**How to Win a Professional Soccer Game**

(Bribe the Refs)

**Soccer: One Goal = 1 Red Card = 2 Yellow Cards**

Good Acting, Bad Refereeing, or Something Else?

Soccer players are known for being great actors. Their facial expressions and body language don’t equal the severity of their pain, but a first-time viewer can’t tell the difference. Even experienced fans and referees can be fooled. In the flurry of a scramble, with arms flailing and legs twisting, it can be difficult to see what happened. The player slides to a dramatic fall, grimaces, and clutches his hamstring in sheer agony. One minute later they are back in action. Some viewers consider them prima-donnas, sissies, or fakers.

In the end, soccer players get a bad rep for their acting.

Why do they do it? Why do summer league kids act tough after rough play, while grown men, professional athletes, strive for the opposite?

This analysis doesn’t strive to define the acting abilities of professional soccer players. Instead, it attempts to codify a phenomenon that soccer players seem to know intuitively – a single card can change the outcome of a match. On a broader scale, this analysis attempts to gauge the impact of other events over the course of a season, such as number of shots on goal, corner kicks, and crosses.

**Business Problems**

Accurately gauging these events can affect several industries. Most notably, it can affect a team’s draft picks and contracts for new players by monetizing which skills are most valuable. In the gambling world, it can affect the odds that betting houses and gambling websites place on matches. In a further reaching sense, it might also affect post-game entertaining staffing. A win means more celebration at home after the game.

For the purposes of this study, we will focus on betting houses, as the impact of wins and other game events is most immediately felt by their business.

**Introduction to the Problem**

We have limited ability to accurately forecast the outcomes of soccer matches. Paradoxically, the unpredictability of soccer matches makes the business of betting possible. Gamblers try their luck at predicting outcomes, and betting houses set odds that are profitable.

Fortunately, a lot of data is available for soccer matches, from goals and penalties per match to individual player attributes. Part of the problem will be to determine which data are significant and which are not.

**Introduction to the Data**

The data originally comes as an SQLite database with eight tables, which are converted into eight separate data frames:

Tables

|  |  |  |  |
| --- | --- | --- | --- |
| Player | **team** | **league** | country |
| Playerattributes | teamattributes | **match** | Sqlitesequence |

Our analysis primarily focuses on the bold tables. The other data frames contain useful information, such as which league ID’s belong to which country. However, they are small and require little analysis. The focus data frames are larger and require some restructuring.

The most important of these data frames is ‘match’. It contains individual match information, such as number of goals scored, fouls committed, and length of ball possession.

Data Frame: ‘match’

The variables in match can be divided into four categories – Base Data, Player Data, XML Data, and Betting House Odds. We used the Base Data to do a high-level exploration. We used Base Data and XML Data to do further analysis. Restructured and calculated data is in **bold\***:

|  |  |  |  |
| --- | --- | --- | --- |
| Base Data**\*** | Player Data | XML Data\* | Betting House Odds |

|  |  |
| --- | --- |
| country\_id | league\_id |
| season | stage |
| date | match\_api\_id |
| home\_team\_api\_id | away\_team\_api\_id |
| home\_team\_goal | away\_team\_goal |
| **goalDiff** | **points** |

|  |  |  |
| --- | --- | --- |
| goal | **goals** | **oppgoals** |
| shoton | **shotson** | **oppShotson** |
| shotoff | **shotsoff** | **oppShotsoff** |
| foulcommit | **fouls** | **oppFouls** |
| card | **Ycards** | **oppYcards** |
|  | **Rcards** | **oppRcards** |
| cross | **crosses** | **oppCrosses** |
| corner | **corners** | **oppCorners** |
| possession | **homePoss** | **awayPoss** |

Base Data:

Our highest-level analysis comes from this data frame. The dataframe 'statsMB3' contains only these variables, plus two calculated variables - goalDiff and points. It is used to show that a higher goal differential correlates to more points accumulated in a season. Because a better goal differential results in a better season outcome, we continued by focusing on what causes a better goal differential, which is the difference between goals scored and goals permitted.

