Brandon Croarkin

IST 565 – Data Mining

# Final Project – Soccer Match Analysis

## Introduction

Soccer is the world’s most popular sport. The 2010 FIFA World Cup final in South Africa was shown in every single country and territory on Earth, including Antarctica and the Arctic Circle. The in-home television coverage reached over 3.2 billion people1. Yet, despite the sport’s popularity, the data analytics rise that has impacted sports like baseball and basketball in major ways, has not yet permeated into the sport of soccer to the same degree.

While more data is beginning to be collected, there’s no real consensus on how to use the data and many coaches/managers are still reluctant to use it. Part of this is because soccer is hard to analyze. Unlike baseball and basketball that have a much more structured game, soccer is very free-flowing. This makes converting the data collected to information much harder. While goals and assists can easily be tracked, it is much harder to quantify the overall impact a player makes to a game because a good tackle or dribble can stop or create a goal, but not be reflected in the score sheet. Without this concrete data to quantify the impact of a player, soccer player rating and evaluation has come down largely to the “eye test”.

This project will aim to investigate a collection of soccer datasets including match data, team attributes, and player ratings to understand at a deeper level what influences a soccer match. This project aims to answer questions such as:

1. Do different soccer leagues have different playing styles?
2. What playing styles lead to more goals scored? What playing styles lead to less goals conceded?
3. What has the biggest impact on a soccer player’s rating?
4. How accurately can a soccer match be predicted? How do these predictions compare to the betting odds?

## Analysis

### The Data

The data used for this project was compiled in Kaggle (<https://www.kaggle.com/airback/match-outcome-prediction-in-football/data>), but is actually a collection from a few different datasets. This data includes stats on the match scores, leagues, teams, lineups, team formations, betting odds, player attributes, and team attributes. The match data was compiled across different websites and processed, the betting odds came from (<http://www.football-data.co.uk/>), and the player and team attributes came from the EA Sports FIFA games (<http://sofifa.com/>).

Overall, the dataset contains around 25k+ matches across 8 seasons (2008-2015) from 11 leagues and 2k+ player. The player data contains information for every season the player is active.

Due to the large size of the dataset, to limit the size of the data, analysis was only conducted on five leagues within the dataset: England Premier League, France Ligue 1, Germany 1. Bundesliga, Italy Serie A, and Spain LIGA BBVA. These five leagues are generally regarded as the best five leagues in Europe and arguably the world as well.

The player ratings data contains a lot of missing data due to player data that was not gathered from FIFA. All the analyses below use average ratings for players across all the seasons they played and aggregates team ratings ignoring the missing values.

Within the Match dataset only the home goals and away goals are recorded as match statistics. The result field was created by comparing the home goals to the away goals and results in three categorizations: “Home Win”, “Tie”, “Home Loss”. This variable was converted to an ordered factor variable after creation.

Model evaluation was done by randomly splitting the full match dataset into a training and test set with 90-10 split and applying the model predictions to the test set.

Since the preprocessing steps differed for each of the different analyses performed, the additional analysis-specific data preprocessing needed for each algorithm will be covered in the respective analysis sections below.

### Exploratory Data Analysis

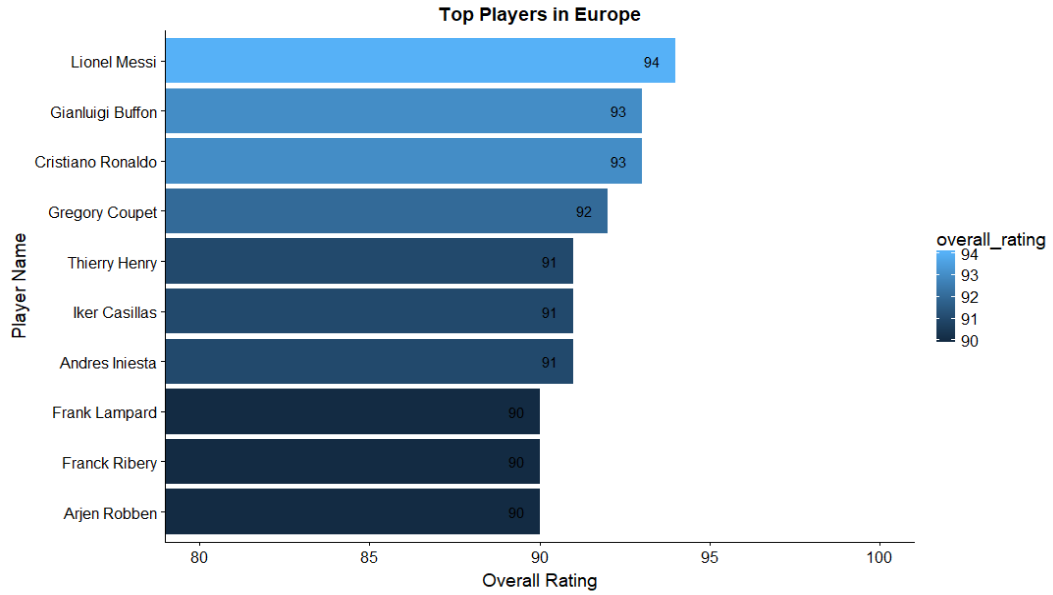
Before trying to answer the primary research questions listed above, it is important to understand the data better. The charts below each provide a new look at the data and seek to answer a specific question to gain more insight.

