English Premier League Exploratory Analysis

Jai D. Bansal

I've played soccer for over a decade, so I thought it would be fun to do some analysis on professional soccer matches. I was able to find English Premier League match data from the 2003/2004 season to the 2015/2016 season (*http://www.football-data.co.uk/englandm.php*)

This document contains results, plots, and my own interpretation. The underlying code is suppressed for clarity but can be viewed in *premier_league_exploratory_analysis.Rmd*.

DATA SAMPLE:

Here's the first few rows of data and what all the columns stand for. This is a pretty rich data set containing not just final scores but also some deeper metrics.

I've removed columns that aren't used in the analysis (half time goals and results, total shots, and some betting odds). These might be used in future iterations.

```
##          Date   HomeTeam  AwayTeam FTHG FTAG FTR  Referee HS AS HST AST HC
## 1: 2003-08-16    arsenal   everton    2    1   h m halsey 11 13   5   7  6
## 2: 2003-08-16 birmingham tottenham    1    0   h r styles 10 15   5   7  1
## 3: 2003-08-16  blackburn    wolves    5    1   h j winter 25  8  13   5  6
##    AC HF AF HY AY HR AR season_num
## 1:  9  8 15  1  3  1  1          1
## 2:  4 20 27  3  5  0  0          1
## 3:  2  8 14  1  1  0  0          1
```

The abbreviated columns stand for:

***FTHG**: home team goals at end of match
* **FTAG**: away team goals at end of match
* **FTR**: match result ([h, a, d] denote [home team victory, away team victory, draw] respectively)
* **HST**: home team shots on target
* **AST**: away team shots on target
* **HC**: home team corner kicks
* **AC**: away team corner kicks
* **HF**: home team fouls
* **AF**: away team fouls
* **HY**: home team yellow cards
* **AY**: away team yellow cards

* **HR**: home team red cards
* **AR**: away team red cards
* **season_num**: added in *data_aggregation_and_cleaning.R* to indicate which season a match occurred in

SUMMARY STATISTICS:

- **First Match Date**: 2003-08-16
- **Last Match Date**: 2016-04-10
- **Number of Seasons**: 13
- **Number of Matches**: 4883
- **Number of Teams**: 38
- **Number of Referees**: 50
- **% of Matches Won by Home Team**: 46.14
- **% of Matches Won by Away Team**: 28.12
- **% of Matches Ending in a Draw**: 25.74

RELEGATION:

At the end of each season, the bottom 3 teams in the standings are relegated to a lower division (the English Football League Championship). The English Premier League (EPL) is the top English league, so teams cannot be "promoted" out of the league. Below is the distribution of how many seasons the teams in the dataset have spent in the EPL.

```
##    Seasons Teams
## 1        1     6
## 2        2     2
## 3        3     3
## 4        4     4
## 5        5     2
## 6        6     3
## 7        7     1
## 8        8     2
## 9        9     3
## 10      10     2
## 11      11     1
## 12      12     1
## 13      13     8
```

Out of 38 teams, 6 (15.8%) only spent 1 year in the EPL. 8 teams (21.1%) spent 13 years, the longest time possible in this data set. The average number of seasons in the EPL is 6.8. It's interesting that relatively many clubs played 13 seasons and 1 season in the EPL, with relatively few clubs playing the number of seasons in between.

There are many teams that break into the EPL from the division beneath...and are promptly relegated the following year. Getting to the EPL is apparently only the first part of the battle.

On the other end, there are a group of dominant teams that are rarely relegated. Looking at the clubs that played all 13 seasons in the EPL (Arsenal, Manchester United, Liverpool, Chelsea, Everton, Manchester City, Tottenham, and Aston Villa) reveals many well-known (and wealthy) teams.

Further Exploration:

- What are the differences between teams that were promoted to the EPL and stayed there as opposed to teams that were promoted and then relegated the next year?
- What features distinguish the teams that are rarely/never relegated? Are they in this data set or external data?

REFEREES:

Let's take a closer look at the people we love to hate. I'm interested in seeing summary statistics about the refs as well as whether some officials are more stringent/lenient than others. Recall that there are 50 referees in this data.

