Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis BSc^2

Home Sweet Home: Investigating Player-based Home Advantage in English Professional Football

Floris Jan van Fraassen (498628)

Supervisor:	J.C. van Ours
Second assessor:	R. Paap
Date final version:	1st July 2023

Abstract

This paper studies seasonal home advantage on an individual player basis in the English Premier League over the period 2009 to 2023. Since there are no standardized metrics to objectively measure individual player performance in football, a proxy is constructed in the form of home rating advantage. Home rating advantage per player is constructed using player ratings per game, which are measured on a scale of 1 to 10. The results show that the average home rating advantage over all players over all seasons equals 0.14 points. Moreover, this paper shows that home advantage on an individual player basis has decreased over the years. Furthermore, players differ substantially in the home advantage they enjoy. To investigate this difference further, possible drivers of home advantage per player per season are investigated using both a linear regression framework and a machine learning approach, specifically random forest regression. Home advantage per player is positively related to average home attendance, player quality and skill level and negatively related to age and being a goalkeeper or defender.

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.

1 Introduction

When a footballer steps onto the pitch, the pressure to perform is immense. Every pass, every shot, every tackle is scrutinized by fans, coaches, and pundits alike. But what happens when the same player steps onto the pitch in front of their home crowd? Does the familiarity of the stadium and the support of the fans boost their performance? And what happens when they play away from home? In this paper, I dive into the effect and causes of home advantage on individual player performance in matches and seasons. By analyzing data from English professional football, I aim to shed light on the elusive phenomenon of home advantage and its impact on the performance of professional footballers. The main research question in this paper is as follows:

What is the impact of home advantage on individual player performance in professional football?

Football is a constantly evolving sport, and in the past decade, we have seen many changes that have affected the game at every level. From the introduction of Video Assistant Referees (VAR) to the impact of COVID-19, football has undergone significant transformations that have altered the way teams play and the factors that contribute to their success. These changes have also had an impact on home advantage, with empty stadiums becoming the norm during the pandemic and VAR potentially influencing referee decisions. As a result, it is interesting to analyze how the effect of home advantage on individual player performance has developed over the years. Therefore the first subquestion is as follows:

How has the impact of home advantage on individual player performance changed over the years?

Showing the existence of home advantage on an individual player basis is one thing, but explaining what causes it is another. There are many factors at play that may contribute to the effect of home advantage. Is it the familiarity of the stadium, the support of the fans, the position of the player in the field or the age of the player in question? For this subquestion, I dive into what factors contribute to the home advantage effect. Therefore the final subquestion of this paper is as follows:

What factors and player characteristics contribute to home advantage on an individual player basis?

The remainder of this paper is structured as follows. Section 2 discusses the relevant literature for the research problem. Section 3 introduces the data used in the quantitative analysis of home advantage. Section 4 deals with the methodology of analyzing home advantage on an individual player basis. Section 5 contains the results of the study. Finally, Section 6 summarizes and concludes this paper.

2 Literature Review

A substantial amount of research has been conducted on home advantage in professional football. The general consensus from this research is that home advantage is both real and statistically significant. Pollard (1986) investigated the historical progression of home advantage in the English Football League since its inception in 1888 and revealed a striking level of consistency in home advantage until the year 1984. The study also revealed that local derbies and the FA Cup exhibited less of a home advantage effect. Additionally, the research indicated that factors such as crowd size and travel fatigue were not significant contributors to home advantage. Finally, the impacts of team tactics, bias among referees and familiarity with local conditions were inconclusive. After Pollard (1986), several papers have examined home advantage in English professional football. Barnett and Hilditch (1993) researched the impact of playing on a pitch made from artificial grass rather than natural grass, reporting an extra home advantage of 0.28 points and 0.31 goals per match.[1] Clarke and Norman (1995) investigated seasonal home advantage for all English football teams between 1981 and 1990, their study reveals substantial variation among teams and over time. Furthermore, Bray et al. (2003) discovered that, on average, teams won 22% more games at home than away in their study on all the top four divisions of the English football league over the time period 1981 to 2000. Carmichael and Thomas (2005) argue that home advantage relates to differences in playing style, with teams playing more aggressively in home games compared to away games. Dawson et al. (2007) concluded that underdogs are prone to face disciplinary sanctions more frequently compared to title-favorites in the English Premier League, and home teams play more aggressively in front of large crowds, but they do not receive more disciplinary sanctions due to a home team bias. Johnston (2008), however, did not find evidence of referee bias affecting the home advantage. Boyko et al. (2010) claimed that home advantage in the English Premier League is affected by spectator attendance and referee decisions about penalties and yellow cards, while Buraimo et al. (2010) concluded that there is a referee bias favoring home teams. Allen and Jones (2014) investigated Premier League matches from season 1992/1993 to season 2011/2012 and found no upward or downward trend in average home advantage over time. Furthermore, they found that teams positioned towards the lower ranks of the league table exhibited a more pronounced home advantage. Contrary to Attrill et al. (2008), shirt color did not appear to affect home advantage.

The phenomenon of home advantage in professional football has also been investigated in various other countries and in competitions and tournaments of international scope. For instance, studies conducted by Buraimo et al. (2010) focused on the Bundesliga in Germany and Armatas and Pollard (2014) on the Superleague in Greece. Garicano et al. (2005) find that referees in Spain tend to provide more additional playing time to the home team when they are closely losing. Furthermore, Pollard and Armatas (2017) explore home advantage in the qualification stage leading up to the FIFA World Cup, discovering that the degree of home advantage was most prominent in Africa and South America and least prominent in Europe. Ponzo and Scoppa (2018) focused on same-stadium derbies in Europe to eliminate the effects of travel distance and familiarity with the stadium. They found that home advantage relies on the crowd's support,

[1] In 1995 the use of artificial pitches was banned in the English Premier League.

which biases referee decisions in benefit of the home team. Krumer and Lechner (2018) investigated the German Bundesliga and discovered that midweek matches had less home advantage than weekend games due to smaller crowds and players' perception of the game's importance. Similarly, Goller and Krumer (2020) noted that the day of play affected home advantage in top European leagues. Meanwhile, kick-off times influenced home advantage in the group stage games of the UEFA Europa League, according to Krumer (2020). Van Ours (2019) found a home advantage of 0.33 points and 0.42 goals per match in Dutch professional football, with artificial pitches providing additional advantage. Furthermore, Van Damme and Baert (2019) studied various distance measures in European international football and revealed that altitude and crowd sizes affected home advantage. Amez et al. (2020) concluded that the second match of a knock-out clash did not provide a bigger home advantage, this is noteworthy for competitions such as the UEFA Europa League and UEFA Champions League.

Finally, Peeters and Van Ours (2021) studied seasonal home advantage in English professional football from season 1973/1974 to season 2017/2018. In their paper Peeters and Van Ours (2021) differentiate between absolute home advantage, experienced identically by all teams in a league, and relative home advantage, which varies between teams. Their analysis shows that absolute home advantage is significant, ranging from 0.59 to 0.64 points per game or 0.44 to 0.46 in terms of goal difference. Moreover, there are considerable differences in the relative home advantage enjoyed by clubs, which are positively linked to the variation in attendance within the team and the use of an artificial pitch. Despite large absolute attendance differences across divisions, absolute home advantage remains consistent across divisions. Lastly, Peeters and Van Ours (2021) observe a considerable decrease in absolute home advantage over time that affects all divisions similarly.