### 

*Figure 1:* Histogram of the age players peak.

The above graph is a histogram of the age that players peak. Answering this question has a very significant value as knowledge of this number can greatly affect the transfer value of players. Generally, most experts tend to say that this number is between 26-30, although there can be large variations between players. This is an age range where the player still maintains much of their athletic abilities, while also benefitting from years of training and match experience.

To answer this question, the player data was filtered down to those who had a minimum age less than 26 and a maximum age over 29. This was done to filter down the player data to only those who were playing during this theorized prime period. For each player, the age at where they hit their peak was recorded and then summarized in the histogram. From the histogram, the peak is evident at 30, meaning that this is the most frequent age where player’s reach their max rating. However, it is also interesting to see the wide spread and an apparent bimodal distribution with a local peak seen around 25.



*Figure 2:* Top 10 players in Europe.

The top 10 players in Europe are shown above, based on their top overall rating reached in the dataset. Lionel Messi is at the top of the list with a 94 overall rating. Interestingly, three out of these top 10 are goalkeepers (Gianluigi Buffon, Gregory Coupet, and Iker Casillas)

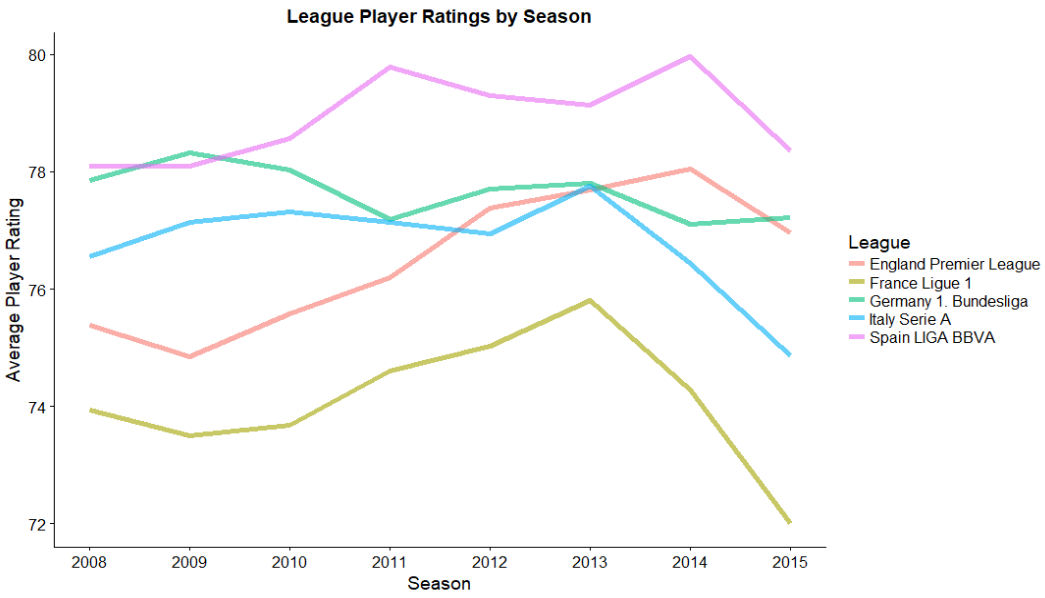
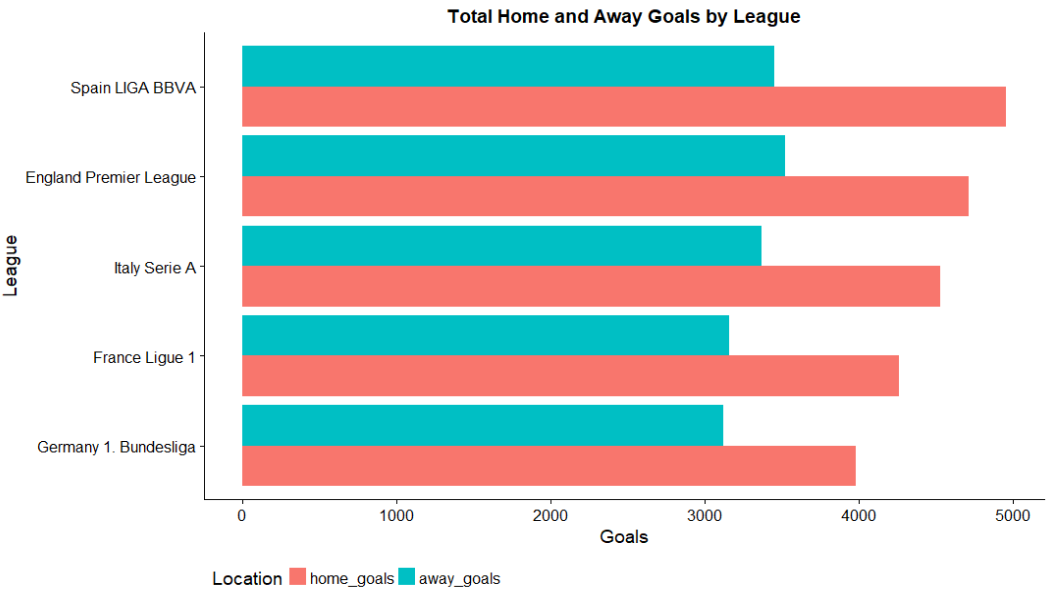


