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

Mike Woodward

2021-01-06

Contents

1 Executive summary

I have submitted this report as part of the requirements for the HarvardX PH125.9X Capstone course. It explains how I built a machine learning system to predict the scores of English Premier League soccer matches.

I evaluated nine factors as potential modeling features and selected eight: home team advantage, team value difference, foreign player count difference between teams, mean age difference between teams, week in season, mean points difference, mean yellow card difference, and mean red card difference. Surprisingly, I found home

team advantage contributed little to my random forest model, suggesting the other features may have already accounted for it.

I tested a baseline and four models. The baseline was a simple goals scored means, the other models were glm, glmnet, random forest, and svm. The best performing model was glmnet with a RMSE of 1.1754744. This is disappointingly high and may be due to insufficient data, not enough modeling features, or the inherent randomness and unpredictability of soccer matches.

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 using mean results alone. The prediction target in this project was the number of goals scored in a match by the home team and by the away team.

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, however, I will describe some features relevant for modeling.

The league was founded by a group of clubs who wanted a larger share of TV rights money and wanted bigger 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 larger clubs score more?
- Does having a higher number of foreign-born players lead to more goals?

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 or bottom of the league has substantial financial and reputational implications. The team that finishes top of the league are the league champions, and the top few teams qualify for European competition. European competition is 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:

- Do teams play differently as the season progresses? Are there more wins and goals as the season progresses?

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 receives a red card, or two yellow cards in the same match, is sent off (can't play in the rest of the match) and can't 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 between 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. This data is available per season in CSV files.

Market value, team size, and foreign players. This data comes from TransferMarkt and is only fully available for the 2011-2012 season and onwards.

- 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 transfer 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 the team size 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.

I scrapped the TransferMarkt data from its website using rvest.

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 matches 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.

Season week turns out to be an important modeling parameter, but it required some effort to calculate. It's easy to calculate the week of the year, but the EPL 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 of the offset date, which will yield season week number in a consistent way.

I stored the data used for analysis and modeling in a data frame called match_results. In the Appendix, I explain the meaning of the field names in this data frame.

2.6 Running the software

Instructions for how to run the software are in the Appendix.

3 Data analysis and feature selection

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. The error bars are the 95% confidence interval (see the Appendix for the formula for the standard error of a proportion).

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, for 2020-2021, the fraction of home wins, 0.491, 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 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).