Minimum # of Matches Officiated: 1
Median # of Matches Officiated: 52.5
Maximum # of Matches Officiated: 340

In terms of experience, there are a wide variety of officials. An EPL season for a team is roughly 38 games. Using this metric, the median referee has worked for 1.38 seasons and the most experienced referee has been through 8.95 seasons.

For the rest of this section, I'll only look at referees with 10 or more officiated matches to avoid outlier officials with little experience. There are a few officials who only have 1 match in the data but are near the top or bottom in terms of fouls/yellow cards/red cards handed out.

Minimum Avg. Fouls Called per Match (Experienced Officials): 20.74
Median Avg. Fouls Called per Match (Experienced Officials): 23.36
Maximum Avg. Fouls Called per Match (Experienced Officials): 29.12

The variation in average fouls per match doesn't look too extreme to me. Soccer fouls are quite common and often tactically necessary.

Minimum Avg. Yellow Cards Given per Match (Experienced Officials): 1.73
Median Avg. Yellow Cards Given per Match (Experienced Officials): 3.16
Maximum Avg. Yellow Cards Given per Match (Experienced Officials): 3.68

This distribution is, in absolute terms, much smaller than fouls called. However, considering that 2 yellow cards gets you ejected from a match, these results are more noteworthy. I find it surprising that the official at the top of the distribution is often handing out enough yellow cards to fully eject 2 players per game.

Minimum Avg. Red Cards Given per Match (Experienced Officials): 0
Median Avg. Red Cards Given per Match (Experienced Officials): 0.15
Maximum Avg. Red Cards Given per Match (Experienced Officials): 0.3

Red cards, are perhaps reassuringly, pretty rare. Note that I'm not sure how this red card data is tallied. Particularly egregious fouls get an automatic red. But I'm unsure if a player's second yellow card is counted as a red.

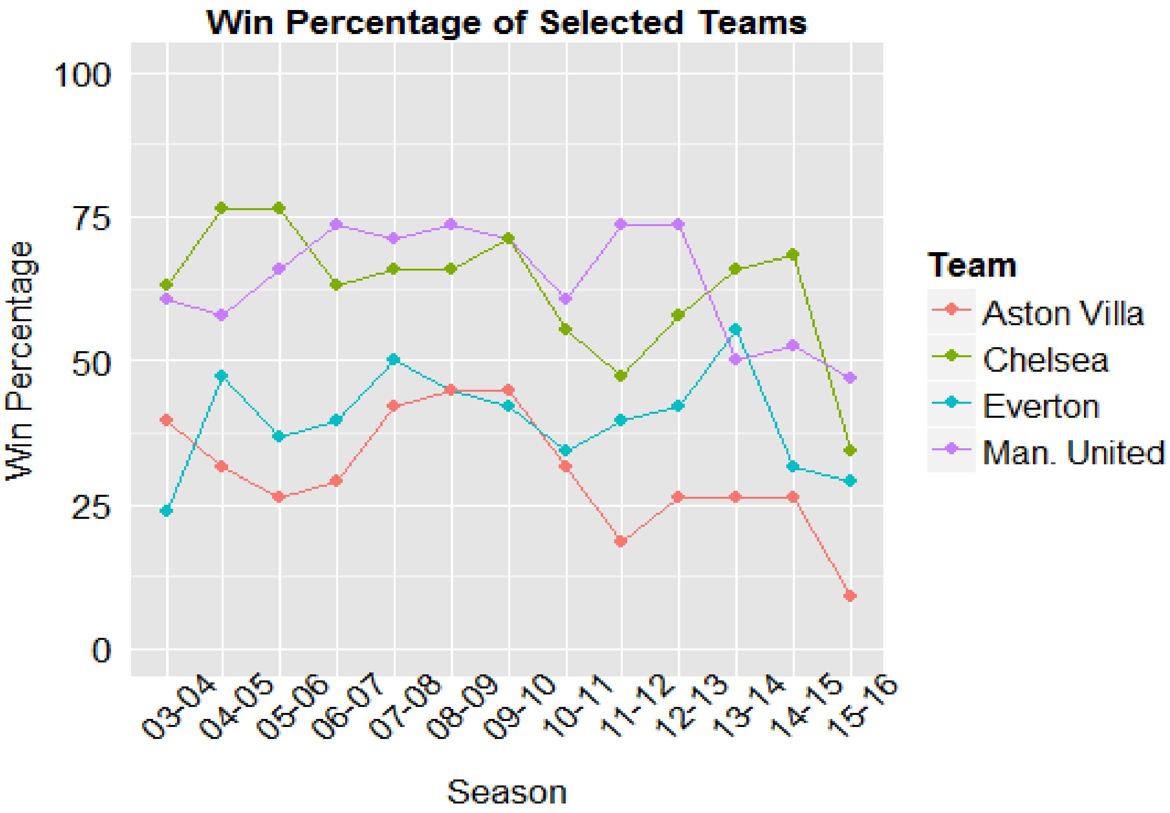
Further Exploration:

