territory. Although the data suggests equal chances for both teams to win the match several unexplained factors, such as luck, cause that this match will not always end in a draw in reality. For each possible match this ordered probit model will assign probabilities to all three outcome possibilities. Fig. 4.1 visualizes the normally distributed y_i^* and points out that the value of y_i^* might be different per match depending on club and match characteristics. This way, the values of the latent variable y_i^* varies per match whereas the values of μ_1 and μ_2 are fixed for a season. This causes that outcome predictions differ per match.

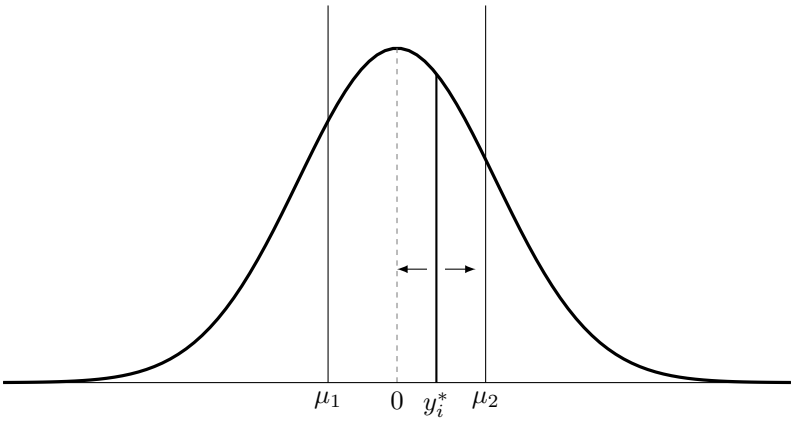


Figure 4.1: This figure visualizes the relation between the latent variable y_i^*, which follows a normal distribution, and the treshold values μ_1 and μ_2.

Irrespective of the value of the latent variable, y_i^*, the treshold values, μ_1 and μ_2, separate the density of the normal distribution into three parts. The values of μ_1 and μ_2 indirectly incorporate the division of chances assigned to a home win, draw and an away win for a specific competition season, and so home advantage is incorporated in this model. The specific estimation of these parameters is based on historic data and will be discussed later in this subsection. Although the treshold values, μ_1 and μ_2, are fixed within a season, the chances for each possible outcome, home win, draw and away win, vary per match. This is caused by different match and club characteristics, x_i, resulting in different values of y_i^*. To come back at the previous mentioned example of a match between two clubs with equal characteristics. The corresponding latent variable, y_i^*, follows a normal distribution with equal chances assigned to the occurrence of a home win and an away win. For the basic ECI model, the only explanatory variable affecting the latent variable is the difference in ECI value between the home and away club, denoted by $ECIHome_i$ and $ECIAway_i$, respectively. This results in the following latent variable y_i^*;

$$y_i^* = \beta \left[ECIHome_i - ECIAway_i\right] + \epsilon_i. \tag{4.1}$$

Fig. 4.1 visualizes the relation between the treshold values, μ_1 and μ_2, and the unobserved latent variable y_i^*. The following equations illustrate how the latent variable and the treshold values are linked to the ordinal response of the match outcome $\{-1, 0, 1\}$;

$$y_i = 1 \quad \text{(home win)} \qquad \text{if} \quad y_i^* + \epsilon_i > \mu 2, \tag{4.2}$$

$$y_i = 0 \quad \text{(draw)} \qquad \text{if} \quad y_i^* + \epsilon_i < \mu 2 \quad \text{and} \quad y_i^* + \epsilon_i > \mu 1, \tag{4.3}$$

$$y_i = -1 \quad \text{(away win)} \qquad \text{if} \quad y_i^* + \epsilon_i < \mu 1. \tag{4.4}$$

Ones the parameters, μ_1, μ_2 and β, are estimated the match-specific value of the unobserved latent variable y_i^* can be used to calculate the probability of occurrence for the three match outcome possibilities. More specifically, the probability of a home win can be calculated using Eq. 4.5, whereas Eq. 4.6 shows the calculation to obtain the probability of an away win. The probability of a draw can be retrieved from Eq. 4.7, which can also be derived ones the probability of a home win and an away win are calculated.

$$P[y_i = 1] = P(\epsilon_i > \mu_2 - y_i^*) = 1 - \Phi_i(\mu_2 - y_i^*), \tag{4.5}$$

$$P[y_i = -1] = P(\epsilon_i < \mu_1 - y_i^*) = 1 - \Phi_i(\mu_1 - y_i^*), \tag{4.6}$$

$$\begin{aligned} P[y_i = 0] &= P(\mu_1 - y_i^* < \epsilon_i < \mu_2 - y_i^*) \\ &= \Phi_i(\mu_2 - y_i^*) - \Phi_i(\mu_1 - y_i^*) \\ &= 1 - P[y_i = -1] - P[y_i = 1], \end{aligned} \tag{4.7}$$

where Φ_i represents the cumulative distribution function (CDF) of a standard normal distribution for observation i.

Estimation of the parameters, μ_1, μ_2 and the variable coefficient(s), β, of the latent variable, is performed using the method of maximum likelihood estimation(MLE). Define M dummy variables for each possible outcome ({-1,0,1}), as that dummy $result_{i,m} = 1$ in case match i ends in outcome $m \in M$ and $result_{i,m} = 0$ if not. Per match the product over all M outcome probabilities raised by $result_{i,m}$ provides the individual likelihood. The product over the individual likelihood of all matches gives the likelihood of the sample, resulting in

$$\begin{aligned} \mathcal{L} &= \prod_{i=n}^{N} \prod_{M=-1}^{1} P[y_i = m]^{result_{i,m}} \\ &= \prod_{i=n}^{N} P[y_i = -1]^{result_{i,-1}} \prod_{i=n}^{N} P[y_i = 0]^{result_{i,0}} \prod_{i=n}^{N} P[y_i = 1]^{result_{i,1}}. \end{aligned} \tag{4.8}$$

In case Eq. 4.5, Eq. 4.6 and Eq. 4.7, defining the calculations of the outcome probailities, are incorporated into the sample likelihood it can be rewritten to

$$\begin{aligned}\mathcal{L} \;=\; & \prod_{i=n}^{N}[1-\Phi_i(\mu_1-y_i^*)]^{result_{i,-1}} \\ & \prod_{i=n}^{N}[\Phi_i(\mu_2-y_i^*)-\Phi_i(\mu_1-y_i^*)]^{result_{i,0}} \\ & \prod_{i=n}^{N}[1-\Phi_i(\mu_2-y_i^*)]^{result_{i,1}}. \quad (4.9)\end{aligned}$$

In case MLE will be performed it is convenient to use the log-likelihood function to estimate the parameters. Transformation of the likelihood function provides the following log-likelihood function;

$$\begin{aligned}l = log\ \mathcal{L} \;=\; & \sum_{i=n}^{N} result_{i,-1}\ log\ [1-\Phi_i(\mu_1-y_i^*)] \\ & + \sum_{i=n}^{N} result_{i,0}\ log\ [\Phi_i(\mu_2-y_i^*)-\Phi_i(\mu_1-y_i^*)] \\ & + \sum_{i=n}^{N} result_{i,1}\ log\ [1-\Phi_i(\mu_2-y_i^*)]. \quad (4.10)\end{aligned}$$

The optimal parameters are obtained by maximizing the log-likelihood with respect to the parameters, μ_1, μ_2 and the variable coefficient(s), β, of the latent variable using the Quadratic Hill Climbing method [13].

This ordered probit model, the basic ECI model, is constructed to predict the outcome of three consecutive seasons, 2013-14, 2014-15 and 2015-16 for winter competitions (and 2013, 2014 and 2015 for summer competitions) for the selected European competitions. In the remainder of the paper 2013(-14) refers to season 2013 for the summer competitions and season 2013-14 for the considered winter competitions. The parameters for the prediction model of a certain season are estimated based on historic data about matches from 2007-08 through the season prior to the season under consideration. So, imagine a model for the season 2013-14 of the English competition, the parameter values are estimated based on matches from season 2007-08 through 2012-13. Using this procedure the obtained match outcome predictions are out-of-sample forecasts.

4.2.2 Simplified prediction model

For the second model, an ordered probit regression model is constructed where the difference in ECI values of the competing clubs remains the main input. The treshold values of the basic ECI model, described in Subsect. 4.2.1, are estimated based on matches since the start of the season 2007-08. The model introduced in this subsection incorporates information of the last three seasons to predict the outcome of soccer matches. The focus with this model is more on the recent past compared to the basic ECI model. The main difference compared to the basic ECI model is in the estimation of the parameters. Less parameters have to be estimated. In the remainder of the paper this second model is therefore called the simplified model. The only parameters that remain for estimation are the treshold values, μ_1 and μ_2, as the coefficient for $ECIHome_i$ - $ECIAway_i$ is fixed at a value of $\beta = \frac{1}{1000}$. This specific fixed value is chosen as that the parameter values of μ_1 and μ_2 are of a similar magnitude as for the basic ECI model. The calculation of the latent variable y_i^* for the simplified model is different and given by

$$y_i^* = \frac{[ECIHome_i - ECIAway_i]}{1000} + \epsilon_i, \tag{4.11}$$

where $\epsilon_i \sim N(0,1)$ and so y_i^* remains normally distributed.

The treshold values are estimated in a different way than for the basic ECI model, where MLE provides the treshold estimations based on the sample of matches since the start of the season 2007-08. Instead, this model makes use of the balance of home wins, draws and away wins of the last three seasons to estimate the treshold values. For each of these three seasons, the treshold values are set in such a way that the amount of matches ending in a home win, draw and away win equals the amounts in real-life. The three obtained values of $\mu_1(\mu_2)$ estimated for the three preliminary seasons are averaged resulting in the estimation of $\mu_1(\mu_2)$ for the considered season. So, the simplified prediction model incorporates the outcome balance of the last three seasons instead of since the start of season 2007-08. The outcome balance over the recent seasons seems more useful for the match outcome estimation of the successor season.

Similar as for the basic ECI model, described in Sect. 4.2.1, this model is utilized for the match outcome prediction of matches from season 2013(-14), 2014(-15) and 2015(-16). As the parameter values are obtained from the three preliminary seasons, the obtained outcome predictions remain out-of-sample forecasts.

4.2.3 Extended ordered probit models

In this subsection a third and fourth model are introduced that both build upon the ordered probit regression model described in Subsect. 4.2.1. The only difference is in the calculation of the latent variable y_i^*, which is more complicated for these extended models than for the basic ECI model. For the basic ECI model the latent variable solely depends on $ECIHome_i - ECIAway_i$, whereas for the extended models many more other variables are considered and included in the estimation of the latent variable. The obtained models vary per season and competition caused by both the difference in coefficients for the variables incorporated in the latent variable as the difference in the season-specific treshold parameters.

For these extended ordered probit models many match and club characteristics will be considered alongside the difference in ECI value of the competing clubs. These characteristics contain information about several factors that might influence the outcome of soccer matches. The fields of information will be briefly covered in the remainder of this subsection.

- In 2014, Yezus [43] suggested to include a dummy for a match featured as a regional derby in models that predict soccer match outcomes. Instead of only regional derbies, dummies for all national club combinations are considered in the models constructed in this paper. The only restriction is that the competing clubs must have played against each other in the national competition at least six times since the start of season 2007-08. This minimum of six matches, or three seasons, is introduced to prevent a lack of observation.

- Blundell [2] uses recent match results as shape indicator in his model, whereas Dixon and Coles [6] specified recent club form based on the last couple of matches played. In both papers, where models are proposed to predict soccer match outcomes, is attempted to express the shape of clubs purely based on recent results. In this paper several variables are constructed that can be categorized as shape-related variables. The average number of goals scored and received in (home or away) matches, the amount of points, the current rank and the percentage of possible points obtained for each club enclose a small selection of considered shape indicators. Besides these 'direct' shape indicators, more variables are constructed that can be seen as shape indicators. Since the season of 2007-08 the changes in ECI value of clubs after one, three and five consecutive matches are gathered. From these ECI increment values five equally sized groups are formed using percentiles, for all three categories. These groups are formed for each competition separately. Subsequently, these shape indicators after one, three and five matches are assigned to the clubs of all matches the model will predict based on their shape at that specific moment.

- Goddard and Asimakopoulos suggest in [12] to incorporate the effect of national cup matches. They found varying effects for the successive match, both positive and negative. The influence of national cup matches as well as European matches will be considered in the models during this study. Dummies will be constructed to indicate the national cup matches and European matches before national league matches.

- As Vergin and Sosik proposed in [40], the home team in sport obtains an advantage over the away team. Besides the predicted season-specific outcome balance covered by the values of μ_1 and μ_2, the home advantage of each specific club is considered. Dummies are constructed for each club in the national competition based on data since the season 2007-08. In [2] it is suggested to incorporate the stadium capacity of both the home and away team into the model. Furthermore, the number of spectators and the distance between stadiums are factors that are considered in the proposed extended ordered probit models.

- In [28], Balmer and Williams found that referees respond different to crowd noise and stadium size. All national combinations of club and referee since the season 2007-08 are considered to investigate club preferences of referees. The only restriction is that the two parties must have crossed at least six times in the national competition since the season 2007-08. This minimum of six matches is introduced to prevent a lack of observation. In consultation with Hypercube Business Innovation B.V. is decided to keep the specific referees anonymous. The influence of club preferences from referees contributes to the model predictions but the specific referee is not mentioned by name.

- Statistics regarding the performance of Manchester City for each month in the English 'Premier League' for season 2007-08 through 2015-16 are shown in Fig. 4.2. Although these statistics do not correct for the strength of the opponents, it looks like the performance of Manchester City is better in Augustus than their performance in for example February or March over the past nine seasons. Although the patterns of performance per month are not similar for all seasons, the performance of Manchester City seems to differ over the months. Similarly, this could be the case for the time a match is scheduled and the day of the week. In this paper dummies are constructed for the day of the week and the month of the year. For the time of the match, competition-specific dummies are constructed. These dummies represent matches played in a certain time span based on equally sized groups of historic matches since the season 2007-08, for example between 4PM and 8PM. The combination of a club and a time span, day of the week or month of the year is only considered in case this combination has occurred at least five times in the national league since the start of season 2007-08.

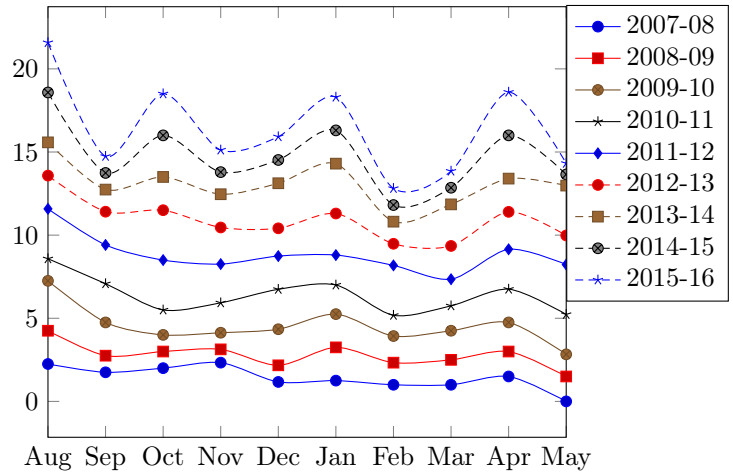


Figure 4.2: This figure shows the stacked plots of the average amount of gathered points per match by Manchester City per month for the soccer seasons 2007-08 through 2015-16 in the English highest national league. Each of the plots represents a certain season.

- Available data regarding artificial turf is limited to the Dutch competition. In the season 2015-16 six Dutch clubs played their home matches on artificial turf, of whom Heracles Almelo and PEC Zwolle played on artificial turf for several years. Both of these clubs ended somewhat surprising in the subtop of the league table in the season 2015-16 of the Dutch national league. Fig. 4.3 presents the average amount of goals scored in home and away matches by the clubs ended at the positions four to nine in the season 2015-16. The clubs with the biggest difference in the average amount of goals scored in home and away matches are Heracles Almelo and PEC Zwolle. This statistic could be coincidence, however the influence of artificial turf is investigated and considered in the models for the Dutch competition. The influence of matches on artificial turf is considered for each of the clubs in competition based on historic matches of different time spans.

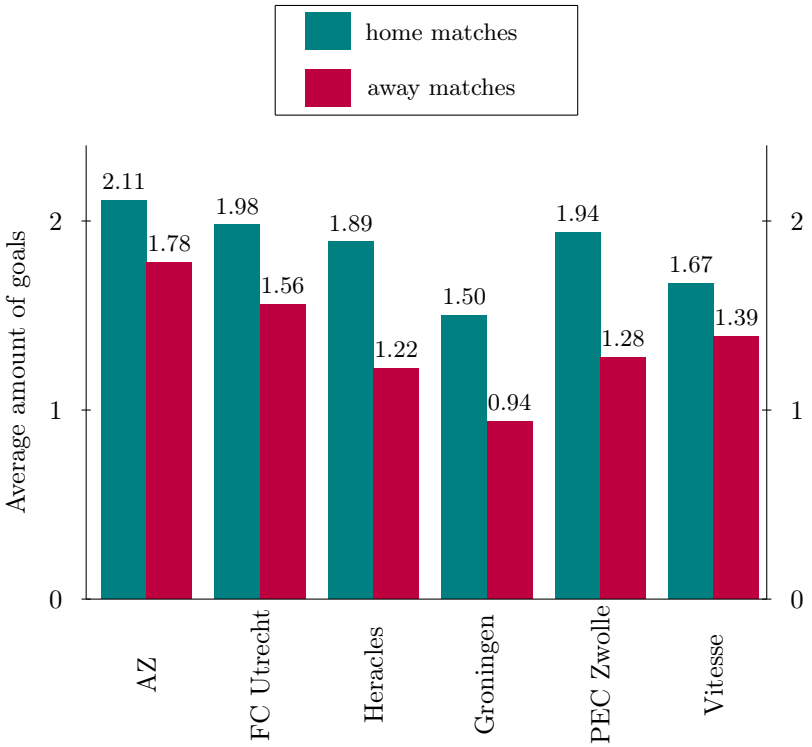


Figure 4.3: This figure shows the average amount of goals scored in home and away matches for the subtop (positions four to nine) of the Dutch highest division in the season 2015-16 (including PEC Zwolle and Heracles Almelo who play on artificial turf for several years).

The procedure to construct the models will be explained here. In case the constructed attributes in the group of 'angstgegner' are considered, one can imagine that the amount of attributes is large. This is also the case for the constructed variables in other fields of information. In order to overcome the problem of a massive amount of variables in the ordered probit models, a pre-selection step is performed. For this filtering procedure a simple ordered probit model is formed for all variables, expressed by $Attribute_i$, as is given in Eq. 4.12. The variable $ECIHome_i$ - $ECIAway_i$ is included to correct for the difference in strength of the competing clubs, in order to obtain an accurate estimate of β_2.

$$y_i^* = \beta_1 + \beta_2 Attribute_i + \beta_3[ECIHome_i - ECIAway_i] + \epsilon_i, \tag{4.12}$$

where $\epsilon_i \sim N(0, 1)$.

Considering the coefficient β_2 in Eq. 4.12, only the attributes with a p-value lower than 0.075 are kept. The resulting selection of remaining variables is used to construct the models. In order to do so, first all remaining variables are included in the model. Subsequently, reduction steps are performed using backward elimination [17]. Another model is constructed using the Akaike information criterion (AIC) [3]. This third model is referred to as the extended AIC model from now on. Stepwise is checked whether excluding the least significant variable, based on p-value, will improve the model according to the model fit of AIC. The model fit of AIC is calculated by

$$AIC = \frac{-2l}{T} + \frac{2k}{T}, \tag{4.13}$$

where l indicates the loglikelihood value, T represents the number of observation and k equals the amount of parameters included in the model.

A fourth model is constructed using an almost similar approach. Instead of AIC the inclusion of variables for the fourth model depends on the Bayesian information criterion (BIC) [4]. AIC and BIC are both useful measures to estimate the fit of the constructed model with a penalty assigned for each included parameter. Overall, the AIC prefers more complex models, whereas BIC chooses the model that contains the least number of parameters. The fourth model is referred to as the extended BIC model in the remainder of the paper. The formula of this information criterion is the following

$$BIC = \frac{-2l}{T} + \frac{klog(T)}{T}. \tag{4.14}$$

Each season the composition of clubs playing in a certain competition is different due to seasonal promotion and relegation. Besides, the selection of referees operating in the highest national division changes over time. As some of the constructed variables depend on the specific clubs and referees in a competition, the considered variables vary per competition and even per season. This makes the extended AIC models and the extended BIC models competition- and season-specific.

4.3 | Comparison of match predictions

In the previous section four prediction models are introduced. In this section methods are discussed to compare the outcome of these models with each other and to the probabilities assigned by bookmakers. In Subsect. 4.3.1 a method to convert bookmakers odds to probabilities is given. Subsequently, Subsect. 4.3.2 enumerates several evaluation measures to compare the prediction accuracy of the models.

4.3.1 Converting bookmakers odds

As is mentioned in Ch. 3 data from *www.oddsportal.com* is used to gather the average assigned odds by multiple bookmakers for all matches. These bookmakers odds can be converted to predicted outcome probabilities, also called bookmakers probabilities. The conversion is performed by the following two steps of equations according to Graham and Scott [14].

$$R = \frac{1}{\Theta_{Home}} + \frac{1}{\Theta_{Draw}} + \frac{1}{\Theta_{Away}}, \tag{4.15}$$

$$\begin{aligned} P[y_i = 1] = \Omega_{Home} &= \frac{(1/\Theta_{Home})}{R}, \\ P[y_i = 0] = \Omega_{Draw} &= \frac{(1/\Theta_{Draw})}{R}, \\ P[y_i = -1] = \Omega_{Away} &= \frac{(1/\Theta_{Away})}{R}, \end{aligned} \tag{4.16}$$

where Θ_{Home}, Θ_{Draw} and Θ_{Away} represent the bookmakers odds for a home win, a draw and an away win, respectively. Furthermore, Ω_{Home}, Ω_{Draw} and Ω_{Away} represent the assigned probabilities for a home win, a draw and an away win, respectively.

