EPL Exploratory Analysis & Model Creation

Juan G. & Andrew H.

2024-12-03

Abstract

The English Premier League has seen a noticeable rise in US viewership over recent years - contributing to the already wide scale popularity of the sport. No matter the sport, refereeing seems to be a major speaking point across fan bases whether that be the call inconsistencies, missed calls, or speculative fouls. The purpose of this report and analyses is to understand whether there is an inconsistency in refereeing and whether or not it has a significant impact on the result of any given match - We also decided to add a second part to our project, where we wanted to look at the predictive power of our data when looking at the match outcome compared to what the expected outcome was.

Methodology

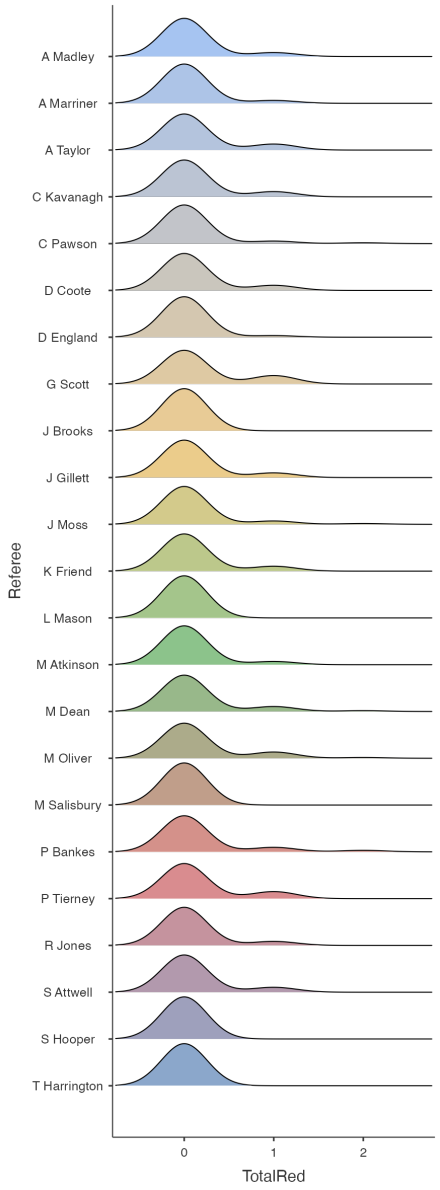
The model creation process in this report started the final_dataset.csv file dowloaded from Kaggle.com with variables regarding Home/Away team, total booking points, Away/Home Cards, Referee, etc. Using this data we generated some variables seen on the EPL_New.csv some of these are Predicted Outcome Success (POS), Expected Result (ExpectedR), Expected Result Odds (ERO), Home/Away/Total Booking Points (HBP,ABP,TBP). Other variables that were taken into consideration were Bet365 Closing Odds (B365CH & B365CA).

POS and ExpectedR were both treated as binary variables. Although in english football matches can result in a draw - bookies release closing odds based on whether a team wins or losings. Fitting a logistic regression model with a draw wouldn't make much sense.

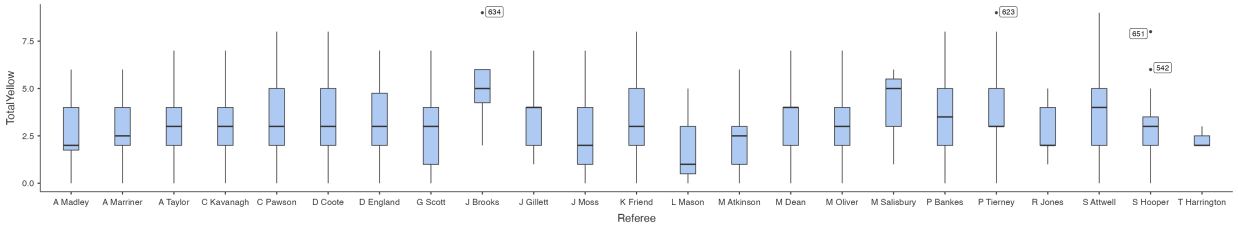
This report contains initial data analysis and exploration, logistic regression, Lasso variable selection and model prediction. The analyses and visualizations done within this report was created with a combination of RStudio, Jamovi, Jupyter Lab.

Discussion

Initially the question we wanted to answer was did booking points - points calculated by yellow/red cards given during the match - significantly impacted a match's outcome. Following this curiosity we wanted to know if there was consistency between the yellow/red cards given throughout a game and the referee's issuing them.



