PH125.9X Capstone project - Predicting the results of English Premiership Soccer Matches

Mike Woodwward

2021-01-01

Contents

1 Abstract **1**

2 Introduction **1**
2.1 Project goals 1
2.2 EPL background and modeling features 2
2.3 Prior work 2
2.4 Data sources and data definitions 3
2.5 Data preparation 3

3 Data analysis and feature selection **4**
3.1 Home field advantage 4
3.2 Team value advantage 5
3.3 Foreign players 5
3.4 Mean age 5
3.5 Squad size 5
3.6 Season effects 8
3.7 Discipline 8
3.8 Points 9
3.9 Feature selection 9

4 Modeling **11**
4.1 Holdout, test, and training data sets 11
4.2 Baseline 11

5 Discussion and conclusion **12**

6 References **12**

7 Appendix **12**

1 Abstract

2 Introduction

2.1 Project goals

The aim of this project was to predict English Premier League (EPL) soccer match results with an accuracy better than random guessing. The prediction target in this project means home team goals and away team goals.

2.2 EPL background and modeling features

Other authors have extensively described the origins and operation of the EPL (see for example [Robinson]), so I won't repeat them here. I will however describe some features relevant for modeling.

The league was founded by a group of clubs who wanted a higher revenue share of TV rights money and wanted larger TV deals [Butler]. This commercial focus has continued over the last twenty-five years, with the league becoming one of the most commercially successful sports organizations in the world [Robinson]. A key part of the league's success has been its ability to attract overseas talent, in fact, the EPL is know for the very high number of foreign players. This suggests two areas for investigation:

- Do financially bigger clubs win more matches?
- Does having a higher number of foreign-born players lead to greater success?

In common with nearly all soccer leagues worldwide, the EPL operates as the top-tier league, with a system of promotion and relegation from the league below it (now called the Champions League). Each year, several clubs are promoted and several relegated, with the rules for promotion and relegation changing over time (e.g at the 1994-1995 season, the league dropped from 22 teams to 20 via relegation). Finishing top and bottom of the league have substantial financial and reputational implications. The team that finishes top of the league are the league champions, and the top teams qualify for European competition. European competitions are very lucrative, both from match attendance and from TV rights. The bottom teams may be relegated, which means a very large drop in revenue and may cause good players to leave the club. Therefore, at the end of the season, teams near the top or bottom of the league may have stronger motivations to win. This suggests another modeling feature:

- At the end of the season, do teams at the top and the bottom of the league play differently?

Home field advantage has been extensively discussed in the literature (e.g. [Pollard], [Leard], [Thomas]). In the EPL, teams play each other twice, once at home and once away (giving 380 games for a league of 20 teams). If there were no home team advantage, we would expect the number of home wins to be about the same as the number of away wins.

- Is there evidence of a home team advantage?

On-field fair-play has been an important issue for the EPL, and for English soccer as a whole. Players receive a yellow card as a warning, with a red card for dangerous play or serious rule-breaking. A player who received a red card or two yellow cards in the same game is sent off (can't play in the rest of the game) and cannot be substituted. His team has to play with one less player, a substantial disadvantage. However, yellow and red cards might also be associated with the kind of risk-taking that wins matches.

- Do the number of red cards and yellow cards affect a team's goal scoring ability?

2.3 Prior work

Mathematicians have studied gambling for hundreds of years; in fact, the whole discipline of probability theory was largely created to understand gambling [Epstein]. Unlike other areas of math, those who are successful at analyzing gambling may chose not to publish, instead becoming wealthy themselves [Mezrich, Meloche]! Despite the the disincentive to publish, researchers have released a large number of studies analyzing soccer matches.

- Home field advantage has been extensively studied (e.g. [Leard], [Pollard], [Thomas], [Vergina]) and has been found to exist in many sports, including the EPL [Allen].
- Dawson *et al* [Dawson] studied consistency of red and yellow cards. They found that referees penalized away teams more (which may contribute to the home team effect). Oberstone [Oberstone] found a weak link between yellow cards and team performance.
- Several researchers ([Plumley], [Barros]) have examined the link between financial performance and on-field performance, but financial performance has been measured using company financial reports (e.g. incomes statements), not the transfer value of the team.

- Surprisingly little has been written about seasonal effects. Allen *et al* [Allen] found a relationship betwen the size of the home advantage effect and final league position.

2.4 Data sources and data definitions

EPL data is widely available on the internet, but much of it is in summary form. I used two detailed sources.

Match results. These came from Football-data. The data goes back to the foundation of the EPL in the 1992/1993 season, but the data before the 2000-2001 has fewer fields, for example, the red card data only appears from 2000-2001 onwards. I have chosen to only use data from 2000-2001 onwards in this analysis.

Market value, team size, and foreign players. This data comes from TransferMarkt.

- The market value for a club is the transfer value of its players, for example, if a team buys a new player for £200mn, then the value of the team goes up by £200mn. Transfermarkt update this value twice a month.
- A soccer team fields 11 players in a match, but substitutions are allowed and of course players get sick or may have to miss games due to births, deaths, marriages, etc. A team will typically have a roster of 20+ players they will choose between (the team size). Transfermarkt update this value twice a month.
- 'Foreign' players here means any player born outside of England. There are some complexities with this definition (for example, an immigrant child may have grown up in England and be English in all but birth, but this definition still counts them as 'Foreign'), but it's the best available definition. TransferMarkt update this data at the start of the season.

Although the amount of data isn't large, there a several hundred files downloaded. Download times are of the order of half an hour.

Please note: continually downloading data from websites (instead of downloading it once and caching it) is considered bad practice. Websites pay for bandwidth and many websites ban users for too many downloads. In this project, I was careful to download the data as little as possible.