Table 4.1 gives the bookmakers odds assigned for the match Arsenal - Tottenham Hotspur in the season 2013-14. The aforementioned equations are applied on these odds resulting in the bookmakers probabilities, given in the same table.

Table 4.1: The bookmakers odds and bookmakers probabilities assigned for the match Arsenal - Tottenham Hotspur in the season 2013-14.

Arsenal	vs.	Tottenham
Θ_{Home}	Θ_{Draw}	Θ_{Away}
2.12	3.41	3.62
Ω_{Home}	Ω_{Draw}	Ω_{Away}
0.453	0.282	0.356

4.3.2 Comparing the models

In order to compare the outcome predictions of the models several evaluation methods are considered. The models have in common that a probability is assigned to all three possible outcomes for each match. These assigned probabilities can be considered in two ways. On the one hand one can see the outcome with the highest assigned probability as the predicted outcome. On the other hand one can interpret all three assigned probabilities together. Based on these two interpretations, six different eval-

uation methods are introduced. For some of the evaluation methods the calculation is illustrated with an example regarding three matches of the English 'Premier League' in September 2013, given in Table 4.2. The assigned predicted probabilities for these matches are obtained from the extended AIC model described in Subsect. 4.2.3.

Table 4.2: The outcome probabilities for three matches in the English 'Premier League' seasons 2013-14 calculated by the extended AIC model.

Match	Ω_{Away}	Ω_{Draw}	Ω_{Home}	Result
Arsenal - Aston Villa	0.101	0.230	0.669	1-3
Liverpool - Stoke City	0.143	0.266	0.591	1-0
Norwich - Everton	0.429	0.316	0.255	2-2

For each of the six evaluation methods the corresponding mathematical formulation will be given in this subsection. Input for these formulas are matrices representing the predicted match outcomes, real-life results and assigned probabilities. For the matches in Table 4.2 these matrices are constructed and given by the following matrices

$$A = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} \tag{4.17}$$

$$B = \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} \tag{4.18}$$

$$\Omega = \begin{bmatrix} 0.101 & 0.230 & 0.669 \\ 0.143 & 0.266 & 0.591 \\ 0.429 & 0.316 & 0.255 \end{bmatrix} \tag{4.19}$$

where A represents the predicted match outcomes, B gives the real-life results and Ω shows the assigned probabilities. The columns of this matrix Ω give the assigned probabilities of an away win, draw and home win, respectively.

Direct outcome evaluation

For this method the outcome with the highest assigned probability is seen as the predicted outcome. This method counts the well predicted matches and divides this by the total number of matches in the season per competition. The *direct outcome evaluation value* (DO value) is calculated using the following formula

$$DO(A, B) = \frac{|A \cap B|}{|A|}. \tag{4.20}$$

Direct probabilities evaluation

For this method the assigned probabilities are used directly. The method sums up the assigned probability of the real-life outcome and divides this by the total number of matches in the season per competition expressed. The *direct probabilities evaluation value* (DP value) is calculated using this formula

$$DP(B,\Omega) = \frac{\sum_i \sum_{k=1}^{3} \Omega_{i,k}\ I[B_i = k-2]}{|B|}, \tag{4.21}$$

where $I[\cdot]$ represents an indicator function that equals 1 in case the statement between brackets holds, and 0 otherwise.

Jaccard index

The third evaluation measure to compare the predicted outcomes from the models is performed based on the *Jaccard similarity index* [19]. This measure compares the predicted outcomes with the real-life results. In order to make this possible the predicted outcome equals the match outcome possibility with the largest assigned probability. The *Jaccard similarity value* (J value) is calculated using the following equation

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A|+|B|-|A \cap B|}. \tag{4.22}$$

The obtained Jaccard value is bounded between 0 and 1. The closer the *Jaccard similarity index* reaches 1, the more similar the vectors and the more accurate the prediction. The matches in Table 4.2 result in the following *Jaccard similarity values* $J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A|+|B|-|A \cap B|} = \frac{1}{3+3-1} = \frac{1}{5}$.

Ordered probit penalty index

The *ordered probit penalty index* evaluates the distribution of chances over the outcome possibilities. This method depends on the division of the assigned probability and the real-life outcome expressed in {-1,0,1}, and penalizes the amount of wrong assigned probabilities. In case the real-life score equals a home win, the assigned probability of an away win will be penalized twice as hard as the assigned probability for a draw. The calculation of this evaluation method is the following:

$$OPP(B,\Omega) = \sum_i \begin{cases} \Omega_{i,3} - \Omega_{i,2} - 2\Omega_{i,1}, & \text{if } B_i = 1 \\ \Omega_{i,2} - \Omega_{i,3} - \Omega_{i,1}, & \text{if } B_i = 0 \\ \Omega_{i,1} - \Omega_{i,2} - 2\Omega_{i,3}, & \text{if } B_i = -1 \end{cases}, \tag{4.23}$$

To illustrate this measure the matches from Table 4.2 are considered once more. The OPP value of the first match equals 0.101 - 0.230 - $2 * 0.609 = -1.338$. The OPP value for the three matches combined equals the average OPP value, calculated by $\frac{-1.388+0.038-0.368}{3} = -0.327$. The model containing the highest OPP value has most accurately assigned probabilities to the possible match outcomes according to this evaluation measure.

Rank probability score

In [39], Štrumbelj and Šikonja employed the *rank probability score* (RPS) to evaluate the accuracy of odds set by several bookmakers. Eq. 4.24 presents this method to evaluate the probability forecasts of ranked categories that was introduced in 1969 by Epstein [9].

$$RPS(B, \Omega) = \frac{1}{J} \sum_{j}^{J} \sum_{i=1}^{3} \left(\sum_{k=1}^{i} \Omega_{j,k} - B_j \right)^2, \tag{4.24}$$

where J is the amount of matches under consideration.

In order to illustrate the method it will be applied on the matches given in Table 4.2. For the first match, the calculation equals $(0.101 - 1)^2 + ((0.101 + 0.230) - 1)^2 + (1 - 1)^2 = 1.256$. Similar calculation for the other matches results in a RPS value of $\frac{1.256+0.188+0.249}{3} = 0.564$ applying the *rank probability score*. The model containing the lowest RPS value indicates the most accurate model according to the *rank probability score*.

Mean squared error

The *mean squared error method* [31] is a widely used evaluation method that compares the predicted outcome with the real-life outcome. The *mean squared error value* (MSE value) is calculated using the following equation

$$MSE(B, \Omega) = \frac{1}{J} \sum_{j}^{J} \left(\frac{\sum_{i=1}^{3} \left(\Omega_{j,i} - B_j \right)^2}{l} \right), \tag{4.25}$$

where l represents the number of ordered categories. An illustration of this method is given by applying it on the matches given in Table 4.2. For the first match of this table the individual MSE value equals $\frac{(-0.669-0)^2+(0.230-0)^2+(0.101-1)^2}{3} = 0.436$.

The MSE value of the three matches combined is given by $\frac{0.436+0.086+0.239}{3} = 0.254$. For the *mean squared error method* the model with the lowest MSE value is the most accurate.

The six aforementioned evaluation methods are applied to compare the models with each other and the bookmakers predictions, obtained from the converted bookmakers odds.

4.4 | Proposed betting strategies

In this section various betting strategies will be proposed depending on the bookmakers odds and the match outcome probabilities assigned by the models. At first, a strategy is considered where one bets that all matches from a season will end in a similar outcome, independent on the assigned outcome probabilities. This 'naive' strategy is studied for home wins, draws or away wins in order to study whether one of the three outcome possibilities is systematically under- or overpriced. Another considered betting strategy is that bettors bet a unit of stake on the outcome possibility containing the largest assigned probabilities per match. Statistically more advanced betting strategies seem more likely to result in sustainable profit. In [23], Langseth suggests that with betting on soccer matches it is more important to gain money than to be right about the outcome. Imagine the following three possible independent betting options;

I Probability of an away win is ($\Omega_j =$) 0.1 with and odd ($\Theta_j =$) 12

II Probability of an away win is ($\Omega_j =$) 0.4 with and odd ($\Theta_j =$) 3

III Probability of a home win is ($\Omega_j =$) 0.8 with and odd ($\Theta_j =$) 1.5

The gross of bettors will choose different bets based on their personal believes and risk attitude. Betting option I contains a small probability of occurrence compensated by a relatively large potential payout, whereas the potential payout of option III, 1.5, has a relatively large probability. Clearly, a bettor does not necessarily have to choose between the betting options. The betting budget C can be divided over $N > 1$ betting options with varying weights c_i for different matches i, where $\sum_i^N c_i \leq C$.

Each match contains three possible betting options, with corresponding bookmakers odds Θ_j and assigned probability Ω_j resulting from one of the constructed models, where $\Theta_j \in \{\Theta_{Home}, \Theta_{Draw}, \Theta_{Away}\}$ and $\Omega_j \in \{\Omega_{Home}, \Omega_{Draw}, \Omega_{Away}\}$. The odds or potential payouts can be seen as a compensation for the uncertainty of the outcome and the risk taken by the bettor. The key-step is to approach the potential profit Π_j of betting option j per unit of stake as a random variable with $\mathrm{E}[\Pi_j] = (\Omega_j\Theta_j$ - 1) and $\mathrm{Var}[\Pi_j] = \Omega_j^2\Theta_j(1 - \Theta_j)$. Each playing round of league matches multiple betting options might have an positive expected profit ($\mathrm{E}[\Pi_j] > 0$). The aforementioned three

betting options, which vary in assigned odds and probabilities, have equal expected payout. For the betting strategies introduced in the remainder of this section, the set of betting options is restricted to the ones with an expected profit that is strictly positive. The betting budget, C, for each of these betting strategies equals the amount of betting options with a positive expected profit per playing round. This way the budget is similar for all betting strategies per prediction model. The resulting weights for the betting options will be used to split the betting budget, C, over the N betting options per playing round as that $\frac{c_j}{\sum_{i=1}^{N} c_i} C$ equals the specific amount bet on betting option j. Using these notations several betting strategies will be introduced.

Equal weights

The N betting options with positive expected profit receive similar weights independent of the corresponding odds and assigned probability, equal to $c_i = 1$.

Equal payout

This betting strategy assigns weights to the betting options as that the potential output is similar for all betting options in one specific playing round. Betting options containing a relatively small assigned outcome probability will get smaller weights. On the other hand larger weights will be assigned to betting options with larger assigned probability. The weights equal $c_i = \frac{1}{\Theta_i}$.

Variance adjusted betting strategy

Rue and Salvesen [34] propose to take the variance of the betting options into account to obtain optimal weights for the betting options. It is suggested to maximize the expected profit whilst the variance of the profit is minimized. This equals a minimization of the expected profit minus the variance and results in the following minimization

$$\min_{c_i} \; c_i \Theta_i \Omega_i \; - \; \Omega_i (1 - \Omega_i)(c_i \Theta_i)^2. \tag{4.26}$$

The weights resulting from this minimization equal

$$c_i = \frac{1}{2\Theta_i \; (1 - \Omega_i)}. \tag{4.27}$$

Kelly's betting strategy

As a fourth betting strategy the proposed betting approach of Kelly [20] is implemented. In this strategy the betting budget C is incorporated to obtain optimal weights. The utility of having amount C after a bet is determined at $\ln C$, whereas the utility of going bankrupt equals -∞. The expected utility is calculated by maximizing the following equation

$$\max_{c_i} \ \Omega_i \ \ln(C + c_i\Theta_i) + (1 - \Omega_i) \ \ln(C - c_i). \tag{4.28}$$

This equation generates the following weights

$$c_i = C \ \frac{\Omega_i\Theta_i - 1}{\Theta_i - 1}. \tag{4.29}$$

A drawback for all of these four betting strategies is that each betting option with positive expected profit is bet on. Although for some strategies the weights depend on the assigned probability, each betting option with a positive expected profit is bet on. Imagine a betting option with an assigned outcome probability of 0.01 and an odd equal to 101. Despite a positive expected profit one can imagine that a rational bettor is not likely to bet on this betting option assuming that the predicted outcome probability is assigned by an accurate prediction model. For this reason an extra restriction on the set of betting options is performed prior to applying one of the four introduced betting strategies. This restricts the set of betting options to the ones with $\Omega_j > \phi$, where ϕ presents a treshold value between 0 and 0.80. The proposed betting strategies are applied for various values of ϕ. Hence, each of the betting strategies is applied on a set of betting options with positive expected profit and $\Omega_j > \phi$.

4.5 | Season predictability

This section proposes three methods that attempt to measure the predictability of a certain season. It is investigated whether the obtained profit share per bet euro is related to the season predictability. The predictability of a season is indicated using three different measures. In Subsect. 4.5.1 a simulation algorithm is introduced in order to predict the season outcome. Subsect. 4.5.2 describes a method to compare the resulting season tables of two consecutive seasons. As a third indicator the profit obtained from betting on the outcome possibility with the largest assigned probability is considered.

4.5.1 Season simulation

In Sect. 4.2 is explained that the difference in strength of the competing clubs, embodied by the ECI value of the clubs, is the main seed for each of the four constructed models. In this subsection the ECI value for clubs is used to simulate the outcome of competition seasons. For each team the ECI value at the start of the competition is input for the simulation algorithm. The outcome probabilities for the considered matches are assigned using the simplified model, described in Subsect. 4.2.2. For each match the assigned outcome probabilities are used to calculate the expected result by

$$\begin{aligned} ExpectedResult_i &= 1 * \Omega_{Home} + 0 * \Omega_{Draw} - 1 * \Omega_{Away} \\ &= \Omega_{Home} - \Omega_{Away}. \end{aligned} \tag{4.30}$$

Once $ExpectedResult_i$ is calculated, $Result_i \in \{-1, 0, 1\}$ is simulated from the assigned outcome probabilities using a random number generator. Given the simulated outcome the exact score is simulated based on the occurrence frequency since season 2000-01. The exact score indicates the goals scored by both teams. The simulated $Result_i$ and $ExpectedResult_i$ lead to an update of the ECI value for both clubs. This change in ECI value, the increment value, is determined by the following equation

$$Increment_i = \kappa \, [Result_i - ExpectedResult_i], \tag{4.31}$$

where κ is called the update factor.

This update factor is calculated for each competition season and remains fixed for the season. The value of κ depends on the ECI value of the clubs at the start of the competitions; namely the difference of the maximum and minimum ECI value of clubs within the competition. Table 4.3 shows the value of κ that belongs to specific value ranges of $max\ ECI - min\ ECI$. This parameter estimation is encountered from the dataset.

Table 4.3: This table presents the parameter value of κ for different values of $max\ ECI - min\ ECI$.

max ECI - min ECI	κ
2000+	35
1500-2000	30
1000-1500	25
1000-	20

Subsequently, the ECI value of both clubs will be updated using the value of $Increment_i$. Consider a match played at time t, the updated ECI value for both the home and away club at time $t + 1$ is calculated using

$$\begin{aligned} ECIHome_{t+1,i} &= ECIHome_{t,i} + Increment_i, \\ ECIAway_{t+1,i} &= ECIAway_{t,i} - Increment_i, \end{aligned} \tag{4.32}$$

where $ECIHome_{t,i}$ and $ECIAway_{t,i}$ represent the ECI values before the match and $ECIHome_{t+1,i}$ and $ECIAway_{t+1,i}$ represent the updated ECI values after the match.

Eq. 4.32 indicates that [$ECIHome_{t+1,i}$ - $ECIHome_{t,i}$] = - [$ECIAway_{t+1,i}$ - $ECIAway_{t,i}$], as that a similar amount of ECI points is maintained within the ranking system after each match. Later in this subsection an example is given concerning the update of the ECI value for several matches.

The described procedure is performed for all matches in the season schedule, providing a season simulation. This season simulation is performed 10,000 times which results in a prediction of the resulting season table. Alg. 1 summarizes the performed simulation algorithm.

Algorithm 1 Simulation

Initialize the match schedule of the competition season, containing K matches.
Initialize the ECI starting values for each club in the competition season.
Initialize the competition season parameters μ_1, μ_2 and κ.
Initialize the exact result probabilities for each outcome possibility gathered from season 2000-01 through the prior season.

for Run $m \in$ 10,000 **do**
  for (match $i = 1$ to K) **do**
    Calculate the outcome probabilities for match i (Eq. 4.5, Eq. 4.7 and Eq. 4.6)
    Calculate *Expected Result* (Eq. 4.30).

    Simulate the outcome of match i using random number generation.
    Simulate the exact result of match i using random number generation.

    Calculate the increment of the ECI value based on the result (Eq. 4.31).
    Update the ECI values of both clubs of match i (Eq. 4.32).

  end for
  Determine the winner of league simulation run m.

end for
Determine the probabilities to become champion for each club based on the m performed simulation runs.

For simplicity, the following two assumptions are made for the proposed simulation algorithm;

- The results of European matches and national cup matches have no influence on the performance in the national league, nor on the ECI value of the clubs in the league. This causes a simplified representation of the reality because in reality the results of clubs in other competitions most certainly can have influence on the performance in the national league.

- The strength difference between the competing clubs is related to the outcome of the match {home win, draw, away win} via the models. The exact score of the matches does not depend on the strength difference of the competing clubs within this algorithm. For the simulation algorithm, the exact score of matches has only small impact on the outcome of the season. Solely in case two or more clubs lead the season table with an equal amount of points at the end of the simulated season, the goal difference gets determinative. In the end, goal difference is worth less than one point. The simulation output points out that goal difference is crucial in defining the champion for only two to nine percent of the simulation runs. This percentage varies per competition and season, depending on the number of club sin the competition and the starting ECI values of the corresponding clubs.

Next, examples are given regarding updating the ECI value of the competing clubs for three matches of the English 'Premier League' season 2013-14. The three considered matches are given in Table 4.4 with the corresponding ECI values.

Table 4.4: This table presents three matches played in the 'Premier League' in the season 2013-14.

Home team	ECI	Away team	ECI
Manchester City	3499	Norwich City	2244
Fulham	2419	Manchester United	3692
Arsenal	3474	Liverpool	3067

Applying the simplified ordered probit model, described in Subsect. 4.2.2, the estimated outcome probabilities for the three matches are given in Table 4.5. The expected results of the matches, calculated by Eq. 4.30, are given as well. For the English 'Premier League' season 2013-14 the value of κ equals 30. Subsequently, the increment value for the three possible outcomes for each match can be calculated using Eq. 4.31.

Table 4.5: This table presents the outcome probabilities and the expected result for each of three matches accompanied with the increment value for each of the possible outcomes.

Home team	Away team	Ω_{Home}	Ω_{Draw}	Ω_{Away}	Expected Result
Manchester City	Norwich City	0.776	0.267	0.560	0.720
Fulham	Manchester United	0.148	0.266	0.586	-0.439
Arsenal	Liverpool	0.561	0.275	0.163	0.398
		Δ ECI	Δ ECI	Δ ECI	
Manchester City	Norwich City	9.8	-25.2	-60.2	
Fulham	Manchester United	50.3	15.3	-19.7	
Arsenal	Liverpool	21.1	-13.9	-48.9	

4.5.2 Comparison of season table vectors

In this subsection the predictability of the competition season tables is studied by another measure. The similarity of resulting season tables of two consecutive seasons is investigated. In order to do so, the resulting season tables of two consecutive seasons are transformed into two vectors. The first vector represents the resulting season table for the clubs in the successive season, whereas the second vector contains the resulting rank for the clubs in the season table of the prior season. The included clubs per season will be different caused by promotion and relegation. For this reason the comparison is performed on the clubs in the highest league of the successive season. The clubs that were relegated in the prior season are removed from the season table. These clubs are replaced by the clubs that earned a promotion to the highest league during this season. In case multiple clubs got promoted, they get ranked based on their results in the second division. Imagine club p (1^{st} in second division) and club q (2^{nd} in second division) got promoted to the English 'Premier League', that contains twenty club. As two clubs got promoted obviously two clubs got relegated as well. For eighteen clubs the ranking in both seasons can be retrieved directly. For the two remaining clubs, p and q, determination of the rank in the successive season is straightforward. Furthermore, for the the prior season club p gets rank nineteen and club q gets rank twenty assigned based on the results in the second division.

To illustrate this method, assume an imaginary European competition containing six clubs from the four considered competitions. Table 4.6 represents the resulting competition table of the seasons 2019-20 and 2020-21. For this competition the club that ended last will be replaced by the champion of the second division, presented with an asterisk in the table.

Table 4.6: This table presents the competition table of the seasons 2019-20 and 2020-21 of the imaginary competition. 1* represents the champion of the second division, which is promoted to the highest division for the consecutive.

Club	2019-20
Atlético Madrid	1
Barcelona	2
Feyenoord	3
Arsenal	4
Liverpool	5
Ajax	6
Malmö FF	1*

Club	2020-21
Arsenal	1
Barcelona	2
Atlético Madrid	3
Feyenoord	4
Malmö FF	5
Liverpool	6
PSV Eindhoven	1*

Performing the proposed method on this competition results in the following two vectors;

$$C = \begin{bmatrix} 4 \\ 2 \\ 1 \\ 3 \\ 6 \\ 5 \end{bmatrix} \quad (4.33) \qquad D = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{bmatrix}, \quad (4.34)$$

where the clubs playing in the highest division in season 2020-21 are considered for both vectors. Vec. 4.33 represents the resulting season table of the season 2019-10, whereas Vec. 4.34 represents the resulting season table for the season 2020-21.

A similarity measure is performed on the two vectors to indicate the predictability of the successive season. This predictability of a season is indicated by the cosine similarity measure, given by

$$cos(C, D) = \frac{C \cdot D}{||C||\ ||D||} = \frac{\sum_{i=1}^{n} C_i\ D_i}{\sqrt{\sum_{i=1}^{n} C_i^2}\ \sqrt{\sum_{i=1}^{n} D_i^2}}, \quad (4.35)$$

where C_i and D_i are observations of the vectors C and D, respectively.

In case the cosine similarity equals 1, the resulting season tables are similar in the two seasons. The minimum value of the cosine similarity for this competition comparison is around 0.5. This occurs in case the resulting seasons table of the successive season is in no way similar to the resulting season table of the prior season. Hence, in case the champion of the prior season finishes as last in the successive seasons, whereas the promoted club becomes champion in the successive season, and so on. The season end ranks are normalized before the cosine similarity measure is performed. This measures provides a similarity compared to the previous season indicating the relative predictability of the considered season.

