Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis

Nonlinear effects of seasonal home advantage and quality

Tom van Amstel (547102)

Supervisor:	Richard Paap
Second assessor:	JC van Ours
Date final version:	1st July 2023

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Abstract

As one of the most popular sports, football generates enormous amounts of money. Knowing how your team can gain an advantage is thus very important. Playing at home is seen as such an advantage. In numerous studies the home advantage is predicted using linear models, resulting in low R-squared values. In this research we therefore investigate whether nonlinear models are more effective in estimating the seasonal home advantage and quality of football teams in England. We use a data set, introduced in the paper by (Peeters and van Ours, 2021), containing information of all matches in the period from 1974 until 2018, including seasonal summary statistics of clubs in the first four leagues of England. We perform a Random Forest and analyse the pattern of the partial dependence plots of the drivers of seasonal home advantage and team quality. Next, we add nonlinear transformations of the explanatory variables and perform a LASSO regression to remove the less important variables. Lastly, we compare the predictive performance of the LASSO and Random Forest to the performance of the linear model. The results of our LASSO regressions, show that the nonlinear transformations of the variables do not drive the seasonal home advantage, but do drive the team quality. While the R-squared improves when including these transformations, the predictive performance does not increase when performing the LASSO regression. The Random Forest does also not result in a better predictive performance.

1 Introduction

Football is one of the most popular sports in the world. With 3.5 Billion fans all around the globe the amount of money that is involved in the sport is huge. Winning some games will generate a revenue of millions of euros. Knowing what causes your team to have an advantage is thus very important nowadays. Playing a match at home is seen as such an advantage. There are several studies on the causes of a so called "home advantage". These studies assume that the relation between home advantage and its drivers is linear. In this paper we relax this assumption and allow for potential nonlinear relations between the home advantage and its drivers. We estimate the seasonal home advantage and quality of football teams in England using models that allow for nonlinear relationships, like a Random Forest and adding nonlinear transformations of variables in combination with LASSO.

The drivers of home advantage, is a subject discussed in several papers. Peeters and van Ours (2021) discuss this subject in their paper. In their research the authors analyze the home advantage generated by teams in English football. By calculating seasonal home advantage for teams separately, the authors find that some teams do have a higher seasonal home advantage than others. By performing linear regressions, the authors analyse the effects of potential drivers of seasonal home advantage and team quality. They conclude that the home advantage is driven by the crowd attendance and being in the possession of an artificial pitch. They also found in their research that the quality of a team is positively driven by the wage of the players and negatively driven by being relegated recently.

The paper by (Pollard, 1986) shows a review of the drivers of home advantage. By investigating the crowd attendance, referee bias and other psychological drivers this paper gives a clear detailed overview of the home advantage in several leagues. The authors do not provide convincing evidence that the crowd attendance is a driver of home advantage. They conclude that the familiarity with conditions is a important driver of home advantage. The paper by (Clarke, 1995) matches the statements concerning the influence of crowd attendance. The authors did not find evidence that the home advantage is driven by crowd size. Conversely the paper by (A.M Nevill a, 2002) found evidence that the referee's decisions are driven by the crowd in favour of the home-playing team.

In the paper by (Clarke, 1995) evidence is found that having an artificial pitch drives the home advantage positively. The authors proved the positive influence on a 0.1% level. A different paper by (Werlayne Leitea, 2022) also investigates the familiarity with the conditions by the home-playing team. The paper discusses the effect of a new stadium on the home advantage of teams. The authors conclude that there is not a significant difference in the home advantage when a team changes its stadium.

The papers mentioned above investigate the factors that potentially drive the home advantage and quality of a team in football using linear regressions. Nonlinear estimation methods are not yet used to estimate home advantage, but there are some papers that estimate the performance of football teams using estimation models that capture nonlinear relationships. In the paper by (Yoel F. Alfredo, 2019) a Random Forest is used to estimate the results of football games in the premier league. The Random Forest is also analysed in the paper by (Jaemin Lee and Lee, 2022). In this paper the authors investigate the change in home advantage after the COVID-break. By performing different machine learning models, including Random Forest, they analyse the effect of the COVID pandemic on expected score and goal difference. They concluded that the machine learning methods are not accurate in predicting football match outcomes. Finally, in a paper by (Charles South, 2020) the authors predicted the goal difference in college football using Random Forest and Lasso regression. They concluded that the predictive performance of these models is similar and outperforms other machine learning methods.

The nonlinear relationships between drivers of home advantage and quality are not yet discussed in previous literature, which makes this research a useful addition to the existing literature. An disadvantage of the linear regressions performed by (Peeters and van Ours, 2021) is that the R-squared of the models is very low. This means that the fit of the linear models is not very good. This can be explained by the fact that the seasonal home advantage and team quality could actually be predicted by nonlinear models and that the nonlinear relationships can thus not be captured by the models that are applied in previous studies. By allowing for these relationships the models could possibly generate a higher R-squared, improving the fit of the models. Analysing nonlinear models also allows us to investigate new potential drivers of seasonal home advantage and quality like the cross effects of existing variables. When these new drivers have a significant effect this is useful for football clubs, because they can then increase their advantage relative to opposing teams. Also for regulators of competitions it is useful to know what drives the home advantage of clubs, so they can change the rules such that the competition is as fair and exciting as possible.

Like in the paper by (Peeters and van Ours, 2021), we analyze the seasonal home advantage based on the goal difference and points earned in home and away games. We discuss some potential drivers of home advantage and quality and include these in a linear and nonlinear estimation framework.

In this paper we answer the main question of our research:
Is the utilization of a nonlinear model more effective in estimating the seasonal home advantage and quality of English football clubs when compared to a linear model?

We answer this question by using prediction models that allow for nonlinear relationships. In these models a lot of different possible nonlinear variables can be included. For this reason it is important to analyse which variables cause the model to be nonlinear. To analyse these variables we perform a Random Forest regression. A Random Forest is able to capture nonlinear relationships between variables, which makes it very useful in our research. We compare the predictive performance of the Random Forest with the widely used linear regression. To analyse the relation of the potential drivers with the seasonal home advantage and quality, we make partial dependence plots. These show the impact of a single explanatory variable on the predicted variable. By looking at the shape of these graphs we investigate whether these nonlinear relationships could possibly exist. Because these plots do not prove the nonlinearity of the models, we include numerous nonlinear transformations of explanatory variables in the regression. This leads to a large amount of explanatory variables. We therefore apply the LASSO regression in order to select the appropriate ones. The LASSO regression penalizes the predictors that are less important and shrinks their coefficients to zero. We then compare the predictive performance of the LASSO model to the linear model. By looking at the coefficients and their significance we can conclude whether the nonlinear explanatory variables actually improve the predictive models.

We conclude that the Random Forest does not outperform the linear model when estimating the seasonal home advantage and quality. Some of the Partial Dependence Plots do show a nonlinear pattern, indicating nonlinear relationships between the variables. Lastly, the results of the LASSO regression show that the nonlinear transformations of the variables do not really affect the seasonal home advantage, but they do affect the quality of the football teams.

The rest of this paper is structured as follows. Section 2 discusses the structure of our data set and how it is modified so it fits with our research. Section 3 explains the derivation of the seasonal home advantage and team quality, which models we use to estimate the home advantage and team quality and how we compare them to one another. Section 4 depicts our results and their interpretations. Section 5 presents our conclusions and discusses further possible extensions to our research.

2 Data

In this section we describe which data set we use in this research and how we have changed the data to make it useful for our research. In the first part of this section we discuss the data that is used in the original paper by (Peeters and van Ours, 2021). In the extended research we modify the data set of the original paper, which we discuss in the second part of this section.