2.5 Data preparation

English soccer teams are often known by several names, for example, Manchester United is also known as:

- Man Utd
- Man United
- Manchester United
- Manchester United FC
- MUFC

and various derivates and combinations. To join data from different sources, I needed a consistent naming convention. I used the EPL codes for teams and mapped name variations to the code, for example, I mapped 'Man United' and 'Manchester United' to the code MUN.

TransferMarkt calculates team values and team sizes twice a month, but games are held many times a month on different days. To map team value (and team size) to matches, I used r's fill function to interpolate team values for the day of the match.

Using a simple join on team and season, I use the foreign player count at the start of the season for all matches in the season. If a team purchases or sells a foreign player during the season, the foreign player count will no longer be accurate. However, EPL teams are known for having a very high number of foreign players, in which case, adding or removing a small number of foreign players might have a small effect.

I stored the data used for analysis in a variable called!!!!!!!!!! In the Appendix, I explain the fields used.

3 Data analysis and feature selection

As explained above, I only used match data from 2000-2001 onwards for this analysis.

3.1 Home field advantage

If there were no home field advantage, we would expect the number of home and away wins to be roughly equal. More formally, we might expect:

$$\frac{count\ of\ home\ wins}{count\ of\ home\ wins + count\ of\ away\ wins} \approx 0.5$$

and the test for equivalence would be a z-test or a t-test as appropriate.

Here are the results for the EPL from the 2000-2001 season onwards. The error bars are the 95% confidence interval.

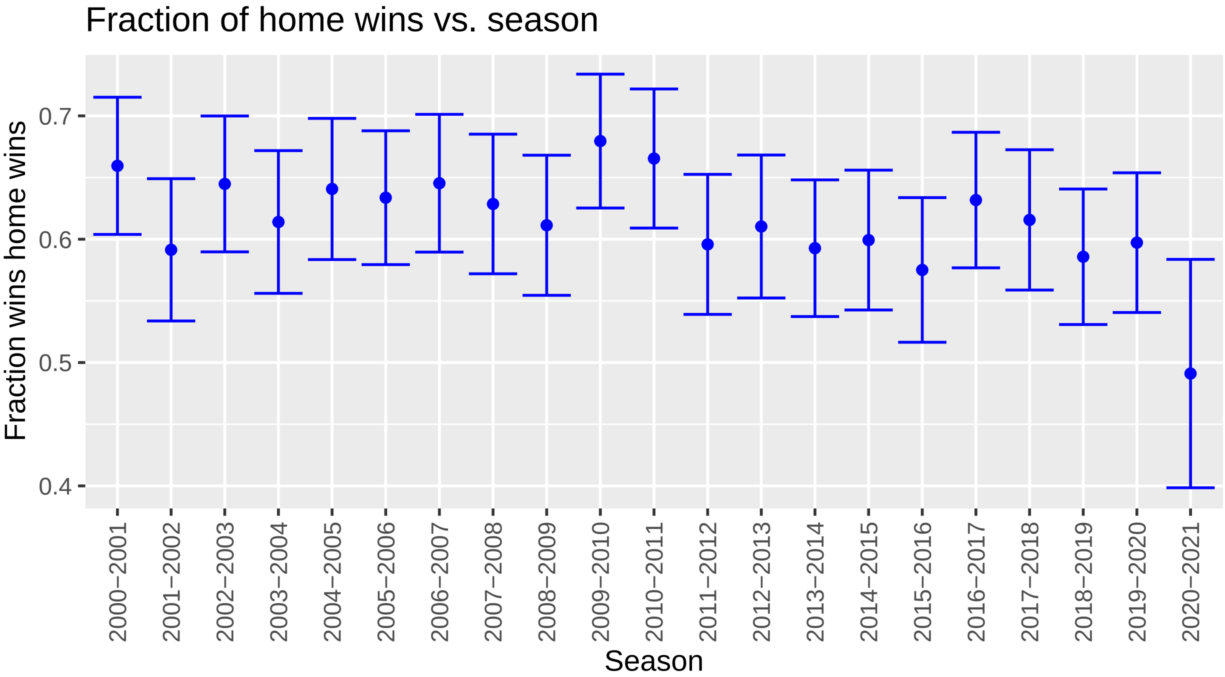

Figure 1: Fraction of home wins vs. Season

Even without running a formal statistical test, it's obvious there's a very strong home field advantage and therefore this is a feature I need to include in my model. The magnitude of this result is consistent with the literature [Allen].

Interestingly, the 2020-2021 results suggest a mechanism for home field advantage. Due to COVID-19, this season is running entirely without spectators, teams are playing in empty stadiums. Notably, the fraction of home wins, 0.4910714, is close to 0.5 (the larger error bars are because the season is only part way through at the time of analysis). It seems like that the home field advantage may be due to home spectators.

The home team effect is also apparent if we look at goal difference. Goal difference is the difference between the number of goals scored by the home team and the away team. In Figure 2, I've plotted the mean goal difference (over all games in the season) against the season. Clearly, home team advantage is worth about 0.35 goals, except for 2020-2021 (more evidence of a COVID effect).