- Is there a correlation between referee experience and the number of fouls/yellow cards/red cards handed out?
- Are the referees at the top/bottom of the foul/yellow card/red card distribution actually more strict/lenient? Or did they just end up officiating more matches betweens teams that were more likely to commit fouls?
- Do officials besides the main referee have any effect on these foul and red/yellow card values?

BEST OF THE BEST AND WORST OF THE BEST:

I want to look at the winning percentages over time of a subset of teams. In particular, I use 4 teams so the plot is readable. I also only use teams that have been in the data for the entire 13 years. This is an important bias. Teams that have never been relegated are consistently above a certain quality. Simply put, they are the best teams. Out of these best teams, I look at the top (best of the best) and bottom (worst of the best) 2.

Win percentage is defined as (# of wins / # of matches).



Given that none of these 4 teams are ever relegated, I find it surprising that Everton and Aston Villa (especially Aston Villa) consistently win less than 50% of their matches. It seems that these clubs haven't done well recently, but haven't done poorly enough to be relegated. Finally, it's interesting that all 4 teams do worse, in some cases much worse, in the 2015-2016 season than the 2014-2015 season.

Further Exploration:

- Conduct this analysis with more teams.
- Why do these teams all do worse in 2015-2016 than in 2014-2015? What teams did well instead and why?

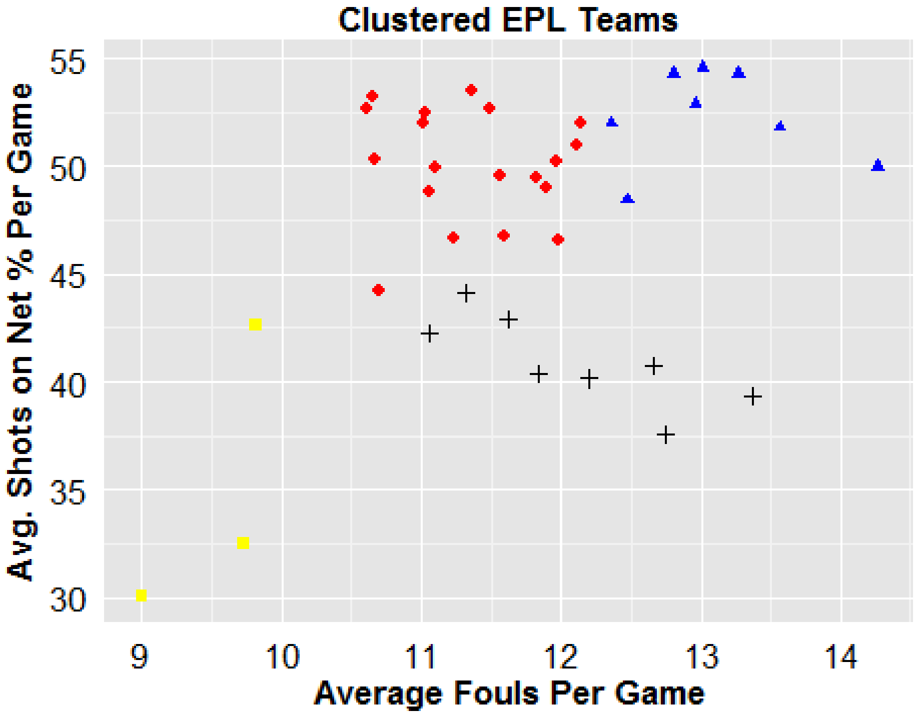
CLUSTERING BY PLAYING STYLE:

I think it's interesting to think about what the data tells us about how different teams play. There are a variety of features relating to what could be called a team's playing style.