2.1 Scientific Relevance

The scientific relevance of this paper lies in its focus on the home advantage for individual players on a team, as opposed to the team as a whole. Previous studies have mainly focused on the impact of home advantage on the performance of teams (in the form of points or goals). This study however, looks at the potential benefits that individual players may have in terms of their performance when playing at home. Furthermore, this study seeks to explain which determinants are important in explaining home advantage on an individual player basis. By filling this gap in the literature, the paper provides a more comprehensive understanding of the concept of home advantage in professional football. This research is significant because it adds to the current knowledge on the impact of the home advantage, and may have important implications for how sports teams approach home and away games.

2.2 Economic Relevance

The economic and practical relevance of this paper is threefold. Firstly, understanding the factors that contribute to home advantage can help teams and coaches develop strategies to maximize their chances of winning games. For example, if the size of the crowd is found to be a significant factor, teams could try to increase ticket sales or create a more lively atmosphere in their stadium to boost the home advantage. Secondly, home advantage can have financial implications. For instance, teams that perform well at home are more likely to attract fans, which can result in increased ticket sales, merchandise sales, and sponsorship deals. Therefore, understanding the determinants of home advantage can be important for the financial success of a professional football club. Thirdly, home advantage can also have implications for tournament organizers and policymakers. For example, if the type of pitch or the size of the crowd is found to benefit particular players or teams, policymakers could consider implementing changes to ensure fairness and equality in competition.

3 Data

The aim of this study is to investigate the effect of home advantage on individual player performance in professional football. Previous research has been able to calculate team quality and home advantage using end of season league tables through the methods of Clarke and Norman (1995). However, since the focus of this study is on individual player performance rather than team performance, this method is not applicable. To overcome this challenge, a proxy is constructed in the form of *home rating advantage* (HRA). Throughout this paper, the abbreviation HRA is used to represent the home rating advantage proxy. HRA per player is constructed using player ratings per game, which are measured on a scale of 1 to 10. The player ratings are obtained from the popular football data center WhoScored.[2] WhoScored calculates player ratings based on live automated algorithms using over 200 raw statistics, with each event valued based on its perceived impact on the match outcome. Positive events are weighed against negative events to determine each player's rating.

Player ratings are gathered on an individual match basis. That means that for any given match at least 22 ratings are gathered (2 teams, 11 players). Next, two averages are calculated. Firstly, the average home rating for player i in season t is calculated. Secondly, the average away rating for player i in season t is calculated. Then, HRA for player i in season t is calculated by subtracting the average away rating from the average home rating for the given player.

Ideally, the same 11 players would play every match in the season for any given team. This, together with the fact that football competitions are played in a round-robin format, would ensure that the HRA metric is not biased by factors like the quality of the opponents a certain player has faced over the course of the season. However, in reality this is not the case. To illustrate this issue further, consider this example.

A player from a top four team plays only 6 home matches, all against weak teams. Furthermore, that same player plays 7 away matches, all against top ten teams. What could happen is that the player receives high ratings for his home matches (since these matches were played against weak opponents) and relatively lower ratings for his away games (since these matches were played against strong opponents). This would lead to a high average home rating and a low average away rating. This would in turn bias the HRA metric for that player.

[2] https://www.whoscored.com/

To mitigate this problem, the data is wrangled. Observations are only kept if they have a pairwise counterpart. Meaning if player 1 played against team A at home, player 1 must also have played against team A away. This ensures unbiasedness in the HRA metric.

In addition to gathering player ratings, data about player characteristics is gathered from FUTBIN.[3] FUTBIN is a company that primarily focuses on providing information and services related to the popular video game franchise FIFA. It offers a range of features for FIFA player characteristics. Some examples include: the team to which the player belongs, the age of the player and the nationality of the player. This data is required for analyzing the factors that contribute to home advantage at an individual player level. A description of all the raw data gathered for this study can be found in Table 1.

All the data that is used comes from the top division in English professional football (The English Premier League). Furthermore, the time span of the data ranges from the 2009/2010 season to the 2022/2023 season. So, in total 14 seasons are analyzed. In the English Premier League (EPL) 20 teams compete. Overall 195,789 player ratings on a match level are considered.

All of the data discussed above is obtained using webscraping. For this purpose the Python programming language is used, as well as several libraries such as Selenium, BeautifulSoup4 and TOR. More details about the code can be found on GitHub.[4]

[3] https://www.futbin.com

[4] https://github.com/florisjanvf/Bachelor-Thesis

Table 1: Variable names and their description

Variable	**Description**	**Granularity**	**Source**
Rating	Rating the player was given for the match	Per match	WhoScored
Player	Name of the player	Per match	WhoScored
Season	Season in which the match took place	Per match	WhoScored
Team	Team for which the player plays	Per match	WhoScored
Opponent	Opponent in the given match	Per match	WhoScored
Home	Boolean variable indicating whether the match was played at home	Per match	WhoScored
Starter	Boolean variable indicating whether the player came on as a substitute or not	Per match	WhoScored
Quality	Metric for the quality of a player (given on a scale of 1-100)	Per season	FUTBIN
Nationality	Nationality of the player	Per season	FUTBIN
Age	Age of the player	Per season	FUTBIN
Position	Playing position in the field (e.g., attacker or defender)	Per season	FUTBIN
Length	Length of the player in CM	Per season	FUTBIN
Weight	Weight of the player in KG	Per season	FUTBIN
Attacking work rate	Categorical variable indicating player's activity and effort when attacking (low, medium and high)	Per season	FUTBIN
Defensive work rate	Categorical variable indicating player's activity and effort when defending (low, medium and high)	Per season	FUTBIN
Skill moves	Ordinal variable indicating player's skill and flair (5>4>3>2>1)	Per season	FUTBIN
Weak foot quality	Ordinal variable indicating player's weak foot quality (5>4>3>2>1)	Per season	FUTBIN
Average home attendance	Average home attendance for the team to which the player belongs	Per season	WorldFootball
Promoted	Boolean variable indicating whether the team to which the player belongs was promoted at the start of the season	Per season	Wikipedia
Captain	Captain of the player's team	Per season	Wikipedia
Manager	Manager of the player's team	Per season	Wikipedia

4 Methodology

4.1 Calculating Home Advantage per Player

As discussed in Section 3, the home advantage for player i in season t is measured using the HRA metric. HRA is calculated by subtracting the average away rating from the average home rating for the given player in the given season:

$$HRA_{it} = \frac{1}{\mid HG_{it} \mid} \sum_{m \in HG_{it}} Rating_{mit} - \frac{1}{\mid AG_{it} \mid} \sum_{m \in AG_{it}} Rating_{mit} \tag{1}$$

Here, HG_{it} denotes the set of home games for player i in season t. Similarly, AG_{it} denotes the set of away games for player i in season t. Finally, $Rating_{mit}$ denotes the rating player i was given in match m in season t.

For any given season t, the weighted average HRA over all the players is calculated as follows:

$$HRA_t = \frac{1}{\sum_{i \in P_t} W_{it}} \sum_{i \in P_t} HRA_{it} W_{it} \tag{2}$$

Here, P_t denotes the set of players that were given ratings in season t. Furthermore, W_{it} denotes the weight of the HRA observation for player i in season t. This weight is chosen as the number of home-away game pairs the player has played in the season. This approach takes into account the varying number of games played by different players. By assigning higher weights to players who have played more games, the weighted average gives greater importance to their observations, considering them to be more reliable and representative of the overall HRA for the season.