2.1 Original Data

In the paper by (Peeters and van Ours, 2021), the authors make use of panel data. They merged three different data sets together into one data set, which they use to analyse the seasonal home advantage in English football. The three data sets can be divided into different groups: games, managers and finance. The first group consists of information of all football games of the first 4 leagues in England in the period from 1974 until 2018. It provides the date of the match; the result; the goals scored by both teams; which manager was coaching the team and the crowd attendance at the particular game. The second data set consists of all managers. It provides the manager's names and a corresponding unique number, called the manager id. The last data set consists of seasonal information for each club. This contains information about: the division in which the club played that particular season; an overview of all match statistics; total wage of the players; average attendance; stadium capacity and the year in which the stadium was build. Merging these three sets together results in one data set, containing 182,938 observations, which provide detailed information of every match played in the period from 1974 until 2018 with seasonal summary statistics for every team.

The authors analyze the seasonal home advantage, so they need to modify the data set. They calculate the total results per season. This results in a data set of 4140 observations, which provide seasonal information about all clubs in the first four leagues of England. The authors also remove the managers name from an observation when the manager did not finish the particular season. Lastly, they add a dummy variable that indicates whether a team has an artificial pitch.

2.2 Construction of alternative variables

In our extended research we use the modified data set discussed above. The data set contains a lot of information we do not need in our research, so we first filter this information. We use 4 dependent variables, which are 2 measures for seasonal home advantage and 2 measures for the quality of teams. The derivation of those variables is shown in the methodology section. The explanatory variables from the original data set we use to estimate the seasonal home advantage and quality of teams are listed below.

- $relatt_{ijt}$, this variable shows a ratio of the average attendance of club i in league j in season t and the average attendance of all clubs in league j and season t.

- $stadage_{it}$, this variable shows how much time has past in years since the stadium has been build.

- ap_{it}, a 0/1 dummy variable, which shows whether the home playing team plays on artificial pitch or not.

- $lagprom_{it}$, a 0/1 dummy variable, which shows if a team is promoted to a higher division previous season.

- $lagrel_{it}$, a 0/1 dummy variable which shows if a team is relegated to a lower division previous season.

- $relwage_{ijt}$, this variable shows a ratio of the wage of a club i and the average wage of the clubs in league j in season t.

- $manobs_{it}$, this variable shows the total amount of experience the manager of team i in season t has. The experience of a manager is the total amount of seasons he has finished completely. When the manager leaves a club mid-season it is not counted as an extra year of experience.

After filtering these variables out of the original data set, we modified the data set such that it is useful in our research. We added the variable $manwins_{it}$. This variable shows the total amount of competition wins the manager of team i in season t has achieved in the period from 1974 until 2018. A competition win is counted when a team finishes on rank 1 of its division at the end of the season. We also added three 0/1 dummy variables for the divisions: $Div2_{it}$, $Div3_{it}$ and $Div4_{it}$. This results in a total of 11 explanatory variables. The original data set has some observations with missing values regarding the wage of players. When estimating a prediction model these missing values can cause misleading results. We therefore delete all observations where the information about the wage is missing. The data for our research now consists of 3337 observations.

3 Methodology

In the methodology section we first discuss how we derive the values of seasonal home advantage and team quality, which are similar to the ones derived in the paper by (Peeters and van Ours, 2021). We also discuss how they formulate their regressions. In the remaining part of the methodology we discuss our own research and the models we use. We discuss the measures we use to test the predictive performance of our models. Next, we discuss how we estimate the Random Forest and how we use this estimation method to show signs of nonlinearity in the estimation of the seasonal home advantage and team quality. To actually show the effects of implementing nonlinear transformations of variables, we perform a LASSO regression. We perform the linear regression and LASSO regression using Eviews. The Random Forest is performed in R studio.

3.1 Calculating Seasonal Home Advantage and Quality

In this part of the methodology we discuss how the seasonal home advantage and quality are derived. These calculations are the same as in the paper by (Peeters and van Ours, 2021).

In the paper by (Peeters and van Ours, 2021), the authors use two kind of measures to calculate the seasonal home advantage. The first measure is based on the values of Home Point Difference and Away Point Difference.

$$HPD_i = (\#homewins_i - \#homelosses_i) \times 3. \tag{1}$$

$$APD_i = (\#awaywins_i - \#awaylosses_i) \times 3. \tag{2}$$

In Equation1 and 2 the calculations of Home Point Difference and Away Point Difference are shown, respectively. When calculating these values the points earned by a draw are not included. The other measure is based on the values of Home Goal Difference and Away Goal difference.

$$HGD_i = HomeGoPro_i - HomeGoAg_i. \tag{3}$$

$$AGD_i = AwayGoPro_i - AwayGoAg_i. \tag{4}$$

The derivation of the Home Goal Difference and Away Goal Difference is shown in Equation 3 and 4, respectively. These Equations are used to calculate the difference between the goal difference in home and away matches, separately. Using these measures we can calculate the seasonal home advantage. The home point/goal difference of a team depends on its quality and home advantage. The formula for both the point and goal difference is similar, so in the upcoming equations we refer to both home differences in Equation 1 and 3 as HD_i and the away differences in Equation 2 and 4 as AD_i. The HD_i and AD_i depend on the quality of team i and of the $N-1$ other teams it plays against. The other factor that drives them is the seasonal home advantage of team i. After rewriting some equations explained in the paper by (Peeters and van Ours, 2021) we get the following equations:

$$HD_i = Nq_i + (N-1)h_i. \tag{5}$$

$$AD_i = Nq_i + h_i - N\overline{h}. \tag{6}$$

In Equation 5 and 6 the calculation of HD_i and AD_i are shown, respectively. Lastly, we derive the total home advantage of all teams, H. After rewriting Equation 5, we derive the formula of H.

$$H = \frac{\sum_{i=1}^{N} HD_i}{N-1}. \tag{7}$$

In Equation 7 the calculation of the total home advantage is shown. These formulas are rewritten again to find the derivation of quality and seasonal home advantage of team i.

$$h_i = \frac{HD_i + AD_i - H}{N - 2}. \tag{8}$$

$$q_i = \frac{HD_i - (N-1)h_i}{N}. \tag{9}$$

In Equation 8 and 9 the calculations of seasonal home advantage and team quality are shown, respectively. Equation 8 and 9 can be calculated for both the point differences and the goal differences.

3.2 Quantitative Analysis

In this section we describe how (Peeters and van Ours, 2021) analyse the drivers of home advantage. They perform a linear regression with the relative seasonal home advantage as dependent variable. This relative home advantage is calculated by subtracting the average league seasonal home advantage from the seasonal home advantage.

$$h^r_{ijt} = \alpha_i + \beta x_{ijt} + \gamma_j. \tag{10}$$

In the regression in Equation 10 the α_i captures the fixed club effects. x_{ijt} Contains the explanatory variables, in the paper the authors use the following: Relative attendance, artificial pitch, promotion, relegation, relative wage. An explanation of these variables can be found in the paper by (Peeters and van Ours, 2021).

The authors perform a total of 8 regressions relating to the seasonal home advantage, 4 using the point difference and 4 using the goal difference as dependent variable. For both home advantage measures they perform regressions including and excluding relative wage and fixed club effects.

In the paper the authors also perform 4 regressions relating to the quality of the teams. Some of the explanatory variables are similar to the variables in the regression relating to home advantage. In the regression with team quality as dependent variable they left out the relative attendance as a possible driver of the quality of the teams. This is due to the fact that the authors think the causality is more likely to work the other way around. In the first 2 regressions the quality is based on the goal difference in the particular season and in the other 2 regressions the quality is based on the points difference. The regressions for both cases are done including and excluding fixed club effects.

3.3 Fixed Team Effects

In this part of the paper the fixed team effects are discussed. In some of the regressions in the paper by (Peeters and van Ours, 2021) the fixed effects of the teams are included. To make a good comparison between the different estimation models, we want to include the fixed effects for every model. We perform a partial regression to include the fixed effects in our data. The Frisch-Waugh theorem, as stated in (Heij et al., 2004), describes how the variables are affected by these effects. By applying this theorem on the data, both the dependent variables and the explanatory variables are modified. We estimate the models in this paper using this data including fixed effects.

