African Institute for Mathematical Sciences

African Masters in Machine Intelligence

Foundations of Machine Learning

Authors:
Belona Sonna
bsonna@aimsammi.org
Lionel N. Tondji
tngoupeyou@aimsammi.org
Sara .E.M Elkafrawy
selkafrawy@aimsammi.org
Keziah Naggita
knaggita@aimsammi.org
Montaser Mohammedalamen
fmontaser@aimsammi.org

Date: December 7, 2018

Abstract

Football is one of the most popular sports in the world. It is enjoyed by both the young and the elderly, the poor and the rich and the black and the white. Among different races, regions and communities, football is a source of entertainment and brings people together. Most of the spectators and supporters love predicting the outcome of matches prior to watching the game and the coaches and teams occasionally buy players. They are mainly interested in players that complement the team. A lot of research focused on analyzing and modelling professional football has been conducted. On Kaggle, an online community of data scientists and machine learners to find and publish data sets, there is a data set concerned with football events and many data scientists analyzed and modelled it. Most of the engineers basically analyze the professional football data sets and visualize the data set contents without having a clear statement about what the major goal is. In addition, other researchers have reviewed the cause and prevention of injury in football matches. In this project, we focus on identifying the strengths and weaknesses of the team, find the most plausible actions to take and to improve the team performance in the leagues. We propose solving prediction problems with supervised machine learning algorithms such as: logistic regression, random forests, and artificial neural networks to achieve the main objective. We describe the ways of dealing with the unbalanced classes, missing data and outliers before we perform analysis, modelling and predictions. We evaluate a football data set of 9,074 games, totaling 941,009 events from the biggest 5 European football (soccer) leagues: England, Spain, Germany, Italy, France from 2011/2012 season to 2016/2017 season as of 25.01.2017 with over 90% of the played games during these seasons having event data. Our neural network predicts whether an event will become a goal or not and the match final result based on the odds of the match. We strongly believe that use of machine learning algorithms will make the decision making, planning and future predictions easier for the players, coaches, funding organizations and even supporters who might be betting on the teams.

Contents

List of Figures

List of Tables

1. Introduction

Football is a sport involving two teams of eleven players. Each of the players strives to kick the spherical ball into the opponent's goal by kicking and directing the ball mainly with their feet, and sometimes the head. Players are not allowed to use their hands during the game, only the goal keeper gets this privilege. According to Wikipedia's football history [2], football emerged in the mid-19^{th} century in Britain which later on, in the following century, became the most popular practiced in every corner in the world. Compared to other sports, football is cheaper and more accessible. Football can be played on all surfaces; sand, concrete and grass, in pairs and no pairs, with a fancy ball and a rag ball, with goal cages and cans. Since most of the material used in the game is cheaper unlike in games like Tennis where you have to buy equipment like tennis rackets, football has become more and more popular over the years. Football rules can be considered as actions aimed at social integration, team building, respect for values such as friendship, respect for others and the ability to accept the final result with an emphasis on the positive aspects of defeat and victory [4].

1.1 Problem Statement

Football data analysis has become a vital field for sports companies that invest heavily in analyzing the performance of their teams as well as the opposing teams [8]; as it is well known that football sport is a huge industry that many businesses invest money in. There are different aspects can be used for performance analysis such as: analyzing what features induce a goal to be scored, what players tend to make foul to be careful when playing against his team, who are the most offensive players (in case a team wants to buy new players) ... etc. Such analysis helps the coach and the management of a team to build a better team which benefit them in gaining more investments. While some data sets are not available for the public use, the dataset we are using in this report is publicly available on Kaggle website.[1]

[1] https://www.kaggle.com/secareanualin/football-events/home

1.2 Objectives

Our main objective is to provide football decision makers with useful insights by spotting weaknesses and strengths in the teams/players in order to take decisions aiming to improve the team performance. In a second time we are going to look at the performance of each player because it is related to the performance of the team and see which kind of game piece the best team are using to win.

1.3 Data Set Description

Our data set has two files of data and one dictionary. The first file gives information on all the recorded events with 941,009 events for a 9,074 games. The second file gives the details of the odds for the games recorded in the first file. for each league we have information on the seasons from 2012 to 2017 except the English league whose information only starts in the season 2014. The dictionary helps us to understands the values in some of the columns of the events table.

We will first try to have a look at the different features of our dataset.