4.2 Investigating the Drivers of Home Advantage

To study the drivers of home advantage at an individual player level, the HRA of player i in season t is modelled as a function of a set of player characteristics x as follows:

$$HRA_{it} = f(x_{it}) + \epsilon_{it} \tag{3}$$

Here x_{it} is a vector of time-varying player-specific variables. The error term is represented by ϵ_{it}. Finally, $f(\cdot)$ denotes an unknown function that requires estimation through a specific method. In the model the following explanatory variables are considered:

- Average home attendance. This variable shows how many fans are attending the homes games of player i. More fans could give rise to a higher home advantage.

- Promotion. This variable indicates whether the team for which player i plays was promoted to the EPL at the beginning of the season. A player who plays for a team that was recently promoted may have an advantage if the opposition is unfamiliar with the grounds they are playing on

- Captain. This is a boolean variable indicating whether player i is captain of the team he plays on.

- Position. This categorical variable reflects the position player i has on the field: goalkeeper, defender, midfielder or attacker
- British. This boolean variable reflects whether player i is of British nationality. The logic being here that British players may have more home advantage than foreign players since they speak the English language and are more familiar with the country.
- Quality. This variable is an indication of the quality of a player. A high quality score implies that a player is of high quality. The logic being here, that high quality players tend to experience higher levels of jeering and targeting not only in away games but also in their home games. The recognition they receive intensifies the scrutiny and reactions from both home and away crowds.
- Skill moves. This variable shows how much flair and skill a player possesses.
- Age. This variables hows the age of player i in season t. The logic being here that as players get older they are less sensitive to fan behavior and more experienced with away team grounds. As such home advantage may be lower for older players.
- Trend. A trend variable, ranging from 1 to 14. Included in the model to examine potential linear changes in home advantage over the years
- Team. This variable indicates the team for which player i plays. Included in the model to control for team-fixed effects, since teams may differ inherently in the home advantage they enjoy.

More details on these variables are included in Table 1.

The unknown function $f(\cdot)$ in Equation 3 is estimated using two different methods: *Linear Regression* and *Random Forest Regression.*

4.2.1 Method 1: Linear Regression

In the linear regression framework the final model from Equation 3 now takes on the following form:

$$HRA_{it} = \alpha + \beta x_{it} + \epsilon_{it} \tag{4}$$

Here β is a vector of parameters. The *weighted least squares* (WLS) criterion is used to estimate the parameters in Equation 4. The weights are chosen as W_{it}, the number of home-away game pairs player i has played in season t. The reason for choosing WLS over *Ordinary Least Squares* (OLS) is similar to the reasoning behind calculating weighted average HRA for season t in Equation 2. This approach takes into account the varying numbers of games played by different players and assigns a higher weight to observations of players who have played more games, since these observations are considered to be more reliable and representative of overall HRA.

4.2.2 Method 2: Random Forest Regression

To explore if machine learning is a valuable tool in evaluating the drivers of home advantage on an individual player basis, the *random forest regression* (RFR) method is also used to estimate $f(\cdot)$ in Equation 3. Similarly to the linear regression method all observations are weighted using W_{it}.

RFR has several advantages over more traditional model-estimation methods like linear regression. Firstly, RFR is less susceptible to multicollinearity issues among explanatory variables compared to linear regression. Secondly, RFR has the ability to capture intricate non-linear relationships between the dependent variable and explanatory variables, which linear regression may not be able to capture effectively. One drawback of RFR is that the resulting model is often less interpretable compared to linear regression.

To address the aforementioned drawback of RFR regarding its interpretability, Explainable AI (XAI) tools are employed. These tools help provide insights into the model's behavior and enhance its explainability. One such tool is *variable importance* (VI), which measures the contribution of individual explanatory variables in predicting the target variable. It quantifies the impact or relevance of each variable, allowing researchers to identify the most influential factors. Additionally, *partial dependence* (PD) plots are utilized to depict the marginalized effect of individual explanatory variables. These plots illustrate how the target variable changes when a specific explanatory variable is varied while keeping other variables constant. By visualizing these relationships, researchers gain a better understanding of how each variable influences the model's predictions.

Another drawback of random forest regression, compared to linear regression, is its susceptibility to overfitting. To mitigate this issue, a method called cross-validation is employed. Specifically, the model is cross-validated using a 3-fold cross-validation approach. This technique partitions the data into three subsets, training the model on two subsets and evaluating its performance on the remaining subset. By iteratively rotating the subsets, the model's performance is assessed more reliably. Furthermore, a randomized grid search is used to optimize the model's hyperparameters and improve its generalization capabilities. The results of this randomized grid search process can be found in Appendix B: Hyperparameter Tuning Results.

In summary, to address the challenges of interpretability, XAI tools such as variable importance and partial dependence plots are utilized. To mitigate overfitting, the model undergoes 3-fold cross-validation and is optimized using a randomized grid search.

5 Results

The results section of this paper is split up in two parts. Firstly, HRA per player is computed, analyzed and aggregated considering various factors, including time, team and manager. Secondly, the drivers of home advantage at an individual player level are investigated using linear regression and random forest regression.

5.1 Home Rating Advantage

5.1.1 Home Advantage over the Seasons

Table 2 shows home advantage over the full sample. As a quick reminder, the concept of home advantage in this study is based on the utilization of home rating advantage (HRA). A distinction is made between the unweighted and weighted average HRA. The unweighted HRA average does not take into account the number of match-pairs an observation for HRA is based on. The weighted HRA does take into account the number of match-pairs an observation is based on. As illustrated in Table 2, the two averages are very similar. HRA over the whole sample is around 0.14 points. This is a modest result, given that player ratings are given on a continuous scale of 1–10. However, this result is still a clear indication that home advantage exists on an individual player basis.

Table 2: Player-based home advantage in the English Premier League, 2009–2023

Average type	**HRA**
Unweighted Average	0.143751
Weighted Average	0.143750

Figure 1 shows the development of home advantage in the English Premier League over the seasons. Home advantage was highest in the 2009/2010 season and lowest in the 2020/2021 season. The latter result is interesting, as this was the season which was played under strict COVID-19 crowd restrictions. Only a small number of spectators were allowed into the stadiums during this season. This is a first indication that the presence of fans has a large impact on the home advantage experienced by players. Furthermore, we see that there seems to be a downward trend in home advantage over the years right up to the 2020/2021 season. This result is in line with Peeters and Van Ours (2021), who also showed that there has been a secular decline in home advantage in English professional football over the years.