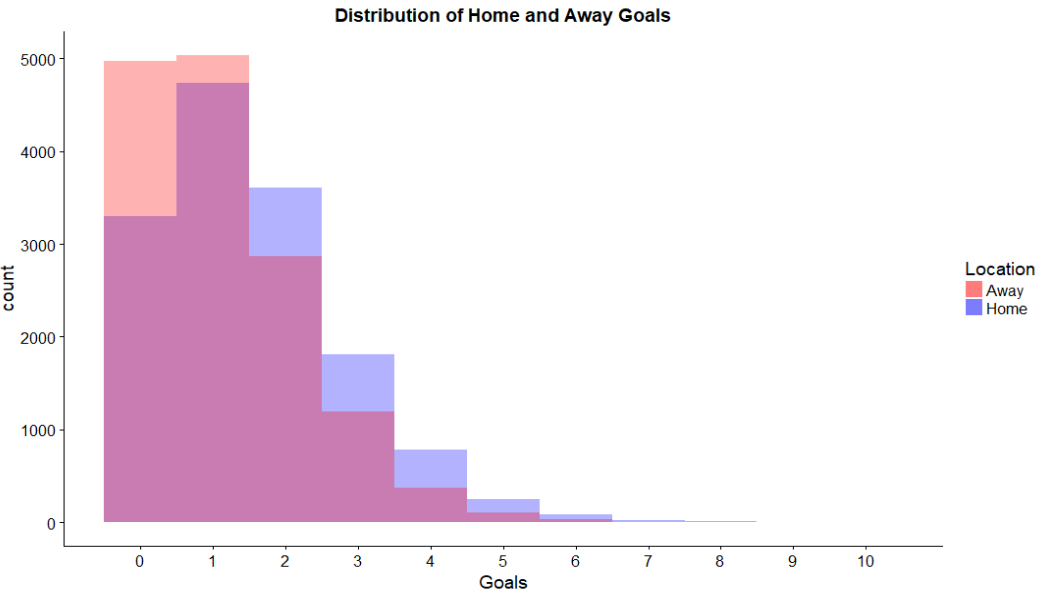
Figure 3: Average league player ratings by season.

Based on the chart above, Spain contains the best players in Europe. This was the case in every season, except for 2009, where Germany was at the top. France consistently is the lowest of these five leagues. There is also an apparent rise in player ratings from 2008 to 2013/2014 with a drop right after. England looks to be the biggest risers in the graph, going from the 4th ranked league, up to 2nd in 2014.



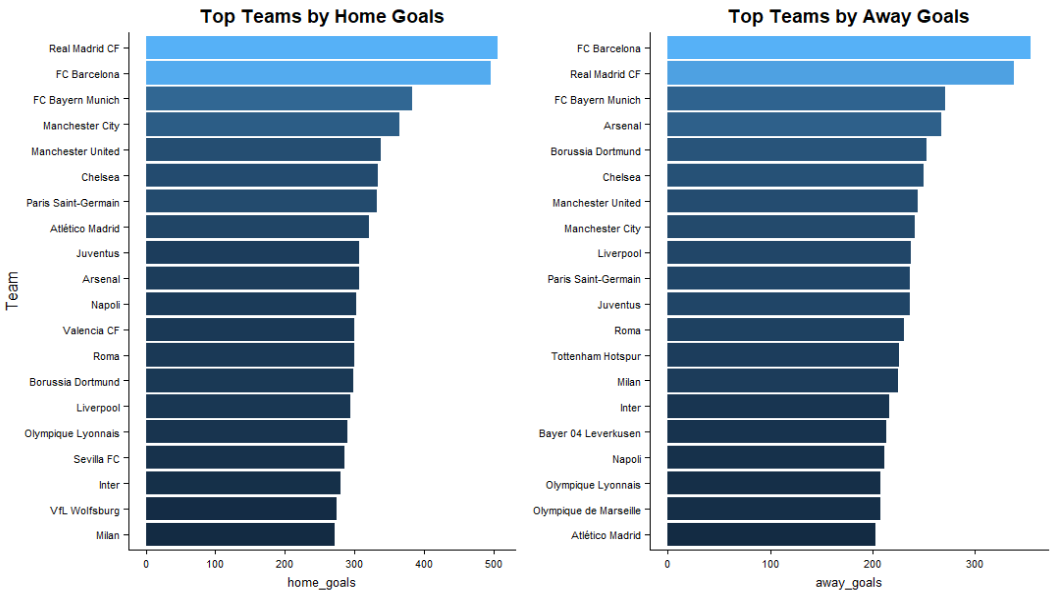
*Figure 4:* Total home and away goals by league.