Since red cards issuance is rare in the sport it's to no surprise that the density plots show that most referees don't issue red cards much - this was expected given what we know. It is worth noting that some referee's density at one red card given stood out though, for example Graham Scott and Paul Tierney.



This photo depicts the differences between referees and their yellow card issuance rate. We noticed that yellow card issuance (on a game-to-game basis) highly varied between referees. This visualization peaked our interest and sparked our initial research question. We figured that there was going to be variability between referees; however, we didn't expect this much variable among yellow card issuance.

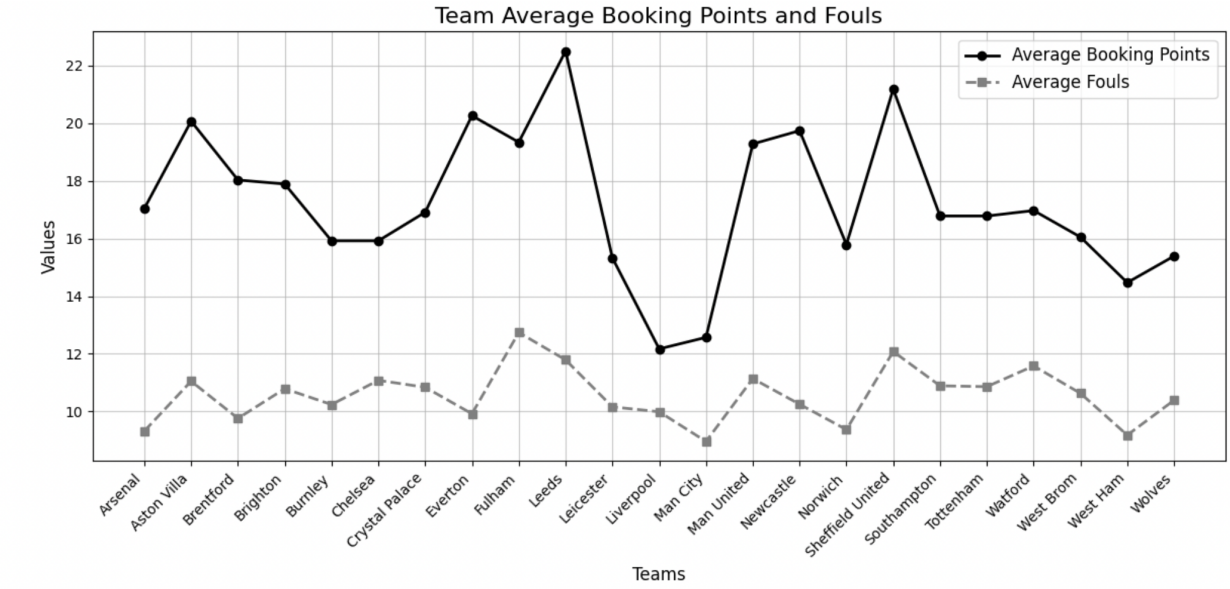
ANOVA

ANOVA - TBP

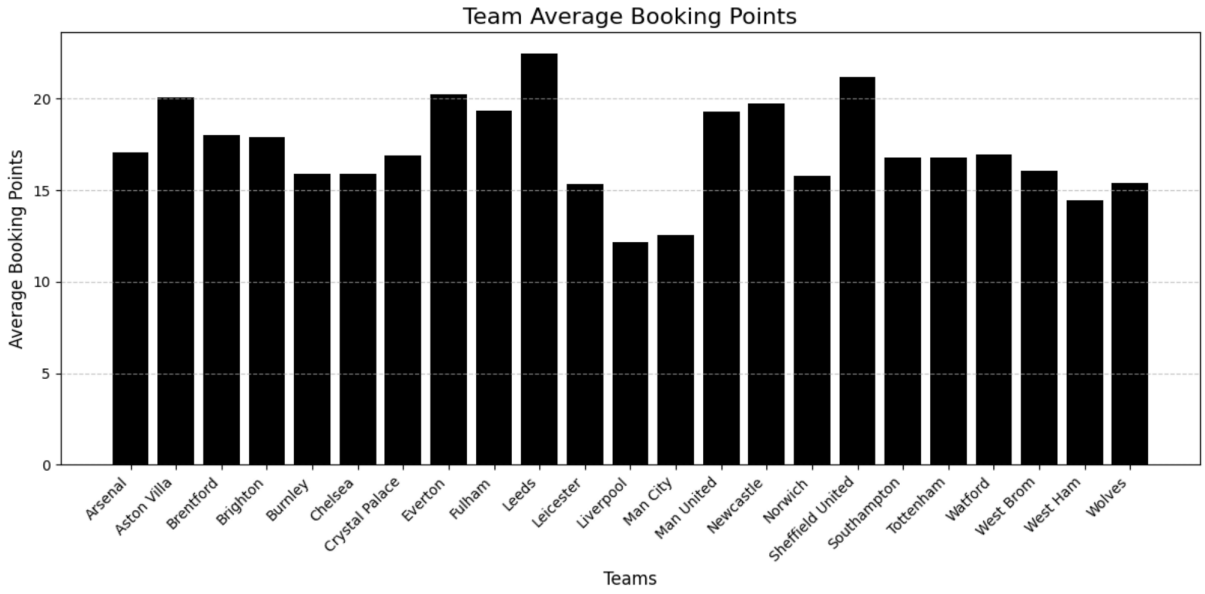
	Sum of Squares	df	Mean Square	F	p
Referee	26255	22	1193	2.96	<.00001
Residuals	297364	737	403		