1. id odsp : unique identifier of game (odsp stands from oddsportal.com)
2. id event : unique identifier of event (id odsp + sort order)
3. sort order : chronological sequence of events in a game
4. time : minute of the game
5. text : text commentary
6. event type : primary event. 11 unique events (1-Attempt(shot), 2-Corner, 3-Foul, 4-Yellow Card, 5-Second yellow card, 6-(Straight) red card, 7-Substitution, 8-Free kick won, 9-Offside, 10-Hand Ball, 11-Penalty conceded)
7. event type2 : secondary event. 4 unique events (12 - Key Pass, 13 - Failed through ball, 14-Sending off, 15-Own goal)
8. side : 1-Home, 2-Away
9. event team : team that produced the event. In case of Own goals, event team is the team that benefited from the own goal
10. opponent : team that the event happened against
11. player : name of the player involved in main event (converted to lowercase and special chars were removed)
12. player2 : name of player involved in secondary event

13. player in :player that came in (only applies to substitutions)
14. player out : player substituted (only applies to substitutions)
15. shot place : placement of the shot (13 possible placement locations, available in the dictionary, only applies to shots)
16. shot outcome : 4 possible outcomes (1-On target, 2-Off target, 3-Blocked, 4-Hit the post)
17. is goal : binary variable if the shot resulted in a goal (own goals included)
18. location : location on the pitch where the event happened (19 possible locations, available in the dictionary)
19. bodypart : (1- right foot, 2-left foot, 3-head)
20. assist method : in case of an assisted shot, 5 possible assist methods (details in the dictionary)
21. situation : 4 types: 1-Open Play, 2-Set piece (excluding Direct Free kicks), 3-Corner, 4-Free kick

1.3.1 events.csv

In this file, we have records of matches of 5 leagues namely: the French league, the English league, the Italian league and the Spanish league. For each record, we have many features that will exploit and explain in details in the Statistical analysis.In general, this file gives us the information on the teams participating in the match, by specifying who receives the match, and some of the events during the match.

1.3.2 ginf.csv

In this file, we find information about each game in the events file such as: the country, season, league, odds on home win, odds on away win and the two teams played the game.

1.4 Distribution and Exploration of the Features

Figure 1.1, 1.2 and 1.3 shows a detailed distribution of the different features, mainly the events. This is helpful in further analysis of the variables.