There is a clear difference in the number of home goals vs. away goals in all the leagues, with the home team scoring ~35% more goals. Spain is at the top of the list with the most goals scored, while Germany is at the bottom.



*Figure 5:* Distribution of home and away goals.

This chart reinforces the difference in home vs. away goals that is show in Figure 4. While the away team only exceeds the home team in scoring 0 or 1 goal, the home team is significantly higher in all other goal counts. This is evidence that there is a clear home advantage that takes place in a soccer match.



*Figure 6:* Top 20 teams by home and away goals.

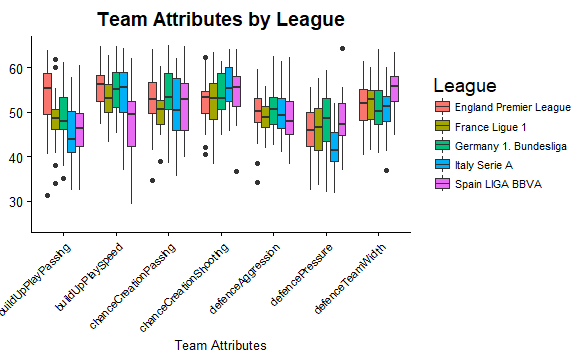
Two teams standout far ahead of the others at the top of the graphs above on total home and away goals: Real Madrid and Barcelona. This makes sense as historically, these are two of the biggest clubs in the world with the financial backing to attract and buy the best players in the world. Unlike the NBA and MLB, soccer has no salary cap or luxury taxes to ensure league parity. As a result, the big clubs consistently dominate and can buy the strong players from weaker sides.

### Answering Research Questions

#### Do different soccer leagues have different playing styles?

Many soccer watchers tend to associate different leagues across Europe with different playing styles. Spain is known to have very technical players and their playing style involves a lot of short, quick passes as opposed to direct, long balls. German playing style tends to mimic that of their country’s culture, being focused on movement, team spirit, and efficiency. These games involve less of the possession style play of Spain and include more fast attacks with organized defensive structures. Italy is known for their excellent defensive structure and use of the counter attack. Their games are known to be lower scoring games as teams often prefer to defend and keep a clean sheet than go all-out in attack. England soccer is known for the competitive, athletic style of play. Size and strength is valued much more greatly than it is in Spain and the play tends to focus on getting crosses into the box. Lastly, French football has failed to create a notable style of play. Instead their teams tend to be a mix of the styles of the leagues listed above. However, are these styles of play listed above actually reflected in the playing styles of the teams, or merely media stories? Using the Team Attributes table and cluster analysis, these hypotheses can be tested.

The Team Attributes table contains information on the build-up play speed, build-up play passing, chance creation passing, chance creation shooting, defensive pressure, defensive aggression, and defensive team width. Each of these variables includes a column with a numeric value and another with a categorical class.



*Figure 7*: Team attributes by league.

The above figure shows a boxplot of the numeric team attributes broken out by league. Just from a quick view of this graph, some trends are evident. England has the highest average build-up play passing and build-up play speed. Spain is high in chance creation shooting and defensive team width, but low in build-up play speed and build-up play passing. Italy has the lowest defensive pressure, while Germany has the highest defensive pressure. These attributes already seem to support the hypothesis of league’s having distinctive playing styles that match media descriptions.

A cluster analysis was run on each of the team’s in the dataset using the k-means clustering algorithm. The k-means algorithm assigns each of the examples to one of *k* clusters. The goal is to minimize the differences within each cluster and maximize the differences between the clusters. It does this by first assigning examples to an initial set of *k* clusters, and then updating the assignments by adjusting the cluster boundaries according to the examples that currently fall into that cluster. This is repeated until changes no longer improve the cluster fit.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Cluster Means** | | | | | | | |
| **Cluster** | **Build Up Play Speed** | **Build Up Play Passing** | **Chance Creation Passing** | **Chance Creation Shooting** | **Defense Pressure** | **Defense Aggression** | **Defense Team Width** |
| 1 | 52 | 49 | 49 | 51 | 40 | 49 | 49 |
| 2 | 49 | 43 | 51 | 57 | 50 | 49 | 56 |
| 3 | 59 | 55 | 57 | 54 | 46 | 52 | 52 |