5.1.2 HRA per Team

In terms of fairness of the competition, total average home advantage over the seasons is not very useful. The round robin format of the EPL ensures that the total effect of home advantage is cancelled out across clubs. Instead, home advantage per team is analyzed. Table 3 gives an overview of the aggregated HRA per team in the EPL over the period 2009–2023. Note that

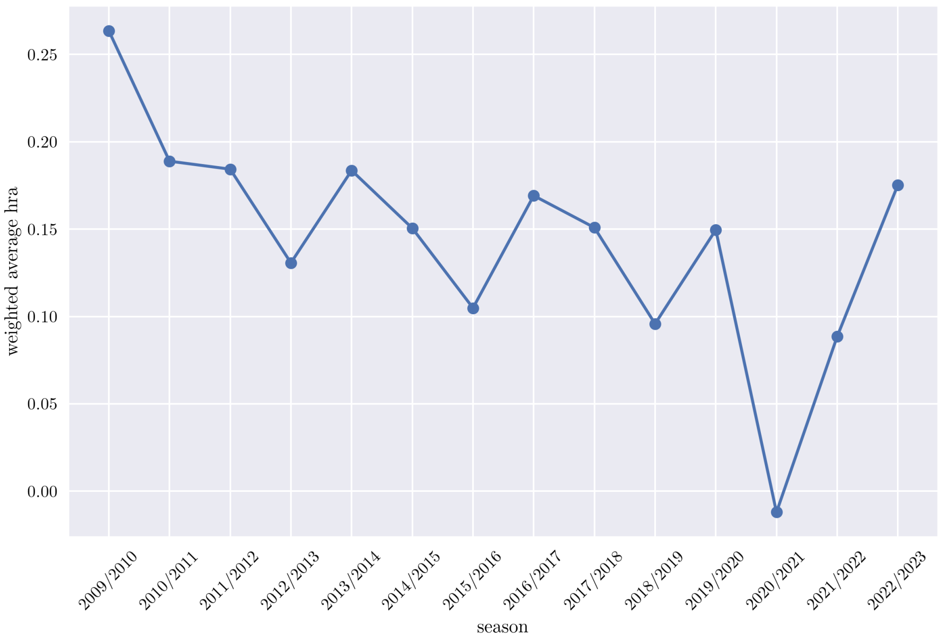

Figure 1: Home advantage in the English Premier League over the years, 2009–2023

this Table only includes teams that played at least seven seasons in the EPL over the given time period. There are stark differences in the HRA enjoyed by players on different teams. Liverpool has the highest weighted average HRA with a value of 0.212 and Crystal Palace the lowest with a value of 0.033. So, the home advantage that is on average enjoyed by Liverpool players is more than six times as high as that of Crystal Palace players in terms of home rating advantage.

5.1.3 Players and HRA

In addition to examining heterogeneity among teams, heterogeneity in home advantage among players is also investigated. This analysis is possible due to the fact that the HRA metrics are all player based. Table 4 shows HRA per player for all players that have played at least three seasons and thirty home-away game pairs. When comparing panel a. and panel b. in Table 4 big differences in HRA per player are observed. Furthermore, when looking at panel c. there is no clear indication that experienced players enjoy more or less HRA compared to the (unweighted) average HRA over the whole sample of 0.14. In Section 5.2 the underlying factors relating to heterogeneity in HRA among players will be investigated.

5.1.4 Managers and HRA

In a similar fashion to the previous subsection, HRA is analyzed but now per manager. Table 5 shows HRA per manager for all managers that have at least three seasons of experience in the EPL. From panel c. it becomes clear that there is no direct relation between the experience of a manager and the home advantage the manager experiences with his players. The weighted average HRA over the whole sample equals 0.14. For some managers in panel c. HRA is higher than 0.14 and for other managers HRA is lower than 0.14.

Table 3: Home advantage per team in the EPL, 2009–2023

Team	**Season count**	**Average HRA**	
		Unweighted	**Weighted**
Arsenal	14	0.202	0.194
Aston Villa	11	0.051	0.045
Burnley	8	0.067	0.068
Chelsea	14	0.189	0.189
Crystal Palace	10	0.036	0.033
Everton	14	0.174	0.171
Fulham	8	0.093	0.104
Leicester City	9	0.130	0.124
Liverpool	14	0.214	0.212
Manchester City	14	0.189	0.187
Manchester United	14	0.173	0.175
Newcastle United	12	0.163	0.167
Southampton	11	0.087	0.090
Stoke City	9	0.139	0.137
Sunderland	8	0.148	0.159
Swansea City	7	0.194	0.180
Tottenham Hotspur	14	0.210	0.200
West Bromwich Albion	9	0.087	0.085
West Ham United	13	0.134	0.133
Wolverhampton Wanderers	8	0.103	0.095

Note. Teams are only included if they have competed in at least 7 out of the 14 seasons under consideration.

Table 4: Home advantage per player in the EPL, 2009–2023

Player	**Experience**		**Average HRA**	
	Season count	**Game pairs**	**Unweighted**	**Weighted**
a. Top 10				
Leroy Fer	4	31	0.857	0.717
Benoit Assou-Ekotto	4	38	0.631	0.640
Mesut Ozil	7	57	0.654	0.593
Samir Nasri	6	49	0.559	0.575
Kevin De Bruyne	7	69	0.554	0.561
Joey Barton	3	35	0.540	0.547
Olivier Giroud	5	40	0.333	0.503
Dirk Kuyt	3	36	0.374	0.501
Clint Dempsey	4	47	0.355	0.466
Sadio Mane	8	93	0.530	0.464
b. Bottom 10				
Shane Duffy	4	37	-0.818	-0.549
Emiliano Martinez	3	53	-0.342	-0.340
Mame Diouf	4	31	-0.295	-0.263
Yves Bissouma	5	34	-0.305	-0.249
Adrian	5	38	-0.750	-0.246
Artur Boruc	4	44	-0.118	-0.238
Nick Pope	5	82	-0.263	-0.237
Tom Cleverley	8	40	-0.308	-0.234
Gabriel Magalhaes	3	39	-0.303	-0.213
Mathew Ryan	3	52	-0.197	-0.201
c. Most experienced				
Jordan Henderson	14	123	0.216	0.281
Jonny Evans	13	95	-0.029	-0.026
James Milner	13	82	0.370	0.247
David de Gea	12	188	-0.084	-0.049
Kyle Walker	12	131	0.081	0.073
Seamus Coleman	12	107	0.231	0.230
Hugo Lloris	11	156	0.117	0.102
Cesar Azpilicueta	11	136	0.121	0.135
Mark Noble	11	117	0.239	0.278
Gary Cahill	11	103	-0.041	0.054

Note. Players are only included if they have competed in at least three seasons and played at least thirty home-away game pairs.

Table 5: Home advantage per manager in the EPL, 2009–2023

Manager	**Season count**	**Average HRA**	
		Unweighted	**Weighted**
a. Top 10			
Ronald Koeman	3	0.237	0.283
Carlo Ancelotti	4	0.416	0.257
Arsene Wenger	9	0.234	0.256
Alex McLeish	3	0.212	0.236
Manuel Pellegrini	4	0.267	0.232
Steve Bruce	6	0.161	0.209
Roberto Mancini	4	0.257	0.208
Harry Redknapp	4	0.074	0.204
Brendan Rodgers	8	0.178	0.186
Rafael Benitez	5	0.180	0.183
b. Bottom 10			
Dean Smith	4	0.021	0.005
Graham Potter	4	0.075	0.018
Paul Lambert	4	0.010	0.018
Marco Silva	3	0.008	0.028
Sean Dyche	8	0.029	0.045
Ole Gunnar Solskjaer	4	0.102	0.091
Roy Hodgson	9	0.103	0.093
Nuno Espirito Santo	3	0.119	0.094
Eddie Howe	7	0.141	0.095
David Moyes	11	0.063	0.102
c. Most experienced			
David Moyes	11	0.063	0.102
Arsene Wenger	9	0.234	0.256
Sam Allardyce	9	0.158	0.171
Roy Hodgson	9	0.103	0.093
Tony Pulis	8	0.142	0.151
Jurgen Klopp	8	0.152	0.177
Sean Dyche	8	0.029	0.045
Brendan Rodgers	8	0.178	0.186
Roberto Martinez	7	0.164	0.136
Pep Guardiola	7	0.125	0.152

Note. Managers are only included if they have coached a team in at least three seasons.