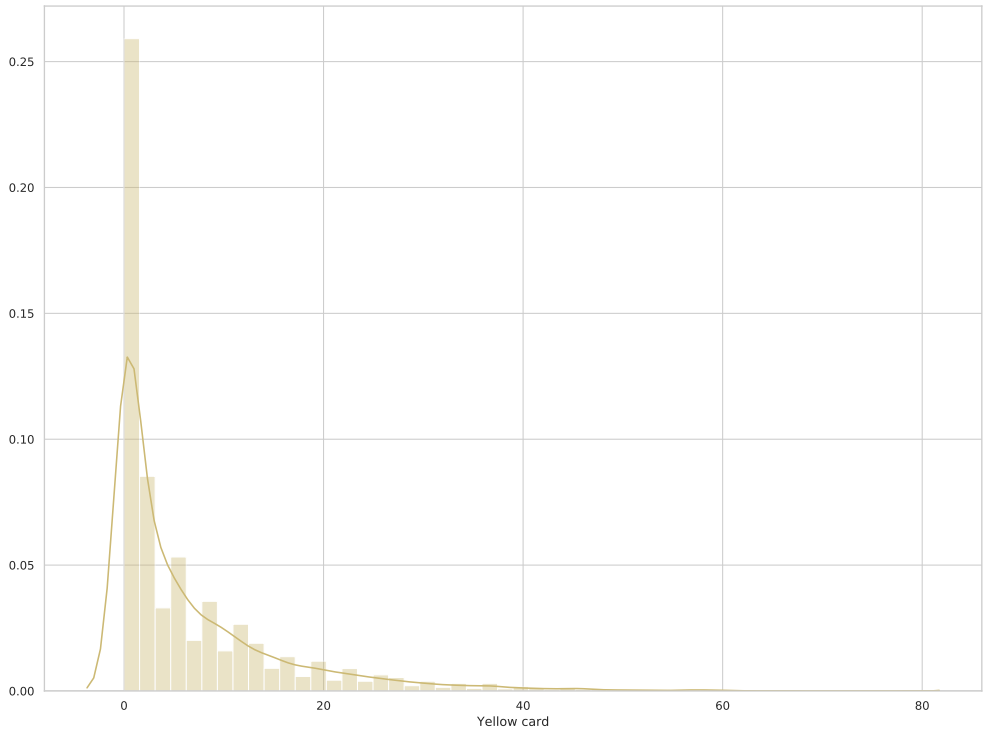


Figure 1.1: A kernel density estimate and histogram

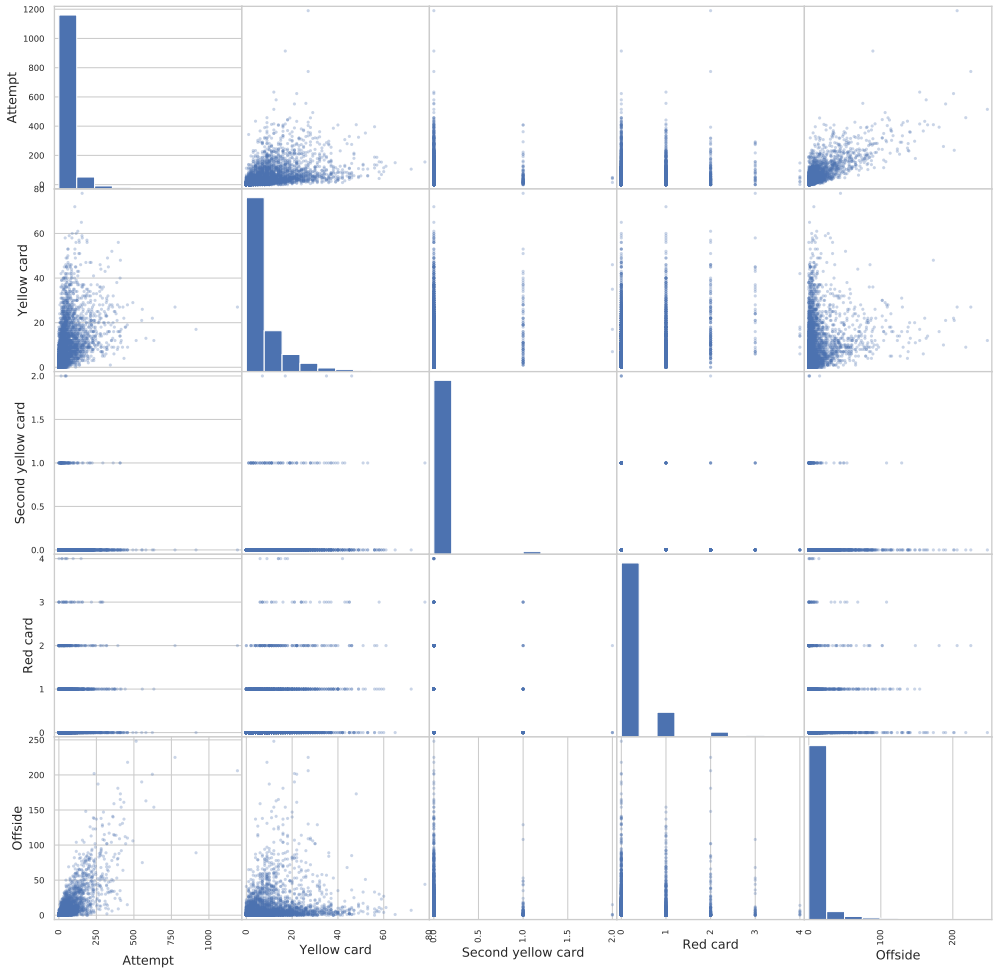


Figure 1.2: Relationship between players in the different leagues shows how many players played in a given league, how many played in two or more leagues and how many played in all