Below, I use K-means to cluster all 38 teams in the data by average fouls per game and average shots on net percentage per game. The latter is computed by dividing the shots on net by total shots. I only use 2 dimensions to allow easy visualization. Since different teams have spent different amounts of time in the EPL, it's important to use average quantities as opposed to totals.

Numerical and visual descriptions of the clusters are below.

```
##    Average Fouls Avg. Shots on Net %
## 1:          11.4                50.1
## 2:          13.1                52.3
## 3:           9.5                35.1
## 4:          12.1                40.9
```



The blue cluster represents teams with many fouls and high shot accuracy. The red cluster commits less fouls than the blue cluster and has a slightly lower shot accuracy. The black cluster is characterized by a lower shot accuracy. The yellow cluster admittedly looks like a miscellaneous clean-up cluster, but is generally characterized by low shot accuracy and fouls.

Further Exploration:

- Add more dimensions to the clustering.
- See if certain playing styles are correlated with winning/losing records.
- Do teams that never get relegated tend to have similar styles?
- Do teams that are promoted into the EPL tend to have similar styles?

MATCH OUTCOME PREDICTION - CAN I PREDICT WINNERS?:

This might be the most interesting question that can be asked with sporting event data: can I predict the winner?

I'll use a random forest to predict match outcomes (home win, away win, draw).
I use features (wins, average goals per match, average on-target shots per match, average corner kicks per match, average fouls per match, average yellow cards per match, average red cards per match) computed for both teams from the **same season** only.

For example, suppose team A plays team B in season 2. I would compute all features for both teams using only previous matches **in season 2**.
This method means I cannot make predictions for a team's 1st match of the season.
For a team's 2nd match of the season, I will make predictions using only data from that team's 1st match. For a team's 10th match, I obtain predictions using data from the 1st 9 matches.
So, each team's features are being updated every match.

Using random forest out of the box, I obtain:
* training set error of **0%**
* OOB (out-of-bag) sample error of **49.94%** (I use this as a cross-validation error)
* test set error of **51.47%**

The baseline model I can compare against is predicting a home team win every time (yielding a test set error rate of **53.51%**).

So, the random forest model is slightly better than always predicting the home team wins.
Now, I'll try to beat the initial model using feature selection and parameter tuning. The training set error (0%) indicates I have an overfitting problem.

Backwards recursive feature selection indicates that using 12 features is slightly better than using the entire feature set.
I'll use this 12 feature model for parameter tuning.

Note that I ran feature selection code and noted the results, but it is now commented out and included for reference. It makes the generation of this document take way too long.

The 2 excluded features are *away_avg_foul* and *home_avg_yellow*.

I tune 3 parameters: # of trees, # of variables considered at each split, and the minimum number of observations in terminal nodes.
Parameter tuning shows that the best values for [ntree, mtry, nodesize] are [251, 3, 20] respectively. This specification results in a test error of **47.82%**, a modest improvement over the initial model.

The training error for this specification is ~23% (as opposed to 0% for the original model), but obviously the refined model generalizes better.

Note that I ran parameter tuning code and noted the results, but it is now commented out but included for reference. It makes the generation of this document take way too long.

So, what's the answer to the question that started this section? Yes, technically. The refined model does better than the null model, chance, and the initial model. But would I bet my life savings on it? Definitely not.

Further Exploration:

- Run parameter tuning with more parameters
- Other modeling approaches
- Incorporate new features and domain knowledge (possibly betting odds, features that represent injuries, team wealth, etc.)
- Fix matches to ensure 100% accuracy