Modelling yellow cards in football games

word count: 13346

MSc Data Science (Statistics), University College London

Candidate Number XPBN7 (Supervised by Dr Abourashchi Niloufar)

Summer 2023

Abstract

In this project, we look at the highly stochastic football yellow card variable and aim to model them using statistical and machine learning methods. The dataset provided by Smartodds, a betting consultancy, allows us to use predictors such as market odds for goal supremacy and total goals, as well as others such as attendance, stadium capacity, etc. Through data exploration we notice generally weak but possibly linear relationships between the outcome and the predictors. We first employ 'simpler' linear models, then with Poisson log-link models, before incorporating mixed effects to account for repeated measurements of teams and referees. In a bid to improve predictability, we use random forests and gradient boosted models, and finally tested a combination of the latter with mixed effects through Sigrist's (2020) GPBoost model. In terms of predictive power on a test set, random forests and GPBoost performed the best, although the linear models LM and LMM (Linear Mixed Effects Model) were very close behind, while the 'Generalised' log-link versions and gradient boosted models vastly underperformed. Upon pre-processing the numeric predictors with Principal Component Analysis (PCA) to reduce the dimensionality of the feature space, and to mitigate multicollinearity, linear models had slight improvements which allowed them to overtake random forests. An analysis of residuals however show that the models systematically overpredict low values and underpredict higher values of yellow card, most likely as a result of omitted variables and the highly random nature of yellow cards.

Contents

1 Introduction

1.1 Football as a sport (a.k.a soccer, association football)

Football is a global sport and is enjoyed by 3.5 billion fans across the globe and 250 million players across 200 countries (Review 2023). Thus it has unsurprisingly churned out a huge betting market that is worth billions of dollars - ex-Footballer Joey Barton even quipped that 'half of pro-footballers' gamble, and that it is 'culturally ingrained' in the sport (BBC 2018). With such a big market comes many different betting options, and betters have evolved from betting on match outcomes and scores to various aspects of the game, such as events surrounding specific players, such as yellow and red cards. Betters in the UK can access an abundance of betting websites, with a lot of companies, such as BET365, sponsoring football teams.

Football is famously very stochastic and hard to predict - as described by (Eryarsoy and Delen 2019), football is very much in 'analytical backwaters', due to the sport's 'highly dynamic nature', with a seemingly infinite variables and their interactions that could determine the outcome of a match between two teams of 11 players, on a vastly 105m by 68m pitch. In a bid to capture as much information as possible, companies such as Opta and Prozone have spurned, with the latter producing video-technology software that tracks footballers on the pitch, generating thousands of datapoints per player per game (The Wired UK, 2017). Thus the avenue for data science in football is huge, with emerging clubs like Brighton known for their 'money-ball' strategies for recruitment.

Apart from betting, whereby betting houses need to predict odds as precisely as possible to generate profits to bettors who simply want to predict the card counts, analysing cards is of great interest to teams and coaches, especially since it could drastically affect the scoring ability of teams and outcome of matches (Badiella et al. 2022). Coaches may be interested in the different team and referee effects on the card counts, and possibly use such information to adjust their tactics and strategies.

Although supplied with both card data, this paper chooses to focus on modelling and predicting yellow cards. As we will see, red cards distribution that is zero-inflated and requires a different approach for modelling. The much wider distribution of yellow cards provide much more breadth in terms of model selection and warrants a lot of focus in analysis.

1.2 Football Rules, Referee, Cards

Under the *Laws of the Game* (contributors 2023), which are the codified rules for association football and are maintained by the International Football Association Board, yellow and red cards fall under the 'Fouls and Misconduct' section of the rule book. Fouls are actions that go against the game's laws and interferes with the progression of the game, and mostly involve overly-aggressive actions on the pitch. On the other hand, misconduct are more disciplinary in nature can occur anytime during the game, not only during play.

To track offenses and misconducts, yellow cards are handed out from list of actions, such as delaying the restart of the play (misconduct) or a reckless tackle with the intention of stopping the opponent's progression (foul). For graver offenses, a red card is shown to a player who is to be sent off from the pitch, forcing the team to play on with one fewer teammate. Possible misconducts include deliberately denying a goal scoring opportunity with a foul or a handball, and highly dangerous fouls such as tackling with the studs or two-legged slide tackles warrant a direct red card. Players can also be sent off if they accumulate two yellow cards (equivalent to a red card).

Logically and as we will come to see, red cards occur very rarely and much less often compared to yellow cards.

2 Surrounding literature

2.1 Literature on yellow/red cards

2.1.1 Referee bias

When it comes to cards, referees take a lot of heat. Referee bias is often seen as a large factor around the awarding of red and yellow cards, with football fans often holding unfavourable opinions of certain referees - the revered manager Jurgen Klopp was fined following comments questioning referee Paul Tierney's judgement and impartiality towards Liverpool Football Club (The Guardian, 2023).

In an ideal world, the awarding of cards is objective, and solely dependent on the specific footballing foul incident on the pitch. Unfortunately (or fortunately for our research purpose), there are many external factors that affect the refereeing decision. Referees are forced to make quick decisions that come with huge ramifications for the teams involved, and this stressful environment forces them to resort to heuristics and make them prone to cognitive biases (A. Nevill, Pearson, and Webb 2022).

Referee Idiosyncracies

Research by (Boyko, Boyko, and Boyko 2007) strongly purports the different effect referees have on games, such as goal advantage and more crucially yellow cards and penalties, aligning with the idea that a lot of outcomes are the result of idiosyncrasies of referee decisions and disciplinary types. Furthermore, a research on almost 4 decades of Australian rubgy reveals that home advantage was the most likely form of 'unconscious bias', and the advantage of playing at home had a large variance, with clubs faring differently under different referees (O'Brien and Mangan 2021).

A paper in BMC Psychology by (McCarrick et al. 2020) that focused on referee height as the main predictor reveals that while there was no significant correlation between height and number of fouls, across the top 4 leagues in the UK, shorter referees tended to award more yellow cards, possibly due to the psychological phenomenon of the "Napolean Syndrome" whereby shorter men compensate for the 'lack of height and social dominance'. Interestingly however, although there was a correlation with red cards, the direction differed between the lower and upper leagues, with taller referees awarding more in the Premier League and fewer in the lower leagues.

2.1.2 Home advantage

As with the familiar home advantage in terms of goals, studies have indicated a home team favoritism in referee decisions, which include yellow and red cards. A study by (Buraimo, Forrest, and Simmons 2010) on the English and German top-flight league revealed that despite controlling for variables such as within-game events and other pre-game factors, via the use of a minute-by-minute bivariate probit model, there was a systematic bias in treatment of home and away teams in favour of the former. Crucially, they controlled for main confounders of bookings, which were aggressive tactics by players, that would otherwise be masked as referee favouritism. In Spain, where (Picazo-Tadeo, Gónzalez-Gómez, and Wanden-Berghe 2011) examined home-bias data from seasons 2002/3 and 2009/10, they discovered that referees favoured the home team not in terms of free kicks/fouls but in terms of bookings, which comes after blowing the whistle to call for the foul.

Home crowd advantage

A common cause of home advantage lies in the home crowd effect. In an interesting psychological-experiment type set up, (A. M. Nevill, Balmer, and Mark Williams 2002) let a group of referees watch videotaped football incidents but muted the crowd noise for half of them. The group with noise called for 15.5% less fouls on the home team, resembling the decisions of the actual referees. Interestingly, referees award more fouls for the away team as the number of years of experience increases, but drops after a certain point (16 years), possibly indicating **referee experience** has a (non-linear) effect on the home-bias.

As in the aforementioned studies by (Picazo-Tadeo, Gónzalez-Gómez, and Wanden-Berghe 2011) and (McCarrick et al. 2020), referees seem to be objective and quick in calling fouls, but the decision to book the player afterwards may be heavily affected by crowd pressure.

Other research on ice-hockey by (Agnew and Carron 1994) argue that crowd *density* is the most important predictor, and proxy factors such as distance from field-to-stand, e.g. through the presence of a running track, ultimately affect the number of YC through crowd-influence in Germany football matches (Unkelbach and Memmert 2010).

2.1.3 Effects of covid

Even before the impact of covid-19, research has shown that the sudden lack of fans when matches are played behind closed doors in Italy (due to regulation following a violent incident) drastically reduced home advantage in terms of punishments for the away side (Pettersson-Lidbom and Priks 2010).

Covid-19 then provided a unprecedented opportunity of a natural experiment to observe the effects of empty stadiums across different countries. (Bryson et al. 2021) examined the 2019/20 season and found that difference in home and away yellow cards was reduced by a third of a card which was not accounted for by regular crowd variation. The authors concluded via the use of linear regression (found no significant differences with an equivalent Poisson regression) that the effect on the card differential was most likely due to referee effects since empty stadiums had no significant influence on the match outcomes (home win etc).

2.1.4 Big team bias

Research also shows that referees are influenced by the 'size' of the club, which may be correlated or confounded by factors such as crowd pressure. (Audrino 2020) found strong evidence in favour of a 'strong team bias' in all top flight leagues except that of France, with slightly different variations: English referees award more yellow cards to the strong team's opponent, while elsewhere (Spain, Italy, Germany) they penalise the strong team less.

2.1.5 Team playing style

As the research on the Spanish league by (Picazo-Tadeo, Gónzalez-Gómez, and Wanden-Berghe 2011) have shown, a statistically significant indicator predicting referee bias in terms of differential fouls and bookings were the difference in ball possession and the number of shots taken between the teams. When teams held onto the ball more, and were more direct in attacking, they drew fewer fouls. Hence we can expect teams to have a varying level of `team effects` on yellow cards.

2.2 Literature on goal modelling

Most statistical modelling research papers on football center around goal modelling. Teams aim to win matches in order to progress in tournaments (e.g. knock-out stages in League cups) or to gain points to climb league positions, and they do this by outscoring the opposition. Both goals and cards are count variables, and we will touch briefly on the research on the former.

Among popular research, one of the earlier papers came from (Maher 1982) who used Poisson distribution to model football goals, incorporating attack and defence from each team, via both independent and Bivariate Poisson distributions which allowed for some correlation between the home and away team goals. Next, the (Dixon and Coles 1997) model tried to improve on the prediction power of goal models by adjusting for low scoring matches, which they believed was being underestimated by standard bivariate Poisson models, by reformulating the likelihood function. (Karlis and Ntzoufras 2003) also used a bivariate Poisson distribution and considered an 'inflation factor' to improve the prediction of draws and overdispersion. Lastly, (Baio and Blangiardo 2010) replaced the frequentist framework with a Bayesian hierarchical model, which naturally accounts for goal covariance, and was shown to not be inferior, without requiring a specific algorithm curated for goal modelling.

3 Exploratory Analysis

3.1 The data as supplied by SmartOdds

The data, which has been supplied by SmartOdds(TM), consists originally of the following columns/variables: X, V1, fixture_id, season, season_start_date, season_end_date, country, league, competition_level, kick_off_datetime, team1_name, team2_name, referee, sup_implied, tg_implied, team1_yc, team2_yc, team1_rc, team2_rc, team1_bk, team2_bk, bksup, bktot, attendance_value, limited_audience, match_date, season_end_year, team2_stadium_dist, stadium_capacity, stadium_runningtrack, stadium_altitude, crowd_density, with a total of 27255 observations. It consists of the top 2 leagues across the "Big Five" European countries for football, from the seasons 2014/15 up till 2021/22. Each row consists of one single match, reflecting the home team and away team, and the relevant statistics (YC, RC) for both teams.

To convert the dataset to one where there is only one single YC column, we reformulate the dataset into one that is twice as large where each match now becomes two - each from the point of view of both teams, and an additional column for the `home` indicator variable is introduced. We adjusted the `sup_implied` for the away team to be negative, while that of the home team is untouched, as per the original definition. This allows us to model the yellow card for a particular team, with everything else (including home) used as predictors.

For reference,

- sup_implied : Implied goal supremacy
- tg_implied : Implied total goals
- YC : Yellow Cards

The definition, given by Smartodds:

- sup_implied and tg_implied : a translation of the available Asian handicap odds for these matches into what we believe the market was expecting for (home) goals supremacy and total goals in the match

These 'implied' data are uniquely supplied by Smartodds, a betting consultancy, and potentially provide interesting insights into card prediction.

Most other columns are self-explanatory. Some were introduced later as an updated dataset, and to prevent confusion, they include:

- stadium_dist : Distance traveled by the away team for the match, i.e. distance from both stadiums of the home and away team. If small, this could represent geographical rival clubs.
- stadium_runningtrack : Presence of a running track around the football pitch, representing a buffer between the pitch and the spectators.

3.2 Our predictors

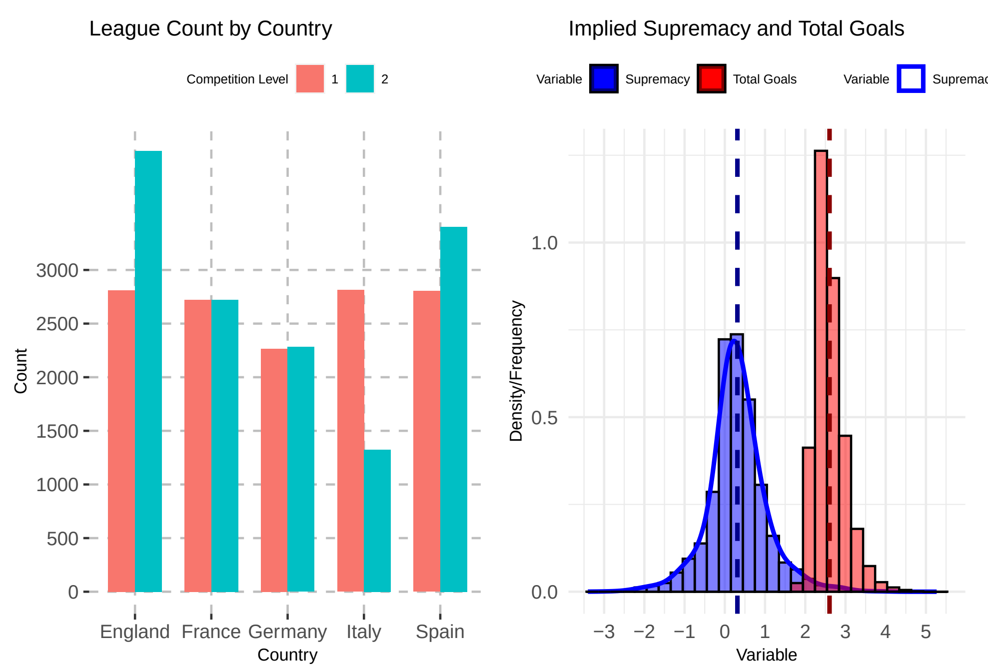


Figure 1: Countries, leagues, implied data

Country and league

From Fig 1, we have data on 5 countries: France, England, Germany, Spain, Italy, each with their top two leagues (FraL2, FraL1, EngCh, EngPr, GerBL2, GerBL1, SpaPr, SpaSe, ItaSA, ItaSB). Of note: There is considerably less data for the 2nd Italian league (Serie B) compared to other leagues, and the most number of data points from the UK Championship.

Implied Supremacy and Total Goals

We are provided with `sup_implied` and `tg_implied`, which are what the markets expect the goal supremacy (how much home team wins by) and total goals to be. The range of supremacy and total goals are : -3.4, 4.21 and 1.6, 5.27

- The `sup_implied` closely resembles a normal distribution centered around $0.3128169 > 0$ with a standard error of 0.7284965, indicating a slight home advantage.
- The `tg_implied` resembles more of an exponential type distribution, since it cannot be negative (sample minimum around 1.6). The mean total goals is 2.6.

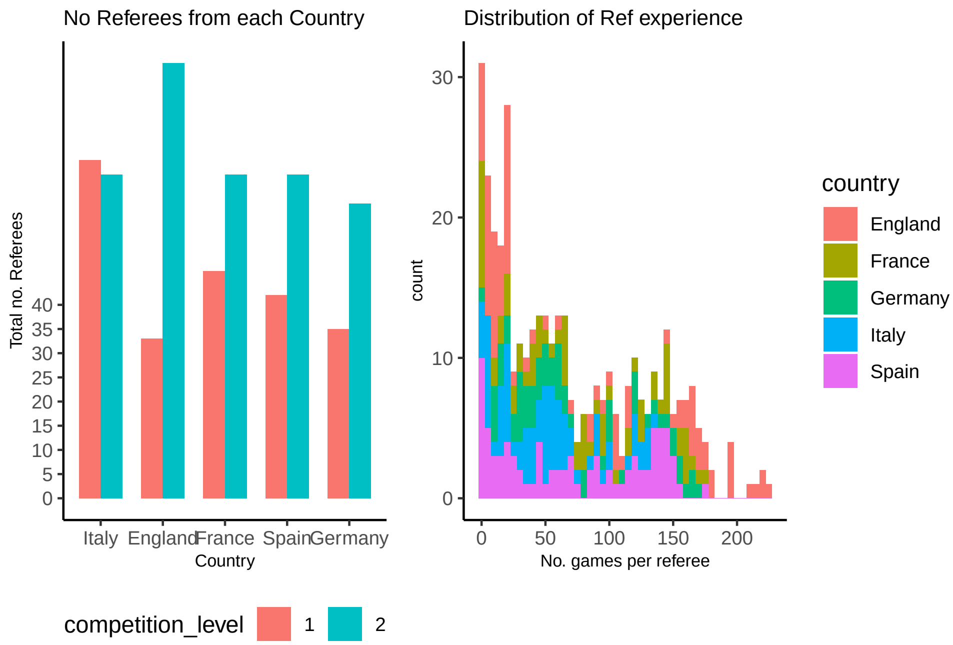


Figure 2: Referees

Referees

There are a total of 217 referees, of which a large proportion of them are from the UK (35+%), followed by Germany and France (~20%), and then Spain and Italy (~12%). Referees with a lot of games come from England, but overall there is a large spread, with some having fewer than 20 or 30 matches, while other with more than 200. Hence we have a large number of referees with a stark imbalance of samples between them, which will guide our modelling approach down the line.

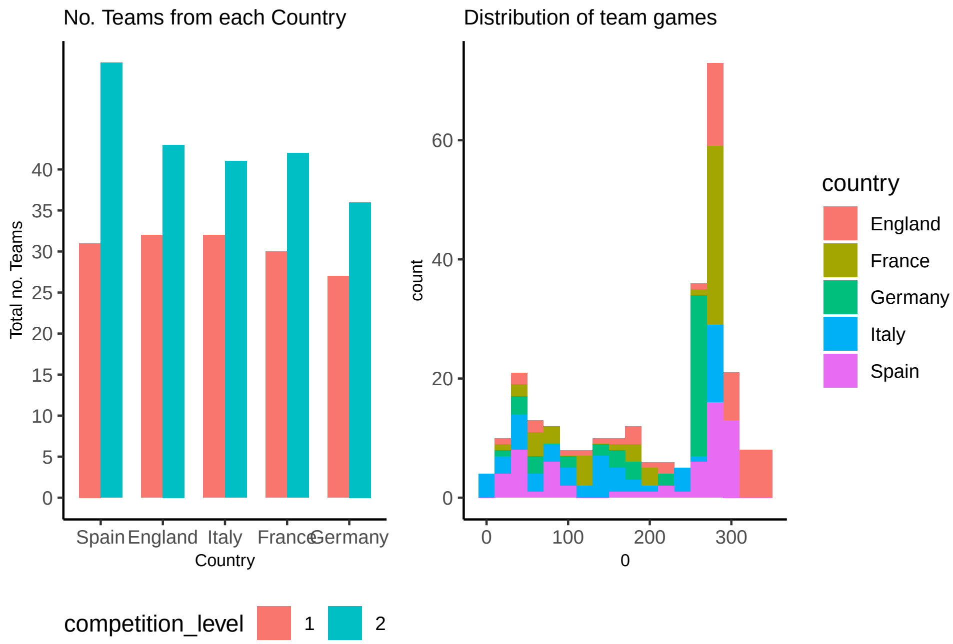


Figure 3: Teams

Team

There are a total of 271 teams across all competitions. Across countries, there are more teams in the second league than first, potentially due a higher turnover from teams being promoted upwards and relegated downwards, the former being not possible in the top league. Teams effects are vital as some may employ tactics that lead to more fouls and hence cards than others. As with referees, we have some teams with few games and many with more than 250. Again, England has the largest spread of number of team games.

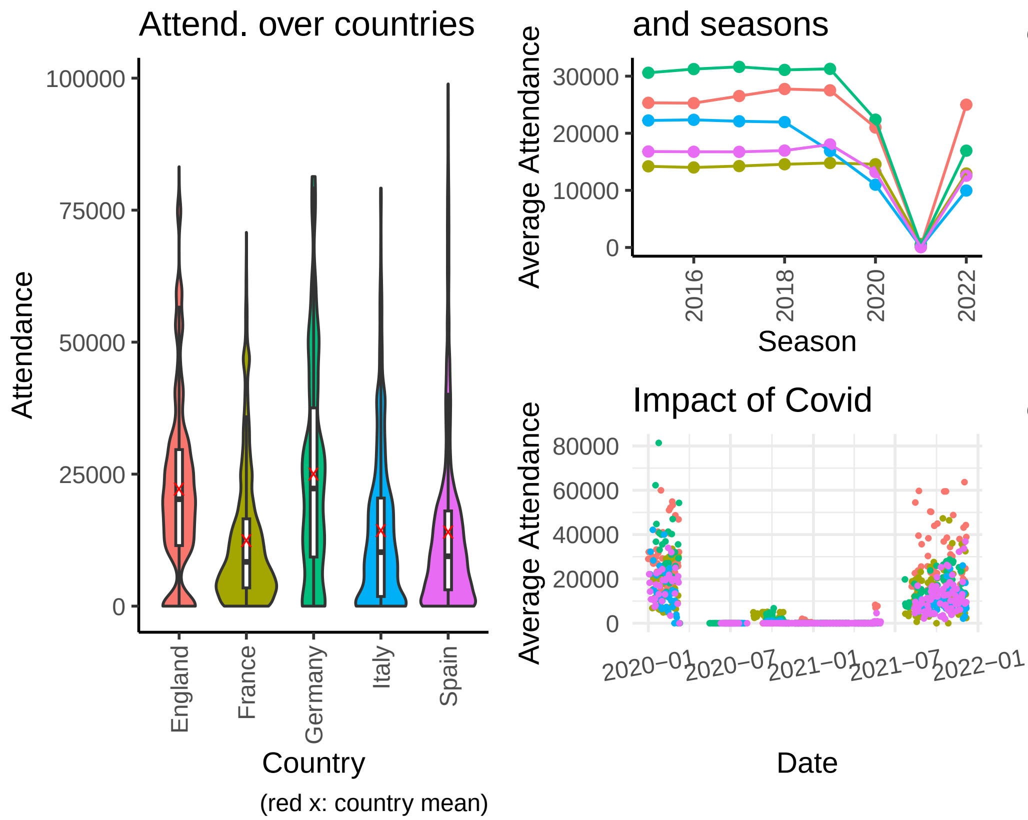


Figure 4: Attendance

Attendance and covid effects

A first glance at the distribution of attendance rates across the countries show a significantly varying number of fans in stadiums, with countries like England and Germany having the higher averages and right skew, with a relatively larger proportion of attendance rates above 50 k. One the other hand, France, Italy and Spain are a lot heavier on the bottom of the attendance scale, with much lower averages and a lot of mass below 25000, although it should be noted that Spain has the highest upper limit (most right skew), possibly due to their renowned high-capacity Camp Nou stadium that attracts the largest crowds in the world.

The moving season average plot shows the comparison of means more clearly - it is almost always the case that Germany has the highest averages, follows by England, Italy, Spain and France. There is a falling trend towards 2021 due to Covid, before bouncing back.

The final plot zooms in on the Covid period where attendance was banned for safety reasons, and attendance remained 0 as games were played indoors (apart from some limited numbers occasionally). For simplicity, we will presume that the covid-ban lasted from 2023 March up till the start of the 2021/2 Season.

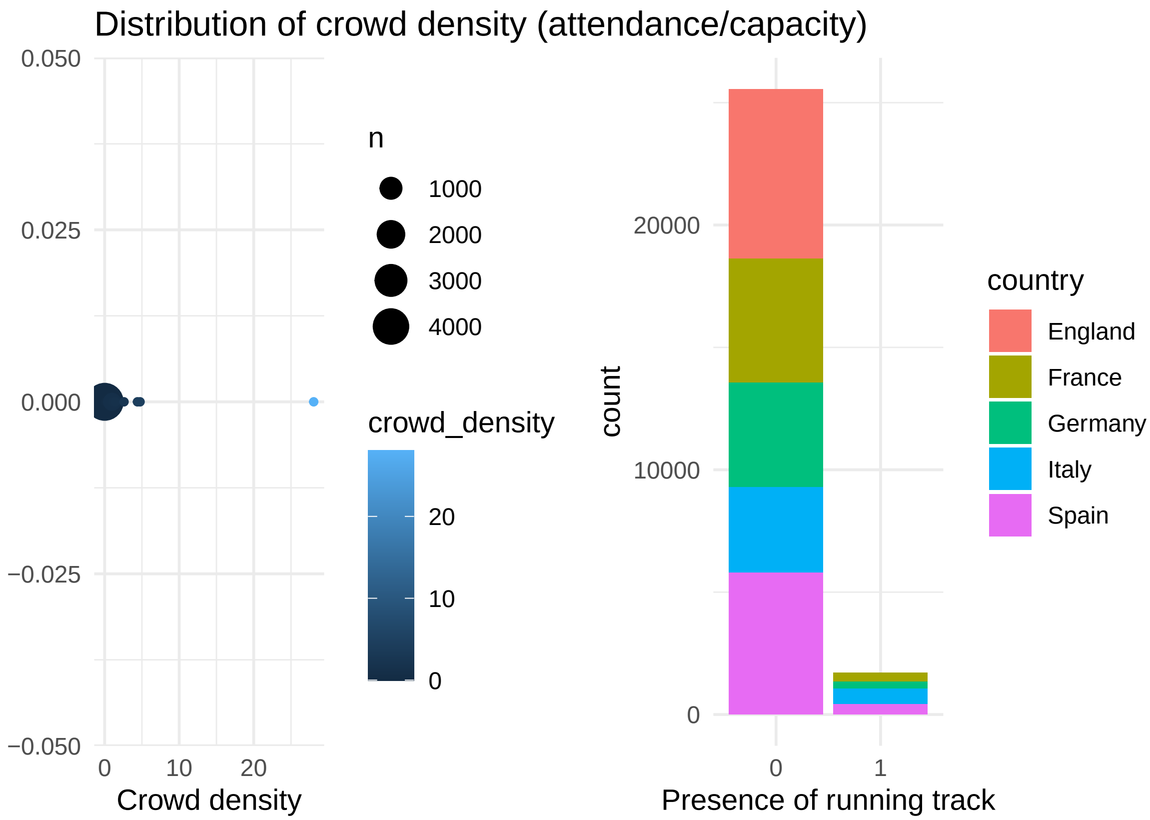


Figure 5: Extra predictors: crowd density, running track

Crowd density

One of the important aspects of data visualisations is to spot outliers that may not be so salient in summary statistics. Logically all values of crowd density should fall between 0 and 1, hence we will set the values >1 to an upper limit of 1, which would be especially important for the anomalous observation of 28x stadium capacity attendance, which would otherwise be an high leverage point that could give biased estimates.

Running track

Most stadiums do not, but a small amount from each country (apart from England) contains a track that provides a buffer between the spectators and the football pitch.

3.3 Summary of predictors:

As a summary, we have 3 main types of predictors: numeric, categorical, and high cardinality categorical. The presence of the latter will influence our choices of models.

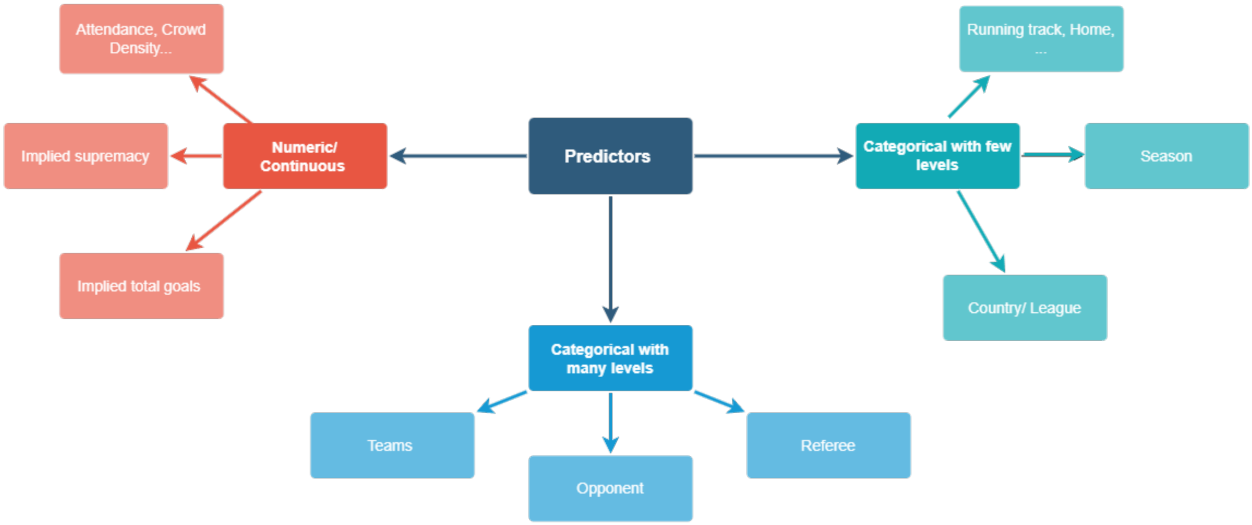


Figure 6: Summary of datatypes for predictors

3.4 Missing data

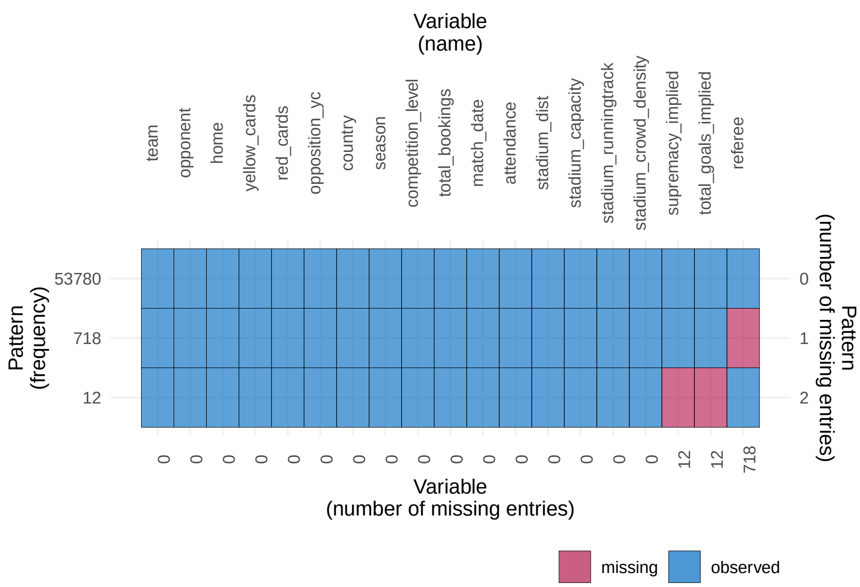


Figure 7: Missingness by factors

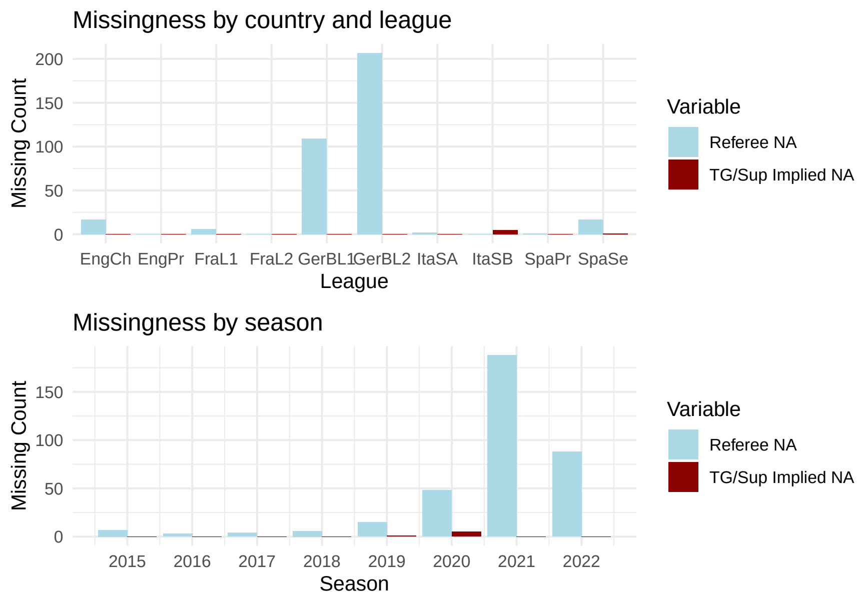


Figure 8: Source of missingness

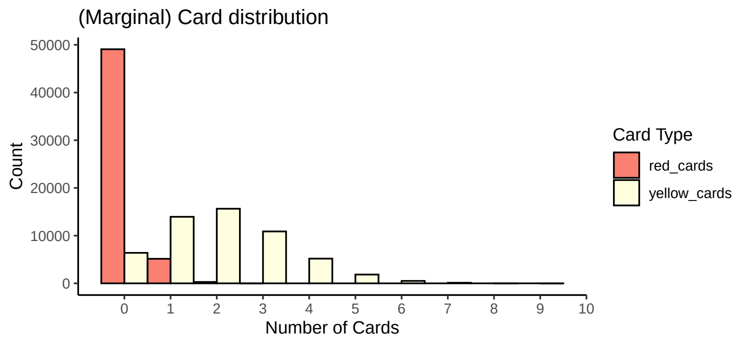
From Fig 7, 718 instances of datapoints are missing for the `Referee` column, while 12 are missing for `tg_implied` and `sup_implied`, which all come from the same observations (i.e. 'implied' NAs overlap fully).

The bulk of the missingness comes from referee NAs in the German leagues, and mostly during the more recent years: from 2020 to 2022. The missing implied data is isolated to the Italian second league in 2020.

Due to the relatively small proportion of rows that contain missing data (less than 1.5%), it should not present an issue to our analysis. Since there are about 700+ missing referee data, we will replace it with a 'missing' placeholder that is further segmented by country (so as to capture some variance of referee behavior) and league, via "Missing - - ". This implicitly assumes there will be some variation between countries and leagues in terms of referee effects. For the implied columns, we will impute them with means of the specific league and country.

3.4.1 Distribution of cards

Cards represent a count variable which is discrete and non-negative. In most cases, a Poisson regression is used, with special extensions/adjustments in the presence of overdispersion.