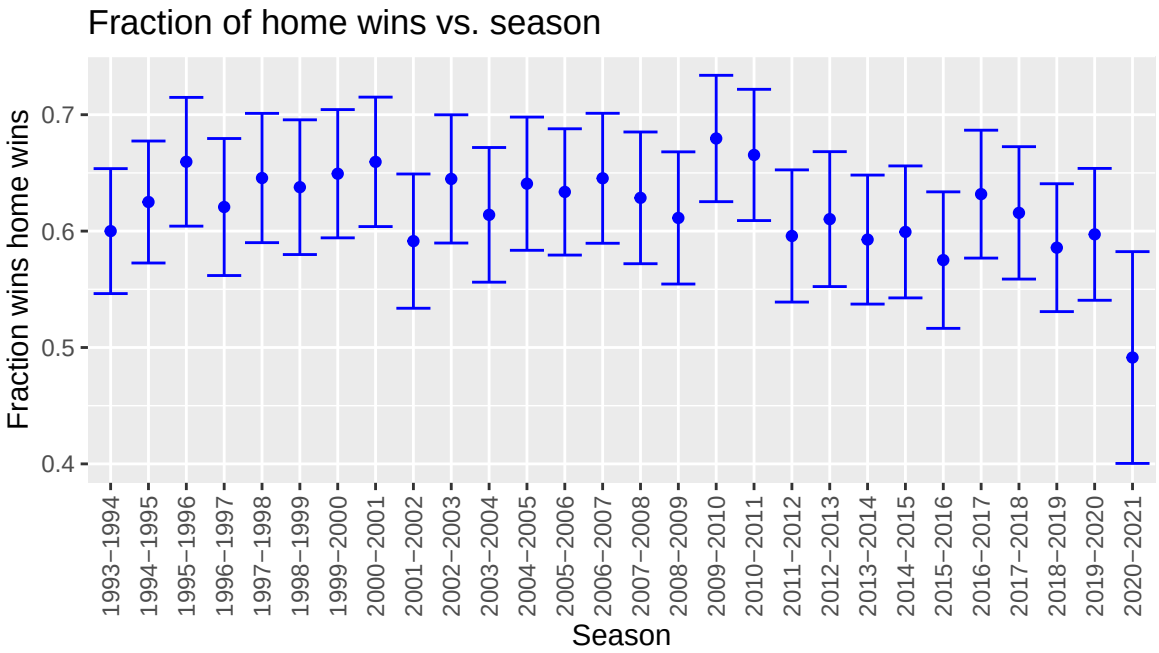

Figure 1: Fraction of home wins vs. Season

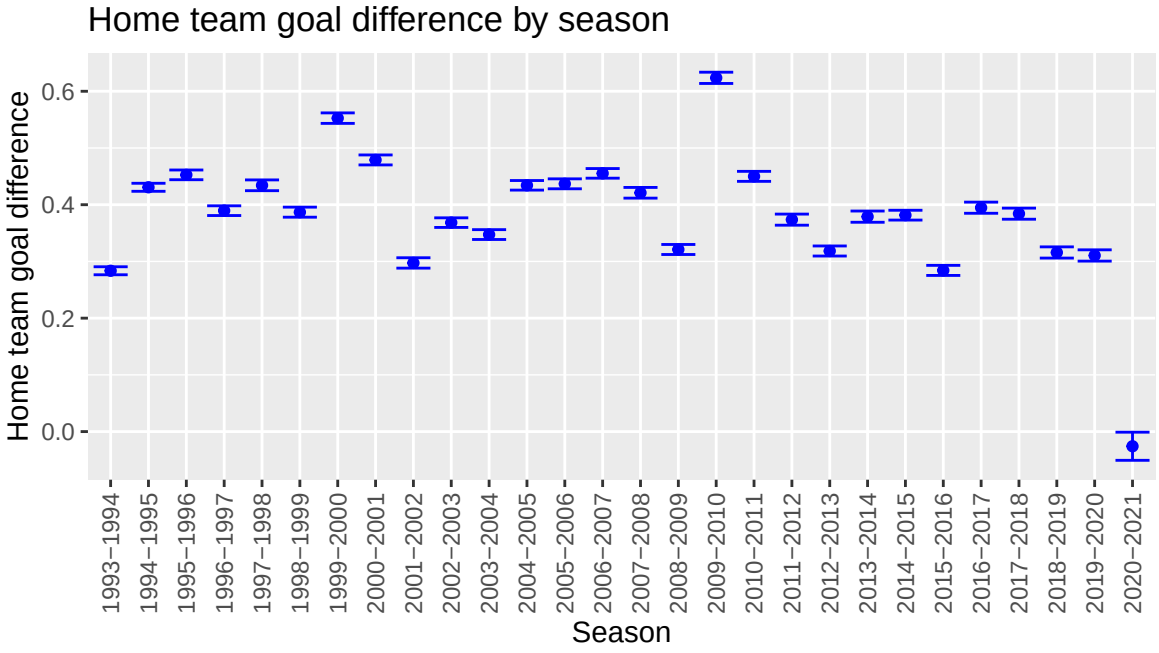

Figure 2: Mean(Home goals - away goals) vs. Season

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 (value difference is the difference in transfer value of the two teams and arranged so it's *most expensive team − cheaper team*, the goal difference is *most expensive team's goals − cheapest team's goals*). Figure 3 shows the result, each point is a match (with an alpha of 0.3 to show where matches overlap), the (red) straight line is a linear fit, with the light red 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.

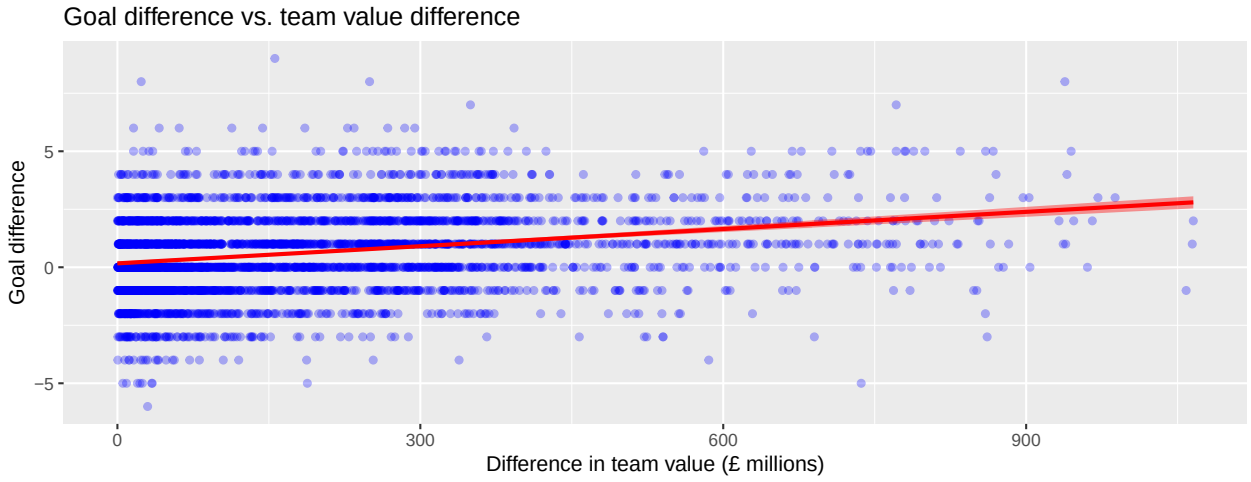

Figure 3: Goal difference vs. value difference

3.3 Foreign players

The EPL is famous for having large numbers of foreign players, TransferMarkt notes that for the 2020-2021 season, 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, which is currently 25 players. For the 2020-2021 season, every team has a squad of 25 players, but that hasn't always been the case and there was much more variability in the past. Is squad size a useful feature? I plotted goal difference against squad size difference for every match in Figure 6.

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.