5 | Evaluation

This chapter contains the evaluation and discussion of the results retrieved from the proposed methods in Ch. 4. In Sect. 5.1 the results for the applied PCA and cluster analysis are discussed. The constructed models will be evaluated in Sect. 5.2 previous to the comparison of the models in Sect. 5.3 based on several evaluation methods. Consequently, several betting strategies are performed using the results of the constructed models. The profitability of these betting strategies will be discussed in Sect. 5.4. Eventually, robustness checks of the betting strategies are performed and studied in Sect. 5.5.

5.1 | PCA and Clustering

In order to select a number of competitions from a total of sixteen European competitions, of which the data is complete, is decided to perform a cluster analysis. Caused by high correlation between the attributes of competition characteristics most clustering methods are not applicable. To overcome this issue a combination of PCA followed by K-means clustering is performed. The first five components resulting from PCA on a normalized dataset of competition characteristics are given in Table A.1. The first component reflects the strength of the competitions, whereas the second component seems to cover the difference in strength of clubs within the competitions. A biplot of the first two components, that is given in Fig. A.1, shows this. The other three presented components obey a clear declaration. Table 5.1 shows the amount of variance each of the principal components declare as well as the eigenvalue of these components. The number of resulting components to use for the K-means clustering is three, as only the first three components have an eigenvalue larger than one. Besides, each of these three components explain more than $1/Q$ of the total variance, where Q indicates the number of included variables. For this study Q equals 15. An overview of the exact included variables can be found in Appendix B. The three suggested components explain cumulatively 77.5% of the total variance.

Table 5.1: This table gives the amount of variance each of the principal components (PCs) declare and the eigenvalues of the components.

	PC1	PC2	PC3	PC4	PC5
Proportion of variance	0.545	0.129	0.101	0.059	0.056
Cumulative proportion	0.545	0.674	0.775	0.835	0.890
Eigenvalue	8.179	1.930	1.520	0.890	0.835

Using the first three constructed principal components resulting from PCA, the K-means clustering algorithm is performed to split the sixteen competitions into groups. A pseudo-F measure points out that two clusters provide optimal clustering for this selection of competitions. Table 5.2 presents the division of the sixteen competitions over the two clusters. Cluster one contains seven European competitions that can be seen as the most appealing competition. The other competitions are assigned to the second cluster. It seems like the European rank of the competitions has large influence on the cluster classification. The obtained cluster means, in case the number of clusters (K) equals two, can be found in Appendix B.

Table 5.2: This table gives the competitions per cluster.

Cluster 1	Cluster 2
England	Austria
France	Belgium
Germany	Denmark
Italy	Netherlands
Portugal	Norway
Russia	Scotland
Spain	Sweden
	Switzerland
	Ukraine

For the remainder of the paper a number of European competition has to be selected in order to study the accuracy of soccer match outcome prediction. Multiple competitions and seasons will be considered to ensure the prediction accuracy of the models for several European competitions. When the selected competitions are different, based on characteristics, the strength of the profitable proposed models is reinforced. From each of the two obtained clusters two competitions will be selected that will be studied in more detail. From the first cluster the competitions of England and Spain are selected, which are the most watched and followed soccer competitions worldwide with the most money involved. Also, these competitions vary a lot from each other as the Spanish competition is yearly dominated by the same small amount of clubs, whereas the English competition has been less predictable over the seasons 2007-08 through 2015-16. From the second cluster the Dutch competition is selected because extra data concerning artificial turf is available for the Dutch competition. The Swedish competition is chosen as fourth competition such that one of the considered competitions is a summer competition.

5.2 | Model predictions

For each of the four selected European competitions models are constructed for three consecutive seasons. These four different prediction models are proposed in Ch. 4. The models are formed for the seasons 2013(-14), 2014(-15) and 2015(-16) for the competitions of England, Netherlands, Spain and Sweden based on data since the start of the season 2007-08.

The basic ECI model, described in Subsect. 4.2.1, can be seen as a straightforward model. The latent variable for each match i depends solely on the difference in ECI value of the competing clubs, $ECIHome_i$ - $ECIAway_i$. Table 5.3 contains the estimated parameters for the three seasons of the four selected competitions. The table presents that β is positive for all competitions and seasons, meaning that if the difference in ECI value ($ECIHome_i$ - $ECIAway_i$) becomes larger, the probability of a home win gets larger. The coefficients can not be compared among the seasons and competitions because the other parameters, μ_1 and μ_2, vary as well. The table indicates that all three parameters (β, μ_1 and μ_2) do not differ largely over the seasons, whereas the difference between competitions is larger.

Table 5.3: The table gives the treshold values from the seasons 2013-14, 2014-15 and 2015-16 of The English, Dutch, Spanish and Swedish competition estimated by the Basic ECI model. Between brackets the standard errors are given.

England	β	μ_1	μ_2
2013-14	6.55E-4 (3.24E-5)	-0.696 (0.030)	0.102 (0.028)
2014-15	6.58E-4 (3.00E-5)	-0.672 (0.028)	0.099 (0.026)
2015-16	6.52E-4 (2.80E-5)	-0.659 (0.026)	0.104 (0.024)
Netherlands	β	μ_1	μ_2
2013-14	9.33E-4 (5.04E-5)	-0.650 (0.033)	0.043 (0.030)
2014-15	8.95E-4 (4.67E-5)	-0.656 (0.030)	0.042 (0.028)
2015-16	8.91E-4 (4.34E-5)	-0.644 (0.028)	0.054 (0.026)
Spain	β	μ_1	μ_2
2013-14	6.29E-4 (3.52E-5)	-0.671 (0.030)	0.003 (0.027)
2014-15	6.22E-4 (3.19E-5)	-0.660 (0.027)	0.012 (0.025)
2015-16	6.36E-4 (2.91E-5)	-0.653 (0.026)	0.028 (0.024)
Sweden	β	μ_1	μ_2
2013	6.83E-4 (6.78E-5)	-0.595 (0.038)	0.100 (0.035)
2014	7.05E-4 (6.17E-5)	-0.592 (0.035)	0.110 (0.033)
2015	7.20E-4 (5.78E-5)	-0.594 (0.032)	0.110 (0.030)

For the simplified model the coefficient β is fixed at a value of $\frac{1}{1000}$. The other parameters, μ_1 and μ_2, are retrieved from the optimal treshold values of the three preliminary seasons, depending on the division of match outcomes in these seasons. Table 5.4 presents the resulting treshold values, μ_1 and μ_2, for the considered seasons of the four competitions.

Table 5.4: This table gives treshold values for the seasons 2013-14, 2014-15 and 2015-16 of The English, Dutch, Spanish and Swedish competition estimated by the Simplified model.

Competition	2013(-14)		2014(-15)		2015(-16)	
	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2
England	-0.690	0.136	-0.596	0.138	-0.601	0.136
Netherlands	-0.692	0.019	-0.669	0.072	-0.648	0.126
Spain	-0.688	-0.011	-0.661	0.036	-0.613	0.080
Sweden	-0.631	0.077	-0.609	0.126	-0.607	0.187

For the extended AIC model and the extended BIC model, the latent variable depends on a lot of other variables alongside the difference in ECI value of the competing clubs. The resulting extended models for three seasons of the four considered competitions can be found in Appendix C and Appendix D, respectively. In those tables all included variables are presented along with the corresponding coefficients and standard errors. An explanation of the included variables can be found in Appendix F. The models for a specific competition vary widely per season, indicating that the prediction of match outcomes is not straightforward. As expected, the models where the incorporated variables are selected based on BIC contain significantly less variables compared to the models where the variables are selected based on AIC.

In Table 5.5, Table 5.6, Table 5.7 and Table 5.8 the variables are presented that are consistently significant, and so included, in the extended AIC models for the prediction of the seasons 2013(-14) through 2015(-16) of the English, Dutch, Spanish and Swedish competition, respectively. Explanation of these included variables can be found in Appendix F. For each of the competitions the most striking significant coefficient(s) will be discussed.

Table 5.5: The table contains the variables that are significant in all three extended AIC models (2013-14, 2014-15 and 2015-16) for the English competition. '+' means a positive coefficient, whereas '-' means a negative coefficient in the models for the three considered seasons.

Variables	Sign
$5MatchesShapeHteam - 3_i$	+
$5MatchesShapeHteam + 3_i$	+
$AwaymatchesHteam_i$	-
$ECIHome_i$ - $ECIAway_i$	+
$EUAwayDrawPostWinter_i$	+
$EvertonShape - 1_i$	+
$HomeManCity_i$	+
$ManCityENGReferee4_i$	-
$ManCityVS.Newcastle_i$	+
$ManUnitedShape - 3_i$	+
$Shape + 1VS.Shape + 1_i$	+

The negative coefficient for $AwaymatchesHteam_i$ in Table 5.5 shows that data from preliminary seasons of the English highest league indicates that the probability of a home win declines with the growth of the amount of away matches played by the home team. In other words, the probability of a home win is smaller at the end of the season than a home win of the similar match at the begin of the seasons, ceteris paribus. Other notable coefficient signs are found for $ManCityENGReferee4_i$ and $ManUnitedShape - 3_i$. The negative coefficient for $ManCityENGReferee4_i$ indicates that Manchester City performs less in case English Referee 4 leads their league match, ceteris paribus. The positive sign for $ManUnitedShape - 3_i$ shows that in case Manchester United performed bad in recent league matches, the data of preliminary seasons assert a larger probability of a win for the next league match of Manchester United.

Table 5.6: The table contains the variables that are significant in all three extended AIC models (2013-14, 2014-15 and 2015-16) for the Dutch competition. '+' means a positive coefficient, whereas '-' means a negative coefficient in the models for the three considered seasons.

Variables	Sign
$\Delta ECIOnArtificialTurfAteam_i$	-
$AjaxVS.Utrecht_i$	-
$AverageGoalsAgainstAteam_i$	+
$AZNovember_i$	+
$ECIHome_i$ - $ECIAway_i$	+
$FeyenoordJanuary_i$	-
$GroningenShape-2_i$	-
$PSVBefore2PM_i$	-
$Shape-3VS.Shape+1_i$	+
$Shape+1VS.Shape+1_i$	+
$Shape+1VS.Shape+3_i$	-
$Spectators_i$	+
$UtrechtBetween6PMAnd8PM_i$	-
$VitesseNETReferee2_i$	-

Table 5.6 shows that the models built for the seasons 2013-14, 2014-15 and 2015-16 of the Dutch soccer competition all find a negative coefficient for the variable $\Delta ECIOnArtificialTurfAteam_i$. This indicates that data from preliminary seasons finds that the larger the amount of gathered ECI points in away matches on artificial turf, the larger the probability of an away win against another opponent playing home matches on artificial turf. The negative sign for $AjaxVS.Utrecht_i$ shows that historic data from seasons since 2007-08 indicates that FC Utrecht is a so-called 'angstgegner' for Ajax, meaning that the probability of a win for Ajax is smaller with FC Utrecht as opponent compared to similarly performing other teams, ceteris paribus. The negative signs for $PSVBefore2PM_i$ and $UtrechtBetween6PMAnd8PM_i$ indicate a significant worse performance for PSV and FC Utrecht to play before 2 PM and between 6 PM and 8 PM, respectively. The probability of a win for FC Utrecht is smaller in case the match is played before 2 PM and a win for PSV gets less likely in case they play a match between 6 PM and 8 PM, ceteris paribus.

Table 5.7: The table contains the variables that are significant in all three extended AIC models (2013-14, 2014-15 and 2015-16) for the Spanish competition. '+' means a positive coefficient, whereas '-' means a negative coefficient in the models for the three considered seasons.

Variables	Sign
$BilbaoSPAReferee1_i$	-
$ECIHome_i$ - $ECIAway_i$	+
$EspanyolShape-3_i$	-
$EULossPreWinterAteam_i$	-
$EUAwayWinAteam_i$	-
$GetafeSPAReferee5_i$	+
$GetafeSPAReferee6_i$	-
$HomeBetis_i$	-
$SevillaSPAReferee4_i$	+
$SevillaSPAReferee11_i$	+
$Shape+1VS.Shape+1_i$	+
$Shape+2VS.Shape-1_i$	+
$ValenciaMarch_i$	-

The extended AIC models for the seasons 2013-14, 2014-15 and 2015-16 of the Spanish competition all find a negative sign for $ValenciaMarch_i$. This indicates that historic data from Spanish seasons since 2007-08 show that the results of matches from Valencia played in March are significantly worse than other combinations of equally strong clubs and months, ceteris paribus. Furthermore, Table 5.7 presents that preliminary seasons point out that an European away win followed by an away league match gives rise to the probability of an away victory. This can be concluded from the negative coefficient sign for $EUAwayWinAteam_i$.

Table 5.8: The table contains the variables that are significant in all three extended AIC models (2013-14, 2014-15 and 2015-16) for the Swedish competition. '+' means a positive coefficient, whereas '-' means a negative coefficient in the models for the three considered seasons.

Variables	Sign
$3MatchesShapeAteam+2_i$	-
$AIKShape2_i$	+
$CupfighterAteam_i$	-
$DjurgårdensSWEReferee2_i$	-
$ECIHome_i$ - $ECIAway_i$	+
$MatchesLeftHteam_i$	+

Table 5.8 presents the variables that are consistently significant in the prediction models for the seasons 2013, 2014 and 2015 of the Swedish competition. The negative coefficient sign for $CupfighterAteam_i$ indicates that in case the away team of a Swedish league match is still playing in the national cup, the probability of a home win is lower than when the away team is already eliminated from the national cup. The positive sign for $MatchesLeftHteam_i$ shows that the probability of a home win gets smaller as the home team has played more matches.

In Table 5.9, Table 5.10, Table 5.11 and Table 5.12 the variables are presented that are consistently significant, and so included, in the extended BIC models for the prediction of the seasons 2013(-14) through 2015(-16) of the English, Dutch, Spanish and Swedish competition, respectively. An explanation of these included variables can be found in Appendix F.

Table 5.9: The table contains the variables that are significant in all three extended BIC models (2013-14, 2014-15 and 2015-16) for the English competition. '+' means a positive coefficient, whereas '-' means a negative coefficient in the models for the three considered seasons.

Variables	Sign
$ECIHome_i$ - $ECIAway_i$	+
$HomeManCity_i$	+
$ManCityVS.Newcastle_i$	+

The positive sign for $HomeManCity_i$ in Table 5.9 indicates that, based on preliminary seasons, Manchester City gets a larger probability for a home win than when equally strong teams compete with each other under similar circumstances. Besides, the positive sign for the variable $ManCityVS.Newcastle_i$ indicates that a match between Manchester City and Newcastle United gets larger probability assigned for a win of Manchester City than when other teams, of equal strength, play against each other, ceteris paribus.

Table 5.10: The table contains the variables that are significant in all three extended BIC models (2013-14, 2014-15 and 2015-16) for the Dutch competition. '+' means a positive coefficient, whereas '-' means a negative coefficient in the models for the three considered seasons.

Variables	Sign
$\Delta ECIOnArtificialTurfAteam_i$	-
$ECIHome_i$ - $ECIAway_i$	+
$Spectators_i$	+
$VitesseNETReferee2_i$	-

Similar to the findings of the extended AIC models in Table 5.6 a positive significant coefficient is found for $\Delta ECIOnArtificialTurfAteam_i$ for all three considered Dutch seasons by the extended BIC models, given in Table 5.10. Furthermore, a larger amount of spectators leads to a larger probability for a home victory, according to the positive sign of $Spectators_i$.

Table 5.11: The table contains the variables that are significant in all three extended BIC models (2013-14, 2014-15 and 2015-16) for the Spanish competition. '+' means a positive coefficient, whereas '-' means a negative coefficient in the models for the three considered seasons.

Variables	Sign
$ECIHome_i$ - $ECIAway_i$	+
$GetafeSPAReferee6_i$	-

The constructed BIC models for the seasons 2013-14, 2014-15 and 2015-16 of the Spanish competition show a negative sign for $GetafeSPAReferee6_i$. This means that the results of Getafe CF in a match leaded by Spanish Referee 6 are significantly worse than for another combination of competing clubs and referee, ceteris paribus.

Table 5.12: The table contains the variables that are significant in all three extended BIC models (2013-14, 2014-15 and 2015-16) for the Swedish competition. '+' means a positive coefficient, whereas '-' means a negative coefficient in the models for the three considered seasons.

Variables	Sign
$AIKShape2_i$	+
$ECIHome_i$ - $ECIAway_i$	+

In case the shape of AIK can be considered as Swedish shape 2, the probability of a win for AIK is larger compared to the combination of other equally strong teams classified in Swedish shape group 2. This is concluded from the positive sign of $AIKShape2_i$ in Table 5.12.

5.3 | Model comparison

In order to find out which of the four constructed models most accurately predicts the outcome of soccer matches six different evaluation methods are introduced in Subsect. 4.3.2. These evaluation methods all differ and evaluate the outcome predictions from the models in another way, concentrating on different aspects of the prediction. Mainly, one of two different approaches is followed by each of these comparison methods.

One approach is to take the outcome possibility with the highest assigned probability as the predicted outcome, whereas with the other approach the division of probabilities among all three outcome possibilities is evaluated. At first, the most straightforward evaluation methods are studied. Table 5.13 presents the results from the evaluation method introduced as the *direct outcome evaluation method* in Subsect 4.3.2. This method follows the first approach and takes the outcome possibility with the highest assigned probability as the predicted outcome. Table 5.14 presents the results from the *direct probabilities evaluation method* introduced in Subsect 4.3.2. This method follows the second approach and takes the division of probabilities assigned to all three outcome possibilities into consideration.

Table 5.13: The table gives the fraction of correctly predicted match outcomes per season for each of the four constructed ordered probit models. The score of the best predicting proposed model per competition seasons is underlined and bold in this table. An asterisk (*) is added in case the best predicting proposed models performs better than the bookmakers probabilities.

Model	Competition	2013(-14)	2014(-15)	2015(-16)	Average
Basic ECI model	England	**0.588***	0.533	**0.464**	**0.528***
	Netherlands	0.466	0.556	0.550	0.524
	Spain	**0.528**	**0.570***	**0.525**	**0.541**
	Sweden	0.510	0.473	0.556	0.513
Simplified model	England	0.586	0.525	**0.464**	0.525
	Netherlands	0.462	**0.557***	0.551	0.523
	Spain	0.525	**0.570***	0.520	0.538
	Sweden	**0.515***	0.498	**0.573***	**0.529***
Extended AIC model	England	0.555	**0.537***	0.450	0.514
	Netherlands	**0.490***	0.546	**0.572***	**0.536***
	Spain	0.524	0.555	0.516	0.532
	Sweden	0.483	**0.513***	0.538	0.511
Extended BIC model	England	0.568	0.529	0.458	0.518
	Netherlands	0.464	0.542	0.552	0.519
	Spain	0.524	0.558	0.516	0.533
	Sweden	0.492	0.500	0.550	0.514
Bookmakers probabilities	England	0.558	0.529	0.468	0.518
	Netherlands	0.484	0.510	0.487	0.494
	Spain	0.547	0.568	0.553	0.556
	Sweden	0.492	0.496	0.483	0.490

One can see in Table 5.13 that over the twelve considered seasons, divided over four competitions, the *direct outcome evaluation method* finds varying models as the best performing model. From the table can be concluded that the *direct outcome evaluation method* does not clearly indicate a preferred model. The bottom part of the table contains the results of the evaluation method applied on the bookmakers probabilities.

For only three of the twelve seasons, the *direct outcome evaluation method* prefers the bookmakers probabilities over each four proposed models. The preferred proposed model is underlined and bold in Table 5.13. Per competition the preferred model is compared to the bookmakers odds based on the *direct outcome evaluation method*. In case this proposed model is preferred over the bookmakers probabilities, an asterisk (*) is placed by the specific obtained profit share.

Table 5.14: The table gives the mean of correct assigned outcome probabilities averaged over the amount of matches per season for each of the four constructed models. The profit of the best predicting proposed model per competition seasons is underlined and bold in this table. An asterisk (*) is added in case the best predicting proposed models performs better than the bookmakers probabilities.

Model	Competition	2013(-14)	2014(-15)	2015(-16)	Average
Basic ECI model	England	0.433	0.419	0.387	0.413
	Netherlands	0.405	0.426	0.429	0.420
	Spain	0.437	0.457	0.444	0.446
	Sweden	0.390	0.389	0.399	0.393
Simplified model	England	**0.482***	0.423	0.389	**0.431***
	Netherlands	0.409	0.431	**0.443***	0.428
	Spain	0.438	0.457	0.439	0.445
	Sweden	**0.410***	**0.404**	0.409	**0.408***
Extended AIC model	England	0.441	**0.437***	**0.396***	0.425
	Netherlands	**0.419***	**0.435***	0.440	**0.431***
	Spain	**0.448***	**0.461***	0.441	**0.450***
	Sweden	0.386	0.400	**0.414***	0.400
Extended BIC model	England	0.434	0.424	0.389	0.416
	Netherlands	0.415	0.425	0.430	0.423
	Spain	0.437	0.456	**0.445***	0.446
	Sweden	0.386	0.387	0.400	0.391
Bookmakers probabilities	England	0.429	0.420	0.392	0.414
	Netherlands	0.413	0.421	0.411	0.415
	Spain	0.442	0.460	0.443	0.448
	Sweden	0.397	0.410	0.398	0.402

The extended AIC model can be seen as the preferred model by the *direct probabilities evaluation method* in Table 5.14. From the twelve considered seasons, this evaluation method ranks the extended AIC model as the best of the proposed models for seven seasons. For the remaining five seasons this method is four times ranked as second best. The simplified model can be seen as a reasonable alternative model. According to the *direct probabilities evaluation method* the extended AIC model performs better in predicting the outcome of soccer matches than the bookmakers probabilities. This extended AIC model is preferred over the bookmakers probabilities in nine of twelve considered seasons. The alternative simplified model is preferred over the book-

makers probabilities for six of the twelve considered seasons. Also, the last column shows that the preferred model, averaged over the three considered seasons, performs better than the bookmakers probability for each of the four competitions.

