BIOS 611 Project Report

Haiyi Chen

2024-12-07

Background

The home Field advantage is believed to exists in many sports such as soccer, football, and basketball. It refers to the benefit that the home team is said to gain over the visiting team. Some of the factors that results in home field advantage are familiarity with the playing environment, reduced travel fatigue, referee bias, and also the presence of home supporters, whose vocal encouragement and psychological impact can motivate the home team and pressure the opposition. The English Premier League (EPL) is regarded as one of the most popular and competitive soccer leagues around the world. We are interested in verifying if the home field advantage significantly influence the results of matches in. In the meantime, we are also interested in the relationship between the standing and defending/passing/possession/shooting performance of teams. Thus, we are going to use datasets of season 2021-2022 of EPL to investigate the following questions:

1. Is the home team more likely to get a win or draw than the away team?

2. Is the home team less likely to get yellow orred cards than the away team?

3. Is the home team more likely to have more goals than the away team?

4. If we cluster teams with their defending/passing/possession/shooting performance, are teams within a cluster close in the standing?

5. Which of the above performance is most important to predict the standing of a team?

Data sets

We used 2 data sets downloaded from kaggle. The first one contains The results of the total 380 matches as well as some match statistics including goals, shots, shots on target, corners, fouls, yellow cards, red cards. The second data set contains a series of statistics to measure the defending/passing/possession/shooting performance of each team in this season such as team goals, team total shots.

Link of the data sets:

https://www.kaggle.com/datasets/mechatronixs/20212022-season-england-premier-league-team-data?select=england_premier_league_squad_shooting_22.csv

1. Match Results and Home Field

The first data set is used for this question. Figure 1 is a box plot of the match statistics for home teams and way teams

According to this graph, home teams tends to have more shots and less yellow cards than way teams.

Figure 2 is a pie plot of the match results. According to this graph, without adjusting for other variables home teams has a higher chance to win the game.

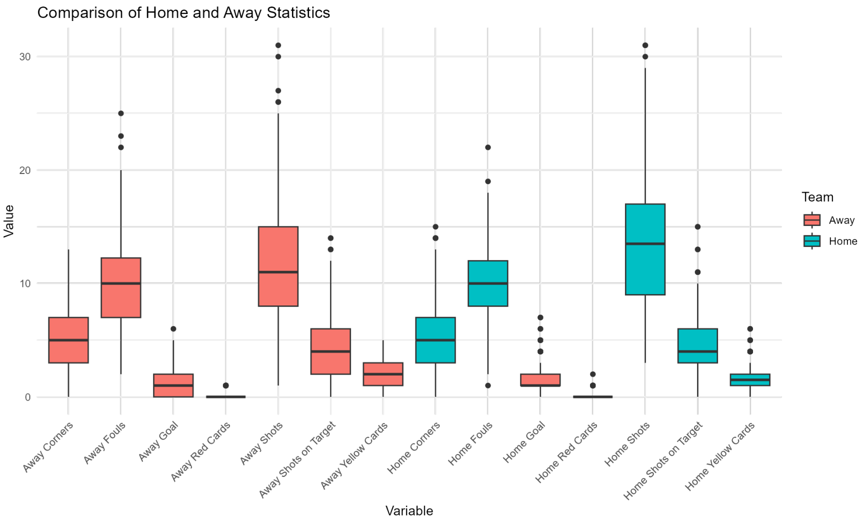

Figure 1: Box Plot of Match Statistics

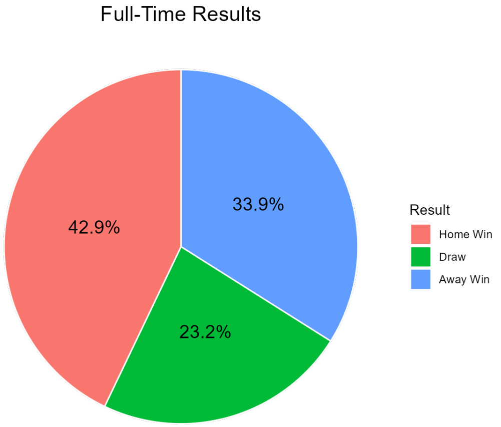

Figure 2: Pie Plot of Match Results

We separated the first data set into 200 data sets by the team. Each new data sets contains 38 games a team played in this season. We fitted logistic regression on each data set. We didn't use the whole data set to fit a single logistic regression because the match results of home teams and away teams are strongly negatively correlated. If the home team wins, the away team can only draw of lose. The response of the logistic regressions is whether the team got win or draw in that match.