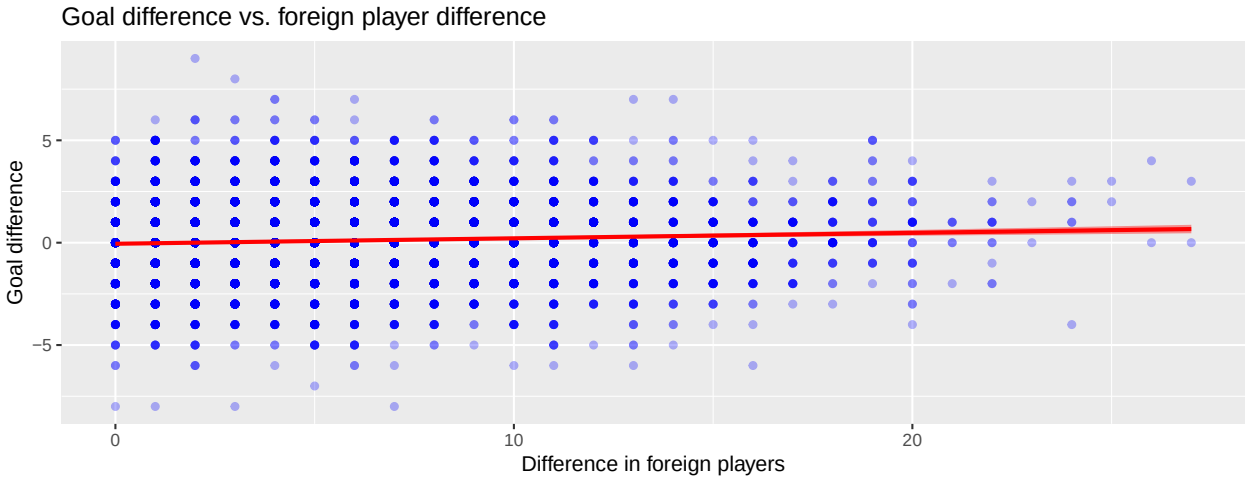

Figure 4: Goal difference vs. foreign player difference

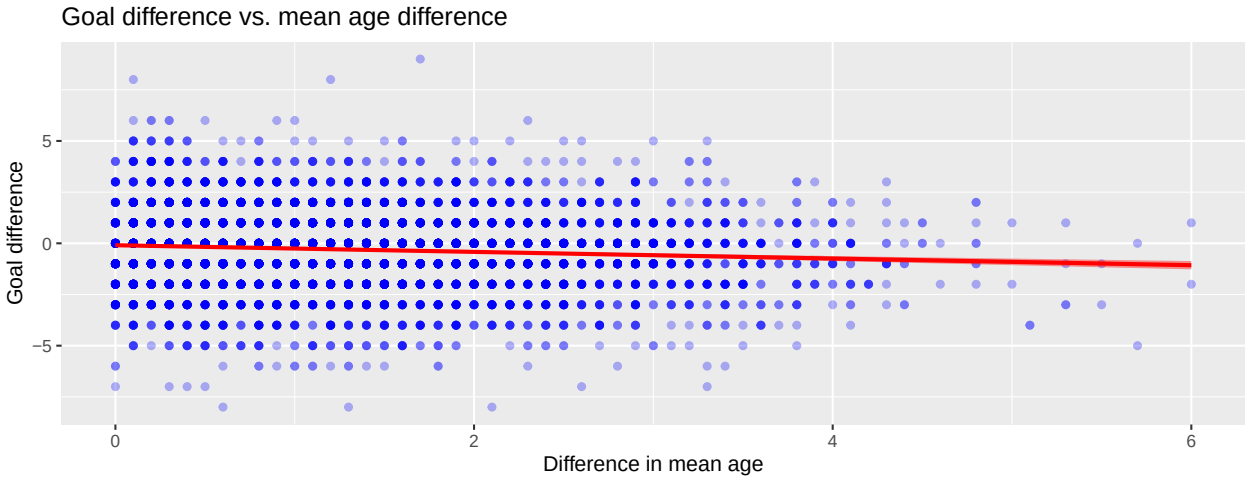

Figure 5: Goal difference vs. difference in mean age

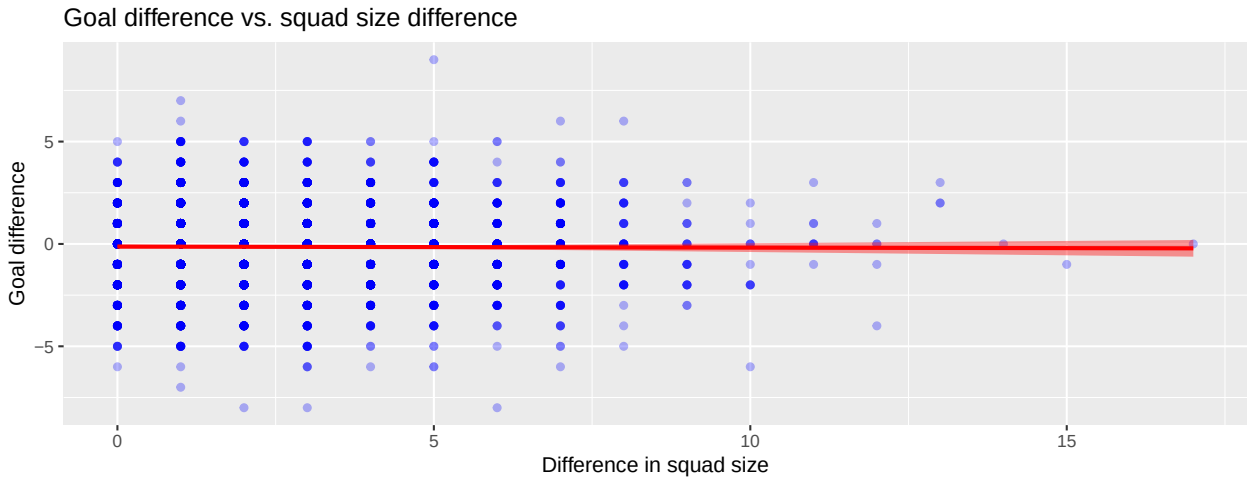

Figure 6: Goal difference vs. difference in squad size

3.6 Season effects

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

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.