3.4 Performance Measures

An important part of our research is comparing the predictive probability of our different models. A model with a better predictive performance contains more useful coefficients, which results in a more precise interpretation of the estimation model. To measure the performance of our prediction models we split our data into a training sample and test sample. We use a commonly used distribution of observations of 80% for our training sample and 20% for our test sample. We modify this distribution slightly so the end of the training sample matches the end of a season. This results in the training sample covering the period from 1974 until 2007, including 2699 observations. The test sample covers the period from 2008 until 2018, including 638 observations.

We use two performance measures, which are derived from the standard errors of our forecasts. The first performance measure we use is the Root Mean Squared Error (RMSE). The RMSE is calculated by taking the square root of the mean of the difference between the actual value of the dependent variable and the prediction.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y_i} - y_i)^2}. \tag{11}$$

The second measure based on which we compare predictive performance of the estimation models is the Mean Absolute Error (MAE). This measure is calculated by taking the mean of the absolute value of the difference between the actual value of the dependent variable and the estimated value.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y_i} - y_i|. \tag{12}$$

3.5 Linear regression

The first and most simple regression we perform in this research is a linear regression. We perform two regressions with different kind of dependent variables. The first regression has the seasonal home advantage from Equation 8 as dependent variable, which is based on the goal difference and point difference. Like in the paper by (Peeters and van Ours, 2021), the average seasonal home advantage of the corresponding league is subtracted from the seasonal home advantage. The second regression has the quality of the teams from Equation 9 as dependent variable, which again is based on the goal and point differences. Like in the paper by (Peeters and van Ours, 2021), the regressions include different explanatory variables.

$$h^r_{ijt} = c + \beta x_{ijt} + \epsilon_{ijt}. \tag{13}$$

$$q^r_{ijt} = c + \beta x_{ijt} + \epsilon_{ijt}. \tag{14}$$

The regression with seasonal home advantage as dependent variable is shown in Equation 13. This Equation is similar to Equation 10, but this regression does include extra explanatory variables. The explanatory variables included in this regressions are explained in the Data section. In Equation 14 the regression with team quality as a dependent variable is shown. Like in the original paper we also left the relative attendance out of this regression. The causality is more likely to work the other way around, so attendance is a result of higher quality. The team fixed effects are included in data set x_{ijt} and the dependent variables h^r_{ijt} and q^r_{ijt}. We estimate the coefficients β_{ijt} of this model using the training sample. The remaining observations are used to measure the performance of the linear model.

3.6 Random Forest

The Random Forest is a famous method in machine learning introduced by (Breiman, 2001). Before explaining the Random Forest we need to clarify the definition of a decision tree.

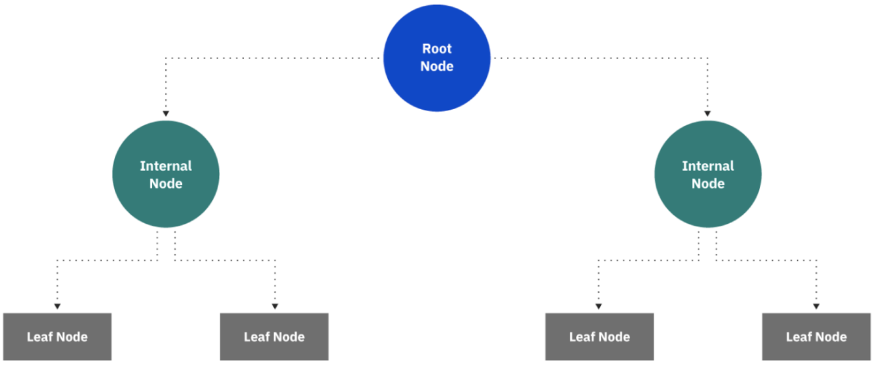


Figure 1: A representation of simple decision tree

In Figure 1 a simple representation of a decision tree is shown. A decision tree can be used to estimate a dependent variable, for instance y_i , where $i = 1, ..., N$. Suppose we estimate y_i using explanatory variable X_{1i}. A decision tree starts with a statement in the root node. For instance $X_1 < m_1$, where m_1 is a value that splits the values of X_1. Depending on if this statement is true or false we go down the left or the right node, respectively. If the statement is true we go down to the left node, which contains another statement, $X_1 < m_2$, where $m_2 < m_1$. Again based on this statement we go down to the left or the right node. We go down like this until we reach the end of the tree, which is called the leaf. This is done for all N values X_{1i}, resulting in each leaf containing the values of X_{1i} and the corresponding y_i in a certain interval of m. In these leafs a final prediction value is derived by calculating the mean of all y_i in the particular leaf. When we add an extra variable X_{2i}, we make several decision trees, like the one discussed above, starting in each leaf of the existing tree.

In order to make good predictions we need to implement values in the statements that split the nodes in an optimal way. We base these split-points on the optimal value for the Gini-coefficient. The Gini-coefficient shows the impurity of a node in the decision tree. The impurity is the chance that a random variable is classified incorrectly.

$$Gini(t) = 1 - \sum_{i=1}^{n}(p_{i|t}^2) \tag{15}$$

In the Equation 15 above, the Gini-coefficient of node t is calculated. In this formula n is total number of classes, which represent a specific value of the dependent variable. $p_{i|t}$ Is the fraction of observations that belong to class i. When a node has a low impurity it means that the observations are divided well. We want to minimize the impurity of the nodes in the decision tree.

When multiple variables are included in a decision tree it becomes more complex. The decision tree runs very deep, which results in the model being sensitive to the training data. This leads to overfitting and high variance. To lower the variance of these prediction models Breiman (2001) introduced "Bootstrap aggregation", also known as bagging. We take B random samples from the data set, with replacement. This means that we do not include all observations of the training sample in the so called "bootstrap samples". The samples can also include duplicate observations. Using these bootstrap samples B, different decision trees are trained on the data independently. When predicting the dependent variable, we take the mean of the outcomes of all bootstrap samples.

In a Random Forest for each bootstrap sample a decision tree is made. For every leaf node in the tree we take a random subset of variables $m \leq p$, where p is the total amount of explanatory variables. Among these m variables we choose the optimal variable and split-point. We repeat this until the leaf nodes have the minimum node size. The node size is the amount of values a leaf contains. The random selection of variables results in low correlation between the trees. By taking the average of the different decision trees with low correlation, the variance is reduced substantially (Trevor Hastie, 2009).

To estimate the Random Forest we use the training sample. The remaining 20% of the observations is used to test the performance of the Random Forest and is therefore called the test sample. It is useful to apply a Random Forest when estimating the seasonal home advantage and team quality, because a Random Forest regression captures possible nonlinear relationships between variables. We estimate the Random Forest regression using the first 2699 observations. We then calculate the root mean squared error and mean absolute error using the test sample.

To investigate which factors cause the nonlinearity in a model for estimating seasonal home advantage and team quality, we analyse the Partial Dependence Plots (PDP) of the explanatory variables. A Partial Dependence Plot shows the relation between a predictor variable and a target variable. The PDP shows how a change in the predictor variable changes the target variable, while keeping other variables in the regression at a fixed value. For a linear model the PDP has the pattern of a straight line. The PDP's do not give any proof for nonlinearity, but the plots can give an indication of the nonlinear relations between the dependent and explanatory variables.

3.7 LASSO Regression

In order to investigate the nonlinear effects on seasonal home advantage and team quality. We analyse the effect of nonlinear transformations of the explanatory variables included in the linear regression in Equation 13 and 14. By including the quadratic variables and the cross effects the total amount of explanatory variables increases substantially. To decrease this large amount of variables we perform the LASSO regression.