XML data: The XML data allowed us to focus on what causes a better goal differential. These variables possess details of shots and gameplay. Because the number of shots and plays can vary from game to game, it is stored in nested data frames which are not easily read by R. We later explain the packages and techniques necessary to extract the desired data from here.

Player Data: This data shows the position on of first, second, and third string players. The variables are coordinates on the plane of a soccer field. This was not used for our analysis.

Betting house odds: These numbers show the odds placed on games by various betting houses. Because we want to find probabilities of wins ourselves, we don’t use these numbers in our analysis.

Limitations

This is a relatively involved data set. It has 198 variables between 8 data frames, not including calculated fields or data within nested XML data frames. The nested XML data frames contain rich information such as ball possession by team, goal types, and players who scored.

While any data set’s limitations can be defined by the data it does not have, this data set is a case where the limitations are better ascribed to the analyst’s creativity and by data that exists but isn’t complete for all observations. For example, the match data frame has ~25,000 rows. However, only 14,217 of those rows contain XML data, and 8,125 of those rows contain complete XML nested data [[check how many more rows were removed with the “/corners “ calculation]]. In general, the earlier seasons in this data set do not contain the nested XML data. This limits how far back we can go in our analysis, which may restrict seeing higher-level trends.

Furthermore, this data set doesn’t contain information on league rules, match referees, or other external events, such as league policy changes or weather during the soccer matches. While this data has the benefit of already being rich, it could be further improved by including these data.

Cleaning and Wrangling

Despite being rich, this data set is intended to thwart practitioners of R. It is stored in SQLite format, and many interesting variables are stored in nested XML data frames within individual observations. This requires a combination of R packages - RSQLite, DBI, XML, dplyr, and magrittr.

The initial transformation involves reading the SQLite data into R and creating a variable for each table. R plays nicer with csv files, so I wrote each table to a csv file that was stored in my project folder. I then created a variable for each csv file and read it back into R. This allowed me to move forward with base R and common R packages.

At this point, it was easy to select the non-XML data with dplyr to create simple plots and regressions. Extracting the desired data from the XML data was more involved. It required converting the XML data into a data frame and extracting the desired data with the following process:

1. Use dplyr to select the desired columns
2. Load library(XML)
3. Remove all rows with NA or incomplete data:

# removes all rows in matchAD1$possession with NA

test <- test[complete.cases(test$possession),]

# remove all rows where test$possession contains only "<possession />"

test <- test[!(test$possession == "<possession />"),]

# removeall rows where test$possession contains only "<possession />"

test <- test[!(test$card == "<card />"),

testALL <- testALL[!(testALL$corner == "<corner />"),]

1. Create for loops that extracts the necessary data into a new variable. All for loops contain elements of the dplyr package, the XML package, and base R conditional coding.
   1. All XML columns do not convert into clean data frames. Many data values are in the wrong column. This may or may not be a result of the parser. However, I mitigated this within the for loop for all variables except “possession” by using “|”. Given additional time, it would be prudent to find a more efficient parser.
   2. The “possession” variable also produced incorrect variables, and the output of the for loop contains excess characters that need to be removed. I mitigated both of these issues outside of the for loop using substr() and magritrr().
   3. Six for loops required a conditional statement, because some observations have a missing column

These for loops take about 12 hours to run.

Note: In the early stages, I also relied heavily on plyr while wrangling the XML data, because much of the helpful online documentation provided plyr examples. I later found and implemented dplyr equivalents to prevent running both packages simultaneously.