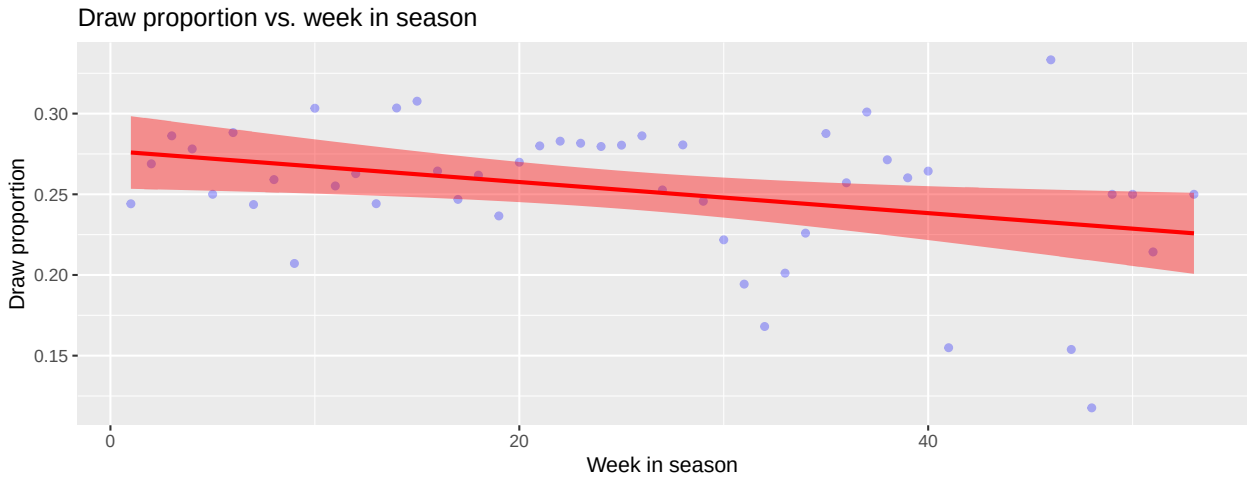

Figure 7: Draw proportion vs. week in season

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). Interestingly, the number of goals scored per match doesn't change very much during the season, suggesting it's the split of goals between the team that changes.

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

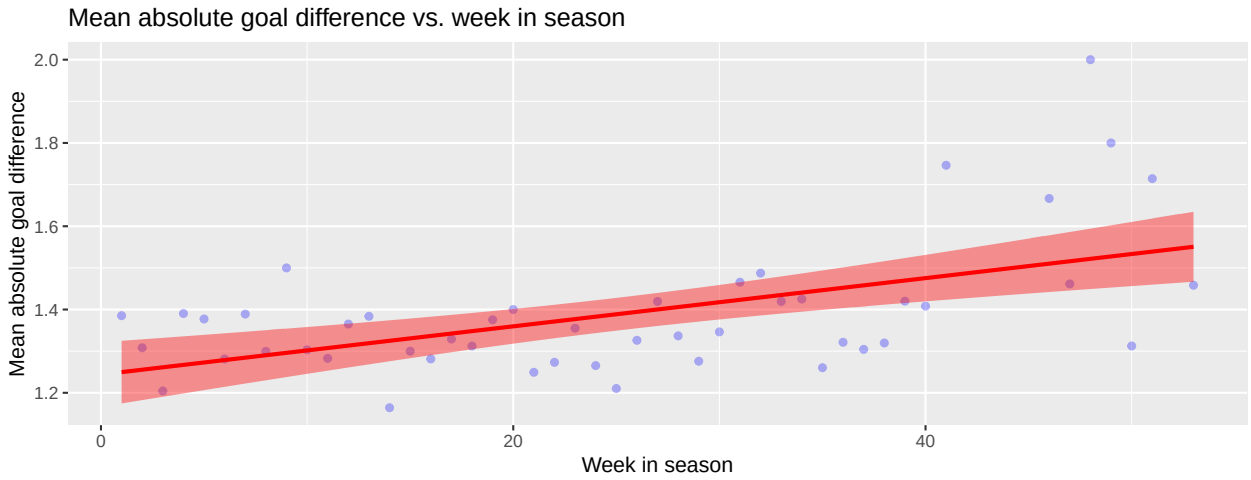

Figure 8: Mean absolute goal difference vs. week in season

red cards and yellow cards prior to the game. For game i between teams A and B, my red card difference was:

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

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

The plot below shows a small effect for yellow cards, but it is there.