The LASSO-regression is a machine learning method that selects parameters in a prediction model, introduced by (Tibshirani, 1994). It consists of a tuning parameter λ, which eliminates unnecessary explanatory variables.

$$min_{\tilde{\beta}}\left(\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{I}(\tilde{h}^{r}_{ijt}-\sum_{k=1}^{K}\tilde{x}'_{ijtk}\tilde{\beta}_k)^2+\lambda_L\sum_{k=1}^{K}|\tilde{\beta_k}|\right). \tag{16}$$

Equation 16 shows the LASSO regression formula. In this formula $\tilde{h}^{r}_{ijt}$ is the seasonal home advantage of club i in league j in season t in terms of points per game or goal difference. $\tilde{x}_{ijtk}$ Is the value of the predictor k for club i in season t. β_k Is the parameter of predictor k. In this equation T is the total number of seasons in the training sample, I the total amount of clubs in season t, K the total number of predictors and λ_L is the tuning parameter, which penalizes the the unnecessary variables. N Is the total amount of observations in the training sample, which is in our case 2699. This formula is also used for the LASSO regression with team quality as a dependent variable.

The value of the tuning parameter has impact on the amount of explanatory variables that are included in the LASSO regression. A higher λ_L results in more coefficients being shrunken to zero. In this research we use k-fold cross validation to calculate the optimal λ_L, as explained by (Efron and Tibshirani, 1993). This method prescribes that we define a set of values for λ_L. Next, we divide the data into k parts. These parts consist of k-1 training parts and 1 part for testing. Next, we calculate the mean squared error using this training and testing sample for one of the values of λ_L. We repeat this k times, using every part as the testing sample once. We calculate the cross validation mean squared error for all values of λ_L. The λ_L, which has the lowest mean squared error is used in our LASSO regression. It takes a lot of time to do these calculations, so we perform this method in Eviews. To estimate the tuning parameter we select the variable selection method with k-fold cross validation as an extra option. In our research we use $k = 5$.

The LASSO method penalizes the variables that have small influence on the dependent variable. This means the variables that are most important are included in the regression. The coefficients are estimated again using the first 2699 observations. Also for this method the predictive performance of the regression is analysed and compared to the other models discussed in the previous sections.

4 Results

In this section we discuss the results of our research. In the first part we discuss the results of the original paper. In the second part we discuss the results of our own research and discuss whether it improves the models used in the paper by (Peeters and van Ours, 2021).

4.1 Replication

In this part we discuss the relevant results of the original research by (Peeters and van Ours, 2021).

Table 1: This table shows a summary of the relevant results from the paper by (Peeters and van Ours, 2021). Column 1 and 2 represent the results of the regressions regarding the seasonal home advantage and 3 and 4 show the results of quality regressions. In all regressions fixed club effects are included.

	Home advantage		**Quality**	
	Goal diff.	*Points diff.*	*Goal diff.*	*Points diff.*
Relative Attendance	0.071	0.073		
Artificial Pitch	0.350***	0.510***	-0.008	0.037
Promoted	0.036	0.044	0.042	0.049
Relegated	-0.028	-0.061	-0.088***	-0.112***
Relative Wage	-0.028	-0.053	0.609***	0.828***
Observations	3337	3337	3337	3337
R-Squared	0.005	0.005	0.209	0.186

[1] The significance of the coefficients is shown by *,** and ***, representing a 10%,5% and 1% significance level.

Table 1 shows the results of the regressions of the possible drivers of seasonal home advantage and team quality. In the original paper the authors performed 8 regressions for seasonal home advantage and 4 regressions for team quality. The tables for these results are shown in the appendix in Table 6 and 7. In Table 1 we only displayed the results that are comparable to our own research. In the first two columns, the regressions of the seasonal home advantage are shown including relative wage as a variable and including the club fixed effects. These results are including the division dummies, for which the coefficients are not shown in the table.

It appears that only having an artificial pitch has a positive significant effect on the seasonal home advantage. When a team has been relegated or has a high relative wage, this affects the home advantage in a negative way. Except for the variable indicating the presence of an artificial pitch, all coefficients are insignificant. The R-squared values of the regressions are very low, indicating that the models do not fit well to the data.

Column 3 and 4 of Table 1 show the results of the regressions with the quality based on goal difference and points difference as dependent variable. It appears that relative wage has a positive significant effect on the quality of teams on a 1% significance level in both cases. Again when a team has been relegated this has a negative and significant effect on the quality of a team, also on a 1% level. The R-squared increased a lot in comparison to the seasonal home advantage regressions. This means the fit of these models is much better than the first two models.

4.2 Adding extra explanatory variables

Table 2: This Table shows the results of the linear regression. The Table shows the coefficients of the explanatory variables. In the last row the R-squared of the regression is shown.

	Home advantage		**Quality**	
	Goal diff.	Points diff.	Goal diff.	Points diff.
C ($\times 10^{-2}$)	-0.113	-0.214	-1.646**	-2.3575**
RELATT ($\times 10^{-1}$)	0.853*	0.690		
RELWAGE	-0.060	-0.062	0.608***	0.806***
STADAGE ($\times 10^{-3}$)	0.769	0.784	-0.607	-0.828
AP	0.366***	0.546***	-0.086	-0.078
LAGREL	-0.015	-0.051	-0.085***	-0.102***
LAGPROM	0.025	0.031	0.027	0.026
MANOBS ($\times 10^{-2}$)	0.061	0.189	2.074***	3.024***
MANWINS ($\times 10^{-1}$)	-0.026	-0.210	0.212**	0.334**
DIV2	-0.027	-0.027	0.153***	0.209***
DIV3	-0.050	-0.043	0.150***	0.208***
DIV4 ($\times 10^{-1}$)	-0.422	-0.072	2.240***	2.978***
R squared	0.008	0.007	0.269	0.243

[1] The significance of the coefficients is shown by *,** and ***, representing a 10%,5% and 1% significance level.

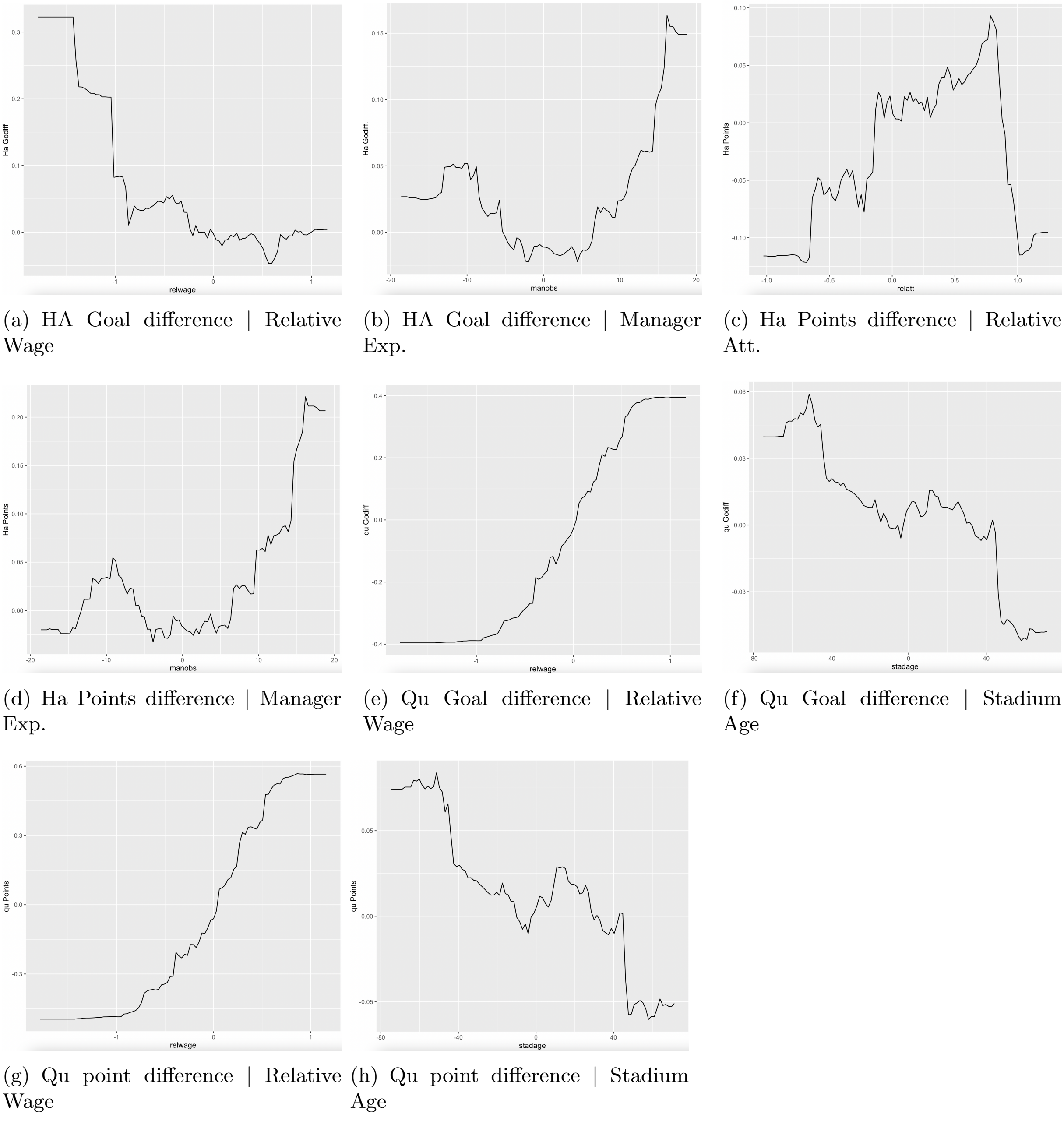
The results of the linear regression are shown in Table 2. In this table the coefficients of the explanatory variables are displayed including their significance. It appears that the amount of significant coefficients is low when the seasonal home advantage is regressed on the explanatory variables. Only the dummy variable AP, which shows whether the home team has an artificial pitch or not, does have a significant positive coefficient at a 1% level. When we compare this to the results in Table 1, it appears that the same coefficients are significant. The coefficients we added do not have a significant effect on the seasonal home advantage. Despite the insignificance of those coefficients the R-squared has increased slightly in our regressions, which could be a result of a higher amount of variables.