Card Count	0	1	2	3	4	5	6	7	8	9
Yellow	6390	13947	15635	10880	5180	1840	505	119	12	2
Red	49072	5143	285	10	0	0	0	0	0	0

Figure 9: Red and Yellow Card Histogram

As seen from Fig 9, the range and spread of YC are much wider than that of the red cards (RC), which is 0-inflated and tapers off quickly. The unconditional distribution of YC resembles a Poisson type. We fit and compare a Poisson and a Negative Binomial to the marginal distributions, and see little difference in Fig 10.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    1.00    2.00    2.05    3.00    9.00
```

```
##       size         mu
## 850.559676   2.049915
```

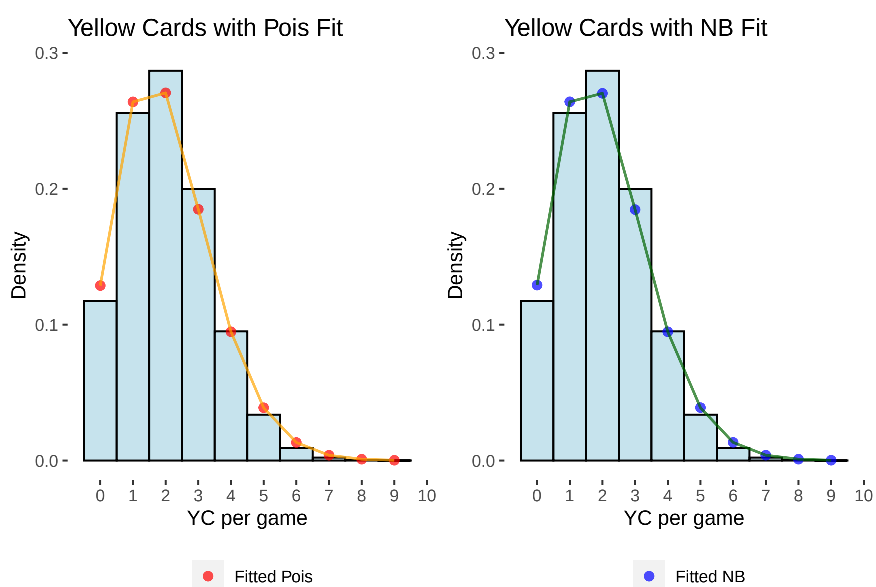


Figure 10: Histogram of Yellow Cards, against Poisson and Negative Binomial fits

3.5 Pairwise plots

3.5.1 Explore some pairwise relationships : Numeric

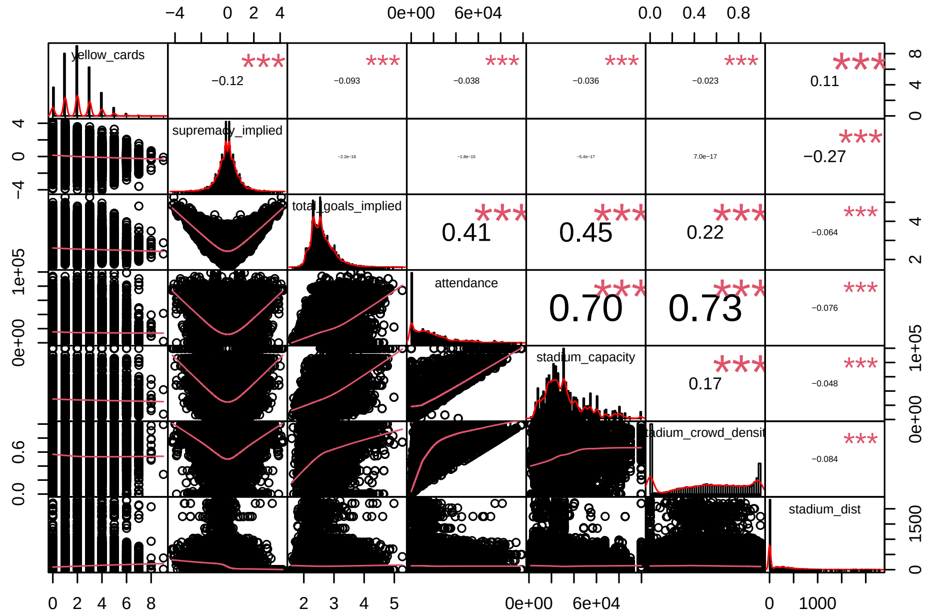


Figure 11: Correlation plot by PerformanceAnalytics

A brief look at the correlation plot in Fig 11 reveals that while significant, the correlation with yellow cards are mostly small in magnitude. We also see high correlation between similar predictors, such as attendance, stadium capacity, etc, a precursor to section 12.

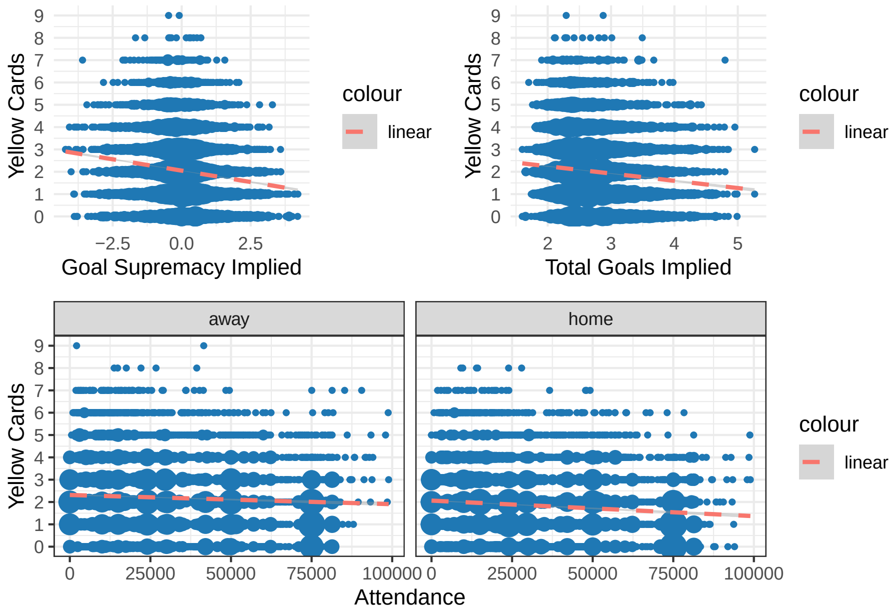


Figure 12: Implied data, attendance

Implied columns

There is a slight negative correlation between goal supremacy and yellow cards - a possible interpretation is that the higher the goal supremacy, the 'stronger' the home team is, and hence less intensity that would usually lead to yellow cards. Total goals shares the same direction of correlation, although the interpretation is trickier in this case, and could be confounded through goal supremacy.

Attendance

Attendance has a very small correlation with YCs. However, what is interesting is that has a relatively high correlation with total goals implied, both of which could be confounded by match intensity. When we separate out the effects of attendance on YC by home/away, we find that as suggested, there is a general crowd variation effect that influences the home advantage - YCs seem to drop faster for the home side, albeit marginally.

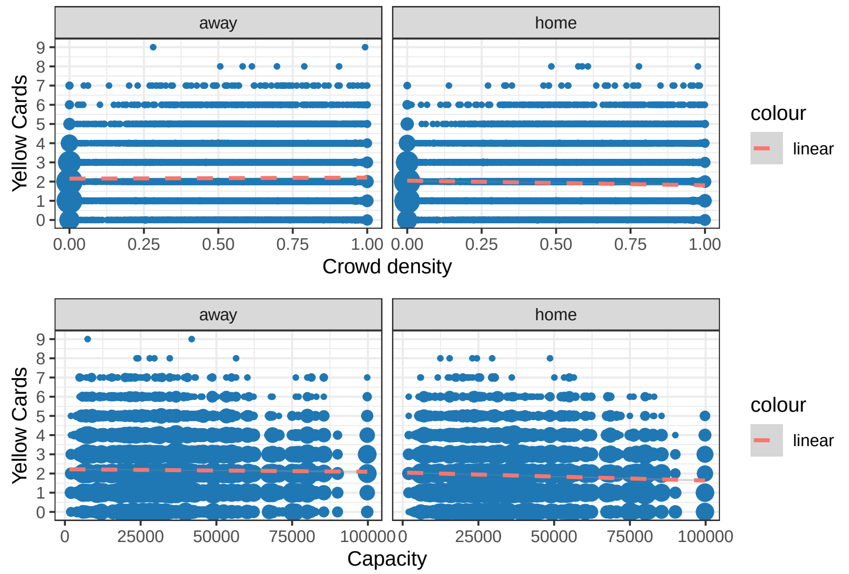


Figure 13: More numeric bivariate plots: crowd density, distance travelled, stadium capacity

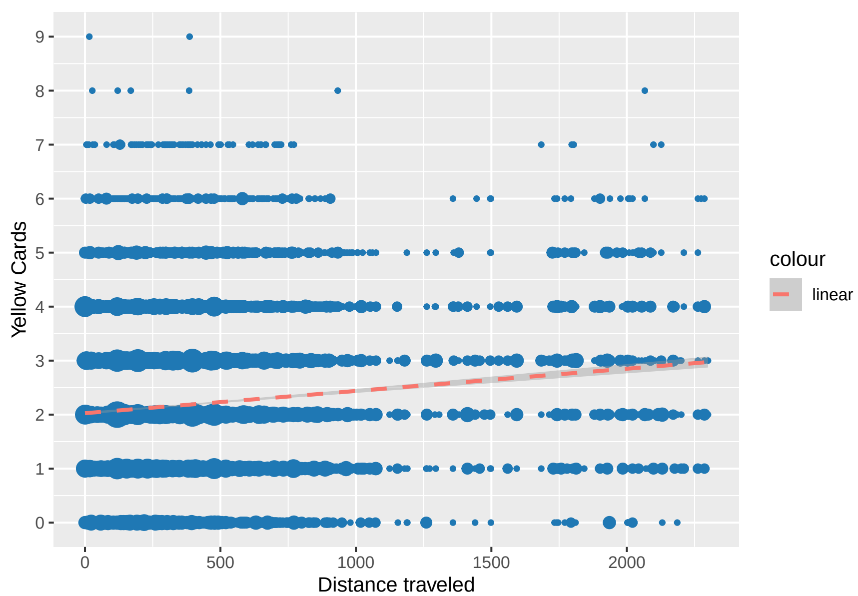


Figure 14: More numeric bivariate plots: crowd density, distance travelled, stadium capacity

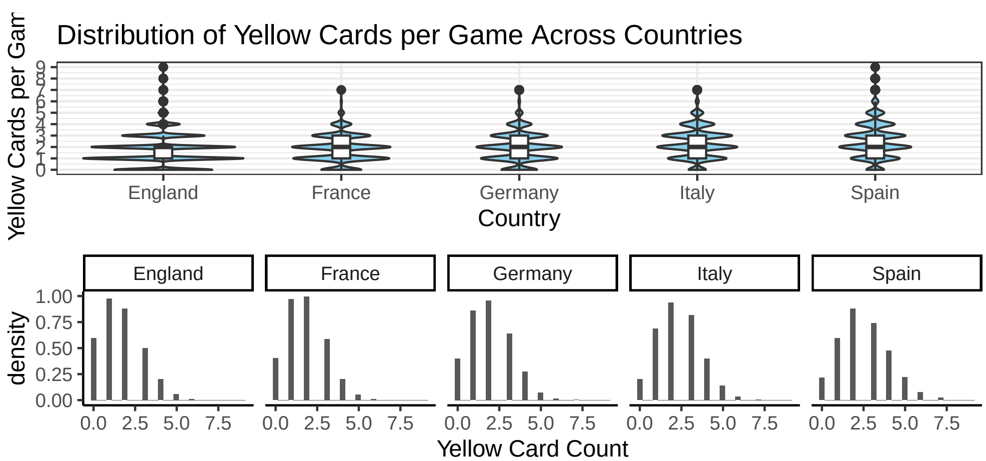
Stadium Capacity and Crowd Density

$Crowd\ density = Attendance/Capacity$, and hence it is reasonable that they are highly collinear, and hence also correlate strongly with total goals implied and. As a standalone predictor, capacity could be an indication of the wealth and budget of a club, is related to the prestige and popularity of the club, which could affect refereeing decision (Bose*, Feess, and Mueller 2021). And interesting note is that the crowd density seems to not affect home advantage (equal slopes), although other factors are not controlled for.

Stadium Distance

The further the away team has traveled, the more likely they are to get YCs, potentially due to travel fatigue (Schwartz and Barsky 1977) or due to a greater home advantage effect that the crowd would have on the referee (stronger in-group out-group bias). This has been the largest correlation thus far, and goes against the perception that closer clubs have stronger rivalries and are marred by a higher number of cards. Another interesting finding is that the further a team travels, the fewer away fans tag along with the away team, possibly diminishing the already inferior crowd presence and pressure for the visitors (Humphreys et al. 2022).

3.5.2 Explore some pairwise relationships : Factorial



country	England	France	Germany	Italy	Spain
mean	1.68	1.82	1.95	2.35	2.54
median	2	2	2	2	2
se	1.26	1.19	1.27	1.30	1.47

Figure 15: YC across countries

Countries

Coincidentally, from the violin/boxplot of countries, the distribution of yellow cards seem to be flattening and moving upwards/rightwards as we move from England to Spain, meaning that there is more weight towards higher card counts, although England and Spain both have many upper outliers. The summary table shows the gradual increase in means more clearly - Spain has almost 1 more YC on average than England, and together with Italy, have more mean YC than median (2). The median is the same across all countries, and from the histograms it is clear that the most common number of cards is 2, apart from England (1).

Home

From Fig 16, we see home advantage manifesting in terms of fewer YCs for home teams, across all countries.

League

The YC difference between competition levels varies across countries: imperceptible for England and France, but higher for the others, with Germany having the most pronounced increase with relegation. A possible reason for this could be skill or tactical differences between stronger and weaker teams in the respective leagues.

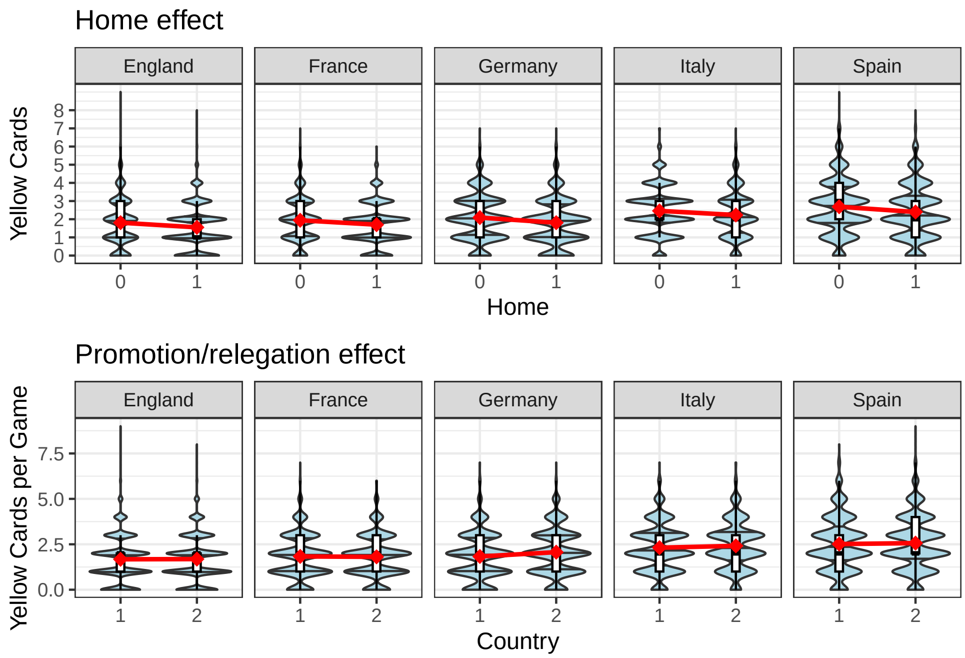


Figure 16: Home and League effects

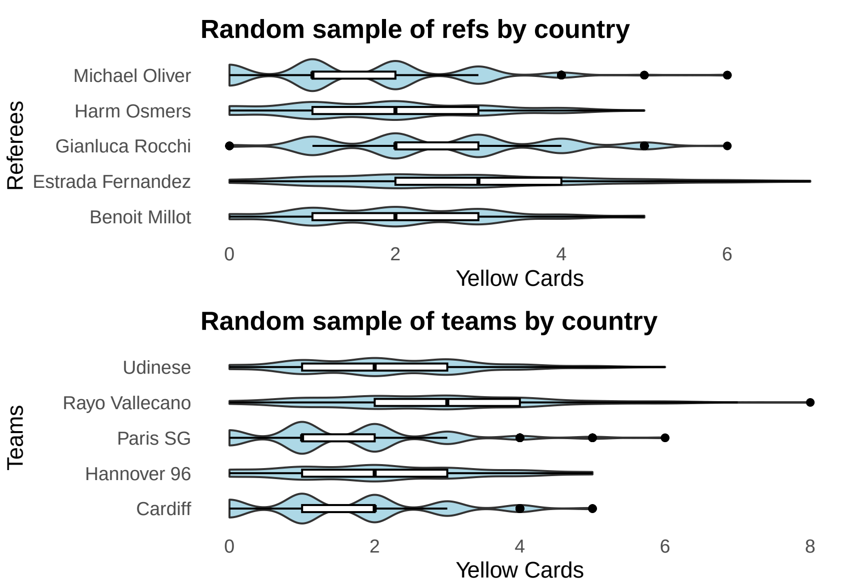


Figure 17: Referee and Team effects

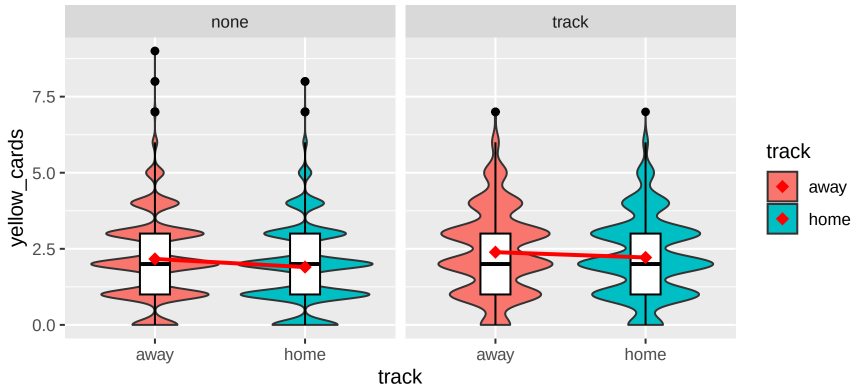
Referees

Referees have a huge influence on number of cards awarded, and it is expected that the distribution varies across individuals. Due to the large number of unique referees, we pick a random sample (1 from each country) in Fig 17.

Teams

We take the same approach for teams, and once again observe unique distributions.

In both variables, YCs are likely to be correlated within levels, and we will address this within-cluster dependence later.



	home	track	mean	se
1	0	0	2.167	1.365
2	0	1	2.389	1.377
3	1	0	1.9	1.316
4	1	1	2.216	1.29

Figure 18: Separation by running track

Track

The YC distributions seem to become less right skewed with a track, although the sample imbalance could be a reason. It is hard to say if home advantage is really reduced in the presence of a running track, as the mean for both sides increase as a result.

3.6 Explore time factor: Any seasonality within seasons?

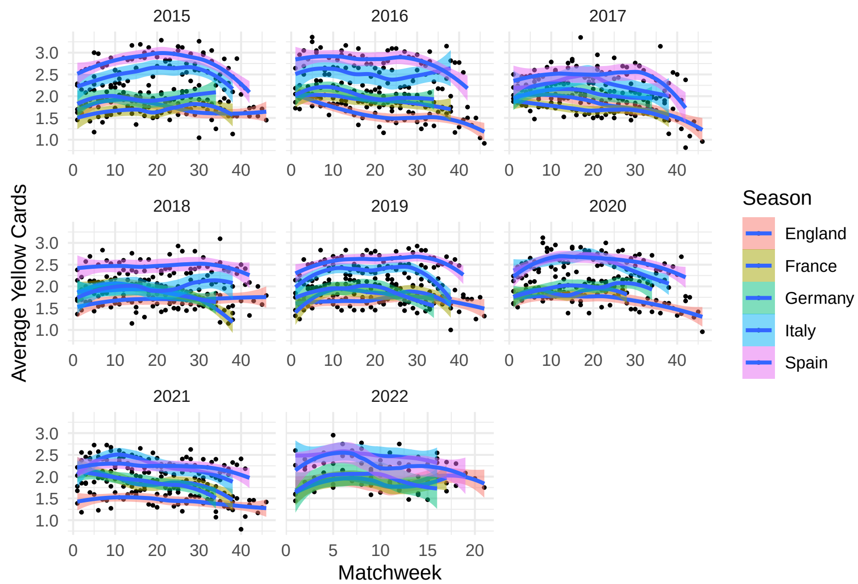


Figure 19: YC Seasonality

We see that there is a general seasonality where the mean number of YCs increase first, then decrease towards the end of the season, although the curvature varies by country. The curves in the plots were fitted with local regression methods and suggest polynomial effects, clustered based on countries.

4 Feature engineering

We "engineer" some new columns that may help in our predictions.

4.1 Adding lag variables - yellow cards

A common method used to model the number of goals in football is to a proxy for the team's current performance, via a weighted average of the previous games' number of goals scored. In the same vein, we will test if adding the lagged count of YCs helps to predict the team's propensity for YCs in a present game. We introduce this as `yellow_card_lag1`.

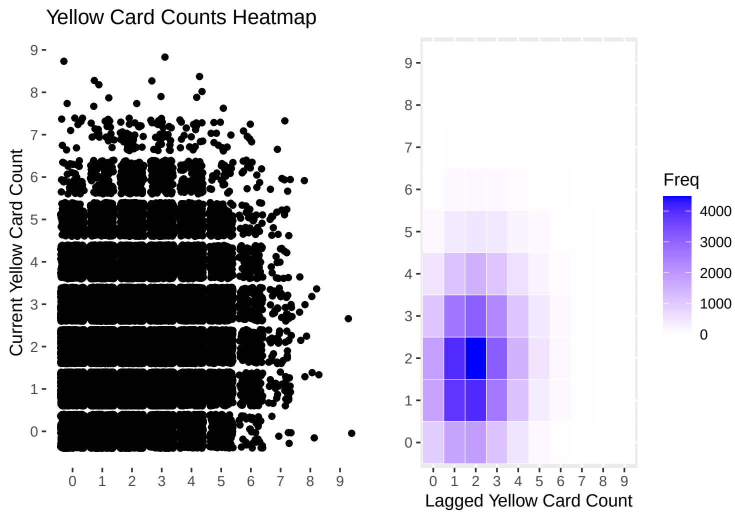


Figure 20: Adding a lagged YC variable

The heatmap gives us a better view since the variables are discrete. It looks safe to assume that there is a linear relationship between the lag and the current yellow card.

4.2 Adding more weight to recent observations for model training

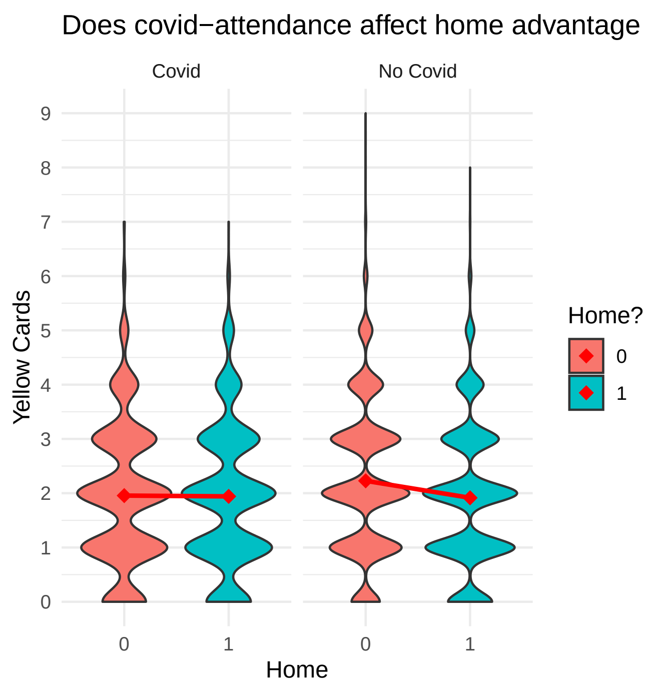
For the purposes of forward predictions (predicting games in the future), we add a new column containing the `days_elapsed` which captures the difference in days between a the specific match (observation) and the most recent date found in the dataset. To give greater weights to observations that are more recent, we use `caseweights_by_recency` which is a strictly decreasing function of number of days, with the formulation taken from ("Creating Case Weights Based on Time," n.d.).

4.3 Add referee experience and covid_impacted period

For engineered features that could be useful, we add an indicator variable for empty stadiums due to covid restrictions, by subsetting for dates where the measures were enacted. To get a proxy for referee experience, we generate a new column that measures the number of games a particular referee has already officiated prior to the observation. [1]

[1] There are obvious limitations - the `referee_experience` is truncated at the beginning of the dataset, and all referees will start at 0 whether they have had a lot or no experience prior.

4.3.1 Does covid affect home-advantage for YCs?

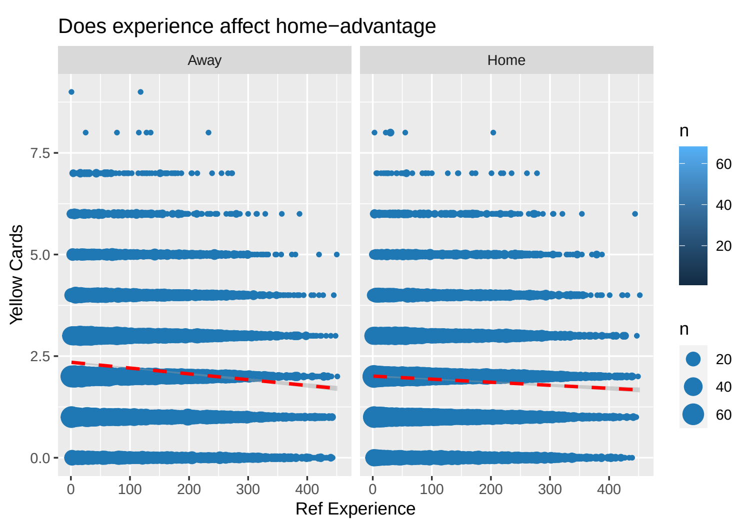


	covid	home	mean	median
1	0	0	2.23	2
2	0	1	1.92	2
3	1	0	1.96	2
4	1	1	1.94	2

Figure 21: Home advantage in ghost games (covid)

As with the literature, referees are indeed more likely to treat both sides more equally in covid-impacted matches, as seen from the fig 21 where the distribution becomes even between home and away.

4.3.2 Does referee experience affect home-advantage for YCs?



Referees with more experience seem to award fewer YCs for both home and away.

5 High Cardinality data

Table 1: High cardinality dataset

Number of referees	Number of teams	Number of opponents	n
394	271	271	54510

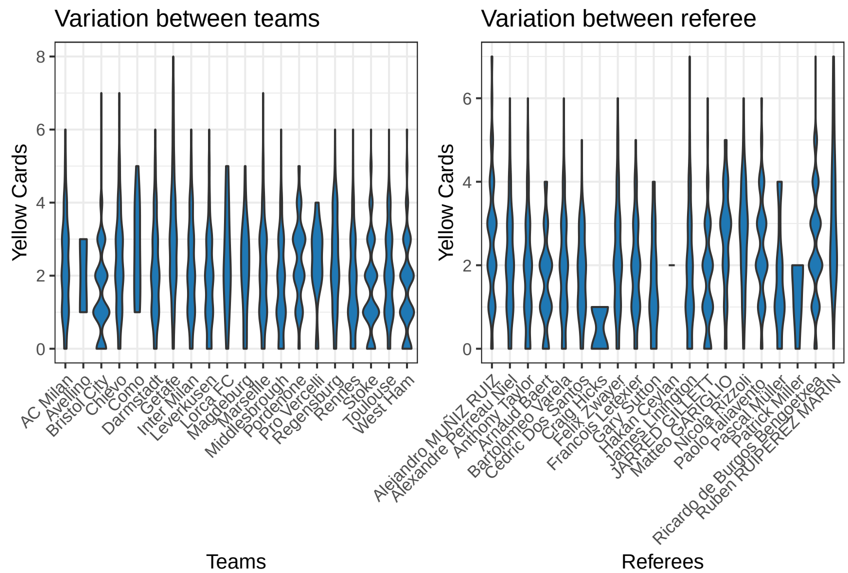


Figure 22: High cardinality categorical variables

From 1 and 22, we struggle with high cardinality for `referee`, `team` and `opponent`. Treating it as a factor results in one-hot encoding, but this would be unfeasible as the feature space becomes too overwhelmed, and the large number of levels violate the orthogonality assumption of each level and suffers the curse of dimensionality as a result, giving further computational difficulties. As we have seen, high-cardinality usually comes with uneven exposure (Fig 2, 3), where levels with fewer observations are poorly estimated (Avanzi et al,2023). Furthermore, we may run into the issue of overfitting, giving the model a high variance, and many models will have an issue finding reasonable regularisation (Sigrist 2023).

Combining with the issue of repeated measurements and uneven sampling per categorical level, the suggested methods include the following

- Adding mixed effects to linear models (good for cardinality where levels are have uneven sample sizes)
- Tree boosting (good for cardinality, high dimensions)
- Combining boosting and random effects via GPBoost (Sigrist, 2022, 2023)

Due to repeated measurements, alongside mixed effects we test Generalised Estimating Equations (GEE), but show its limitations for our use-case.

6 Approach for modelling

Table 2: Models and their focus

Type	Model	Description	Focus	Dataset
Linear	Linear Model	Standard Multiple Linear Model	Interpretation	Country by Country
GLM	Generalized Linear Model (GLM)	Poisson regression for count data	-	.
.	Standard GLM	Referee, team random effects	-	.
.	Generalized Estimating Equations (GEE)	Correlation structure for referee, teams	-	.
Tree-based	Random Forest	Ensemble of decision trees	Prediction	Full data
.	XGBoost	Gradient boosting algorithm	-	.
.	GPBoost	Tree boosted models with mixed effects	-	.

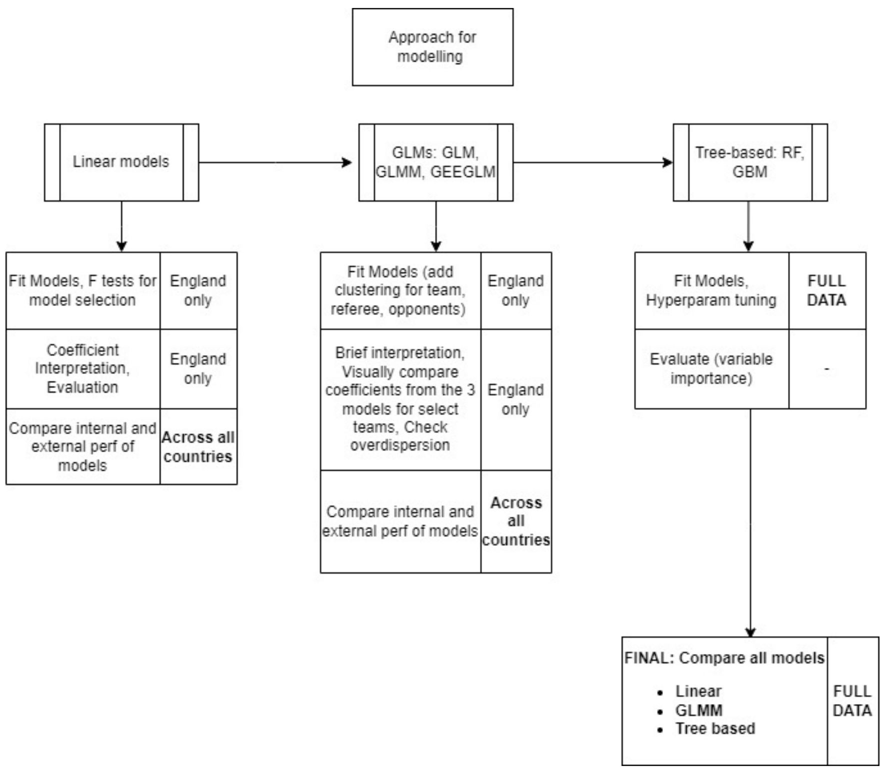


Figure 23: Methodology

Linear models allow us to get a sense of how each predictor affects the outcome directly. For simplicity, we run the models country by country but take a deeper look at the coefficients for the England dataset. On many metrics the English Premier League is the most highly-anticipated in the world, leading by broadcasting income per club of 143.2 million pounds, substantially higher than the next best league which is the La Liga from Spain (66.6 million pounds), with the other top-flight leagues trailing behind (UEFA Benchmarking Report 2019).

Nevertheless, we do not lose out on generality as at the end, we compare all models on the full datasets with all countries, and the associated linear model regression outputs can be found in the appendix.

[2]

[2] Readers who are only interested in models trained with the entire dataset can skip to the last section

7 Model 1: Linear Models

7.1 Motivation for Linear Models

Following the exploratory analysis on the pairwise plots (e.g. YC against Implied Supremacy), we detect linear relationships that were weak in magnitude. The obvious choice for Poisson regression unfortunately assumes a non-linear, exponential relationship between the outcome and the predictors (more on this later), which may be a tough ask. Linear models are simple and highly interpretable (estimates can be interpreted directly in the same scale as the outcome) and offers a good starting point in our analysis.

7.1.1 Assumptions:

- Linear relationship between predictors and outcome
- Normal distribution of errors
- Independence between observations
- No multicollinearity
- Homoskedasticity

7.2 Theory for Linear Models

We regress

$$\begin{aligned} YC \sim \text{Season ,Competition Level, Referee, Team, Opponent, Implied Supremacy} \\ \text{Implied Total Goals, YC Lag 1, Matchweek} \end{aligned}$$

To model the equation

$$\begin{aligned} YC = \beta_0 + \beta_1 * \text{Season} + \beta_2 * \text{Competition Level} + \beta_3 * \text{Referee} + \beta_4 * \text{Team} + \beta_5 * \text{Opponent} \\ + \beta_6 * \text{Implied Sup} + \beta_7 * \text{Implied Total Goals} \\ + \beta_8 * \text{Matchweek} + \beta_9 * \text{YC Lag 1} + \epsilon \\ \text{where } \epsilon \sim N(0, \sigma^2) \end{aligned}$$

excluding the possible interactions between predictors, and the newer variables that were supplied later on by Smartodds, such as attendance, capacity, etc.

Since `referee`, `team` and `opponent` are one hot encoded vectors, $\beta_3, \beta_4, \beta_5$ are actually vectors of coefficients themselves, where $\beta_{\mathbf{j}} = [\beta_{j_1}, \beta_{j_2}, \ldots, \beta_{j_{n_j - 1}}]$ are coefficients for categorical variable j which has n_j levels. In this case, each level for the `referee` represents a particular referee, and each level for `team` or `opponent` represents a team from the league.

7.3 Data Preprocessing

We standardise `tg_implied` and `sup_implied`. To still allow for easier interpretation for attendance and capacity, we now define it in terms of 10 000 people (i.e. divide by 10 000).

R uses **reference level coding** which means that a certain level from every categorical variable is used as a point of reference and is captured by the intercept, allowing all other level coefficients to be interpreted with respect to the reference level. We change the relevel the `team` variable to have *Arsenal* as the reference, and *Aston Villa* for that of the `opponent`.

Train test split

Since the logical use of a football model would be one for predicting the future instead of the pass, we select a random sample for the test set from the most recent season (2022) in the dataset. The models are built with special `weights` by recency which can be conveniently specified in the built in `lm` function in R, as well as most other model fitting algorithms.

7.4 Running the Linear models

7.4.1 England results

The simplicity and efficiency of linear models allow for us to do simple hypothesis testings and exploratory analysis for the determinants of YCs. Of interest would be whether there are still interaction effects when all other variables are held constant, i.e. via the 'ceteris paribus' assumption which basic plots would not have informed us on.

The base linear model `model_1_linear` includes no interaction and treats `season` as numeric. In `model_1.2_linear` we allowed for `seasons` to be factorial, which agreed with the data visualization that every season was unique rather than following some general time trend, season * matchweek interactions. For `model_1.3_linear` we added quadratic matchweek effects. We then added interactions between `home` and `team` since clubs may have different home-advantages for `model_1.4_linear`, which was quickly proven to be overfitted and hence dropped the interactions. Finally, we added `home` related interactions with `attendance`, covid, capacity etc for `model_1.5_linear`.

[3]

The F tests overwhelming stood in favour of the fullest model (`model_1.5_home_intns`) - the p-values from comparing 1.5 to the nested models are all essentially 0. Using the `compare_performance` function from the package `performance`, we get a rank with `model_1.5_linear` being the best in terms of a combination of model metrics, such as R^2, AIC, (although BIC naturally favoured the base model with no interactions at all).

Table 3: Comparison of MSE (Linear Models)

Name	AIC	BIC	RMSE	test_RMSE
model_1.1_linear	47973.76	49598.90	1.213684	1.641204
model_1.2_linear	47958.09	49635.89	1.211236	1.632218
model_1.3_linear	47952.81	49638.14	1.210971	1.631244
model_1.4_linear_home_itns	47913.70	49614.07	1.209582	1.636457
model_1.5_linear_home_itns	47913.48	49628.90	1.209803	1.638482

On both training fit and test performance, model 1.5 had the best overall results, allowing an increase in model fit without a substantial increase in test MSE (although differences in metrics across models are small).

7.4.1.1 Interpreting linear model (England)

Table 4: Linear model output table

term	estimate	std.error	statistic	p.value

[3]Recall that since England only has stadiums without running track, we won't include it as a predictor.

teamSwansea	-0.36	0.114	-3.167	0.002
teamAston Villa	-0.015	0.105	-0.146	0.884
teamMillwall	-0.225	0.12	-1.887	0.059
teamSheffield United	-0.072	0.111	-0.646	0.518
teamWest Ham	-0.263	0.104	-2.533	0.011
teamHuddersfield	-0.187	0.115	-1.624	0.104
teamWatford	0.149	0.109	1.363	0.173
teamWigan	0.038	0.141	0.273	0.785
teamMan United	0.244	0.103	2.361	0.018
teamLiverpool	-0.4	0.104	-3.857	0
...	...	...	...	...
refereeIain Williamson	0.463	0.4	1.158	0.247
refereeMathew Buonassisi	-0.941	1.467	-0.642	0.521
refereeDean Whitestone	-0.26	0.201	-1.291	0.197
refereeCraig Pawson	0.234	0.087	2.686	0.007
refereeJARRED GILLETT	0.119	0.238	0.499	0.618
refereeMichael Salisbury	-0.108	0.301	-0.36	0.719
refereeChris Foy	0.164	0.387	0.423	0.673
refereeMike Jones	0.318	0.145	2.189	0.029
refereeJames Adcock	-0.043	0.22	-0.197	0.844
refereeJohn Brooks	0.373	0.192	1.945	0.052
...	...	...	...	...
opponentNorwich	-0.263	0.099	-2.645	0.008
opponentLeicester	-0.43	0.103	-4.187	0
opponentCoventry	0.096	0.13	0.744	0.457
opponentWest Bromwich	-0.565	0.099	-5.703	0
opponentPreston	-0.5	0.102	-4.884	0
opponentBurton	-0.907	0.185	-4.91	0
opponentHuddersfield	-0.305	0.101	-3.008	0.003
opponentHull	-0.6	0.11	-5.463	0
opponentMiddlesbrough	-0.372	0.101	-3.693	0
opponentLiverpool	-0.634	0.113	-5.625	0
(Intercept)	-3.808	88.687	-0.043	0.966
home	-0.143	0.146	-0.974	0.33
season	0.003	0.044	0.068	0.946
competition_level2	0.203	0.06	3.416	0.001
attendance	-0.079	0.025	-3.145	0.002
stadium_dist	0	0	-1.751	0.08
stadium_capacity	-0.02	0.021	-0.94	0.347
stadium_crowd_density	0.502	0.164	3.061	0.002
matchweek	0.001	0.006	0.243	0.808
yellow_card_lag1	-0.007	0.008	-0.817	0.414
covid_impacted	-0.141	0.133	-1.057	0.291
total_goals_implied_scaled	-0.044	0.019	-2.387	0.017
supremacy_implied_scaled	-0.112	0.024	-4.66	0
factor(season)2016	-0.061	0.146	-0.415	0.678
factor(season)2017	0.251	0.123	2.034	0.042
factor(season)2018	-0.267	0.104	-2.571	0.01
factor(season)2019	-0.223	0.087	-2.544	0.011
factor(season)2020	-0.119	0.078	-1.521	0.128
factor(season)2021	-0.196	0.109	-1.792	0.073

factor(season)2022	NA	NA	NA	NA
I(matchweek^2)	0	0	-2.657	0.008
matchweek:factor(season)2016	-0.005	0.006	-0.758	0.449
matchweek:factor(season)2017	-0.008	0.006	-1.447	0.148
matchweek:factor(season)2018	0.01	0.005	1.868	0.062
matchweek:factor(season)2019	0.007	0.005	1.402	0.161
matchweek:factor(season)2020	0.004	0.005	0.722	0.47
matchweek:factor(season)2021	0.003	0.005	0.687	0.492
matchweek:factor(season)2022	0.005	0.007	0.68	0.497
home:attendance	0.098	0.035	2.805	0.005
home:covid_impacted	0.177	0.132	1.34	0.18
home:stadium_capacity	-0.013	0.026	-0.486	0.627
home:stadium_crowd_density	-0.426	0.209	-2.036	0.042

The full summary table contains too many rows, and hence we have taken a snippet for display. [4]

The R^2 and adjusted R^2 are 0.101 and 0.086 respectively, which is clearly low, potentially signifying the amount of noise and unobservables captured in a football game that determines the number of cards. As expected, there is a strong overall significance of the model, with an F statistic, 'num_df' and 'den_df' of 7, 226, 1.3454×10^4, which gives a p value of essentially 0.

Team, opponent, referee effects

For the categorical variables, coefficients are interpreted with respect to the base category. The teams coefficients are in relation to Arsenal (all else constant) - for example, Liverpool has a statistically significant **0.40** ($p < 0.001$) less YCs than Arsenal, all else equal. Likewise, the opponent coefficients are in reference to Aston Villa, and the referee coefficients to Andre Marriner. Interestingly, few of the `team` effects are statistically significant[5], but many of `opponent` effects are - suggesting that a specific team can expect to get a varying number of YCs when playing against teams. E.g. A team is likely to get **-0.63** ($p < 0.001$) YCs playing against Liverpool than playing against Aston Villa. The team you are up against strongly determines your YC.

Fewer levels of referees achieved low p values[6], suggesting that, holding all else constant, the perception that referees differ in their disciplinary style and card-awarding behavior may be overblown, although it could be an issue of power of the test.

Others

Home, competition level

The standalone home effect is negative at **-0.14**, as expected, but did not achieve significance ($p = 0.33$). This is possibly due to the confounding effects through attendance, crowd density etc, that have been included in the regression. In terms of league, relegating to the championship sees an increase in YC of **0.20** ($p = 0.005$).

Implied

An increase of 1 standard error of implied supremacy corresponds to a decrease of **0.11** ($p < 0.001$) YC, while the same for implied total goals leads to a drop of **0.04** ($p = 0.002$), both of which are highly statistically significant. As mentioned, implied supremacy could be a proxy for the level of competition or intensity in a football game - when the match is uneven (one side is much stronger and hence higher supremacy), there may be less intensity on the pitch and a lower need for YCs, although more research is needed to prove this correlation or causality.

Matchweek

[4] The full table can be found in the appendix.
[5] See full table
[6] See full table

The effects related to matchweek turn out to be either weak or statistically insignificant or both, including interactions with season. Lagged YCs failed to achieve significance as a predictor. (This could be unique to England, as seen in the seasonality plot)

Covid, attendance, and others

The `covid_impacted` period saw a reduction of **.14** YC for the away side, but an increase for home side (about 0.177 YC), reaffirming the reversal in home advantage in empty stadiums, albeit statistically insignificant.

For general variation in attendance, a 10k increase in spectators results in an decrease of **0.079** YC ($p < 0.01$) for the away team, and is nullified for the home team (-0.079+0.098 = 0.19 YC), an unexpected result. In this case, perhaps more granular split of the home and away attendance would be more useful, although in most cases the home fans take up a much larger proportion of the attendance.

Stadium distance also ultimately proved to have no effect on YC. Seating capacity failed to gain any (statistically) significant effect, for both home and away.

Finally, in agreement with contextual knowledge, crowd density is instrumental for home advantage, increasing the YC for away team by 0.5 on average, and which drops by 0.426 to 0.074 for the home club with statistical significance.

Top (and bottom) referees, teams, opponents

Table 5: Teams/Opps/Refs with most YCs

Group	Name	Coefficients
team	Man United	0.244
team	Charlton	0.193
team	Milton Keynes	0.157
opponent	Coventry	0.096
opponent	Milton Keynes	0.089
opponent	Blackpool	-0.028
referee	Antony Coggins	0.955
referee	Phil Dowd	0.683
referee	Missing-England-2	0.568

Ignoring p-values and focusing on conditional means, we look for teams, opponents and referees that are involved with the most number of YCs. Manchester United takes the cake when it comes to getting cautioned, with 0.25 cards more than the base team (Arsenal), followed by two Championship teams Charlton and Milton Keynes. For opponents that one might draw the most YCs against, Conventry is top, and Milton Keynes appears again. The 3rd place for Opponents would be Aston Villa, being the base level (Blackpool as 4th). Finally, teams would love to avoid Antony Coggins and Phil Dowd, who are likely to punish bad behavior on the pitch with a yellow (base refereee = Andre Marriner), and the unknown referees in the Championship tend to be stricter as well. On the other hand, the lowest awarding/ receiving referees/teams are as follows (Table 5):

Table 6: Teams/Opps/Refs with least YCs

Group	Name	Coefficients
team	Peterborough	-0.469
team	Liverpool	-0.400
team	Burton	-0.398
opponent	Rotherham	-0.993
opponent	Burton	-0.907
opponent	Burnley	-0.795
referee	Scott Mathieson	-1.355
referee	Mathew Buonassisi	-0.941
referee	Craig Hicks	-0.695

7.4.1.2 Evaluation of linear model

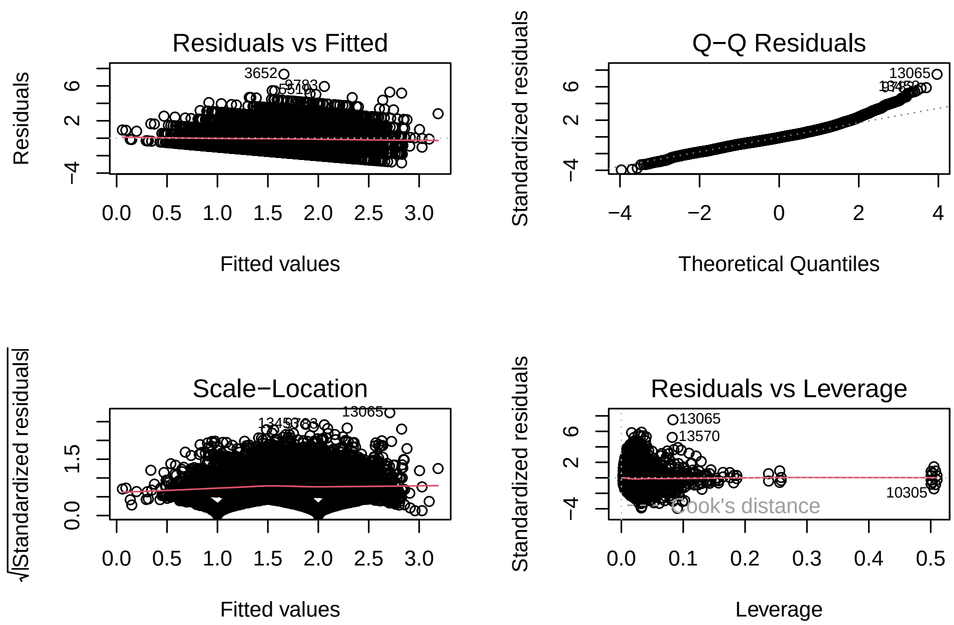


Table 7: Snippet of Outliers/High leverage points

team	opponent	home	yellow_cards	referee	season	competition_level	attendance
Middlesbrough	Fulham	0	3	Keith Stroud	2022	2	1.6058
Crystal Palace	Man City	0	3	Andre Marriner	2022	1	5.3014
Barnsley	Swansea	0	4	Antony Coggins	2020	2	1.7097

We do not expect to see pleasant diagnosis results from fitting a linear model to a count variable, from the diagnostic plots.

The Residuals-Fitted plot shows systematic downward sloping parallel lines, likely due to the discrete nature of YC. At low fitted values, the model overestimates the true YC (more positive residuals), but as the fitted values the residuals get steadily more negative.