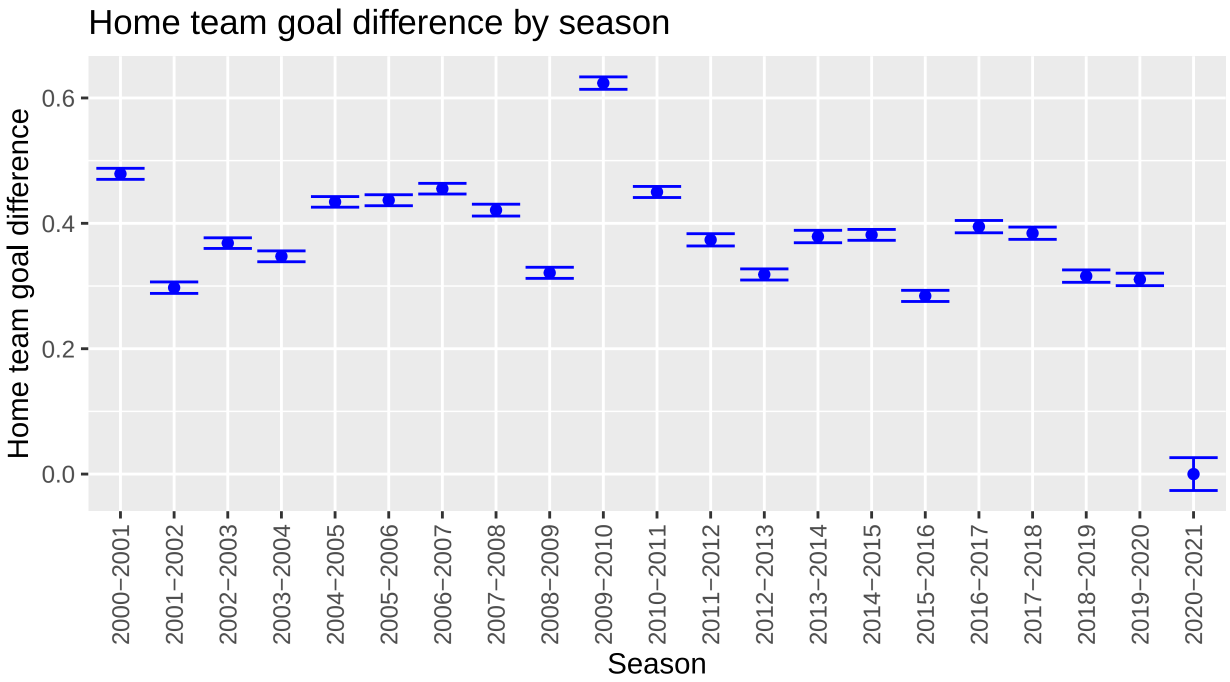

Figure 2: Mean(Home goals - away goals) vs. Season

3.2 Team value advantage

Is having a more valuable team than your opponents an advantage? As a reminder, team value in this project is the notional transfer value of the team as reported by TransferMarkt.

For each match in each season, I calculated a value difference and a goal difference. Figure 3 shows the result, each point is a match (with an alpha of 0.3 to show where matches overlap), the straight line is a linear fit, with the gray zone a 95% confidence interval. The chart clearly shows the aggregate effect of a value difference between teams, with a £700mn difference worth about 2 goals for the more valuable team.

3.3 Foreign players

The EPL is famous for having large numbers of foreign players, TransferMarkt notes that about *63%* of players in the league are foreign born. The obvious question is, does having foreign born players give a team an advantage?

For each match in each season, I plotted goal difference vs. the difference in foreign player count. Figure 4 shows there's a small effect.

3.4 Mean age

It may be true that younger players have more energy, older players have more experience, but what about at the team level? Does the mean age of the team make a difference? For each match, I plotted the goal difference vs the difference in mean age for the teams (Figure 5). There is an effect, worth maybe a goal for a 5 year difference, to put it simply, older teams appear to be at a disadvantage.

3.5 Squad size

There are league rules on the maximum size of squads, current 25 players. For the 2020-2021 season, every team has a squad of 25 players, but that hasn't always been the case. Is squad size a useful feature? I plotted goal difference against squad size difference for every match below.