5.2 Drivers of Home Advantage at the Player Level

In this section the drivers of home advantage at an individual player level are investigated. This analysis is based on the model presented in Equation 3. The model is estimated using two different methods: linear regression and random forest regression. Further elaboration and detailed information can be found in Section 4.2. Additionally, some sensitivity analysis is performed. The model-estimation methods are applied to two different datasets. Firstly, the methods are applied to the full dataset. Secondly, the methods are applied to a filtered dataset. Here the filtered dataset only contains observations where the number of home-away match pairs per player is at least 10. This dataset may give more accurate results, since now only players that have played at least half of the season are included in the analysis. This could potentially mitigate issues like the form of the player or injuries, since players with only one or two home-away match pairs have these matches spread out across the year.

5.2.1 Linear Regression

Table 6 shows the parameter estimates for Equation 4 estimated using the weighted least squares (WLS) criterion.

When considering the filtered sample parameter estimates by looking at Table 6 we see that being promoted has a small positive effect on home advantage for a player, however it is only significant at a 15% level. Being the captain on a team does not seem to have any contribution to the home advantage enjoyed by a player. Furthermore, it is interesting to see that goalkeepers and defenders enjoy significantly less home advantage than midfielders and attackers. There is no indication that British players experience more home advantage than foreign players. Average home attendance has a highly significant positive effect on the home advantage enjoyed by a player. There is no indication that higher quality players experience more home advantage than lower quality players. Additionally, it appears that skill moves and flair are positively related to a player's home advantage. However, this result is only significant at a 15% level. The age of a player does not seem to have any effect on the home advantage of a player. Finally, there is strong evidence that home advantage on an individual player basis has decreased over the period 2009–2023. The trend variable is highly significant and has a negative coefficient.

Table 6 also contains the full sample parameter estimates. Two differences are observed compared to the filtered sample parameter estimates. Firstly, the promoted variable is no longer significant. Secondly, the British boolean variable is now significant at a 15% level, and carries a positive effect to home advantage. The variables for the playing position in the field, the average home attendance, skill moves and the trend remain highly significant over both samples.

Finally, as a robustness check, the model in Equation 4 is also fitted using season-fixed effects instead of the trend variable, to analyze the stability of the parameter coefficients and their significance. The parameter estimates obtained from the model with season-fixed effects are not significantly different from those presented in Table 6. However, one notable change is that the promotion variable is no longer statistically significant for the filtered sample at a 15% level.

Table 6: Linear regression results for the relationship between seasonal home advantage per player and several player characteristics in the English Premier League, 2009–2023

Variable	**Sample**	
	Full	**Filtered**
Promoted	0.0194	0.0332*
	(0.0170)	(0.0213)
Captain	0.0165	0.0151
	(0.0194)	(0.0228)
Goalkeeper	−0.1550****	−0.1695****
	(0.0316)	(0.0403)
Defender	−0.0638****	−0.0895****
	(0.0196)	(0.0269)
Midfielder	-0.0022	-0.0197
	(0.0181)	(0.0250)
British	0.0167*	0.0041
	(0.0113)	(0.0146)
Average home attendance (x10,000)	0.0302****	0.0316****
	(0.0045)	(0.0057)
Quality	0.0013	0.0024
	(0.0016)	(0.0021)
Skill moves	0.0213***	0.0186*
	(0.0092)	(0.0118)
Age	0.0006	-0.0008
	(0.0015)	(0.0020)
Trend	−0.0085****	−0.0086****
	(0.0016)	(0.0020)
Observations	4438	1619
R^2	0.084	0.164

Note. Standard errors are in parentheses; HC1 robust standard errors are used; parameters for team-fixed effects and constants are not included; * p<0.15, ** p<0.10, *** p<0.05, **** p<0.01.

5.2.2 Random Forest Regression

Table 7 shows a comparison of the in-sample fit of between the LR-based model and the RFR-based model. Three metrics are used for this purpose: *mean squared error* (MSE), *mean absolute error* (MAE) and R^2. The results are interesting for two reasons. Firstly, the results clearly show that the explanatory power of the model becomes larger when switching to the filtered sample. This is shown by the higher R^2 values and lower MSE and MAE values. Secondly, Table 7 shows that the RFR-based model beats the LR-based model in terms of in-sample fit for every performance metric for both datasets. This shows that the explanatory power of the RFR-based model is higher than that of the more traditional LR-based model.

Table 7: In-sample fit comparison, linear regression vs random forest regression

Metric	**LR**		**RFR**	
	Full sample	**Filtered sample**	**Full sample**	**Filtered sample**
MSE	0.111	0.062	0.105	0.056
MAE	0.242	0.193	0.236	0.187
R^2	0.084	0.164	0.132	0.236

Table 8 and Table 9 show the variable importance (VI) of the ten most important explanatory variables for the RFR-based model for the full and filtered dataset respectively. For both datasets the average home attendance variable is the most important, accounting for about 18% of the total explanation power of the model. Furthermore, the set of top ten variables is the same in both Tables, however the order is a bit different. The five most important variables have a total explanation power of about 60% relative to the total model explanation power. While the order of the variables is slightly different between Table 8 and Table 9, the actual magnitudes of the variable importance per variable are not that different between the Tables. Notably, in the RFR-based model, player quality and age emerge as significant factors, both ranking within the top five most important explanatory variables in terms of variable importance. This is in contrast to the LR-based model results in Table 6, here player quality and age are both not statistically significant on any level.

Now that the variable importance of the explanatory variables in the RFR-based model have been investigated, the actual effect of the explanatory variables on home advantage is analyzed. Figure 2 and Figure 3 show the partial dependence plots for the most important variables in terms of variable importance. Additionally, the effect of the playing position of a player is examined. Both the full dataset and the filtered dataset are considered. Just like in the LR-based model, average home attendance is positively related to home advantage. The difference between a near empty stadium and a stadium packed with nearly 70,000 fans is roughly 0.08 HRA points. Furthermore, the relationship between average home attendance and home advantage exhibits a monotonically increasing pattern with a decreasing marginal effect. Moreover, higher player quality is related to higher home advantage for a player. This is especially true for the

Table 8: Variable importance of the random forest regression based model (full sample)

Variable	**Variable importance**	**Cumulative importance**
Average home attendance	0.178	0.178
Skill moves	0.112	0.290
Trend	0.111	0.402
Quality	0.109	0.511
Age	0.077	0.588
Goalkeeper	0.077	0.665
Midfielder	0.035	0.699
Defender	0.033	0.732
Attacker	0.025	0.757
British	0.021	0.779

Note. Variable importance per explanatory variable is calculated as the total reduction in the sum of squared errors brought by that variable (Gini importance).

Table 9: Variable importance of the random forest regression based model (filtered sample)

Variable	**Variable importance**	**Cumulative importance**
Average home attendance	0.171	0.171
Quality	0.110	0.281
Trend	0.106	0.387
Skill moves	0.096	0.483
Age	0.085	0.568
Goalkeeper	0.064	0.632
Attacker	0.033	0.665
Midfielder	0.032	0.697
Defender	0.028	0.725
British	0.019	0.744

Note. Variable importance per explanatory variable is calculated as the total reduction in the sum of squared errors brought by that variable (Gini importance).

top twenty percent players, here we see a steep increase in HRA from about 0.13 to 0.21. It seems that the rare players of exceptional quality enjoy a lot more home advantage than average quality players. The relation between the home advantage a player enjoys and his quality is also strongly non-linear, which may be one of the reasons the LR-based model did not find it to be statistically significant. The trend variable is negatively related to home advantage per player. This result is similar to the LR-based model. Furthermore, more skilled players enjoy more home advantage according to the RFR-based model. This result was also found in the LR-based model. Figure 3 also shows that as players get older, the home advantage they enjoy decreases. This relation is not monotonically decreasing for the whole sample however. The negative effect of player age on home advantage was also not found in the LR-based model. Finally, the partial dependence plot of the playing position reveals that attackers and midfielders enjoy more home advantage than defenders and goalkeepers. This result was also found in the LR-based model.