The normal qq-plots indicate departures from normality at both extremes. From both Residuals-Plot and the Scale-Location we see an increasing variance and a few cases of high leverage points whose rows are shown. These rows could potentially be due to a rare combination of predictors or contain abnormally high number of cards (Table 7). The arc shapes in the Scale-Location point to the discreteness of the outcome, where the model occasionally acheives perfect predictions for YC = 1, 2 or 3.

Multicollinearity is likely and will be discussed in the last section with the full models.

Other limitations

It is possible that a lot more complex interactions exist in determining the number of YCs than we were able to specific in our linear model without ending up with too many parameters, such as team*opponent interactions for rivalries. According to talksport.com (2019), the Tottenham-Arsenal derby has seen 227 YCs since the beginnings of the league, trailing behind Chelsea - Manchester United with 251, and just in front of Arsenal - Manchester united (225). It is clear that some clubs have animosities that lead to more above average number of cards. `stadium_dist` would be a decent proxy but limited to geographical rivalries.

Furthermore, the linear model lacks a nesting/hierarchical structure, where some teams (or referees) only appear in one league and not the other.

7.4.1.3 Repeat Linear Model for other countries We now repeat for the other countries, for completeness, and focus a little less on interpretation and more on model performance and generalisability. Predictions are very sensitive to rank deficiency for the other countries (and the models further down the line), so we stick with choosing statistically significant predictors (a method with limitations but for simplicity and consistency for the rest of the analysis).

- Basic linear model
- Add polynomial effects for matchweek
- Add interactions

Table 8: Linear Models for all Countries

Linear Models		Internal Metrics		External Metric
Country	Model	AIC	RMSE	test_RMSE
Spain	Basic	46298.13	1.394566	2.117418
Spain	+Polynomial Matchweeks	46299.84	1.394647	2.117111
Spain	+Interactions	46212.28	1.382816	2.169076
Germany	Basic	32003.03	1.220168	1.344677
Germany	+Polynomial Matchweeks	31986.39	1.219713	1.355140
Germany	+Interactions	31903.65	1.212373	1.346848
Italy	Basic	28882.90	1.234887	1.969837
Italy	+Polynomial Matchweeks	28857.98	1.235376	1.964408
Italy	+Interactions	28759.81	1.219058	1.998082
France	Basic	37028.11	1.150181	1.070566
France	+Polynomial Matchweeks	36989.69	1.148690	1.082803
France	+Interactions	36940.81	1.141324	1.068231

8 Model 2: Poisson Regression (GLM)

8.1 Standard Poisson GLM

Contextual knowledge of the outcome of interest and the data visualization have pointed us towards the use of Poisson regression models. Generalised Linear Models (GLMs) extend the idea of modelling the mean of a normal distribution in the case of ordinary linear models to the mean of any exponential family distribution, allowing different types of outcomes to be modeled, such as binary or count data.

The Poisson GLM uses the log-link function to connect the conditional mean of the outcome to the linear estimator. This is shown as below:

$$YC \sim Poisson(\lambda), \quad Pr(YC = yc) = \frac{\lambda^{yc} * e^{-\lambda}}{yc!}$$
$$\text{where } \lambda = g^{-1}(\mathbf{B^T X})$$
$$\text{where } \; g = \text{link function}$$
$$\text{and } \; g(*) = log(*)$$

$$\text{hence } \; \lambda = exp(\mathbf{B^T X}) = E(YC|\mathbf{X}) = Var(YC|\mathbf{X})$$

Hence, we can model the conditional mean of the Poisson distribution using the same linear predictor as in the previous case.

- Good for count data (discrete and non-negative)
- As seen in exploratory analysis, YC distributions look Poisson
- Allow for non-linearity (exponential)

Notably, vanilla GLMs still assume independence of observations, which may be violated due to repeated exposure of referee and team levels. This is common in longitudinal studies, where there are repeated observations over time for repeated individuals (in our case, teams or referees), and is prevalent in healthcare or psychological studies, where causal effects (of treatments) are of interest. As noted by (Zhang et al. 2012), where GLMMs (Generalised Linear Mixed Models) account for such intra-cluster correlations with random effects, the Generalised Estimating Equations use a sandwich type variance estimator (more on this later).

8.2 Random effects (mixed models) theory

Random effects recognise that there is some hierarchical nature to the data.

From (Bolker 2015), random effects can be seen as a way to combine information from different groups. It is a compromise between 'complete pooling' (e.g. for ecological studies, when we have very few observations for levels, and do not mind ignoring the differences between site groups) and 'no pooling' (small number of sites and large number of site observations), the latter which functions as regular fixed effects.

Mathematically, this means that we assume that the effects come from an underlying distribution, whose mean and variance can be estimated from the data, usually using a normal distribution with mean 0 (without loss of generality) and unknown variance. This way, for high cardinality categorical variables, the model saves a lot of degrees of freedom by reducing the effects to draws of a distribution, instead of n_j levels and number of fixed effect estimates. For example, for a variable which we treat as a random effect $\mathbf{X}_{\text{random effect}}$, the coefficients are:

$$\beta_{\text{random effect, } j} \sim N(0, \sigma^2) \quad \text{for} \quad j = 1, 2....n_j, \quad n_j = \text{number of levels}$$

Since we assume that the random effects come from parent distribution with a well-defined population mean, then the predicted effect for each level is a *weighted average* of the level-average and the overall average across levels, with the weight inversely related to the amount of variance. This 'shrinkage' effect will thus be larger for a level with less data and more noise, being pulled more to the overall mean (Bolker 2015).

The shrinkage effect for a particular level j is given as the following (Bolker 2015) :

$$\frac{\mu_{levelj}/\sigma^2_{levelj} + \mu_{overall}/\sigma^2_{overall}}{1/\sigma^2_{levelj} + 1/\sigma^2_{overall}}$$

where, continuing from an ecological perspective, level j could represent a particular site of research.

8.2.1 Motivation for Mixed Models

As mentioned, we need random effects to factor in the likely within-cluster correlation within teams and referees (see 22. Random effects are also good for dealing with high cardinality (Laird and Ware 1982).

According to (Bolker 2015), random effects are of interest when we are willing to trade testing for differences between levels (referees etc) for making inferences about the distribution of level effects, allowing us to quantify the variance within and among groups/levels, and to generalise to levels that were not measured in the data. They are particularly useful when we have a lot of levels (388 referees, 271 teams), relatively little data on each level and even level sample sizes. (Refer to 2 and 3)

A common interpretation for taking a categorical variable as 'random effects' occurs when we view the selected levels as a 'sample from a larger population' of levels.

Therefore we combine mixed effects and GLMs to get GLMMs (Generalised Linear Mixed Models).

The paper by (Badiella et al. 2022) on the effects of cards on team performance also utilised GLMM (Poisson) to model goal count, accounting for 'experimental units that are sampled repeatedly', to provide more valid inferences.

8.3 Theory, Motivation for GEE GLM (Poisson)

A strong cause for Generalised Estimatin Equations (GEEs) come from the mispecification of models, such as in standard GLMs. (Chandler 2023) argues that the basic likelihood-based approach commonly fails with real world data, where bog-standard probability density functions with specified forms $f(x)$ do not match reality, providing wrong standard errors and thus inferences.

As such, the sandwich estimator (LIANG and ZEGER 1986), used to give a better estimate of variances regardless of specification, is the following :

$$\left[X^T\hat{W}X\right]^{-1}\left[\sum_i X_i^T(y_i-\hat{\mu}_i)(y_i-\hat{\mu}_i)^T X_i\right]\left[X^T\hat{W}X\right]^{-1}$$

Figure 24: Sandwich Estimator (Eberly College of Science, PennState, online.stat.psu.edu)

As with GLMMs, we are able to tackle (unknown) interdependencies within observations due to clustering effects. GEE-GLMs allow us to specify the specific correlation type (e.g. exchangeable, or AR1, or unstructured - see below) that we think might be appropriate for the particular use case. The main benefit is being able to produce good estimates of parameters without having to specify a full likelihood function for the model which can be difficult in some situations. GEE can operate on fewer assumptions and give population level estimates for easier interpretability, on top of producing asymptotically correct standard errors for inferences, even with correlation misspecification (Pekár and Brabec 2018).

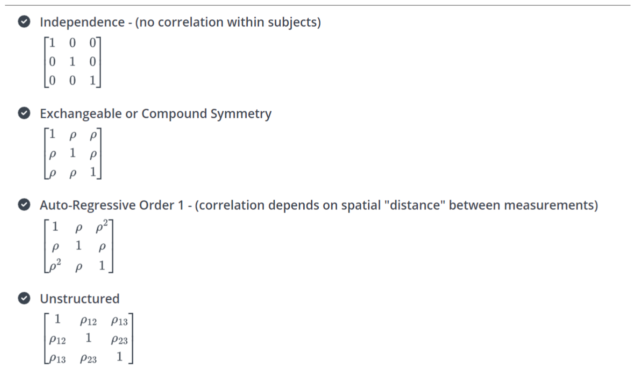


Figure 25: Types of correlation structures for clustering (Eberly College of Science, PennState, online.stat.psu.edu)

Differences between GEE and GLMM

Both GEE and GLMM seek out to factor in correlation for within cluster observations, which otherwise would violate the independence assumption in standard GLMs, albeit in different ways. From (Barrett 2019), the key difference lies in the interpretation: where GEE finds a population-level model, GLMM is subject specific, i.e. GEE looks at the average responses across the entire 'population' of teams and referees, while mixed effects focus on the individual levels. As mentioned where GLMM falls on normality assumptions for random effects, GEE avoids that altogether, providing consistent estimates regardless of correlation structure (Zhang et al. 2012).

8.4 Running the GLM models

The focus on these models will be mainly on clustering of observations.

We carry on with most predictors from the linear counterpart, but omitting predictors that cause singularity issues, due to the sensitivity of `lme4` models. First, we use a standard base R GLM to fit a `model_2.1_standardglm` model. Then, we add mixed effects, fitting a GLMM model `model_2.2_glm` via `lme4`, adding random effects on the intercept on referee and team. Finally, we fit a GEE GLM via `geepack` to get `model_2.3_geeglm_exch_1_ref`, using an interaction between referees and team.

For the GLMM, we use a crossed design (Bolker 2015) where 'at least one of the levels of each effect is repersented in multiple levels of the other effect', here being `referee` and `team`.

For the GEEGLM, however, we run into difficulties: to our best knowledge, the algorithm is unable to include separate clustering variables, which is usually reserved for one variable (e.g. patient ID). An option would be to add in an `interaction` between the variables, which would unfortunately now be clustering the variables together rather than separately. We decided to compromise by providing an interaction between `team` and `referee`. We note that the comparisons with GLMM would be much less direct now.

8.4.1 Interpreting Poisson GLM

Comparing the coefficients of the fixed effects linear model and the corresponding Poisson, we can see that the coefficients are different but are of very similar scale/order of magnitude, since the coefficient of the Poisson GLM is not directly interpretable on the same scale of the outcome but rather on the exponential scale, i.e. represents the *change in the expected log count of YC*, due to the log-linear relationship. E.g. for a 1 unit increase in X_1 with coefficient β_1, we move from mean λ_1 to λ_2

$$\lambda_2 = exp(\mathbf{b^T X}) * exp(b_1) = \lambda_1 * exp(b_1)$$

Hence there is an exponential multiplicative effect on the mean via the coefficient, rather than the additive effect of linear models. For example, for coefficient of `teamLiverpool`, the coefficients are -0.29 and -0.4 for GLM and LM, where the exponential multiplicative effect of Liverpool (with respect to Arsenal, all other variables equal) is 0.748 on the current $E(YC|\mathbf{X})$, while it is a decrease in $E(YC|\mathbf{X})$ of -0.4 in the linear model.

Table 9: GLM output table (England)

term	estimate	std.error	statistic	p.value
teamSwansea	-0.237	0.13	-1.826	0.068
teamAston Villa	-0.022	0.116	-0.189	0.85
teamMillwall	-0.151	0.133	-1.13	0.258
teamSheffield United	-0.053	0.122	-0.435	0.663
teamWest Ham	-0.173	0.119	-1.455	0.146
teamHuddersfield	-0.126	0.128	-0.983	0.326
teamWatford	0.066	0.118	0.563	0.574
teamWigan	-0.007	0.152	-0.045	0.964
teamMan United	0.134	0.112	1.19	0.234
teamLiverpool	-0.29	0.127	-2.279	0.023
...	...	...	...	...
refereeIain Williamson	0.277	0.411	0.674	0.501
refereeMathew Buonassisi	-0.606	2.099	-0.289	0.773
refereeDean Whitestone	-0.183	0.129	-1.418	0.156
refereeCraig Pawson	0.149	0.099	1.506	0.132
refereeJARRED GILLETT	0.078	0.127	0.614	0.539
refereeMichael Salisbury	-0.078	0.171	-0.454	0.65
refereeChris Foy	0.104	0.417	0.249	0.803
refereeMike Jones	0.197	0.156	1.259	0.208
refereeJames Adcock	-0.019	0.245	-0.077	0.939
refereeJohn Brooks	0.219	0.114	1.918	0.055
...	...	...	...	...
opponentNorwich	-0.14	0.103	-1.356	0.175
opponentLeicester	-0.24	0.11	-2.182	0.029
opponentCoventry	0.047	0.13	0.364	0.716
opponentWest Bromwich	-0.332	0.108	-3.076	0.002
opponentPreston	-0.29	0.11	-2.634	0.008
opponentBurton	-0.553	0.221	-2.509	0.012
opponentHuddersfield	-0.172	0.106	-1.626	0.104
opponentHull	-0.346	0.12	-2.882	0.004
opponentMiddlesbrough	-0.209	0.106	-1.977	0.048
opponentLiverpool	-0.361	0.125	-2.885	0.004
(Intercept)	0.89	0.246	3.617	0
home	-0.101	0.144	-0.699	0.485
competition_level2	0.11	0.065	1.703	0.089
covid_impacted	-0.149	0.143	-1.046	0.296
total_goals_implied_scaled	-0.037	0.021	-1.783	0.075
supremacy_implied_scaled	-0.072	0.027	-2.665	0.008
attendance	-0.043	0.025	-1.735	0.083
stadium_crowd_density	0.255	0.176	1.452	0.147
factor(season)2016	-0.026	0.171	-0.151	0.88
factor(season)2017	0.135	0.157	0.862	0.389
factor(season)2018	-0.141	0.154	-0.917	0.359

factor(season)2019	-0.113	0.148	-0.761	0.447
factor(season)2020	-0.055	0.146	-0.38	0.704
factor(season)2021	-0.044	0.165	-0.266	0.79
factor(season)2022	-0.017	0.147	-0.112	0.911
matchweek	-0.005	0.005	-0.909	0.363
stadium_capacity	-0.016	0.019	-0.826	0.409
stadium_dist	0	0	-1.014	0.31
home:covid_impacted	0.095	0.145	0.657	0.511
home:attendance	0.05	0.028	1.772	0.076
home:stadium_crowd_density	-0.231	0.223	-1.033	0.301
factor(season)2016:matchweek	-0.003	0.007	-0.496	0.62
factor(season)2017:matchweek	-0.004	0.006	-0.708	0.479
factor(season)2018:matchweek	0.006	0.006	0.974	0.33
factor(season)2019:matchweek	0.004	0.006	0.744	0.457
factor(season)2020:matchweek	0.003	0.006	0.505	0.614
factor(season)2021:matchweek	0.001	0.005	0.253	0.8
factor(season)2022:matchweek	0.006	0.007	0.808	0.419

Since coefficients operate on exponential scale, $\beta_i < 0$ corresponds to a drop in $E(YC|\mathbf{X})$ since the multiplicative effect $exp(\beta_i) < 1$, conversely if $\beta_i > 0$ then $E(YC|\mathbf{X})$ will increase. Hence, the direction operates similarly for Poisson GLMs and LMs, and as seen from the coefficient table the signs are generally the same for both.

However, an interesting note is that compared to the LM equivalent (with the same formula/ set of predictors), many numeric predictors lost statistical significance (using 5% level), such as `competition_level2`, `covid_impacted`, `total_goals_implied_scaled`, `stadium_crowd_density`, and so on.

8.4.2 Comparing Poisson Models

We take a look at how using different variants of the GLMs result in similar or different coefficients for the same selected teams and referees. This could provide an insight into whether the assumptions of the variants are strong enough or not.

Table 10: GLMs comparisons

Model	Internal Metrics AIC	Internal Metrics BIC	External Metric Test_RMSE
glm	14660.43	16338.23	1.731690
glmerMod	14512.09	14745.32	1.719791
geeglm	14660.43	16338.23	1.731690
lm	47913.49	49598.81	1.279555

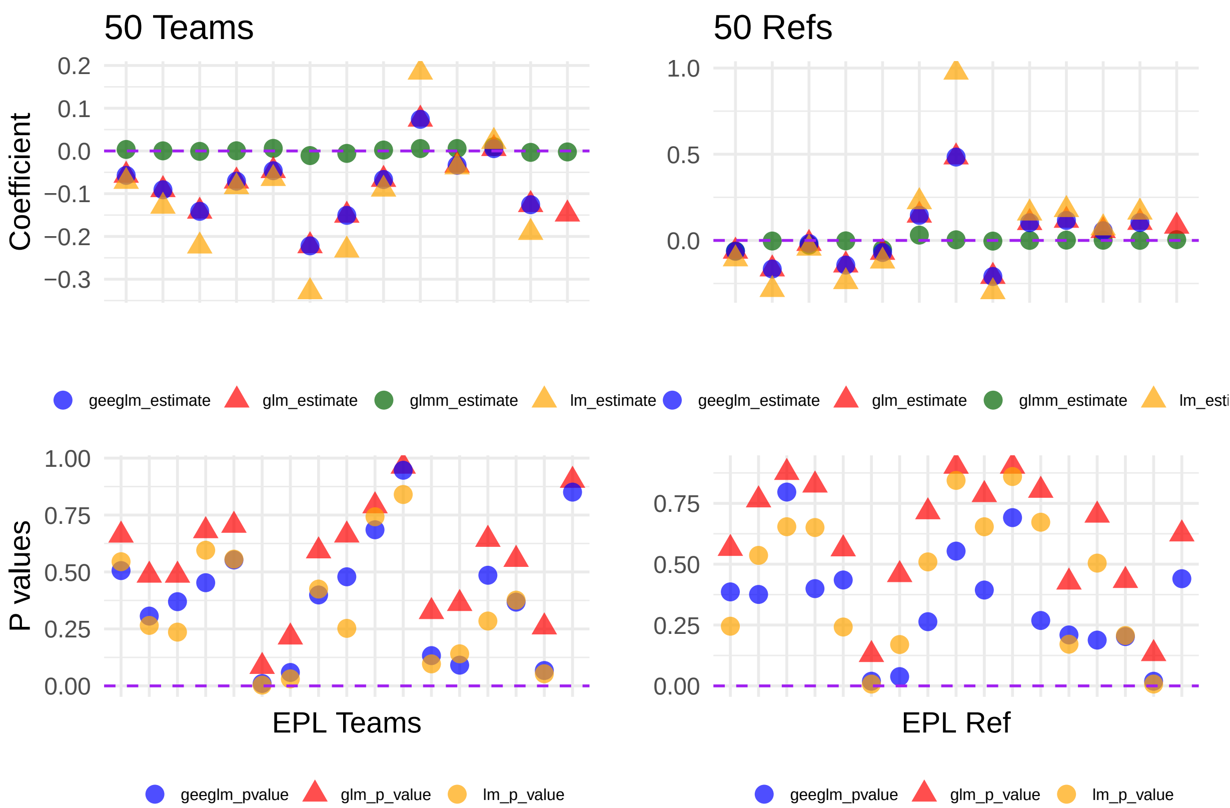


Figure 26: Coefficients for GLM, GLMM, GEEGLM and LM

GLM vs LM As confirmed in the literature review, the GLM and LM coefficients generally differ to some degree but have the same direction (sign), thus having the same qualitative interpretations. The GLM p values are generally higher than that of LM, likely due to the difference in distributional assumptions (Poisson in GLM, Normal in LM) and log-linearity vs regular linearity. As we will see later on, the AIC is much smaller for the GLM, possibly suggesting that the distributional assumptions of Poisson GLM is much stronger.

GLMM The coefficients for GLMM come from the `ranef` coefficients, which represent each level's deviation from the population mean, and we can clearly see that they have been pooled/undergone shrinkage.

GEE The GEEGLM coefficients are very similar to that of GLM, suggesting that the impact of adding a correlation structure within the clusters ('exchangable') was negligible.

Out of the **Generalised** models (that allow for a exponential link function), the GLMM gives the best results on model fit (AIC, BIC), as well as generalisation, possibly indicating using random effects for the clusters (referee, teams, etc) and the associated hierarchical shrinkage helps to explain the within cluster heterogeniety for YCs and improves the model prediction.

Nevertheless, although linear model greatly suffers in AIC and BIC, possibly due to distributional misspecifications, it excels on prediction power, significantly lower than the generalised models.

8.4.3 Other evaluation:

We tested for dispersion using `DHARMa` package (Hartig, 2022), which provides a simulation based methods to create residuals for generalised linear mixed models.

Although the models clearly do not suffer from overdispersion, commonly the first check, they suffer from the opposite, with p values of essentially 0 for all underdispersion tests. Subsequently, Negative Binomial models were fitted, but revealed to have not much difference in prediction power to Poisson (refer to appendix), hence we carry on with the Poisson models.

Another limitation is that the variances of the random effects for all 3 categorical variables are very low, at 0.0033145, 0.0004137, 0.0005821 respectively. This implies that variability between levels are low, and perhaps not well captured by a mixed effects model.

Table 11: Linear Models for all Countries

Linear Models		Internal Metrics		External Metric
Country	Model	AIC	RMSE	test_RMSE
Spain	glm	14817.393	1.387281	2.015109
Spain	glmerMod	14675.190	1.402827	2.020756
Spain	geeglm	14817.393	1.387281	2.015109
Spain	lm	46224.950	1.385882	1.471411
Germany	glm	10095.836	1.215432	1.685082
Germany	glmerMod	9954.052	1.227985	1.669000
Germany	geeglm	10095.836	1.215432	1.685082
Germany	lm	31923.198	1.213620	1.162139
Italy	glm	11452.972	1.223263	2.131799
Italy	glmerMod	11246.399	1.246155	2.110841
Italy	geeglm	11452.972	1.223263	2.131799
Italy	lm	28778.115	1.221925	1.417898
France	glm	11384.610	1.142579	1.531257
France	glmerMod	11211.749	1.161133	1.521788
France	geeglm	11384.610	1.142579	1.531257
France	lm	36941.256	1.142788	1.036368

Note:

Upon latest fit, Germany and Italy returned a boundary singular fit, likely for glmer - proceed with caution

8.4.3.1 Repeat Poisson GLMs for all other countries Across the board from Table 11, LM had the best prediction power on test set despite (or because of) a lower fit to the training data, especially AIC. Amongst the generalised models, GLMM tended to have the best AIC fit as well as prediction power. GEEGLM failed to improve on GLM, hence will be omitted henceforth.

[7]

[7] For consistency, we used the same formula for the models across all countries - which resulted in ‘singularity fit’ warnings for Germany and Italy, most likely when using the GLMM fitting algorithm. Future research could perform better model selection to reduce multicollinearity, which is the suspected reason for the issue.

9 Model 3: Random forests/ Tree based models (full data models)

9.1 Motivation for tree-based methods

As we proceed to non-parametric, non-linear models in hopes of achieving better predictability, a particular machine learning algorithm stands out - tree based models.

- High dimensionality

Tree based models are more resistant to additional 'non-informative predictors' than linear and non-linear models, growing in RMSE much less slowly than the others when the number of predictors increases. Thus they are less susceptible to dimensionality issues, and do not attempt to partition the input space (predictors) across all values/levels. (Boehmke and Greenwell 2020)

Moreover, trees efficiently disregard unnecessary variables or levels, only considering those that aid in prediction ("University College London STAT0030 Statistical Computing, Lab 7" 2022-2023). This could be particularly beneficial for high-cardinality.

Furthermore, from (Hastie, Tibshirani, and Friedman 2009),

- Tree-based models can capture interactions between predictors
- Handle missing data (*implementation subject to the specific R package*) [8]
- Naturally handle both continuous and categorical type predictors

9.1.1 Decision Tree Theory

The commonly used method is the Classification and Regression Trees (CART) model from (Leo Breiman et al. 1984) and relies on the partitioning of input feature space recursively, via binary splits. For the case of classification, the split is done to minimise the gini impurity or entropy, while for regression the criterion could be MSE as a measure of goodness of fit. The tree searches through the values of the predictors and picks the best split to separate the observations. This is done until some stopping mechanism, such as reaching a minimum node size, or a minimum level of entropy, etc. This is considered 'greedy' - the algorithm optimises for the best criterion (gini/entropy/MSE) at the split level, instead of the entire tree. After the tree has been fitted, a prediction will be determined by the leaf node it falls into, where the output will be the most common class (classification, as shown in the image below) or the mean value (regression, see below) of the leaf node.

[8] Initially, but no longer useful for our data after updated version

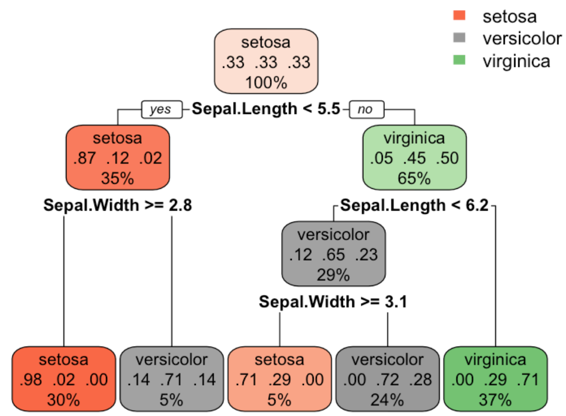


Figure 27: Decision Tree on Iris Data (Boehmke and Greenwell, 2020)

For the regression case,

$$f(x) = E[Y|\mathbf{x}] = \sum_{1}^{K} \mu_k * I(\mathbf{x} \in R_k)$$

, where $R_1....R_K$ are the partitioned space of the input space, and μ_k represents the mean value of the particular partitioned space R_k ("University College London STAT0030 Statistical Computing, Lab 7" 2022-2023).

However, it is well known that decision trees are prone to overfitting, and the internal structure can be unstable. Hence, they are mostly used as units in ensemble methods that draw on their benefits while improving generalisability.

9.2 Random Forest Theory and Benefits

Random forests (RF) rely on the 'wisdom of the masses', being an ensemble of independently trained trees, from which we take the aggregate decision. From (Hastie, Tibshirani, and Friedman 2009), RFs use the concept of bagging (bootstrap aggregation) to enhance generalisability of the overall model. This works by using bootstrap sampling to train trees on different subsets of the original dataset, and further decorrelates the trees by considering a random subset of predictors to pick from at each split. After fitting the trees, a linear combination of each of the individual 'votes' gives the final prediction, where each tree is given equal weight. The algorithm ((Hastie, Tibshirani, and Friedman 2009)) is the following:

Algorithm 17.1 *Random Forest for Regression or Classification.*

1. For $b = 1$ to B:
   (a) Draw a bootstrap sample $\mathbf{Z}^*$ of size N from the training data.
   (b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
      i. Select m variables at random from the p variables.
      ii. Pick the best variable/split-point among the m.
      iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$.

To make a prediction at a new point x:

Regression: $\hat{f}_{\rm rf}^B(x) = \frac{1}{B}\sum_{b=1}^B T_b(x)$.

Classification: Let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}_{\rm rf}^B(x) = \textit{majority vote } \{\hat{C}_b(x)\}_1^B$.

Figure 28: Fitting Algorithm for Random Forests (ESL)

More can be read from (L. Breiman 2001).

There are numerous strengths of using RFs. (Hastie, Tibshirani, and Friedman 2009) claims that the model does competitively well in general, such as on the canonical `spam` dataset example, and requires 'very little tuning', in agreement with (Probst, Bischl, and Boulesteix 2018) who show little differences when tuned. Furthermore, they can be efficiently trained through parallelisation with modern implementations, such as `ranger` and `h2O` (Boehmke and Greenwell 2020).

9.3 Data Preprocessing (full data)

The full data is used now. We use the same train test split as before, which focuses on testing most recent games for better relevance. Models are still trained with case weights that prioritise more recent observations.

9.4 Running the RF models

For RandomForests, we use the `ranger` package (Wright and Ziegler 2017) - RANdom forest GEneRator(ranger) - which is very applicable to our high dimensional dataset, being built for genomic data such as GWAS (genome-wide association studies. The authors tested `ranger` against the other most commonly used random forests packages, such as `randomForest` (Liaw and Wiener 2002), on simulated data, and `ranger` was shown to outperform the rest in terms of runtime for an increase in number of trees, features, sample size and `mtry` i.e. number of variables considered at each split.