The outcome of the logistic regressions are if the team got a win or draw in this match. The covariates includes number of goals, fouls, shots, shots on target, yellow cards and red cards.

Table 1 show the estimated coefficients of home field for each team adjusted p-values. P-values are adjusted since this analysis involves multiple testing.

Table 1: Home Coefficient by Team

term	estimate	std.error	statistic	p.value	Team	p_adjusted
LocationHome	0.1533562	1.059106e+00	0.1447977	0.8848706	Brentford	0.9831895
LocationHome	3.6468257	2.354528e+00	1.5488565	0.1214162	Man United	0.8645206
LocationHome	0.3159042	1.734881e+00	0.1820899	0.8555122	Burnley	0.9831895
LocationHome	-0.4011160	1.216314e+00	-0.3297801	0.7415661	Chelsea	0.9831895
LocationHome	0.9400341	1.188693e+00	0.7908129	0.4290532	Everton	0.9534515
LocationHome	1.1852905	1.031461e+00	1.1491370	0.2504995	Leicester	0.9534515
LocationHome	-2.5967155	1.309111e+00	-1.9835715	0.0473036	Watford	0.8645206
LocationHome	0.7112241	1.154204e+00	0.6162029	0.5377606	Norwich	0.9831895
LocationHome	1.0348135	2.307455e+00	0.4484653	0.6538174	Newcastle	0.9831895
LocationHome	-2.6208672	1.729520e+00	-1.5153726	0.1296781	Tottenham	0.8645206
LocationHome	98.0371934	9.228264e+04	0.0010624	0.9991524	Liverpool	0.9991524
LocationHome	2.0325260	1.561956e+00	1.3012700	0.1931660	Aston Villa	0.9534515
LocationHome	0.8125379	9.213285e-01	0.8819199	0.3778201	Crystal Palace	0.9534515
LocationHome	-0.2444864	1.098692e+00	-0.2225250	0.8239052	Leeds	0.9831895
LocationHome	-0.0625335	2.158135e+00	-0.0289757	0.9768840	Man City	0.9991524
LocationHome	-1.4607350	1.811399e+00	-0.8064124	0.4200051	Brighton	0.9534515
LocationHome	1.0848329	1.085125e+00	0.9997311	0.3174407	Southampton	0.9534515
LocationHome	0.7139576	1.241822e+00	0.5749273	0.5653405	Wolves	0.9831895
LocationHome	0.6153642	1.195166e+00	0.5148774	0.6066387	Arsenal	0.9831895
LocationHome	-0.2698920	1.239412e+00	-0.2177581	0.8276176	West Ham	0.9831895

This results shows that after adjusting for goals, fouls, shots, shots on target, yellow cards and red cards, the association between home field and match results is not significant for every team. This is a surprising result, but this may due to that the home field advantage affects the match results by mediation of goals, fouls, shots, shots on target, yellow and red cards. Thus, after adjusting them, the direct effect of home field on the match results is not significant.

2. Home Field and Yellow/Red Cards

The first data set is used for this question. We fitted 2 zero-inflated Poisson mixed effect models to investigate the association between home field and yellow cards/red cards. Zero-inflated Poisson model is used because in most matches there is no red card. We also included a random effect for teams since the yellow and red cards in games of the same team might be correlated. The outcomes are the number of yellow cards and the number of red cards in each match. The covariates includes number of goals, fouls, corners, shots, and shots on target.