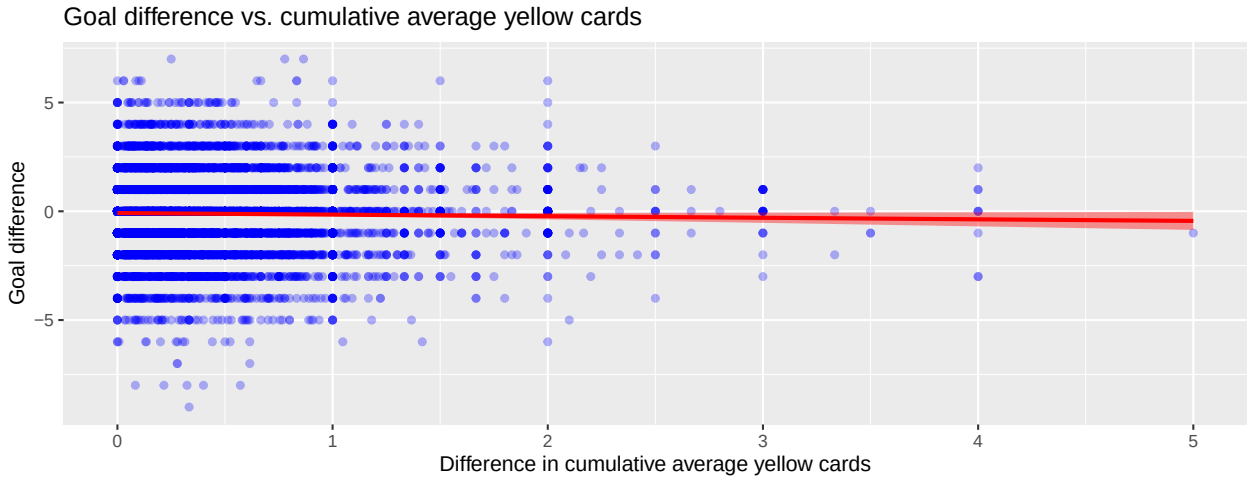

Figure 9: Goal difference vs. yellow card difference

The red card effect is small, and oddly, it's a positive 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 team's 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

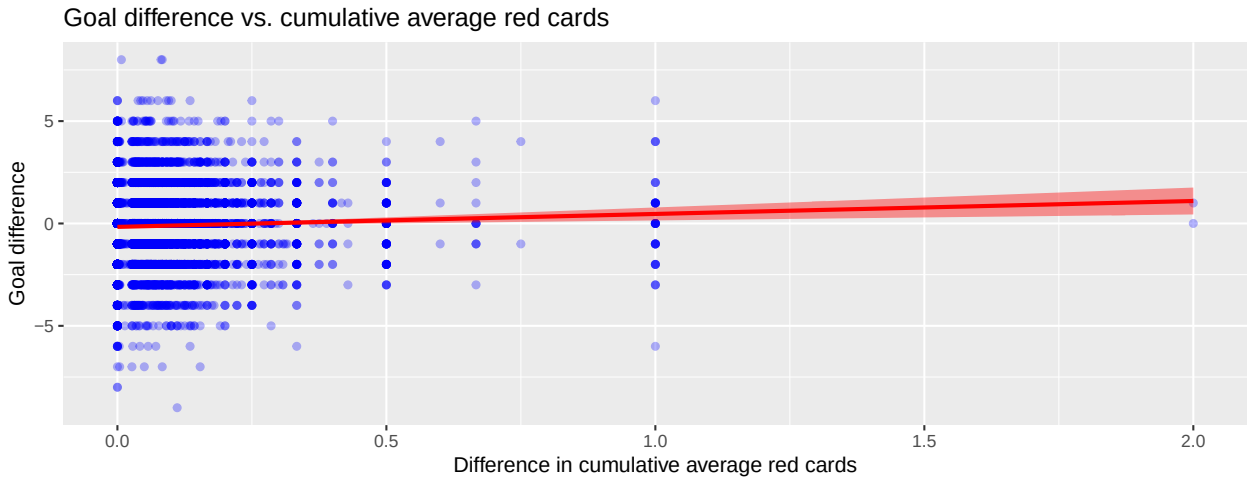

Figure 10: Goal difference vs. red card difference

remove time effects, I calculated the mean number of points (per match) each team has prior to the match I'm predicting.

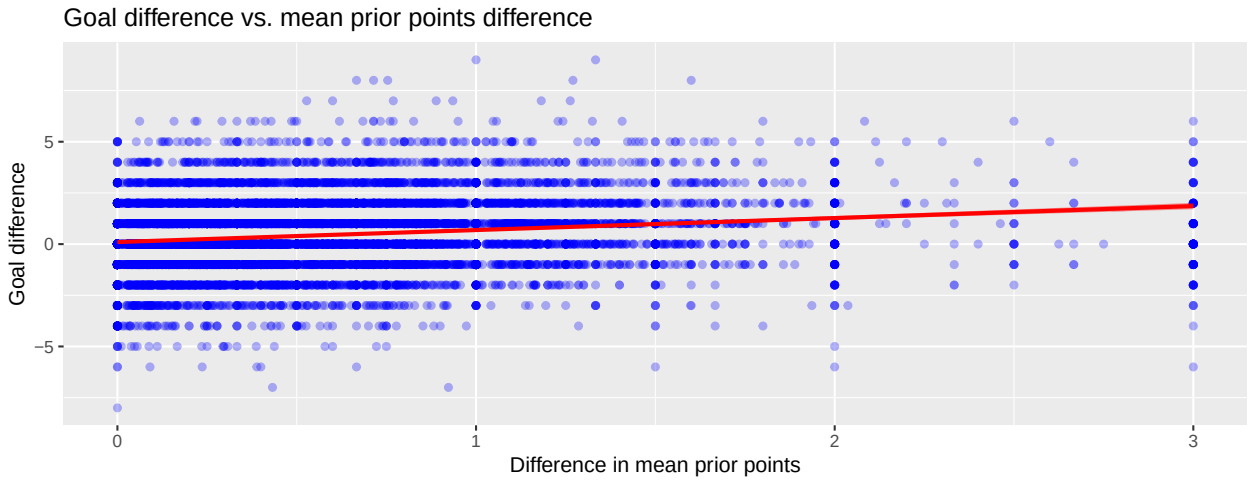

Figure 11: Goal difference vs. mean point difference

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

3.9 Feature selection

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

- Home team advantage
- Team value difference
- Foreign player count difference between teams
- Mean age difference between teams
- Week in season

- Mean points difference
- Mean yellow card difference
- Mean red card difference

But the following feature is not worth including:

- Team size

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

4 Machine learning modeling

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

- Goal difference between the teams. Predict the goal difference between the teams rather than match score. The easiest way to do this is to calculate $home\ goals - away\ goals$, however, this is approach is less satisfying from a prediction viewpoint because it doesn't predict match scores.
- One model, split games. In this approach, every game is split in two, with a 'Home' version and an 'Away' version. One model is trained on both Home and Away data, and a prediction is made for the Home version and the Away version of ech match. This version makes predictions of match scores, which is more satisfying.