[9]

Ranger Results

The benefit of bagging is the left-out subset of data that is not used to train the trees i.e. the out-of-bag (OOB) data that the model can test on and give an estimate on test error. Here, we get an OOB MSE of 1.63, with an R^2 of 0.10. The number of trees used was 500, and the `mtry` was at a default value of 4.

We set a hyperparameter grid space of different combinations of `mtry` and `num.trees` and tested their OOB RMSE (Boehmke and Greenwell 2020), and found little variance in the RMSE. Hence we stick to the default values. Here, the OOB RMSE serves as a proxy for validation error.

[9] Limitation: this regression-based random forest algorithm optimises on RMSE, which is unideal for count data (better to use deviance as metric). We sacrifice this for the computational ability of the ranger for our large dataset.

Table 12: Results from hyperparameter tuning for ranger

mtry	num.trees	OOB_RMSE
2	700	1.273567
4	700	1.274880
2	500	1.274979
4	500	1.275404
2	300	1.275992
6	700	1.276391

9.4.1 RF evaluation: Variable importance

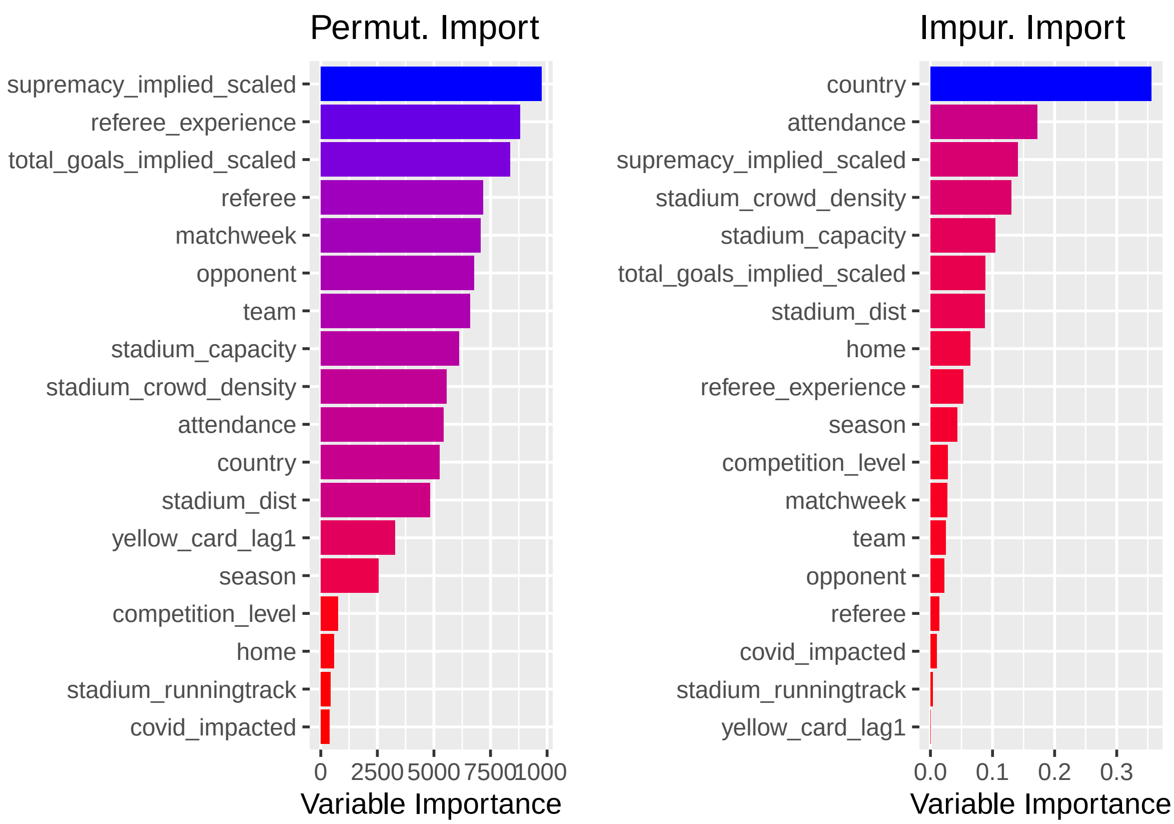


Figure 29: Variable Importance: Permutation vs Impurity

We plot the variable importance (Sam 2018) in 29, which shows us the decreasing order of 'amount of contribution' each predictor goes into determining the final outcome. Usually, for bagged decision trees, feature importance corresponds to the amount of variance explained (SSE) by each predictor in a tree across all splits, aggregated across all trees, which is the case with `importance = 'impurity'`. With `importance = permutation`, we allow for each OOB sample to be passed down a tree, and then we randomly permutate each feature to see how much the final outcome is affected (Boehmke and Greenwell 2020), which serves as a proxy for how influential a variable is in determining the outcome. According to the sklearn documentation, '*Permutation feature importance overcomes limitations of the impurity-based feature importance: they do not have a bias toward high-cardinality features and can be computed on a left-out test set*'. Hence impurity alone is not sufficient and we need permutation to allow for some generalisability, and we may expect our high cardinality variables (referee, team, opponent) to be over-represented.

Interestingly, for **impurity**, we do not see the aforementioned categorical variables on the leaderboards - instead, we have `country`, followed by `attendance` and `sup_implied`, with `country` leading by a large margin. Instead, they are found close to the bottom, above `covid_impacted`, `stadium_runningtrack` and `yellow_card_lag1`.

As for **permutation**, the top variables are `sup_implied`, `referee_experience` and `tg_implied`, having one common variable in the top three (`sup_implied`). Here, the highly cardinal variables take a much higher position (4th, 6th and 7th places). This might be a better indication of their performance. `covid_impacted` was of low importance for both metrics, and `running_track` was relegated to the bottom as well.

Unfortunately, the importance metrics represent 'global importance' for the variables and we are unable to see the interesting interactions that we used the model for.

9.5 Gradient Boosting Theory

Gradient boosting, like random forest, is an ensemble method that commonly makes use of trees as the base learners. The idea of boosting is to convert individual weak learners that are only slightly correlated with the outcome.

Unlike random forests, here we iteratively add trees sequentially, boosting the performance by optimising the latest tree on the most recent residuals of the overall model ("University College London STAT0030 Statistical Computing, Lab 7" 2022-2023). For the regression case, we can represent the ensemble via

$$E(Y|\mathbf{x}) = f(\mathbf{x}) \equiv \sum_{m=1}^{M} \beta_m b(\mathbf{x}; \gamma_m)$$

whereby each $\beta_m b(\mathbf{x}; \gamma_m)$ represents a tree with its own parameter, and are fitted via the following algorithm (UCL 2023)

1. First, define $\hat{f}_0(\cdot) \equiv 0$.
2. Repeat, for $m = 1, 2, \ldots, M$, the following two stages: i) solve for $(\hat{\beta}_m, \hat{\gamma}_m)$ by minimising $\sum_{i=1}^{n} \left\{ y_i - (\hat{f}_{m-1}(\mathbf{x}_i) + \beta_m b(\mathbf{x}_i; \gamma_m)) \right\}^2$ and ii) set $\hat{f}_m(\cdot) \equiv \hat{f}_{m-1}(\cdot) + \hat{\beta}_m b(\cdot; \hat{\gamma}_m)$.
3. Return $\hat{f}_M(\cdot)$ as an estimate of $f(\cdot)$.

Figure 30: Gradient Boosting Fitting (UCL 2023)

The second step illustrates the fitting process: adding a new unit that minimises current residual (SSE) from the fitted model, then adding it into the ensemble. This optimisation process is done via gradient descent, which gives the name in "gradient boosted models".

9.5.1 Gradient Boosting Benefits

Notably, for data outside the realm of image and natural language processing, boosting attains similar standards as neural networks on tabular data, with XGBoost being the mainstay for practitioners (Kossen et al. 2021). Gradient boosted models (GBMs) have been popular across many domains, and is highly used on Kaggle, which famously had 17 challenge winning that used gradient boosted methods (XGBoost) for model training (Chen and Guestrin 2016). [From the Elements of Statistical Learning, Friedman et al.(Hastie, Tibshirani, and Friedman 2009) show how boosting outperforms random forests over a specific number of trees on the `California housing data`, where the latter stabilised at around n.trees = 200, meanwhile the boosting method continued to improve, in terms of average absolute error.

Table 13: Results from hyperparameter tuning for gbm

interaction.depth	n.trees	shrinkage	OOB_RMSE
5	1200	0.1	0.8561255
5	1000	0.1	0.8881735
4	1200	0.1	0.8894688
4	1000	0.1	0.9189933
5	800	0.1	0.9228099
4	800	0.1	0.9583580

9.6 Running the GBM

For our particular task, we use the `gbm` package (Greenwell et al. 2022) that implements Friedman's generalisation of Adaboost back in the 1990s (Freund and Schapire 1999) to regression problems and other loss functions (Friedman 2001). It allows for the appropriate Poisson distribution to be specified, and hence potentially a better fit (the random forest approach via `ranger` lacked this).

For hyperparameter tuning, we use the same strategy as before and calculate the OOB RMSE as an indication of validation performance. Higher (3 onwards) interaction depths were chosen to allow for interactions between predictors, which should not overfit too much considering our high dimensionality with a decent sample size.

From the gridsearch calculations, the first thing we notice is that we have finally crossed below the RMSE of 1 mark. The optimal hyperparameters are: 5, 1200 and 0.1 for the `interaction.depth`, `n.trees` and `shrinkage`, which is similar to learning rate - notably, we are choosing the most complex models (highest number of most complex trees).

9.6.1 gbm evaluation: Variable importance

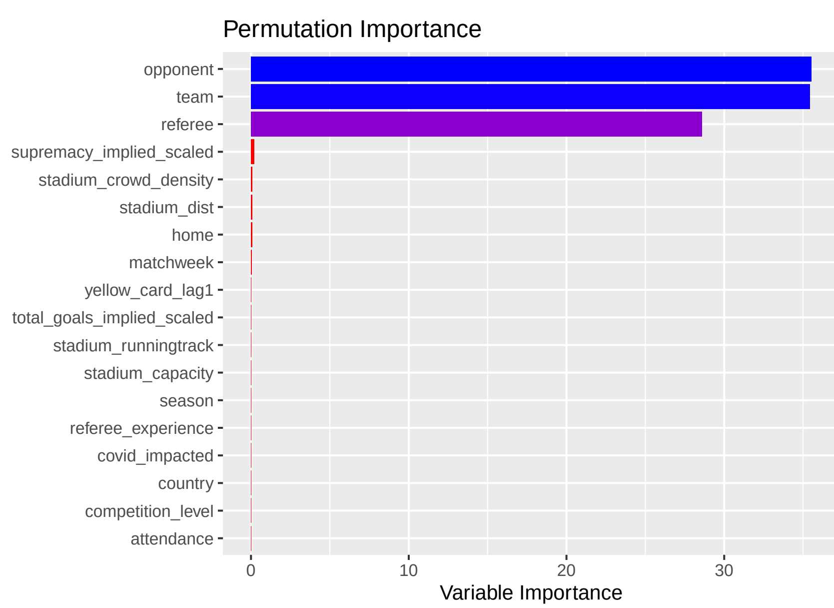


Figure 31: gbm Variable Importance

Here we see the obvious shift away from the other variables towards the 3 high cardinality variables, with a distant 4th being `sup_implied`. This `gbm` model emphasises that the number of cards are most commonly determined by the two teams who are playing against each other, and the referee officiating, though significantly less important.

If the `gbm` is indeed the appropriate model, it is possible to make sense of the results. Cards may be strongly determined by the two teams playing against each other, and they capture the other information such as crowd density, goals likely to be scored, rivalries, etc. Referees ostensibly play a big part, with some being stricter or more lenient than others, and referees may be primed to behave in a certain way given the teams they are disciplining.

There is an obvious sparsity here compared to the random forest version - gradient boosted trees are made of relatively simpler base learners (decision stumps) compared to that of random forests which are deeper, although we had allowed an interaction depth of 5 in the former. Thus it is natural that boosting ends up avoiding some variables completely, unlike random forests (Friedman 2001).

However, a simple check on the test set (shown later in model comparison) reveals that our strategy of tuning results in a very high test RMSE of over 4. To attempt to reduce this, we try the tuning method of (Boehmke and Greenwell 2020). Unlike Random Forests, GBM is sensitive to changes in hyperparameters (Probst, Bischl, and Boulesteix 2018).

The authors suggest a strategy of first finding the learning rate, then optimal number of trees, before tuning the tress specific parameters. They also used a `cv.fold` value of 10, which should provide a decently robust estimate of errors. Their method includes tuning `n.minobsinnode`, which further controls tree complexity. The result of the gridsearch is below, and since the top three have the same RMSE, we choose the least complex (3rd) option, to refit another gbm model `trees_boosted_poisson_fit_2` or `gbm2`

We also apply the authors' tuning method for the ranger random forest model to produce `RF2`, which tuned more parameters such as `sample.fraction` and `min.node.size`, which control the fraction of observations from the training set to use for fitting and tree complexity respectively.

So far we have explored separately: mixed effects and tree-based models to deal with high dimensionality, namely with the high number of teams and referees, which also accommodates for levels with fewer levels of data (at least for mixed effects models).

9.7 GPBoost theory, benefits

Now, we combine the two to get a tree-boosted random effects model via work from (F. Sigrist 2020). As the author explains, although tree-boosted models are generally regarded as the 'most effective off the shelf non-linear learning method for a wide range of application problems' (Johnson and Zhang 2014), they are not perfect where the response variables are dependent and with high-cardinality.

The R package `gpboost`, using the GPBoost (Gaussian Process Boosting) (F. Sigrist 2020), introduces a novel method of combining boosting with Gaussian processes and mixed effects. They are essentially flexible, non parametric models for regression and classification and are able to achieve high prediction power. They are a method of defining a distribution over functions with a multivariate Gaussian distribution, fully defined by mean and covariance functions. In GPBoost, an ensemble of trees are used to model the non-linear prediction function (a relaxation of the linear assumption in Gaussian processes). We are also able to specify the likelihood function for the outcome variable as Poisson.

From (F. Sigrist 2023), the algorithm has also been shown to outperform other similar variants, such as basic linear mixed effects models, to deep neural networks with random effects, on an array of datasets chosen due to the precense of high-cardinality categorical variables, from Airbnb, IMDb, etc.

9.8 Running GPBoost

Hence, we fit a `gpboost` model, following the model building steps and hyperparameter tuning from (Fabio Sigrist 2023), via a gridsearch and a 10% validation split from within the training data, using MSE as the criteria. The optimal parameters are given in 14

Table 14: Best parameters for GPBoost (from tuning)

	unlist(opt_params)
best_params.learning_rate	0.100000
best_params.max_depth	5.000000
best_params.min_data_in_leaf	100.000000
best_params.lambda_l2	10.000000
best_iter	17.000000
best_score	1.599799

9.8.1 GPboost evaluation: Variable importance

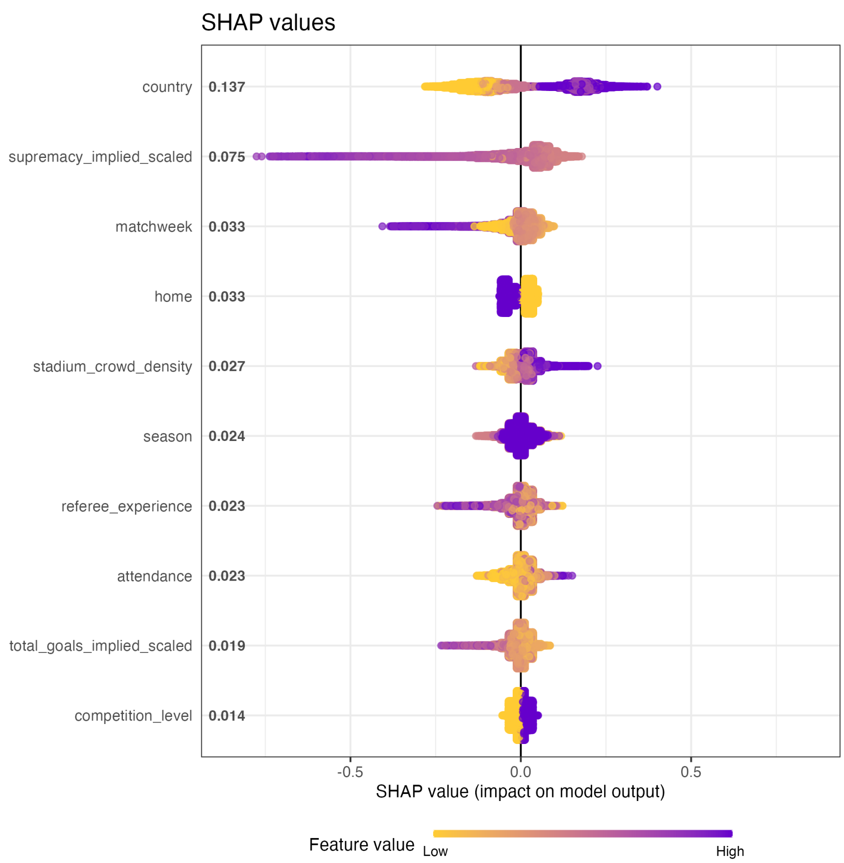


Figure 32: SHAP values for GPBoost

SHAP (SHapley Additive exPlanation) values are derived from game theory and represents a type of feature importance. In this case, this could only be used to examine non-cardinal predictors. We can see that `country` was the most important, followed by goal supremacy, then jointly `matchweek` and `home`. From the wide range of SHAP values of goal supremacy, we can see that this feature has a varied effect on the outcome on the different observations, more so than e.g. `matchweek`. The colour shades also tell us that generally, higher goal supremacy leads to a negative effect on YC, which was corroborates with our exploratory analysis.

10 Testing and comparing all models: LM, GLM, RF, GBM

To compare all models, we refit a linear model (LM), a generalised linear model (GLM), a mixed effects linear model (LMM) and generalised linear mixed model (GLMM) with the full dataset, adding in some interactions involving `home` and `country` predictors. The summary tables will be shown in the appendix for curious readers (e.g. country effects previously not included, or to compare coefficients with the England only case).

Nested vs non-nested mixed effects models

With each country and league having its own unique set of teams (and referees), this introduces a hierarchical structure, where the original high-cardinality variables are now nested within countries. If we assume that each country has its own idiosyncracies pertaining to referee and team effects, this would be captured by the nesting structure. We fit LMMs and GLMMs with and without the nested structure, and test with an ANOVA test to see if the nesting effect is significant.

Once again, the mixed effects models suffered from singularity when all predictors were chosen and a simple method of model selection via choosing significant predictors from the full linear model was employed (a method that understandably has limitations), as the `lme4` documentation recommended against overly complex models in such situations.

10.1 Results

The ANOVA likelihood test revealed the nesting by country to be insignificant which also added complexity that resulted in singularity warning. Hence, we go ahead with the non-nested versions.

10.2 Interesting results (full linear model)

Table 15: Some results from full linear regression

term	estimate	p.value
(Intercept)	-58.528	0.022
home	-0.234	0.000
competition_level2	0.106	0.000
stadium_crowd_density	0.360	0.000
covid_impacted	-0.119	0.000
total_goals_implied_scaled	-0.107	0.000
supremacy_implied_scaled	-0.121	0.000
poly(matchweek, 2)2	-5.736	0.008
home:covid_impacted	0.262	0.000
countryItaly:poly(matchweek, 2)1	-7.831	0.053
countrySpain:poly(matchweek, 2)1	9.746	0.007
countryItaly:poly(matchweek, 2)2	-13.480	0.001
countrySpain:poly(matchweek, 2)2	-7.486	0.032
countryGermany:stadium_runningtrack	0.232	0.022
countryItaly:stadium_runningtrack	0.200	0.014

[a] High Cardinality categorical variables filtered out, alongside non-significant coefficients

We look at the results from the linear model (fixed effects), filtering out the highly cardinal categorical variables (which are too rows many to display), variables with p values greater than 0.1 and variables with coefficients that are too small (less than 0.1).

We now see that `home` effect has finally gained significance and the polynomial effects of `matchweeks` become much clearer now with the full data instead of the England-only subset. The shape of the polynomial differs between countries, as with the running track effect.

10.3 Full comparison

Table 16: Final Model Comparisons

Model	Internal		External
model	AIC	Internal.RMSE	Test.RMSE
RF2	NA	1.6160979	1.731010
gpboost	NA	1.5997987	1.731747
RF1	NA	1.6259491	1.733802
LMM	192976.54	1.2510770	1.741221
LM	192474.47	1.2491650	1.793032
GLMM	61487.81	0.5686205	3.803085
GLM	62267.36	0.6496406	3.884857
gbm2	NA	1.0301570	3.962545
gbm1	NA	0.8561255	4.242806

[a] Internal RMSE estimates vary in precision - e.g. gbm2 (Boehmke 2020) used CV, likely more precise than gbm1, and are OOB estimates for tree-based models

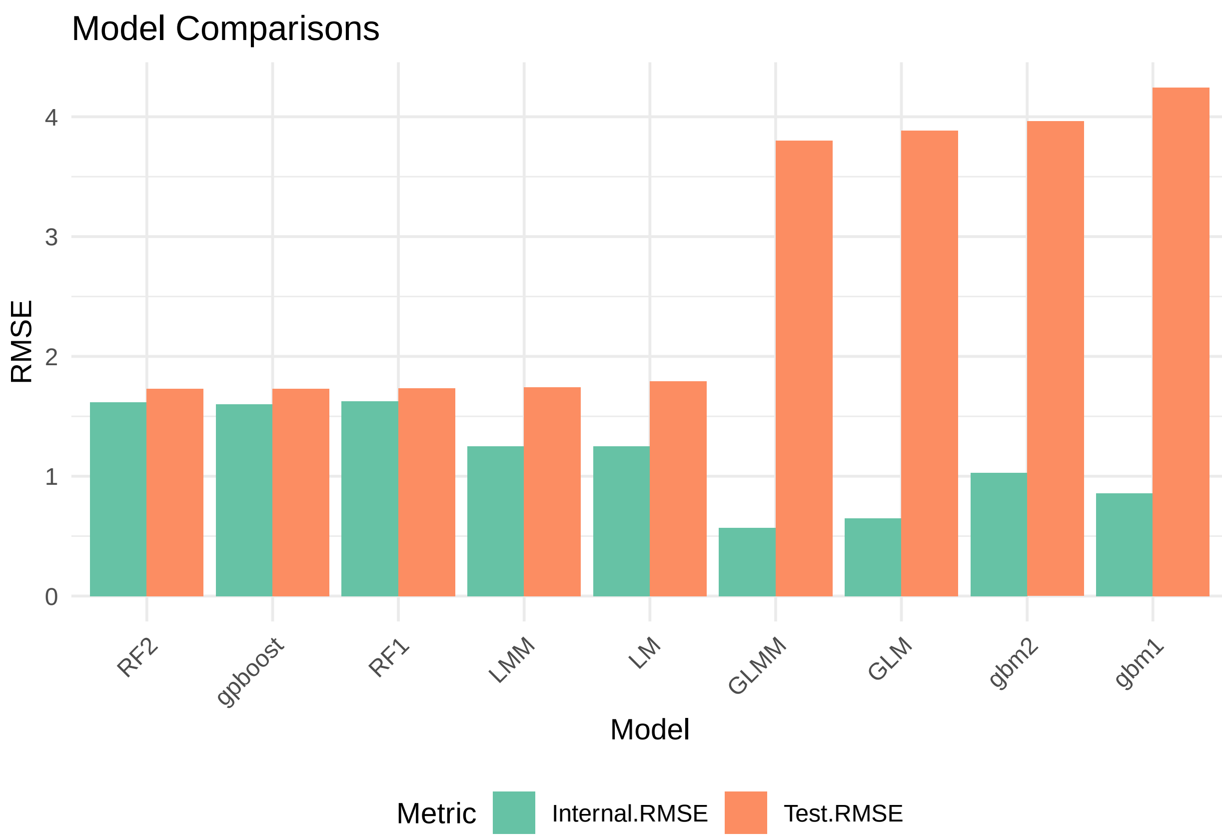


Figure 33: Comparison across all models

We arrange both the final comparison Table 16 and Fig 33 in order of 'Test RMSE'. We can see that RF models do have the best fit, they also generalise the best to unseen data, although the slight increases in test error the linear models make up quite a lot in terms of model fit (much lower 'Internal RMSE'). The GPBoost model, although more similar to the gbm model as they are tree-boosted algorithms, fares very closely to the random forest models. The generalised linear models and gbm ones fared the worst, with more than double of test error - this is could be related to the fact that the train error is so low, and are likely attributed to overfitting. In terms of the best train fit, the Poisson log-link GLM takes the top spot.

We moved from simple models, to mixed effects, to tree-based models, each time increasing complexity to take advantage of some characteristics of the data. Adding mixed effects to (generalised) linear models do not seem to drastically change results, but did for tree boosted models, as with the gpboost model. The closeness in performance of linear models to that of random forests could suggest that the ability of the latter to handle non-linearity, interactions and outliers was less fruitful than expected. For ensemble models, training independent trees instead of sequential ones seem to fit this dataset much better by ostensibly discouraging overfitting.

10.3.1 Evaluation of model : analysing residuals

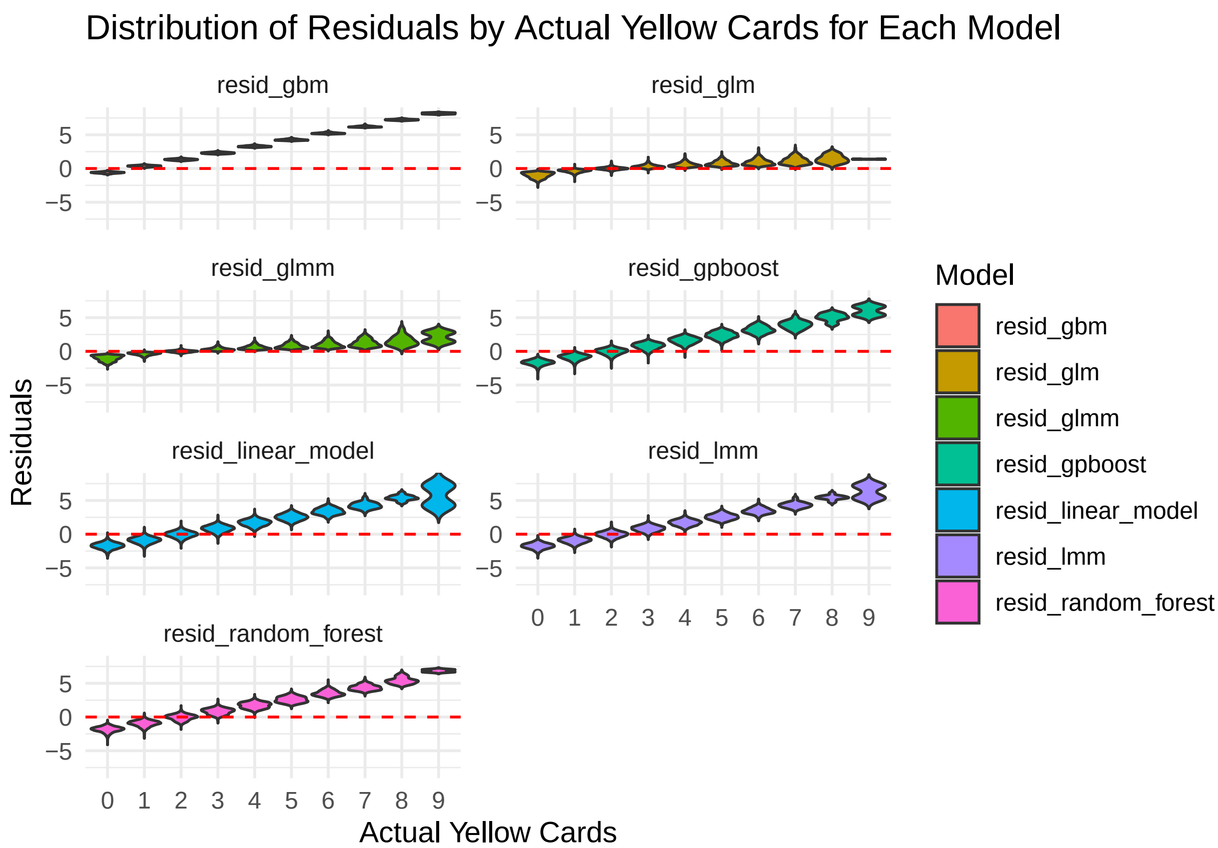


Figure 34: Residual Analysis across models (Training set)

There is a very clear pattern across all models, whether the model accounts for random effects or uses tree-based models, or both: for low values of YCs, models tend to overpredict for small values of YC, and underpredict for large values of YC. As with the previous table comparing the performances 16, the simpler, parametric models achieved the best fit, barring the linear mixed model - but at least for the generalised versions that used log-links (Poisson family), the generalisability was badly affected.

The obvious systematic increase residuals as the number of YC increased could be the result of many possibilities:

- model mispecification: the model may not be capturing the right relationship between the predictors and the outcome. However, the pattern is very similar across all models, which are capturing different relationships here (e.g. linear model captures linearity between the predictors and the outcome, glmm captures log-linearity, tree based capture complex non linearity) suggests that it may not be the issue with any one model specification.
- unobservables: there are likely to be omitted variables that are missing from our dataset that are correlated with YC and possibly with some of the the predictors. This could also be due to uncaptured interactions, as we were unable from including all possible interactions in the linear models (too much memory required to include home team and away team interactions).

It is most likely that the outcome in question, YCs, is too noisy and the available variables from the dataset cannot capture the variability in the outcome. For example, it is likely that an unpredicted flare up or bust up that occurred between the two teams caused a series of YCs, which cannot be captured by the goal supremacy, teams, attendance etc.

It is also very likely that the model has been biased towards the high imbalance in the dataset, where most of the data come from games with 1 or 2 YCs, where the latter is the mode of the distribution. Collectively, they make up 54 % of the number of teams (per match).

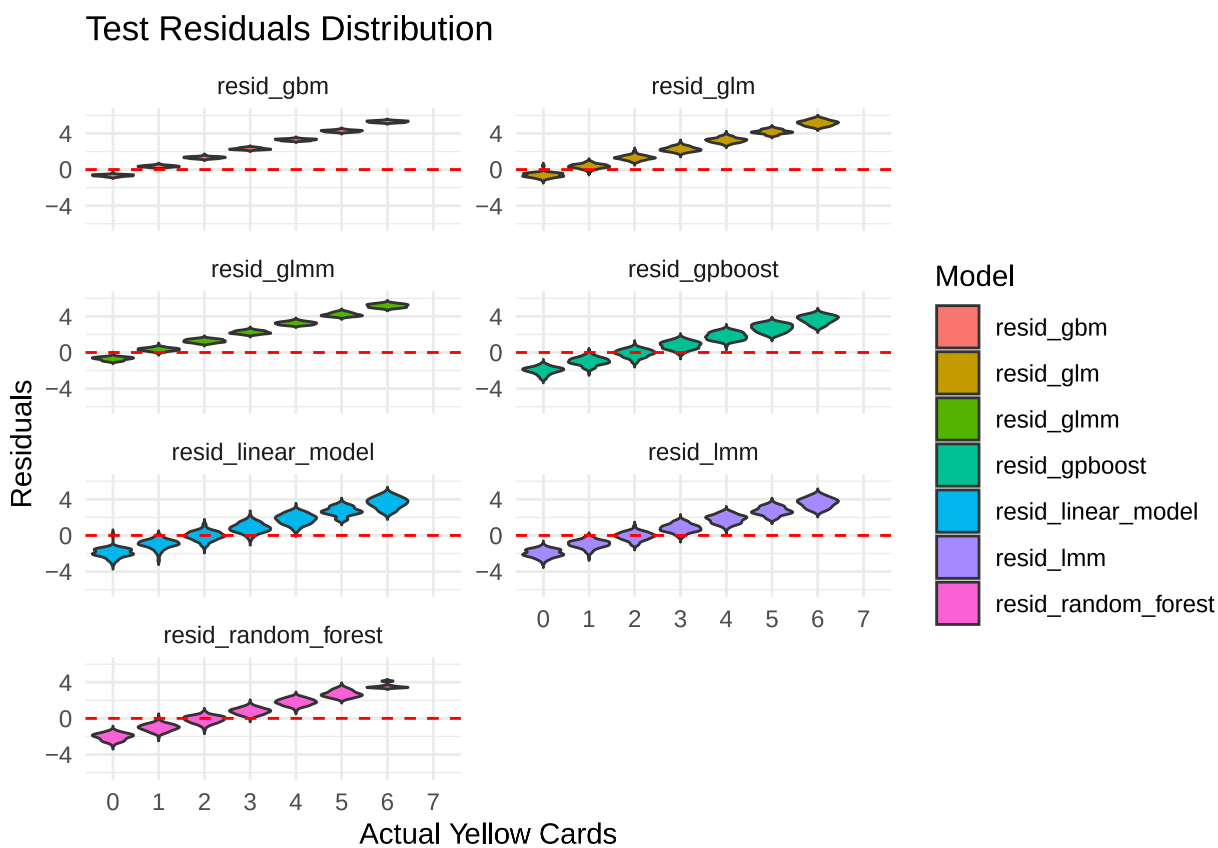


Figure 35: Residual Analysis across models (Test set)

11 Limitations

If goals are considered hard to predict, cards are arguably much more random. Goals are a function of ability and are positive actions that help achieve the game objective, whereas cards are generally undesirable. Bust-ups can happen in a game, where an anomalously high number of cards are awarded which may be hard to fully capture with the current predictors.

Poisson models also give non-0 distributions to all positive integers, where in this case we clearly have an upper limit to the number of possible YCs in a game.

There are also card accumulation rules that we fail to consider in this research. For example, if a player has already accumulated a number of YCs, he could be suspended for the a certain number of subsequent games (rules vary between leagues). A player who has already collected a few YCs is likely to be more cautious to refrain from hitting the threshold for a suspension. To factor this in, we would require player-specific data on card accumulation.

12 High dimensionality and multicollinearity

One particular area of multiple linear regression (LMs and GLMs) that we have not sufficiently looked into is multicollinearity, which can affect linear regression fundamentally, from the standard sums of squares analysis to estimated coefficients and to predictions (Lexi V. Perez, 2017).

Previously, we were unable to use standard multicollinearity checks,such as the variable inflation factor, due to the presence of high cardinality categorical variables. As in our case, when trying to use the `vif` function from `car` package, it returns an error specifying 'aliased coefficients in the model' which is a common result of including highly cardinal variables. Nevertheless, we went ahead with the models also in part to preserve interpretability, in the form of coefficients of predictors from summary output tables, which would otherwise be lost with dimensionality reduction techniques such as PCA.

Multicollinearity has particularly detrimental effects, namely:

- large standard errors for coefficients, which makes them highly unstable and unreliable (Alin, A. 2010)
- 'partial regression coefficient' interpretation, refers to the effect of one unit increase in the predictor on the outcome variable, when all other variables are held constant, becomes not applicable due to collinearity between the problematic predictor and others (Alin, A. 2010)
- usual inferences, such as hypothesis testing, become unreliable (Alin, A. 2010)
- modelling packages (such as with lme4 mixed effect models) become inaccurate due to singular and ill-conditioned information matrices and having to invert them (Aguilera et al. 2006)

Even after removing the 3 categorical variables (referee, team, opponent), we are left with 19 predictors of which 14 are numeric. Thus we have high dimensionality and possible multicollinearity, which we can investigate for numeric variables.

12.1 Examining the correlation coefficients

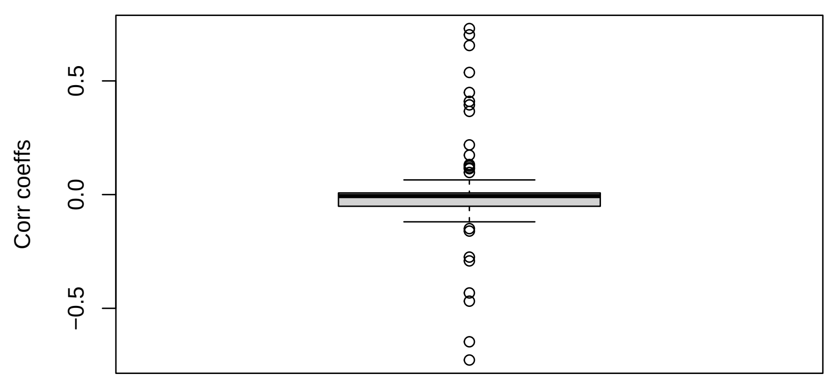


Figure 36: Distribution of correlation coefficients

From Fig 36, we can the spread of correlation coefficients. Most of them are close to 0, which is desirable, however we have outliers close to the extremes of 1 and -1. Of the 91 possible pairs of predictors, 6 of them had a correlation coefficient of above 0.5, and 10 above 0.4.

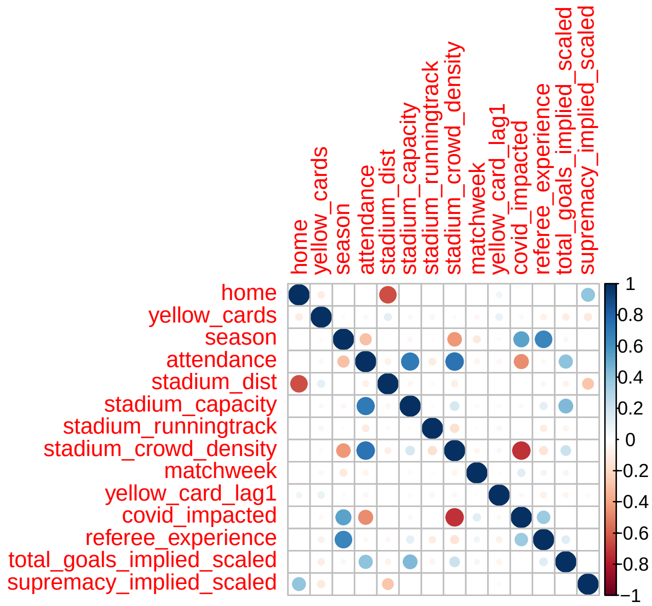


Figure 37: Distribution of correlation coefficients

The correlation table in Fig 37 displays the magnitude and direction of correlation of all pairs of predictors, with the darkest colours having the highest magnitutde, and we can already see some conspicuous dots. Some of the (obvious) pairs include: stadium distance and `home` (only relevant for away team), stadium capacity is positively correlated with attendance, etc.

12.2 Principal Component Analysis (PCA) and Principal Component Regression (PCR)

PCA is a dimensionality reduction technique used to reduce a large set of variables that have some degree of collinearity to a smaller and less correlated subset, which are called principle components (PCs). Ideally, the first few PCs are enough to capture most of the information of the original data, as the PCs themselves are a set of linear combinations of predictors (Lexi V. Perez, 2017). We then regress the outcome on the new set of regressors (PCs) to perform PCR, with the assumption that the directions of most variance of the original predictors are associated with the outcome.

The main ideas are (taken from S. Jackson (2023)) :

$$Z_k = \sum_{j=1}^{p} \phi_{jk} X_j$$

for $Z_k, k = 1, ..., q$ are the q PCs, $X_1....X_p$ are the original p predictors, where $q < lep$, and $\phi_{1k}, ...\phi_{pk}$ are constant coefficients to the original p predictors, thus representing a linear combination

Then, we fit the model for PCR :

$$y_i = B_0 + \sum_{k=1}^{q} B_k z_{ik}$$

for $i = 1, ..., n$ (all observations), representing a reduction in number of parameters from p+1 to q+1

The PCs are found via solving for the $\phi_{1k}, ...\phi_{pk}$ coefficients for the *kth* PC, which involves maximising its variance and ,via Lagrange multiplier methods, leads to the solving of eigenvectors and eigenvalues of the sample correlation matrix. More theory on PCA can be read from Lexi V. Perez (2017).

12.3 Running the PCA

We try to reduce the dimensions and resolve the multicollinearity issues with the help of PCA, for the numeric columns. We then try to rebind the output of the PCA with the other variables (such as the categorical variables), and re-run the regression models to see if there have been any improvements in prediction results.

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.7264 1.4130 1.3648 1.04407 1.03424 0.98368 0.88436
## Proportion of Variance 0.2293 0.1536 0.1433 0.08385 0.08228 0.07443 0.06016
## Cumulative Proportion  0.2293 0.3829 0.5261 0.60999 0.69228 0.76671 0.82687
##                            PC8    PC9    PC10    PC11    PC12    PC13
## Standard deviation     0.86596 0.7783 0.57840 0.50877 0.48705 0.25381
## Proportion of Variance 0.05768 0.0466 0.02573 0.01991 0.01825 0.00496
## Cumulative Proportion  0.88455 0.9312 0.95689 0.97680 0.99504 1.00000
```