Just for further confirmation, we then wanted to run an ANOVA model to see if we really had a significant difference in means across referees, this would eliminate any worries of certain referees officiating more or less matches than others. When choosing the variable to run ANOVA on - we realized we wanted to grasp a referees' card impact through one variable, total booking points. This is a computed variable that football-data.co.uk (the original source of the data) mentions on their website; It gives each yellow card per match a value of 10 points, and a red card 25 points. For the situation in which a player received a red card through two yellow cards, the first yellow card was counted as 10 points, while the second was just counted as a red card, 25 points. Back to our ANOVA model, the p-value for referees was < 0.00001. This tells us that referees impact on total booking points per game is statistically significant and can be attributed to individual referee's decision-making.

Continuing our in our exploratory process we wanted to take a closer look at booking points through a couple of models made in Jupyter Lab:



The three graphs shown above depict our thought process as we started exploring. First, we started with a simple bar graph of average booking points per team. We thought this would be useful as we push for a case of favoritism among the league towards specific teams. It's often something fans and even professionals in the sport theorize about as they feel a preference is shown to teams by giving them less cards. The second graph above then paired each team's value for average booking points with average fouls. At this point we began to see if we could find any difference between the two, because if we consider what we're looking for - refereeing consistency - then we would expect to see that for more fouls, more booking points were given. That's where we step into our third graph and see that this is not the case. We switched to line graphs here to enhance the variation in both values across the teams. For this graph we would have expected lines which were more parallel than not, in order to depict consistency in booking points by foul.


Team Average Booking Points
Average Booking Points
0
5
10
15
20
Arsenal
Aston Villa
Brentford
Brighton
Burnley
Chelsea
Crystal Palace
Everton
Fulham
Leeds
Leicester
Liverpool
Man City
Man United
Newcastle
Norwich
Sheffield United
Southampton
Tottenham
Watford
West Brom
West Ham
Wolves
Teams