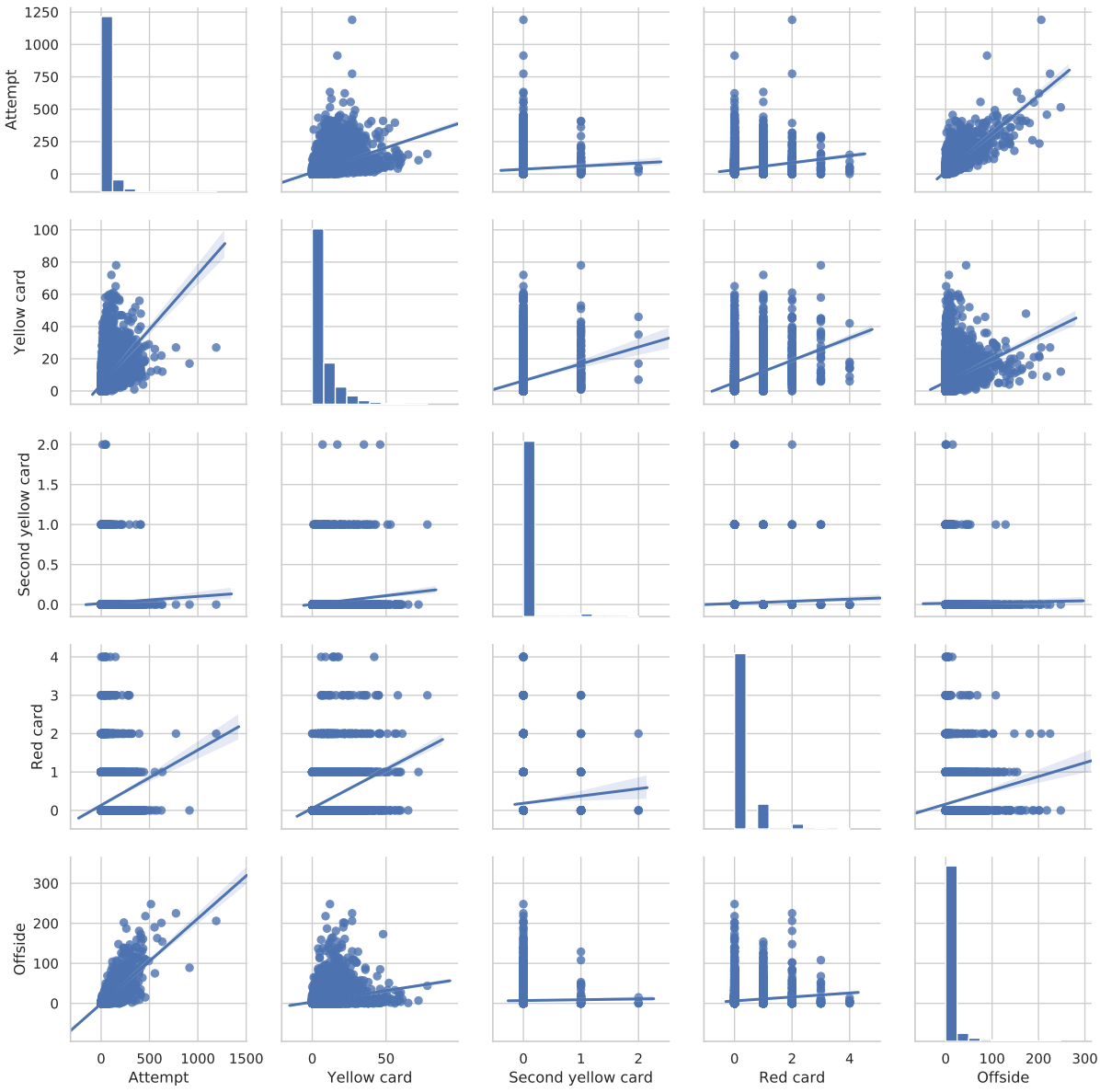


Figure 1.3: Relationship between players in the different leagues shows how many players played in a given league, how many played in two or more leagues and how many played in all

2. Statistical Data Analysis

Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. In today's business, data analysis is playing a role in making decisions more scientific and helping the business achieve effective operation[7]. The aims of this part is by using data analysis techniques, extract as much information from the football events data set to achieve the main objective. There are several steps to follow in a iterative ways in order to analyze our data. These steps are :

- ☐ Data collection : Our data was collected from kaggle website
- ☐ Data processing : Here we are placing our data into rows and columns in a table format using Python libraries for data analysis called **Pandas**
- ☐ Data cleaning : In this step we are trying to remove duplicate data and outliers
- ☐ Exploratory data analysis : After cleaning the data, we will summarize the main characteristics by using visual methods. We will use information graphic types such as : Line chart, Bar chart, Histogram, Scatterplot, Boxplot, pie chart
- ☐ Modeling and algorithms : In this part we part we try to develop some predictive models that can be use in order to predict odds.

From this enumeration given in the Data Set Description part, we can see that we have more than 21 features presents in our data set. This data set provide a view game from the biggest five European football (soccer) leagues: England (Premier League), Spain (La Liga), Germany (Bundesliga), Italy (Serie A), France (League One) from 2011/2012 season to 2016/2017 season as of 25.01.2017.

2.1 General Analysis of Players, Events and Leagues

2.1.1 Analysis of players in all leagues

Figure 2.1 shows the league composition in terms on players. Here, we are interested in knowing how many players played in a league. This is important in knowing which leagues are comparatively bigger than others to facilitate a further analysis of teams.

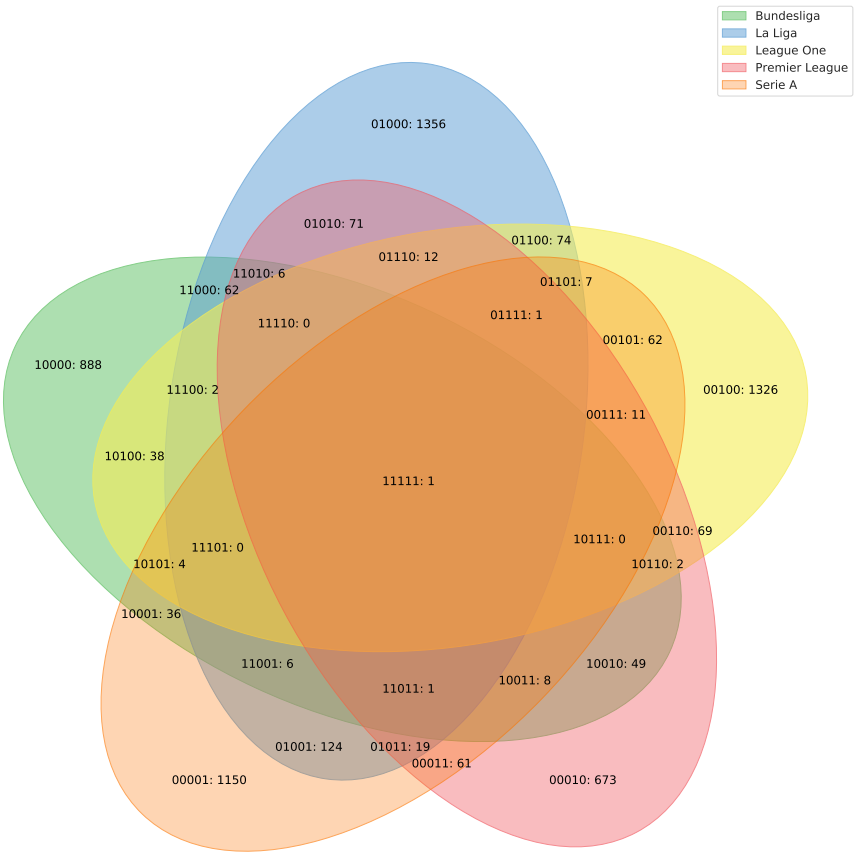


Figure 2.1: Relationship between players in the different leagues shows how many players played in a given league, how many played in two or more leagues and how many played in all