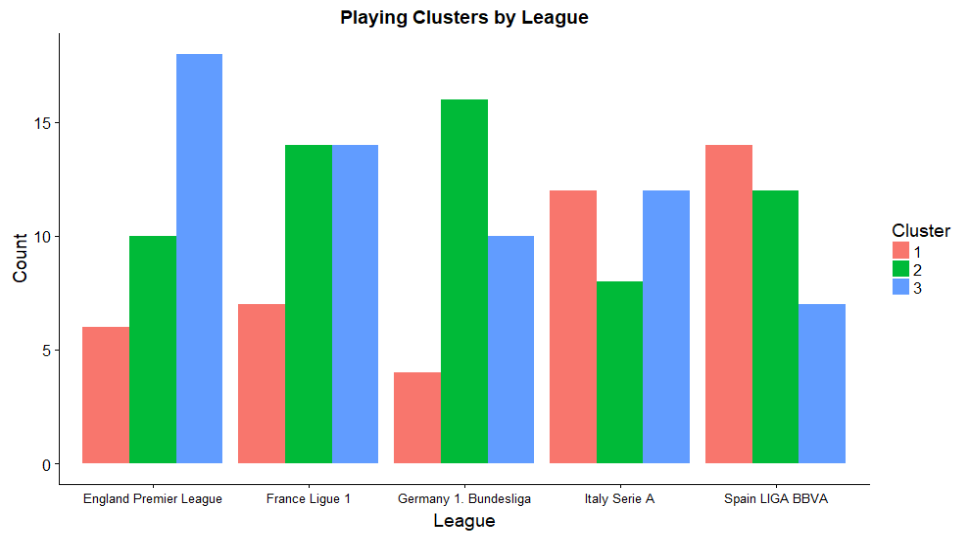
*Figure 8:* Cluster means from knn algorithm, k = 3

After running the algorithm with multiple values of k, the best algorithm result appeared to be with a k = 3. The results gave good divisions between the groups, while also making enough groups to draw interesting insights from the data.

By examining the cluster means, patterns become noticeable that distinguish the clusters from each other. Cluster 1 is characterized by narrow defensive team width, low defensive pressure and aggression, and low chance creation passing and shooting. This cluster consists of teams that are compact and have a low defensive line. This is typically characteristic of weaker teams that sit back against stronger opposition and test them to break down their defensive lines. They typically do not get many chances in a game but hope to convert on their few chances while not allowing the other team to score. Cluster 1 is “Compact and Defensive”.

Cluster 2 has low build up play and passing, but high chance creation shooting, defensive team width, and defensive pressure. This cluster consists of direct, counter-attacking teams. They put high pressure on the ball in the opponent’s half and hope to intercept the ball and quickly create opportunities on goal. Cluster 2 consists of “Direct, counter-attacking” teams.

Cluster 3 has high build up play and passing, high chance creation passing, and high defensive aggression. This style of play involves a lot of passing and likes to maintain possession and slowly move the ball up the field. They use this possession to methodically break down the opposition’s defensive structure and wait for good opportunities to strike. Cluster 3 appears to be a “Possession” cluster.



*Figure 9:* Team cluster assignments grouped by league.

By analyzing the distribution of team cluster assignments by league, some interesting insights can be made. England has a high concentration of the “Possession” cluster. This high number is interesting, given that England is not frequently associated with being a very possession-oriented league. However, this number does also make sense given that since England soccer does like to use build up passing to get the ball to the corner to put in crosses. The small number of teams in the “Compact and Defensive” could also be an indication of the competitive nature of the English league where there is not a large distribution of skill from top to bottom in the league.

Spain has the highest concentration of “Compact and Defensive” teams. This is an interesting finding given that Spain is known for their passing-style of play. However, this could potentially be explained by the large skill gap between the top teams in Spain and the rest of the league. While England is known to be a very competitive league, Spain is known to have a few powerhouse teams that consistently win the league. Consequently, many teams may be adopting this defensive mindset when playing against these stronger teams.

Germany has the most teams in the “Direct, counter-attacking” cluster. This finding makes a lot of sense and supports the theories of Germany being a very direct, fast style of play. Germany also has a very low number of the “Possession” cluster teams. These cluster results support the notion of Germany having a very fast-paced and direct style of play.

Italy has an equal number of teams in the “Company and Defensive” cluster and the “Possession” cluster. This finding also supports some of the stories of Italy’s style of play. They are compact defensive teams that prefer to play with a defensive mindset. When they attack, it is a methodical build-up from the back. Their games are not very fast paced, and this is supported by the low number of “Direct, counter-attacking” teams.

France has an equal number of teams in the “Direct, counter-attacking” cluster and the “Possession” cluster. Their low number of teams in the “Compact and Defensive” is interesting and seems to suggest that a lot of the games in France are open with a lot of passing and quick attacks.

Overall, the cluster analysis tends to support a lot of the stories on the different leagues styles of play.