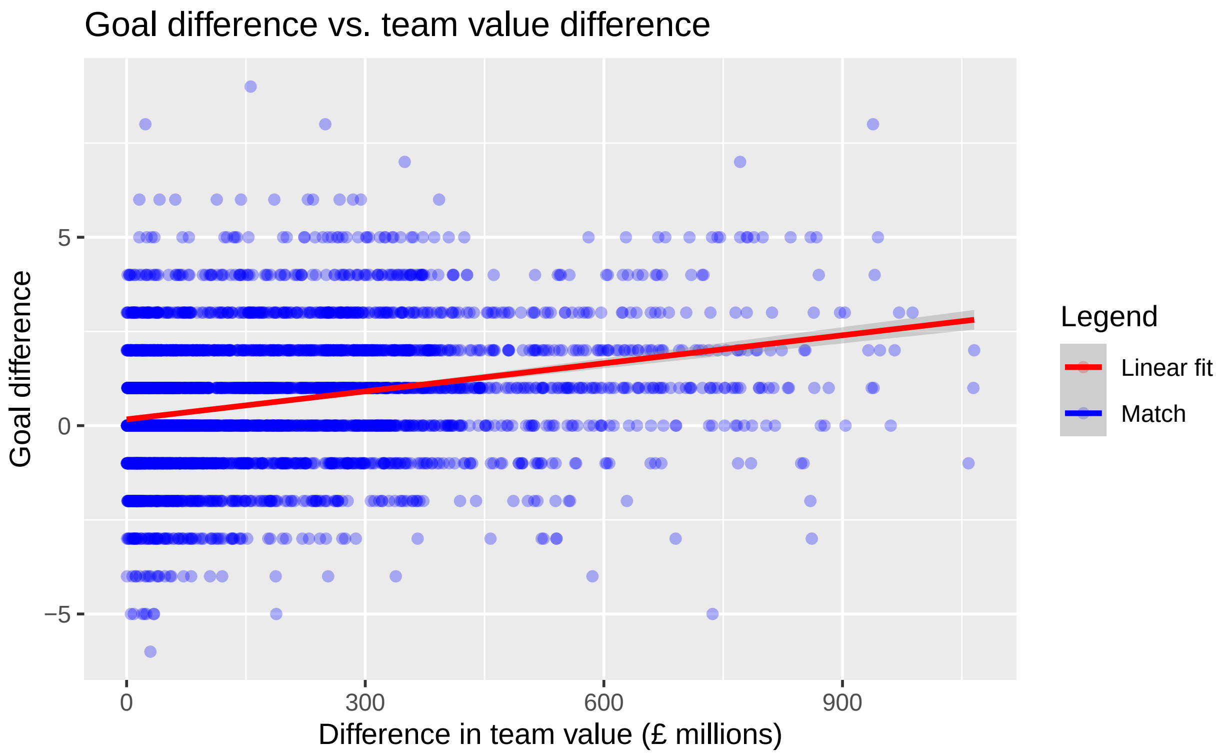

Figure 3: Goal difference vs. value difference

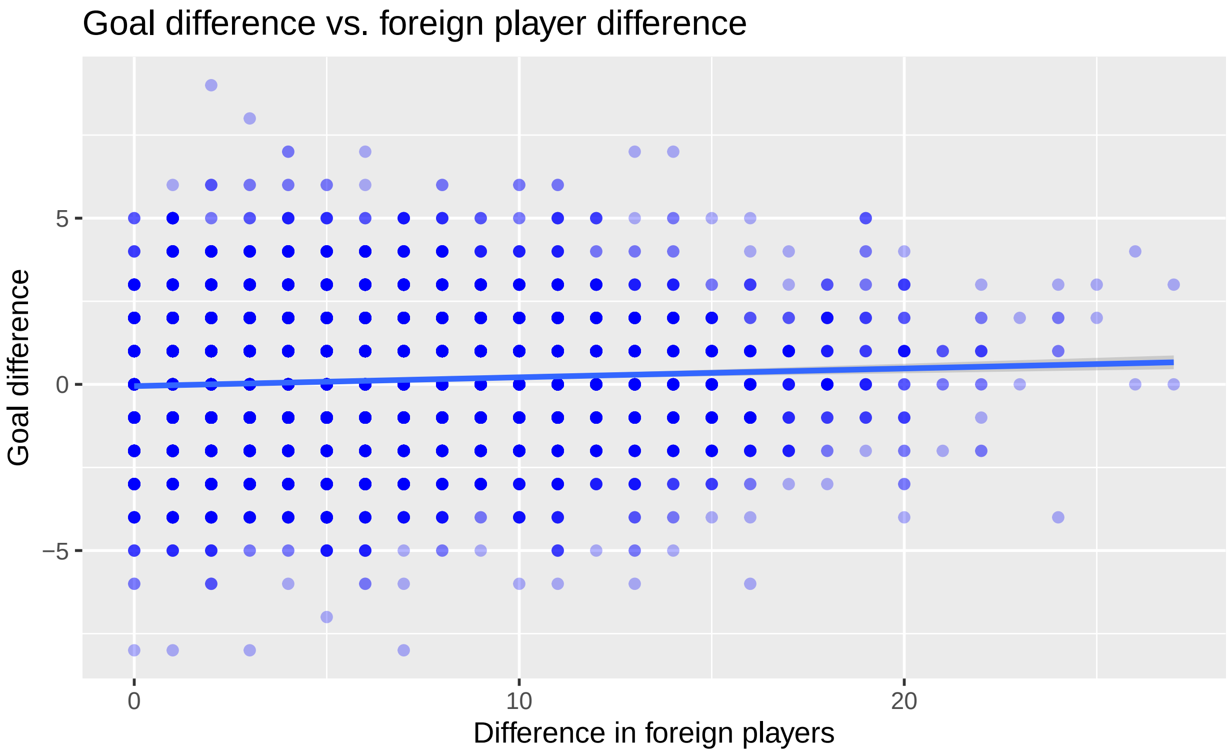

Figure 4: Goal difference vs. foreign player difference

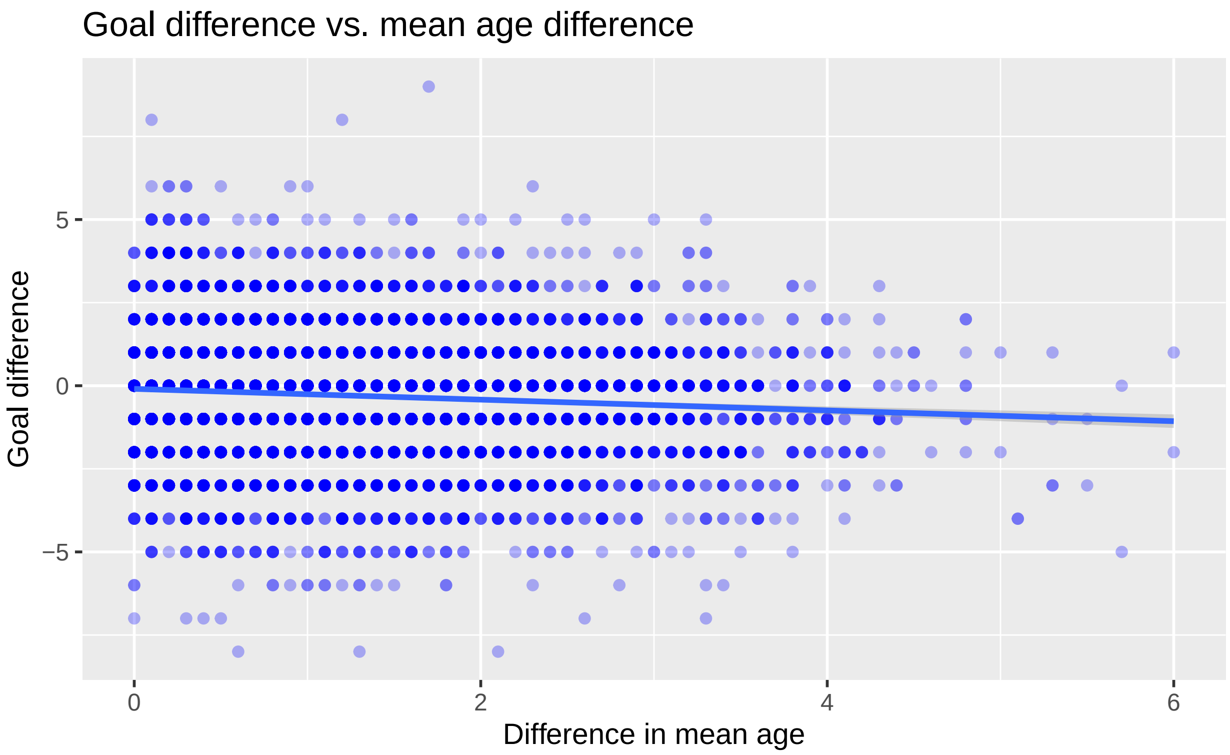

Figure 5: Goal difference vs. difference in mean age

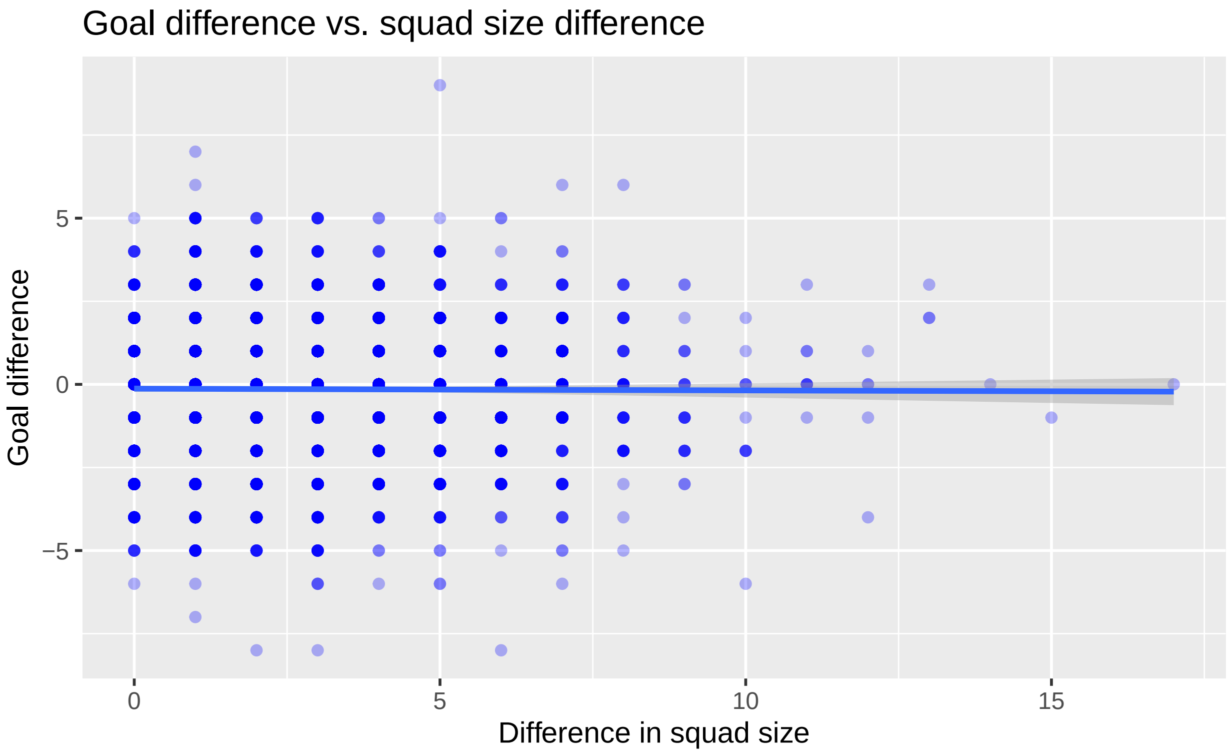

Figure 6: Goal difference vs. difference in squad size

If there is an effect, it's very small. As squad sizes are now limited to 25 and every team has a squad of 25, I've excluded squad size as a modeling feature.

3.6 Season effects

As I explained earlier, some teams at the end of the season may well face additional incentives to win. If we look at the season week, we should see the proportion of matches that ended in a draw go down as the season progresses.

(Season week effects required more data preparation than usual. It's easy to calculate a week of year, but the season typically starts in August and runs to May of the next year, meaning the week number will rollover to 1 in the first week of the new year. However, we can work out an offset and apply it to all match dates and then take the week number - this will yield season week number in a consistent way.)

For each season week, I calculated the proportion of matches that were draws and plotted them in Figure 7. There's a clear trend downward as the season progresses. Looking at the absolute difference in goals scored per match, and taking a mean per season week, I have Figure 8. This clearly shows goal difference increases as the season progresses.