The results from the other four proposed evaluation methods can be found in Appendix E. An overview of these results will be discussed here. Similar to the *direct outcome evaluation method* the *Jaccard index evaluation method* sees the outcome possibility with the highest assigned probability as the predicted outcome. The outcome of the *Jaccard index evaluation method* is given in Table E.1. The other evaluation methods, the *ordered probit penalty index*, the *rank probability score* and the *mean squared error method*, follow the other approach in which the division of probabilities over the three outcome possibilities is evaluated. Their results can be found in Table E.2, Table E.3 and Table E.4, respectively. The results of the *Jaccard index*, the *rank probability score* and the *mean squared error method* are comparable to the results of the *direct outcome evaluation method*; the evaluation method does not clearly prefer one model. For almost all of the twelve considered seasons either the basic ECI, the simplified or the extended AIC model is the preferred model according to each of these three evaluation methods. The extended BIC model is in almost no situation the preferred model. The results of the *ordered probit penalty index*, which can be found in Table E.2, indicate a clear preference for the extended AIC model. For eight of the twelve considered seasons the extended AIC model is preferred over the other three proposed models by the *ordered probit penalty index*. Overall, one can conclude that the preference of a certain model not only depends on the specific season, but also on the evaluation method. Taken all the six evaluation methods into consideration, the extended BIC model is the least preferred model. The *direct probability evaluation method* and the *ordered probit penalty index method* indicate that the extended AIC model is the best prediction model. The other four evaluation methods have no clear preference for one of the following three models; the basic ECI, the simplified and the extended AIC model.

Comparing the proposed models with the bookmakers probabilities using the same evaluation methods leads to varying insights. The *direct outcome evaluation method* and the *direct probability evaluation method*, given in Table 5.13 and Table 5.14, indicate that the preferred proposed model performs better than the bookmakers probabilities for nine and eleven of the twelve considered seasons, respectively. According to the *rank probability score method* and the *mean squared error method* the bookmakers provide better prediction probabilities for match outcomes. From the twelve seasons only three and one season(s) the preferred model, for that specific season, is able to perform better than the bookmakers probabilities. Once more the findings here depend largely on the evaluation method. The specific results from the other four evaluation methods can be found in Appendix E. Considering all evaluation methods one can conclude that none of the proposed models is able to systematically beat the bookmakers probabilities.

Two of the six evaluation methods point out that the extended AIC model beats the bookmakers in more than half of the considered seasons, but this finding is totally different for other evaluation methods. One might conclude that the extended AIC model is best able to compete with the bookmakers probabilities. The basic ECI model and simplified model follow the extended AIC model, whereas the extended BIC model is worst able to compete with the bookmakers probabilities. This suggests that the extra included variables contain important imformation concerning the prediction of soccer match outcomes.

5.4 | Evaluation of betting strategies

It is attempted in this paper to come up with a betting strategy that is able to systematically beat the bookmakers and generate a positive profit with betting on soccer matches. At first a 'naive' betting strategy is considered to check if a certain match outcome (home win, draw, away win) is systematically under- or overpriced for one of the considered competitions. The results of betting based on this strategy over the period of three seasons, 2013(-14) through 2015(-16) for each competition, are given in Table 5.15. One-sided student's t-tests [17] are applied to check whether the average profit share is significantly different than zero using a 5% significance level. These results indicate that betting on one specific possible outcome (home win, draw or away win) for all matches does not systematically generate a significant positive profit share for one of the four considered competitions. In only three out of twelve considered betting situation, given in the table, the profit share is (insignificantly) positive. This betting strategy makes us conclude that none of the possible match outcomes is systematically under- or overpriced for one of the four competitions.

Table 5.15: The table gives the portion of profit per bet euro of the 'naive' betting strategy over a period of three seasons; 2013(-14) through 2015(-16). An asterisk (*) indicates an average profit that is significantly different than zero.

	Home win	Draw	Away win
England	-0.021	-0.080	0.023
Netherlands	-0.066	0.048	-0.125*
Spain	0.000	-0.089*	0.045
Sweden	-0.053*	-0.064*	-0.089*

Next, the predictions resulting from each of the four proposed models are utilized to bet on the matches of the 12 considered soccer seasons. Table 5.16 presents the profit share per bet euro of betting unit stake on the match outcome with the highest assigned probability for each of the four proposed models and the converted bookmakers odds. This table shows that the profit share can vary several percentages per model and not

one of the models consistently outperforms the other models. However, the profit share per model seems to be correlated, suggesting that certain seasons can be seen as harder to predict than other seasons. In a certain way this betting strategy can also be seen as an evaluation method to compare the proposed models. For each considered season the model with the highest profit is underlined. Although the preferred model varies largely over the seasons, the average portion of profit per bet euro over the twelve considered seasons points out a clear preference. Where the basic ECI model, the simplified model and the extended BIC model generate an average profit share of -0.027, -0.038 and -0.038 per bet euro over the twelve considered competitions, respectively, the extended AIC model generates a positive average profit of 0.03 per bet euro. According to this 'evaluation method' the extended AIC model outperforms the other proposed models. Also, the bookmakers probabilities are beaten by the extended AIC model, as the bookmakers probabilities obtain an average profit share of even -0.039 per bet euro. In fact, the average profit share of all four proposed models is higher than that of the bookmakers probability.

Table 5.16: The table gives the portion of profit per bet euro for the seasons 2013-14 through 2015-16 in case a bettor bets on the highest assigned outcome possibility of three models and the bookmakers predictions. The model that provides the highest profit share using this strategy is underlined for each considered season.

	Basic ECI model	Simplified model	Extended AIC model	Extended BIC model	Bookmakers
England					
2013-14	0.063	0.066	0.020	0.036	0.089
2014-15	-0.031	-0.011	0.043	-0.018	-0.035
2015-16	-0.105	-0.109	-0.081	-0.122	-0.111
Netherlands					
2013-14	-0.200	-0.213	-0.073	-0.193	-0.175
2014-15	0.018	0.018	0.014	-0.006	-0.035
2015-16	0.004	0.023	0.108	0.012	-0.002
Spain					
2013-14	-0.055	-0.062	-0.063	-0.064	-0.032
2014-15	0.008	0.001	0.007	-0.021	-0.017
2015-16	-0.054	-0.072	-0.021	-0.075	-0.004
Sweden					
2013	-0.014	-0.026	-0.042	-0.030	-0.095
2014	-0.066	-0.150	0.039	-0.048	-0.109
2015	0.112	0.077	0.085	0.075	0.055

Even though the aforementioned betting strategy, where the outcome possibility with the largest assigned probability is bet on, gives an average seasonal profit share of 0.03 per bet euro for the extended AIC model, only eight of the twelve considered seasons generate a positive portion of profit per bet euro. In Sect. 4.4 four more so-

phisticated betting strategies are introduced. Table 5.3 shows the performance of these strategies applied on the match outcome predictions of the four proposed models. The table gives the average seasonal portion of profit per bet euro over the twelve considered seasons for each combination of betting strategy and proposed model accompanied with the restriction of a minimum assigned probability given by ϕ. Once more, one-sided student's t-tests are used to check whether the average profit share per bet euro is significantly different than zero.

Table 5.17: The table average portion of profit per bet euro of the betting strategies applied on the outcome of the four models. underline means positive profit in each season and model. An asterisk (*) indicates that the average profit share over the twelve considered seasons is significantly different than zero. An underlined average profit indicates that the profit share of this specific is positive for each of the twelve considered seasons.

Equal weights betting strategy

ϕ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Basic ECI model	-0.006	-0.007	-0.008	-0.019	-0.019	0.020	0.030	0.073	0.171*
Simplified model	-0.061*	-0.058*	-0.036*	-0.028	-0.044	0.028	0.033	0.137*	0.168*
Extended AIC model	0.008	0.013	0.019	0.028	0.078*	0.114*	0.101*	0.180*	0.233*
Extended BIC model	-0.035	-0.034	-0.023	-0.032*	-0.039	0.009	0.047	0.090	0.044

Equal payout betting strategy

ϕ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Basic ECI model	-0.003	-0.004	-0.009	-0.011	-0.009	0.026	0.030	0.073	0.171*
Simplified model	-0.041*	-0.040*	-0.036*	-0.028	-0.034	0.034	0.030	0.138*	0.168*
Extended AIC model	0.018	0.020	0.019	0.024	0.066*	0.099*	0.098*	0.166*	0.226*
Extended BIC model	-0.011	-0.011	-0.011	-0.015	-0.015	0.024	0.061*	0.095*	0.047

Variance adjusted betting strategy

ϕ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Basic ECI model	0.018	0.018	0.015	0.018	0.023	0.046	0.046*	0.085	0.173*
Simplified model	-0.014	-0.013	-0.011	-0.002	0.001	0.062*	0.048	0.154*	0.171*
Extended AIC model	0.055*	0.056*	0.055*	0.063*	0.097*	0.127*	0.118*	0.177*	0.226*
Extended BIC model	0.018	0.018	0.017	0.017	0.021	0.043*	0.081*	0.111*	0.045

Kelly's betting strategy

ϕ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Basic ECI model	0.041	0.042	0.040	0.045	0.041	0.061*	0.061*	0.096*	0.205*
Simplified model	0.004	0.007	0.028	0.043	0.033	0.095*	0.079	0.164*	0.169*
Extended AIC model	0.183*	0.174*	0.163*	0.166*	0.178*	0.201*	0.151*	0.229*	0.229*
Extended BIC model	0.013	0.012	0.021	0.027	0.022	0.035	0.064	0.125*	0.026

Table 5.17 indicates that both the variance adjusted betting strategy and Kelly's betting strategy applied on the outcome probabilities assigned by the extended AIC model generate a significant positive profit share per bet euro independent of the treshold value ϕ. Only six combinations of proposed model, proposed betting strategy and minimum value of ϕ generate a significant positive profit share per bet euro among all of the twelve considered seasons. These six betting situations, underlined in Table 5.17, all use the probabilities assigned by the extended AIC model. Four of these six betting situations are generated by a combinations of the extended AIC model with either the variance adjusted or Kelly's betting strategy and a treshold value ϕ. Choosing between the variance adjusted betting strategy and Kelly's strategy accompanied with the extended AIC model based on the results of Table 5.17 indicates clear preference for Kelly's betting strategy with the extended AIC model. The average portion of profit per bet euro over the considered seasons is consistently higher for Kelly's betting strategy than for the variance adjusted betting strategy. However, an average obtained profit share over four competitions is considered in this table. In Figure 5.1 the profit share per bet euro for each competition averaged over the three considered seasons is therefore compared for both the variance adjusted betting strategy as Kelly's betting strategy applied on the extended AIC model for two values of ϕ. The values of ϕ that are considered are 0.4 and 0.7. The value 0.7 is chosen as Table 5.17 shows that four of the six betting situations that generate a positive profit share per bet euro among all twelve considered seasons contain a treshold value of ϕ equal to 0.7. As a second value $\phi = 0.4$ is chosen as the other two of the six situations obtain a value of ϕ equal to 0.4 and 0.5.

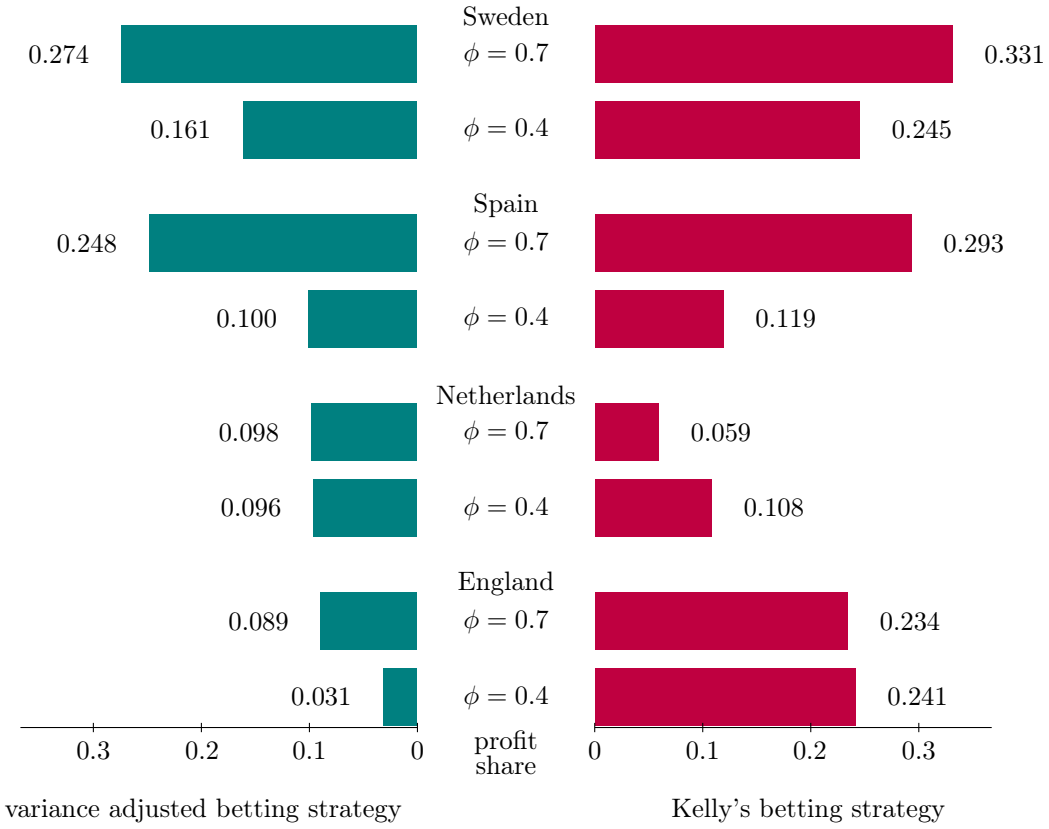


Figure 5.1: This figure shows a comparison of the average profit share per bet euro using the variance adjusted betting strategy and the Kelly's betting strategy accompanied with a minimum assigned probability of ϕ equal to 0.4 and 0.7. This average profit share over three seasons is compared for the four considered competitions.

Figure 5.1 indicates a preference for the combination of Kelly's betting strategy applied on the extended AIC model over the variance adjusted betting strategy applied on the probabilities assigned by the extended AIC model. In case the assigned probabilities are obtained from the extended AIC model, only the Spanish competition with a ϕ-value equal to 0.4 results in a larger profit share for the variance adjusted betting strategy compared to Kelly's betting strategy. The difference between the obtained portion of profit per bet euro performing the variance adjusted betting strategy and Kelly's strategy on the match outcome probabilities retrieved from the extended AIC model is not that large.

Solely the obtained profit for the English competition is much larger in case Kelly's betting strategy is performed compared to the variance adjusted betting strategy. Overall, one can conclude from the figure that Kelly's betting strategy provides a higher portion of profit per bet euro than the variance adjusted betting strategy does for values of ϕ equal to 0.4 and 0.7.

Table 5.18: This table presents the portion of profit per bet euro obtained by performing the Kelly's betting strategy with several values of ϕ for each of the twelve considered seasons. The obtained profit share per bet euro for the optimal value of ϕ is bold.

ϕ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
England									
2013-14	0.103	0.105	0.112	0.121	0.192	0.250	0.260	**0.258**	0.452
2014-15	**0.798**	0.732	0.573	0.493	0.420	0.569	0.392	0.336	0.386
2015-16	**0.450**	0.398	0.273	0.244	0.112	0.109	0.014	0.109	0.039
Netherlands									
2013-14	0.013	0.010	0.002	-0.017	0.019	0.038	**0.073**	0.07	0.002
2014-15	0.071	0.071	0.092	0.103	**0.119**	0.115	0.057	0.026	0.016
2015-16	0.139	0.145	0.158	**0.202**	0.186	0.136	0.115	0.146	0.151
Spain									
2013-14	0.280	0.279	0.268	0.246	0.143	0.157	0.230	0.437	**0.449**
2014-15	0.090	0.089	0.099	0.108	0.080	0.103	0.119	**0.148**	0.101
2015-16	0.080	0.080	0.091	0.077	0.134	0.209	0.222	**0.294**	0.285
Sweden									
2013	-0.060	-0.060	-0.046	-0.021	0.064	**0.152**	0.067	0.141	-0.223
2014	0.054	0.056	0.119	0.148	0.135	0.002	0.022	**0.196**	-0.037
2015	0.178	0.178	0.209	0.286	0.535	0.572	0.471	0.655	**1.125**

Table 5.18 shows the seasonal profit share per bet euro of Kelly's betting strategy applied on the outcome probabilities of the extended AIC model for the four considered competition with different values of ϕ. From these results one can conclude that performing Kelly's betting strategy on the assigned probabilities of the extended AIC models with optimal values of ϕ lead to an positive average seasonal share of profit per bet euro. This average seasonal portion of profit per bet euro for England, Netherlands, Spain and Sweden equals 0.449, 0.131, 0.297 and 0.491, respectively. One can see from the table that the optimal value of ϕ varies largely per season and competition. Although the optimal treshold ϕ-value of the preliminary season is most of the time not optimal for the successive season, it provides a significant positive portion of profit per bet euro. Figure 5.2 shows the obtained share of profit per bet euro for the seasons 2014(-15) and 2015(-16) for the four considered competitions. The figure shows the profit share obtained for the optimal value of ϕ and the obtained profit share in case the value of ϕ is estimated out-of-sample. The optimal value of ϕ from the preliminary seasons is used in this case. One can see that for only two of the eight considered sea-

sons the out-of-sample profit share equals the profit share obtained using the optimal ϕ. Table 5.18 shows that for these two cases the out-of-sample value of ϕ is the optimal value of ϕ. The average seasonal profit share of this method for England, Netherlands, Spain an Sweden equals 0.393, 0.122, 0.198 and 0.329, respectively, over the seasons 2014-(15) and 2015(-16). An overall average seasonal out-of-sample profit share per bet euro of 0.260 over the eight considered seasons for four different competitions is generated with this strategy.

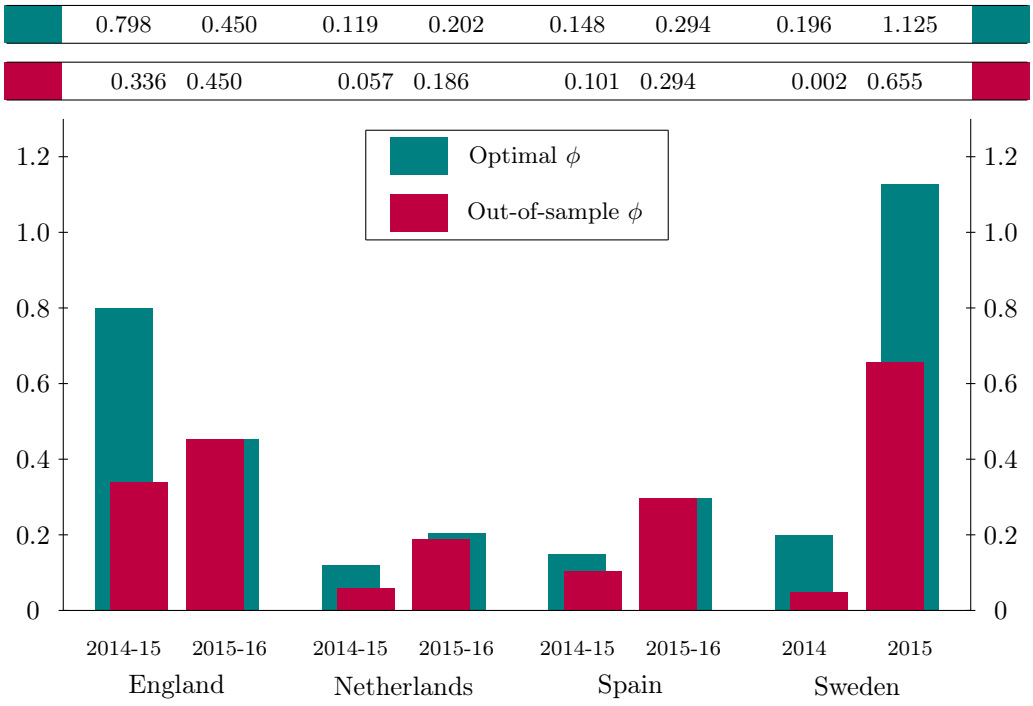


Figure 5.2: This figure visualizes the difference in obtained portion of profit per bet euro applying the variance adjusted betting strategy with the optimal value of ϕ and the out-of-sample value of ϕ. This is performed for the seasons 2014(-15) and 2015(-16) of the English, Dutch, Spanish and Swedish competition.

Overall, one can conclude that applying Kelly's betting strategy on the match outcome probabilities retrieved from the extended AIC model systematically provides a positive share of profit per bet euro. This positive profit is obtained for both optimal values as out-of-sample values of ϕ. Table 5.18 and Fig. 5.2 show that the obtained profit share per bet euro varies largely per season and competition, both for optimal values of ϕ and out-of-sample values of ϕ.

5.5 | Predictability of the seasons

Previous sections elaborate upon the results of several prediction models and betting strategies performed on the matches of three seasons from four European competitions. Kelly's betting strategy applied on the outcome predictions obtained from the extended AIC models turned out to perform best and delivers the highest profit share per bet euro. As this method is applied on the twelve different seasons, the usability of the method, expressed in obtained profit share per bet euro, can be studied. Among these twelve considered seasons, there are seasons in which a lot of outcome matches were unforeseen. For other seasons the majority of matches ended in the expected outcome. A link between obtained profit and the predictability of a specific season is studied. Classification of seasons based on their predictability is not straightforward. Three different measures are applied to indicate the predictability of a certain season.