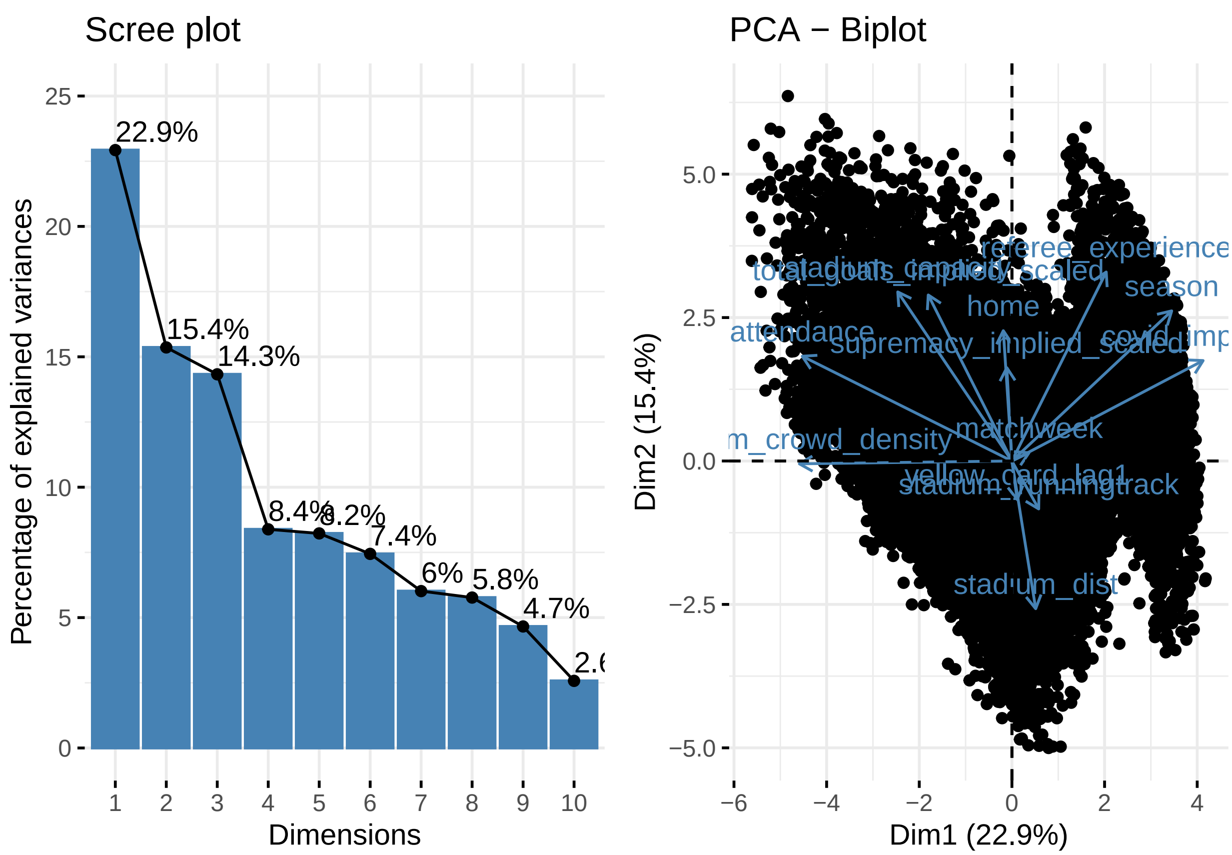


Figure 38: Principle Components

We choose the first 6 principle components, which captures about 76% total variance, as displayed in the Scree Plot in Fig 38. This represents a reduction of 14 numeric columns to 6 principle components, which is a 0.6 proportion reduction in number of numeric predictors. The Biplot in Fig 38 shows the association between the first two PCs and the predictors, and we can see that the first PC correlates positively with `season`, `covid_impacted`, negatively with stadium crowd density and attendance, while the second PC correlated positively with `home`, `supremacy_implied_scaled` (standardised version of `sup_implied`), while varying negatively with `stadium_dist`. This gives us an intuitive sense of the information that each PC is trying to capture.

12.4 Refitting the linear models

With the reduced numeric space, we refit the models again and check the results of the performance on the test set.

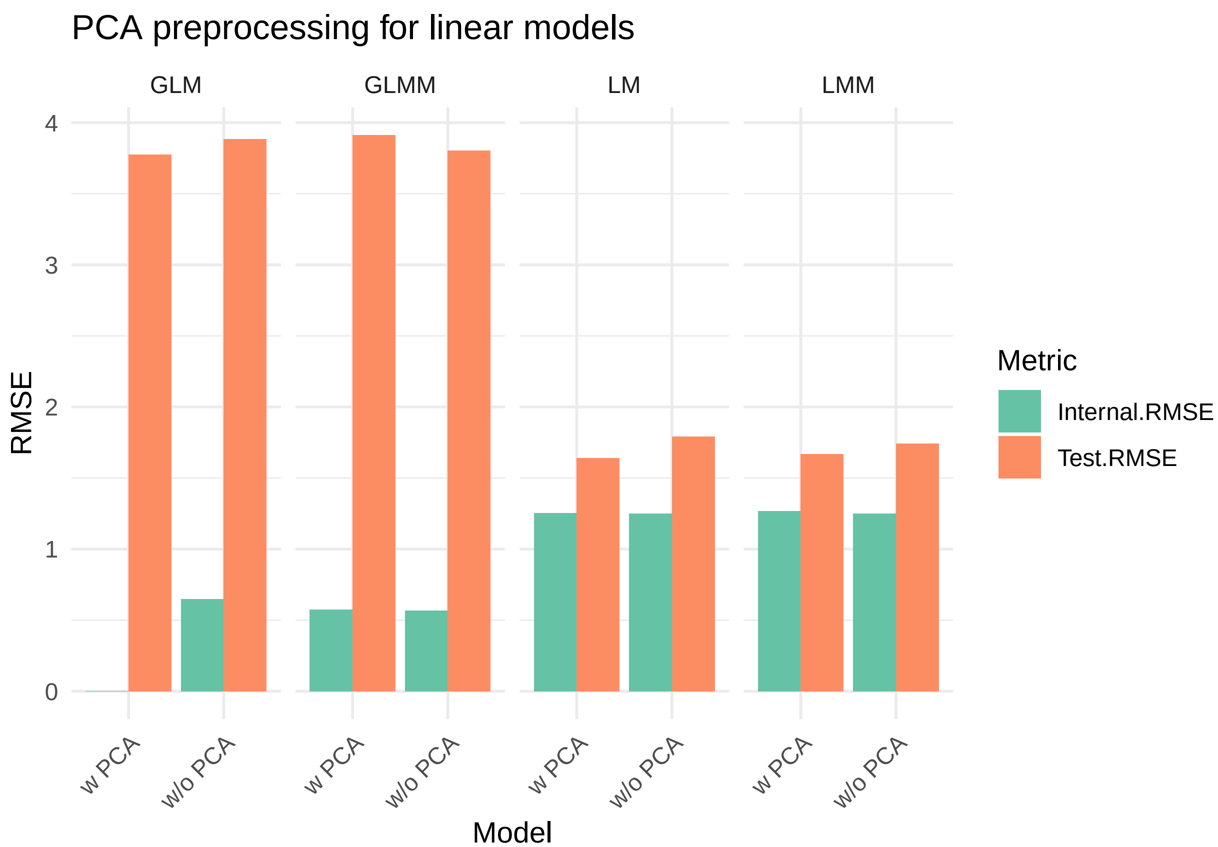


Figure 39: Does PCA-preprocessed data help with training linear models?

There is noticeable change in the test error when the linear regression models were first preprocessed with PCA (PCR models). In Fig 39, we see that across the board, test error went down, although training error was more stable. This could suggest that PCA helps to regularise the parameters which helps in generalising.

When comparing across all models including the tree-based ones, we see that the linear models that benefited from PCA overtook the random forest models as the best performing models on test data, but did not affect the relative positions of the GLMs.

Despite the improvements, there is still clearly room for improvement on the lowest test error of 1.64, which effectively means that the best model still is off by that number of cards per team, per game, on average.

Table 17: Comparison with new PCR models

	Internal		External
model	AIC	Internal.RMSE	Test.RMSE
(PCA) LM	195228.39	1.2552472	1.642137
(PCA) LMM	196643.29	1.2686187	1.667827
RF2	NA	1.6160979	1.731010
gpboost	NA	1.5997987	1.731747
RF1	NA	1.6259491	1.733802
LMM	192976.54	1.2510770	1.741221
LM	192474.47	1.2491650	1.793032
(PCA) GLM	64400.45	0.0010889	3.775334
GLMM	61487.81	0.5686205	3.803085
GLM	62267.36	0.6496406	3.884857
(PCA) GLMM	64134.88	0.5767841	3.912801
gbm2	NA	1.0301570	3.962545
gbm1	NA	0.8561255	4.242806

This might reaffirm the possibility that the main issue is not the data that we have, but rather the data that we do not. In order to effectively predict YCs, it seems that much more data is still required.

13 Conclusion

Predicting YCs is still a very difficult task, with so much unexplained variation due to variables that we do not have, not least from the dataset that we have been provided. It is very likely that YCs are determined on the pitch itself, and any game would find itself with hundreds of game events that would go into determining the cards handed out by the referee. Hence despite the dataset we are presented with, predicting YCs may require more variables and possibly more granular data.

From our models we see that simplicity still wins. Linear models, with the benefit of interpretability, narrowly lost out to the random forest models in test error while having significantly less train error. Generalised linear models that account for the distribution of the outcome allowed for a much better fit, with both AIC and internal error, but suffered with around double of the test error. Standard gradient boosted models were more ill-fitted than the already poor GLMs, but when mixed effects were taken into account in Sigrist's model, the performance improved drastically to that of the random forests.

14 Appendix

14.1 Negative Binomial models

14.1.1 England only

We repeat the Poisson GLM analysis in section 'Model 2', but with Negative Binomial distributions instead. Unfortunately, NB models are not supported in `geepack`, hence was omitted.

Table 18: NB GLMs comparisons

	Internal Metrics		External Metric
Model	AIC	BIC	Test_RMSE
negbin	1.466155e+04	1.633935e+04	1.289180
glmerMod	-1.170992e+10	-1.170992e+10	1.259121

Table 19: Linear Models for all Countries

Linear Models		Internal Metrics		External Metric
Country	Model	AIC	RMSE	test_RMSE
Spain	negbin	1.481745e+04	1.387275	2.015109
Spain	glmerMod	-2.284063e+10	1.392595	2.018726
Spain	lm	4.622296e+04	1.385876	1.471191
Germany	negbin	1.009677e+04	1.215524	1.684853
Germany	glmerMod	-1.202495e+10	1.218381	1.673481
Germany	lm	3.192552e+04	1.213711	1.161365
Italy	negbin	1.145307e+04	1.223280	2.132104
Italy	glmerMod	-1.697304e+10	1.232250	2.111289
Italy	lm	2.877624e+04	1.221938	1.418234
France	negbin	1.138591e+04	1.142794	1.530143
France	glmerMod	-1.672390e+10	1.147188	1.519373
France	lm	3.694416e+04	1.142832	1.034741

14.2 England dataset models: full LM and GLM output

Table 20: Full dataset: Linear Model

term	estimate	std.error	statistic	p.value
(Intercept)	-3.808	88.687	-0.043	0.966
teamBlackburn	-0.124	0.116	-1.065	0.287
teamCardiff	-0.084	0.112	-0.743	0.457
teamBirmingham	-0.063	0.118	-0.537	0.591
teamBlackpool	-0.227	0.188	-1.205	0.228
teamBolton	-0.077	0.158	-0.489	0.625
teamBournemouth	-0.069	0.111	-0.620	0.535
teamBrentford	-0.332	0.113	-2.943	0.003
teamBrighton and Hove Albion	-0.230	0.108	-2.135	0.033
teamCharlton	0.193	0.161	1.201	0.230
teamDerby	-0.033	0.114	-0.286	0.775
teamFulham	0.022	0.109	0.206	0.837
teamHuddersfield	-0.187	0.115	-1.624	0.104
teamIpswich	-0.208	0.146	-1.426	0.154
teamLeeds	0.116	0.105	1.100	0.271
teamMiddlesbrough	-0.092	0.112	-0.823	0.411
teamMillwall	-0.225	0.120	-1.887	0.059
teamNottingham Forest	0.005	0.115	0.044	0.965
teamReading	-0.310	0.117	-2.652	0.008
teamRotherham	-0.118	0.135	-0.876	0.381

teamSheffield Wednesday	-0.128	0.117	-1.093	0.274
teamWatford	0.149	0.109	1.363	0.173
teamWigan	0.038	0.141	0.273	0.785
teamNorwich	-0.104	0.109	-0.953	0.341
teamWolverhampton	-0.190	0.107	-1.779	0.075
teamAston Villa	-0.015	0.105	-0.146	0.884
teamCrystal Palace	-0.147	0.109	-1.349	0.177
teamEverton	-0.090	0.105	-0.859	0.390
teamHull	-0.278	0.124	-2.250	0.024
teamLeicester	-0.301	0.105	-2.852	0.004
teamMan United	0.244	0.103	2.361	0.018
teamQueens Park Rangers	0.029	0.117	0.246	0.806
teamStoke	-0.129	0.112	-1.148	0.251
teamSunderland	0.153	0.166	0.919	0.358
teamSwansea	-0.360	0.114	-3.167	0.002
teamTottenham	0.022	0.102	0.211	0.833
teamWest Bromwich	-0.091	0.109	-0.836	0.403
teamWest Ham	-0.263	0.104	-2.533	0.011
teamLiverpool	-0.400	0.104	-3.857	0.000
teamMan City	-0.040	0.109	-0.364	0.716
teamNewcastle	-0.129	0.108	-1.199	0.231
teamSouthampton	-0.186	0.105	-1.759	0.079
teamBurnley	-0.196	0.111	-1.776	0.076
teamChelsea	-0.152	0.104	-1.461	0.144
teamBristol City	-0.355	0.117	-3.036	0.002
teamMilton Keynes	0.157	0.328	0.478	0.633
teamPreston	-0.217	0.116	-1.865	0.062
teamBarnsley	-0.066	0.122	-0.537	0.591
teamBurton	-0.398	0.196	-2.027	0.043
teamSheffield United	-0.072	0.111	-0.646	0.518
teamLuton	-0.343	0.132	-2.600	0.009
teamCoventry	-0.392	0.139	-2.827	0.005
teamWycombe	-0.210	0.164	-1.281	0.200
teamPeterborough	-0.469	0.222	-2.114	0.035
opponentArsenal	-0.054	0.106	-0.514	0.607
opponentBarnsley	-0.634	0.108	-5.876	0.000
opponentBirmingham	-0.618	0.102	-6.086	0.000
opponentBlackburn	-0.383	0.103	-3.723	0.000
opponentBlackpool	-0.028	0.189	-0.148	0.882
opponentBolton	-0.668	0.145	-4.611	0.000
opponentBournemouth	-0.277	0.101	-2.744	0.006
opponentBrentford	-0.290	0.102	-2.860	0.004
opponentBrighton and Hove Albion	-0.463	0.101	-4.598	0.000
opponentBristol City	-0.462	0.102	-4.528	0.000
opponentBurnley	-0.795	0.103	-7.708	0.000
opponentBurton	-0.907	0.185	-4.910	0.000
opponentCardiff	-0.709	0.101	-7.042	0.000
opponentCharlton	-0.216	0.149	-1.445	0.148
opponentChelsea	-0.251	0.108	-2.330	0.020
opponentCoventry	0.096	0.130	0.744	0.457
opponentCrystal Palace	-0.245	0.103	-2.388	0.017

opponentDerby	-0.567	0.100	-5.647	0.000
opponentEverton	-0.291	0.102	-2.863	0.004
opponentFulham	-0.464	0.099	-4.671	0.000
opponentHuddersfield	-0.305	0.101	-3.008	0.003
opponentHull	-0.600	0.110	-5.463	0.000
opponentIpswich	-0.660	0.134	-4.938	0.000
opponentLeeds	-0.211	0.098	-2.146	0.032
opponentLeicester	-0.430	0.103	-4.187	0.000
opponentLiverpool	-0.634	0.113	-5.625	0.000
opponentLuton	-0.685	0.117	-5.868	0.000
opponentMan City	-0.527	0.119	-4.411	0.000
opponentMan United	-0.256	0.109	-2.347	0.019
opponentMiddlesbrough	-0.372	0.101	-3.693	0.000
opponentMillwall	-0.603	0.105	-5.734	0.000
opponentMilton Keynes	0.089	0.322	0.275	0.783
opponentNewcastle	-0.338	0.102	-3.306	0.001
opponentNorwich	-0.263	0.099	-2.645	0.008
opponentNottingham Forest	-0.350	0.100	-3.487	0.000
opponentPeterborough	-0.667	0.197	-3.385	0.001
opponentPreston	-0.500	0.102	-4.884	0.000
opponentQueens Park Rangers	-0.594	0.103	-5.789	0.000
opponentReading	-0.181	0.102	-1.773	0.076
opponentRotherham	-0.993	0.120	-8.285	0.000
opponentSheffield United	-0.680	0.102	-6.661	0.000
opponentSheffield Wednesday	-0.703	0.104	-6.771	0.000
opponentSouthampton	-0.429	0.102	-4.210	0.000
opponentStoke	-0.606	0.100	-6.050	0.000
opponentSunderland	-0.402	0.160	-2.515	0.012
opponentSwansea	-0.361	0.101	-3.577	0.000
opponentTottenham	-0.033	0.107	-0.306	0.760
opponentWatford	-0.509	0.101	-5.053	0.000
opponentWest Bromwich	-0.565	0.099	-5.703	0.000
opponentWest Ham	-0.467	0.102	-4.594	0.000
opponentWigan	-0.201	0.128	-1.569	0.117
opponentWolverhampton	-0.164	0.100	-1.645	0.100
opponentWycombe	-0.542	0.151	-3.592	0.000
home	-0.143	0.146	-0.974	0.330
refereeAndrew Madley	-0.109	0.095	-1.151	0.250
refereeAndy D'Urso	-0.283	0.462	-0.613	0.540
refereeAndy Davies	-0.045	0.117	-0.381	0.703
refereeAndy Haines	-0.245	0.531	-0.461	0.645
refereeAndy Woolmer	-0.120	0.152	-0.790	0.429
refereeAnthony Taylor	0.224	0.090	2.480	0.013
refereeAntony Coggins	0.955	0.746	1.279	0.201
refereeBrendan Malone	-0.304	0.455	-0.669	0.503
refereeCarl Berry	0.141	0.806	0.176	0.861
refereeCarl Boyeson	0.175	0.400	0.438	0.661
refereeCharles Breakspear	0.058	0.414	0.139	0.889
refereeChris Foy	0.164	0.387	0.423	0.673
refereeChris Kavanagh	0.124	0.101	1.218	0.223
refereeChris Sarginson	0.481	0.735	0.655	0.513

refereeCraig Hicks	-0.695	0.638	-1.090	0.276
refereeCraig Pawson	0.234	0.087	2.686	0.007
refereeDarren Bond	0.072	0.104	0.692	0.489
refereeDarren Deadman	0.045	0.345	0.130	0.896
refereeDarren Drysdale	-0.333	0.365	-0.913	0.361
refereeDarren England	0.188	0.163	1.153	0.249
refereeDarren Handley	0.133	0.671	0.199	0.842
refereeDarren Sheldrake	0.140	1.845	0.076	0.940
refereeDavid Coote	0.398	0.109	3.658	0.000
refereeDavid Webb	-0.125	0.164	-0.764	0.445
refereeDean Whitestone	-0.260	0.201	-1.291	0.197
refereeEddie Ilderton	-0.084	0.357	-0.236	0.813
refereeFred Graham	0.125	0.395	0.318	0.751
refereeGary Sutton	-0.155	0.666	-0.233	0.816
refereeGavin Ward	0.152	0.172	0.888	0.375
refereeGeoff Eltrigham	0.104	0.108	0.964	0.335
refereeGraham Horwood	0.437	0.776	0.563	0.573
refereeGraham Salisbury	0.025	0.416	0.059	0.953
refereeGraham Scott	-0.009	0.109	-0.078	0.938
refereeIain Williamson	0.463	0.400	1.158	0.247
refereeJames Adcock	-0.043	0.220	-0.197	0.844
refereeJames Linington	0.051	0.103	0.498	0.618
refereeJARRED GILLETT	0.119	0.238	0.499	0.618
refereeJeremy Simpson	0.131	0.105	1.246	0.213
refereeJohn Brooks	0.373	0.192	1.945	0.052
refereeJohn Busby	0.397	0.341	1.164	0.244
refereeJonathan Moss	0.109	0.088	1.238	0.216
refereeJosh Smith	0.029	0.315	0.092	0.927
refereeKeith Hill	0.126	0.361	0.350	0.726
refereeKeith Stroud	0.112	0.095	1.180	0.238
refereeKevin Friend	0.106	0.087	1.224	0.221
refereeKevin Johnson	-0.400	0.673	-0.594	0.553
refereeKevin Wright	0.030	0.383	0.079	0.937
refereeLee Collins	-0.543	0.947	-0.574	0.566
refereeLee Mason	-0.007	0.098	-0.069	0.945
refereeLee Probert	-0.190	0.167	-1.141	0.254
refereeLeigh DOUGHTY	0.554	0.299	1.850	0.064
refereeMark Brown	0.162	0.348	0.464	0.643
refereeMark Clattenburg	0.048	0.174	0.278	0.781
refereeMark Heywood	0.199	0.296	0.672	0.502
refereeMartin Atkinson	-0.045	0.089	-0.507	0.612
refereeMathew Buonassisi	-0.941	1.467	-0.642	0.521
refereeMatt Donohue	0.133	0.269	0.496	0.620
refereeMichael Bull	0.265	0.518	0.511	0.609
refereeMichael Oliver	0.057	0.088	0.650	0.516
refereeMichael Salisbury	-0.108	0.301	-0.360	0.719
refereeMike Dean	0.458	0.092	4.961	0.000
refereeMike Jones	0.318	0.145	2.189	0.029
refereeMike Russell	-0.672	0.537	-1.251	0.211
refereeMissing-England-2	0.568	0.287	1.978	0.048
refereeNeil Swarbrick	0.073	0.147	0.494	0.621

refereeNigel Miller	-0.282	0.352	-0.803	0.422
refereeOliver Langford	0.244	0.105	2.329	0.020
refereePatrick Miller	-0.451	1.032	-0.438	0.662
refereePaul Tierney	0.272	0.096	2.819	0.005
refereePeter Bankes	0.370	0.105	3.511	0.000
refereePhil Dowd	0.683	0.390	1.749	0.080
refereePhil Gibbs	0.028	0.566	0.049	0.961
refereeRichard Clark	0.174	0.657	0.265	0.791
refereeRobert Atkin	-0.033	1.571	-0.021	0.983
refereeRobert Jones	-0.023	0.170	-0.135	0.893
refereeRobert Lewis	-0.368	0.522	-0.706	0.480
refereeRobert Madley	0.149	0.141	1.053	0.292
refereeRoger East	0.243	0.132	1.836	0.066
refereeRoss Joyce	-0.188	0.635	-0.296	0.767
refereeScott Duncan	0.009	0.130	0.067	0.946
refereeScott Mathieson	-1.355	1.836	-0.738	0.460
refereeSebastian Stockbridge	-0.082	1.015	-0.081	0.935
refereeSimon Hooper	0.009	0.096	0.099	0.921
refereeStephen Martin	-0.286	0.108	-2.634	0.008
refereeStuart Attwell	0.350	0.105	3.327	0.001
refereeT J Robinson	0.123	0.102	1.207	0.227
refereeThomas Bramall	-0.435	0.358	-1.215	0.224
refereeTony Harrington	-0.010	0.110	-0.086	0.931
refereeTrevor Kettle	-0.459	0.511	-0.898	0.369
season	0.003	0.044	0.068	0.946
competition_level2	0.203	0.060	3.416	0.001
attendance	-0.079	0.025	-3.145	0.002
stadium_dist	0.000	0.000	-1.751	0.080
stadium_capacity	-0.020	0.021	-0.940	0.347
stadium_crowd_density	0.502	0.164	3.061	0.002
matchweek	0.001	0.006	0.243	0.808
yellow_card_lag1	-0.007	0.008	-0.817	0.414
covid_impacted	-0.141	0.133	-1.057	0.291
referee_experience	0.000	0.001	0.065	0.948
total_goals_implied_scaled	-0.044	0.019	-2.387	0.017
supremacy_implied_scaled	-0.112	0.024	-4.660	0.000
factor(season)2016	-0.061	0.146	-0.415	0.678
factor(season)2017	0.251	0.123	2.034	0.042
factor(season)2018	-0.267	0.104	-2.571	0.010
factor(season)2019	-0.223	0.087	-2.544	0.011
factor(season)2020	-0.119	0.078	-1.521	0.128
factor(season)2021	-0.196	0.109	-1.792	0.073
I(matchweek^2)	0.000	0.000	-2.657	0.008
matchweek:factor(season)2016	-0.005	0.006	-0.758	0.449
matchweek:factor(season)2017	-0.008	0.006	-1.447	0.148
matchweek:factor(season)2018	0.010	0.005	1.868	0.062
matchweek:factor(season)2019	0.007	0.005	1.402	0.161
matchweek:factor(season)2020	0.004	0.005	0.722	0.470
matchweek:factor(season)2021	0.003	0.005	0.687	0.492
matchweek:factor(season)2022	0.005	0.007	0.680	0.497
home:attendance	0.098	0.035	2.805	0.005

home:covid_impacted	0.177	0.132	1.340	0.180
home:stadium_capacity	-0.013	0.026	-0.486	0.627
home:stadium_crowd_density	-0.426	0.209	-2.036	0.042