I chose to use the one model, split games approach in this work.

Goals are an integer numbers ≥ 0, however, there is value in real number model predictions. 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:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i}(\hat{y}_i - y_i)^2}$$

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

Not all features are available for all seasons. For this reason, I selected seasons 2011-2012 and onwards for modeling. As I showed in feature selection, the 2020-2021 season shows a strong COVID effect, so I'll exclude this season from my modeling and prediction work.

4.1 Holdout, test, and training data sets

I split the matches data set 90%-10%, with 10% of matches as a holdout used for *final evaluation only* and 90% of the data used for training (in a variable called epl). The epl 90% I split again, 90%-10% as a train and test set. As I'll show, the only place where I used the train and the test data sets was to find the number of trees in the random forest model. In all other models I used cross-validation in train control in the caret train function using the epl variable, for the RMSE calculation, I used the holdout dataset.

4.2 One model, split games

Description of model

Each game is between two teams, a team playing at home and a team playing away. The data is organized on a per match basis. To create data suitable for this model, I split every game into a home and away version and calculated the appropriate feature value. For example, for the 2015-02-22 Tottenham vs. West Ham match, I have this data:

Date: 2015-02-22

Season: 2014-2015

WeekNumber: 28

HomeTeam: Tottenham

AwayTeam: West Ham

FTHG (Full time home goals): 2

FTAG (Full time away goals): 2

HomeTeamValue: 236.48

AwayTeamValue: 106.88

HomeTeamForeignPlayers: 25

AwayTeamForeignPlayers: 21

HomeTeamMeanAge: 23.7

AwayTeamMeanAge: 24.8

HomeMeanPoints: 1.72

AwayMeanPoints: 1.52

HomeMeanRedCards: 0.12

AwayMeanRedCards: 0.08

HomeMeanYellowCards: 2.2

AwayMeanYellowCards: 1.68

I split this games as shown, introducing a new field called Home. The Home value I set to 1 for the home team and 0 for the away team. Note that I transformed the feature values to *differences*. Here's the transformed data, note the away version is the 'mirror' of the home version.

This row of data is for the home team:

Date: 2015-02-22

Season: 2014-2015

WeekNumber: 28

Home: 1

Goals: 2

ValueDifference: 129.6

ForeignDifference: 4

MeanAgeDifference: 1.1

PointsDifference: 0.2

RedCardDifference: 0.04

YellowCardDifference: 0.52

This row of data is for the away team, note it's the 'mirror image' of the home team.

Date: 2015-02-22

Season: 2014-2015

WeekNumber: 28

Home: 0

Goals: 2

ValueDifference: -129.6

ForeignDifference: -4

MeanAgeDifference: -1.1

PointsDifference: -0.2

RedCardDifference: -0.04

YellowCardDifference: -0.52

I used this dataset for modeling.

Baseline

My baseline model is a mean goals model, I calculated the means using the epl set and evaluated the result with the holdout data set. The baseline RMSE result is 1.2489922.

Generalized linear model

This used the simple generalized linear model as shown below. Note the use of cross validation using the epl data set, removing the need for a separate train and test data set.

```
train(Goals ~ Home + WeekNumber + PointsDifference +
                          ValueDifference + ForeignDifference +
                          MeanAgeDifference + RedCardDifference,
      method = "glm",
      data = epl,
      trControl = trainControl(method = "repeatedcv",
                               number = 10,
                               repeats = 3,
                               p = 0.9),
      metric='RMSE',
      maximize=FALSE)
```

The improvement here was of the order of 10%, the RMSE with the holdout data was 1.1768333.

Glmnet model

The glmnet model is an evolution of the glm model. I used alpha values of 0 and 1, corresponding to ridge regression and lasso regression. The lambda values I set by experimentation to yield the lowest RMSE.

```
lambdas <- seq(0, 0.2, by=0.001)
fit_glmnet <- train(Goals ~ Home + WeekNumber + PointsDifference +
                        ValueDifference + ForeignDifference + MeanAgeDifference +
                        RedCardDifference,
              method = "glmnet",
              data = epl,
```

```
              metric='RMSE',
              maximize=FALSE,
              trControl = trainControl(method = "repeatedcv",
                                       number = 10,
                                       repeats = 3,
                                       p = 0.9),
              tuneGrid = expand.grid(alpha = 0:1,
                                     lambda = lambdas))
```

With the holdout data, this model gave about the same results as the glm model, 1.1754744.

Support Vector Model (SVM)

SVM is another machine learning method for regression. I used the linear form of the SVM as below, once again, note the use of cross-validation.

```
train(Goals ~ Home + WeekNumber + PointsDifference +
                  ValueDifference + ForeignDifference +
                  MeanAgeDifference + RedCardDifference,
      method = "svmLinear",
      data = epl,
      metric='RMSE',
      maximize=FALSE,
      trControl = trainControl(method = "repeatedcv",
                               number = 10,
                               repeats = 3,
                               p = 0.9),
      tuneGrid = expand.grid(C = seq(from=0.5, to=10, by=0.1)),
      preProcess = c("center","scale"))
```

With the holdout data, this model gave about the same results as the previous models, 1.185541. This model might not be an appropriate choice for the data set, as we can see from Figure 12. The 'svmLinear' model includes a tuning parameter C, and RMSE *should* vary with C, but it doesn't here.