At first, a simulation algorithm is applied in which a competition season is simulated 10,000 times based on the ECI value of the clubs at the start of the season. Fig. 5.3 shows the simulated probability for the real-life league winner to become the champion according to Alg. 1 in the seasons 2013-(14) through 2015(-16) for each of the four competitions. The dotted line presents the average amount of simulated probability assigned to the league winner for the seasons 2007(-08) through 2015(-16). The figure suggests that the champion of the Spanish league is better predictable than the Swedish league winner, based on simulations, for the seasons 2007(-08) through 2015(-16). The simulation algorithm suggests that season 2015 of the Swedish competition, season 2013-14 of the Spanish competition and season 2015-16 of the English competition were relatively hard to predict. On the other hand, season 2014 of the Swedish competition and season 2013-14 of the Dutch competition were relatively easy to predict according to this measure. Table 5.18 and Fig. 5.2 find a higher portion of profit per bet euro for the seasons that contain high prediction correctness than for the seasons that contain small prediction correctness. This suggests that hard predictable seasons lead to higher profit share. However, this link is not present for all considered seasons. Furthermore, no clear link between season predictability and the optimal value of ϕ can be found.

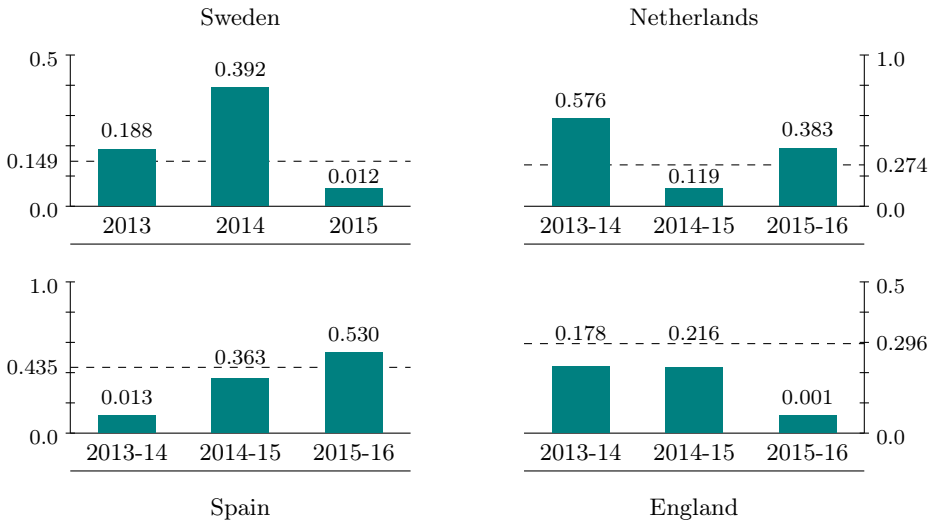


Figure 5.3: This figure presents the simulated probability of the league winner for the seasons 2013(-14) through 2015(-16) for the four considered seasons. The dotted line presents the average amount of simulated probability assigned to the champion of the seasons 2007(-08) through 2015(-16).

Taking this simulation algorithm as an indicator for the predictability of a certain season contains a drawback. The simulation algorithm uses the ECI value of the clubs at the start of the season as the only input. In case the real-life league winner is a club that performed disappointing in the foregoing season, the simulation will predict bad. An unexpected league winner does not necessarily imply that the total gradient of the competition has to be unpredictable.

Another measure that is used as an indicator measure for season predictability is the cosine similarity between the vectors of the resulting season tables of two consecutive seasons. If the resulting season table varies a lot with the resulting season table of the preliminary season this measure finds low similarity, whereas a resulting season table that is relatively similar to the preliminary season will get high similarity. Table 5.19 shows for each of the twelve considered seasons the similarity of the resulting season table compared to the resulting season table of the preliminary season using the cosine similarity measure. Besides, the average cosine similarity for the seasons 2000(-01) through 2015(-16) is given for each of the four competitions. The table suggests that the resulting season table for season 2013-14 of the Spanish competition and season 2015-16 of the Dutch competition had an relatively unexpected outcome, whereas the resulting season table for season 2014-15 of the English, seasons 2013-14 and 2014-15

of the Dutch competition, season 2014-15 of the Spanish competition and season 2013 of the Swedish competition were relatively easy to predict. No clear correlation can be found between the similarity of the resulting season tables with that of the preliminary season and the obtained profit share.

Table 5.19: This table gives the similarity of the considered season with the preliminary season using cosine similarity on the vectors of the resulting season tables. The last column contains the average value of the cosine similarity measure applied on the seasons 2000(-01) through 2015(-16) for each of the four considered competitions.

Model	2013(-14)	2014(-15)	2015(-16)	2000(-01) - 2015(-16)
England	0.921	0.973	0.932	0.928
Netherlands	0.970	0.980	0.931	0.946
Spain	0.895	0.978	0.938	0.912
Sweden	0.936	0.915	0.902	0.901

A drawback of this method is that it only compares the season tables at the end of the season. Besides, this method solely compares the similarity of the current season with the preliminary season. A low similarity between these two season vectors only suggests an unexpected resulting season table compared to the resulting season table of the preliminary season.

As a third indicator measure the obtained profit from the betting strategies, given in Table 5.16, is studied. This table shows the profit of betting unit stack on the outcome possibility with the highest assigned probability for each of the four proposed models and the bookmakers probabilities. Table 5.20 presents the average profit share per bet euro of these four proposed models and the bookmakers probabilities for each of the twelve considered seasons. This betting strategy bets per match on the outcome possibility with the highest assigned probability according to the model at hand. This indication measure evaluates the proportion of matches that have an expected outcome according to the models. The measure suggests that season 2015-16 of the English competition and season 2013-14 of the Dutch competition were the seasons with the most unpredictable match outcomes. To a lesser extent season 2013-14 and 2015-16 from the Spanish competition and season 2013 and 2014 from the Swedish competition can be classified as unpredictable seasons. On the other hand, season 2013-14 of the English competition and season 2015 of the Swedish competition are seasons containing a lot of matches with an expected outcome. Also for this measure of seasons predictability Table 5.18 and Fig. 5.2 find no clear link between the average profit of betting on the outcome with the highest assigned probability, given in Table 5.20, and the obtained profit share.

Table 5.20: This table gives the average profit share per bet euro of the four proposed models and the bookmakers odds betting unit stack on the outcome possibility with the highest assigned probability.

Model	England	Netherlands	Spain	Sweden
2013(-14)	0.055	-0.171	-0.055	-0.041
2014(-15)	-0.010	0.002	-0.004	-0.067
2015(-16)	-0.106	0.029	-0.045	0.081

A drawback of this third indication measure is that it depends on the proposed models. The results of this method rely largely on the assigned bookmakers odds and the accuracy of the four proposed models.

Although each of the three measures consists of mentioned shortcomings, all these measures focus on another aspect of season predictability. The results of the three measures present that these indicator measures are not consistent in classifying seasons as predictable or unpredictable. For only one of the measures a relation between the predictability and the obtained profit share is discovered. The seasons for which the simulation algorithm assigned low probability to the real-life league winner generated a larger profit share per bet euro than seasons where the simulation algorithm assigned high probability to the real-life league winner. However, this link is not applicable for all twelve considered seasons. This makes us conclude that Kelly's betting strategy applied on outcome probabilities assigned by the extended AIC models can be seen as a robust betting strategy. Both seasons in which the majority of matches ends as expected and seasons in which a lot of unexpected match outcomes occur generate a positive share of profit per bet euro. The size of the profit share and the value of the optimal value of ϕ are neither correlated with the predictability of the season indicated by one of the three considered indication measures.

6 | Conclusion

One of the main goals of this paper is to come up with prediction models that accurately predict the outcome of soccer matches. In addition, it is attempted to construct a betting strategy that is able to defeat the bookmakers and generate profit from betting on soccer matches.

For this research, the focus is on European national competitions. PCA accompanied with a K-means clustering algorithm is applied on a dataset of competition characteristics to select a number of European leagues to study in more detail. This method led to selection of the Dutch, English, Spanish and Swedish competition. For each of these four selected competitions various ordered probit forecasting models are constructed to predict a probability of occurrence for all three outcome possibilities. Two of the four prediction models solely depend on the strength of the competing clubs, expressed in ECI value, and the division of home wins, draws and away wins in recent seasons. Two other prediction models incorporate a widely range of club-specific explanatory variables alongside the ECI value for clubs. The extended AIC model uses the Akaike information criterion to decide on the inclusion of variables. The extended BIC model uses Bayesian information criterion. The performance of the proposed models varies widely over six evaluation measures. Overall, the extended AIC model is the preferred model for most competitions and seasons, whereas the extended BIC model is clearly the least preferred. According to some of the evaluation measures the prediction accuracy of the proposed models outperforms the bookmakers probabilities, whereas the converted bookmakers odds predict more accurately than each of the proposed models according to other evaluation methods.

Consequently, several betting strategies are implemented for which the predicted probability of the possible soccer match outcomes assigned by the prediction models is the input. A minimum assigned probability restriction is imposed to four existing betting strategies. Applying Kelly's betting strategy on the predicted outcome probabilities obtained from the extended AIC model provides the overall best performing betting strategy. This betting strategies generates an average out-of-sample profit share per bet euro of 0.26 over the four considered competitions in 2014(-15) and 2015(-16). It generates a positive profit share per bet euro for all considered seasons.

Eventually, it is investigated whether the obtained profit is related to the predictability of a certain season. A season simulation algorithm, resulting season table vector comparison of consecutive seasons and the obtained profit from betting on the match outcomes with the highest assigned probability are three measures to indicate a seasons' predictability. Overall, the indicated predictability and the obtained profit shares obey a clear link among the considered seasons. This indicates that the proposed strat-

egy is a robust betting strategy that systematically generates profit independent of the predictability of the season.

We can conclude that the extended AIC model is the most accurate model in predicting soccer match outcomes. It is best able to compete with the bookmakers. The preference for this model shows that the included soccer-related variables contain significant information to predict match outcomes. Applying the betting strategy of Kelly on the results of this model provides an average profit share of 0.26 per bet euro.

Future research on this content should investigate if inclusion of player-specific variables into the prediction models is worth it. Injuries and suspensions might have influence on the outcome probabilities and inclusion of these factors might improve the constructed models. For this research the used betting odds are the average match odds assigned by several bookmakers. As the match odds vary per bookmaker one could attempt to construct a betting algorithm that benefits from these discrepancies in odds by different bookmakers. As this paper finds a profitable betting strategy using average odds an even higher retrieved profit could be found in case the most favorable assigned odds per match can be chosen from several bookmakers.

Bibliography

[1] Abdi, H., Williams, L.J.: Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics 2(4), 433–459 (2010)

[2] Blundell, J.: Numerical Algorithms for Predicting Sports Results. Ph.D. thesis, University of Leeds, School of Computer Studies (2009)

[3] Bozdogan, H.: Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. Psychometrika 52(3), 345–370 (1987)

[4] Burnham, K.P., Anderson, D.R.: Multimodel inference understanding aic and bic in model selection. Sociological methods & research 33(2), 261–304 (2004)

[5] Clarke, S.R., Norman, J.M.: Home ground advantage of individual clubs in english soccer. The Statistician pp. 509–521 (1995)

[6] Dixon, M.J., Coles, S.G.: 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)

[7] Dobson, S., Goddard, J.A., Dobson, S.: The economics of football. Cambridge University Press Cambridge (2001)

[8] Elo, A.: The rating of chess players, past and present (arco, new york) (1978)

[9] Epstein, E.S.: A scoring system for probability forecasts of ranked categories. Journal of Applied Meteorology 8(6), 985–987 (1969)

[10] Finnigan, M., Nordsted, P.: The premier football betting handbook 2010/11. Hampshire, Great Britain: Harriman House (2010)

[11] Frey, J.H.: Gambling: A sociological review. The Annals of the American Academy of Political and Social Science 474(1), 107–121 (1984)

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

[13] Goldfeld, S.M., Quandt, R.E., Trotter, H.F.: Maximization by quadratic hill-climbing. Econometrica: Journal of the Econometric Society pp. 541–551 (1966)

[14] Graham, I., Stott, H.: Predicting bookmaker odds and efficiency for uk football. Applied Economics 40(1), 99–109 (2008)

[15] Greene, W.H.: Econometric analysis (international edition) (2000)

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

[17] Heij, C., De Boer, P., Franses, P.H., Kloek, T., Van Dijk, H.K., et al.: Econometric methods with applications in business and economics. OUP Oxford (2004)

[18] Hing, N.: Sports betting and advertising. Australian Gambling Research Centre, Australian Institute of Family Studies (2014)

[19] Jaccard, P.: The distribution of the flora in the alpine zone. New phytologist 11(2), 37–50 (1912)

[20] Kelly, J.: A new interpretation of information rate. IRE Transactions on Information Theory 2(3), 185–189 (1956)

[21] King, B., Reynolds, J., Botta, C., Wardle, S.a.: Dynamic modelling and prediction of english football league matches for betting. Street and Smith's SportsBusiness Journal (www.sportsbusinessjournal.com) 17(ISSUE B) (2014, June)

[22] Koning, R.H.: Balance in competition in dutch soccer. Journal of the Royal Statistical Society: Series D (The Statistician) 49(3), 419–431 (2000)

[23] Langseth, H.: Beating the bookie: A look at statistical models for prediction of football matches. In: SCAI. pp. 165–174 (2013)

[24] Liang, Y., Balcan, M.F., Kanchanapally, V.: Distributed pca and k-means clustering. In: The Big Learning Workshop at NIPS. vol. 2013. Citeseer (2013)

[25] Lloyd, S.P.: Least squares quantization in pcm. Information Theory, IEEE Transactions on 28(2), 129–137 (1982)

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

[27] Markazi, A.: Why leicester city become biggest long shot champion sports history. In: http://espn.go.com/espn/feature/story/_/id/14759409/why-leicester-city-become-biggest-long-shot-champion-sports-history (2016)

[28] Nevill, A.M., Balmer, N.J., Williams, A.M.: The influence of crowd noise and experience upon refereeing decisions in football. Psychology of Sport and Exercise 3(4), 261–272 (2002)

[29] Norris, D.: Football's beauty is its unpredictability. In: %http://www.irishexaminer.com/sport/soccer/footballs-beauty-is-its-unpredictability-says-arsene-wenger-321805.html (2015)

[30] Peason, K.: On lines and planes of closest fit to systems of point in space. Philosophical Magazine 2, 559–572 (1901)

[31] Pohlman, J.T., Leitner, D.W.: A comparison of ordinary least squares and logistic regression (2003)

[32] Pollard, R., Pollard, G.: Home advantage in soccer: A review of its existence and causes. International Journal of Soccer and Science Journal 3(1), 31–44 (2005)

[33] Prabhu, P., Anbazhagan, N.: Improving the performance of k-means clustering for high dimensional data set. International journal on computer science and engineering 3(6), 2317–2322 (2011)

[34] Rue, H., Salvesen, O.: Prediction and retrospective analysis of soccer matches in a league. Journal of the Royal Statistical Society: Series D (The Statistician) 49(3), 399–418 (2000)

[35] Sauer, R.D.: The economics of wagering markets. Journal of economic Literature 36(4), 2021–2064 (1998)

[36] Smith, R.W., Preston, F.W.: Vocabularies of motives for gambling behavior. Sociological Perspectives 27(3), 325–348 (1984)

[37] Snyder, J.A.L.: What actually wins soccer matches: Prediction of the 2011-2012 premier league for fun and profit (2013)

[38] Spann, M., Skiera, B.: Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters. Journal of Forecasting 28(1), 55–72 (2009)

[39] Štrumbelj, E., Šikonja, M.R.: Online bookmakers' odds as forecasts: The case of european soccer leagues. International Journal of Forecasting 26(3), 482–488 (2010)

[40] Vergin, R.C., Sosik, J.J.: No place like home: an examination of the home field advantage in gambling strategies in nfl football. Journal of Economics and Business 51(1), 21–31 (1999)

[41] Wilkinson, L., Engelman, L., Corter, J., Mark, C.: Systat 10 statistics i. 4. Cluster Analysis pp. I–77 (2000)

[42] Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemometrics and intelligent laboratory systems 2(1-3), 37–52 (1987)

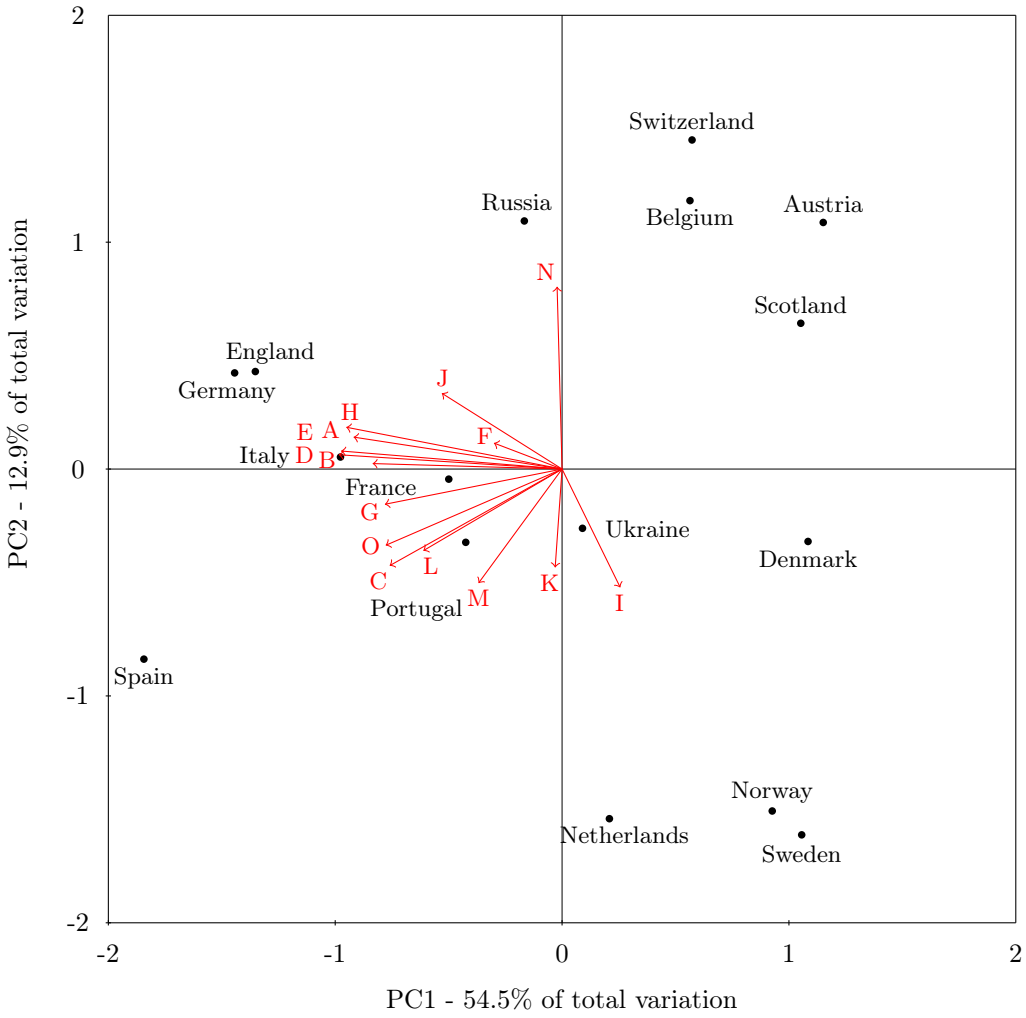
[43] Yezus, A.: Predicting outcome of soccer matches using machine learning. Ph.D. thesis, Saint-Petersburg State University, Mathematics and Mechanics Faculty (2014)

Appendices

A | Output PCA

Table A.1: The table gives the values of the principal component for each competition.

Country	PC1	PC2	PC3	PC4	PC5
Netherlands	-0.616	-2.211	-0.852	-0.885	1.451
Belgium	-1.665	1.698	0.606	0.519	0.874
England	3.993	0.617	1.220	0.309	-0.272
Italy	2.885	0.076	-0.452	0.966	-0.854
France	1.475	-0.064	1.316	0.890	0.951
Scotland	-3.108	1.042	-1.765	1.971	0.655
Spain	4.263	0.608	1.139	-0.866	1.120
Germany	5.444	-1.202	-0.804	-0.256	-0.204
Portugal	1.254	-0.463	-1.574	-0.348	0.0817
Denmark	-3.202	-0.458	-0.059	-0.114	-0.457
Norway	-2.736	-2.163	1.012	-0.380	-1.432
Ukraine	-0.266	-0.375	-2.647	-0.164	-0.350
Austria	-3.400	1.560	0.535	-0.880	1.032
Russia	0.491	1.569	0.601	0.785	-1.637
Sweden	-3.120	-2.313	1.887	0.590	0.008
Switzerland	-1.692	2.081	-0.164	-2.147	-0.966



A| Average stadium capacity
B| Average number of spectators
C| Competition size
D| UEFA competition points rank
E| Number of CL tickets
F| Number of teams in Europe at 1 Dec.
G| Number of teams in Europe at 1 Apr.
H| Average ECI value
I| Different champions last 10 seasons
J| Average age of players
K| Average amount of goals per match
L| Number of relegated teams
M| Perc. of possible points champion
N| Δ points first and last club
O| Δ ECI value first and last club

Figure A.1: This figure shows the biplot of the first two principal components using scaled data. The 16 observations are visualized by dots. The variables are given by vectors. In the legend is specified which variable belongs to each of three vectors. The remaining 13 variables are not visualized as the length of these vectors was insignificant compared to the three given in the figure.

B | Output cluster method

Table B.1: The table gives the coefficients of the cluster means per competition characteristic.

Variable	Cluster 1	Cluster 2
Average stadium capacity	36250.14	18015.06
Average number of spectators	24516.02	9618.38
League size	18.86	13.89
UEFA competition points rank	70658.71	28132.78
CL tickets	3.29	1.44
European active teams 1 dec	4.86	1.11
European active teams 1 apr	2	0.11
Average value of ECI	2504.20	1786.03
Different winners last 10 seasons	3.57	4.11
Average age of players	27.68	26.66
Average goals per match	2.62	2.69
Number of relegation per year	2.5	1.5
% of points from the winner	0.78	0.76
Point difference first and last	0.31	0.28
Difference ECI value first and last	1855.80	1266.03

C | Extended AIC models

Table C.1: The table gives the extended ordered probit model based on AIC for England in the season 2013-14.