#### What playing styles lead to more goals scored? What playing styles lead to less goals conceded?

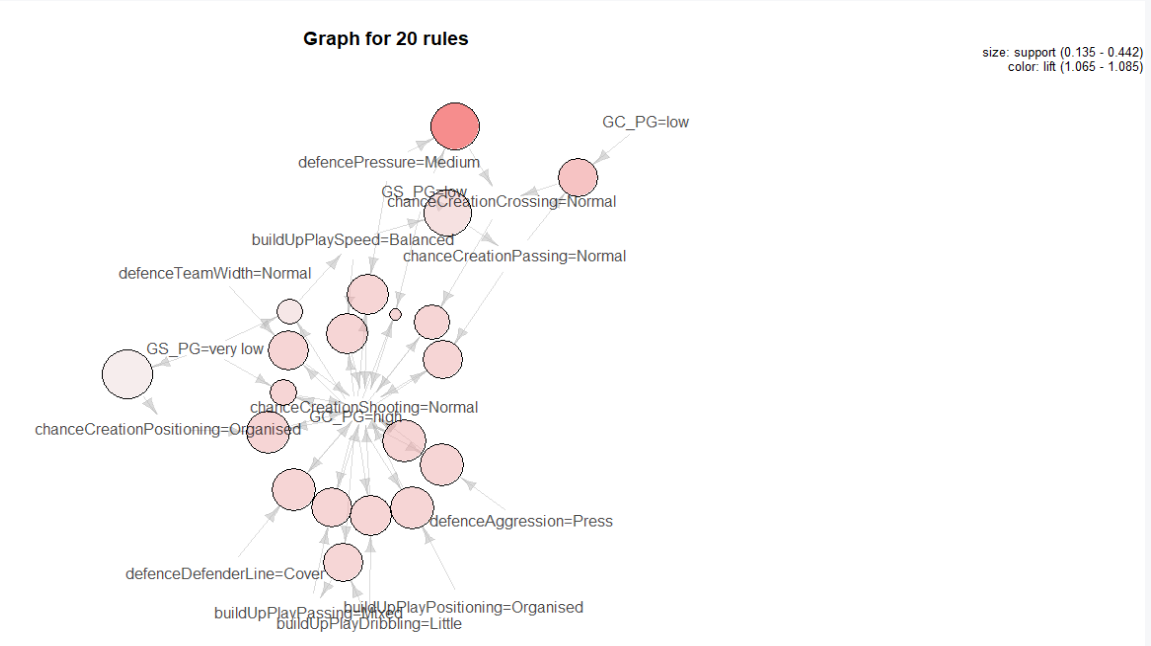
To determine what playing styles lead to more goals scored and less goals conceded, a data frame was created with each of the teams and their average goals scored per game, average goals conceded per game, and all of the categorical team attributes. However, to perform association rule mining, all data needs to be in a categorical format. To accomplish this, the goals scored and goals converted columns were converted to categorical data by breaking them into bins. This was done using 4 equally-spaced intervals from the maximum value to the minimum value. Splitting based on intervals versus bin frequency makes the most sense for this data since the size difference in goals scored/conceded is important, so by using an interval division more of this is captured. However, this also means that the bins had uneven numbers of examples.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **LHS** | **RHS** | **Support** | **Confidence** | **Lift** | **Count** |
| {GS\_PG=low, defencePressure = Medium} | {chanceCreationCrossing = Normal} | 0.42 | 0.96 | 1.1 | 69 |
| {GC\_PG=low, chanceCreationPassing = Normal} | {chanceCreationCrossing = Normal} | 0.36 | 0.95 | 1.1 | 58 |
| {GC\_PG=high} | {chanceCreationShooting = Normal} | 0.39 | 1 | 1.1 | 63 |
| {GC\_PG=high, chanceCreationPositioning = Organised} | {chanceCreationShooting = Normal} | 0.39 | 1 | 1.1 | 63 |
| {GC\_PG=high, buildUpPlayPositioning = Organised} | {chanceCreationShooting = Normal} | 0.39 | 1 | 1.1 | 63 |

*­Figure 10:* Top 5 association rules with maximum length of 3, sorted by lift.

The table above shows the top 5 association rules created, sorted by lift. Rules were filtered down to only include rules with a maximum length of 3 so rules were more understandable. The minimum support level was set at 0.1 and the minimum confidence was set at 0.95.

These rules do not provide that much information as all the right-hand side is “chanceCreationShooting = Normal”. Additionally, while they all have strong support and confidence, none of them have a lift above 1.1. This suggest that these rules do not actually have much impact and are mainly due to a large frequency of teams that “chanceCreationShooting = Normal”.



*Figure 11:* Plot of the top 20 rules.

The plot above seems to confirm these suspicions as “chanceCreationShooting = Normal” is firmly in the middle of many different attributes. Furthermore, the main rules of interest “GC\_PG=low” and “GS\_PG” =high” are all connected to this. When tuning the rules to have these rules of interest in the right-hand side, no rules were outputted.

In the end, the association rule mining was not very successful at isolating what playing styles lead to more goals scored or less goals conceded. This may be a limitation in the size of the dataset and the large frequency of certain attributes like “chanceCreationShooting = Normal”. More data may have been able to shed more information on the subject. The decision to split the bins by interval instead of frequency could also have affected the results since it meant less teams were in the low goals conceded and high goals scored categories. However, while switching this binning to frequency may have improved the number of rules output, it may have also given rules that were less reflective of the truth. Or, it may just be that the specific style of play of a team does not have a major impact on the goals scored or conceded. A team can adopt a defensive mindset, yet still concede a lot of goals. Conversely, some of the more attacking style teams may not have the player quality to score the goals.