Table 2: Home Coefficient for Yellow Cards

effect	component	term	estimate	std.error	statistic	p.value
fixed	cond	(Intercept)	-1.9513846	0.5685669	-3.4321105	0.0005989
fixed	cond	Goal	-0.3700119	0.2181877	-1.6958422	0.0899158
fixed	cond	ShotsOnTarget	-0.0656027	0.1225391	-0.5353612	0.5924001
fixed	cond	Shots	-0.0565943	0.0520603	-1.0870912	0.2769965
fixed	cond	Corners	-0.1964866	0.0810756	-2.4234976	0.0153719
fixed	cond	Fouls	0.0969639	0.0392856	2.4681784	0.0135803
fixed	cond	LocationHome	0.1149960	0.3134023	0.3669276	0.7136730
fixed	zi	(Intercept)	-18.7132443	8145.8220882	-0.0022973	0.9981670

Table 3: Home Coefficient for Red Cards

effect	component	term	estimate	std.error	statistic	p.value
fixed	cond	(Intercept)	-1.9513846	0.5685669	-3.4321105	0.0005989
fixed	cond	Goal	-0.3700119	0.2181877	-1.6958422	0.0899158
fixed	cond	ShotsOnTarget	-0.0656027	0.1225391	-0.5353612	0.5924001
fixed	cond	Shots	-0.0565943	0.0520603	-1.0870912	0.2769965
fixed	cond	Corners	-0.1964866	0.0810756	-2.4234976	0.0153719
fixed	cond	Fouls	0.0969639	0.0392856	2.4681784	0.0135803
fixed	cond	LocationHome	0.1149960	0.3134023	0.3669276	0.7136730
fixed	zi	(Intercept)	-18.7132443	8145.8220882	-0.0022973	0.9981670

Table 2 and table 3 are the estimated coefficents for the two models.After adjusting for other variables. According to the results, the yellow cards and red cards are not significantly associated with home fields. This implies the referee bias is not significant with respect to the red and yellow cards.

Home Field and Goals

Again the first data set is used for this question. We fitted a Poisson Mixed effect model. The outcome is the number of goals the team scored in each game. The covariates includes number of fouls, corners, shots, and shots on target, yellow and red cards

Table 4: Home Coefficient for Goal

effect	component	term	estimate	std.error	statistic	p.value
fixed	cond	(Intercept)	-0.4548838	0.1361066	-3.3421145	0.0008314
fixed	cond	ShotsOnTarget	0.1844683	0.0148562	12.4168907	0.0000000
fixed	cond	Shots	0.0000831	0.0086054	0.0096574	0.9922946
fixed	cond	Corners	-0.0292980	0.0123781	-2.3669122	0.0179372
fixed	cond	Fouls	0.0042467	0.0094705	0.4484102	0.6538572
fixed	cond	YellowCards	-0.0343095	0.0276086	-1.2427100	0.2139747
fixed	cond	RedCards	-0.3645292	0.1838673	-1.9825667	0.0474158
fixed	cond	LocationHome	0.0617795	0.0626526	0.9860657	0.3241009

Table 4 are the estimated coefficents for the models.After adjusting for other variables, the number of goals is not significantly associated with home field. This might be because we adjusted for shots and number of shots. The presence of home supporters might have slight psychological impact on the away goalkeeper, but the impact is usually not that big for professional goalkeepers.

4. Cluster by Team Season Statistics

The second data set is used for this and the next question. We included 89 team season statistics that measures the defending, passing, possession, and shooting performances of the teams.

We first applied PCA on the 89 variables and made a scatter plot of the 20 teams with the first 2 principal components. Figure 3 contains this scatter plot.