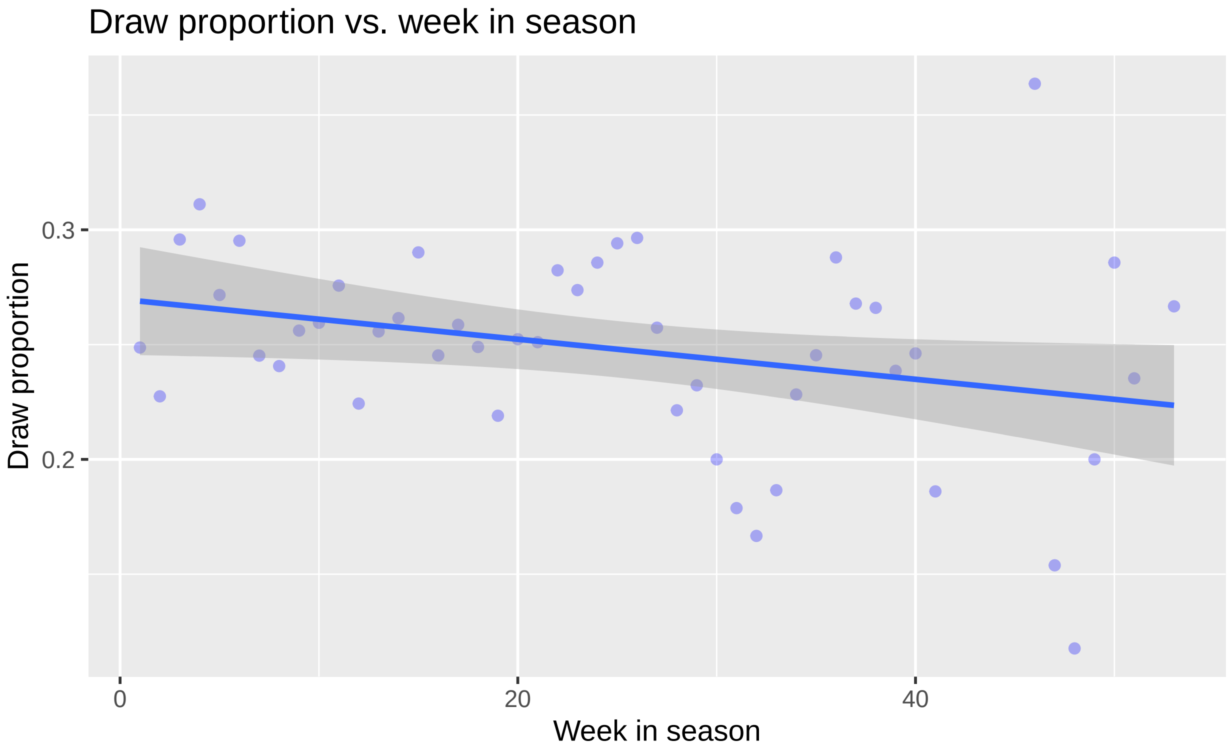

Figure 7: Draw proportion vs. week in season

This effect is also apparent in the mean absolute goal difference as the season progresses, meaning games are won by a larger goal margin (Figure 8).

3.7 Discipline

Are the number of red cards or yellow cards a team has a useful predictor? As the season progresses, teams will accumulate more red cards and yellow cards, so to remove season time effects, I took the mean number of red cards and yellow cards prior to the game, so for game i between teams A and B, my red card difference was:

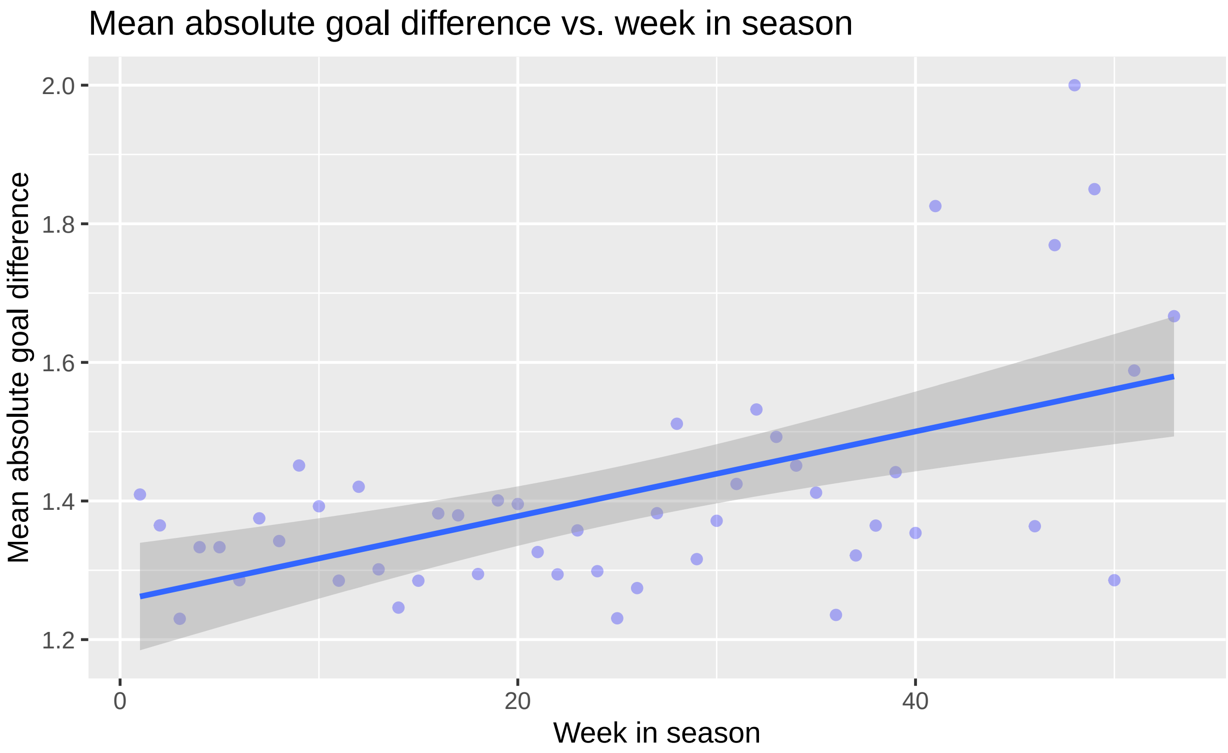

Figure 8: Mean absolute goal difference vs. week in season

$$\left(\frac{1}{i-1}\sum_{g=1}^{i-1} red\ cards\right)_{TeamA} - \left(\frac{1}{i-1}\sum_{g=1}^{i-1} red\ cards\right)_{TeamB}$$

where A team A was the team with the higher number of red cards.

The plots below shows very little effect for red cards and yellow cards. Even countin the number of red/yellow cards instead of taking an average shows very little effect.

3.8 Points

Like most soccer leagues, the EPL awards 3 points for win, 1 for a draw, and none for a loss. Each team will have a number of points before a match. Points encode a teams track record, the more successful a team is, the more points it will have. Does the difference in points before a match predict goal difference? To remove season effects, I calculated the average number of points.