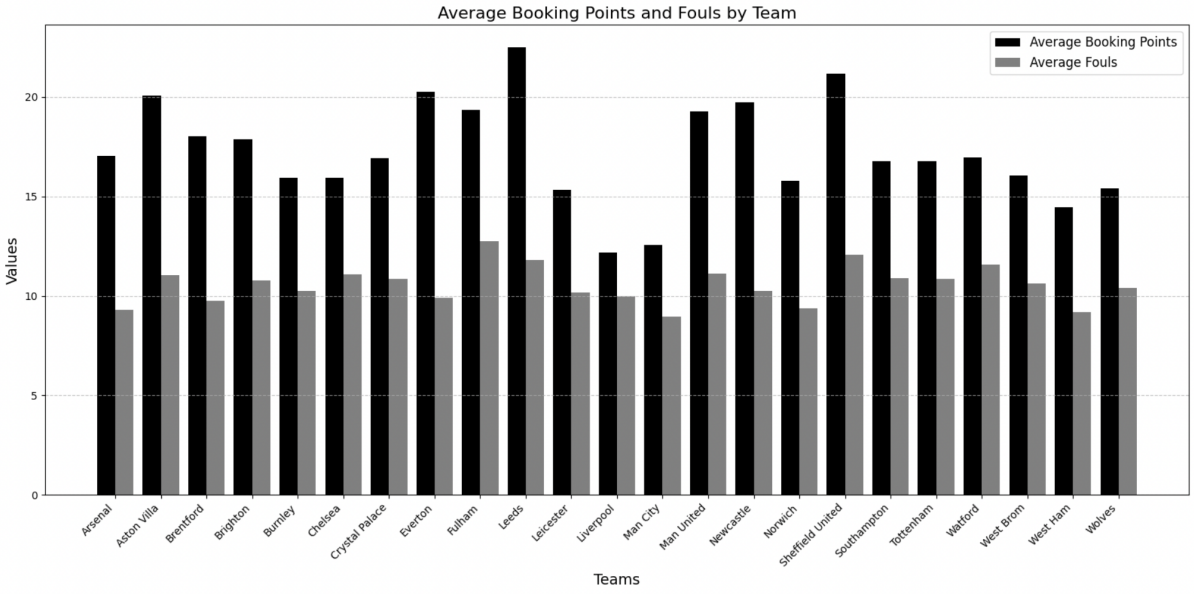
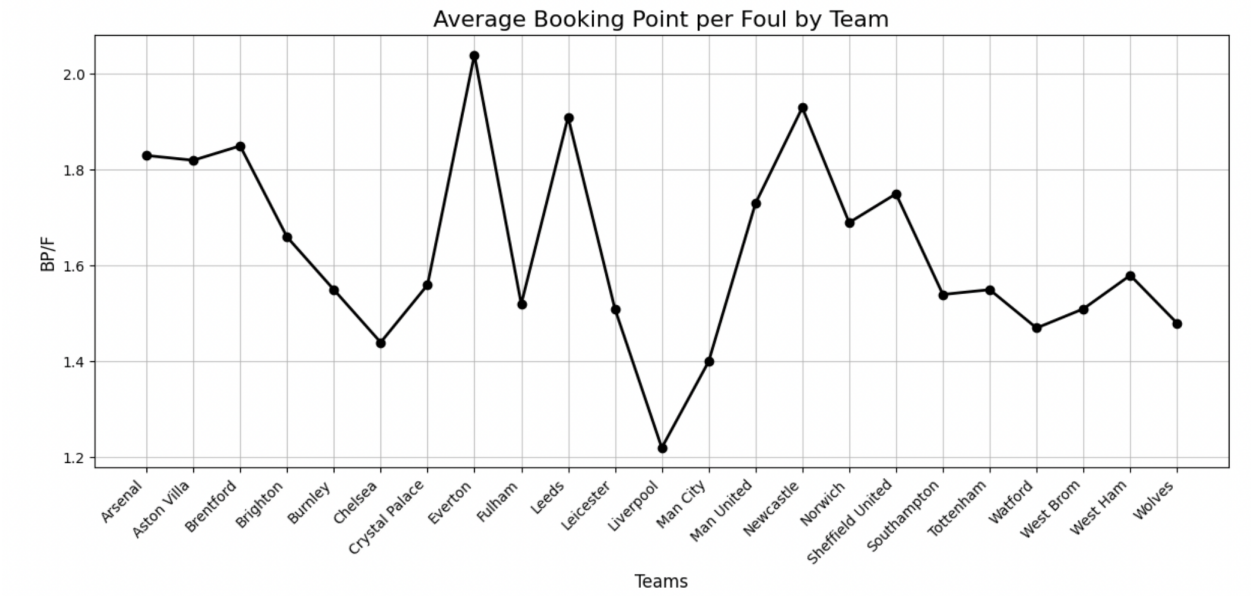

Average Booking Points and Fouls by Team
Average Booking Points
Average Fouls
Values
0
5
10
15
20
Arsenal
Aston Villa
Brentford
Brighton
Burnley
Chelsea
Crystal Palace
Everton
Fulham
Leeds
Leicester
Liverpool
Man City
Man United
Newcastle
Norwich
Sheffield United
Southampton
Tottenham
Watford
West Brom
West Ham
Wolves
Teams

Figure 1:



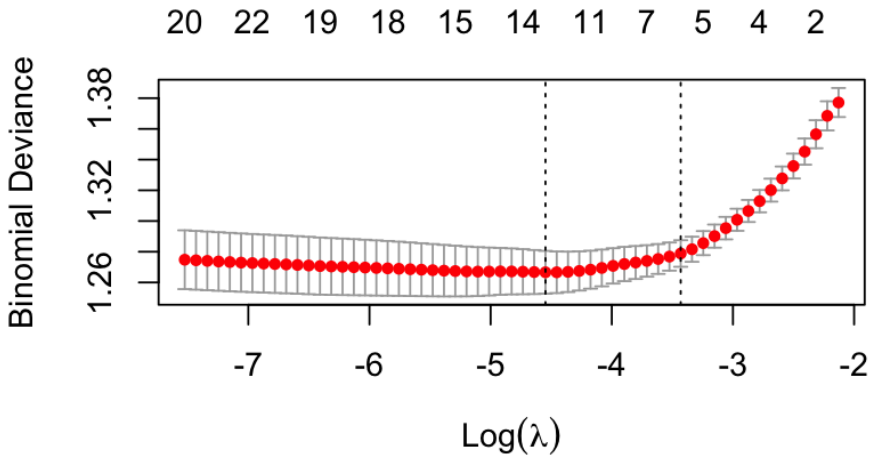
After pulling the values observed in the graphs previously shown, we calculated a new value: booking points per foul. We found this was something quite interesting to both visualize and consider as we think about refereeing consistency. This line graph showed us how much booking points per foul varies by team, with Everton seeing their fouls being more penalized than others, and Liverpool being penalized less by others. To put why this is an issue into perspective - Liverpool saw a yellow card (10 booking points) for about every 10 fouls they committed, while Everton faced the same amount of booking points in half the amount of fouls committed. Once we reached this point, we were satisfied with our exploration in whether or not there is refereeing consistency. We were now moving on to part two of our project; Are there other statistically significant variables in relation to the predicted outcome success (POS). In this next portion we wanted moved away from referees and focused more on the numerical values in our model. This following analyses fit a logistic regression model and performed variable selection to determine which numerical predictors had the held the most significance. The fitted logistic regression model was fit to give us a general understanding of what variables might be significant - although this approach differed greatly from Least Absolute Shrinkage and Selection Operator (LASSO).

The initial model created was a logistic regression model with the Predicted Outcome Success variable being our binary reponse and using all other numeric variables as predictors. The intention was to understand if in this darts at the wall model would yield any statistically significant variables - Full-time Result Home Win ($p < 0.00556$), Away Yellow ($p < 0.01577$), Bet 365 Away Closing Odds ($p < 0.03799$).This generally didn't tell us much about our initial research question outside of that Away Yellow was significant.

After the initial logistic regression model was performed, LASSO was used for model selection. The model was created and first performed on the entire EPL_New.csv data set. LASSO is an extension of OLS that performs shrinkage and variable selection simultaneously. LASSO shrinks standardized coefficients exactly to 0 by applying a L1 penalty - the portion responsible for removing non-significant coefficients from the model. This shrinkage depends on the selected λ - a higher λ results in greater model shrinkage.

Coefficient Name <chr>	Coefficient Value <dbl>
(Intercept)	0.830182483
FTHG	0.376037018
HTAG	0.111224327
HS	-0.012847305
AS	-0.024525779
AF	0.006706909
HC	-0.025502389
ABP	-0.003360767
TBP	-0.009311264
B365D	0.019457954
B365CH	0.082135922
B365CD	0.086090523
ERO	-0.518314470

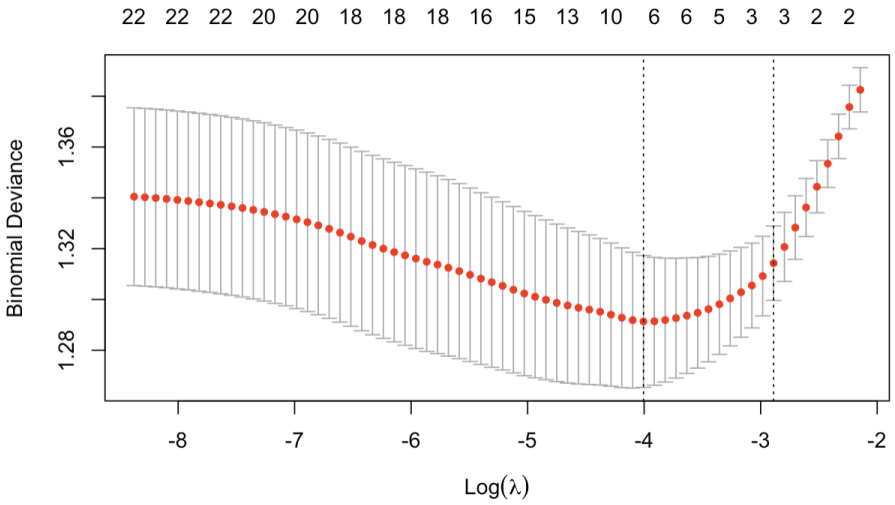
Metric <chr>	Value <dbl>
Accuracy	0.6644737
Predicted 0	292.0000000
Predicted 1	468.0000000