2.1.2 Analysis of events, cards and assist methods

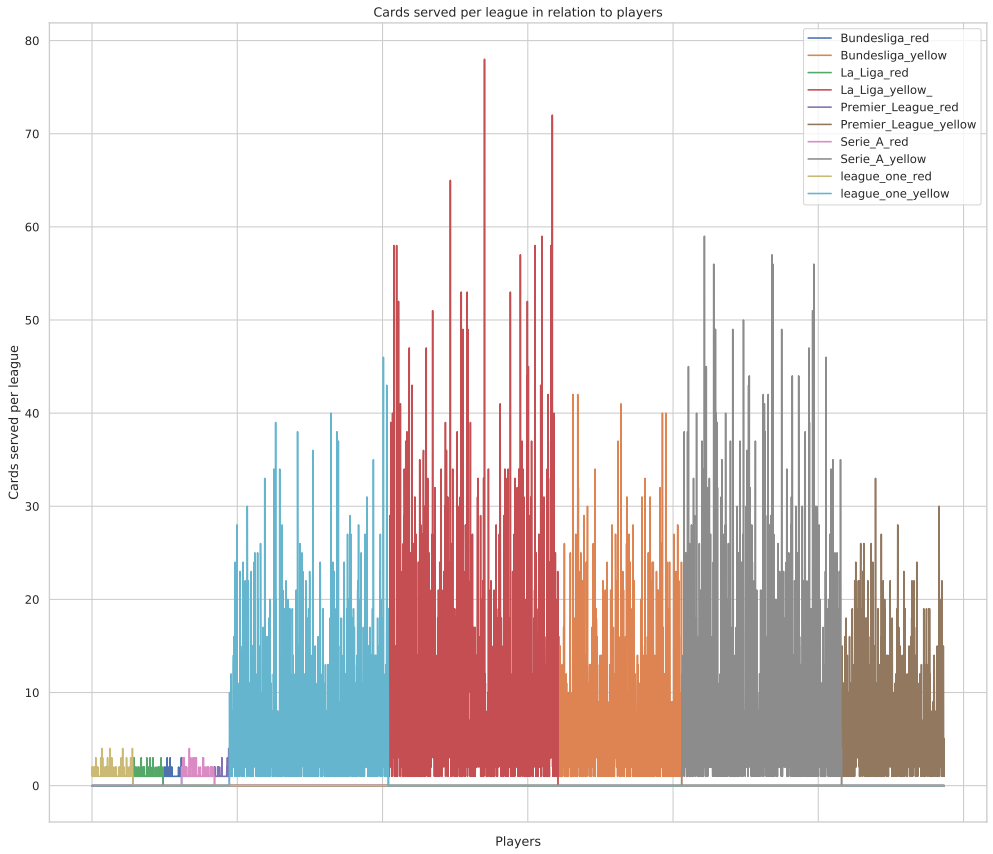


Figure 2.2: Distribution of cards served per league highlights the leagues in which the players were more aggressive and which weren't.