There's a strong effect. To put it simply, teams with a stronger track record tend to score more goals against teams with a weaker track record.

3.9 Feature selection

My analysis has shown the following features are worth including in machine learning:

- Home team
- Team value difference
- Foreign player count difference between teams
- Mean age difference between teams

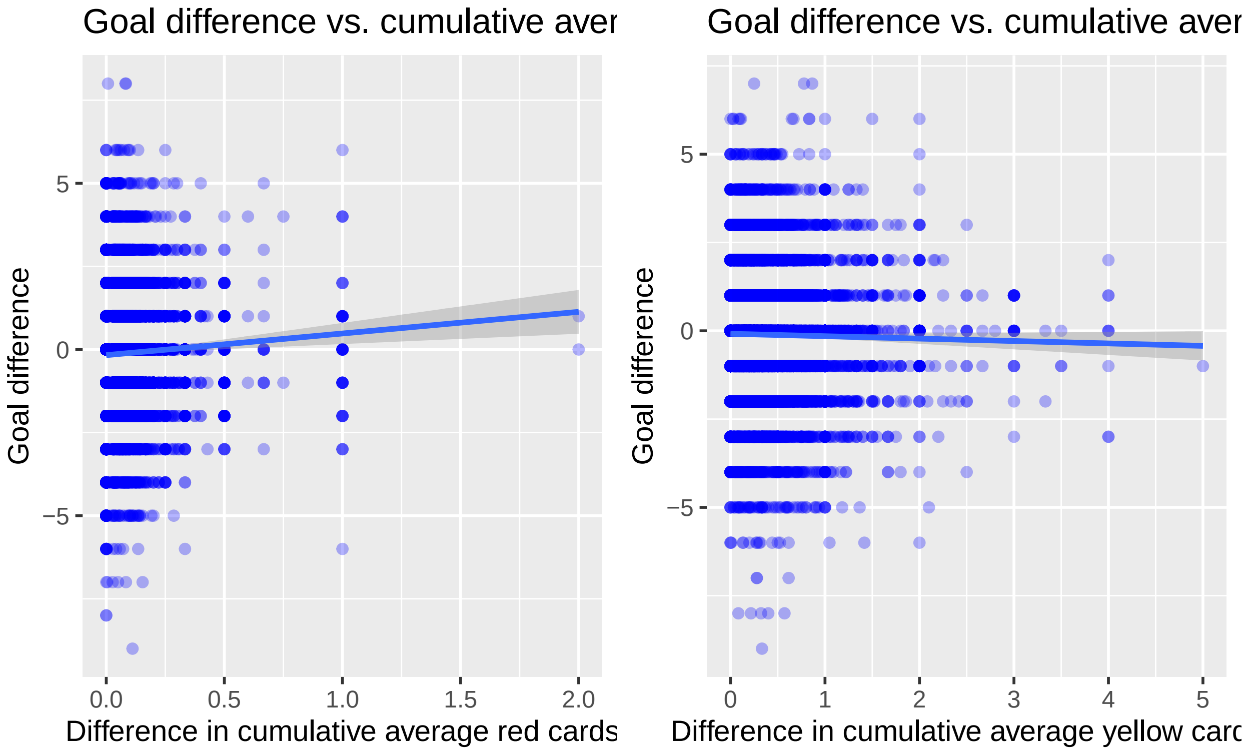

Figure 9: Goal difference vs. red card difference

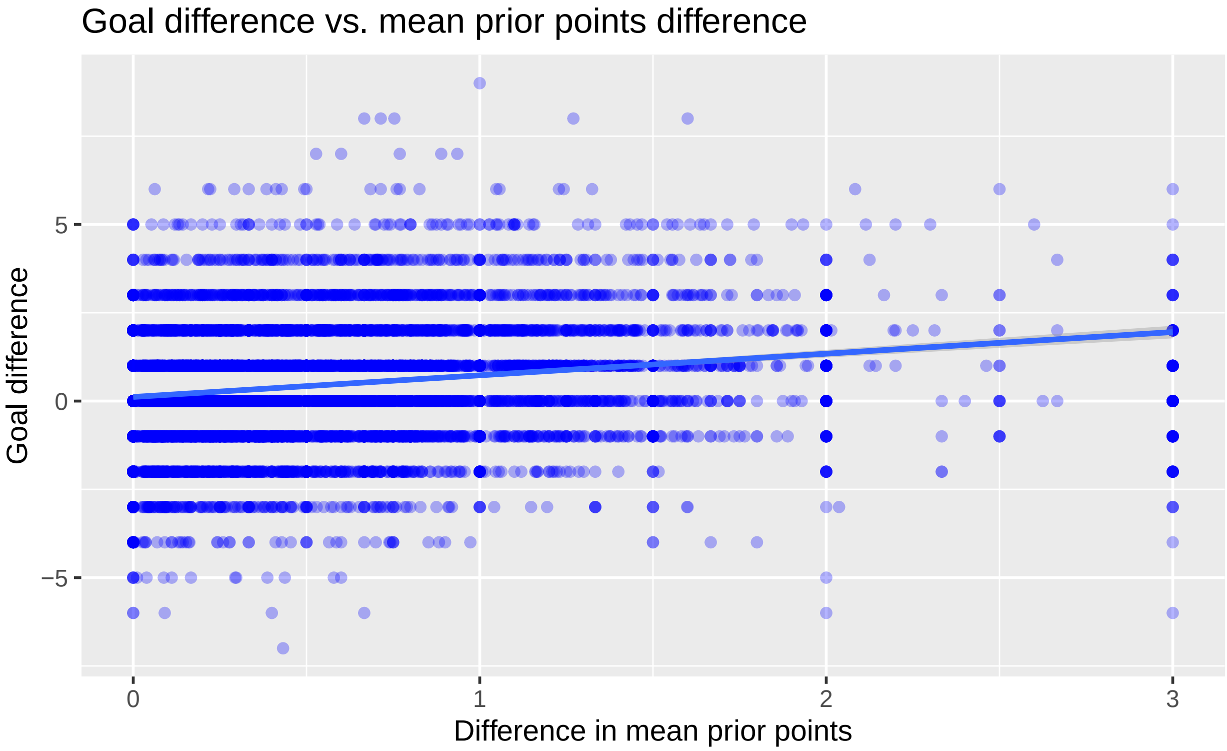

Figure 10: Goal difference vs. mean point difference

- Week in season
- Mean points difference

But the following features are no worth including:

- Team size
- Red card/yellow card count.