The first lasso created was untrained and predicted over the entire dataset the resulting minimal $\lambda = 0.01059308$ and an accuracy of 66% - marginally better than a coin flip. As a result we created another lasso model, training this time on matches from years 2020 - 2021 and testing it it on matches played throughout 2022. This resulted in better overall prediction accuracy over the 2022 matches at 75% and a $\lambda = 0.01821493$

Coefficient Name <chr>	Coefficient Value <dbl>
(Intercept)	0.830182483
FTHG	0.376037018
HTAG	0.111224327
HS	-0.012847305
AS	-0.024525779
AF	0.006706909
HC	-0.025502389
ABP	-0.003360767
TBP	-0.009311264
B365D	0.019457954

Metric <chr>	Value <dbl>
Accuracy	0.7461929
Predicted 0	75.0000000
Predicted 1	122.0000000



Conclusions & Limitations

The exploration of our question "is there consistency in refereeing in the Premier League?" was one that we enjoyed as we not only explored data we found, but computed our own variables.

From the start, looking at the graphs about cards by referees, we knew we had a case for there not being consistency. This led us to consider all possibilities we could think of for why we were seeing differences in cards by referees.

In our discussion, we both agreed that there were factors that played into card issuance that we did not have information about in our data. Rather than touching on those implications in the body of our project, we felt it was more appropriate to mention them in our conclusion with a focus on limitations. Factors like crowd attendance, match importance, and most importantly, discipline by team were some of the most important ones that came to mind.

We saw this precisely in our ANOVA model, which followed the card graphs. The sum of squares was 26,255

for referees, while the residual sum of squares was 297,364. We found that referees' impact on booking points was indeed significant at a p-value less than 0.00001, but there was still so much more variability to account for.

I mentioned discipline by team as most important to what we discussed because the exploration we did through bar graphs and line graphs after the ANOVA model confirmed what the majority of Premier League fans will tell you about certain teams being dirtier. Newcastle, for example, wasn't up there with the highest average booking points, but they were second highest when we looked at booking points per foul.

Would we like to call it a day and say all this means is that referees are biased towards Newcastle and that's that? Of course. But we must consider that Newcastle could just be a team which plays rougher and, as a result, are penalized more often due to dirtier fouls. The same could be said about Everton, which was first in booking points per foul. Both are teams that are notoriously known by fans of the league for being rough.

This didn't mean what we found was meaningless. We found some worth through these graphs as we found that Liverpool and Manchester City were teams which received the least booking points per foul. City were the champions in both seasons our data covered, while Liverpool were contenders in both seasons and also the winners the season before our data started. There will always be arguments for why these things may be, like they just played clean and this attributed to their success, but we feel it is a clear indication of favoritism by the referees.

We'd be going in circles if we kept describing what was observed and why there exist other possibilities, which is precisely why we felt that the lack of data for some of these known variables in the sport was a huge limitation to our project. It didn't take away from our project, but it is data that we know would've made our research far more insightful.

After training our LASSO model using 2020 - 2021 matches the final model performed better at predicting the 2022 matches; however, I suspect that a few of our coefficients might have been penalized because the Bet365 predictors were highly correlated. LASSO often struggles when predictors are highly correlated because it will arbitrarily select one variable then shrinks the remaining to zero. Another limitation that should be kept in mind when performing LASSO is that the L1 penalty applied can be too aggressive and will lead to variables that still yield some predictive power. Although our final model tested on 2022 yielded better results than the initial LASSO model - from 66% to ~ 75% accuracy - there could have been predictors forced to zero that may have predictive power. This could use some further refinement though for now it's a step in the right direction. Returning to our original question, our model did select the variables Away Fouls (AF), Away Booking Points (ABP), and Total Booking Points (TBP) as predictors that yielded significant predictive power in the final model. Moving forward, I'm curious if AF and ABP had been arbitrarily selected given the aggressiveness of the L1 penalty applied or if this lends into the notion that there is an inherent "home pitch" advantage.