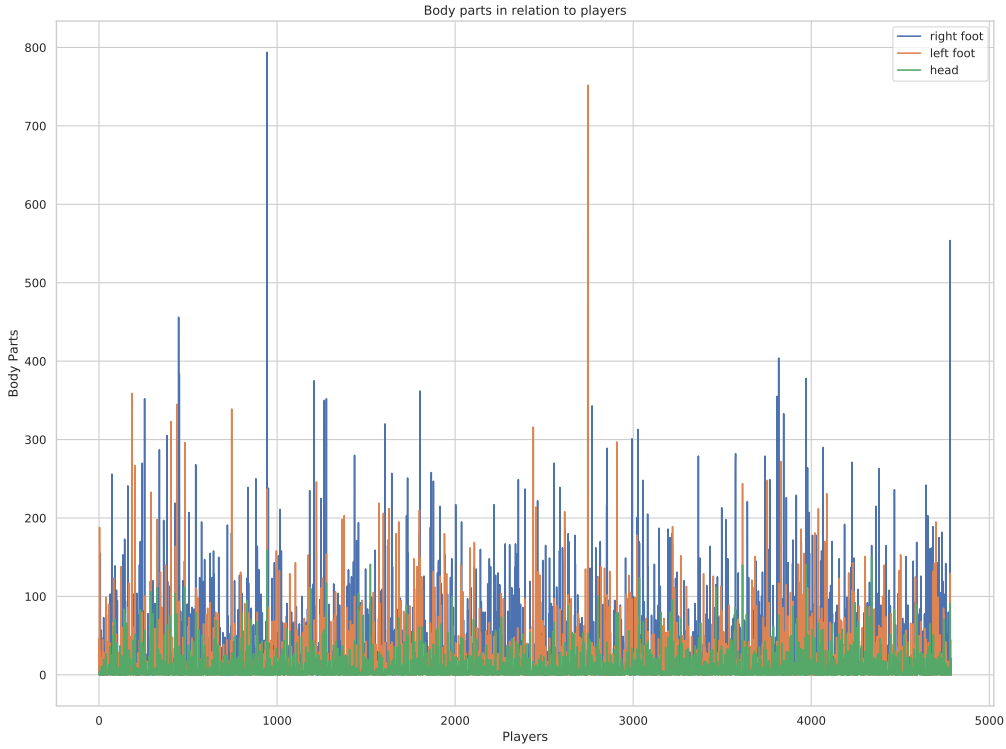


Figure 2.3: A detailed analysis of the body parts: left leg, right leg and head highlight the most and least used body parts by the players

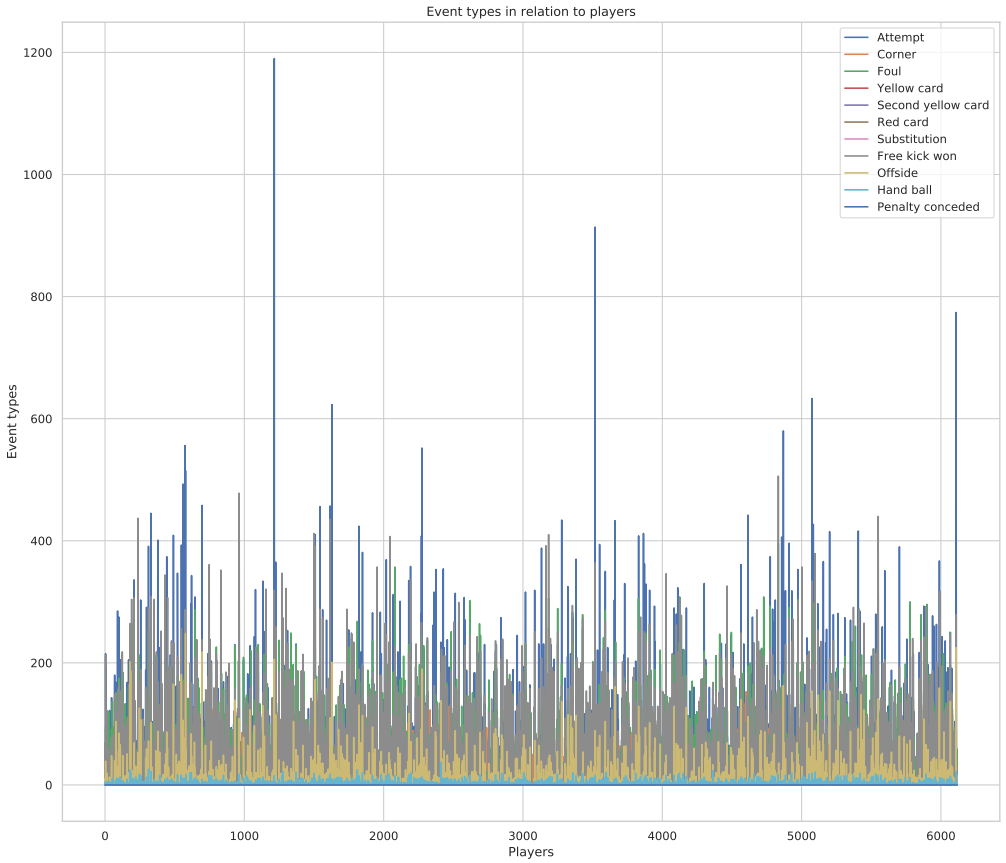


Figure 2.4: There are several events in a game; foul, corner, free kick, goal, and so on. The distribution of events frequency for different players in different teams and leagues is highlighted

2.2 Detailed Analysis of Goals

Now we will try to answer some questions related to the performance study of each team and each player.

2.2.1 Time when teams are most likely to score

Here we are interested, at which key period of the game, the players are more trying to score a goal in all the data gathering all the seasons of all the leagues.