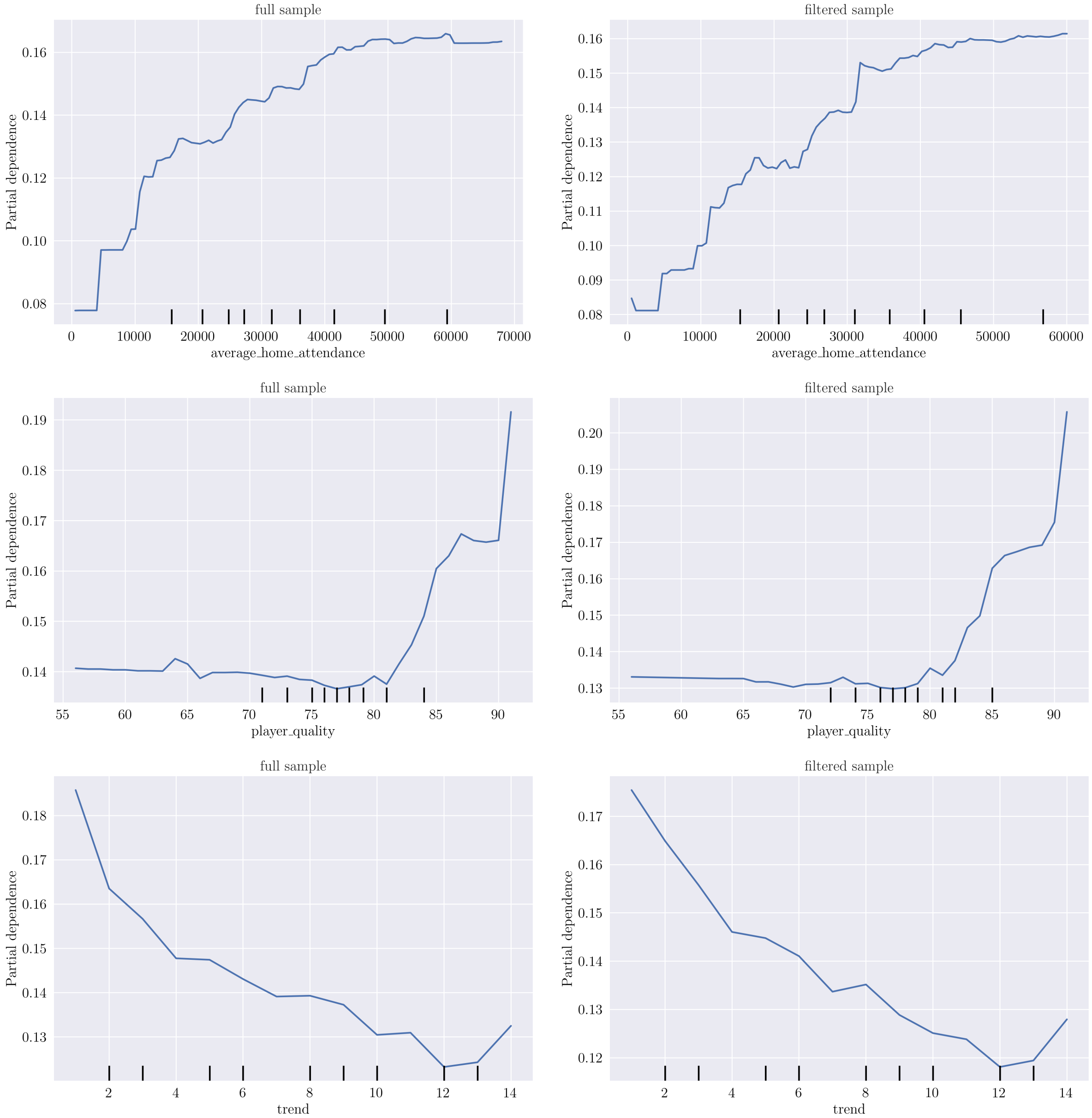

Figure 2: Partial dependence plots for the effect of average home attendance, player quality and trend on the home advantage a player enjoys