The coefficients of the regression with team quality as a dependent variable are significant at a 1% level for 6 variables in both cases and for one variable on a 5% level. It stands out that the coefficient for the variable AP is not significant anymore. Another difference between the regression of seasonal home advantage and the regression of team quality relates to the coefficients of the divisions. A lower division seems to have a negative impact on the seasonal home advantage but a significantly positive impact on the team quality. The R-squared of the regression with the quality as a dependent variable is much higher than the regression with home advantage as dependent variable. The R-squared has also increased substantially compared to the value in Table 1. A higher R-squared means a higher fraction of the total variance of the dependent variable is explained by the model. Thus a higher R-squared is desirable for a prediction model. This means the model regarding the seasonal home advantage as a dependent variable is probably bad in predicting the actual home advantage.

4.3 Random Forest

Because a Random Forest captures nonlinear relationships between variables, we use this method to analyze the nonlinearity in the prediction models for seasonal home advantage and team quality. Only a fraction of the PDP's shows a nonlinear relationship with the home advantage and quality of teams in England.

Figure 2: These Figures show the Partial Dependence Plots of the explanatory variables, which have a non-linear pattern, in relation to the home advantage.



(a) HA Goal difference | Relative Wage

(b) HA Goal difference | Manager Exp.

(c) Ha Points difference | Relative Att.

(d) Ha Points difference | Manager Exp.

(e) Qu Goal difference | Relative Wage

(f) Qu Goal difference | Stadium Age

(g) Qu point difference | Relative Wage

(h) Qu point difference | Stadium Age

These PDP's that show nonlinear patterns are shown in Figure 2. In Figure 3 - 6 in the Appendix the complete Partial Dependence Plots of our explanatory variables are shown.

In Figure 2 the PDP's of some of the explanatory variables in relation to the seasonal home advantage and team quality are shown. The graphs do not show a straight line, which we would expect when the variables would have a linear relationship with the seasonal home advantage. We can thus conclude that these graphs show a nonlinear pattern. Graph (a) shows the relation between relative wage and the home advantage. This graph follows the pattern of a quadratic function. This is also the case for Graph (b), which represents the relation of manager experience and the home advantage. The other graphs are harder to interpret, because they do not seem to follow a certain pattern. Graph (c) and (d) show the PDP's of the home advantage based on the points earned in home and away games. Graph (c) shows the pattern of a quadratic function. The PDP regarding the relative wage of the home advantage based on the point difference shows a similar pattern as in Graph (a). This is not the case for the PDP regarding manager experience in Graph (d), which shows the pattern of a different kind of nonlinear function.

Graph (e) and (f) show the PDP's of the quality based on the goal difference. These graphs are harder to interpret, because they show a pattern that looks more like a linear function. For these graphs it is hard to say whether the graph shows an exponential or linear growth. By implementing both the quadratic variable and the normal variables in the LASSO function we can analyse if the relation between those variables and the team quality is linear or not. In Graph (g) and (h) the PDP's of the quality based on the points difference are shown. Like in Graph (e) and (f) it is difficult to conclude what pattern the PDP's follow. Because these PDP's only give us an indication what the relation between the variables is, we cannot yet conclude whether the relations are nonlinear.

4.4 LASSO regression

The results of the LASSO regression are shown in Table 3. In this table the results of the LASSO regression including the explanatory variables, their squared values and their cross effects are displayed.

The LASSO regression with seasonal home advantage as dependent variable only contains a few coefficients that have not been shrunken to 0. For the home advantage based on goal difference only the coefficient of AP is significant at a 1% level. It is also the only non-transformed variable that is included in the regression. The other coefficients that have not been shrunken to zero, are some of the cross effects of the explanatory variables. The only cross effect that is significant is the $LAGREL \times MANOBS$ variable, which is significant on a 5% level. Because of the significant coefficients, it is possible for the model to be nonlinear. The second column of table 3 shows the coefficients of the LASSO regression including the seasonal home advantage based on the points difference. Again the coefficient of the variable AP is significant at a 1% level. In this regression also $LAGREL$ is included as a non-transformed variable. The remaining variables in this regressions are squared variables or cross effects. In this regression the coeffi-

Table 3: This Table shows the results of the LASSO regression. In this regression the squared variables and the cross effects of the variables are included. In the last row the R-squared of the regression is shown.

	Home advantage		**Quality**	
	Goal diff.	Points diff.	Goal diff.	Points diff.
C ($\times 10^{-2}$)	-0.132	0.056	-2.019**	1.649
RELWAGE	0	0	0.609***	0.857***
AP	0.497***	0.933***	-0.082	0
LAGREL	0	-0.076*	-0.058	0
LAGPROM	0	0	0.032	0
MANOBS	0	0	0.021***	0.034***
MANWINS	0	0	0.028***	0.031**
DIV2	0	0	0.144***	0.106***
DIV3	0	0	0.128***	
DIV4	0	0	0.201***	0.090**
RELATT*RELWAGE	0	-0.136		
RELATT*AP	-0.349	0		
RELATT*MANOBS	0	-0.010		
RELWAGE_SQUARED	0	0.088	0.058	0.173**
RELWAGE*AP	0	-0.675	-0.288	0
RELWAGE*LAGREL	0	0	-0.105	-0.277***
RELWAGE*LAGPROM	0	0	-0.086	-0.121
RELWAGE*MANOBS ($\times 10^{-1}$)	-0.040	-0.052	0.168***	0.325***
RELWAGE*MANWINS	-0.055	0	0.114**	0.223***
STADAGE*AP	0.020	0.060**	0	0
LAGREL*MANOBS	-0.014**	-0.017*	0	0
LAGREL*MANWINS	0	0	-0.031	0
LAGPROM*MANOBS ($\times 10^{-2}$)	0	0	-0.752*	0
LAGPROM*MANWINS	0	0	0.057	0.087
MANOBS_SQUARED ($\times 10^{-2}$)	0	0	-0.089***	-0.156***
Rsquared	0.009	0.011	0.286	0.264

[1] The significance of the coefficients is shown by *,** and ***, representing a 10%,5% and 1% significance level.

cients of $STADAGE \times AP$ and $LAGREL \times MANOBS$ are significant at a 5% and 10% level, respectively. This is again a sign of nonlinear relationships in the estimation of home advantage. When comparing this LASSO regression to the linear regression in Table 2, it appears that the R-squared of the LASSO models is higher than the linear regression while containing a lower amount variables for both the cases. This means the goodness of fit of the LASSO models is slightly better than the fit of the linear models. This means this model captures the underlying relationships of the variables better, improving the predictive performance. The R-squared is still very low. As stated in the paper by (Ozili, 2022), a model needs at least a R-squared of 0.1 on the condition that a large fraction of explanatory variables has significant coefficients. This is not the case for the models with seasonal home advantage as a dependent variable. Therefore we cannot prove the nonlinearity based on this LASSO regression.