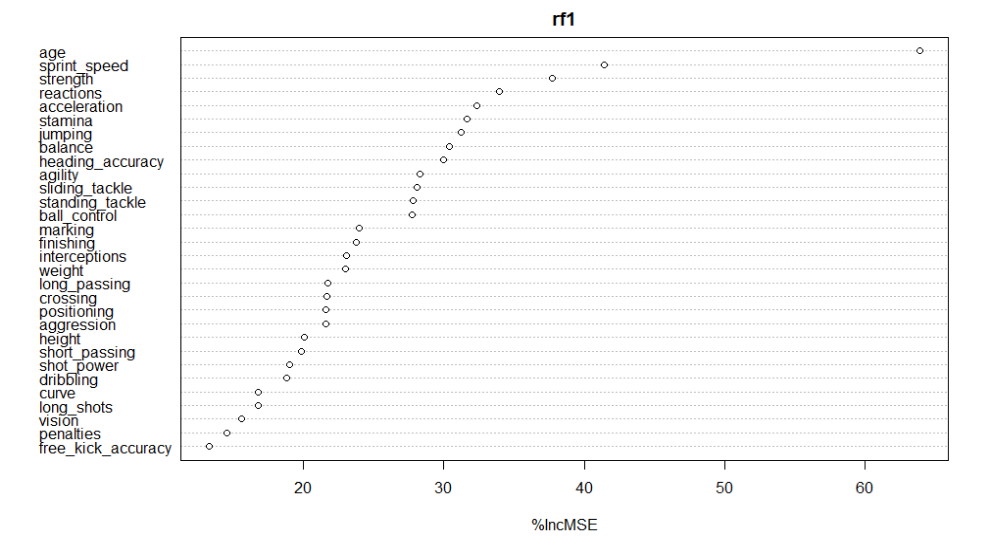
#### What has the biggest impact on a soccer player’s rating?

All aspiring soccer players want to know what it takes to be a great player. What should be the focus of training to make a step to the next level? To answer, this question, a random forest analysis was performed on the player ratings table to find what attributes have the biggest impact on overall player ratings.

Random forest combines the principles of bagging with random feature selection to add additional diversity to decision tree models. Bagging is the process of generating new training sets using samples and then combining the model outputs. By using bagging and random feature selection, random forest models are great at determining what features are the most important in a dataset for reaching an output.

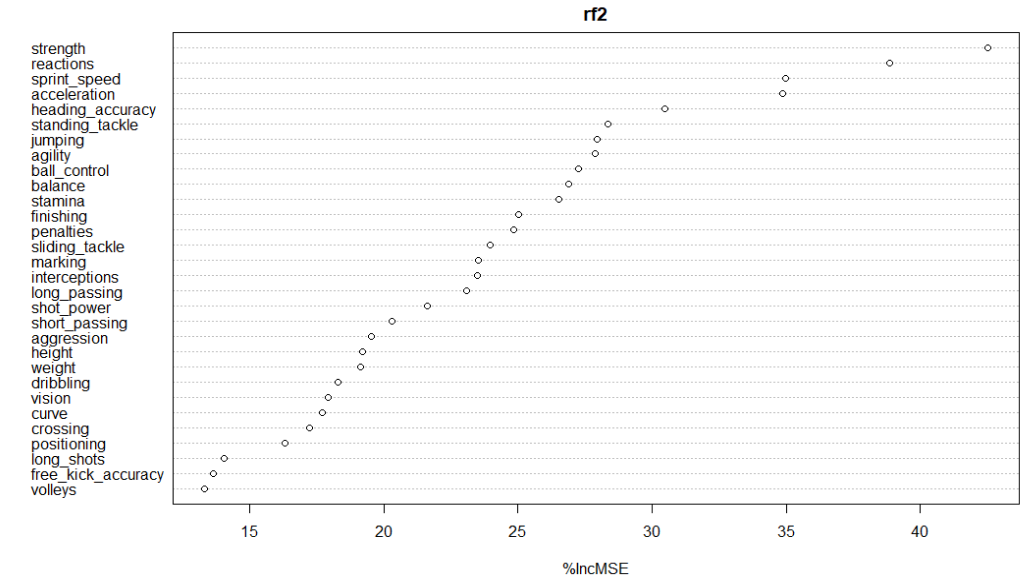
The variable importance plot generated from the Random Forest model sorts the features using %IncMSE, which is the % increase in the mean squared error (MSE). The MSE is an indication of how well the model’s prediction fit the actual values observed. The %IncMSE is thus a value that measures the increase in the MSE of predictions because of variable *x* being shuffled. The higher this number is, the more important the feature is to predict the output.

The Player table includes 32 attributes that can be roughly divided into soccer-specific skills like “ball\_control”, “marking”, and “short\_passing” and more athletic traits like “height”, “sprint\_speed”, and “strength”.



*Figure 12*: Random forest variable importance plot.