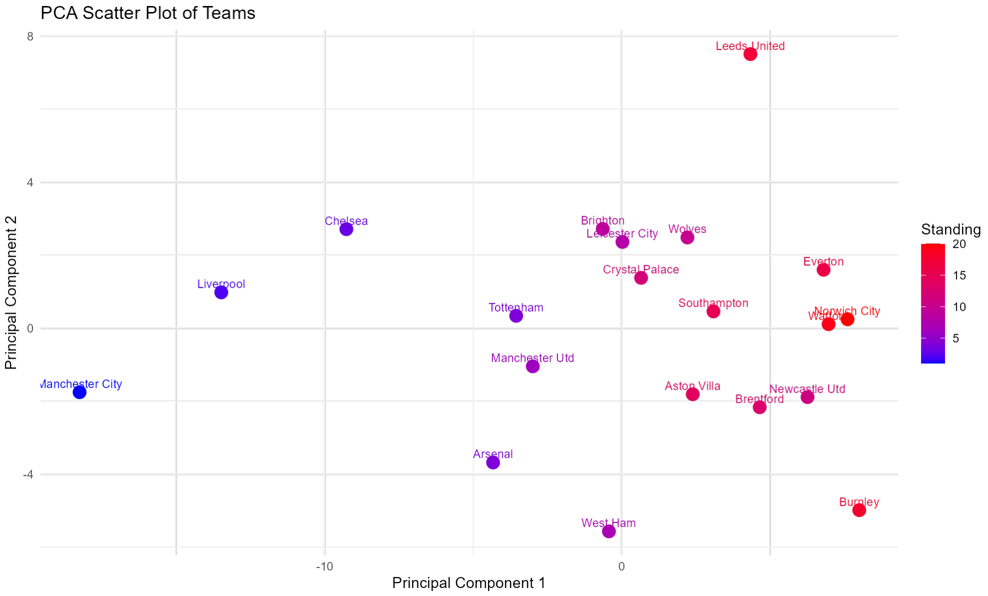

Figure 3: Scatter Plot of Teams with First 2 Principal Components

From Figure 3, the first component is strongly correlated with the standing. Teams on the left tends to have higher ranking and teams on the right have lower ranking. From left to right, the first 6 teams, “Man City”, “Liverpool”, “Chelsea”, “Arsenal”, “Tottenham”, and “Man United”, are exact in the same order as the standing.

We then use KNN to cluster the 20 teams into 4 clusters. Since KNN does not work well in high dimension, we still use the PCA to reduce the dimension. Figure 4 is a plot of the variance explained by each principal component and the cumulative variance explained. The elbow is on the second component. Thus, we use the first 2 components for the clustering, and the cluster is plotted in figure 5.

The cluster works pretty good for the top 2 clusters. Teams with ranking 1, 2, 3 are in the same cluster, and teams with ranking 4, 5, 6, 7 are in the same cluster. However, the third and forth clusters does not perfectly separate the teams. These might be because those teams are close to each other in the first component. Overall, the cluster perform well. Thus, those team in the same cluster are closer to each other.

5. Random Forest for Performance Importance.

We first use PCA on each of defending variables, passing variables, possession variables, and shooting variables to get the first principal component, totally 4 principal components. Then, we fitted a random forest with classes to be the top 10 in standing or the last 10. The number of features in each tree is set to be $\sqrt{4} = 2$. The reduced Gini Index is used as a measure of the importance in Figure 6.