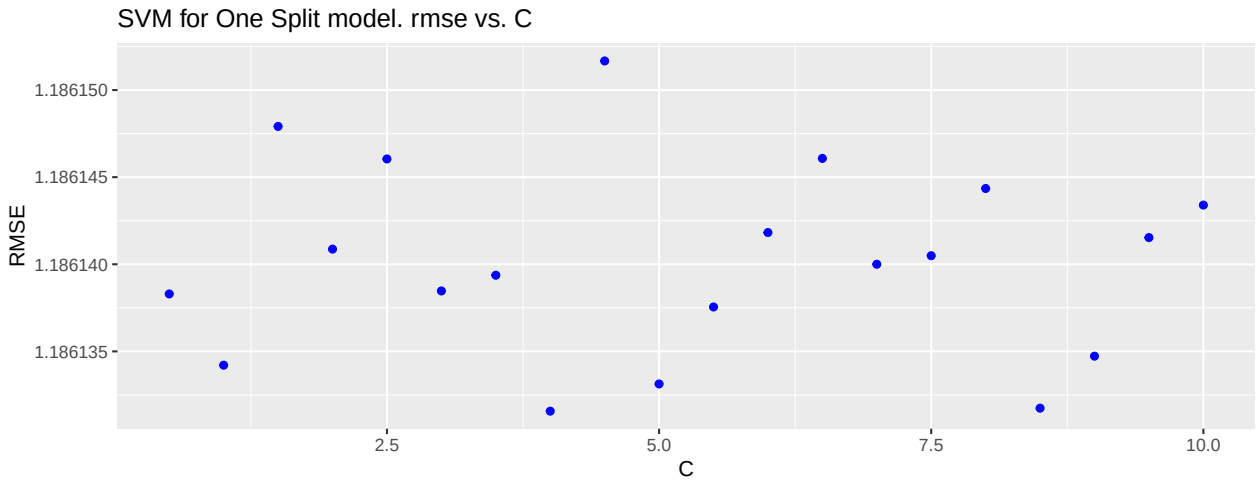

Figure 12: RMSE vs. C for One Split model - SVM

Random forest

Random forests are very popular for machine learning, but despite claims to the contrary, they don't always work with all datasets [Tang].

When using the random forest method in the caret train method, the number of trees is a parameter that the user must be supply; it isn't tuned within the model. To find an appropriate value of ntree, I used the train and test datasets. Here's my model code.

```
rf_model <- function(ntree) {
  print(sprintf("In rf_model function, ntrees=%d", ntree))
  fit_rf <- train(Goals ~ Home + WeekNumber + PointsDifference +
                          ValueDifference + ForeignDifference +
                          MeanAgeDifference + RedCardDifference,
                  method = "rf",
                  data = train,
                  metric='RMSE',
                  maximize=FALSE,
                  trControl = trainControl(method = "repeatedcv",
                                           number = 10,
                                           repeats = 3,
                                           p = 0.9),
                  tuneGrid = expand.grid(.mtry=seq(1:7)),
                  ntree=ntree)

  predict_rf <- predict(fit_rf, newdata=test)
  RMSE_rf <- RMSE(test$Goals, predict_rf)

  c(ntree, varImp(fit_rf), fit_rf$bestTune$mtry, RMSE_rf)
}
```

Using sapply, I tried various values of ntree from 1 to 200 and recorded the RMSE. Here's the chart of RMSE vs. ntree (Figure 13). The optimum ntree value was 47.

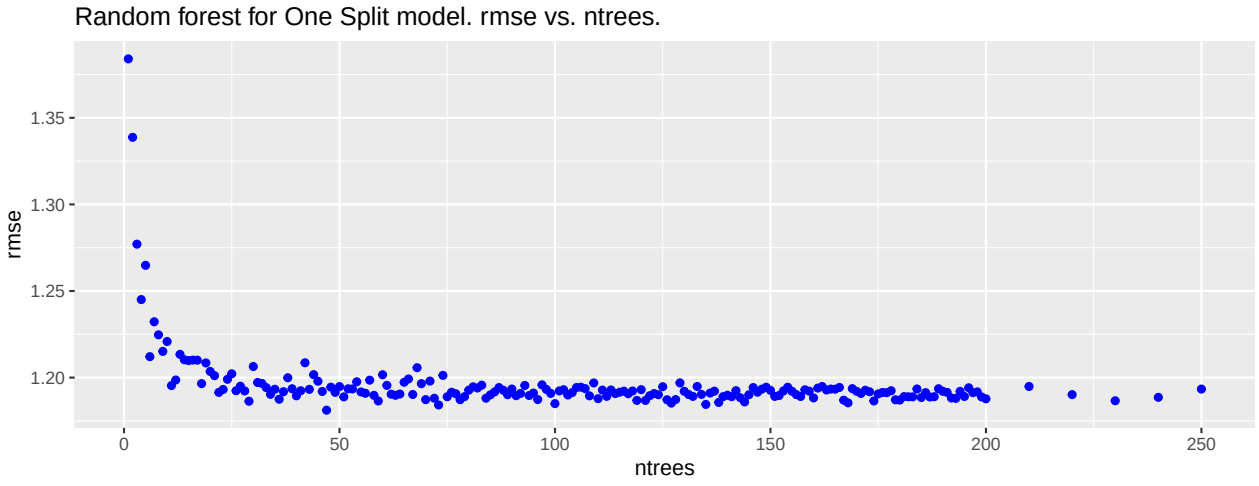

Figure 13: RMSE vs. ntrees for One Split model - random forest

At this point, we have enough data to see how features are contributing to the random forest model using the varImp function:

	Overall
Home	0.00000
WeekNumber	32.30450
PointsDifference	88.32720
ValueDifference	100.00000
ForeignDifference	31.26553
MeanAgeDifference	45.44257
RedCardDifference	32.71474
YellowCardDifference	40.70996

The Home parameter is either TRUE or FALSE depending on whether we're modeling the home team or the away team. The fact that this has an importance of zero suggests that home advantage is completely coded into the other parameters. In future modeling work, it could be removed altogether. It's no surprise that team value has such a strong influence, and as points difference encodes a team's track record for the season, this is also as expected. The mean age difference seems to have a larger impact than feature selection suggested.

Using an ntree value of 47 I calculated a final random forest model result using the holdout data. The final result was 1.1954222. This is a disappointing result as I'd expected a lower RMSE, but a clue to the performance of the model lies in the contribution of the features

Here are the summary results for the one split model.

Method	RMSE
baseline	1.248992
glm	1.176833
glmnet	1.175474
random forest	1.185541
svm	1.195422

5 Discussion and conclusion

To find model features, I evaluated nine factors and selected eight features. It's likely that there are more factors that might influence a match, for example:

- ground capacity - if spectators have an effect on performance, then more spectators ought to have a larger effect
- weather - it's possible that some weather conditions favor some teams
- injuries - if 'star' players are injured and can't play, that may well impact performance

Future work should consider these factors and more.

The data sets I used had some limitations. The foreign players data is only updated at the start of the season and may well change during the season. Team value is updated twice a month, but is only a notional value; a player's value changes over time due to age and performance. It may well be that some players are substantially over- or under-valued. Mean age is calculated on a per-team basis, but it may be that the ages of players in certain positions are more important, goalkeepers for example tend to be older; future work might consider modeling age by position (striker, defender, goalkeeper etc.). The two most obvious limitations were the number of fields and the length of the data set. Data availability limited the data set to the seasons 2011-2012 to 2019-2020, a total of 3,420 matches. There were also a limited number of fields available. Future work could search for data prior to the 2011-2012 season, and find other fields, such as ground capacity etc.

The linear models I used (glm, glmnet, svmLinear) didn't perform well with this data set. It may well be that the underlying system is non-linear, so linear approximations will fail. The random forest approach might not have had enough features to perform well. Future work should consider models which are not linear, including neural networks and fuzzy inference etc. However, a more fruitful approach might be to consider multi-dimensional target modeling which may include unsupervised methods such as archetype analysis [Cutler] and XGBoost.

The top team of the 2015-2016 Premier League Season was Leicester City, a surprising result given the relative cheapness of the team and their unimpressive previous track record. At the start of the season, bookmakers were quoting odds of 5000-1 for Leicester City to win. Leicester's win points out that soccer is not deterministic and surprising results do happen, even over the course of an entire season. This suggests there may well be limits to the predictive power of any machine learning model. In other words, there may be an RMSE floor for EPL predictions.

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

[Cutler] Adele Cutler, Leo Breiman, Archtype analysis, Technometrics, 36(4), pp3380347

[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

[Tang] Cheng Tang, Damien Garreau, Ulrike von Luxburg, When do random forests fail?, 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montreal, Canada

[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 the mean is:

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

and the standard error is:

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

the 95% confidence interval is:

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

Fields used in data analysis

Analysis and modeling is done with the data frame match_results. Here are the fields in the match_results data frame.

Field	Explanation
Date	Date of the match in IS08601 format
AwayTeam	Away team name
HomeTeam	Home team name
HomeTeamAbbreviation	Home team abbreviation
AwayTeamAbbreviation	Away team abbreviation
FTHG	Full time home goals. How many goals the home team scored at the end of the match.
FTAG	Full time away goals. How many goals the away team scored at the end of the match.
FTR	Full time result. Home win 'H', draw 'D', away win 'A'.
HR	Home team red cards
AR	Away team red cards
HY	Home team yellow cards
AY	Away team yellow cards
HomeTeamValue	Transfer value of home team players
AwayTeamValue	Transfer value of away team players
HomeTeamSquadSize	How many players on the home team
AwayTeamSquadSize	How many players on the away team
HomeTeamForeignPlayers	How many foreign players on the home team
AwayTeamForeignPlayers	How many foreign players on the away team
HomeTeamMeanAge	The mean age of the home team squad
AwayTeamMeanAge	The mean age of the away team squad
HGD	The goal difference for the home team
AGD	The goal difference for the away team

For modeling, I add more fields. Here are the added fields.

Field	**Explanation**
WeekNumber	The week number of the season (starts from 1)
AwayPriorMeanCumPoints	The points the away team has prior to the match
HomePriorMeanCumPoints	The points the home team has prior to the match
AwayMPCR	The mean red cards for the away team before the match
HomeMPCR	The mean red cards for the home team before the match

How to run the software

The software *must* be run in the order below.

1. Run EPL-Downloads.R. This will create the necessary folders and download and scrape data from the web. It downloads several hundred files and takes about 30 minutes to run. Note: it's good practice to download data as little as often as a courtesy to website owners. Downloading data too frequently may result in a ban.
2. The next step is cleaning. Run the file EPL-Cleaning.R.
3. Data analysis is done by the file EPL-DataAnalysis.R.
4. Finally, modeling is performed by EPL-Model-OneSplit.R - note this model takes over 30 hours to run.