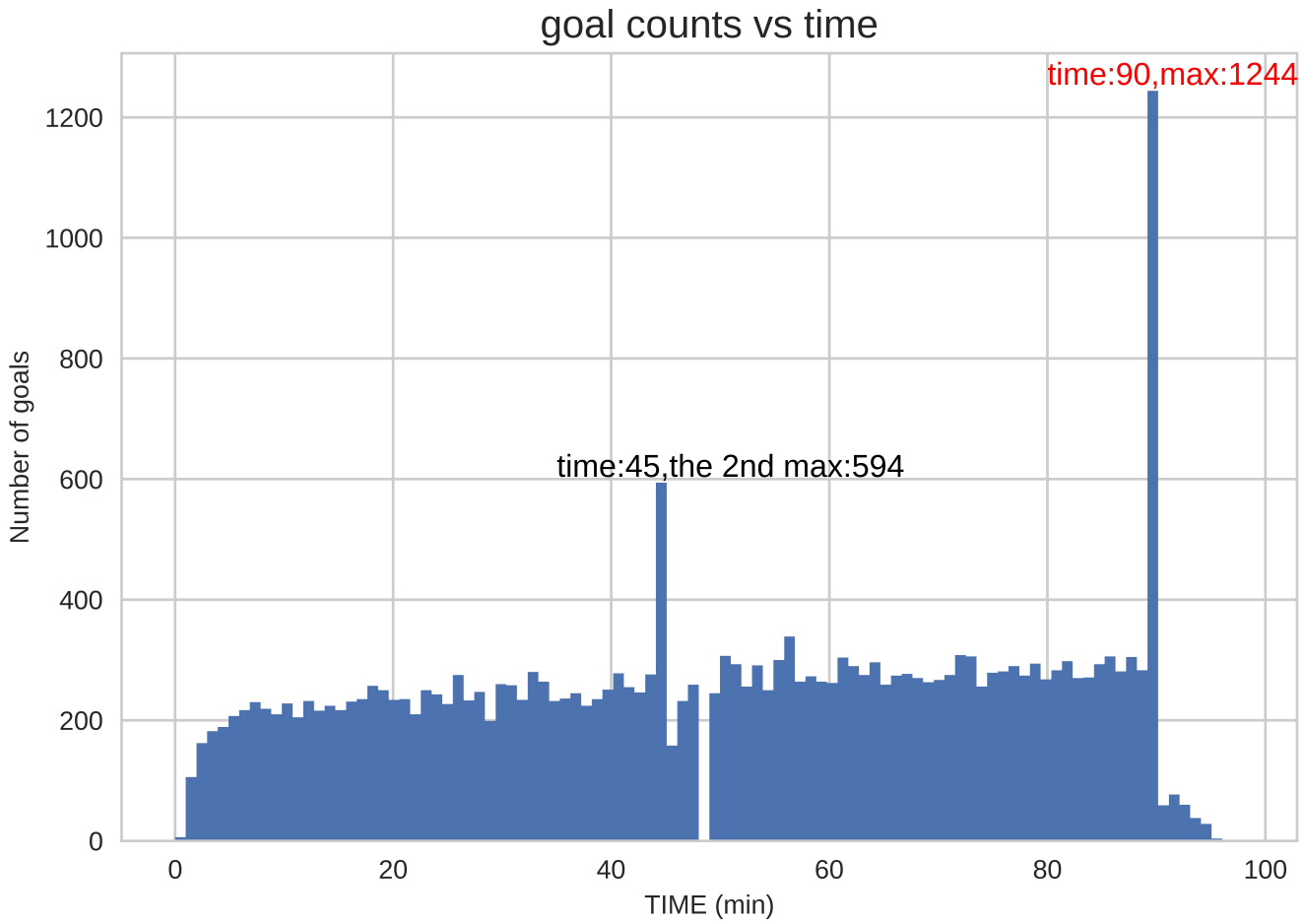


Figure 2.5: Number of Goals by the time for all the leagues from 2012 to 2017

From Figure 2.5, it can be seen that most goals are scored at half-time (45 minutes + overtime) and at half-time (90 minutes + overtime). In general the number of goals scored in the second half is higher than the number of goals scored in the first half.

2.2.2 Determination of the most offensive team

Here we are interested at what are the most offensive and the less offensive teams in terms of the number of goals they scored during all the season. We choose La Liga to look at the characteristics of team involved in this league.