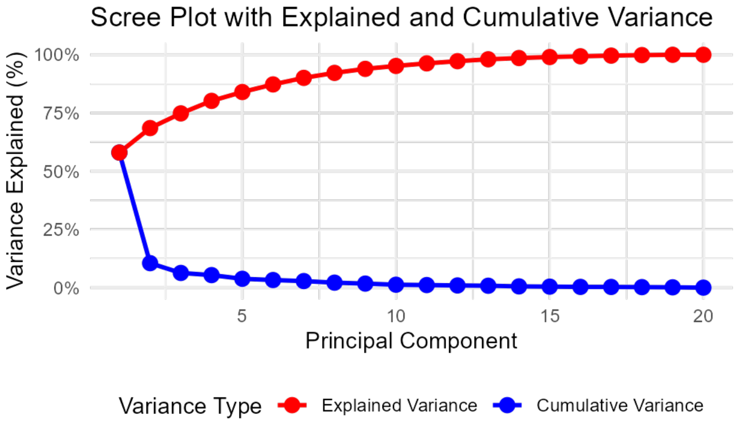

Figure 4: Variance Explained by Each Principal Components

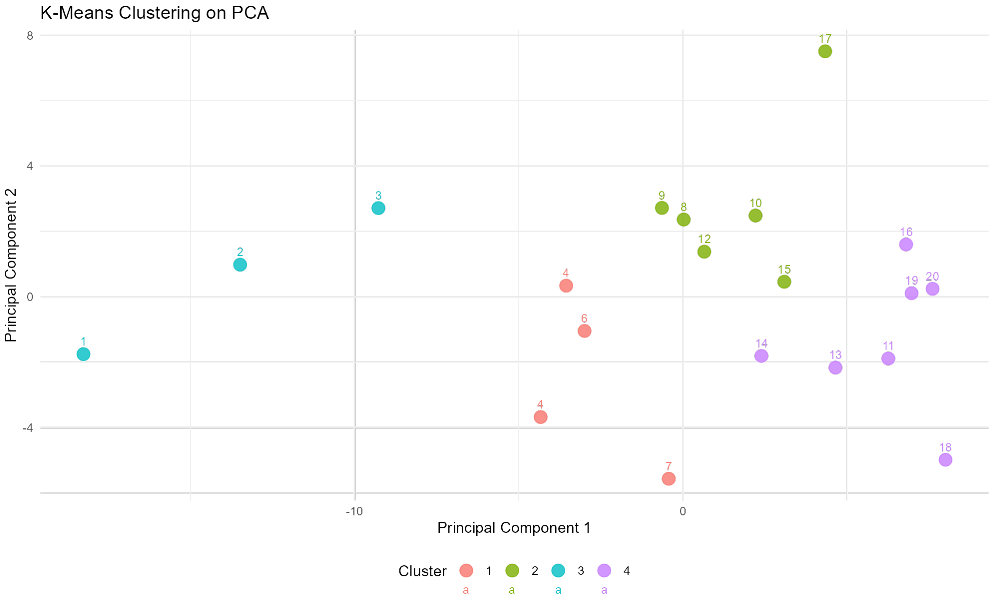

Figure 5: Clusters in Scatter Plot with First 2 Principal Components

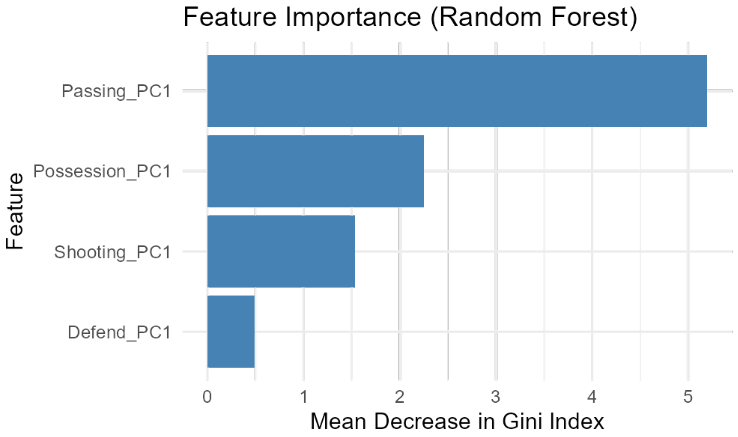

Figure 6: Importance of Performance

From the results, sassing performance is the most important to predict if the team is among top 10, followed by possession performance, shooting performance and defending performance.

Conclusion

From this project, we don't have enough evidence to show the association between match results, yellow and red cards, and goals and home field advantage, after adjusting for other match statistics. We also showed that after clustering teams by season defending, passing, possession, and shooting performance, the teams in the same cluster are close to each other. Passing performance is most important to predict the standing of a team.

Limitation is we after adjusting for other match statistics, we may ignored the effect of home field on the match results by mediation of these match statistics. Thus, a future work can be to investigate those effect.