The last 2 columns show the coefficients of the LASSO regression with team quality as dependent variable. It appears that a lot less coefficients are shrunken to zero. In the two cases with different dependent variables, the variables that are included in the regression are similar to each other. In both cases the same non-transformed variables are significant at a 1% level as in the linear regressions. Apart from these non-transformed variables also some of the squared and cross effects have significant coefficients. $RELWAGE\_SQUARED$ Is significant at a 5% level for the points difference based quality. $MANOBS\_SQUARED$ Has a significant coefficient at a 1% level in both cases, but the coefficient is very low, which means it does not have a large influence on the team quality. Furthermore, a few cross effects of variables do have significant coefficients. Especially the variables $RELWAGE \times MABOBS$ and $RELWAGE \times MANWINS$, which are significant on a 1% level. Due to the high number of significant coefficients of the squared variables and cross effects of variables, it appears that there are multiple nonlinear relationships between variables. Again the R-squared of these models is higher than the R-squared of the linear models. Again this is caused by a better goodness of fit in the LASSO models. The R-squared of the models is 0.286 and 0.264, combined with the fact that over half of the explanatory variables has significant coefficients, we can conclude that this model does fit the data well.

4.5 Forecast Performance of the Models

Table 4: This table shows the values of the Mean Absolute Error of the different estimation models. The prediction sample contains the last 638 of the observations in the original data set.

	Home advantage		**Quality**	
	Goal diff.	*Points diff.*	*Goal diff.*	*Points diff.*
Linear Regression	0.381	0.582	0.331	0.493
Random forest	0.340	0.602	0.336	0.498
LASSO	0.381	0.583	0.330	0.493

Table 5: This table shows the values of the Root Mean Squared Error of the estimation models. Like the MAE these are calculated using the prediction sample of the last 638 observations of the data-set.

	Home advantage		**Quality**	
	Goal diff.	*Points diff.*	*Goal diff.*	*Points diff.*
Linear Regression	0.481	0.735	0.422	0.622
Random forest	0.501	0.765	0.429	0.625
LASSO	0.480	0.736	0.424	0.624

In Table 4 and 5 the values of the MAE and RMSE are shown, respectively. Table 4 shows that the Mean Absolute Error of the linear regression and the LASSO regression are approximately the same for all 4 of the cases. The Random Forest model has a higher MAE in all cases, which means that the Random Forest performs worse in predicting the seasonal home advantage and quality of football teams, but the differences are small. This is also the case for the values in Table 5, which are also higher for the Random Forest predictions. Based on the performance measures we can thus not conclude that the Random Forest outperforms the standard linear model. Because the Random Forest captures nonlinear relationships when predicting variables, better performance could be a sign of nonlinearity. Because the MAE an RMSE of the LASSO regression are so similar to the performance measures of the linear regression, we do not have prove that the nonlinear coefficients do have a large effect on the estimation of the seasonal home advantage. So based on the performance of the three models we cannot conclude that the estimation model of home advantage and quality is nonlinear.

In the paper by (Peeters and van Ours, 2021), the authors state that overtime the seasonal home advantage has a downward sloping trend. This would imply that the values of our training sample differ from the values in our testing sample. The prediction models do not account for such a downward sloping trend. This could affect the performance of our prediction models and explain why the higher R-squared of the LASSO regression does not imply a higher MAE and RMSE.

5 Conclusion

In this report, we have investigated the effectiveness of a nonlinear model when estimating the seasonal home advantage and quality of football teams in England. We extend the research by (Peeters and van Ours, 2021) and include several extra potential drivers of seasonal home advantage and team quality. To investigate the effectiveness of a nonlinear model we perform a Random Forest and discuss the Partial Dependence Plots of the drivers in relation to the home advantage and quality. In order to prove the effectiveness of nonlinear models we use a LASSO regression, including several nonlinear transformations of variables. By analysing the coefficients of the variables and comparing the performance of this model to the linear model, we have reached conclusion about the nonlinearity of seasonal home advantage and quality of teams in English football.

In our linear regression the three variables we have added do not seem to improve the model introduced by (Peeters and van Ours, 2021). Like in the original paper, only the variable indicating if the home-playing team had an artificial pitch is significant on a 1% level when we regress the home advantage on it's potential drivers. The stadium age, manager experience and manager wins do not have a significant effect on the seasonal home advantage. The R-squared of our model exceeds the R-squared of the linear model in the original paper, but it is still very low. When we regress the quality of the teams on the potential drivers we find that manager experience and manager wins have a significant effect on the team quality. The R-squared of these models also exceeds the R-squared of the linear models in the original paper. Because of the high amount of significant coefficients and the high R-squared, this model has a good fit to the data.

The Random Forest does not seem to outperform our linear model in predicting the seasonal home advantage. The MAE and RMSE of the Random Forest are slightly higher than their values for the linear regression. For some of the variables the Partial Dependence Plots have a nonlinear pattern, but combined with the bad performance of the model we cannot prove that estimating seasonal home advantage and quality is more effective using a nonlinear model.

When including the nonlinear variables in the LASSO regression, we find that the squared and cross effects of the drivers of home advantage do not generate significant coefficients. The R-squared does increase a small amount, but it is still to low to draw a conclusion about the influence of the variables included in the regression. The results of the regression of quality have shown that quality has several nonlinear variables that have significant influence on the quality of a team. The cross effects of relative wage with manager experience and manager wins seem to have a significant positive effect on the quality of a team. The goodness of fit also improves when including nonlinear variables to the model. The predictive performance of the LASSO regression does not improve for both seasonal home advantage and quality. We therefore conclude that, despite nonlinear variables having a significant effect on seasonal home advantage and team quality, the utilization of nonlinear variables when estimating the seasonal home advantage and team quality is not more effective than using a linear model.

Estimating seasonal home advantage and quality of football teams using a nonlinear model is a complex problem. Because there are so many kinds of nonlinear variables, it is difficult to include the right ones in your estimation model. In further research other kinds of polynomial regressions can be performed. In our linear regression of seasonal home advantage the R-squared appeared to be very low. This means there are other factors that drive the seasonal home advantage of football clubs. In further research these other potential drivers of home advantage can be included in prediction models, resulting in a higher R-squared and thus better models. Some nonlinear transformations of variables did have a significant effect on the quality of football teams. In further research it would interesting to analyse how this effect can be explained. Lastly, future researches could focus on the nonlinearity in estimating seasonal home advantage and quality in competitions in different countries. By combining the results of several competitions we can reach a stronger conclusion.

References

A.M Nevill a, N.J Balmer b, A. M. W. b. (2002). The influence of crowd noise and experience upon refereeing decisions in football. *Psychology of Sport and Exercise*, pages 261–272.

Breiman, L. (2001). Random forests. *Machine Learning*, page 5–32.

Charles South, E. E. (2020). Forecasting college football game outcomes using modern modeling techniques. *Journal of Sports Analytics*, 6(1):25–33.

Clarke, S.R. Norman, J. (1995). Home ground advantage of individual clubs in english soccer. *The Statistician*, pages 509–521.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Oxford University Press.

Heij, C., de Boer, P., Franses, P. H., Kloek, T., and van Dijk, H. K. (2004). *Econometric Methods with Applications in Business and Economics*. Oxford University Press.

Jaemin Lee, Juhuhn Kim, H. K. and Lee, J.-S. (2022). A bayesian approach to predict football matches with changed home advantage in spectator-free matches after the covid-19 break. *Entropy*, 366(24).

