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

Mike Woodwward

2020-12-30

Contents

1 Abstract

2 Introduction

2.1 Project 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 to both a higher revenue share of TV rights money and who wanted large TV deals [Bulter]. 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:

- Are financially larger clubs more successful?
- 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 the league dropped from 22 teams to 20 via relegation). The team that finishes top of the league are the league champions, and the top teams qualify for European competition. Finishing top and bottom of the league have substantial financial implications. European competitions are very lucrative, both from match attendance and from TV rights. Relegation means a very large drop in revenue and may cause good players to leave the club. Therefore, teams at the bottom and the top of the league near the end of the season have stronger motivations to win. This suggests another modeling factor:

- 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 leage of 20 teams). Because of this, 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 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.

- Are red cards and yellow cards affect results?

2.3 Prior work

Mathematicians have studied gambling for hundreds of years; 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.

Pollard & Pollard [Pollard] Home win

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 XXXX has fewer fields, for example, the red card data only appears from XXX onwards. I have chosen to only use data from XXX onwardsin 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. Tranfermarkt 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. Tranfermarkt 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, 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.

3 Data 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:

$$\frac{countofhomewins}{countofhomewins + countofawaywins} \approx 0.5$$

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.

ADD CHART

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. Interestingly, the 2020-2021 results suggest a mechanism for home field advantage. Due to COVID-19, this season is running entirely without spectators, the stadiums are empty while matches are played. 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 the XXX, 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.

GOALS WORTH

3.2 Team value advantage

Is having a more valuable team than your opponents an advantage?

For each match between teams A and B, I calculated a value difference and a goal difference. I arranged the math so that team A was always the more valuable team. The results vary by season, but they all show the same general shape, I've chosen the 2017-2018 season as a representative sample (Figure XXXX).

ADD

The curve fit is a linear model fit to the data. The gray shaded area is the 95% confidence interval. I've used an alpha value of 0.3 so you can more easily see higher density data, as expected, small goal differences are more common than large differences.

GOALS WORTH

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 the 2018-2019 season, I plotted goal difference vs. the difference in foreign player count. Figure XXX shows XXXX.

GOALS WORTH

This feature is worth including in my model.

3.4 Discipline

3.5 Season effects

3.6 COVID-19

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

4 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:

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

5 References

[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

[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

[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

5.1 R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

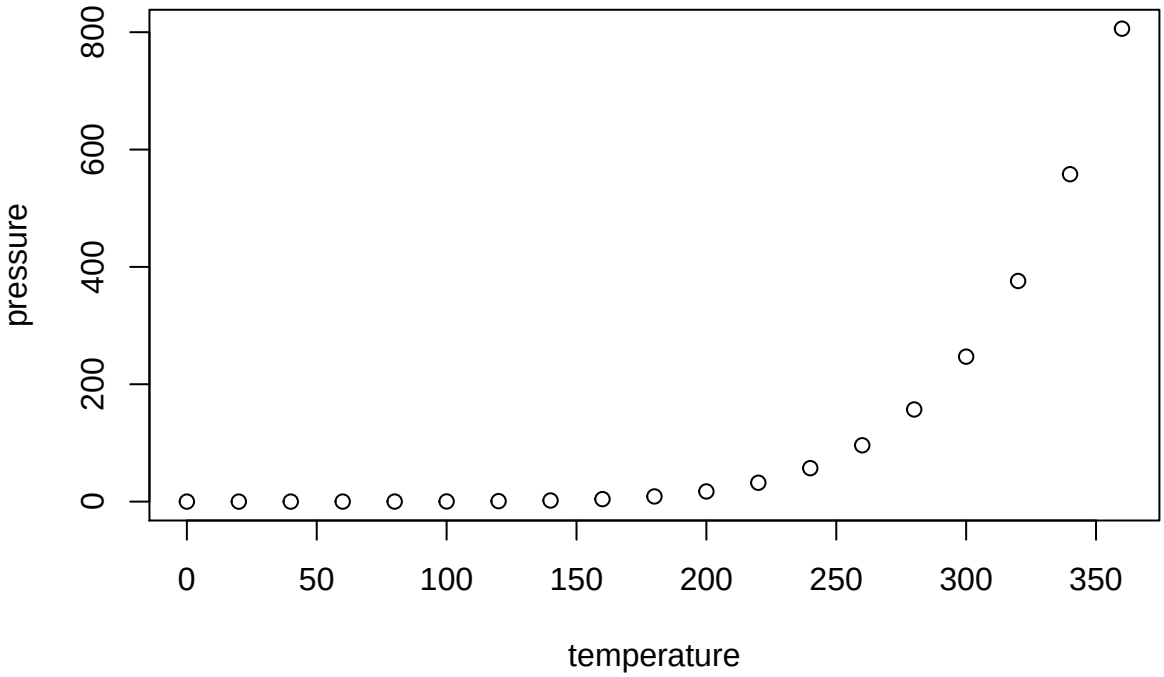
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

5.2 Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.