Variables	Coefficient	Std. Error
$5MatchesShapeHteam - 3_i$	0.192	(0.090)
$5MatchesShapeHteam + 3_i$	0.293	(0.091)
$ArsenalENGReferee1_i$	-1.594	(0.602)
$AverageHomeAgainstHteam_i$	-0.200	(0.075)
$AwaymatchesAteam_i$	0.102	(0.037)
$AwaymatchesHteam_i$	-0.084	(0.036)
$CapacityHteam_i$	8.09E-6	(2.56E-6)
$CapacityAteam_i$	-5.50E-6	(2.46E-6)
$ChelseaApril_i$	0.691	(0.271)
$ECIHome_i$ - $ECIAway_i$	0.50E-3	(5.20E-5)
$EUAwayDrawPostWinter_i$	0.680	(0.337)
$EUAwayWinPostWinterAteam_i$	-0.783	(0.379)
$EvertonShape - 1_i$	0.591	(0.207)
$EvertonVS.Arsenal_i$	-0.967	(0.387)
$HomeManCity_i$	0.432	(0.136)
$LiverpoolENGReferee2_i$	1.098	(0.515)
$ManCityAugust_i$	0.752	(0.327)
$ManCityENGReferee4_i$	-1.006	(0.363)
$ManCityVS.Newcastle_i$	1.592	(0.602)
$ManUnitedShape - 3_i$	0.626	(0.295)
$PercHomepointsAteam_i$	0.430	(0.203)
$PercPointsAteam_i$	-0.595	(0.247)
$SouthamptonShape1_i$	1.377	(0.491)
$Shape + 1VS.Shape + 1_i$	0.449	(0.169)
$Shape + 3VS.Shape + 3_i$	0.346	(0.157)
$StokeENGReferee6_i$	0.768	(0.285)
$SunderlandVS.WBA_i$	-1.262	(0.483)
$TottenhamFebruary_i$	0.618	(0.295)
$VillaMarch_i$	-0.774	(0.248)
μ_1	-0.833	(0.148)
μ_2	0.004	(0.147)

Table C.2: The table gives the extended ordered probit model based on AIC for England in the season 2014-15.

Variables	Coefficient	Std. Error
$3MatchesShapeHteam-2_i$	-0.169	(0.081)
$5MatchesShapeHteam-3_i$	0.209	(0.083)
$5MatchesShapeHteam+3_i$	0.210	(0.088)
$ArsenalENGReferee1_i$	-0.932	(0.444)
$AwaymatchesHteam_i$	-0.111	(0.049)
$ChelseaNovember_i$	-0.526	(0.226)
$CupmatchHteam_i$	0.172	(0.082)
$ECIHome_i$ - $ECIAway_i$	0.64E-3	(4.16E-5)
$EUAwayDrawPostWinter_i$	0.979	(0.317)
$EUPostWinterHteam_i$	0.169	(0.086)
$EUIn6WinPostWinterHteam_i$	0.664	(0.275)
$EvertonShape-1_i$	0.562	(0.186)
$HomeVilla_i$	-0.252	(0.104)
$HomeManCity_i$	0.494	(0.125)
$HomeStoke_i$	0.307	(0.113)
$LiverpoolMarch_i$	0.493	(0.247)
$LiverpoolENGReferee2_i$	1.123	(0.535)
$LiverpoolENGReferee3_i$	-1.458	(0.741)
$ManCityENGReferee4_i$	-0.936	(0.363)
$ManCityVS.Newcastle_i$	1.541	(0.587)
$ManUnitedShape-3_i$	0.519	(0.242)
$MatchesLeftAteam_i$	-0.050	(0.025)
$NewcastleENGReferee5_i$	1.134	(0.539)
$NewcastleVS.Swansea_i$	-1.482	(0.646)
$PercPointsAteam_i$	-0.844	(0.257)
$RankAteam_i$	-0.019	(0.007)
$Shape+1VS.Shape+1_i$	0.460	(0.153)
$Shape+2VS.Shape+2_i$	0.343	(0.152)
$SouthamptonSaturday_i$	0.478	(0.170)
$SunderlandJanuary_i$	0.635	(0.247)
$SwanseaENGReferee7_i$	1.114	(0.463)
$SwanseaENGReferee8_i$	-2.053	(0.814)
$SwanseaShape-1_i$	0.693	(0.303)
$TottenhamENGReferee9_i$	-0.637	(0.081)
$TottenhamVS.Swansea_i$	1.289	(0.647)
$WBAENGReferee9_i$	-0.847	(0.386)
$WBASeptember_i$	0.666	(0.291)
$WestHamENGReferee5_i$	1.398	(0.573)
μ_1	-2.827	(0.874)
μ_2	-2.014	(0.874)

Table C.3: The table gives the extended ordered probit model based on AIC for England in the season 2015-16.

Variables	Coefficient	Std. Error
$3MatchesShapeHteam-2_i$	-0.228	(0.075)
$5MatchesShapeAteam-3_i$	-0.177	(0.076)
$5MatchesShapeHteam-3_i$	0.190	(0.078)
$5MatchesShapeHteam+3_i$	0.240	(0.080)
$AverageHomeAgainstHteam_i$	-0.187	(0.069)
$AverageGoalsAwayHteam_i$	0.272	(0.094)
$ArsenalVS.Swansea_i$	-0.849	(0.421)
$AwaymatchesHteam_i$	-0.099	(0.046)
$CapacityAteam_i$	-4.78E-6	(2.09E-6)
$ChelseaApril_i$	0.583	(0.239)
$ECIHome_i$ - $ECIAway_i$	0.58E-3	(4.38E-5)
$EvertonShape-1_i$	0.570	(0.182)
$EUAwayDrawPostWinter_i$	0.829	(0.317)
$EUHomeIn6WinPreWinter_i$	-0.240	(0.115)
$HomeManCity_i$	0.468	(0.119)
$HomeStoke_i$	0.281	(0.108)
$HomepointsHteam_i$	-0.012	(0.006)
$LiverpoolMarch_i$	0.491	(0.229)
$ManCityENGReferee4_i$	-1.000	(0.362)
$ManCityVS.Newcastle_i$	1.641	(0.571)
$ManUnitedShape-3_i$	0.461	(0.218)
$MatchesLeftAteam_i$	-0.062	(0.024)
$NewcastleVS.Swansea_i$	-1.220	(0.500)
$PercAwaypointsHteam_i$	-0.595	(0.250)
$Shape+1VS.Shape+1_i$	0.365	(0.147)
$SouthamptonJanuary_i$	0.777	(0.342)
$SouthamptonSaturday_i$	0.395	(0.144)
$Spectators_i$	5.92E-6	(2.15E-6)
$StokeENGReferee6_i$	0.554	(0.254)
$SwanseaAugust_i$	0.801	(0.403)
$SwanseaENGReferee8_i$	-1.754	(0.723)
$SwanseaShape-1_i$	0.649	(0.249)
$TottenhamENGReferee9_i$	-0.631	(-0.301)
$TottenhamVS.Swansea_i$	1.508	(0.623)
$WBAENGReferee9_i$	-0.929	(0.379)
$WBASeptember_i$	0.804	(0.268)
μ_1	-2.813	(0.834)
μ_2	-2.013	(0.833)

Table C.4: The table gives the extended ordered probit model based on AIC for Netherlands in the season 2013-14.

Variables	Coefficient	Std. Error
$\Delta ECIOnArtificialTurfAteam_i$	-0.010	(0.003)
$AjaxMarch_i$	0.832	(0.336)
$AjaxVS.Utrecht_i$	-1.074	(0.342)
$AverageGoalsAgainstAteam_i$	0.177	(0.065)
$AwaymatchesAteam_i$	-0.019	(0.009)
$AwaypointsHteam_i$	0.018	(0.006)
$AZNovember_i$	0.690	(0.271)
$AZShape-2_i$	-0.645	(0.242)
$ECIHome_i$ - $ECIAway_i$	0.73E-3	(7.41E-5)
$EULossAwayPostWinter_i$	-1.057	(0.318)
$FeyenoordJanuary_i$	-0.807	(0.309)
$FeyenoordShape1_i$	0.457	(0.227)
$GroningenNETReferee2_i$	0.829	(0.354)
$GroningenShape-2_i$	-0.558	(0.210)
$NACNETReferee5_i$	-1.226	(0.618)
$PSVBefore2PM_i$	-0.752	(0.299)
$RKCReferee6_i$	-1.459	(0.733)
$Shape-3VS.Shape+1_i$	0.649	(0.217)
$Shape+1VS.Shape+1_i$	0.541	(0.183)
$Shape+1VS.Shape+3_i$	-0.568	(0.199)
$Spectators_i$	8.05E-6	(3.26E-6)
$TwenteFebruary_i$	-0.542	(0.245)
$UtrechtBetween6PMAnd8PM_i$	-0.449	(0.222)
$VitesseNETReferee2_i$	-1.276	(0.342)
μ_1	-0.212	(0.134)
μ_2	0.523	(0.134)

Table C.5: The table gives the extended ordered probit model based on AIC for Netherlands in the season 2014-15.

Variables	Coefficient	Std. Error
$\Delta ECIOnArtificialTurfAteam_i$	-0.011	(0.002)
$AjaxVS.Utrecht_i$	-0.855	(0.313)
$AverageGoalsAgainstAteam_i$	0.151	(0.060)
$AwaymatchesAteam_i$	-0.020	(0.008)
$AwaypointsHteam_i$	0.017	(0.006)
$AZNovember_i$	0.595	(0.246)
$CapacityAteam_i$	-7.88E-6	(2.66E-6)
$ECIHome_i$ - $ECIAway_i$	0.53E-3	(8.09E-5)
$EUAwayLossPostWinterHteam_i$	-0.833	(0.347)
$FeyenoordJanuary_i$	-0.775	(0.278)
$GroningenNETReferee2_i$	0.821	n (0.349)
$GroningenShape-2_i$	-0.499	(0.194)
$PSVBefore2PM_i$	-0.808	(0.288)
$Shape-3VS.Shape+1_i$	0.435	(0.186)
$Shape+1VS.Shape+1_i$	0.460	(0.170)
$Shape+1VS.Shape+3_i$	-0.494	(0.178)
$Spectators_i$	1.30E-5	(3.22E-6)
$UtrechtBetween6PMAnd8PM_i$	-0.461	(0.199)
$UtrechtNETReferee3_i$	0.723	(0.353)
$VitesseMay_i$	-0.977	(0.415)
$VitesseNETReferee2_i$	-1.017	(0.282)
$WillemIINETReferee7_i$	-0.882	0.385)
μ_1	-0.416	(0.141)
μ_2	0.328	(0.141)

Table C.6: The table gives the extended ordered probit model based on AIC for Netherlands in the season 2015-16.

Variables	Coefficient	Std. Error
$\Delta ECIOnArtificialTurfAteam_i$	-0.013	(0.002)
$AjaxMarch_i$	0.590	(0.262)
$AjaxVS.Utrecht_i$	-0.728	(0.293)
$AverageGoalsAgainstAteam_i$	0.128	(0.057)
$AZNovember_i$	0.575	(0.227)
$ECIHome_i$ - $ECIAway_i$	0.70E-3	(6.15E-5)
$EUAwayLossPostWinterHteam_i$	-0.791	(0.337)
$FeyenoordJanuary_i$	-0.514	(0.257)
$FeyenoordNETReferee4_i$	0.959	(0.456)
$GraafschapShape1_i$	-0.509	(0.241)
$GroningenNETReferee1_i$	0.870	(0.398)
$GroningenNETReferee2_i$	-0.587	(0.263)
$GroningenShape-2_i$	-0.419	(0.174)
$PSVBefore2PM_i$	-0.669	(0.263)
$Shape-3VS.Shape+1_i$	0.460	(0.174)
$Shape-2VS.Shape+1_i$	-0.331	(0.140)
$Shape-1VS.Shape+2_i$	-0.365	(0.167)
$Shape+1VS.Shape+1_i$	0.429	(0.159)
$Shape+1VS.Shape+3_i$	-0.475	(0.165)
$Shape+2VS.Shape-3_i$	0.357	(0.175)
$Spectators_i$	1.05E-5	(2.69E-6)
$TwenteFebruary_i$	-0.553	(0.200)
$UtrechtBetween6PMAnd8PM_i$	-0.399	(0.190)
$UtrechtNETReferee3_i$	0.814	(0.320)
$VitesseNETReferee2_i$	-0.732	(0.243)
$WillemIINETReferee7_i$	-0.903	(0.372)
μ_1	-0.255	(0.117)
μ_2	0.486	(0.117)

Table C.7: The table gives the extended ordered probit model based on AIC for Spain in the season 2013-14.

Variables	Coefficient	Std. Error
$BilbaoSPAReferee1_i$	-1.018	(0.431)
$CapacityAteam_i$	-6.50E-6	(1.74E-6)
$CapacityHteam_i$	7.57E-6	(1.86E-6)
$ECIHome_i$ - $ECIAway_i$	0.45E-3	(6.30E-5)
$EspanyolShape-3_i$	-0.611	(0.199)
$EUAwayWinAteam_i$	-0.519	(0.197)
$EUIn6PreWinter_i$	0.260	(0.090)
$EULossPreWinterAteam_i$	-0.870	(0.282)
$GetafeSPAReferee5_i$	0.877	(0.419)
$GetafeSPAReferee6_i$	-0.885	(0.325)
$HomeBetis_i$	-0.360	(0.141)
$LevanteSPAReferee7_i$	0.916	(0.412)
$RankAteam_i$	0.012	(0.006)
$RealMadridWeek_i$	0.808	(0.366)
$SevillaSPAReferee4_i$	1.085	(0.383)
$SevillaSPAReferee11_i$	1.222	(0.586)
$Shape+1VS.Shape+1_i$	0.469	(0.164)
$Shape+2VS.Shape-1_i$	0.431	(0.174)
$SociedadSPAReferee10_i$	1.580	(0.721)
$ValenciaMarch_i$	-0.651	(0.227)
μ_1	-0.486	(0.128)
μ_2	0.215	(0.127)

Table C.8: The table gives the extended ordered probit model based on AIC for Spain in the season 2014-15.

Variables	Coefficient	Std. Error
$\Delta ECI3MatchesAteam_i$	0.001	(0.001)
$\Delta ECI5MatchesAteam_i$	-0.001	(0.001)
$AlmeríaSPAReferee12_i$	0.830	(0.412)
$AtléticoSPAReferee2_i$	-0.633	(0.282)
$BilbaoSPAReferee1_i$	-0.934	(0.384)
$BilbaoVS.Espanyol_i$	-0.826	(0.340)
$CapacityAteam_i$	-5.74E-6	(1.66E-6)
$CeltaSPAReferee3_i$	1.322	(0.623)
$DeportivoSPAReferee4_i$	-0.945	(0.455)
$DeportivoNovember_i$	0.781	(0.286)
$DeportivoVS.Atlético_i$	-1.090	(0.486)
$DeportivoVS.Gijón_i$	-1.109	(0.484)
$ECIHome_i$ - $ECIAway_i$	0.51E-3	(5.44E-5)
$EspanyolShape-3_i$	-0.604	(0.187)
$EUIn6PreWinter_i$	0.230	(0.084)
$EULossPreWinterAteam_i$	-0.957	(0.267)
$EUPostWinterAteam_i$	-0.237	(0.069)
$EUAwayWinAteam_i$	-0.589	(0.177)
$GetafeJanuary_i$	-0.567	(0.225)
$GetafeSPAReferee5_i$	1.081	(0.434)
$GetafeSPAReferee6_i$	-0.869	(0.305)
$HomeBetis_i$	-0.413	(0.125)
$HomepointsHteam_i$	-0.010	(0.003)
$LevanteSPAReferee7_i$	0.900	(0.410)
$RealMadridSPAReferee9_i$	1.180	(0.598)
$SevillaAfter10PM_i$	-0.322	(-0.412)
$SevillaSPAReferee4_i$	1.058	(0.383)
$SevillaSPAReferee11_i$	1.245	(0.583)
$SevillaVS.Getafe_i$	-0.755	(0.347)
$Shape+1VS.Shape+1_i$	0.466	(0.155)
$Shape+2VS.Shape-1_i$	0.418	(0.161)
$SociedadVS.Barcelona_i$	1.016	(0.443)
$Spectators_i$	8.10E-6	(1.93E-6)
$ValenciaMarch_i$	-0.723	(0.206)
$ValenciaVS.Sociedad_i$	-1.260	(0.480)
μ_1	-0.685	(0.069)
μ_2	0.025	(0.068)

Table C.9: The table gives the extended ordered probit model based on AIC for Spain in the season 2015-16.

Variables	Coefficient	Std. Error
$Atlético SPAReferee2_i$	-0.548	(0.267)
$BarcelonaAfter10PM_i$	0.509	(0.190)
$BarcelonaSunday_i$	0.321	(0.147)
$BetisVS.Rayo_i$	-1.501	(0.633)
$BilbaoSPAReferee1_i$	-0.763	(0.340)
$BilbaoNovember_i$	0.771	(0.240)
$BilbaoVS.Deportivo_i$	-0.840	(0.366)
$BilbaoVS.Espanyol_i$	-0.681	(0.326)
$CapacityHteam_i$	5.34E-6	(1.59E-6)
$CeltaSPAReferee3_i$	1.042	(0.497)
$CeltaShape1_i$	0.630	(0.234)
$DeportivoSPAReferee4_i$	-0.916	(0.451)
$DeportivoNovember_i$	0.573	(0.267)
$DeportivoVS.Atlético_i$	-1.092	(0.464)
$ECIHome_i$ - $ECIAway_i$	0.57E-3	(4.66E-5)
$EspanyolShape-3_i$	-0.539	(0.177)
$EspanyolSPAReferee8_i$	0.616	(0.302)
$EUAwayWinAteam_i$	-0.398	(0.167)
$EULossPreWinterAteam_i$	-0.849	(0.243)
$EUPostWinterAteam_i$	-0.272	(0.066)
$GetafeJanuary_i$	-0.487	(0.213)
$GetafeSPAReferee5_i$	1.004	(0.376)
$GetafeSPAReferee6_i$	-0.911	(0.308)
$HomeBetis_i$	-0.370	(0.128)
$HomepointsHteam_i$	-0.010	(0.003)
$RayoVS.Getafe_i$	1.213	(0.541)
$RealMadridApril_i$	0.569	(0.254)
$RealMadridOctober_i$	0.609	(0.297)
$RealMadridSPAReferee9_i$	1.202	(0.571)
$SevillaSPAReferee3_i$	1.110	(0.549)
$SevillaSPAReferee4_i$	1.054	(0.386)
$SevillaSPAReferee11_i$	1.193	(0.585)
$Shape+1VS.Shape+1_i$	0.421	(0.143)
$Shape+2VS.Shape-1_i$	0.505	(0.154)
$SociedadVS.Barcelona_i$	1.111	(0.414)
$ValenciaMarch_i$	-0.481	(0.196)
$ValenciaVS.Granada_i$	1.175	(0.592)
$ValenciaVS.Sociedad_i$	-0.777	(0.388)
μ_1	-0.488	(0.067)
μ_2	0.235	(0.067)

Table C.10: The table gives the extended ordered probit model based on AIC for Sweden in the season 2013.

Variables	Coefficient	Std. Error
$3MatchesShapeAteam + 2_i$	-0.259	(0.117)
$5MatchesShapeAteam - 2_i$	0.271	(0.128)
$5MatchesShapeAteam - 5_i$	0.378	(0.117)
$AIKShape2_i$	0.732	(0.252)
$AIKSWEReferee1_i$	0.822	(0.337)
$AverageGoalsHomeAteam_i$	-0.258	(0.075)
$BrommaAugust_i$	-1.135	(0.401)
$BrommaSWEReferee5_i$	1.140	(0.410)
$CupfighterAteam_i$	-0.223	(0.073)
$DjurgårdensSWEReferee2_i$	-0.640	(0.299)
$ECIHome_i$ - $ECIAway_i$	0.69E-3	(8.25E-5)
$HomeHalmstads_i$	-0.343	(0.158)
$HomepointsHteam_i$	0.041	(0.011)
$KalmarSWEReferee7_i$	-0.648	(0.261)
$MatchesLeftHteam_i$	0.028	(0.007)
$MjälbySWEReferee1_i$	1.046	(0.381)
$MjällbyMarch_i$	1.822	(0.778)
$NorrköpingSWEReferee5_i$	1.096	(0.437)
$PercAwaypointsHteam_i$	0.706	(0.295)
μ_1	0.042	(0.229)
μ_2	0.780	(0.230)

Table C.11: The table gives the extended ordered probit model based on AIC for Sweden in the season 2014.

Variables	Coefficient	Std. Error
$\Delta ECI5MatchesHteam_i$	-0.003	(0.001)
$3MatchesShapeAteam - 3_i$	0.243	(0.110)
$3MatchesShapeAteam + 2_i$	-0.216	(0.109)
$3MatchesShapeHteam + 4_i$	0.244	(0.113)
$AIKSWEReferee1_i$	0.690	(0.334)
$AIKShape2_i$	0.824	(0.232)
$BrommaSWEReferee5_i$	0.811	(0.387)
$BrommaSWEReferee6_i$	-1.048	(0.475)
$CupfighterAteam_i$	-0.212	(0.063)
$DjurgårdensSWEReferee2_i$	-0.667	(0.303)
$DjurgårdensShape - 3_i$	-1.048	(0.475)
$ECIHome_i$ - $ECIAway_i$	0.65E-3	(2.88E-5)
$ElfsborgMay_i$	0.545	(0.251)
$HomeHalmstads_i$	-0.336	(0.142)
$HomepointsHteam_i$	0.031	(0.010)
$HalmstadsSWEReferee4_i$	1.119	(0.497)
$MatchesLeftHteam_i$	0.026	(0.006)
$MjälbySWEReferee1_i$	1.057	(0.382)
$MjällbyMarch_i$	1.625	(0.815)
$MjälbySWEReferee7_i$	1.101	(0.510)
$ÖrebroVS.Helsingborgs_i$	0.821	(0.379)
$PercAwaypointsHteam_i$	0.715	(0.274)
$PercHomepointsAteam_i$	-0.600	(0.204)
$Pos.\Delta ECILast5MatchesHteam_i$	0.304	(0.102)
μ_1	0.087	(0.215)
μ_2	0.828	(0.215)

Table C.12: The table gives the extended ordered probit model based on AIC for Sweden in the season 2015.