Ozili, P. K. (2022). The acceptable r-square in empirical modelling for social science research. *Social Research Methodology and Publishing Results*.

Peeters, T. and van Ours, J. (2021). Seasonal home advantage in english professional football. *De Economist*, pages 107–126.

Pollard, R. (1986). Home advantage in soccer: A retrospective analysis. *Journal of Sports Sciences*, pages 237–248.

Tibshirani, R. (1994). Regression selection and shrinkage via the lasso. *Journal of the Royal Statistical Society*, pages 267–288.

Trevor Hastie, R. T. . J. F. (2009). Random forests. *The Elements of Statistical Learning*, page 587–604.

Werlayne Leitea, b, M. G. C. H. A. R. P. (2022). Home advantage in football after moving to a new stadium:evidence from european professional teams. *INTERNATIONAL JOURNAL OF SPORT AND EXERCISE PSYCHOLOGY*.

Yoel F. Alfredo, S. M. I. (2019). Football match prediction with tree based model classification. *Intelligent Systems and Applications*.

6 Appendix

6.1 Results of linear regression without extra variables

Table 6: This table shows the results of the 8 regressions of the possible drivers of home advantage. For all variables the value of the coefficients is shown. Column 1-4 show the regressions with goal difference as measure of home advantage and column 5-8 show the results of the regression with point difference as a measure. The results of column 3,4,7 and 8 show the results of the regression including the relative wage as a variable. Column 2, 4, 6 and 8 show the results of the regression including club fixed effects. Furthermore the amount of observations used in the regression and the R-squared are displayed.

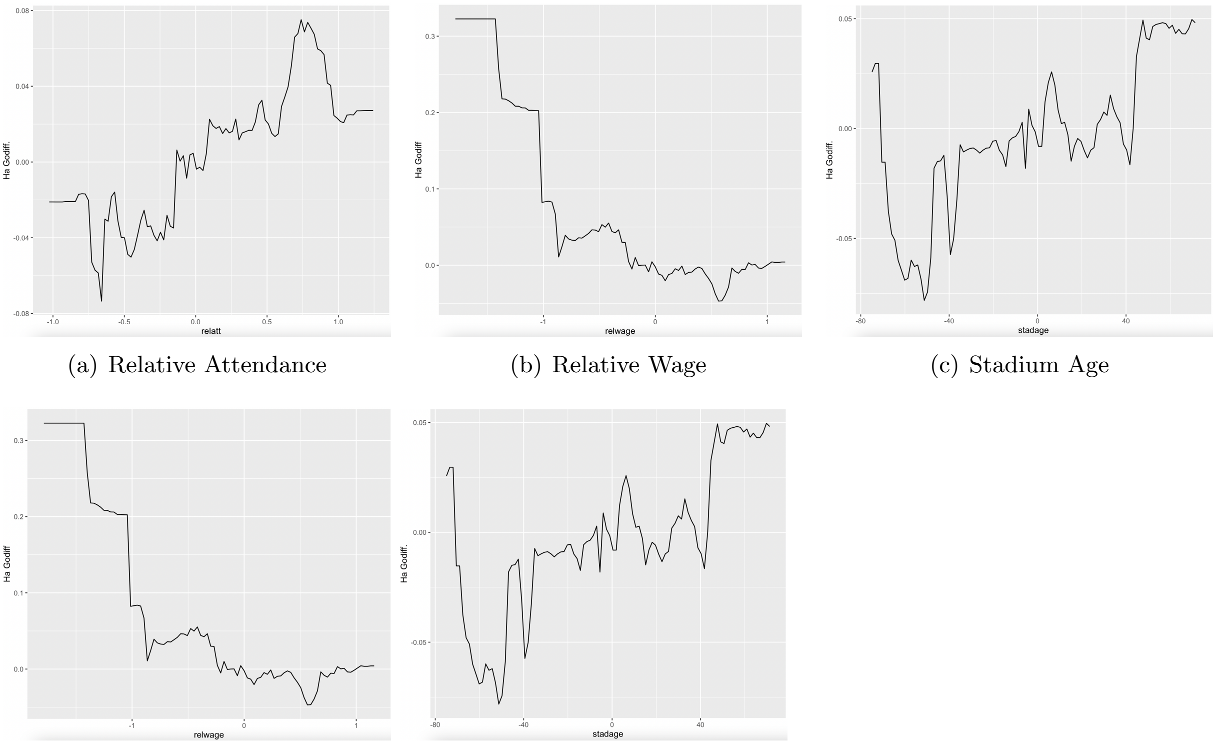
	Relative home advantage goal difference				Relative home advantage points difference			
	1	2	3	4	5	6	7	8
Relative attendance	0.0473**	0.0842**	0.0694**	0.0712	0.0458	0.1177**	0.0895**	0.0732
Artificial pitch	0.3976***	0.3511***	0.4256***	0.3495***	0.6611	0.5353	0.6819***	0.5103***
Promoted	0.0484	0.0456	0.0316	0.0360	0.0602	0.0534	0.0336	0.0440
Relegated	-0.0498*	-0.0425**	-0.0341	-0.02827	-0.0851**	-0.0729**	-0.0615	-0.0608
Relative wage			-0.0391	-0.0281			-0.0848*	-0.0532
Observations	4140	4140	3337	3337	4140	4140	3337	3337
R-squared	0.0071	0.0067	0.0073	0.0053	0.0069	0.0062	0.0079	0.005

Table 7: This table shows the results of the regression of the possible drivers of quality of teams in English football. In the table the coefficients and their significance are shown. The first 2 columns show the results of with quality based on goal difference as dependent variable and column 3 and 4 show the results of quality based on points as dependent variable. Column 2 and 4 show the results of the regression when including club fixed effects.

	Quality goal difference		Quality points	
	1	2	3	4
Relative wage	0.6823***	0.6089***	0.9210***	0.8276***
Artificial pitch	-0.1270	-0.0082	-0.1909	0.0365
promoted	0.0307	0.0418	0.0367	0.0487
relegated	-0.1079***	-0.0879***	-0.1406***	-0.1122***
observations	3337	3337	3337	3337
R-squared	0.2545	0.2089	0.2298	0.1857

6.2 Partial dependence plots of Random Forest

Figure 3: These Figures show the Partial Dependence Plots of the explanatory variables on the Home Advantage based on the goal difference.