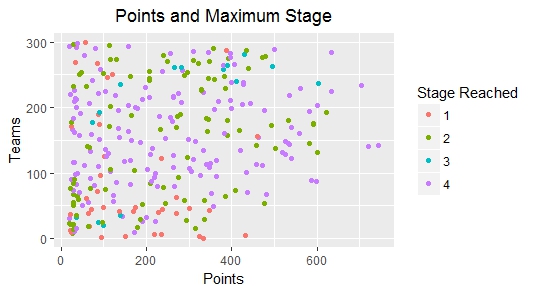
Preliminary exploration

Before I unpacked the XML data, I created various plots to see if teams with more points or goals proceeded to further stages.

ggplot(data = statsMB3, aes(spoints, statID, col = factor(maxStage))) + geom\_point() +

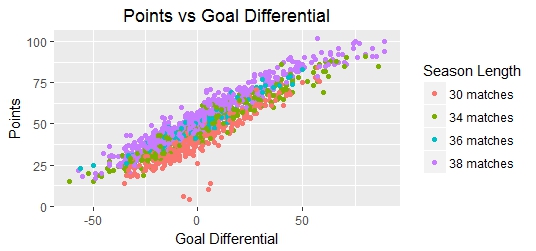
labs(title = "Points and Maximum Stage") + theme(plot.title = element\_text(hjust = 0.5)) +

labs(x = "Points", y = "Teams")



This plot suggests a very loose association between number of points acquired and stage reached. On further analysis, I found that the “stage” variable represented the match number in the season. For example, the teams in “Stage 3” were playing their third match of the season. The seasons varied in length from 30 matches to 38 matches (half at home and half away) for each league.

While this nullified any stage-progression style analysis, it still allowed to compare leagues with shorter seasons vs leagues with longer seasons. The following chart shows that teams with longer seasons tend to score more points (as expected), but it also shows that the dispersion of points accumulated by teams across leagues appears similar:



This chart shows expected data – more goals scored and less goals allowed lead to more wins, which directly leads to more points acquired.

The interesting part begins here. Why do certain teams score more goals, win more games, and acquire more points? This is where the nested XML data comes in handy.

Starting with summary statistics, we can see how many games are played, how often home teams are carded versus away teams, and how often winning teams are carded versus losing teams.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Poss% | Games | YCards | RCards | Fouls | Goals | Wins | Ties | Loss | Crosses | CKicks | ShOn | ShOff |
| Home | 48.77 | 8124 | 14371 | 442 | 99654 | 12697 | 3741 | 2054 | 2329 | 149923 | 47374 | 50028 | 50739 |
| % T | 48.7% | 50% | 45.5% | 46.8% | 48.9% | 57.3% | 61.6% | 50% | 39% | 56.1% | 56.3% | 55.6% | 55.5% |
| Away | 51.23 | 8124 | 17220 | 503 | 104221 | 9471 | 2329 | 2054 | 3641 | 117108 | 36819 | 39995 | 40662 |
| % T | 51.2% | 50% | 54.5% | 53.2% | 51.1% | 42.7% | 38.4% | 50% | 61% | 43.9% | 43.7% | 44.4% | 44.5% |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | #Matches | Home | Away | Ycards | Rcards | Fouls | Goals | Crosses | c-kicks |
| Wins | 6070 | 3741 | 2329 | 10706 | 185 | 73779 | 14610 | 96495 | 32558 |
| % of total | 74.7% | 46.0% | 28.7% | 33.9% | 19.6% | 36.2% | 65.9% | 36.1% | 38.7% |
| Draws | 2054 | 2054 | 2054 | 8317 | 202 | 52918 | 4104 | 69993 | 21657 |
| % of total | 25.3% | 25.3% | 25.3% | 26.3% | 21.4% | 26.0% | 18.5% | 26.2% | 25.7% |
| Losses | 6070 | 2329 | 3741 | 12568 | 558 | 77178 | 3454 | 100543 | 29978 |
| % of total | 74.7% | 28.7% | 46.0% | 39.8% | 59.0% | 37.9% | 15.6% | 37.7% | 35.6% |

These tables show some clear information. With some simple arithmetic, we find the following:

Home Team wins

an away team is 20% more likely to be carded than a home team. The away team, ceteris parabus, is also 20% more likely to lose the match. The converse is true for a home team. They are xx% more likely to win, and xx% less likely to be carded.