The initial variable importance plot generated from the random forest model shown above shows that “age”, “sprint\_speed”, and “strength” are the top three most influential features in a high player rating. This finding is interesting as all the attributes are physical traits and are not directly related to soccer skills. However, the “age” feature in this output may confound the results. While age is an important factor in a player’s skill, as more years of experience result in a player improving over time, this is not a feature that is particularly interesting as this is general knowledge. By removing this feature, a clearer view of the other features can be gleaned.



*Figure 13:* Random forest variable importance plot (age variable removed).

By removing the “age” variable, we still reach a very similar plot. Physical attributes like “strength”, “reactions”, “sprint\_speed”, and “acceleration” all are ranked at the top of the plot. More soccer specific skills like “ball\_control”, “long\_passing”, and “short\_passing” are further down the plot. Although these ratings come from the FIFA video games and are not necessarily a direct indication of what it takes to be a top soccer player, these are still interesting findings. Athleticism is clearly key to becoming a professional soccer player.

#### How accurately can a soccer match be predicted?

The ultimate goal of sports analytics for many people is being able to predict the outcome of a game. This match outcome is divided into three classes (“Home Win”, “Draw”, “Home Loss”). The bookies can predict this outcome correctly around 53% of the time. Although this may not seem high, given the random nature of the sport and the three possible outcomes, 53% is actually a very strong result. However, it also should be noted that the home team does win about 46% of the time, which means that their predictions are only able to add 7% of accuracy to predicting “Home Win” for every game. This statistic highlights the difficulty in predicting soccer outcomes. The full distribution of match results for this dataset is shown below.



*Figure 14:* Table of match outcomes

The attributes used for this analysis to predict match outcome was the home team’s win percentage, the away team’s win percentage, the average player rating on the home team, and the average player rating on the away team. To calculate the teams’ win percentages, a rolling count of wins was calculated for each team for their home and away games. This rolling count was reset at the beginning of each new season to account for the many changes that occur within each team during the offseason. Furthermore, by having a rolling count of wins each team’s win percentage for home and away would fluctuate throughout the season. By doing this, the form of the teams could be partly accounted for.

Results

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **Accuracy** | **Parameters** |
| Random Forest | 48.4% | Default |
| Naïve Bayes | 51.5% | Default. Laplace = 1 |
| SVM | 52.4% | Radial kernel, C = 1 |

*Figure 15*: Table of match prediction result accuracy.

None of the classification techniques were able to surpass the 53% accuracy of the bookies; SVM was the highest with 52.4%. However, despite not surpassing the bookies, 52.4% is a very good accuracy, especially given the random nature of soccer games. All the supervised learning techniques were also able to surpass the 46% accuracy of picking “Home Win” for every game. This demonstrates the ability of the models to find patterns in the data and demonstrates a good prediction result.

The SVM model used a radial kernel and a cost constraint of 1. It outperformed other SVM models that used linear and polynomial kernels. Increases to the C value hurt the model outputs in all trials. Although the SVM had the best overall results, it also did the worst at predicting ties as most of its predictions were either “Home Win” or “Home Loss”.

To use the Naïve Bayes method, all the numeric fields used were converted to categorical variables before running the algorithm. The Laplace estimator was included in the model to ensure that each feature has a nonzero probability of occurring with each class.

Despite the ensemble method of Random Forest that combines model outputs to create one stronger prediction, the Random Forest model performed the worse. The model was tuned to find the optimal number of features to split on but found that the default parameter matched this already. Adding additional trees also did not improve performance. The weaker results may be a result of the small numbers of features in the model. Without many features to tune on, it loses some of the benefits that are implicit to random forest.

## Conclusion

Given these results, the created match prediction models are not ready to take to Las Vegas yet, but these results are encouraging. Further feature engineering and tuning of the models could likely get these models above the 53% threshold to make these models profitable against betting odds. However, this is also based on the assumption that the bookie accuracy has not improved since this Kaggle dataset was released.

This analysis also supports many of the stories of the different leagues’ styles of play. The defining characteristics of each cluster gave interesting insight into how each league plays. However, we can still see that within every league there is variation in playing styles, as all clusters were found in all the leagues.

The random forest analysis also provided interesting insights into what it takes to be a professional soccer player. Although it is necessary to possess many of the soccer specific skills, having athleticism is vital to make it at the top level. While many players can learn these soccer skills, it may be that having the unteachable traits of speed and strength are what it takes to differentiate players at this level. However, this should also not be taken to mean that athleticism is all it takes to succeed at the professional level, as many soccer players, such as Lionel Messi, have been able to excel despite being one of the smallest players on the field.

Overall, these analyses have shed a lot of light on the game of soccer but have also illuminated a lot of the difficulties implicit in the analysis. Quantifying a player’s ability is still not an exact science and generally has come down to an “eye-test” evaluation. Furthermore, the lack of many trackable events during the game makes many insights currently impossible to do. In order to take soccer analytics to the next level, finding a way to quantify more of the details of the game is critical.

## References

1. <https://www.fifa.com/worldcup/news/almost-half-the-world-tuned-home-watch-2010-fifa-world-cup-south-africat-1473143>