Table 21: Full dataset: Generalised Linear Model (Poisson, log-link)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.890	0.246	3.617	0.000
home	-0.101	0.144	-0.699	0.485
competition_level2	0.110	0.065	1.703	0.089
covid_impacted	-0.149	0.143	-1.046	0.296
total_goals_implied_scaled	-0.037	0.021	-1.783	0.075
supremacy_implied_scaled	-0.072	0.027	-2.665	0.008
opponentArsenal	-0.015	0.110	-0.136	0.891
opponentBarnsley	-0.371	0.120	-3.097	0.002
opponentBirmingham	-0.364	0.111	-3.285	0.001
opponentBlackburn	-0.211	0.109	-1.926	0.054
opponentBlackpool	-0.049	0.184	-0.267	0.790
opponentBolton	-0.406	0.169	-2.402	0.016
opponentBournemouth	-0.156	0.105	-1.479	0.139
opponentBrentford	-0.170	0.105	-1.616	0.106
opponentBrighton and Hove Albion	-0.268	0.109	-2.459	0.014
opponentBristol City	-0.261	0.108	-2.414	0.016
opponentBurnley	-0.518	0.122	-4.255	0.000
opponentBurton	-0.553	0.221	-2.509	0.012
opponentCardiff	-0.428	0.113	-3.782	0.000
opponentCharlton	-0.117	0.157	-0.741	0.459
opponentChelsea	-0.139	0.113	-1.235	0.217
opponentCoventry	0.047	0.130	0.364	0.716
opponentCrystal Palace	-0.133	0.109	-1.220	0.222
opponentDerby	-0.326	0.108	-3.011	0.003
opponentEverton	-0.155	0.108	-1.437	0.151
opponentFulham	-0.268	0.107	-2.511	0.012
opponentHuddersfield	-0.172	0.106	-1.626	0.104
opponentHull	-0.346	0.120	-2.882	0.004
opponentIpswich	-0.391	0.151	-2.582	0.010
opponentLeeds	-0.118	0.101	-1.160	0.246
opponentLeicester	-0.240	0.110	-2.182	0.029
opponentLiverpool	-0.361	0.125	-2.885	0.004
opponentLuton	-0.418	0.132	-3.176	0.001
opponentMan City	-0.285	0.130	-2.193	0.028
opponentMan United	-0.128	0.116	-1.105	0.269
opponentMiddlesbrough	-0.209	0.106	-1.977	0.048
opponentMillwall	-0.355	0.114	-3.107	0.002
opponentMilton Keynes	0.050	0.328	0.154	0.878
opponentNewcastle	-0.182	0.111	-1.645	0.100
opponentNorwich	-0.140	0.103	-1.356	0.175
opponentNottingham Forest	-0.197	0.105	-1.880	0.060
opponentPeterborough	-0.371	0.221	-1.682	0.093

opponentPreston	-0.290	0.110	-2.634	0.008
opponentQueens Park Rangers	-0.347	0.112	-3.099	0.002
opponentReading	-0.104	0.105	-0.991	0.322
opponentRotherham	-0.687	0.150	-4.568	0.000
opponentSheffield United	-0.413	0.114	-3.616	0.000
opponentSheffield Wednesday	-0.426	0.116	-3.656	0.000
opponentSouthampton	-0.244	0.111	-2.200	0.028
opponentStoke	-0.355	0.108	-3.280	0.001
opponentSunderland	-0.217	0.178	-1.217	0.223
opponentSwansea	-0.203	0.106	-1.915	0.055
opponentTottenham	-0.010	0.110	-0.093	0.926
opponentWatford	-0.290	0.109	-2.651	0.008
opponentWest Bromwich	-0.332	0.108	-3.076	0.002
opponentWest Ham	-0.264	0.113	-2.336	0.019
opponentWigan	-0.120	0.133	-0.904	0.366
opponentWolverhampton	-0.106	0.102	-1.031	0.302
opponentWycombe	-0.328	0.181	-1.811	0.070
refereeAndrew Madley	-0.063	0.110	-0.576	0.564
refereeAndy D'Urso	-0.166	0.552	-0.301	0.764
refereeAndy Davies	-0.018	0.117	-0.155	0.876
refereeAndy Haines	-0.143	0.646	-0.221	0.825
refereeAndy Woolmer	-0.070	0.121	-0.579	0.563
refereeAnthony Taylor	0.146	0.096	1.521	0.128
refereeAntony Coggins	0.484	0.650	0.744	0.457
refereeBrendan Malone	-0.209	0.572	-0.365	0.715
refereeCarl Berry	0.103	0.836	0.123	0.902
refereeCarl Boyeson	0.116	0.426	0.272	0.785
refereeCharles Breakspear	0.056	0.457	0.123	0.902
refereeChris Foy	0.104	0.417	0.249	0.803
refereeChris Kavanagh	0.082	0.103	0.796	0.426
refereeChris Sarginson	0.280	0.729	0.384	0.701
refereeCraig Hicks	-0.901	1.148	-0.785	0.433
refereeCraig Pawson	0.149	0.099	1.506	0.132
refereeDarren Bond	0.055	0.112	0.490	0.624
refereeDarren Deadman	0.046	0.386	0.118	0.906
refereeDarren Drysdale	-0.225	0.469	-0.480	0.631
refereeDarren England	0.119	0.114	1.042	0.298
refereeDarren Handley	0.094	0.744	0.126	0.900
refereeDarren Sheldrake	0.058	1.874	0.031	0.975
refereeDavid Coote	0.235	0.101	2.321	0.020
refereeDavid Webb	-0.073	0.118	-0.622	0.534
refereeDean Whitestone	-0.183	0.129	-1.418	0.156
refereeEddie Ilderton	-0.038	0.420	-0.092	0.927
refereeFred Graham	0.092	0.439	0.209	0.835
refereeGary Sutton	-0.075	0.781	-0.097	0.923
refereeGavin Ward	0.102	0.115	0.888	0.375
refereeGeoff Eltrigham	0.075	0.109	0.684	0.494
refereeGraham Horwood	0.275	0.777	0.354	0.723
refereeGraham Salisbury	0.030	0.454	0.066	0.948
refereeGraham Scott	-0.003	0.109	-0.027	0.979
refereeIain Williamson	0.277	0.411	0.674	0.501

refereeJames Adcock	-0.019	0.245	-0.077	0.939
refereeJames Linington	0.039	0.111	0.351	0.726
refereeJARRED GILLETT	0.078	0.127	0.614	0.539
refereeJeremy Simpson	0.093	0.109	0.855	0.393
refereeJohn Brooks	0.219	0.114	1.918	0.055
refereeJohn Busby	0.209	0.187	1.112	0.266
refereeJonathan Moss	0.076	0.100	0.764	0.445
refereeJosh Smith	0.011	0.173	0.061	0.951
refereeKeith Hill	0.088	0.384	0.230	0.818
refereeKeith Stroud	0.082	0.107	0.765	0.444
refereeKevin Friend	0.070	0.102	0.692	0.489
refereeKevin Johnson	-0.275	0.895	-0.307	0.759
refereeKevin Wright	0.040	0.419	0.095	0.924
refereeLee Collins	-0.559	1.543	-0.362	0.717
refereeLee Mason	-0.004	0.115	-0.035	0.972
refereeLee Probert	-0.124	0.168	-0.737	0.461
refereeLeigh DOUGHTY	0.310	0.151	2.044	0.041
refereeMark Brown	0.113	0.378	0.300	0.764
refereeMark Clattenburg	0.041	0.197	0.209	0.835
refereeMark Heywood	0.132	0.320	0.414	0.679
refereeMartin Atkinson	-0.026	0.102	-0.257	0.797
refereeMathew Buonassisi	-0.606	2.099	-0.289	0.773
refereeMatt Donohue	0.090	0.135	0.671	0.502
refereeMichael Bull	0.169	0.551	0.307	0.759
refereeMichael Oliver	0.046	0.098	0.473	0.636
refereeMichael Salisbury	-0.078	0.171	-0.454	0.650
refereeMike Dean	0.266	0.095	2.797	0.005
refereeMike Jones	0.197	0.156	1.259	0.208
refereeMike Russell	-0.427	0.705	-0.606	0.544
refereeMissing-England-2	0.293	0.190	1.538	0.124
refereeNeil Swarbrick	0.055	0.167	0.329	0.742
refereeNigel Miller	-0.158	0.419	-0.378	0.706
refereeOliver Langford	0.156	0.108	1.442	0.149
refereePatrick Miller	-0.281	1.271	-0.221	0.825
refereePaul Tierney	0.173	0.101	1.713	0.087
refereePeter Bankes	0.219	0.104	2.109	0.035
refereePhil Dowd	0.369	0.374	0.987	0.324
refereePhil Gibbs	0.040	0.651	0.062	0.951
refereeRichard Clark	0.141	0.713	0.198	0.843
refereeRobert Atkin	-0.051	1.595	-0.032	0.975
refereeRobert Jones	-0.016	0.116	-0.135	0.893
refereeRobert Lewis	-0.262	0.681	-0.385	0.700
refereeRobert Madley	0.099	0.158	0.626	0.531
refereeRoger East	0.153	0.143	1.070	0.285
refereeRoss Joyce	-0.072	0.711	-0.101	0.920
refereeScott Duncan	0.012	0.141	0.089	0.929
refereeScott Mathieson	-1.280	3.719	-0.344	0.731
refereeSebastian Stockbridge	-0.030	1.128	-0.026	0.979
refereeSimon Hooper	0.015	0.110	0.140	0.889
refereeStephen Martin	-0.189	0.118	-1.604	0.109
refereeStuart Attwell	0.213	0.101	2.103	0.035

refereeT J Robinson	0.084	0.106	0.791	0.429
refereeThomas Bramall	-0.313	0.277	-1.131	0.258
refereeTony Harrington	0.005	0.112	0.047	0.962
refereeTrevor Kettle	-0.300	0.665	-0.452	0.651
teamBlackburn	-0.091	0.130	-0.698	0.485
teamCardiff	-0.067	0.125	-0.536	0.592
teamBirmingham	-0.057	0.130	-0.438	0.661
teamBlackpool	-0.141	0.202	-0.698	0.485
teamBolton	-0.071	0.172	-0.412	0.681
teamBournemouth	-0.046	0.122	-0.379	0.704
teamBrentford	-0.222	0.129	-1.723	0.085
teamBrighton and Hove Albion	-0.151	0.121	-1.242	0.214
teamCharlton	0.074	0.169	0.437	0.662
teamDerby	-0.033	0.126	-0.266	0.790
teamFulham	0.005	0.120	0.045	0.964
teamHuddersfield	-0.126	0.128	-0.983	0.326
teamIpswich	-0.148	0.162	-0.913	0.361
teamLeeds	0.053	0.115	0.464	0.643
teamMiddlesbrough	-0.073	0.124	-0.591	0.555
teamMillwall	-0.151	0.133	-1.130	0.258
teamNottingham Forest	-0.016	0.126	-0.123	0.902
teamReading	-0.207	0.132	-1.569	0.117
teamRotherham	-0.080	0.149	-0.537	0.591
teamSheffield Wednesday	-0.094	0.131	-0.718	0.473
teamWatford	0.066	0.118	0.563	0.574
teamWigan	-0.007	0.152	-0.045	0.964
teamNorwich	-0.070	0.122	-0.575	0.566
teamWolverhampton	-0.129	0.120	-1.081	0.280
teamAston Villa	-0.022	0.116	-0.189	0.850
teamCrystal Palace	-0.099	0.122	-0.811	0.417
teamEverton	-0.064	0.117	-0.546	0.585
teamHull	-0.180	0.138	-1.303	0.193
teamLeicester	-0.197	0.121	-1.632	0.103
teamMan United	0.134	0.112	1.190	0.234
teamQueens Park Rangers	0.004	0.128	0.029	0.977
teamStoke	-0.094	0.124	-0.757	0.449
teamSunderland	0.056	0.173	0.325	0.745
teamSwansea	-0.237	0.130	-1.826	0.068
teamTottenham	0.004	0.114	0.038	0.970
teamWest Bromwich	-0.063	0.120	-0.527	0.598
teamWest Ham	-0.173	0.119	-1.455	0.146
teamLiverpool	-0.290	0.127	-2.279	0.023
teamMan City	-0.025	0.127	-0.199	0.842
teamNewcastle	-0.089	0.120	-0.740	0.460
teamSouthampton	-0.120	0.119	-1.010	0.313
teamBurnley	-0.128	0.123	-1.044	0.297
teamChelsea	-0.098	0.120	-0.818	0.413
teamBristol City	-0.230	0.133	-1.736	0.083
teamMilton Keynes	0.074	0.348	0.214	0.831
teamPreston	-0.147	0.130	-1.129	0.259
teamBarnsley	-0.057	0.135	-0.422	0.673

teamBurton	-0.254	0.220	-1.155	0.248
teamSheffield United	-0.053	0.122	-0.435	0.663
teamLuton	-0.225	0.149	-1.509	0.131
teamCoventry	-0.273	0.164	-1.663	0.096
teamWycombe	-0.124	0.187	-0.662	0.508
teamPeterborough	-0.275	0.242	-1.136	0.256
attendance	-0.043	0.025	-1.735	0.083
stadium_crowd_density	0.255	0.176	1.452	0.147
factor(season)2016	-0.026	0.171	-0.151	0.880
factor(season)2017	0.135	0.157	0.862	0.389
factor(season)2018	-0.141	0.154	-0.917	0.359
factor(season)2019	-0.113	0.148	-0.761	0.447
factor(season)2020	-0.055	0.146	-0.380	0.704
factor(season)2021	-0.044	0.165	-0.266	0.790
factor(season)2022	-0.017	0.147	-0.112	0.911
matchweek	-0.005	0.005	-0.909	0.363
stadium_capacity	-0.016	0.019	-0.826	0.409
stadium_dist	0.000	0.000	-1.014	0.310
home:covid_impacted	0.095	0.145	0.657	0.511
home:attendance	0.050	0.028	1.772	0.076
home:stadium_crowd_density	-0.231	0.223	-1.033	0.301
factor(season)2016:matchweek	-0.003	0.007	-0.496	0.620
factor(season)2017:matchweek	-0.004	0.006	-0.708	0.479
factor(season)2018:matchweek	0.006	0.006	0.974	0.330
factor(season)2019:matchweek	0.004	0.006	0.744	0.457
factor(season)2020:matchweek	0.003	0.006	0.505	0.614
factor(season)2021:matchweek	0.001	0.005	0.253	0.800
factor(season)2022:matchweek	0.006	0.007	0.808	0.419

14.3 Full dataset models

Table 22: Output for Linear model with entire dataset

term	estimate	std.error	statistic	p.value
(Intercept)	-58.5283991	25.5465721	-2.2910471	0.0219646
team1FC Koln	-0.3500791	0.2288754	-1.5295619	0.1261312
teamAalen	-0.9777790	0.5010681	-1.9513892	0.0510160
teamAC Milan	-0.7725627	0.4020807	-1.9214122	0.0546851
teamAJ Auxerre	-1.2076947	0.4041996	-2.9878673	0.0028106
teamAjaccio	-0.9391287	0.4049654	-2.3190347	0.0203969
teamAlaves	-0.9495566	0.4475898	-2.1214886	0.0338853
teamAlbacete	-0.7205783	0.4478504	-1.6089710	0.1076287
teamAlcorcon	-0.9860117	0.4467165	-2.2072426	0.0273014
teamAlessandria	-0.6223441	0.4647997	-1.3389511	0.1805924
teamAlmeria	-0.7135042	0.4458282	-1.6004015	0.1095155
teamAmiens	-1.0922079	0.4047463	-2.6985004	0.0069675
teamAmorebieta	-1.4155136	0.4942174	-2.8641517	0.0041829
teamAngers	-1.2430775	0.4061137	-3.0609099	0.0022078
teamArles	-0.9337934	0.5819329	-1.6046411	0.1085788
teamArminia Bielefeld	-0.7580961	0.2277667	-3.3283893	0.0008741

teamArsenal	-1.5483031	0.6565394	-2.3582790	0.0183635
teamAS Beziers	-0.8557112	0.4467479	-1.9154232	0.0554439
teamAscoli	-0.4102999	0.4026808	-1.0189210	0.3082451
teamAston Villa	-1.5793839	0.6552219	-2.4104565	0.0159360
teamAtalanta	-0.8790264	0.4027695	-2.1824552	0.0290803
teamAthletic Bilbao	-1.1588254	0.4471502	-2.5915800	0.0095563
teamAthletic Bilbao B	-1.0977887	0.5520130	-1.9887007	0.0467394
teamAtletico Madrid	-0.8682594	0.4473834	-1.9407501	0.0522939
teamAugsburg	-0.4412143	0.2295866	-1.9217774	0.0546391
teamAvellino	-0.8165704	0.8711688	-0.9373274	0.3485944
teamBarcelona	-0.9604876	0.4480240	-2.1438307	0.0320510
teamBarcelona B	-1.1084346	0.4841552	-2.2894198	0.0220589
teamBari	-0.8867423	0.6896185	-1.2858447	0.1985029
teamBarnsley	-1.6347358	0.6552328	-2.4948934	0.0126025
teamBastia	-1.0148697	0.4216777	-2.4067425	0.0160989
teamBayern Munich	-0.4950339	0.2347595	-2.1086856	0.0349764
teamBenevento	-0.7329966	0.4017448	-1.8245332	0.0680771
teamBetis	-0.8587050	0.4470382	-1.9208760	0.0547527
teamBirmingham	-1.6824140	0.6538034	-2.5732718	0.0100769
teamBlackburn	-1.6913923	0.6546027	-2.5838456	0.0097732
teamBlackpool	-1.8928815	0.6791831	-2.7869975	0.0053218
teamBochum	-0.5236796	0.2270150	-2.3068061	0.0210695
teamBologna	-0.6399572	0.4018031	-1.5927134	0.1112305
teamBolton	-1.6859476	0.6603201	-2.5532276	0.0106758
teamBordeaux	-0.7921808	0.4062409	-1.9500273	0.0511781
teamBourg Peronnas	-1.1090817	0.4299208	-2.5797347	0.0098903
teamBournemouth	-1.6637271	0.6547950	-2.5408367	0.0110616
teamBraunschweig	-0.4851608	0.2391958	-2.0283001	0.0425346
teamBrentford	-1.9000211	0.6546057	-2.9025428	0.0037030
teamBrescia	-0.6856111	0.4022844	-1.7042946	0.0883319
teamBrest	-1.1351491	0.4056997	-2.7980030	0.0051438
teamBrighton and Hove Albion	-1.8693800	0.6549748	-2.8541249	0.0043172
teamBristol City	-1.9250921	0.6544002	-2.9417658	0.0032649
teamBurgos	-1.6094461	0.4782916	-3.3649893	0.0007660
teamBurnley	-1.8438406	0.6555585	-2.8126257	0.0049157
teamBurton	-1.9888048	0.6732606	-2.9539894	0.0031383
teamCadiz	-1.2910883	0.4473453	-2.8861110	0.0039019
teamCaen	-1.0427792	0.4039580	-2.5814050	0.0098426
teamCagliari	-0.7003452	0.4020302	-1.7420213	0.0815105
teamCalcio Padova	-0.6272489	0.4460690	-1.4061701	0.1596795
teamCardiff	-1.6693169	0.6536518	-2.5538319	0.0106572
teamCarpi 1909	-0.7603179	0.4305021	-1.7661188	0.0773817
teamCartagena	-0.5260881	0.4557431	-1.1543523	0.2483610
teamCastellon	-1.4112393	0.4607934	-3.0626289	0.0021951
teamCelta Vigo	-0.7566456	0.4486377	-1.6865404	0.0916977
teamCesena	-0.7777655	0.5084796	-1.5295906	0.1261241
teamChambly	-0.9579518	0.4131704	-2.3185390	0.0204238
teamCharlton	-1.4061046	0.6611137	-2.1268728	0.0334353
teamChateauroux	-0.9771802	0.4072875	-2.3992396	0.0164326
teamChelsea	-1.7787420	0.6566551	-2.7087918	0.0067550
teamChievo	-0.9196475	0.4001814	-2.2980763	0.0215613

teamCittadella	-0.1296404	0.4022289	-0.3223052	0.7472228
teamClermont	-1.2829324	0.4047978	-3.1693168	0.0015288
teamComo	-0.6064674	0.4493274	-1.3497227	0.1771107
teamCordoba	-0.9045231	0.4557464	-1.9847070	0.0471822
teamCosenza 1914	-0.9861335	0.4026639	-2.4490238	0.0143276
teamCoventry	-2.0771962	0.6589396	-3.1523318	0.0016206
teamCremonese	-0.8920461	0.4021181	-2.2183685	0.0265339
teamCreteil	-0.9285980	0.4816466	-1.9279653	0.0538648
teamCrotone	-0.8837722	0.4010924	-2.2034132	0.0275699
teamCrystal Palace	-1.7803141	0.6557199	-2.7150526	0.0066287
teamDarmstadt	-0.5289910	0.2273914	-2.3263459	0.0200039
teamDeportivo La Coruna	-1.2186375	0.4504545	-2.7053510	0.0068254
teamDerby	-1.6367789	0.6541880	-2.5020008	0.0123524
teamDijon	-0.9196762	0.4050685	-2.2704214	0.0231860
teamDortmund	-0.5437718	0.2318992	-2.3448626	0.0190378
teamDunkerque	-1.1071669	0.4137726	-2.6757856	0.0074577
teamDusseldorf	-0.3627148	0.2273481	-1.5954157	0.1106253
teamDynamo Dresden	-0.4334096	0.2362481	-1.8345529	0.0665776
teamEibar	-1.4048311	0.4472779	-3.1408465	0.0016855
teamEintracht Frankfurt	-0.0625966	0.2297555	-0.2724490	0.7852779
teamElche	-0.7769536	0.4475815	-1.7358932	0.0825886
teamEmpoli	-1.1207804	0.4005440	-2.7981452	0.0051416
teamErzgebirge Aue	-0.5020677	0.2286722	-2.1955777	0.0281265
teamEspanyol	-1.1613864	0.4464461	-2.6014034	0.0092869
teamEverton	-1.7040014	0.6558722	-2.5980690	0.0093776
teamEvian Thonon Gaillard	-1.1675901	0.4815884	-2.4244564	0.0153347
teamExtremadura	-1.2449072	0.4579606	-2.7183717	0.0065625
teamFiorentina	-0.5835070	0.4015820	-1.4530206	0.1462240
teamFoggia	-0.6973223	0.4421609	-1.5770781	0.1147835
teamFreiburg	-0.8938110	0.2292976	-3.8980394	0.0000971
teamFrosinone	-0.6256782	0.4012949	-1.5591482	0.1189673
teamFSV Frankfurt	-0.5597916	0.3575817	-1.5654928	0.1174734
teamFuenlabrada	-0.7574261	0.4506949	-1.6805740	0.0928516
teamFulham	-1.5605705	0.6541817	-2.3855308	0.0170580
teamGazelec Ajaccio	-0.7197956	0.4155595	-1.7321121	0.0832594
teamGenoa	-0.6728359	0.4014186	-1.6761456	0.0937156
teamGetafe	-0.4835413	0.4474144	-1.0807461	0.2798150
teamGimnastic Tarragona	-0.5081118	0.4572977	-1.1111182	0.2665225
teamGirona	-1.0520811	0.4461180	-2.3583023	0.0183624
teamGranada	-0.8604876	0.4470464	-1.9248283	0.0542562
teamGrenoble	-1.0855670	0.4076515	-2.6629780	0.0077476
teamGreuther Furth	-0.4132478	0.2271561	-1.8192242	0.0688829
teamGuingamp	-1.0677124	0.4048276	-2.6374498	0.0083556
teamHamburger SV	-0.6606914	0.2274379	-2.9049304	0.0036749
teamHannover 96	-0.5182133	0.2267676	-2.2852179	0.0223040
teamHansa Rostock	0.3032849	0.3008125	1.0082192	0.3133539
teamHeidenheim	-0.9044808	0.2271584	-3.9817192	0.0000685
teamHertha BSC	-0.4604596	0.2326961	-1.9788025	0.0478434
teamHoffenheim	-0.2614199	0.2297895	-1.1376495	0.2552720
teamHolstein Kiel	-0.7812413	0.2297626	-3.4002105	0.0006738
teamHuddersfield	-1.8137945	0.6534778	-2.7756023	0.0055119

teamHuesca	-1.0946426	0.4472206	-2.4476572	0.0143821
teamHull	-1.8089387	0.6560009	-2.7575246	0.0058261
teamIngolstadt	-0.6313530	0.2424922	-2.6036012	0.0092276
teamInter Milan	-0.8570420	0.4025521	-2.1290212	0.0332571
teamIpswich	-1.8174493	0.6592588	-2.7568069	0.0058389
teamJuve Stabia	-0.1720137	0.4247545	-0.4049721	0.6854996
teamJuventus	-0.8217914	0.4029094	-2.0396433	0.0413908
teamKaiserslautern	-0.4990633	0.2673108	-1.8669781	0.0619102
teamKarlsruhe	-0.5017428	0.2341600	-2.1427348	0.0321389
teamLaval	-1.3086331	0.4434781	-2.9508404	0.0031705
teamLazio	-0.4420794	0.4025712	-1.0981398	0.2721485
teamLe Havre	-1.1552149	0.4043385	-2.8570494	0.0042777
teamLe Mans	-1.0198984	0.4476434	-2.2783724	0.0227084
teamLecce	-0.7956600	0.4031019	-1.9738433	0.0484048
teamLeeds	-1.4406855	0.6549306	-2.1997531	0.0278287
teamLeganes	-0.9759192	0.4456910	-2.1896766	0.0285520
teamLeicester	-1.9417711	0.6561720	-2.9592410	0.0030853
teamLens	-0.6974548	0.4052355	-1.7211098	0.0852367
teamLeonesa	-0.9949654	0.4982287	-1.9970052	0.0458297
teamLevante	-1.1293877	0.4475397	-2.5235473	0.0116206
teamLeverkusen	-0.4064098	0.2305793	-1.7625597	0.0779806
teamLille	-0.8462668	0.4063481	-2.0826155	0.0372911
teamLiverpool	-1.9086533	0.6570737	-2.9047782	0.0036767
teamLivorno	-0.7510423	0.4130558	-1.8182584	0.0690303
teamLlagostera	-0.5828551	0.5094456	-1.1440969	0.2525886
teamLorca FC	-1.1333155	0.4982158	-2.2747480	0.0229250
teamLorient	-0.9706078	0.4054911	-2.3936599	0.0166847
teamLugo	-1.0275976	0.4457971	-2.3050790	0.0211660
teamLuton	-1.9039471	0.6562174	-2.9013968	0.0037166
teamLyon	-0.9011089	0.4073370	-2.2121954	0.0269574
teamMagdeburg	-0.5325300	0.3032349	-1.7561633	0.0790663
teamMainz 05	-0.4481319	0.2300245	-1.9481926	0.0513972
teamMalaga	-0.9582783	0.4464333	-2.1465206	0.0318360
teamMallorca	-0.9438500	0.4486192	-2.1039000	0.0353919
teamMan City	-1.5304132	0.6579222	-2.3261310	0.0200153
teamMan United	-1.3601495	0.6564775	-2.0718907	0.0382805
teamMarseille	-0.4923518	0.4064300	-1.2114062	0.2257452
teamMetz	-0.8774386	0.4058103	-2.1621892	0.0306081
teamMiddlesbrough	-1.7212996	0.6532454	-2.6349969	0.0084163
teamMillwall	-1.8601519	0.6546190	-2.8415796	0.0044908
teamMilton Keynes	-1.5577925	0.7275300	-2.1412071	0.0322619
teamMirandes	-1.3759932	0.4485301	-3.0677836	0.0021576
teamMonaco	-0.5796770	0.4087652	-1.4181174	0.1561624
teamMonchengladbach	-0.4775745	0.2306896	-2.0702038	0.0384381
teamMontpellier	-1.0419295	0.4063378	-2.5641953	0.0103443
teamMonza	-0.7395773	0.4110228	-1.7993584	0.0719677
teamMSV Duisburg	-0.2023647	0.2612180	-0.7746964	0.4385225
teamNancy	-1.1343617	0.4041941	-2.8064773	0.0050105
teamNantes	-0.9392800	0.4060850	-2.3130132	0.0207257
teamNapoli	-0.9285906	0.4026201	-2.3063694	0.0210938
teamNewcastle	-1.7254598	0.6554122	-2.6326332	0.0084750

teamNice	-0.8522753	0.4062932	-2.0976855	0.0359377
teamNimes	-1.0093960	0.4054735	-2.4894254	0.0127980
teamNiort	-0.9639116	0.4036189	-2.3881729	0.0169359
teamNorwich	-1.6737068	0.6546255	-2.5567394	0.0105686
teamNottingham Forest	-1.6511680	0.6536793	-2.5259604	0.0115411
teamNovara	-0.7009044	0.8715173	-0.8042347	0.4212651
teamNumancia	-1.2096093	0.4507218	-2.6837159	0.0072832
teamNurnberg	-0.6596607	0.2266649	-2.9102908	0.0036124
teamOrleans	-0.8970281	0.4117959	-2.1783317	0.0293857
teamOsasuna	-1.0143747	0.4469157	-2.2697225	0.0232284
teamOsnabruck	-0.6493412	0.2420944	-2.6821818	0.0073166
teamOviedo	-0.7622344	0.4455857	-1.7106347	0.0871544
teamPaderborn	-0.3651177	0.2301820	-1.5862131	0.1126970
teamPalermo	-0.5412759	0.4163899	-1.2999256	0.1936321
teamParis FC98	-1.0780713	0.4066657	-2.6510014	0.0080277
teamParis SG	-0.5353952	0.4087404	-1.3098661	0.1902468
teamParma	-0.8582443	0.4032345	-2.1284000	0.0333085
teamPau	-0.7039852	0.4130745	-1.7042570	0.0883390
teamPerugia	-0.6246602	0.4071740	-1.5341356	0.1250023
teamPescara	-0.8113032	0.4028529	-2.0138944	0.0440257
teamPeterborough	-1.6756020	0.6772647	-2.4740724	0.0133614
teamPisa	-0.4826027	0.4045639	-1.1928960	0.2329154
teamPonferradina	-1.4448415	0.4490549	-3.2175165	0.0012938
teamPordenone	-0.6880950	0.4043437	-1.7017576	0.0888067
teamPreston	-1.7903962	0.6539949	-2.7376301	0.0061904
teamPro Vercelli	-0.8471930	0.6369140	-1.3301529	0.1834736
teamQueens Park Rangers	-1.5543194	0.6544789	-2.3748962	0.0175574
teamQuevilly	-1.0052147	0.4259792	-2.3597741	0.0182897
teamRacing Santander	-0.8562443	0.4648639	-1.8419248	0.0654917
teamRayo Majadahonda	-0.9394282	0.4822389	-1.9480556	0.0514136
teamRayo Vallecano	-0.7034465	0.4463892	-1.5758590	0.1150643
teamRB Leipzig	-0.4492354	0.2305775	-1.9483056	0.0513837
teamReading	-1.8965525	0.6541982	-2.8990490	0.0037445
teamReal Madrid	-1.2815518	0.4482912	-2.8587484	0.0042548
teamReal Sociedad	-1.3009546	0.4473656	-2.9080344	0.0036386
teamReal Sociedad B	-1.4175290	0.4936482	-2.8715367	0.0040864
teamReal Zaragoza	-1.1467859	0.4451484	-2.5761879	0.0099923
teamRecreativo Huelva	-0.5308529	0.5945969	-0.8927945	0.3719713
teamRed Star	-1.0528007	0.4247068	-2.4788880	0.0131824
teamRegensburg	-0.2939713	0.2301297	-1.2774158	0.2014611
teamReggiana Audace	-0.6198893	0.4170766	-1.4862719	0.1372132
teamReggina	-0.3645836	0.4105034	-0.8881379	0.3744706
teamReims	-0.9598646	0.4060933	-2.3636553	0.0180992
teamRennes	-0.9396682	0.4064757	-2.3117449	0.0207956
teamReus	-1.6139504	0.4659040	-3.4641266	0.0005324
teamRodez Aveyron	-0.9914183	0.4082194	-2.4286412	0.0151588
teamRoma	-0.5082019	0.4026557	-1.2621252	0.2069093
teamRotherham	-1.7353932	0.6568055	-2.6421721	0.0082400
teamSabadell	-1.2391829	0.4585367	-2.7024726	0.0068848
teamSalernitana	-0.6573292	0.4009262	-1.6395265	0.1011096
teamSampdoria	-0.6853064	0.4017946	-1.7056139	0.0880859

teamSandhausen	-0.2698570	0.2271432	-1.1880478	0.2348199
teamSassuolo	-0.7244634	0.4015227	-1.8042902	0.0711915
teamSchalke 04	-0.3456890	0.2288899	-1.5102846	0.1309768
teamSevilla	-0.8934592	0.4472890	-1.9974987	0.0457762
teamSevilla B	-0.9146761	0.4756173	-1.9231347	0.0544685
teamSheffield United	-1.6636134	0.6557816	-2.5368406	0.0111886
teamSheffield Wednesday	-1.7446029	0.6541892	-2.6668173	0.0076597
teamSochaux	-1.1012533	0.4042648	-2.7240891	0.0064500
teamSouthampton	-1.8351703	0.6559635	-2.7976711	0.0051491
teamSpal	-0.6586641	0.4018080	-1.6392510	0.1011669
teamSpezia	-0.8353332	0.4015026	-2.0805177	0.0374829
teamSporting Gijon	-1.1486651	0.4462216	-2.5742033	0.0100498
teamSt Etienne	-0.8081767	0.4062802	-1.9892100	0.0466832
teamSt Pauli	-0.7144344	0.2273255	-3.1427824	0.0016744
teamStoke	-1.7634066	0.6542746	-2.6952088	0.0070367
teamStrasbourg	-1.0619083	0.4069064	-2.6097116	0.0090644
teamStuttgart	-0.3253456	0.2285529	-1.4235025	0.1545964
teamSunderland	-1.4688860	0.6675172	-2.2005216	0.0277742
teamSwansea	-2.0142319	0.6543895	-3.0780321	0.0020848
teamTenerife	-0.7347800	0.4469469	-1.6439983	0.1001825
teamTernana	-1.1747650	0.4538273	-2.5885731	0.0096401
teamTorino	-0.8892173	0.4015614	-2.2143996	0.0268056
teamTottenham	-1.5559757	0.6562987	-2.3708346	0.0177515
teamToulouse	-0.7916435	0.4044737	-1.9572186	0.0503270
teamTours	-0.9465037	0.4255209	-2.2243414	0.0261297
teamTrapani Calcio	-0.3221605	0.4265138	-0.7553343	0.4500518
teamTroyes	-0.8219273	0.4048276	-2.0303147	0.0423296
teamUCAM Murcia	-1.3228602	0.5204814	-2.5416090	0.0110372
teamUD Ibiza	-2.0788262	0.4766016	-4.3617690	0.0000129
teamUD Logrones	-1.3410497	0.4607639	-2.9104923	0.0036101
teamUdinese	-1.0758310	0.4016880	-2.6782751	0.0074025
teamUnion Berlin	-0.5734962	0.2280329	-2.5149710	0.0119072
teamUnion Deportivo Las Palmas	-0.6948036	0.4446043	-1.5627458	0.1181184
teamValencia	-0.9378210	0.4474286	-2.0960238	0.0360848
teamValenciennes	-0.8953851	0.4048768	-2.2114999	0.0270055
teamValladolid	-1.0455620	0.4472904	-2.3375464	0.0194145
teamVenezia	-0.6893726	0.4024891	-1.7127735	0.0867601
teamVerona	-0.7534287	0.4027347	-1.8707816	0.0613809
teamVicenza	-0.0118788	0.4101890	-0.0289593	0.9768972
teamVillarreal	-1.1411604	0.4484751	-2.5445346	0.0109452
teamVirtus Entella	-0.5999890	0.4078291	-1.4711774	0.1412491
teamWatford	-1.4491877	0.6544671	-2.2143020	0.0268123
teamWehen	-0.2680820	0.2788072	-0.9615319	0.3362892
teamWerder Bremen	-0.3485872	0.2289201	-1.5227458	0.1278283
teamWest Bromwich	-1.6864369	0.6549880	-2.5747599	0.0100337
teamWest Ham	-1.8567894	0.6557024	-2.8317562	0.0046311
teamWigan	-1.5257086	0.6592295	-2.3143817	0.0206506
teamWolfsburg	-0.4326979	0.2298615	-1.8824286	0.0597833
teamWolverhampton	-1.8012398	0.6544422	-2.7523285	0.0059193
teamWurzburger Kickers	-0.6496406	0.2519513	-2.5784372	0.0099275
teamWycombe	-1.8265796	0.6637880	-2.7517513	0.0059298

opponent1FC Koln	-0.1195693	0.2290561	-0.5220090	0.6016663
opponentAalen	-0.2889029	0.5010541	-0.5765902	0.5642188
opponentAC Milan	0.1125180	0.1409630	0.7982094	0.4247526
opponentAJ Auxerre	-0.6195457	0.1111705	-5.5729342	0.0000000
opponentAjaccio	-0.4301043	0.1108678	-3.8794355	0.0001048
opponentAlaves	0.1391979	0.1084075	1.2840247	0.1991389
opponentAlbacete	0.2969471	0.1176014	2.5250293	0.0115718
opponentAlcorcon	0.1509782	0.1122366	1.3451771	0.1785738
opponentAlessandria	-0.5580338	0.2655788	-2.1011984	0.0356283
opponentAlmeria	0.3604969	0.1156308	3.1176546	0.0018239
opponentAmiens	-0.3248258	0.1144634	-2.8378146	0.0045441
opponentAmorebieta	-0.6247865	0.2311052	-2.7034724	0.0068641
opponentAngers	-0.2631436	0.1161018	-2.2664898	0.0234254
opponentArles	-0.4602579	0.4340890	-1.0602845	0.2890200
opponentArminia Bielefeld	-0.2994409	0.2278896	-1.3139735	0.1888608
opponentArsenal	0.3901475	0.1569515	2.4857846	0.0129297
opponentAS Beziers	0.1334737	0.2193323	0.6085455	0.5428283
opponentAscoli	0.3056864	0.1395108	2.1911304	0.0284467
opponentAston Villa	0.5340413	0.1512489	3.5308782	0.0004145
opponentAtalanta	0.2942947	0.1401901	2.0992537	0.0357993
opponentAthletic Bilbao	0.1851143	0.1074467	1.7228482	0.0849218
opponentAthletic Bilbao B	0.9046987	0.3436121	2.6329069	0.0084682
opponentAtletico Madrid	-0.1015100	0.1103203	-0.9201384	0.3575046
opponentAugsburg	-0.2901190	0.2292903	-1.2652912	0.2057725
opponentAvellino	-0.4989908	0.7862934	-0.6346114	0.5256846
opponentBarcelona	0.1736835	0.1120783	1.5496617	0.1212287
opponentBarcelona B	0.6018679	0.2184091	2.7556904	0.0058588
opponentBari	-0.5741381	0.5820872	-0.9863438	0.3239689
opponentBarnsley	-0.0087776	0.1496161	-0.0586675	0.9532172
opponentBastia	-0.3192976	0.1701516	-1.8765484	0.0605855
opponentBayern Munich	-0.6023140	0.2345423	-2.5680396	0.0102303
opponentBenevento	0.1511492	0.1371780	1.1018470	0.2705332
opponentBetis	0.4111962	0.1078500	3.8126688	0.0001376
opponentBirmingham	-0.1052923	0.1462083	-0.7201531	0.4714339
opponentBlackburn	0.1689519	0.1476616	1.1441833	0.2525528
opponentBlackpool	0.5533787	0.2118719	2.6118553	0.0090078
opponentBochum	-0.3914520	0.2268672	-1.7254673	0.0844491
opponentBologna	0.3196973	0.1382023	2.3132563	0.0207124
opponentBolton	-0.1416171	0.1798573	-0.7873858	0.4310596
opponentBordeaux	-0.1816749	0.1163974	-1.5608162	0.1185731
opponentBourg Peronnas	-0.3622005	0.1850742	-1.9570560	0.0503461
opponentBournemouth	0.3283364	0.1497420	2.1926815	0.0283347
opponentBraunschweig	-0.2283535	0.2391945	-0.9546772	0.3397453
opponentBrentford	0.2802313	0.1473574	1.9017116	0.0572143
opponentBrescia	0.1208000	0.1372239	0.8803131	0.3786937
opponentBrest	-0.2278573	0.1130391	-2.0157392	0.0438323
opponentBrighton and Hove Albion	-0.0261268	0.1527267	-0.1710687	0.8641704
opponentBristol City	0.0706244	0.1468123	0.4810523	0.6304813
opponentBurgos	-0.6288746	0.2311639	-2.7204707	0.0065210
opponentBurnley	-0.2799321	0.1536479	-1.8219058	0.0684749
opponentBurton	-0.2992245	0.2151015	-1.3910852	0.1642054