(a) Relative Attendance

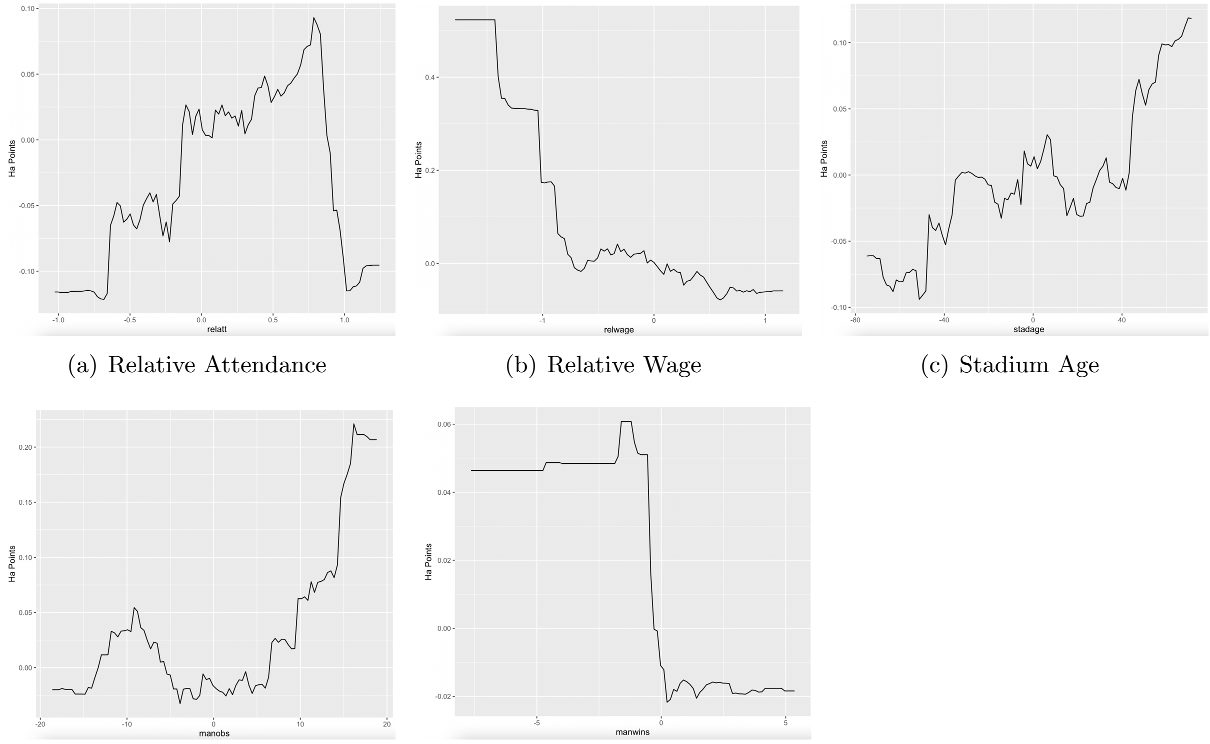
(b) Relative Wage

(c) Stadium Age

(d) Manager Experience

(e) Manager Wins

Figure 4: These Figures show the Partial Dependence Plots of the explanatory variables on the Home Advantage based on the points difference.



(a) Relative Attendance

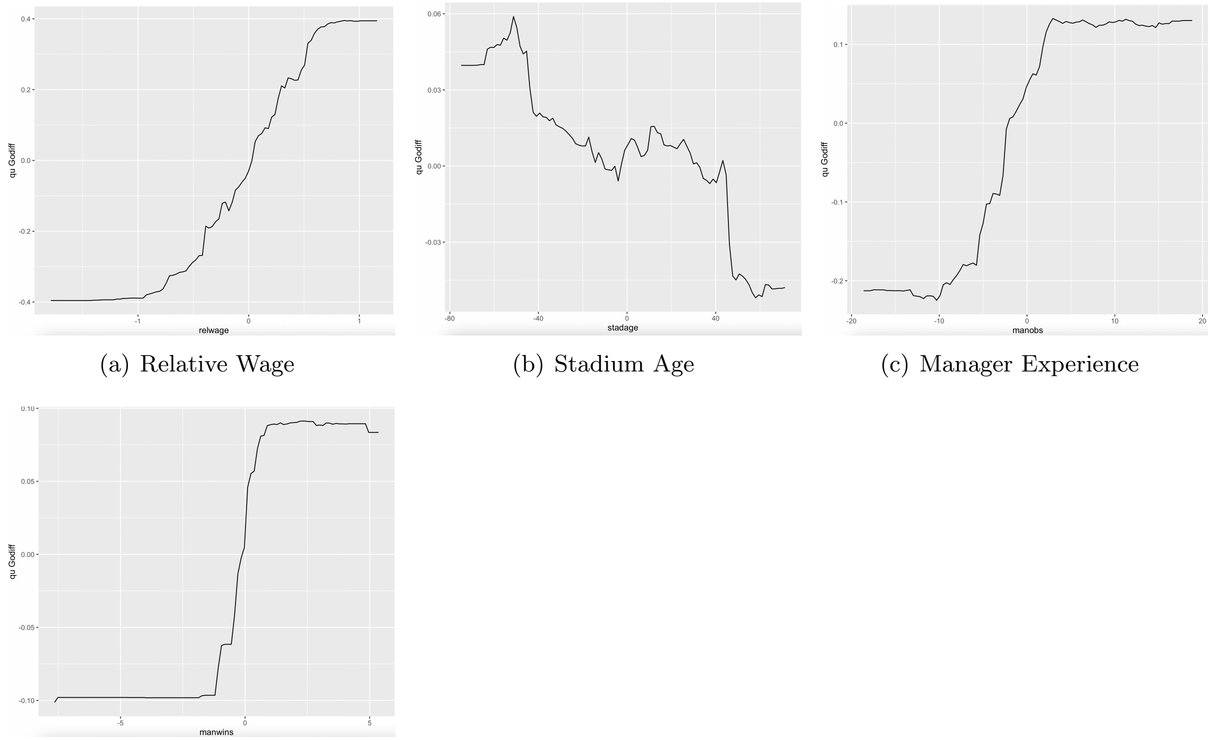
(b) Relative Wage

(c) Stadium Age

(d) Manager Experience

(e) Manager Wins

Figure 5: These Figures show the Partial Dependence Plots of the explanatory variables on the Quality based on the goal difference.



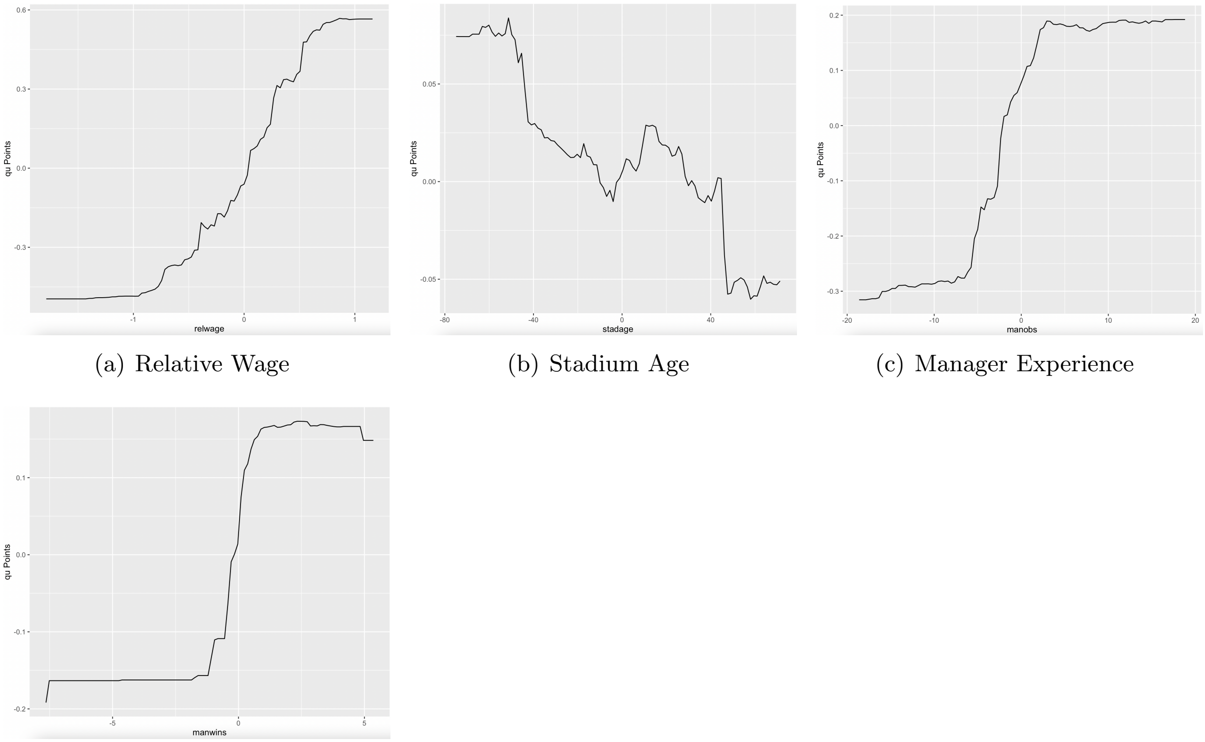
(a) Relative Wage

(b) Stadium Age

(c) Manager Experience

(d) Manager Wins

Figure 6: These Figures show the Partial Dependence Plots of the explanatory variables on the Quality based on the points difference.



(a) Relative Wage

(b) Stadium Age

(c) Manager Experience

(d) Manager Wins