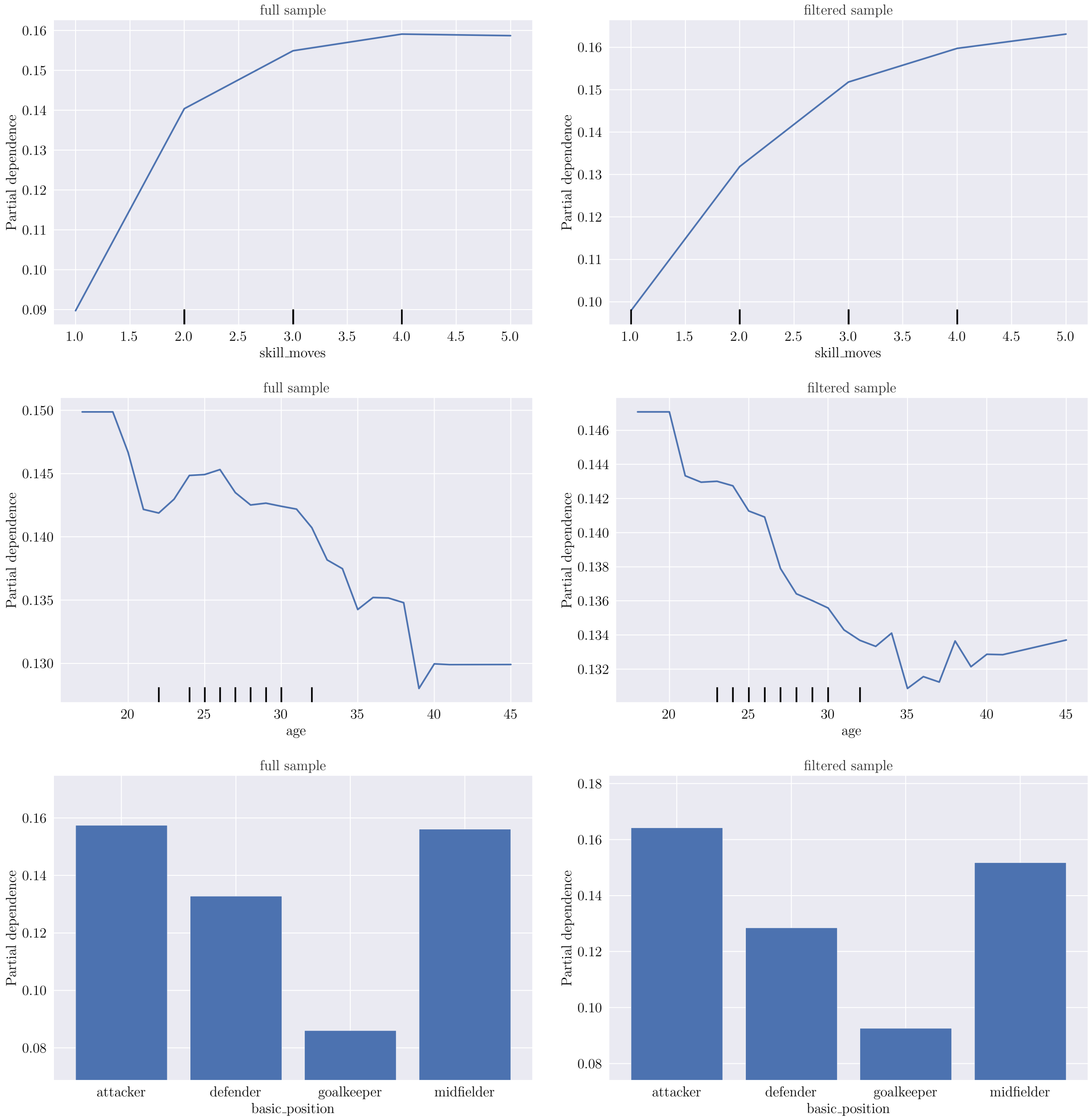

Figure 3: Partial dependence plots for the effect of skill moves, age and playing position on the home advantage a player enjoys

6 Conclusion

The existence of a home advantage has been firmly recognized within the realm of professional sports, with football serving as a prominent example. Prior research has examined diverse factors contributing to this phenomenon, encompassing the impact of crowd pressure on referee decisions, the psychological advantages derived from competing on familiar pitches, as well as the fatigue experienced by away players due to the distance they must travel. However, all of the aforementioned research has focused on home advantage on a team basis. As such, one aspect of home advantage that has not been researched extensively yet is the existence of home advantage on an individual player basis. This paper has investigated the presence and impact of home advantage on an individual player basis. Moreover, this paper has examined how home advantage has evolved over the years. Finally, this study has analyzed the drivers behind home advantage at an individual player level.

In this paper, 14 years of data relating to the performance of individual players from the English Premier League has been analyzed (2009–2023). More specifically, the data focuses on the home rating advantage (HRA) for a given player in a given season. Here home rating advantage was constructed as the difference between the average rating a player was given for his home games and the average rating a player was given for his away games. Ratings are measured on a continuous scale of 1–10. To investigate the drivers of home advantage at an individual player basis a wide variety of player characteristics and team characteristics were gathered and analyzed. All of the data was webscraped from several websites.

The analysis revealed that home advantage exists at an individual player basis. When averaging HRA over all players and seasons, a HRA of approximately 0.14 points was found. This result is modest compared to the study by Peeters and Van Ours (2021), which focused on home advantage at a team level and discovered an average home advantage of 0.64 points per game per team. It is important to note however that this study covered a shorter and more recent timespan. Furthermore, HRA and points gained per game are not directly comparable metrics for measuring home advantage. Still, the notion that *"The whole is greater than the sum of its parts"* seems to hold true in this context. Here the parts refer to the individual players' home advantage and the whole represents the team's home advantage. Next, the findings in this paper indicate a decrease in home advantage over the years. This conclusion is supported by a visual inspection of HRA over the seasons, as well as the inclusion of a trend variable in various models such as linear regression and random forest regression.

Substantial heterogeneity in the home advantage experienced by individual players was observed. Utilizing a linear regression (LR) framework, several factors that correlate with home advantage were identified. Specifically, average home attendance, promotion to the English Premier League and the skill level of a player exhibit positive correlations. On the other hand, being a goalkeeper or defender and the trend over the seasons exhibit negative correlations with home advantage.

Machine learning in the form of random forest regression (RFR) was also employed to investigate the drivers of home advantage per player. The results showed that the RFR-based model outperforms the LR-based model in terms of in-sample fit, as demonstrated by various goodness-of-fit metrics such as R^2, *MSE* and *MAE*.

Variable importance analysis revealed that the RFR-based model identifies mostly the same variables as significant in explaining home advantage per player as the LR-based model. The most influential factor is average home attendance, which positively relates to home advantage. Skill moves also exhibit a positive relationship, while the trend is negatively related to home advantage. However, the RFR-based model showed two significant differences compared to the LR-based model. Firstly, the RFR-based model identifies player quality as the second most important explanatory variable in the variable importance analysis. Secondly, the RFR-based model revealed that a player's age is also in the top five most important variables. Both player quality and age were not found to be significant in the LR-based model. Moreover the RFR-based model showed that the relationship between a player's quality and the home advantage he enjoys is highly non-linear. For the lower 50th quantile of player ratings, no discernible effect is observed on home advantage. However, for the upper 50th quantile, a positive exponential effect emerges. For instance, players in the top 20 percent of player ratings enjoy substantially more home advantage than other players. This non-linearity was not captured by the LR-based model. Additionally, age exhibits a negative relationship with the home advantage enjoyed by a player. This result was also not found in the LR-based model. These findings suggest that employing machine learning methods, such as random forest regression, provides additional value in understanding the drivers of home advantage at an individual player level, compared to the more traditional linear regression framework.

In conclusion, while this study sheds light on player-based home advantage in English professional football, it is important to acknowledge its limitations and identify potential avenues for further research. One notable limitation of this study is the omission of the number of years a player has played for a team in the model. Incorporating this variable could provide valuable insights into the impact of player familiarity with the home environment. Additionally, considering the wage of a player and the transfer fee as factors in the model could offer a deeper understanding of the financial dynamics influencing player performance and the home advantage effect. Furthermore, exploring the robustness of the findings across different countries (e.g., Spain) and competition types (e.g., the UEFA Champions League), would enhance the generalizability of the results and provide a broader perspective on player-based home advantage. Addressing these limitations and pursuing these future research directions will contribute to a more comprehensive understanding of the intricate dynamics underlying home advantage in professional football.

Appendix A: Statistical Significance HRA

The results described below were not included in the main text due to concerns regarding their reliability and robustness. The exclusion is based on the understanding that the Theoretical Framework below relies on the assumption of independent and identically distributed HRA observations per player in any given season. However, it is likely that this assumption does not hold. To exemplify this, consider the nature of football as a team sport, where a striker's performance is influenced by the performance of the wingers and midfielders in delivering crosses and creating chances for the striker. If the midfielders perform better in home games compared to away games, the home advantage of the striker may be contingent upon the home advantage of the midfielder, and vice versa. Therefore, it is plausible that HRA between players on the same team could exhibit dependence.

Theoretical Framework

To test if home advantage is statistically significant in season t, a t-test is used. But first, some statistical assumptions and theorems are introduced. Firstly, it is assumed that HRA for a given season t are independent and identically (i.i.d.) distributed over the players:

$$HRA_{it} \overset{\text{i.i.d.}}{\sim} \mathcal{D}(\mu, \sigma^2) \qquad \text{for } i \in P_t \tag{5}$$

Here $\mathcal{D}$ denotes an unspecified statistical distribution with mean μ and variance σ^2. Testing if HRA is significant then boils down to testing the following null hypothesis:

$$\begin{aligned} H_0 &: \mu = 0 \\ H_A &: \mu \neq 0 \end{aligned} \tag{6}$$

Here H_0 is that of no Home Rating Advantage. Using the Central Limit Theorem (CLT), it follows that the average HRA as in Equation (2) will be asymptotically normally distributed:

$$HRA_t \overset{\text{CLT}}{\approx} \mathcal{N}(\mu, \frac{\sigma^2}{\mid P_t \mid}) \tag{7}$$

In Equation (7), σ^2 is still an unknown variable. To solve, this we estimate it as follows:

$$\hat{\sigma}^2 = \frac{1}{\mid P_t \mid -1} \sum_{i \in P_t} (HRA_{it} - HRA_t)^2 \tag{8}$$

Rewriting Equation (7) and inserting the expression for $\hat{\sigma}^2$ gives our final estimator T and its distribution:

$$T = \sqrt{\mid P_t \mid} * \frac{HRA_t - \mu}{\hat{\sigma}} \sim \mathcal{T}(\mid P_t \mid -1) \tag{9}$$

Here $\mathcal{T}$ denotes the student t distribution with $\mid P_t \mid -1$ degrees of freedom.