opponentCadiz	0.2198640	0.1120025	1.9630275	0.0496482
opponentCaen	-0.2746492	0.1118710	-2.4550521	0.0140896
opponentCagliari	0.0289616	0.1372539	0.2110078	0.8328820
opponentCalcio Padova	0.0274036	0.2381172	0.1150845	0.9083786
opponentCardiff	-0.1890297	0.1465449	-1.2899093	0.1970878
opponentCarpi 1909	-0.0310200	0.2077017	-0.1493489	0.8812789
opponentCartagena	0.3864561	0.1459423	2.6480061	0.0080992
opponentCastellon	-0.0607770	0.1599744	-0.3799172	0.7040084
opponentCelta Vigo	0.6210010	0.1121506	5.5372077	0.0000000
opponentCesena	-0.3736396	0.3451108	-1.0826659	0.2789617
opponentChambly	0.0779763	0.1398850	0.5574313	0.5772351
opponentCharlton	0.3387246	0.1840283	1.8406119	0.0656841
opponentChateauroux	-0.5651392	0.1202019	-4.7015817	0.0000026
opponentChelsea	0.2218809	0.1586597	1.3984702	0.1619778
opponentChievo	-0.0360865	0.1381101	-0.2612880	0.7938715
opponentCittadella	-0.2322760	0.1393348	-1.6670353	0.0955133
opponentClermont	-0.2214611	0.1119905	-1.9774987	0.0479905
opponentComo	-0.1153592	0.2591101	-0.4452130	0.6561676
opponentCordoba	0.1534960	0.1451765	1.0573064	0.2903766
opponentCosenza 1914	0.0570144	0.1414110	0.4031818	0.6868161
opponentCoventry	0.5880142	0.1663532	3.5347340	0.0004085
opponentCremonese	-0.2639221	0.1387869	-1.9016364	0.0572241
opponentCreteil	-0.5173963	0.2860866	-1.8085304	0.0705297
opponentCrotone	0.2156993	0.1350712	1.5969307	0.1102871
opponentCrystal Palace	0.2909816	0.1541509	1.8876413	0.0590796
opponentDarmstadt	-0.0729304	0.2281159	-0.3197075	0.7491914
opponentDeportivo La Coruna	0.0522273	0.1256202	0.4157559	0.6775903
opponentDerby	0.0222884	0.1463827	0.1522612	0.8789815
opponentDijon	-0.3921480	0.1150189	-3.4094236	0.0006515
opponentDortmund	-0.0804185	0.2316254	-0.3471920	0.7284485
opponentDunkerque	-0.2456388	0.1419133	-1.7309084	0.0834739
opponentDusseldorf	-0.0567754	0.2270530	-0.2500535	0.8025469
opponentDynamo Dresden	-0.1770644	0.2366828	-0.7481087	0.4543979
opponentEibar	-0.2783650	0.1072433	-2.5956390	0.0094441
opponentEintracht Frankfurt	-0.0875120	0.2296114	-0.3811311	0.7031075
opponentElche	0.2076905	0.1122253	1.8506569	0.0642245
opponentEmpoli	-0.0279106	0.1343027	-0.2078188	0.8353712
opponentErzgebirge Aue	-0.0774505	0.2281468	-0.3394765	0.7342522
opponentEspanyol	0.2179148	0.1076295	2.0246743	0.0429058
opponentEverton	0.2056680	0.1545097	1.3311007	0.1831616
opponentEvian Thonon Gaillard	-0.3169697	0.2858078	-1.1090310	0.2674219
opponentExtremadura	0.4061588	0.1514667	2.6815057	0.0073314
opponentFiorentina	0.3126560	0.1386767	2.2545674	0.0241646
opponentFoggia	0.5617164	0.2313240	2.4282671	0.0151745
opponentFreiburg	-0.1929444	0.2295105	-0.8406776	0.4005323
opponentFrosinone	0.1831890	0.1357108	1.3498485	0.1770704
opponentFSV Frankfurt	-0.4841520	0.3575752	-1.3539866	0.1757464
opponentFuenlabrada	0.5335333	0.1260845	4.2315525	0.0000232
opponentFulham	0.0587222	0.1488531	0.3944975	0.6932153
opponentGazelec Ajaccio	-0.4815881	0.1473140	-3.2691261	0.0010795
opponentGenoa	0.1079311	0.1371141	0.7871626	0.4311902

opponentGetafe	0.5335443	0.1077243	4.9528702	0.0000007
opponentGimnastic Tarragona	0.2409035	0.1494604	1.6118212	0.1070068
opponentGirona	0.5992272	0.1100252	5.4462702	0.0000001
opponentGranada	0.3996420	0.1083663	3.6878798	0.0002264
opponentGrenoble	-0.1858703	0.1197823	-1.5517341	0.1207319
opponentGreuther Furth	-0.1255291	0.2270498	-0.5528702	0.5803547
opponentGuingamp	-0.1304431	0.1119851	-1.1648250	0.2440951
opponentHamburger SV	-0.2048804	0.2272795	-0.9014471	0.3673548
opponentHannover 96	-0.4165692	0.2270472	-1.8347250	0.0665521
opponentHansa Rostock	-0.3284606	0.2964704	-1.1079034	0.2679086
opponentHeidenheim	-0.4737387	0.2270047	-2.0869108	0.0369010
opponentHertha BSC	-0.3295603	0.2331155	-1.4137210	0.1574497
opponentHoffenheim	-0.2837155	0.2303692	-1.2315683	0.2181159
opponentHolstein Kiel	-0.2335178	0.2305828	-1.0127288	0.3111944
opponentHuddersfield	0.2353722	0.1465366	1.6062352	0.1082282
opponentHuesca	0.0860086	0.1087890	0.7906002	0.4291809
opponentHull	-0.0101611	0.1530541	-0.0663887	0.9470686
opponentIngolstadt	-0.4359255	0.2400322	-1.8161131	0.0693587
opponentInter Milan	0.1773821	0.1427715	1.2424195	0.2140873
opponentIpswich	-0.1494871	0.1717273	-0.8704914	0.3840359
opponentJuve Stabia	0.5254254	0.1987651	2.6434496	0.0082090
opponentJuventus	0.2498715	0.1419696	1.7600346	0.0784077
opponentKaiserslautern	-0.2656178	0.2673125	-0.9936601	0.3203929
opponentKarlsruhe	-0.2721202	0.2350543	-1.1576909	0.2469954
opponentLaval	-0.1787756	0.2156323	-0.8290764	0.4070649
opponentLazio	0.0991515	0.1414943	0.7007457	0.4834648
opponentLe Havre	-0.1874004	0.1103992	-1.6974802	0.0896118
opponentLe Mans	-0.2844895	0.2212728	-1.2856956	0.1985549
opponentLecce	0.1294916	0.1395809	0.9277172	0.3535585
opponentLeeds	0.3101727	0.1498021	2.0705495	0.0384058
opponentLeganes	0.0796904	0.1098658	0.7253434	0.4682446
opponentLeicester	0.0923287	0.1556993	0.5929937	0.5531879
opponentLens	-0.0247654	0.1133309	-0.2185225	0.8270229
opponentLeonesa	0.3736075	0.2478230	1.5075577	0.1316737
opponentLevante	0.3260266	0.1075761	3.0306611	0.0024414
opponentLeverkusen	-0.3135063	0.2307371	-1.3587165	0.1742422
opponentLille	-0.2008272	0.1194704	-1.6809782	0.0927731
opponentLiverpool	-0.1121436	0.1600423	-0.7007124	0.4834856
opponentLivorno	0.3828170	0.1712248	2.2357561	0.0253719
opponentLlagostera	0.4978878	0.2705802	1.8400744	0.0657629
opponentLorca FC	-0.0528498	0.2480088	-0.2130966	0.8312524
opponentLorient	-0.0439724	0.1144717	-0.3841329	0.7008815
opponentLugo	0.2061738	0.1123239	1.8355296	0.0664329
opponentLuton	-0.0335294	0.1560518	-0.2148608	0.8298767
opponentLyon	-0.1265118	0.1208026	-1.0472610	0.2949840
opponentMagdeburg	0.0282450	0.3032559	0.0931392	0.9257933
opponentMainz 05	0.0252936	0.2297732	0.1100809	0.9123456
opponentMalaga	0.2326996	0.1107881	2.1004035	0.0356981
opponentMallorca	0.3749805	0.1174092	3.1937911	0.0014050
opponentMan City	0.0316868	0.1621623	0.1954019	0.8450790
opponentMan United	0.1492295	0.1585905	0.9409741	0.3467224

opponentMarseille	-0.2543158	0.1194569	-2.1289346	0.0332643
opponentMetz	-0.3834546	0.1149412	-3.3360935	0.0008502
opponentMiddlesbrough	0.1805857	0.1467711	1.2303900	0.2185566
opponentMillwall	-0.1016190	0.1476272	-0.6883484	0.4912364
opponentMilton Keynes	0.4966165	0.3502327	1.4179615	0.1562079
opponentMirandes	-0.0149310	0.1185715	-0.1259243	0.8997923
opponentMonaco	-0.2115658	0.1248927	-1.6939813	0.0902747
opponentMonchengladbach	0.0887163	0.2304065	0.3850425	0.7002075
opponentMontpellier	0.0494889	0.1173017	0.4218939	0.6731041
opponentMonza	-0.0155219	0.1600905	-0.0969572	0.9227607
opponentMSV Duisburg	-0.2270435	0.2612363	-0.8691116	0.3847900
opponentNancy	-0.1257013	0.1118927	-1.1234088	0.2612690
opponentNantes	-0.1187248	0.1166815	-1.0175120	0.3089146
opponentNapoli	0.4393783	0.1421434	3.0910916	0.0019953
opponentNewcastle	0.1527777	0.1532596	0.9968558	0.3188391
opponentNice	-0.2344144	0.1164286	-2.0133743	0.0440803
opponentNimes	0.2240249	0.1132701	1.9777938	0.0479572
opponentNiort	-0.3205907	0.1112272	-2.8823050	0.0039494
opponentNorwich	0.3168014	0.1489832	2.1264243	0.0334726
opponentNottingham Forest	0.1398457	0.1469888	0.9514035	0.3414039
opponentNovara	-0.5347689	0.7862121	-0.6801840	0.4963909
opponentNumancia	0.4898851	0.1271641	3.8523863	0.0001171
opponentNurnberg	-0.1489880	0.2265012	-0.6577802	0.5106822
opponentOrleans	-0.5970109	0.1358623	-4.3942367	0.0000111
opponentOsasuna	0.1146883	0.1084611	1.0574140	0.2903275
opponentOsnabruck	-0.0263550	0.2421037	-0.1088583	0.9133153
opponentOviedo	0.1954777	0.1118617	1.7474942	0.0805575
opponentPaderborn	-0.0430011	0.2311209	-0.1860545	0.8524027
opponentPalermo	0.0249370	0.1795744	0.1388675	0.8895554
opponentParis FC98	-0.2144466	0.1205027	-1.7795997	0.0751472
opponentParis SG	-0.5153745	0.1261074	-4.0867890	0.0000438
opponentParma	0.1327792	0.1398238	0.9496179	0.3423108
opponentPau	-0.3829158	0.1389210	-2.7563573	0.0058469
opponentPerugia	0.0073354	0.1568512	0.0467665	0.9626995
opponentPescara	-0.0144596	0.1416783	-0.1020591	0.9187102
opponentPeterborough	0.1365062	0.2270891	0.6011129	0.5477674
opponentPisa	0.3711347	0.1432902	2.5900914	0.0095977
opponentPonferradina	0.5491252	0.1232334	4.4559770	0.0000084
opponentPordenone	0.4202317	0.1424963	2.9490711	0.0031887
opponentPreston	0.0135758	0.1470222	0.0923387	0.9264293
opponentPro Vercelli	-0.7887151	0.5153295	-1.5305064	0.1258974
opponentQueens Park Rangers	-0.0799752	0.1463915	-0.5463105	0.5848548
opponentQuevilly	-0.0225757	0.1730131	-0.1304854	0.8961829
opponentRacing Santander	0.3863564	0.1719464	2.2469575	0.0246468
opponentRayo Majadahonda	0.4716610	0.2140741	2.2032607	0.0275807
opponentRayo Vallecano	0.1776068	0.1089444	1.6302518	0.1030542
opponentRB Leipzig	-0.0648578	0.2307256	-0.2811036	0.7786321
opponentReading	0.4199728	0.1460145	2.8762400	0.0040260
opponentReal Madrid	0.0980445	0.1103376	0.8885867	0.3742293
opponentReal Sociedad	0.3582147	0.1085241	3.3007853	0.0009648
opponentReal Sociedad B	-0.1896897	0.2281107	-0.8315686	0.4056562

opponentReal Zaragoza	0.1587707	0.1110151	1.4301717	0.1526736
opponentRecreativo Huelva	0.0271977	0.4092533	0.0664568	0.9470144
opponentRed Star	-0.5531182	0.1714832	-3.2254954	0.0012583
opponentRegensburg	-0.1518168	0.2305814	-0.6584086	0.5102785
opponentReggiana Audace	-0.0252210	0.1773624	-0.1422001	0.8869225
opponentReggina	0.2298912	0.1621226	1.4180084	0.1561942
opponentReims	-0.4576501	0.1154943	-3.9625351	0.0000743
opponentRennes	-0.0844092	0.1172227	-0.7200752	0.4714819
opponentReus	-0.1157347	0.1737512	-0.6660946	0.5053536
opponentRodez Aveyron	-0.1973162	0.1293175	-1.5258284	0.1270586
opponentRoma	0.0055344	0.1413391	0.0391566	0.9687657
opponentRotherham	-0.4333533	0.1573949	-2.7532868	0.0059020
opponentSabadell	-0.0444912	0.1538050	-0.2892701	0.7723758
opponentSalernitana	0.2011505	0.1375445	1.4624396	0.1436267
opponentSampdoria	0.1627513	0.1369852	1.1880940	0.2348017
opponentSandhausen	-0.1955065	0.2272567	-0.8602891	0.3896336
opponentSassuolo	0.3877843	0.1371781	2.8268684	0.0047023
opponentSchalke 04	-0.1126714	0.2283647	-0.4933836	0.6217437
opponentSevilla	0.4022769	0.1080761	3.7221633	0.0001977
opponentSevilla B	0.2333046	0.1979619	1.1785332	0.2385894
opponentSheffield United	-0.1713043	0.1530088	-1.1195716	0.2629014
opponentSheffield Wednesday	-0.1967216	0.1489405	-1.3208066	0.1865716
opponentSochaux	-0.2666516	0.1111060	-2.3999760	0.0163996
opponentSouthampton	0.0923170	0.1547238	0.5966570	0.5507389
opponentSpal	0.2047453	0.1374532	1.4895640	0.1363449
opponentSpezia	0.1805750	0.1392596	1.2966788	0.1947474
opponentSporting Gijon	0.0200782	0.1103139	0.1820100	0.8555756
opponentSt Etienne	-0.1836660	0.1168580	-1.5717027	0.1160255
opponentSt Pauli	-0.3062623	0.2267858	-1.3504474	0.1768783
opponentStoke	-0.1140325	0.1477995	-0.7715347	0.4403935
opponentStrasbourg	-0.1510668	0.1197519	-1.2614984	0.2071349
opponentStuttgart	-0.1897168	0.2285456	-0.8301051	0.4064831
opponentSunderland	0.0619866	0.1984586	0.3123402	0.7547832
opponentSwansea	0.1603730	0.1471396	1.0899376	0.2757456
opponentTenerife	0.2311602	0.1141675	2.0247457	0.0428985
opponentTernana	-0.1794406	0.2288625	-0.7840542	0.4330118
opponentTorino	0.3871369	0.1378179	2.8090458	0.0049707
opponentTottenham	0.4103833	0.1575216	2.6052506	0.0091833
opponentToulouse	-0.1356768	0.1133752	-1.1967067	0.2314263
opponentTours	-0.2433443	0.1747284	-1.3926999	0.1637164
opponentTrapani Calcio	0.1274653	0.1987014	0.6414915	0.5212062
opponentTroyes	-0.2477100	0.1112720	-2.2261671	0.0260072
opponentUCAM Murcia	-0.0569203	0.2900592	-0.1962369	0.8444255
opponentUD Ibiza	-0.4837694	0.2304748	-2.0990120	0.0358206
opponentUD Logrones	-0.4601964	0.1602848	-2.8711170	0.0040919
opponentUdinese	0.0567381	0.1367324	0.4149572	0.6781749
opponentUnion Berlin	-0.4067352	0.2280456	-1.7835696	0.0744994
opponentUnion Deportivo Las Palmas	0.4674431	0.1125761	4.1522401	0.0000330
opponentValencia	0.4508587	0.1078313	4.1811503	0.0000291
opponentValladolid	0.3680806	0.1076312	3.4198307	0.0006271
opponentVenezia	0.0202850	0.1378853	0.1471146	0.8830421

opponentVerona	0.2369603	0.1402748	1.6892579	0.0911759
opponentVicenza	-0.0161029	0.1581552	-0.1018171	0.9189022
opponentWatford	0.0162592	0.1504619	0.1080621	0.9139468
opponentWehen	-0.3062863	0.2788367	-1.0984433	0.2720160
opponentWerder Bremen	0.0062082	0.2289382	0.0271173	0.9783663
opponentWest Bromwich	-0.0449640	0.1492349	-0.3012966	0.7631895
opponentWest Ham	0.0109401	0.1544646	0.0708258	0.9435366
opponentWigan	0.3601227	0.1662863	2.1656791	0.0303402
opponentWolfsburg	-0.1559229	0.2300488	-0.6777816	0.4979131
opponentWolverhampton	0.3034232	0.1522611	1.9927828	0.0462904
opponentWurzburger Kickers	0.2495957	0.2519522	0.9906469	0.3218626
home	-0.2335822	0.0276659	-8.4429635	0.0000000
refereeAdrian CORDERO VEGA	0.4741143	0.2459655	1.9275642	0.0539147
refereeAfonso Suarez	-0.1739334	0.7442212	-0.2337119	0.8152095
refereeAlain Bieri	0.1460478	0.6000440	0.2433951	0.8077003
refereeAlain Durieux	0.5748304	0.6390458	0.8995137	0.3683832
refereeAlberto SANTORO	-0.2550002	0.1747705	-1.4590576	0.1445552
refereeAlberto Undiano Mallenco	0.4948637	0.2680112	1.8464294	0.0648355
refereeAleandro Di Paolo	-0.0547746	0.2332135	-0.2348689	0.8143114
refereeAlejandro Hernandez Hernandez	1.1227000	0.2494577	4.5005620	0.0000068
refereeAlejandro MUÑIZ RUIZ	0.5031510	0.2428281	2.0720459	0.0382660
refereeAlessandro Dudic	0.3214282	0.6426591	0.5001534	0.6169691
refereeAlessandro PRONTERA	0.2484855	0.1451841	1.7115198	0.0869911
refereeAlexander Sather	-0.2769117	0.2576277	-1.0748523	0.2824458
refereeAlexandre Castro	-0.0897183	0.2178967	-0.4117469	0.6805266
refereeAlexandre Perreau Niel	0.5863540	0.1599185	3.6665798	0.0002461
refereeAlfonso Alvarez Izquierdo	0.7964776	0.2874878	2.7704741	0.0055994
refereeAmaury Delerue	0.6681331	0.1685387	3.9642721	0.0000737
refereeAndre Marriner	0.5372166	0.5341493	1.0057424	0.3145441
refereeAndrea COLOMBO	-0.2976343	0.3101318	-0.9597026	0.3372093
refereeAndrea Gervasoni	0.5599201	0.2964725	1.8886073	0.0589499
refereeAndrew Madley	0.4704803	0.5333756	0.8820806	0.3777372
refereeAndy D'Urso	0.2673118	0.7100566	0.3764654	0.7065724
refereeAndy Davies	0.5079688	0.5338272	0.9515603	0.3413244
refereeAndy Haines	0.2544340	0.7617490	0.3340129	0.7383711
refereeAndy Woolmer	0.3439309	0.5342693	0.6437407	0.5197464
refereeAngelo Cervellera	0.4811912	0.3168707	1.5185729	0.1288760
refereeAnthony Taylor	0.7695356	0.5343775	1.4400599	0.1498564
refereeAntonello BALICE	0.1151622	0.6379798	0.1805107	0.8567524
refereeAntonio DAMATO	0.2922126	0.2054441	1.4223462	0.1549317
refereeAntonio DI MARTINO	0.3599075	0.1417180	2.5396031	0.0111007
refereeAntonio GIUA	0.2455003	0.1417672	1.7317148	0.0833302
refereeAntonio Mateu Lahoz	0.5762176	0.2493827	2.3105756	0.0208601
refereeAntónio Nobre	-0.5535450	0.5582464	-0.9915783	0.3214078
refereeAntonio RAPUANO	0.3891969	0.1411914	2.7565194	0.0058440
refereeAntony Coggins	1.3953709	0.9140991	1.5264985	0.1268917
refereeAntony Gautier	0.1378799	0.1697467	0.8122685	0.4166412
refereeArcerdiano MONESCILLO	0.3623448	0.2477813	1.4623572	0.1436493
refereeARECES FRANCO, Victor	0.6202302	0.2546256	2.4358518	0.0148600
refereeArias Lopez	0.3360922	0.3052885	1.1009003	0.2709451
refereeArnaud Baert	0.0818617	0.1573762	0.5201655	0.6029504

refereeArne Aarnink	-0.2444095	0.2174234	-1.1241179	0.2609681
refereeAurelien Petit	0.2297479	0.1617554	1.4203418	0.1555141
refereeBartolomeo Varela	0.3167309	0.1641471	1.9295554	0.0536673
refereeBastian Dankert	0.1744744	0.1979511	0.8814019	0.3781043
refereeBastien Dechepy	0.2556803	0.1591417	1.6066201	0.1081437
refereeBenedikt Kempkes	-0.0745462	0.2543720	-0.2930599	0.7694775
refereeBenjamin Brand	-0.3655221	0.2192394	-1.6672280	0.0954750
refereeBenjamin Cortus	0.0465315	0.2012294	0.2312360	0.8171323
refereeBenjamin Lepaysant	0.3997654	0.1565031	2.5543607	0.0106411
refereeBenoit Bastien	0.5191208	0.1707706	3.0398723	0.0023679
refereeBenoit Millot	0.6370390	0.1716050	3.7122411	0.0002056
refereeBibiana Steinhaus	-0.2899603	0.2284543	-1.2692266	0.2043658
refereeBrendan Malone	0.0740163	0.7050495	0.1049803	0.9163919
refereeCarl Berry	0.7151219	0.9935573	0.7197591	0.4716765
refereeCarl Boyeson	0.6397707	0.6677187	0.9581440	0.3379945
refereeCarlos Clos Gomez	0.7944358	0.3048256	2.6061980	0.0091579
refereeCarlos del Cerro Grande	0.5479780	0.2499486	2.1923624	0.0283577
refereeCarlos Velasco Carballo	1.3990865	0.3677738	3.8042039	0.0001424
refereeCarmine Russo	0.0447701	0.2458204	0.1821253	0.8554851
refereeCédric Dos Santos	0.3097280	0.2026821	1.5281467	0.1264821
refereeCELI Domenico	0.2261429	0.2673285	0.8459364	0.3975920
refereeCesar Muniz Fernandez	-0.4071929	0.5809758	-0.7008776	0.4833825
refereeCesar SOTO GRADO	0.5189005	0.2420796	2.1435117	0.0320766
refereeCharles Breakspear	0.4058506	0.6759434	0.6004209	0.5482283
refereeChris Foy	0.6664929	0.6610816	1.0081856	0.3133700
refereeChris Kavanagh	0.6585744	0.5334683	1.2345147	0.2170167
refereeChris Sarginson	0.9705280	0.9315326	1.0418615	0.2974807
refereeChristian Bandurski	0.5439635	0.6943368	0.7834289	0.4333788
refereeChristian Dietz	0.2317032	0.2650211	0.8742823	0.3819685
refereeChristian DINGERT	0.1459155	0.1989176	0.7335475	0.4632278
refereeChristof Günsch	0.2084398	0.2519384	0.8273442	0.4080457
refereeClaudio Gavillucci	0.3249110	0.2146860	1.5134238	0.1301780
refereeClement Turpin	0.0362870	0.1702980	0.2130796	0.8312657
refereeCraig Hicks	-0.5353216	0.8007947	-0.6684879	0.5038251
refereeCraig Pawson	0.7832345	0.5340897	1.4664848	0.1425222
refereeDaniel AMABILE	-0.1716953	0.1627303	-1.0550914	0.2913883
refereeDaniel Schlager	0.0120662	0.2562635	0.0470850	0.9624457
refereeDaniel Siebert	-0.0102102	0.1984458	-0.0514510	0.9589663
refereeDaniel Trujillo Suarez	0.3716456	0.2458155	1.5118885	0.1305682
refereeDaniele Chiffi	-0.0253369	0.1386163	-0.1827845	0.8549678
refereeDaniele DOVERI	0.1734149	0.1398044	1.2404108	0.2148290
refereeDaniele MARTINELLI	0.0257257	0.4579815	0.0561719	0.9552051
refereeDaniele MINELLI	0.0656010	0.1575413	0.4164050	0.6771153
refereeDaniele Orsato	0.6436017	0.1407912	4.5713204	0.0000049
refereeDaniele PATERNA	-0.3292716	0.1736505	-1.8961748	0.0579424
refereeDarren Bond	0.6016208	0.5331866	1.1283495	0.2591775
refereeDarren Deadman	0.4396735	0.6326071	0.6950183	0.4870469
refereeDarren Drysdale	0.0809691	0.6448400	0.1255646	0.9000771
refereeDarren England	0.6487392	0.5344275	1.2138956	0.2247930
refereeDarren Handley	0.4747878	0.8750620	0.5425762	0.5874239
refereeDarren Sheldrake	0.7540114	2.0101423	0.3751035	0.7075850

refereeDavid Coote	0.9102678	0.5333271	1.7067720	0.0878703
refereeDavid Fernandez Borbalan	0.9194921	0.2840670	3.2368848	0.0012092
refereeDavid GONZALEZ RASCON	-0.1720018	0.2480104	-0.6935265	0.4879822
refereeDavid Perez Pallas	0.7837422	0.2651267	2.9561044	0.0031169
refereeDavid Webb	0.3362214	0.5338864	0.6297621	0.5288530
refereeDavide Ghersini	-0.2214734	0.1391667	-1.5914245	0.1115200
refereeDavide Massa	0.5169854	0.1384711	3.7335256	0.0001890
refereeDe La Fuente, Oliver	0.6874751	0.2441706	2.8155524	0.0048711
refereeDean Whitestone	0.1092795	0.5358835	0.2039239	0.8384138
refereeDeniz Aytekin	-0.0657303	0.2009254	-0.3271375	0.7435652
refereeDIAZ DE MERA	0.9239676	0.2481167	3.7239241	0.0001964
refereeDidier Falcone	0.3066376	0.4154737	0.7380433	0.4604914
refereeDino Tommasi	-0.0877632	0.4480921	-0.1958597	0.8447207
refereeDominguez CERVANTES	0.2790380	0.3091608	0.9025658	0.3667605
refereeEddie Ilderton	0.3346153	0.6407986	0.5221848	0.6015439
refereeEduardo Prieto Iglesias	0.3361546	0.2454223	1.3696988	0.1707867
refereeEmilio Ostinelli	-0.5779822	1.2538381	-0.4609704	0.6448218
refereeEric Poulat	-0.9205191	0.7361304	-1.2504838	0.2111284
refereeEric Wattellier	-0.1481706	0.1592404	-0.9304835	0.3521250
refereeEstrada Fernandez	0.7685060	0.2510683	3.0609437	0.0022075
refereeEugenio Abbattista	-0.1157948	0.1418698	-0.8162047	0.4143867
refereeFabio MARESCA	0.7907576	0.1379191	5.7334890	0.0000000
refereeFabio PISCOPO	-0.4289441	0.2561782	-1.6743974	0.0940584
refereeFabrizio Pasqua	-0.3512609	0.1461014	-2.4042259	0.0162102
refereeFaouzi Benchabane	0.4352917	0.2074468	2.0983290	0.0358808
refereeFedayi San	0.9903119	0.8563059	1.1564932	0.2474847
refereeFederico DIONISI	0.0659575	0.1448114	0.4554720	0.6487716
refereeFederico La Penna	0.2115412	0.1519156	1.3924917	0.1637794
refereeFelix Brych	-0.1295551	0.1968764	-0.6580529	0.5105070
refereeFelix Zwayer	0.3234404	0.1977678	1.6354555	0.1019595
refereeFernando Teixeira Vitienes	1.1022187	0.4936058	2.2329939	0.0255535
refereeFlorent Batta	0.3868835	0.1670497	2.3159775	0.0205633
refereeFlorian Badstübner	-0.0473728	0.2533624	-0.1869763	0.8516799
refereeFlorian Heft	0.2481057	0.2460349	1.0084165	0.3132592
refereeFlorian Meyer	0.1072646	0.3114750	0.3443762	0.7305647
refereeFloris Aubin	0.6660882	0.3105094	2.1451467	0.0319457
refereeFrancesco COSSO	0.5473440	0.2735964	2.0005528	0.0454457
refereeFrancesco FOURNEAU	0.4936432	0.1395519	3.5373457	0.0004045
refereeFrancesco GUCCINI	-0.1711669	0.2648130	-0.6463689	0.5180433
refereeFrancesco MERAVIGLIA	0.4990843	0.1707574	2.9227689	0.0034708
refereeFrancisco Jose HERNANDEZ MAESO	0.3320689	0.2457963	1.3509921	0.1767037
refereeFrancisco Manuel Arias Lopez	0.5547126	0.6705770	0.8272169	0.4081179
refereeFrancisco Ontanaya Lopez	0.2329362	1.1616297	0.2005254	0.8410705
refereeFrancois Letexier	0.4612480	0.1694329	2.7223040	0.0064850
refereeFrank Schneider	0.5511948	0.1688879	3.2636730	0.0011005
refereeFrank Willenborg	-0.0120609	0.2031404	-0.0593723	0.9526558
refereeFred Graham	0.5000741	0.6635195	0.7536690	0.4510513
refereeFredy Fautrel	0.1492716	0.2734962	0.5457903	0.5852124
refereeGael Angoula	-0.0279173	0.1616914	-0.1726581	0.8629208
refereeGary Sutton	0.3899508	0.8710658	0.4476709	0.6543926
refereeGavin Ward	0.6044215	0.5343620	1.1311087	0.2580145

refereeGeoff Eltrigham	0.5411289	0.5327984	1.0156354	0.3098077
refereeGiacomo CAMPLONE	-0.1030149	0.1539748	-0.6690375	0.5034745
refereeGianluca AURELIANO	0.3688724	0.1395858	2.6426214	0.0082291
refereeGianluca MANGANIELLO	-0.1225110	0.1384661	-0.8847729	0.3762831
refereeGianluca Rocchi	0.5921178	0.1584035	3.7380341	0.0001857
refereeGianpaolo Calvarese	0.1477287	0.1428712	1.0339993	0.3011412
refereeGianpiero MIELE	0.4535208	0.2593571	1.7486348	0.0803600
refereeGiovanni AYROLDI	0.3341371	0.1470448	2.2723496	0.0230694
refereeGorka Gardeazabal Gomez	1.1846608	0.7024929	1.6863669	0.0917311
refereeGorka Sagues Oscoz	0.5054702	0.2461276	2.0536912	0.0400105
refereeGorostegui FERNANDEZ-ORTEGA	0.5743657	0.2398121	2.3950656	0.0166208
refereeGraham Horwood	0.8290207	0.9662006	0.8580213	0.3908846
refereeGraham Salisbury	0.5879714	0.6791956	0.8656879	0.3866653
refereeGraham Scott	0.4805170	0.5336428	0.9004469	0.3678865
refereeGuido Winkmann	0.0175439	0.2023620	0.0866958	0.9309136
refereeGuillaume Paradis	0.1021081	0.1708869	0.5975185	0.5501638
refereeGuillermo CUADRA FERNANDEZ	0.2449360	0.2456019	0.9972888	0.3186289
refereeGünter Perl	-0.1317447	0.2670490	-0.4933356	0.6217776
refereeHakan Ceylan	0.2907159	1.4559254	0.1996777	0.8417334
refereeHakim Ben El Hadj Salem	0.3981800	0.1676594	2.3749336	0.0175557
refereeHamid Guenaoui	0.3252517	0.3532082	0.9208496	0.3571331
refereeHarm Osmers	-0.2159079	0.2010012	-1.0741621	0.2827549
refereeIain Williamson	0.8240968	0.6666487	1.2361784	0.2163977
refereeIgnacio Iglesias Villanueva	0.7650726	0.2632611	2.9061358	0.0036607
refereeInaki Vikandi Garrido	0.0923962	0.2480597	0.3724757	0.7095402
refereeIosu Galech APEZTEGUIA	0.7775732	0.2401612	3.2377140	0.0012057
refereeIsidro ESCUDEROS	0.3546368	0.2538512	1.3970264	0.1624115
refereeIVAN CAPARROS	-0.9509938	0.3460699	-2.7479821	0.0059984
refereeIvan ROBILOTTA	0.4263822	0.1689573	2.5236088	0.0116186
refereeIvano Pezzuto	-0.2812329	0.1397794	-2.0119760	0.0442275
refereeJaime Latre	0.5753648	0.2498239	2.3030814	0.0212781
refereeJames Adcock	0.4962863	0.5696990	0.8711378	0.3836829
refereeJames Linington	0.5846422	0.5330780	1.0967292	0.2727648
refereeJARRED GILLETT	0.4887758	0.5374934	0.9093616	0.3631634
refereeJavier ALBEROLA ROJAS	-0.2155379	0.2466587	-0.8738302	0.3822147
refereeJAVIER Iglesias Villanueva	0.0746012	0.2371521	0.3145710	0.7530886
refereeJean-Charles Cailleux	-0.0971193	0.4693240	-0.2069344	0.8360619
refereeJeremie Pignard	0.1966126	0.1611728	1.2198867	0.2225133
refereeJeremy Simpson	0.6437984	0.5329039	1.2080949	0.2270162
refereeJeremy Stinat	0.3407715	0.1620106	2.1033908	0.0354363
refereeJerome Brisard	0.1200525	0.1631746	0.7357302	0.4618982
refereeJerome Miguelgorry	0.2640228	0.1670530	1.5804734	0.1140045
refereeJesus Gil Manzano	0.7217804	0.2498458	2.8889037	0.0038674
refereeJesus Munoz Mayordomo	0.7641700	0.4938183	1.5474721	0.1217554
refereeJochen Drees	-0.0769973	0.2940759	-0.2618280	0.7934551
refereeJohan HAMEL	0.3269637	0.1699550	1.9238247	0.0543819
refereeJohann Pfeifer	-0.2996096	0.2886458	-1.0379835	0.2992825
refereeJohn Brooks	0.7726063	0.5351307	1.4437713	0.1488092
refereeJohn Busby	0.8174829	0.5599707	1.4598673	0.1443325
refereeJON ANDER GONZALEZ	-0.0126808	0.2431914	-0.0521433	0.9584147
refereeJonathan Moss	0.6724762	0.5341414	1.2589853	0.2080412