Because COVID-19 is such a disruptive event, I will exclude the 2020-2021 season from my modeling work.

4 Modeling

A soccer match result is a two-dimensional variable, it consists of a home team score and an away team score. This is known as multi-dimensional classification [Borchani] and the techniques involved go well beyond the scope of PH125. To tranform this into a one-dimensional problem, there are at least two approaches:

- Predict goal difference rather than score. The easiest way to do this is to calculate *home goals* – *away goals*, however, it becomes much harder to analyze home advantage in this model because it's already factored into the target variable. This is approach is also less satisfying from a prediction view point.
- Have two models, one a home goals model and one an away goals model. This will predict scores, which is more satisfying, and it also enables us to model home advantage. We can introduce a binary variable 'home' which is true for a team playing at home and false for a team playing away. The downside of this approach is I may be missing interactions between home and away teams.

On balance, I decided to use two models, one for home goals and one for away goals.

Goals are are integer numbers ≥ 0, however, there is value in model predictions of real numbers. For example, if 2 away goals were scored in a match, and two models predicted 1.5 and 1.9, we would consider the 1.9 model to be better. My success criteria is minimizing a loss function, and similarly to the MovieLens project, I decided on RMSE as my loss function, where:

$$RMSE = \sqrt{\frac{1}{N}\sum_{u,i}(\hat{y}_{u,i} - y_{u,i})^2}$$

where: N *is the number of matches*, $\hat{y}_i$ is the predicted number of goals for the ith game * y_i is the actual number of goals for the ith game

4.1 Holdout, test, and training data sets

I created a modeling and vlidation data set by partioning the

4.2 Baseline

I calculated home goals per match and away goals per match and obtained the following RMSE results:

- Home goals RMSE: 1.2683192
- Away goals RMSE: 1.3153368

Given that soccer is low scoring game, these RMSE scores are not good.

5 Discussion and conclusion

6 References

[Allen] Mark S. Allen, Marc V. Jones, The home advantage over the first 20 seasons of the English Premier League: Effects of shirt colour, team ability and time trends, International Journal of Sport and Exercise Psychology. Vol. 12, No. 1, 10–18

[Barros] Barros CP, Leach S. Analyzing the Performance of the English F.A. Premier League With an Econometric Frontier Model. Journal of Sports Economics. 2006;7(4):391-407

[Borchani] Borchani, H., Varando, G., Bielza, C. and Larrañaga, P. (2015), A survey on multi-output regression. WIREs Data Mining Knowl Discov, 5: 216-233,

[Butler] Robert Butler, Patrick Massey, Has Competition in the Market for Subscription Sports Broadcasting Benefited Consumers? The Case of the English Premier League, Journal of Sports Economics

[Dawson] Peter Dawson, Stephen Dobson, John Goddard, John Wilson, Are football referees really biased and inconsistent? Evidence on the incidence of disciplinary sanction in the English Premier League, Journal of the Royal Statistical Society: Series A (Statistics in Society), 170: 231-250

[Epstein] Richard A. Epstein, The Theory of Gambling and Statistical Logic, Academic Press, 2nd Edition, 2009

[Leard] Leard B, Doyle JM. The Effect of Home Advantage, Momentum, and Fighting on Winning in the National Hockey League. Journal of Sports Economics. 2011;12(5):538-560.

[Mezrich] Ben Mezrich, Bringing down the house, Atria Books, 2002

[Meloche] Renee Meloche, The High-Tech Gambler: The True Story of Keith Taft & His Astonishing Machines, 2012

[Oberstone] Joel Oberstone, Differentiating the Top English Premier League Football Clubs from the Rest of the Pack: Identifying the Keys to Success, Journal of Quantitative Analysis in Sports, Volume 5, Issue 3, 2009, Article 10

[Plumley] Daniel James Plumley, Robert Wilson and Simon Shibli, A holistic performance assessment of English Premier League football clubs 1992-2013. Journal of Applied Sport Management, 9 (1)

[Pollard] Richard Pollard and Gregory Pollard, Home advantage in soccer: a review of its existence and causes, International Journal of Soccer and Science Journal Vol. 3 No 1 2005, pp28-44

[Robinson] Joshua Robinson, Jonathan Clegg, The Club: How the English Premier League Became the Wildest, Richest, Most Disruptive Force in Sports, Mariner Books, 2019

[Thomas] Thomas S, Reeves C, Bell A. Home Advantage in the Six Nations Rugby Union Tournament. Perceptual and Motor Skills. 2008;106(1):113-116

[Vergina] Roger C.Vergina, John J.Sosika, No place like home: an examination of the home field advantage in gambling strategies in NFL football, Journal of Economics and Business Volume 51, Issue 1, January–February 1999, Pages 21-31

7 Appendix

Standard error of a proportion

If the proportion is $\frac{m}{n}$ where n is the number of samples, then:

$$\hat{p} = \frac{m}{n}$$

is the mean

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

is the standard error

The 95% confidence interval is:

$$\hat{p} \pm 1.96 * SE$$

Fields used in data analysis

ADD!!!!!