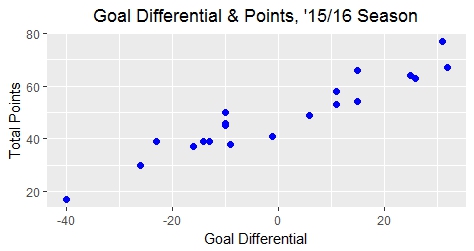
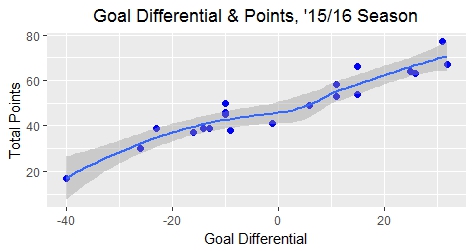
If we look within home games only, winning teams were xx% less likely to have cards against them and xx% less likely to have fouls against them

Within away games, losing teamx xx% and winning teams xx%

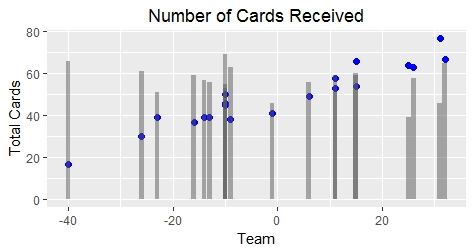
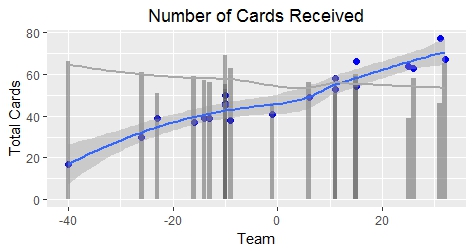
Another interesting point – we can see how many matches are won by one goal. We can then see how many of these matches include yellow and red cards. We can then compare this versus the rest of the matches (ties and differences of more than one goal) to see if those matches proportionally have more or fewer cards.

We also build on the charts that were shown earlier, for example by adding bar charts that represent other factors. An interesting chart superimposes how often a team is carded over how many points they score.

To begin, we use the English Premier league for the 2015/2016 season as an example. This small cross section of data allows us to view how certain factors *might* be related:



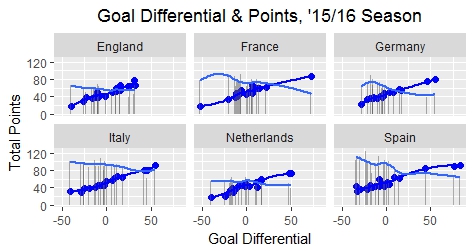
1. **Goal Differential vs Points Scored (2) With trendline**



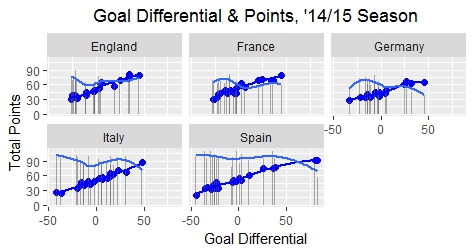
**(3) Superimposed bar chart of cards awarded to each team** **(4) With trendlines**

Charts (1) and (2) show the relationship between goal differential and points acquired. It naturally places the teams in rough order from left to right by those which won least to those which won most.

The power of data allows us to view if this league and season is an isolated occurrence, or if this happens on a broader scale. The following chart shows the 2015/2016 season again. Instead of just the English Premier League, it shows all leagues with available data. The blue trendline shows the relationship between goal differential and wins. Again, these plots naturally arrange the teams roughly from least winningest (left) to most winningest (right). The grey trendline shows the carding trend of teams from least to most winningest.