Variables	Coefficient	Std. Error
$3MatchesShapeAteam - 3_i$	0.210	(0.102)
$3MatchesShapeAteam + 2_i$	-0.293	(0.100)
$3MatchesShapeHteam + 4_i$	0.292	(0.100)
$AIKMarch_i$	-1.089	(0.510)
$AIKShape2_i$	0.703	(0.212)
$CupfighterAteam_i$	-0.131	(0.056)
$DjurgårdensSWEReferee2_i$	-0.785	(0.299)
$DjurgårdensVS.Hammarby_i$	-1.407	(0.664)
$ECIHome_i$ - $ECIAway_i$	0.82E-3	(6.38E-5)
$GefleSWEReferee1_i$	-0.622	(0.308)
$HalmstadsShape3_i$	-0.606	(0.293)
$HalmstadsSWEReferee3_i$	-0.590	(0.282)
$HalmstadsSWEReferee4_i$	1.072	(0.510)
$HammarbyShape1_i$	-1.594	(0.649)
$HammarbySunday_i$	-0.944	(0.344)
$KalmarSWEReferee7_i$	-0.589	(0.235)
$MatchesLeftHteam_i$	0.026	(0.006)
$Neg.\Delta ECILast5MatchesAteam_i$	0.306	(0.102)
$ÖrebroVS.Helsingborgs_i$	0.713	(0.333)
$Pos.\Delta ECILast5MatchesAteam_i$	0.220	(0.101)
μ_1	0.240	(0.217)
μ_2	0.975	(0.217)

D | Extended BIC models

Table D.1: The table gives the extended ordered probit model based on BIC for England in the season 2013-14.

Variables	Coefficient	Std. Error
$5MatchesShapeHteam + 3_i$	0.247	(0.088)
$ECIHome_i$ - $ECIAway_i$	0.65E-3	(3.26E-5)
$HomeManCity_i$	0.408	(0.127)
$ManCityVS.Newcastle_i$	1.454	(0.597)
$SouthamptonShape1_i$	1.427	(0.495)
μ_1	-0.661	(0.314)
μ_2	0.145	(0.029)

Table D.2: The table gives the extended ordered probit model based on BIC for England in the season 2014-15.

Variables	Coefficient	Std. Error
$ECIHome_i$ - $ECIAway_i$	0.64E-3	(3.02E-5)
$EUAwayDrawPostWinter_i$	0.898	(0.308)
$HomeManCity_i$	0.437	(0.120)
$ManCityVS.Newcastle_i$	1.508	(0.583)
$NewcastleVS.Swansea_i$	-1.838	(0.649)
$SwanseaENGReferee8_i$	-2.129	(0.819)
μ_1	-0.661	(0.028)
μ_2	0.119	(0.026)

Table D.3: The table gives the extended ordered probit model based on BIC for England in the season 2015-16.

Variables	Coefficient	Std. Error
$3MatchesShapeHteam - 2_i$	-0.220	(0.073)
$ECIHome_i$ - $ECIAway_i$	0.62E-3	(2.86E-5)
$EUAwayDrawPostWinter_i$	0.896	(0.308)
$HomeManCity_i$	0.417	(0.113)
$ManCityVS.Newcastle_i$	1.566	(0.570)
$ManUnitedShape - 3_i$	0.597	(0.214)
μ_1	-0.669	(0.027)
μ_2	0.102	(0.025)

Table D.4: The table gives the extended ordered probit model based on BIC for Netherlands in the season 2013-14.

Variables	Coefficient	Std. Error
$\Delta ECIOnArtificialTurfAteam_i$	-0.011	(0.002)
$AjaxVS.Utrecht_i$	-0.928	(0.335)
$ECIHome_i$ - $ECIAway_i$	0.84E-3	(6.07E-5)
$EULossAwayPostWinter_i$	-0.946	(0.309)
$Shape + 1VS.Shape + 3_i$	-0.584	(0.197)
$Spectators_i$	8.34E-6	(2.86E-6)
$VitesseNETReferee2_i$	-1.094	(0.335)
μ_1	-0.495	(0.065)
μ_2	0.215	(0.064)

Table D.5: The table gives the extended ordered probit model based on BIC for Netherlands in the season 2014-15.

Variables	Coefficient	Std. Error
$\Delta ECIOnArtificialTurfAteam_i$	-0.012	(0.002)
$ECIHome_i$ - $ECIAway_i$	0.78E-3	(5.64E-5)
$PSVBefore2PM_i$	-0.805	(0.285)
$Shape + 1VS.Shape + 3_i$	-0.494	(0.176)
$Spectators_i$	7.82E-6	(2.64E-6)
$VitesseNETReferee2_i$	-0.977	(0.273)
μ_1	-0.501	(0.060)
μ_2	0.220	(0.059)

Table D.6: The table gives the extended ordered probit model based on BIC for Netherlands in the season 2015-16.

Variables	Coefficient	Std. Error
$\Delta ECIOnArtificialTurfAteam_i$	-0.013	(0.002)
$ECIHome_i$ - $ECIAway_i$	0.76E-3	(5.25E-5)
$Shape - 3VS.Shape + 1_i$	0.492	(0.173)
$Spectators_i$	7.51E-6	(2.46E-6)
$VitesseNETReferee2_i$	-0.725	(0.244)
μ_1	-0.474	(0.056)
μ_2	0.242	(0.055)

Table D.7: The table gives the extended ordered probit model based on BIC for Spain in the season 2013-14.

Variables	Coefficient	Std. Error
$CapacityAteam_i$	-6.61E-6	(1.62E-6)
$CapacityHteam_i$	5.49E-6	(1.68E-6)
$ECIHome_i$ - $ECIAway_i$	0.43E-3	(1.68E-6)
$EspanyolShape-3_i$	-0.580	(0.197)
$GetafeSPAReferee6_i$	-0.895	(0.324)
$SevillaSPAReferee4_i$	1.025	(0.384)
$Shape+1VS.Shape+1_i$	0.451	(0.161)
μ_1	-0.710	(0.073)
μ_2	-0.025	(0.072)

Table D.8: The table gives the extended ordered probit model based on BIC for Spain in the season 2014-15.

Variables	Coefficient	Std. Error
$HomeBetis_i$	-0.376	(0.123)
$CapacityAteam_i$	-6.55E-6	(1.60E-6)
$ECIHome_i$ - $ECIAway_i$	0.44E-3	(5.09E05)
$Spectators_i$	7.69E-6	(1.82E-6)
$GetafeSPAReferee6_i$	-0.858	(0.302)
$EspanyolShape-3_i$	-0.528	(0.184)
$ValenciaMarch_i$	-0.628	(0.203)
μ_1	-0.722	(0.061)
μ_2	-0.040	(0.060)

Table D.9: The table gives the extended ordered probit model based on BIC for Spain in the season 2015-16.

Variables	Coefficient	Std. Error
$CapacityHteam_i$	5.60E-6	(1.47E-6)
$ECIHome_i$ - $ECIAway_i$	0.59E-3	(3.44E-5)
$EUPostWinterAteam_i$	-0.199	(0.061)
$GetafeSPAReferee6_i$	-0.954	(0.298)
$HomeBetis_i$	-0.373	(0.125)
$HomepointsHteam_i$	-0.009	(0.003)
$Shape+1VS.Shape+1_i$	0.431	(0.141)
$Shape+2VS.Shape-1_i$	0.466	(0.151)
μ_1	-0.472	(0.062)
μ_2	0.217	(0.062)

Table D.10: The table gives the extended ordered probit model based on BIC for Sweden in the season 2013.

Variables	Coefficient	Std. Error
$5MatchesShapeAteam - 5_i$	0.370	(0.115)
$AIKShape2_i$	0.757	(0.253)
$AIKSWEReferee1_i$	0.869	(0.337)
$AverageGoalsHomeAteam_i$	-0.283	(0.075)
$AwaymatchesHteam_i$	0.025	(0.007)
$BrommaAugust_i$	-1.110	(0.400)
$BrommaSWEReferee5_i$	1.110	(0.410)
$CupfighterAteam_i$	-0.201	(0.072)
$ECIHome_i$ - $ECIAway_i$	0.61E-3	(8.05E-5)
$HomepointsHteam_i$	0.037	(0.011)
$MjälbySWEReferee1_i$	1.067	(0.379)
$PercAwaypointsHteam_i$	0.899	(0.291)
μ_1	0.023	(0.226)
μ_2	0.747	(0.227)

Table D.11: The table gives the extended ordered probit model based on BIC for Sweden in the season 2014.

Variables	Coefficient	Std. Error
$AIKShape2_i$	0.800	(0.228)
$AwaymatchesHteam_i$	0.018	(0.006)
$CupfighterAteam_i$	-0.188	(0.062)
$ECIHome_i$ - $ECIAway_i$	0.61E-3	(7.46E-5)
$HomepointsHteam_i$	0.026	(0.010)
$PercAwaypointsHteam_i$	0.797	(0.262)
$PercHomepointsAteam_i$	-0.643	(0.119)
μ_1	-0.197	(0.195)
μ_2	0.516	(0.196)

Table D.12: The table gives the extended ordered probit model based on BIC for Sweden in the season 2015.

Variables	Coefficient	Std. Error
$3MatchesShapeAteam + 2_i$	-0.319	(0.098)
$3MatchesShapeHteam + 4_i$	0.283	(0.098)
$AIKShape2_i$	0.697	(0.207)
$ECIHome_i$ - $ECIAway_i$	0.72E-3	(5.80E-5)
μ_1	-0.605	(0.035)
μ_2	0.107	(0.033)

E | Evaluation methods

Table E.1: The table gives the results from the Jaccard index evaluation method for each of the four constructed models. The profit of the best predicting proposed model per competition seasons is underlined and bold in this table. An asterisk (*) is added in case the best predicting proposed models performs better than the bookmakers probabilities.

Model	Competition	2013(-14)	2014(-15)	2015(-16)	Average
Basic ECI model	England	0.413	0.355	**0.301**	**0.356**
	Netherlands	0.299	**0.385***	0.378	0.354
	Spain	**0.357**	**0.400***	**0.357**	**0.371***
	Sweden	**0.345***	0.330	**0.392**	**0.356***
Simplified model	England	**0.415**	0.362	**0.301**	0.359
	Netherlands	0.297	**0.385***	0.385	0.356
	Spain	0.355	0.397	0.350	0.367
	Sweden	0.341	0.304	0.374	0.340
Extended AIC model	England	0.384	**0.367**	0.290	0.347
	Netherlands	**0.325***	0.375	**0.400***	**0.367***
	Spain	0.355	0.397	0.348	0.367
	Sweden	0.319	**0.345***	0.361	0.342
Extended BIC model	England	0.397	0.360	0.297	0.351
	Netherlands	0.302	0.372	0.381	0.352
	Spain	0.355	0.387	0.348	0.363
	Sweden	0.326	0.333	0.374	0.344
Bookmakers probabilities	England	0.431	0.360	0.306	0.366
	Netherlands	0.319	0.375	0.385	0.360
	Spain	0.377	0.397	0.382	0.385
	Sweden	0.326	0.330	0.396	0.351

Table E.2: The table gives the results from the ordered probit penalty index for each of the four constructed models. The profit of the best predicting proposed model per competition seasons is underlined and bold in this table. An asterisk (*) is added in case the best predicting proposed models performs better than the bookmakers probabilities.

Model	Competition	2013(-14)	2014(-15)	2015(-16)	Average
Basic ECI model	England	-0.329	-0.363	-0.440	-0.377
	Netherlands	-0.387	-0.342	-0.329	-0.353
	Spain	-0.346	-0.280	**-0.319**	-0.315
	Sweden	-0.395	**-0.414**	**-0.389**	**-0.399**
Simplified model	England	-0.342	-0.369	-0.442	-0.384
	Netherlands	-0.399	-0.353	-0.337	-0.363
	Spain	-0.345	-0.281	-0.327	-0.318
	Sweden	**-0.441**	-0.451	-0.430	-0.441
Extended AIC	England	**-0.327**	**-0.323***	**-0.422***	**-0.357***
model	Netherlands	**-0.365***	**-0.339***	**-0.319***	**-0.341***
	Spain	**-0.317***	**-0.273**	-0.325	**-0.305**
	Sweden	-0.462	-0.421	-0.394	-0.426
Extended BIC	England	-0.350	-0.365	-0.442	-0.386
model	Netherlands	-0.377	-0.367	-0.347	-0.364
	Spain	-0.348	-0.288	**-0.319**	-0.318
	Sweden	-0.462	-0.460	-0.425	-0.449
Bookmakers	England	-0.326	-0.375	-0.433	-0.378
probabilities	Netherlands	-0.383	-0.364	-0.356	-0.368
	Spain	-0.326	-0.268	-0.310	-0.301
	Sweden	-0.424	-0.396	-0.364	-0.395

Table E.3: The table gives the results from the rank probability score for each of the four constructed models. The profit of the best predicting proposed model per competition seasons is underlined and bold in this table. An asterisk (*) is added in case the best predicting proposed models performs better than the bookmakers probabilities.

Model	Competition	2013(-14)	2014(-15)	2015(-16)	Average
Basic ECI model	England	**0.395**	0.403	0.432	**0.410**
	Netherlands	0.415	0.401	0.381	0.399
	Spain	0.402	0.367	**0.389**	**0.386**
	Sweden	0.413	0.418	**0.396**	**0.409**
Simplified model	England	0.396	0.403	**0.431**	**0.410**
	Netherlands	0.413	**0.400**	0.381	**0.398**
	Spain	0.402	**0.366**	0.390	**0.386**
	Sweden	**0.412***	0.417	0.402	0.410
Extended AIC	England	0.412	**0.401**	0.449	0.421
model	Netherlands	0.417	0.407	**0.378***	0.401
	Spain	**0.398**	0.376	0.398	0.391
	Sweden	0.447	**0.415**	0.405	0.422
Extended BIC	England	0.421	0.406	0.435	0.421
model	Netherlands	**0.407***	0.407	0.385	0.400
	Spain	0.406	0.371	0.393	0.390
	Sweden	0.435	0.422	0.407	0.421
Bookmakers	England	0.381	0.394	0.419	0.398
probabilities	Netherlands	0.408	0.398	0.385	0.397
	Spain	0.389	0.356	0.370	0.372
	Sweden	0.413	0.398	0.381	0.397

Table E.4: The table gives the results from the mean squared error. The profit of the best predicting proposed model per competition seasons is underlined and bold in this table. An asterisk (*) is added in case the best predicting proposed models performs better than the bookmakers probabilities.

Model	Competition	2013(-14)	2014(-15)	2015(-16)	Average
Basic ECI model	England	**0.186**	0.195	0.211	**0.197**
	Netherlands	**0.201**	**0.194**	0.188	**0.194**
	Spain	0.191	**0.181**	**0.189**	**0.187**
	Sweden	0.203	**0.201**	**0.194**	**0.199**
Simplified model	England	0.187	0.195	**0.210**	**0.197**
	Netherlands	0.205	**0.194**	0.188	0.196
	Spain	0.191	**0.181**	0.190	**0.187**
	Sweden	**0.202**	**0.201**	0.195	**0.199**
Extended AIC model	England	0.192	**0.194**	0.217	0.201
	Netherlands	0.207	0.196	**0.186***	0.196
	Spain	**0.189**	0.185	0.193	0.189
	Sweden	0.213	**0.201**	0.197	0.204
Extended BIC model	England	0.195	0.196	0.211	0.201
	Netherlands	0.203	0.196	0.189	0.196
	Spain	0.192	0.183	0.191	0.189
	Sweden	0.209	0.203	0.197	0.203
Bookmakers probabilities	England	0.181	0.192	0.207	0.193
	Netherlands	0.201	0.191	0.188	0.193
	Spain	0.186	0.176	0.183	0.182
	Sweden	0.202	0.194	0.188	0.195

F | Variable explanation

Table F.1: The table provides an explanation for variables included in the extended prediction models, the extended AIC models and the extended BIC models.

Variable	**Definition**
$\Delta ECI3MatchesAteam_i$	The difference in ECI value over the past three matches for the away team
$\Delta ECI5MatchesAteam_i$	The difference in ECI value over the past five matches for the away team
$\Delta ECI5MatchesHteam_i$	The difference in ECI value over the past five matches for the home team
$\Delta ECIOnArtificialTurfAteam_i$	The increment in ECI value in away matches on artificial turf in current and last season away team
$3MatchesShapeAteam + 2_i$	Dummy variable that is 1 in case the away team is in 3 matches shape groups +2 and 0 in all other cases [2]
$3MatchesShapeAteam - 3_i$	Dummy variable that is 1 in case the away team is in 3 match shape groups -3 and 0 in all other cases [2]
$3MatchesShapeHteam + 4_i$	Dummy variable that is 1 in case the home team is in 3 match shape groups +4 and 0 in all other cases [2]
$3MatchesShapeHteam - 2_i$	Dummy variable that is 1 in case the home team is in 3 match shape groups -2 and 0 in all other cases [2]
$5MatchesShapeAteam - 2_i$	Dummy variable that is 1 in case the away team is in 5 match shape groups -2 and 0 in all other cases [2]
$5MatchesShapeAteam - 3_i$	Dummy variable that is 1 in case the away team is in 5 match shape groups -3 and 0 in all other cases [2]
$5MatchesShapeAteam - 5_i$	Dummy variable that is 1 in case the away team is in 5 match shape groups -5 and 0 in all other cases [2]
$5MatchesShapeHteam + 3_i$	Dummy variable that is 1 in case the home team is in 5 match shape groups +3 and 0 in all other cases [2]
$5MatchesShapeHteam - 3_i$	Dummy variable that is 1 in case the home team is in 5 match shape groups -3 and 0 in all other cases [2]
$AIKMarch_i$	Variable that is 1 in case AIK plays a home match in March, -1 in case AIK plays an away match in March and 0 in all other cases
$AIKShape2_i$	Variable that is 1 in case AIK plays a home match in shape 2, -1 in case AIK plays an away match in shape 2 and 0 in all other cases
$AIKSWEReferee1_i$	Variable that is 1 in case AIK plays a home match with SWE referee 1, -1 in case AIK plays an away match with SWE referee 1 and 0 in all other cases
$AjaxMarch_i$	Variable that is 1 in case Ajax plays a home match in March, -1 in case Ajax plays an away match in March and 0 in all other cases

Variable	Definition
$AjaxVS.Utrecht_i$	Variable that is 1 in case Ajax plays home against FC Utrecht, -1 in case FC Utrecht plays home against Ajax and 0 in all other cases
$AlmeríaSPAReferee12_i$	Variable that is 1 in case UD Almería plays a home match with SPA referee 12, -1 in case UD Almería plays an away match with SPA referee 12 and 0 in all other cases
$ArsenalENGReferee1_i$	Variable that is 1 in case Arsenal plays a home match with ENG referee 1, -1 in case Arsenal plays an away match with ENG referee 1 and 0 in all other cases
$ArsenalVS.Swansea_i$	Variable that is 1 in case Arsenal plays home against Swansea City, -1 in case Swansea City plays home against Arsenal and 0 in all other cases
$AtléticoSPAReferee2_i$	Variable that is 1 in case Atlético Madrid plays a home match with SPA referee 2, -1 in case Atlético Madrid plays an away match with SPA referee 2 and 0 in all other cases
$AverageGoalsAgainstAteam_i$	The average amount of goals against by the away team
$AverageGoalsAwayHteam_i$	The average amount of away goals by the home team
$AverageHomeAgainstHteam_i$	The average amount of home goals against by the home team
$AverageGoalsHomeAteam_i$	The average amount of home goals by the away team
$AwaymatchesAteam_i$	The amount of away matches played in the current season by the away team
$AwaypointsHteam_i$	The current amount of away points of the home team
$AZNovember_i$	Variable that is 1 in case AZ Alkmaar plays a home match in November -1 in case AZ Alkmaar plays an away match in November and 0 in all other cases
$AZShape-2_i$	Variable that is 1 in case AZ Alkmaar plays a home match in shape -2, -1 in case AZ Alkmaar plays an away match in shape -2 and 0 in all other cases
$BarcelonaAfter10PM_i$	Variable that is 1 in case FC Barcelona plays a home match after 10 PM, -1 in case FC Barcelona plays an away match after 10 PM and 0 in all other cases
$BarcelonaSunday_i$	Variable that is 1 in case FC Barcelona plays a home match on Sunday, -1 in case FC Barcelona plays an away match on Sunday and 0 in all other cases
$BetisVS.Rayo_i$	Variable that is 1 in case Real Betis plays home against Rayo Vallecano, -1 in case Rayo Vallecano plays home against Real Betis and 0 in all other cases
$BilbaoNovember_i$	Variable that is 1 in case Athletic Club de Bilbao plays a home match in November -1 in case Athletic Club de Bilbao plays an away match in November and 0 in all other cases
$BilbaoSPAReferee1_i$	Variable that is 1 in case Athletic Club de Bilbao plays a home match with SPA referee 1, -1 in case Athletic Club de Bilbao plays an away match with SPA referee 1 and 0 in all other cases