Results

Table 10 shows the statistical significance of home rating advantage averaged over all the seasons and all the players. The p-value of 0.304 for the weighted average HRA is not small enough to conclude that home advantage is significant at the traditional significance levels of 1, 5 and 10 percent. However, the result still serves as an indication that home advantage may exist at an individual player level.

Table 10: Statistical significance player based home advantage in the English Premier League, 2009-2023

Average type	HRA	Standard deviation	t-statistic	p-value
Unweighted	0.144	0.512	0.280	0.390
Weighted	0.144	0.349	0.512	0.304

Table 11 shows the statistical significance of average home rating advantage per season. Based on the traditional significance levels of 1, 5, and 10 percent, there are no seasons in which HRA was statistically significant. However, the results still serve as an indication that home advantage may exist at an individual player level. In the 2009/2010 season for instance, weighted HRA equals 0.264 points with a p-value of 0.246. This result is significant at a 25% level.

Table 11: Statistical significance player based home advantage over the seasons in the English Premier League, 2009-2023

Season	Weighted average HRA	Weighted standard deviation	t-statistic	p-value
2009/2010	0.264	0.383	0.687	0.246
2010/2011	0.189	0.339	0.557	0.289
2011/2012	0.184	0.333	0.554	0.290
2012/2013	0.131	0.325	0.403	0.344
2013/2014	0.184	0.326	0.563	0.287
2014/2015	0.151	0.358	0.420	0.337
2015/2016	0.105	0.325	0.322	0.374
2016/2017	0.169	0.365	0.463	0.322
2017/2018	0.151	0.346	0.436	0.332
2018/2019	0.096	0.330	0.290	0.386
2019/2020	0.150	0.353	0.424	0.336
2020/2021	-0.012	0.333	-0.036	0.514
2021/2022	0.089	0.353	0.251	0.401
2022/2023	0.175	0.328	0.535	0.297

Appendix B: Hyperparameter Tuning Results

Table 12: Random forest regression hyperparameter tuning results (full sample)

Hyperparameter	**Description**	**Search space**	**Selected value**
n_estimators	Number of trees in the forest	[10, 300]	300
max_depth	Maximum depth per tree in the forest	[3, 15]	7
max_features	Maximum number of explanatory variables considered per split in any given tree	sqrt', 'log2', 'all'	sqrt'
max_samples	Number of observations used to train each tree (as a fraction of the total number of observations)	[0.5, 0.7, 0.9, 1.0]	0.7

Note. For the randomized grid search 100 iterations were performed; the random seed for training the random was set at 530; the random seed for the randomized grid search was set at 69.

Table 13: Random forest regression hyperparameter tuning results (filtered sample)

Hyperparameter	**Description**	**Search space**	**Selected value**
n_estimators	Number of trees in the forest	[10, 300]	245
max_depth	Maximum depth per tree in the forest	[3, 15]	7
max_features	Maximum number of explanatory variables considered per split in any given tree	sqrt', 'log2', 'all'	sqrt'
max_samples	Number of observations used to train each tree (as a fraction of the total number of observations)	[0.5, 0.7, 0.9, 1.0]	0.5

Note. For the randomized grid search 100 iterations were performed; the random seed for training the random was set at 530; the random seed for the randomized grid search was set at 69.

References

Allen, M. & Jones, M. (2014). The home advantage over the first 20 seasons of the English Premier League: Effects of shirt colour, team ability and time trends. *International Journal of Sport and Exercise Psychology*, *12*(1), 10–18.

Amez, S., Baert, S., Neyt, B. & Vandemaele, M. (2020). No evidence for second leg home advantage in recent seasons of European soccer cups. *Applied Economics Letters*, *27*(2), 156–160.

Armatas, V. & Pollard, R. (2014). Home advantage in Greek football. *European Journal of Sport Science*, *14*(2), 116–122.

Attrill, M., Gresty, K., Hill, R. & Barton, R. (2008). Red shirt colour is associated with long-term team success in English football. *Journal of Sports Sciences*, *26*, 577–582.

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

Boyko, R., Boyko, A. & Boyko, M. (2010). Referee bias contributes to home advantage in English premiership football. *Journal of Sports Sciences*, *25*(11), 1185–1194.

Bray, S., Law, J. & Foyle, J. (2003). Team quality and game location effects in English professional soccer. *Journal of Sport Behavior*, *26*, 319–334.

Buraimo, B., Forrest, D. & Simmons, R. (2010). The 12th man? Refereeing bias in English and German soccer. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *173*(2), 431–449.

Carmichael, F. & Thomas, D. (2005). Home-field effect and team performance: Evidence from English premiership football. *Journal of Sports Economics*, *6*(3), 264–281.

Clarke, S. R. & Norman, J. M. (1995). Home ground advantage of individual clubs in English soccer. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *44*(4), 509–521.

Dawson, P., Dobson, S., Goddard, J. & Wilson, J. (2007). Are football referees really biased and inconsistent? Evidence on the incidence of disciplinary sanction in the English Premier League. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *170*, 231–250.

Garicano, L., Palacios-Huerta, I. & Prendergast, C. (2005). Favoritism under social pressure. *The Review of Economics and Statistics*, *87*(2), 208–216.

Goller, D. & Krumer, A. (2020). Let's meet as usual: Do games played on non-frequent days differ? Evidence from top European soccer leagues. *European Journal of Operational Research*, *286*(2), 740–754.

Krumer, A. (2020). Testing the effect of kick-off time in the UEFA Europa League. *European Sport Management Quarterly*, *20*(2), 225–238.

Krumer, A. & Lechner, M. (2018). Midweek effect on soccer performance: Evidence from the German Bundesliga. *Economic Inquiry*, *56*(1), 193–207.

Peeters, T. & Van Ours, J. C. (2021). Seasonal home advantage in English professional football. *De Economist*, *169*, 107–126.

Pollard, R. (1986). Home advantage in soccer: A retrospective analysis. *Journal of Sports Sciences*, *4*(3), 237–248.

Pollard, R. & Armatas, V. (2017). Factors affecting home advantage in football World Cup qualification. *International Journal of Performance Analysis in Sport*, *17*, 121–135.

Ponzo, M. & Scoppa, V. (2018). Does the home advantage depend on crowd support? Evidence from same-stadium derbies. *Journal of Sports Economics*, *19*(4), 562–582.

Van Damme, N. & Baert, S. (2019). Home advantage in European international soccer: Which dimension of distance matters? *Economics: The Open-Access, Open-Assessment E-Journal (2007-2020)*, *13*(50), 1–17.

Van Ours, J. C. (2019). A note on artificial pitches and home advantage in Dutch professional football. *De Economist*, *167*, 89–103.