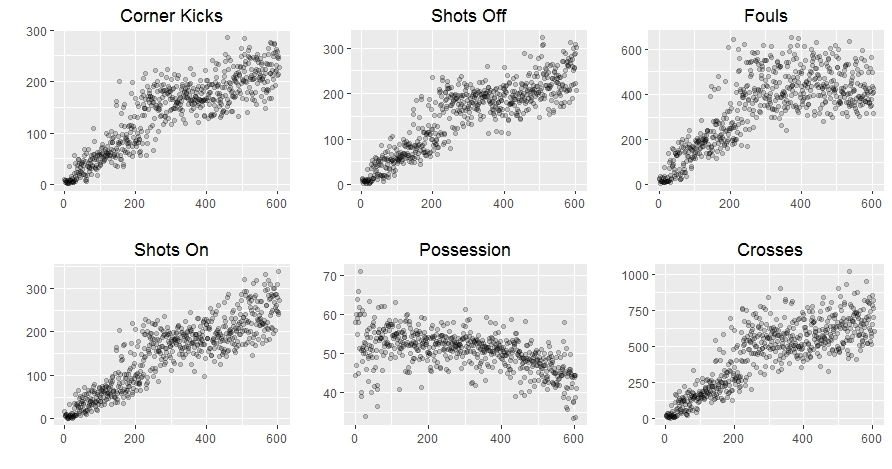


It seems to be a clear relationship. We see this same relationship in other seasons too. Here is the 2014/2015 season:



The plots speak for themselves. The winningest teams are carded the least. We cannot say whether those teams have less aggressive players or referees who call in their favor, but it warrants further investigation. This trend stays consistent across all seasons and leagues. Certain factors seem to magnify the effect, such as whether the match is a home or away match for a team, but the effect itself remains.

Do other factors show a similar relationship? We’ve included charts for all variables. In this case, we’ve ordered the teams from least to most winningest, and plotted the corresponding point.



Each point represents a unique season team. Some of these plots are expected – teams that produce more chances for goals tend to win more games. An interesting plot is possession. It seems to show that teams who win more frequently have less ball possession.

This leads us to the question: By how much do these variables affect match outcome? What is the effect of yellow and red cards, and why does less possession seem to help? This is where additional plotting and regression analysis comes in. It can show which variables are significant, and by how much they affect a match.

Our goal then is to determine how much influence a variable has on match outcome. Because match outcome is measured in the number of goals a team scores minus the number of goals they permit, we chose to run a linear regression where the independent variables are measured by their effect on goal differential. An added benefit of our data is that it is panel or longitudinal in style. This means that we can use an earlier season as our training data and a subsequent season as our test data.

The most complete data exists for these seasons:

1. One
2. Two
3. Three

So we progress as such.

Post Regression Analysis:

Perhaps it’s not surprising that these variables all have some level of contribution to a match. As much as soccer may be a game of skill, it is also a game of odds. The more crosses and shots on goal a team has, the more they score and win. The more a team can get the opponent carded, the more they will win.

**Recommendation**

These stats can be divided into two types – player-based and referee-involved:

Player-based: goals, shotson, shotsoff, crosses, corner kicks, crosses, possession

Some of these may have referee influence. For example, a red card in the goal box may lead to a penalty kick. However, the event itself still depends ultimately on the player.

Referee based: yellow cards, red cards, fouls

These events depend ultimately on the referee.

We recognize that these are not black and white. A corner kick can be a clear call, or it can be based on the referee judgement.

For betting house:

The recommendation would be: study further which players contribute most to goals and contributing factors, and base their betting rates on the team line-up and those factors. It would also be prudent to collect data on the referees and how they tend to make calls.

For the team management: collect data on match attendance and to find what contributes to game attendance. It is likely wins and goals. You can then monetize goals, the factors that contribute to goals, and by extension, the players which contribute most to these factors. This analysis can help determine which players are overvalued and undervalued.

For both betting houses and team – bribe the refs.

We originally mentioned that this analysis can be useful in several businesses. We used as examples professional sports organizations, betting houses, and soccer fan entertainment.

A third noteworthy industry is regulation in both the government and non-government realms.

Finally, it can also affect federal governments, state governments, and private regulatory bodies, which work to protect the consumer. When there is proof of fixing matches or defrauding customers, these regulatory bodies can impose fines on sports organizations or remove a gambling business’s ability to operate.