Variable	Definition
$BilbaoVS.Deportivo_i$	Variable that is 1 in case Athletic Club de Bilbao plays home against Deportivo La Coruña, -1 in case Deportivo La Coruña plays home against Athletic Club de Bilbao and 0 in all other cases
$BilbaoVS.Espanyol_i$	Variable that is 1 in case Athletic Club de Bilbao plays home against RCD Espanyol, -1 in case RCD Espanyol plays home against Athletic Club de Bilbao and 0 in all other cases
$BrommaAugust_i$	Variable that is 1 in case IF Brommapojkarna plays a home match in August, -1 in case IF Brommapojkarna plays an away match in August and 0 in all other cases
$BrommaSWEReferee5_i$	Variable that is 1 in case IF Brommapojkarna plays a home match with SWE referee 5, -1 in case IF Brommapojkarna plays an away match with SWE referee 5 and 0 in all other cases
$BrommaSWEReferee6_i$	Variable that is 1 in case IF Brommapojkarna plays a home match with SWE referee 6, -1 in case IF Brommapojkarna plays an away match with SWE referee 6 and 0 in all other cases
$CapacityAteam_i$	The stadium capacity of the away team
$CapacityHteam_i$	The stadium capacity of the home team
$CeltaShape1_i$	Variable that is 1 in case Celta de Vigo plays a home match in shape 1, -1 in case Celta de Vigo plays an away match in shape 1 and 0 in all other cases
$CeltaSPAReferee3_i$	Variable that is 1 in case Celta de Vigo plays a home match with SPA referee 3, -1 in case Celta de Vigo plays an away match with SPA referee 3 and 0 in all other cases
$ChelseaApril_i$	Variable that is 1 in case Chelsea plays a home match in April, -1 in case Chelsea plays an away match in April and 0 in all other cases
$ChelseaNovember_i$	Variable that is 1 in case Chelsea plays a home match in November, -1 in case Chelsea plays an away match in November and 0 in all other cases
$CupfighterAteam_i$	Dummy variable that is 1 in case the away team is still active in the national cup and 0 in all other cases
$CupmatchHteam_i$	Dummy variable that is 1 in case the home team played an national cup match within six days before the current match and 0 in all other cases
$DeportivoNovember_i$	Variable that is 1 in case Deportivo La Coruña plays a home match in November, -1 in case Deportivo La Coruña plays an away match in November and 0 in all other cases
$DeportivoSPAReferee4_i$	Variable that is 1 in case Deportivo La Coruña plays a home match with SPA referee 4, -1 in case Deportivo La Coruña plays an away match with SPA referee 4 and 0 in all other cases

Variable	Definition
$DeportivoVS.Atlético_i$	Variable that is 1 in case Deportivo La Coruña plays home against Atlético Madrid, -1 in case Atlético Madrid plays home against Deportivo La Coruña and 0 in all other cases
$DeportivoVS.Gijón_i$	Variable that is 1 in case Deportivo La Coruña plays home against Sporting Gijón, -1 in case Sporting Gijón plays home against Deportivo La Coruña and 0 in all other cases
$DjurgårdensShape-3_i$	Variable that is 1 in case Djurgårdens IF plays a home match in shape -3, -1 in case Djurgårdens IF plays an away match in shape -3 and 0 in all other cases
$DjurgårdensSWEReferee2_i$	Variable that is 1 in case Djurgårdens IF plays a home match with SWE referee 2, -1 in case Djurgårdens IF plays an away match with SWE referee 2 and 0 in all other cases
$DjurgårdensVS.Hammarby_i$	Variable that is 1 in case Djurgårdens IF plays home against Hammarby IF, -1 in case Hammarby IF plays home against Djurgårdens IF and 0 in all other cases
$ECIHome-ECIAway_i$	Gives the difference in ECI value between the home team and away team
$ElfsborgMay_i$	Variable that is 1 in case IF Elfsborg plays a home match in May, -1 in case IF Elfsborg plays an away match in May and 0 in all other cases
$EspanyolShape-3_i$	Variable that is 1 in case RCD Espanyol plays a home match in shape -3, -1 in case RCD Espanyol plays an away match in shape -3 and 0 in all other cases
$EspanyolSPAReferee8_i$	Variable that is 1 in case RCD Espanyol plays a home match with SPA referee 8, -1 in case RCD Espanyol plays an away match with SPA referee 8 and 0 in all other cases
$EUAwayDrawPostWinter_i$	Variable that is 1 in case the last match of the home team was a European match after Newyears that ended in a draw, -1 in case the last match of the away team was a European match after Newyears that ended in a draw and 0 in all other cases
$EUAwayLossPostWinterHteam_i$	Dummy variable that is 1 in case the last match of the home team was a European away match after Newyears that ended in a loss and 0 in all other cases
$EUAwayWinAteam_i$	Dummy variable that is 1 in case the last match of the away team was a European away match that ended in a win and 0 in all other cases
$EUAwayWinPostWinterAteam_i$	Dummy variable that is 1 in case the last match of the away team was a European away match after Newyears that ended in a win and 0 in all other cases
$EUHomeIn6WinPreWinter_i$	Variable that is 1 in case the last match of the home team was a European home match before Newyears that ended in a win and was within six days, -1 in case the last match of the home team was a European home match before Newyears that ended in a win and was within six day and 0 in all other cases

Variable	Definition
$EUIn6PreWinter_i$	Variable that is 1 in case the last match of the home team was a European match before Newyears and was within six days, -1 in case the last match of the home team was a European match before Newyears and was within six days and 0 in all other cases
$EUIn6WinPostWinterHteam_i$	Dummy variable that is 1 in case the last match of the home team was a European match after Newyears that ended in a win and was within 6 days and 0 in all other cases
$EULossAwayPostWinter_i$	Variable that is 1 in case the last match of the home team was a European away match after Newyears that ended in a loss, -1 in case the last match of the away team was a European away match after Newyears that ended in a loss and 0 in all other cases
$EULossPreWinterAteam_i$	Dummy variable that is 1 in case the last match of the away team was a European match before Newyears that ended in a loss and 0 in all other cases
$EUPostWinterAteam_i$	Dummy variable that is 1 in case the last match of the away team was a European away match after Newyears and 0 in all other cases
$EUPostWinterHteam_i$	Dummy variable that is 1 in case the last match of the home team was a European match after Newyears and 0 in all other cases
$EvertonShape-1_i$	Variable that is 1 in case Everton plays a home match in shape -1, -1 in case Everton plays an away match in shape -1 and 0 in all other cases
$EvertonVS.Arsenal_i$	Variable that is 1 in case Everton plays home against Arsenal, -1 in case Arsenal plays home against Everton and 0 in all other cases
$FeyenoordJanuary_i$	Variable that is 1 in case Feyenoord plays a home match in January, -1 in case Feyenoord plays an away match in January and 0 in all other cases
$FeyenoordNETReferee4_i$	Variable that is 1 in case Feyenoord plays a home match with NET referee 4, -1 in case Feyenoord plays an away match with NET referee 4 and 0 in all other cases
$FeyenoordShape-2_i$	Variable that is 1 in case Feyenoord plays a home match in shape -2, -1 in case Feyenoord plays an away match in shape -2 and 0 in all other cases
$GefleSWEReferee1_i$	Variable that is 1 in case Gefle IF plays a home match with SWE referee 1, -1 in case Gefle IF plays an away match with SWE referee 1 and 0 in all other cases
$GetafeJanuary_i$	Variable that is 1 in case Getafe CF plays a home match in January, -1 in case Getafe CF plays an away match in January and 0 in all other cases
$GetafeSPAReferee5_i$	Variable that is 1 in case Getafe CF plays a home match with SPA referee 5, -1 in case Getafe CF plays an away match with SPA referee 5 and 0 in all other cases

Variable	Definition
$GetafeSPAReferee6_i$	Variable that is 1 in case Getafe CF plays a home match with SPA referee 6, -1 in case Getafe CF plays an away match with SPA referee 6 and 0 in all other cases
$GraafschapShape1_i$	Variable that is 1 in case De graafschap plays a home match in shape 1, -1 in case De graafschap plays an away match in shape 1 and 0 in all other cases
$GroningenNETReferee1_i$	Variable that is 1 in case FC Groningen plays a home match with NET referee 1, -1 in case FC Groningen plays an away match with NET referee 1 and 0 in all other cases
$GroningenNETReferee2_i$	Variable that is 1 in case FC Groningen plays a home match with NET referee 2, -1 in case FC Groningen plays an away match with NET referee 2 and 0 in all other cases
$GroningenShape-2_i$	Variable that is 1 in case FC Groningen plays a home match in shape -2, -1 in case FC Groningen plays an away match in shape -2 and 0 in all other cases
$HalmstadsShape3_i$	Variable that is 1 in case Halmstads BK plays a home match in shape 3, -1 in case Halmstads BK plays an away match in shape 3 and 0 in all other cases
$HalmstadsSWEReferee3_i$	Variable that is 1 in case Halmstads BK plays a home match with SWE referee 3, -1 in case Halmstads BK plays an away match with SWE referee 3 and 0 in all other cases
$HalmstadsSWEReferee4_i$	Variable that is 1 in case Halmstads BK plays a home match with SWE referee 4, -1 in case Halmstads BK plays an away match with SWE referee 4 and 0 in all other cases
$HammarbyShape1_i$	Variable that is 1 in case Hammarby IF plays a home match in shape 1, -1 in case Hammarby IF plays an away match in shape 1 and 0 in all other cases
$HammarbySunday_i$	Variable that is 1 in case Hammarby IF plays a home match on Sunday, -1 in case Hammarby IF plays an away match on Sunday and 0 in all other cases
$HomeBetis_i$	Dummy variable that is 1 in case Real Betis plays a home match and 0 in all other cases
$HomeHalmstads_i$	Dummy variable that is 1 in case Halmstads BK plays a home match and 0 in all other cases
$HomeManCity_i$	Dummy variable that is 1 in case Manchester City plays a home match and 0 in all other cases
$HomeStoke_i$	Dummy variable that is 1 in case Stoke City plays a home match and 0 in all other cases
$HomeVilla_i$	Dummy variable that is 1 in case Aston Villa plays a home match and 0 in all other cases
$HomepointsHteam_i$	The current amount of home points of the home team
$KalmarSWEReferee7_i$	Variable that is 1 in case Kalmar FF plays a home match with SWE referee 7, -1 in case Kalmar FF plays an away match with SWE referee 7 and 0 in all other cases

Variable	Definition
$LevanteSPAReferee7_i$	Variable that is 1 in case Levante UD plays a home match with SPA referee 7, -1 in case Levante UD plays an away match with SPA referee 7 and 0 in all other cases
$LiverpoolENGReferee2_i$	Variable that is 1 in case Liverpool plays a home match with ENG referee 2, -1 in case Liverpool plays an away match with ENG referee 2 and 0 in all other cases
$LiverpoolENGReferee3_i$	Variable that is 1 in case Liverpool plays a home match with ENG referee 3, -1 in case Liverpool plays an away match with ENG referee 3 and 0 in all other cases
$LiverpoolMarch_i$	Variable that is 1 in case Liverpool plays a home match in March, -1 in case Liverpool plays an away match in March and 0 in all other cases
$ManCityAugust_i$	Variable that is 1 in case Manchester City plays a home match in August, -1 in case Manchester City plays an away match in August and 0 in all other cases
$ManCityENGReferee4_i$	Variable that is 1 in case Manchester City plays a home match with ENG referee 4, -1 in case Manchester City plays an away match with ENG referee 4 and 0 in all other cases
$ManCityVS.Newcastle_i$	Variable that is 1 in case Manchester City plays home against Newcastle United, -1 in case Newcastle United plays home against Manchester City and 0 in all other cases
$ManUnitedShape-3_i$	Variable that is 1 in case Manchester United plays a home match in shape -3, -1 in case Manchester United plays an away match in shape -3 and 0 in all other cases
$MatchesLeftAteam_i$	The amount of matches left to play in the current seasons by the away team
$MatchesLeftHteam_i$	The amount of matches left to play in the current seasons by the home team
$MjälbySWEReferee1_i$	Variable that is 1 in case Mjällby AIF plays a home match with SWE referee 1, -1 in case Mjällby AIF plays an away match with SWE referee 1 and 0 in all other cases
$MjälbySWEReferee7_i$	Variable that is 1 in case Mjällby AIF plays a home match with SWE referee 7, -1 in case Mjällby AIF plays an away match with SWE referee 7 and 0 in all other cases
$MjällbyMarch_i$	Variable that is 1 in case Mjällby AIF plays a home match in March, -1 in case Mjällby AIF plays an away match in March and 0 in all other cases
$NACNETReferee5_i$	Variable that is 1 in case NAC Breda plays a home match with NET referee 5, -1 in case NAC Breda plays an away match with NET referee 5 and 0 in all other cases
$Neg.\Delta ECILast5MatchesAteam_i$	Dummy variable that is 1 in case the increment in ECI value of the away team in negative over the last five matches

Variable	Definition
$NewcastleENGReferee5_i$	Variable that is 1 in case Newcastle United plays a home match with ENG referee 5, -1 in case Newcastle United plays an away match with ENG referee 5 and 0 in all other cases
$NewcastleVS.Swansea_i$	Variable that is 1 in case Newcastle United plays home against Swansea City, -1 in case Swansea City plays home against Newcastle United and 0 in all other cases
$NorrköpingSWEReferee4_i$	Variable that is 1 in case IFK Norrköping plays a home match with SWE referee 4, -1 in case IFK Norrköping plays an away match with SWE referee 5 and 0 in all other cases
$ÖrebroVS.Helsingborgs_i$	Variable that is 1 in case Örebro SK plays home against Helsingborgs IF, -1 in case Helsingborgs IF plays home against Örebro SK and 0 in all other cases
$PercAwaypointsHteam_i$	The percentage of possible awaypoints home team
$PercHomepointsAteam_i$	The percentage of possible homepoints away team
$PercPointsAteam_i$	The percentage of possible points away team
$Pos.\Delta ECILast5MatchesAteam_i$	Dummy variable that is 1 in case the increment in ECI value of the away team in positive over the last five matches
$Pos.\Delta ECILast5MatchesHteam_i$	Dummy variable that is 1 in case the increment in ECI value of the home team in positive over the last five matches
$PSVBefore2PM_i$	Variable that is 1 in case PSV plays a home match before 2 PM, -1 in case PSV plays an away match before 2 PM and 0 in all other cases
$RankAteam_i$	The current rank of the away team
$RayoVS.Getafe_i$	Variable that is 1 in case Rayo Vallecano plays home against Getafe CF, -1 in case Getafe CF plays home against Rayo Vallecano and 0 in all other cases
$RealMadridApril_i$	Variable that is 1 in case Real Madrid plays a home match in April, -1 in case Real Madrid plays an away match in April and 0 in all other cases
$RealMadridOctober_i$	Variable that is 1 in case Real Madrid plays a home match in October, -1 in case Real Madrid plays an away match in October and 0 in all other cases
$RealMadridSPAReferee9_i$	Variable that is 1 in case Real Madrid plays a home match with SPA referee 9, -1 in case Real Madrid plays an away match with SPA referee 9 and 0 in all other cases
$RealMadridWeek_i$	Variable that is 1 in case Real Madrid plays a home match outside the weekend, -1 in case Madrid plays an away match outside the weekend and 0 in all other cases
$RKCReferee6_i$	Variable that is 1 in case RKC Waalwijk plays a home match with NET referee 6, -1 in case RKC Waalwijk plays an away match with NET referee 6 and 0 in all other cases

Variable	Definition
$SevillaAfter10PM_i$	Variable that is 1 in case Sevilla FC plays a home match after 10 PM, -1 in case Sevilla FC plays an away match after 10 PM and 0 in all other cases
$SevillaSPAReferee11_i$	Variable that is 1 in case Sevilla FC plays a home match with SPA referee 11, -1 in case Sevilla FC plays an away match with SPA referee 11 and 0 in all other cases
$SevillaSPAReferee3_i$	Variable that is 1 in case Sevilla FC plays a home match with SPA referee 3, -1 in case Sevilla FC plays an away match with SPA referee 3 and 0 in all other cases
$SevillaSPAReferee4_i$	Variable that is 1 in case Sevilla FC plays a home match with SPA referee 4, -1 in case Sevilla FC plays an away match with SPA referee 4 and 0 in all other cases
$SevillaVS.Getafe_i$	Variable that is 1 in case Sevilla plays home against Getafe CF, -1 in case Getafe CF plays home against Sevilla and 0 in all other cases
$Shape+1VS.Shape+1_i$	Dummy variable that is 1 in case the home team is in shape groups +1 and the awayteam is in shape group +1 and 0 in all other cases [3]
$Shape+2VS.Shape-1_i$	Dummy variable that is 1 in case the home team is in shape groups +2 and the awayteam is in shape group -1 and 0 in all other cases [3]
$Shape+2VS.Shape-3_i$	Dummy variable that is 1 in case the home team is in shape groups +2 and the awayteam is in shape group -3 and 0 in all other cases [3]
$Shape+2VS.Shape+2_i$	Dummy variable that is 1 in case the home team is in shape groups +2 and the awayteam is in shape group +2 and 0 in all other cases [3]
$Shape+3VS.Shape+3_i$	Dummy variable that is 1 in case the home team is in shape groups +3 and the awayteam is in shape group +3 and 0 in all other cases [3]
$Shape-2VS.Shape+1_i$	Dummy variable that is 1 in case the home team is in shape groups -2 and the awayteam is in shape group +1 and 0 in all other cases [3]
$Shape-3VS.Shape+1_i$	Dummy variable that is 1 in case the home team is in shape groups -3 and the awayteam is in shape group +1 and 0 in all other cases [3]
$Shape-1VS.Shape+2_i$	Dummy variable that is 1 in case the home team is in shape groups -1 and the awayteam is in shape group +2 and 0 in all other cases [3]
$Shape+1VS.Shape+3_i$	Dummy variable that is 1 in case the home team is in shape groups +1 and the awayteam is in shape group +3 and 0 in all other cases [3]
$SociedadSPAReferee10_i$	Variable that is 1 in case Real Sociedad plays a home match with SPA referee 10, -1 in case Real Sociedad plays an away match with SPA referee 10 and 0 in all other cases
$SociedadVS.Barcelona_i$	Variable that is 1 in case Real Sociedad plays home against FC Barcelona, -1 in case FC Barcelona plays home against Real Sociedad and 0 in all other cases

Variable	Definition
$SouthamptonJanuary_i$	Variable that is 1 in case Southampton plays a home match in January, -1 in case Southampton plays an away match in January and 0 in all other cases
$SouthamptonSaturday_i$	Variable that is 1 in case Southampton plays a home match on Saturday, -1 in case Southampton plays an away match on Saturday and 0 in all other cases
$SouthamptonShape1_i$	Variable that is 1 in case Southampton plays a home match in shape 1, -1 in case Southampton plays an away match in shape 1 and 0 in all other cases
$Spectators_i$	The number of spectators
$StokeENGReferee6_i$	Variable that is 1 in case Stoke City plays a home match with ENG referee 1, -1 in case Stoke City plays an away match with ENG referee 1 and 0 in all other cases
$SunderlandJanuary_i$	Variable that is 1 in case Sunderland plays a home match in January, -1 in case Sunderland plays an away match in January and 0 in all other cases
$SunderlandVS.WBA_i$	Variable that is 1 in case Sunderland plays home against West Bromwich Albion, -1 in case West Bromwich Albion plays home against Sunderland and 0 in all other cases
$SwanseaAugust_i$	Variable that is 1 in case Swansea City plays a home match in August, -1 in case Swansea City plays an away match in August and 0 in all other cases
$SwanseaENGReferee7_i$	Variable that is 1 in case Swansea City plays a home match with ENG referee 7, -1 in case Swansea City plays an away match with ENG referee 7 and 0 in all other cases
$SwanseaENGReferee8_i$	Variable that is 1 in case Swansea City plays a home match with ENG referee 8, -1 in case Swansea City plays an away match with ENG referee 8 and 0 in all other cases
$SwanseaShape-1_i$	Variable that is 1 in case Swansea City plays a home match in shape -1, -1 in case Swansea City plays an away match in shape -1 and 0 in all other cases
$TottenhamENGReferee9_i$	Variable that is 1 in case Tottenham Hotpur plays a home match with ENG referee 9, -1 in case Tottenham Hotpur plays an away match with ENG referee 9 and 0 in all other cases
$TottenhamFebruary_i$	Variable that is 1 in case Tottenham Hotspur plays a home match in February, -1 in case Tottenham Hotspur plays an away match in February and 0 in all other cases
$TottenhamVS.Swansea_i$	Variable that is 1 in case Tottenham Hotspur plays home against Swansea City, -1 in case Swansea City plays home against Tottenham Hotspur and 0 in all other cases
$TwenteFebruary_i$	Variable that is 1 in case FC Twente plays a home match in February, -1 in case FC Twente plays an away match in February and 0 in all other cases
$UtrechtBetween6PMAnd8PM_i$	Variable that is 1 in case FC Utrecht plays a home match between 6 PM and 8 PM, -1 in case FC Utrecht plays an away match between 6 PM and 8 PM and 0 in all other cases

Variable	Definition
$UtrechtNETReferee3_i$	Variable that is 1 in case FC Utrecht plays a home match with NET referee 3, -1 in case FC Utrecht plays an away match with NET referee 3 and 0 in all other cases
$ValenciaMarch_i$	Variable that is 1 in case Valencia CF plays a home match in March, -1 in case Valencia CF plays an away match in March and 0 in all other cases
$ValenciaVS.Granada_i$	Variable that is 1 in case Valencia CF plays home against Granada CF, -1 in case Granada CF plays home against Valencia CF and 0 in all other cases
$ValenciaVS.Sociedad_i$	Variable that is 1 in case Valencia CF plays home against Real Sociedad, -1 in case Real Sociedad plays home against Valencia CF and 0 in all other cases
$VillaMarch_i$	Variable that is 1 in case Aston Villa plays a home match in March, -1 in case Aston Villa plays an away match in March and 0 in all other cases
$VitesseMay_i$	Variable that is 1 in case Vitesse plays a home match in May, -1 in case Vitesse plays an away match in May and 0 in all other cases
$VitesseNETReferee2_i$	Variable that is 1 in case Vitesse plays a home match with NET referee 2, -1 in case Vitesse plays an away match with NET referee 2 and 0 in all other cases
$WBAENGReferee9_i$	Variable that is 1 in case West Bromwich Albion plays a home match with ENG referee 9, -1 in case West Bromwich Albion plays an away match with ENG referee 9 and 0 in all other cases
$WBASeptember_i$	Variable that is 1 in case West Bromwich Albion plays a home match in September, -1 in case West Bromwich Albion plays an away match in September and 0 in all other cases
$WestHamENGReferee5_i$	Variable that is 1 in case West Ham United plays a home match with ENG referee 5, -1 in case West Ham United plays an away match with ENG referee 5 and 0 in all other cases
$WillemIINETReferee7i$	Variable that is 1 in case FC Utrecht plays a home match with NET referee 7, -1 in case FC Utrecht plays an away match with NET referee 7 and 0 in all other cases

[2]Ten approximately equally sized groups are formed based on percentiles. Five positive and Five negative shape groups

[3]Six approximately equally sized groups are formed based on percentiles. Three positive and three negative shape groups