refereeJorge Figueroa Vazquez	0.1750484	0.2539781	0.6892263	0.4906839
refereeJorge FIGUEROA VAZQUEZ	-0.0547382	0.2550098	-0.2146516	0.8300398
refereeJorge Valdes Aller	0.3019733	0.2780666	1.0859747	0.2774951
refereeJose Antonio LOPEZ TOCA	0.4935957	0.2387490	2.0674251	0.0386990
refereeJosé Antonio Texeira Vitienes	0.9602506	0.5147196	1.8655800	0.0621057
refereeJose Luis Gonzalez Gonzalez	0.3812840	0.2585091	1.4749346	0.1402361
refereeJosé Luis Lesma López	0.6055334	0.4416172	1.3711724	0.1703270
refereeJose Luis Munuera Montero	0.2884413	0.2486062	1.1602338	0.2459589
refereeJose Maria Sanchez Martinez	0.9042938	0.3175746	2.8475001	0.0044081
refereeJosé María Sánchez Martínez	0.8822837	0.2462771	3.5824829	0.0003406
refereeJose Ramon Pineiro Crespo	0.6319487	0.3380523	1.8693816	0.0615753
refereeJose Zorilla	1.0803962	1.2362042	0.8739626	0.3821426
refereeJosh Smith	0.3449482	0.5532959	0.6234426	0.5329964
refereeJOSU Galech Apezteguía	1.2882376	0.8147528	1.5811391	0.1138522
refereeJuan Luca Sacchi	-0.0100825	0.1367626	-0.0737224	0.9412316
refereeJuan Martinez Munuera	0.5508833	0.2499552	2.2039277	0.0275337
refereeKarim Abed	0.2187050	0.1678884	1.3026809	0.1926894
refereeKeith Hill	0.5930394	0.6436885	0.9213143	0.3568906
refereeKeith Stroud	0.7050871	0.5328726	1.3231813	0.1857808
refereeKevin Friend	0.6394145	0.5338651	1.1977081	0.2310360
refereeKevin Johnson	-0.0433124	0.8766002	-0.0494095	0.9605931
refereeKevin Wright	0.5136877	0.6568916	0.7819977	0.4342195
refereeKnut Kircher	-0.5516470	0.3082875	-1.7893912	0.0735575
refereeLasse Koslowski	-0.0146159	0.2526365	-0.0578536	0.9538655
refereeLee Collins	-0.1441108	1.1219841	-0.1284428	0.8977990
refereeLee Mason	0.5023878	0.5351401	0.9387967	0.3478394
refereeLee Probert	0.2176412	0.5441988	0.3999295	0.6892100
refereeLeigh DOUGHTY	0.7959392	0.5474829	1.4538154	0.1460034
refereeLionel Jaffredo	0.3400746	0.2459167	1.3828852	0.1667059
refereeLivio MARINELLI	0.1836732	0.1393393	1.3181720	0.1874518
refereeLopez Acera	0.1957289	0.7289293	0.2685156	0.7883035
refereeLOPEZ AMAYA, Juan Manuel	0.6430976	0.3096161	2.0770807	0.0377989
refereeLorenzo ILLUZZI	0.3527465	0.1549662	2.2762805	0.0228332
refereeLorenzo MAGGIONI	-0.0999489	0.1435227	-0.6963977	0.4861828
refereeLuca Banti	0.1935859	0.1741954	1.1113146	0.2664380
refereeLuca MASSIMI	0.2574761	0.1424258	1.8077907	0.0706448
refereeLuca Pairetto	0.1806495	0.1361224	1.3271111	0.1844777
refereeLuca ZUFFERLI	0.0695466	0.2724714	0.2552438	0.7985358
refereeLuigi Nasca	-0.0623130	0.2277456	-0.2736080	0.7843870
refereeLuigi PILLITTERI	0.0425333	0.2517815	0.1689292	0.8658529
refereeLuis Godinho	0.4933148	0.5636217	0.8752588	0.3814371
refereeLuis Mario MILLA ALVENDIZ	0.9649582	0.2379884	4.0546432	0.0000503
refereeLuis Medina Cantalejo	0.8427850	1.5361774	0.5486248	0.5832653
refereeManuel Angel Perez Lima	0.5698233	1.5681651	0.3633695	0.7163303
refereeManuel Grafe	-0.5342820	0.1976111	-2.7037051	0.0068593
refereeManuel VOLPI	0.1225355	0.1415127	0.8658974	0.3865504
refereeMarc Bollengier	-0.0543365	0.2258542	-0.2405822	0.8098799
refereeMarco Di Bello	0.3548260	0.1407483	2.5209962	0.0117052
refereeMarco Fritz	-0.1275870	0.1991757	-0.6405755	0.5218013
refereeMarco Guida	0.4776562	0.1409498	3.3888394	0.0007024
refereeMarco PICCININI	0.0659632	0.1419996	0.4645311	0.6422692

refereeMarco SERRA	0.2277883	0.1441937	1.5797379	0.1141729
refereeMaria MAROTTA	-0.8134450	0.5967734	-1.3630718	0.1728657
refereeMaria Sole FERRIERI CAPUTI	-0.7560028	0.5503860	-1.3735866	0.1695759
refereeMario Melero Lopez	0.6986483	0.2482308	2.8145112	0.0048870
refereeMark Brown	0.6142310	0.6333641	0.9697912	0.3321550
refereeMark Clattenburg	0.5445501	0.5560107	0.9793879	0.3273928
refereeMark Heywood	0.6952900	0.6062233	1.1469206	0.2514196
refereeMarkus Schmidt	0.1489197	0.2022645	0.7362622	0.4615744
refereeMarkus Wingenbach	-0.3406800	0.6892186	-0.4942989	0.6210972
refereeMartin Atkinson	0.4978913	0.5343146	0.9318317	0.3514278
refereeMartin Petersen	-0.0562701	0.2039407	-0.2759138	0.7826154
refereeMartin Thomsen	-0.1190299	0.2204123	-0.5400331	0.5891765
refereeMassimiliano Irrati	0.3297107	0.1419061	2.3234420	0.0201592
refereeMathew Buonassisi	-0.5797024	1.6271912	-0.3562595	0.7216476
refereeMathieu Vernice	0.1241018	0.1723914	0.7198839	0.4715996
refereeMatt Donohue	0.4181666	0.5407093	0.7733667	0.4393088
refereeMatteo GARIGLIO	0.3713863	0.1833349	2.0257261	0.0427979
refereeMatteo MARCENARO	0.2605611	0.2440476	1.0676653	0.2856764
refereeMatteo MARCHETTI	-0.1441436	0.1798941	-0.8012690	0.4229795
refereeMatthias Jöllenbeck	-0.1826724	0.2499560	-0.7308184	0.4648933
refereeMaurizio Mariani	0.1703605	0.1351256	1.2607563	0.2074022
refereeMedie Jimenez	0.6969055	0.2507641	2.7791278	0.0054524
refereeMehdi Mokhtari	0.3976383	0.1615375	2.4615847	0.0138356
refereeMichael Bacher	0.0293143	0.2949069	0.0994019	0.9208195
refereeMichael Bull	0.6899022	0.7528088	0.9164375	0.3594417
refereeMichael Fabbri	0.2266785	0.1372832	1.6511736	0.0987091
refereeMichael Oliver	0.6296023	0.5342376	1.1785061	0.2386002
refereeMichael Salisbury	0.2511361	0.5489751	0.4574636	0.6473398
refereeMichael Weiner	-0.0957380	0.3566644	-0.2684261	0.7883724
refereeMiguel Angel ORTIZ ARIAS	0.2490122	0.2414571	1.0312895	0.3024098
refereeMikael Lesage	0.3607669	0.1719226	2.0984262	0.0358722
refereeMike Dean	1.0440398	0.5345032	1.9532902	0.0507905
refereeMike Jones	0.8246287	0.5466854	1.5084155	0.1314542
refereeMike Russell	-0.1390363	0.7666457	-0.1813566	0.8560883
refereeMissing-England-2	0.9115964	0.5621520	1.6216189	0.1048909
refereeMissing-France-1	0.0243126	0.6036899	0.0402734	0.9678753
refereeMissing-Germany-1	-0.2417987	0.1999513	-1.2092880	0.2265576
refereeMissing-Germany-2	0.2697778	0.1930379	1.3975385	0.1622576
refereeMissing-Italy-1	-0.8409562	0.4222381	-1.9916633	0.0464131
refereeMissing-Spain-1	-0.2894962	0.6900515	-0.4195284	0.6748317
refereeMissing-Spain-2	-0.1659914	0.3067512	-0.5411271	0.5884222
refereeMORENO ARAGON	0.5937202	0.2375853	2.4989771	0.0124582
refereeNeil Swarbrick	0.5755289	0.5473439	1.0514941	0.2930365
refereeNiccolo' Baroni	0.0347946	0.1647750	0.2111646	0.8327596
refereeNicola Rizzoli	0.7304164	0.2386851	3.0601679	0.0022132
refereeNicolas Rainville	0.5523949	0.1627088	3.3949902	0.0006868
refereeNigel Miller	0.1987868	0.6380633	0.3115471	0.7553860
refereeNorbert Grudzinski	0.0990596	0.6715038	0.1475191	0.8827229
refereeOcon Arraiz	0.2693625	0.2478760	1.0866825	0.2771821
refereeOliver Langford	0.8012316	0.5329100	1.5035026	0.1327155
refereeOlivier Husset	0.5443625	0.4674619	1.1645066	0.2442240

refereeOlivier Thual	0.0566762	0.1661040	0.3412089	0.7329477
refereePablo Gonzalez Fuertes	0.7772901	0.2496229	3.1138578	0.0018476
refereePaolo Saia	0.7037181	0.4953153	1.4207479	0.1553960
refereePaolo Silvio Mazzoleni	-0.0076002	0.1743640	-0.0435881	0.9652329
refereePaolo Taliavento	0.2657443	0.2052667	1.2946295	0.1954537
refereePaolo Valeri	0.3618968	0.1408861	2.5687195	0.0102102
refereeParide TREMOLADA	-1.7230967	0.7699300	-2.2379916	0.0252258
refereePascal Müller	-0.4319027	0.3485379	-1.2391842	0.2152828
refereePatrick Alt	0.0566253	0.2213673	0.2557980	0.7981078
refereePatrick Ittrich	0.0736200	0.2053999	0.3584225	0.7200286
refereePatrick Miller	0.0794290	1.2027378	0.0660402	0.9473461
refereePatrick Schult	-0.8792304	1.1384128	-0.7723300	0.4399224
refereePaul Tierney	0.7526315	0.5337122	1.4101824	0.1584917
refereePedro Jesus Perez Montero	0.4862530	0.3110522	1.5632520	0.1179993
refereePedro Sureda	0.7517167	0.4754173	1.5811724	0.1138446
refereePérez Lasa	0.4856132	1.9653217	0.2470910	0.8048388
refereePeter Bankes	0.9038911	0.5332274	1.6951326	0.0900562
refereePeter Gagelmann	-0.1552901	0.4341261	-0.3577073	0.7205638
refereePeter Sippel	-0.1834303	0.3218219	-0.5699745	0.5686974
refereePhil Dowd	1.1697739	0.6628947	1.7646452	0.0776292
refereePhil Gibbs	0.4315610	0.7891051	0.5468993	0.5844502
refereePhilippe Kalt	-0.2848148	0.4565364	-0.6238601	0.5327222
refereePiero Giacomelli	0.0199847	0.1389536	0.1438226	0.8856411
refereePierre Gaillouste	-0.0039916	0.1571101	-0.0254061	0.9797311
refereePierre Legat	0.3427090	0.2427324	1.4118801	0.1579911
refereePULIDO SANTANA	0.2136110	0.2400885	0.8897177	0.3736216
refereeQUINTERO GONZALEZ	1.5296869	0.3504873	4.3644580	0.0000128
refereeRAFAEL Sanchez Lopez	0.6483624	0.2571760	2.5210846	0.0117023
refereeRamon Hernandez	1.2892691	1.9692664	0.6546951	0.5126669
refereeRAUL MARTINEZ FRANCES	-0.2382773	0.3002715	-0.7935395	0.4274671
refereeRemy Landry	0.3324215	0.1614320	2.0592047	0.0394795
refereeRene Rohde	0.0363976	0.2415489	0.1506841	0.8802255
refereeRicardo de Burgos Bengoetxea	0.5643554	0.2497463	2.2597150	0.0238430
refereeRiccardo Pinzani	-0.1610711	0.4710016	-0.3419757	0.7323706
refereeRiccardo ROS	0.1837950	0.1504587	1.2215647	0.2218777
refereeRichard Clark	0.6126593	0.8641263	0.7089928	0.4783321
refereeRobert Atkin	0.6326712	1.7326045	0.3651562	0.7149963
refereeRobert Hartmann	-0.1752926	0.1995235	-0.8785561	0.3796460
refereeRobert Jones	0.4220227	0.5345878	0.7894356	0.4298610
refereeRobert Kampka	0.2470831	0.2080959	1.1873518	0.2350942
refereeRobert Kempter	0.6462085	0.2433191	2.6558071	0.0079143
refereeRobert Lewis	-0.0551759	0.7521127	-0.0733613	0.9415189
refereeRobert Madley	0.6565685	0.5457368	1.2030863	0.2289483
refereeRobert Schröder	-0.0057066	0.2305677	-0.0247502	0.9802543
refereeRoger East	0.7361453	0.5423510	1.3573227	0.1746845
refereeRomain Delpech	0.2191203	0.1751702	1.2508994	0.2109767
refereeRomain Lissorgue	0.1490788	0.1589753	0.9377481	0.3483782
refereeRosario Abisso	0.3808366	0.1371135	2.7775276	0.0054793
refereeRoss Joyce	0.1789612	0.8266379	0.2164928	0.8286044
refereeRUBEN AVALOS	0.3848950	0.2388186	1.6116628	0.1070413
refereeRUBEN EIRIZ MATA	0.4189701	0.2756096	1.5201577	0.1284773

refereeRuben RUIPEREZ MARIN	0.7915581	0.4365817	1.8130814	0.0698249
refereeRuddy Buquet	0.3928990	0.1722392	2.2811241	0.0225451
refereeSaid Ennjimi	-0.0224762	0.2899965	-0.0775051	0.9382220
refereeSandro Schärer	-0.7602964	1.0507080	-0.7236039	0.4693121
refereeSantiago VARON ACEITON	0.0114467	0.2497700	0.0458289	0.9634468
refereeSascha Stegemann	-0.0202933	0.1980476	-0.1024669	0.9183865
refereeSaul AIS REIG	0.0252312	0.2435274	0.1036074	0.9174813
refereeScott Duncan	0.5435879	0.5388419	1.0088079	0.3130714
refereeScott Mathieson	-0.8131426	2.0008769	-0.4063931	0.6844554
refereeSebastian Stockbridge	0.4708835	1.1872858	0.3966050	0.6916604
refereeSebastiano Peruzzo	0.4542606	0.4916144	0.9240181	0.3554811
refereeSebastien Desiage	0.1695325	0.2201529	0.7700669	0.4412637
refereeSebastien Moreira	0.1104280	0.2299763	0.4801714	0.6311075
refereeSimon Hooper	0.5554886	0.5334879	1.0412393	0.2977692
refereeSimone SOZZA	0.0041964	0.1481068	0.0283335	0.9773963
refereeSören Storks	-0.0213065	0.2310024	-0.0922350	0.9265117
refereeStephan Klossner	-0.2846226	1.0292787	-0.2765263	0.7821450
refereeStephane Jochem	0.0306745	0.2097952	0.1462118	0.8837547
refereeStephane Lannoy	0.2834238	0.2588308	1.0950156	0.2735148
refereeStephanie Frappart	0.3800397	0.1642358	2.3139882	0.0206722
refereeStephen Martin	0.2682052	0.5329838	0.5032146	0.6148155
refereeStuart Attwell	0.8236623	0.5339281	1.5426465	0.1229226
refereeSven Jablonski	-0.3470730	0.2039175	-1.7020268	0.0887562
refereeSven Waschitzki	-0.1222508	0.2613585	-0.4677514	0.6399643
refereeSylvain Palhies	0.5291976	0.1665649	3.1771250	0.0014883
refereeT J Robinson	0.6716395	0.5325271	1.2612306	0.2072313
refereeThierry Bouille	-0.0921571	0.2298522	-0.4009406	0.6884655
refereeThomas Bramall	-0.2458855	0.5688261	-0.4322683	0.6655482
refereeThomas Leonard	0.4294876	0.1669532	2.5725034	0.0100993
refereeThorben Siewer	-0.4014480	0.2234161	-1.7968627	0.0723631
refereeThorsten Kinhöfer	0.1714250	0.4324311	0.3964216	0.6917956
refereeThorsten Schriever	0.2260752	0.4614212	0.4899541	0.6241684
refereeTimo Gerach	-0.0318736	0.2196836	-0.1450888	0.8846413
refereeTobias Reichel	-0.1793555	0.2638631	-0.6797292	0.4966789
refereeTobias Stieler	-0.2542776	0.1994438	-1.2749336	0.2023384
refereeTobias Welz	0.0360237	0.2165410	0.1663599	0.8678743
refereeTony Chapron	0.2711001	0.2201567	1.2313962	0.2181802
refereeTony Harrington	0.4885758	0.5329584	0.9167241	0.3592914
refereeValentin PIZARRO GOMEZ	0.3416687	0.2477347	1.3791718	0.1678476
refereeVicente José Lizondo Cortés	0.1519666	0.5494447	0.2765822	0.7821020
refereeWilfried Bien	0.3436835	0.2915637	1.1787596	0.2384992
refereeWilliam Lavis	0.1598863	0.1909880	0.8371538	0.4025099
refereeWilly Delajod	0.2801679	0.1610780	1.7393299	0.0819826
refereeYohann Rouinsard	0.0433370	0.1971466	0.2198215	0.8260111
season	0.0303941	0.0126891	2.3952935	0.0166105
competition_level2	0.1063472	0.0296084	3.5917857	0.0003287
attendance	-0.0390948	0.0097769	-3.9986950	0.0000638
stadium_dist	-0.0000475	0.0000310	-1.5326575	0.1253662
stadium_capacity	-0.0043592	0.0064512	-0.6757203	0.4992212
stadium_runningtrack	-0.0927377	0.0626762	-1.4796325	0.1389773
stadium_crowd_density	0.3603652	0.0490498	7.3469301	0.0000000

matchweek	-0.0038493	0.0009814	-3.9221732	0.0000879
yellow_card_lag1	0.0132833	0.0042446	3.1294585	0.0017522
covid_impacted	-0.1190957	0.0302094	-3.9423362	0.0000808
referee_experience	-0.0008756	0.0002885	-3.0346707	0.0024092
total_goals_implied_scaled	-0.1067942	0.0089639	-11.9138469	0.0000000
supremacy_implied_scaled	-0.1212017	0.0129390	-9.3671536	0.0000000
poly(matchweek, 2)2	-5.7363209	2.1706234	-2.6427067	0.0082270
home:attendance	0.0051854	0.0088114	0.5884899	0.5562060
home:covid_impacted	0.2619232	0.0287606	9.1070076	0.0000000
countryFrance:poly(matchweek, 2)1	-3.0514519	4.0537311	-0.7527514	0.4516026
countryGermany:poly(matchweek, 2)1	0.3575132	4.9698608	0.0719363	0.9426529
countryItaly:poly(matchweek, 2)1	-7.8307172	4.0481064	-1.9344148	0.0530675
countrySpain:poly(matchweek, 2)1	9.7463941	3.6127130	2.6978047	0.0069821
countryFrance:poly(matchweek, 2)2	-4.8823316	4.0709677	-1.1993049	0.2304148
countryGermany:poly(matchweek, 2)2	0.3571850	5.0785884	0.0703316	0.9439300
countryItaly:poly(matchweek, 2)2	-13.4797082	4.1659366	-3.2356969	0.0012142
countrySpain:poly(matchweek, 2)2	-7.4860653	3.4928151	-2.1432756	0.0320955
countryFrance:stadium_runningtrack	0.1078729	0.0961604	1.1218025	0.2619515
countryGermany:stadium_runningtrack	0.2321486	0.1014798	2.2876341	0.0221628
countryItaly:stadium_runningtrack	0.1995015	0.0808687	2.4669808	0.0136289

	Model 1
(Intercept)	21.72
	(22.05)
caseweights_by_recency	0.02
	(0.07)
season	-0.01
	(0.01)
competition_level2	0.13^{*}
	(0.05)
attendance	-0.05^{***}
	(0.01)
stadium_crowd_density	0.50^{***}
	(0.06)
matchweek	-0.01^{***}
	(0.00)
poly(matchweek, 2)2	-6.04^{**}
	(2.17)
yellow_card_lag1	0.02^{***}
	(0.00)
covid_impacted	-0.05
	(0.03)
total_goals_implied_scaled	-0.10^{***}
	(0.01)
supremacy_implied_scaled	-0.12^{***}
	(0.01)
home	-0.11^{**}
	(0.04)
countryFrance	0.17^{*}
	(0.08)
countryGermany	0.37^{***}

	Model 1
	(0.08)
countryItaly	0.79^{***}
	(0.08)
countrySpain	0.82^{***}
	(0.08)
stadium__runningtrack	-0.02
	(0.06)
attendance:home	0.04^{**}
	(0.01)
covid__impacted:home	0.16^{***}
	(0.04)
stadium__crowd__density:home	-0.27^{***}
	(0.08)
poly(matchweek, 2)1:countryFrance	-3.53
	(4.03)
poly(matchweek, 2)2:countryFrance	-4.92
	(4.06)
poly(matchweek, 2)1:countryGermany	-0.17
	(4.95)
poly(matchweek, 2)2:countryGermany	0.17
	(5.07)
poly(matchweek, 2)1:countryItaly	-6.98
	(4.02)
poly(matchweek, 2)2:countryItaly	-14.32^{***}
	(4.14)
poly(matchweek, 2)1:countrySpain	11.88^{***}
	(3.59)
poly(matchweek, 2)2:countrySpain	-7.41^{*}
	(3.49)
countryFrance:stadium__runningtrack	0.04
	(0.09)
countryGermany:stadium__runningtrack	0.13
	(0.09)
countryItaly:stadium__runningtrack	0.12
	(0.07)
competition__level2:countryFrance	-0.03
	(0.07)
competition__level2:countryGermany	-0.03
	(0.07)
competition__level2:countryItaly	-0.10
	(0.07)
competition__level2:countrySpain	-0.15^{*}
	(0.07)
AIC	192976.54
BIC	193332.29
Log Likelihood	-96448.27
Num. obs.	53846
Num. groups: referee	394
Num. groups: team	271
Num. groups: opponent	271
Var: referee (Intercept)	0.04
Var: team (Intercept)	0.03
Var: opponent (Intercept)	0.03

	Model 1
Var: Residual	0.54

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 23: Output for Linear mixed model with entire dataset

	Model 1
(Intercept)	8.21
	(19.78)
caseweights_by_recency	0.00
	(0.07)
season	-0.00
	(0.01)
competition_level2	0.03
	(0.03)
attendance	-0.02^{*}
	(0.01)
stadium_crowd_density	0.24^{***}
	(0.05)
matchweek	-0.00^{***}
	(0.00)
poly(matchweek, 2)2	-4.32
	(2.33)
yellow_card_lag1	0.01^{***}
	(0.00)
covid_impacted	-0.01
	(0.03)
total_goals_implied_scaled	-0.05^{***}
	(0.01)
supremacy_implied_scaled	-0.06^{***}
	(0.01)
home	-0.03
	(0.03)
countryFrance	0.11^{**}
	(0.04)
countryGermany	0.16^{***}
	(0.04)
countryItaly	0.37^{***}
	(0.04)
countrySpain	0.38^{***}
	(0.04)
stadium_runningtrack	0.04
	(0.04)
attendance:home	0.02
	(0.01)
covid_impacted:home	0.05
	(0.04)
stadium_crowd_density:home	-0.17^{*}
	(0.07)
poly(matchweek, 2)1:countryFrance	-1.37

	Model 1
	(4.18)
poly(matchweek, 2)2:countryFrance	−2.27
	(4.20)
poly(matchweek, 2)1:countryGermany	1.08
	(5.03)
poly(matchweek, 2)2:countryGermany	0.42
	(5.09)
poly(matchweek, 2)1:countryItaly	0.10
	(3.89)
poly(matchweek, 2)2:countryItaly	−4.40
	(3.95)
poly(matchweek, 2)1:countrySpain	9.03**
	(3.46)
poly(matchweek, 2)2:countrySpain	−0.91
	(3.35)
countryFrance:stadium_runningtrack	−0.01
	(0.07)
countryGermany:stadium_runningtrack	0.00
	(0.07)
countryItaly:stadium_runningtrack	−0.01
	(0.05)
competition_level2:countryFrance	−0.03
	(0.04)
competition_level2:countryGermany	0.03
	(0.05)
competition_level2:countryItaly	0.00
	(0.04)
competition_level2:countrySpain	−0.04
	(0.04)
AIC	61487.81
BIC	61834.67
Log Likelihood	−30704.90
Num. obs.	53846
Num. groups: referee	394
Num. groups: team	271
Num. groups: opponent	271
Var: referee (Intercept)	0.00
Var: team (Intercept)	0.00
Var: opponent (Intercept)	0.00

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

Table 24: Output for Generalised linear mixed model with entire dataset

Agnew, G. A., and A. V. Carron. 1994. "Crowd Effects and the Home Advantage." *International Journal of Sport Psychology* 25 (1): 53–62.

Audrino, Francesco. 2020. "Strongest Team Favoritism in European National Football: Myth or Reality?" *International Journal of Sport Finance* 15 (July): 137–52.

Badiella, L., P. Puig, C. Peñas, and M. Casals. 2022. "Influence of Red and Yellow Cards on Team Performance in Elite Soccer." *Annals of Operations Research* 325. https://doi.org/10.1007/s10479-022-04733-0.

Baio, Gianluca, and Marta Blangiardo. 2010. "Bayesian Hierarchical Model for the Prediction of Football Results." *Journal of Applied Statistics* 37: 253–64. https://doi.org/10.1080/02664760802684177.

Barrett, T. S. 2019. "R for Researchers: An Introduction." May 13, 2019. https://tysonbarrett.com/Rstats/.

BBC. 2018. "Joey Barton Claims 50." January 23, 2018. https://www.bbc.co.uk/sport/football/42783527#:~:text=He%20was%20found%20to%20have,out%20because%20it\XeTeXglyph\numexpr\XeTeXcharglyph"0027\relax{}s%20culturally%20engrained.%22.

Boehmke, Bradley, and Brandon Greenwell. 2020. "Hands-on Machine Learning with r." Bradley Boehmke's GitHub. February 2020. https://bradleyboehmke.github.io/HOML/.

Bolker, B. M. 2015. "Linear and Generalised Linear Mixed Models (Chapter 13)." In *ECOLOGICAL STATISTICS: CONTEMPORARY THEORY AND APPLICATION*, edited by G. A. Fox, S. Negrete-Yankelevich, and V. J. Sosa, xxx–. Oxford Academic. https://doi.org/10.1093/acprof:oso/9780199672547.001.0001.

Bose*, Paul, Eberhard Feess, and Helge Mueller. 2021. "Favoritism towards High-Status Clubs: Evidence from German Soccer." *The Journal of Law, Economics, and Organization* 38 (2): 422–78. https://doi.org/10.1093/jleo/ewab005.

Boyko, Ryan H., Adam R. Boyko, and Mark G. Boyko. 2007. "Referee Bias Contributes to Home Advantage in English Premiership Football." *Journal of Sports Sciences* 25 (11): 1185–94. https://doi.org/10.1080/02640410601038576.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45: 5–32. https://doi.org/10.1023/A:1010933404324.

Breiman, Leo, Jerome Friedman, Richard Olshen, and Charles Stone. 1984. "Classification and Regression Trees. Wadsworth Int." *Group* 37 (15): 237–51.

Bryson, Alex, Peter Dolton, J. James Reade, Dominik Schreyer, and Carl Singleton. 2021. "Causal Effects of an Absent Crowd on Performances and Refereeing Decisions During Covid-19." *Economics Letters* 198: 109664. https://doi.org/https://doi.org/10.1016/j.econlet.2020.109664.

Buraimo, Babatunde, David Forrest, and Robert Simmons. 2010. "The 12th Man?: Refereeing Bias in English and German Soccer." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 173 (2): 431–49. https://doi.org/https://doi.org/10.1111/j.1467-985X.2009.00604.x.

Chandler, Richard E. 2023. "Lecture Notes for STAT0017 Part 1: Week 1: Introduction and Motivation."

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." *CoRR* abs/1603.02754. http://arxiv.org/abs/1603.02754.

contributors, Wikipedia. 2023. "Laws of the Game (Association Football)." 2023. https://en.wikipedia.org/w/index.php?title=Laws_of_the_Game_(association_football)&oldid=1163999762.

"Creating Case Weights Based on Time." n.d. Tidymodels. https://www.tidymodels.org/learn/work/case-weights.

Dixon, M. J., and S. G. Coles. 1997. "Modelling Association Football Scores and Inefficiencies in the Football Betting Market." *Appl. Statist.* 46: 265–80.

Eryarsoy, E., and D. Delen. 2019. "Predicting the Outcome of a Football Game: A Comparative Analysis of Single and Ensemble Analytics Methods." In *Hawaii International Conference on System Sciences*.

Freund, Yoav, and Robert E Schapire. 1999. "Adaptive Game Playing Using Multiplicative Weights." *Games and Economic Behavior* 29 (1-2): 79–103.

Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics*, 1189–1232.

Greenwell, Brandon, Bradley Boehmke, Jay Cunningham, and GBM Developers. 2022. *Gbm: Generalized Boosted Regression Models*. https://CRAN.R-project.org/package=gbm.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York. https://doi.org/10.1007/978-0-387-84858-7.

Humphreys, Brad, J. James Reade, Dominik Schreyer, and Carl Singleton. 2022. "Separating the crowds: Examining home and away attendances at football matches." Economics Discussion Papers em-dp2022-11. Department of Economics, University of Reading. https://ideas.repec.org/p/rdg/emxxdp/em-dp2022-11.html.

Johnson, Rie, and Tong Zhang. 2014. "Learning Nonlinear Functions Using Regularized Greedy Forest." https://arxiv.org/abs/1109.0887.

Karlis, Dimitris, and Ioannis Ntzoufras. 2003. "Analysis of Sports Data by Using Bivariate Poisson Models."

Journal of the Royal Statistical Society: Series D (The Statistician) 52 (3): 381–93. https://doi.org/https://doi.org/10.1111/1467-9884.00366.

Kossen, Jannik, Neil Band, Clare Lyle, Aidan N Gomez, Thomas Rainforth, and Yarin Gal. 2021. "Self-Attention Between Datapoints: Going Beyond Individual Input-Output Pairs in Deep Learning." In *Advances in Neural Information Processing Systems*, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan, 34:28742–56. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2021/file/f1507aba9fc82ffa7cc7373c58f8a613-Paper.pdf.

Laird, Nan M., and James H. Ware. 1982. "Random-Effects Models for Longitudinal Data." *Biometrics* 38 (4): 963–74. http://www.jstor.org/stable/2529876.

LIANG, KUNG-YEE, and SCOTT L. ZEGER. 1986. "Longitudinal data analysis using generalized linear models." *Biometrika* 73 (1): 13–22. https://doi.org/10.1093/biomet/73.1.13.

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22. https://CRAN.R-project.org/doc/Rnews/.

Maher, M. J. 1982. "Modelling Association Football Scores." *Statistica Neerlandica* 36 (3): 109–18. https://doi.org/https://doi.org/10.1111/j.1467-9574.1982.tb00782.x.

McCarrick, D., G. Brewer, M. Lyons, et al. 2020. "Referee Height Influences Decision Making in British Football Leagues." *BMC Psychology* 8: 4. https://doi.org/10.1186/s40359-020-0370-4.

Nevill, A. M, N. J Balmer, and A Mark Williams. 2002. "The Influence of Crowd Noise and Experience Upon Refereeing Decisions in Football." *Psychology of Sport and Exercise* 3 (4): 261–72. https://doi.org/https://doi.org/10.1016/S1469-0292(01)00033-4.

Nevill, Alan, Alastair Pearson, and Tom Webb. 2022. "No Crowds, No Home Advantage in Football During the COVID-19 Season: Are Crowds Able to Manipulate All but the Best Referees' Behaviour?" *Journal of Global Sport Management* 0 (0): 1–18. https://doi.org/10.1080/24704067.2022.2136102.

O'Brien, K. A., and J. Mangan. 2021. "The Issue of Unconscious Bias in Referee Decisions in the National Rugby League." *Frontiers in Sports and Active Living* 3: 739570. https://doi.org/10.3389/fspor.2021.739570.

Pekár, Stano, and Marek Brabec. 2018. "Generalized Estimating Equations: A Pragmatic and Flexible Approach to the Marginal GLM Modelling of Correlated Data in the Behavioural Sciences." *Ethology* 124 (2): 86–93. https://doi.org/https://doi.org/10.1111/eth.12713.

Pettersson-Lidbom, Per, and Mikael Priks. 2010. "Behavior Under Social Pressure: Empty Italian Stadiums and Referee Bias." *Economics Letters* 108 (2): 212–14. https://doi.org/https://doi.org/10.1016/j.econlet.2010.04.023.

Picazo-Tadeo, Andrés, Francisco Gónzalez-Gómez, and Jorge Guardiola Wanden-Berghe. 2011. "Referee home bias due to social pressure. Evidence from Spanish football." Working Papers 1119. Department of Applied Economics II, Universidad de Valencia. https://ideas.repec.org/p/eec/wpaper/1119.html.

Probst, Philipp, Bernd Bischl, and Anne-Laure Boulesteix. 2018. "Tunability: Importance of Hyperparameters of Machine Learning Algorithms."

Review, World Population. 2023. "Most Popular Sport by Country 2023." 2023. https://www.worldpopulationreview.com/.

Sam. 2018. "Plot Feature Importance Computed by Ranger Function." Stack Overflow. March 5, 2018. https://stackoverflow.com/questions/49105358/plot-feature-importance-computed-by-ranger-function.

Schwartz, Barry, and Stephen F. Barsky. 1977. "The Home Advantage." *Social Forces* 55 (3): 641–61. http://www.jstor.org/stable/2577461.

Sigrist, F. 2020. "Gaussian Process Boosting." https://arxiv.org/abs/2004.02653.

———. 2023. "A Comparison of Machine Learning Methods for Data with High-Cardinality Categorical Variables." https://arxiv.org/abs/2307.02071.

Sigrist, Fabio. 2023. "Mixed Effects Machine Learning for High-Cardinality Categorical Variables — Part II: GPBoost Library: A Demo of GPBoost in Python & r Using Real-World Data." July 11, 2023. https://towardsdatascience.com/mixed-effects-machine-learning-for-high-cardinality-categorical-variables-part-ii-gpboost-3bdd9ef74492.

"University College London STAT0030 Statistical Computing, Lab 7." 2022-2023.

Unkelbach, C., and D. Memmert. 2010. "Crowd Noise as a Cue in Referee Decisions Contributes to the Home Advantage." *Journal of Sport & Exercise Psychology* 32 (4): 483–98. https://doi.org/10.1123/jsep.32.4.483.

Wright, Marvin N., and Andreas Ziegler. 2017. "Ranger: A Fast Implementation of Random Forests for High Dimensional Data in c++ and r." *Journal of Statistical Software* 77 (1): 1–17. https://doi.org/10.18637/jss.v077.i01.

Zhang, H., Q. Yu, C. Feng, D. Gunzler, P. Wu, and X. M. Tu. 2012. "A New Look at the Difference Between the GEE and the GLMM When Modeling Longitudinal Count Responses." *Journal of Applied Statistics* 39 (9): 2067–79. https://doi.org/10.1080/02664763.2012.700452.