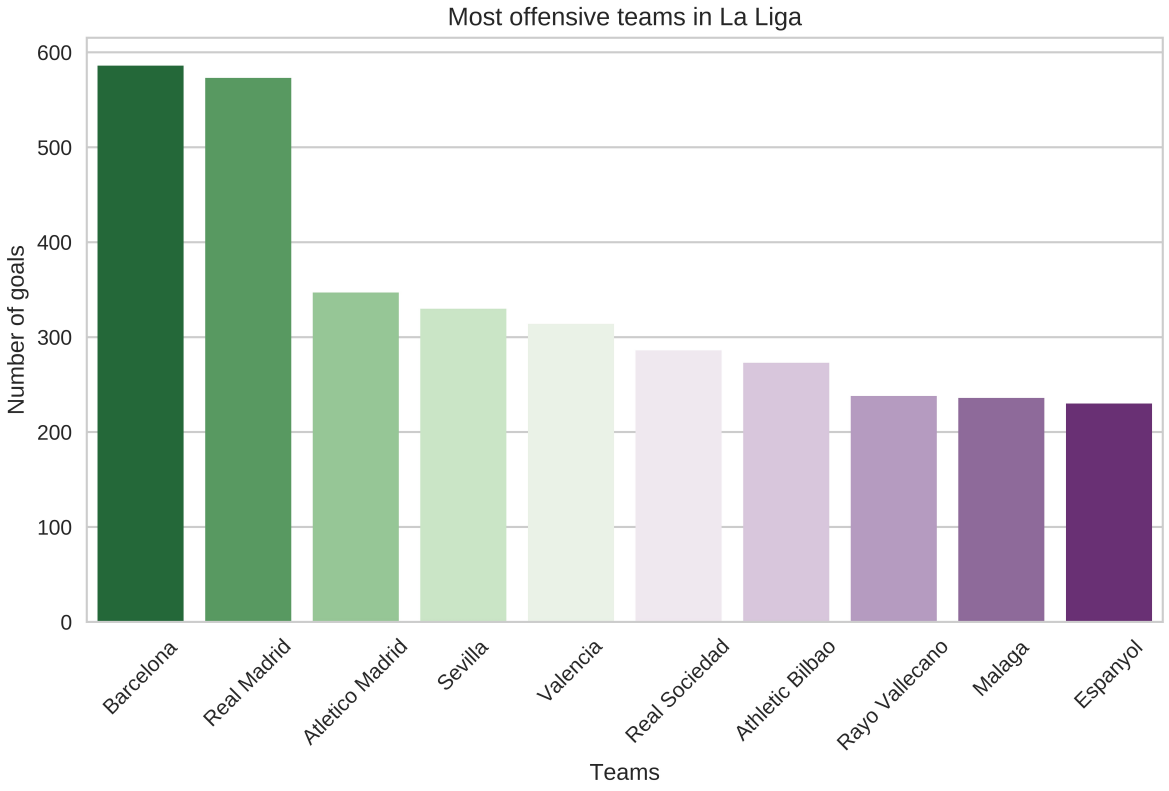


Figure 2.6: The Most offensive team in La Liga from 2012 to 2017

From Figure 2.6, we can notice that the leading teams in term of offensive strategy are Barcelona and Real Madrid, may be because they have each one best player of the world.

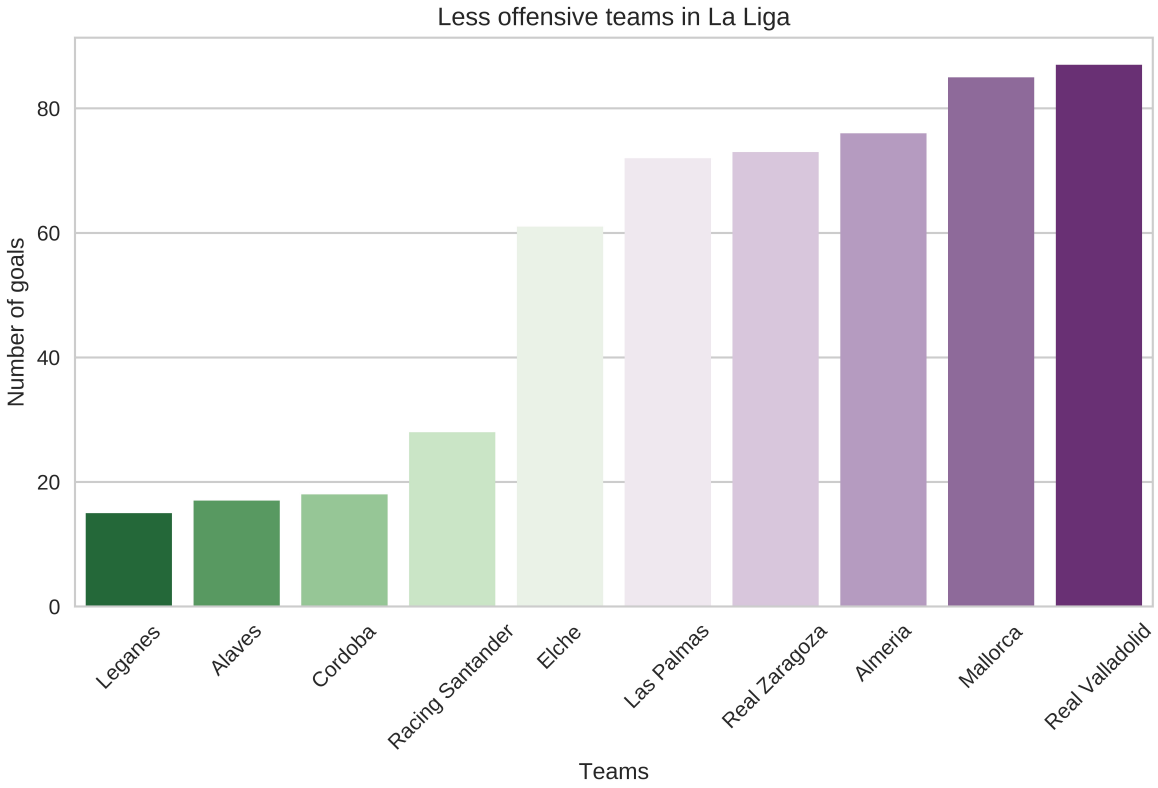


Figure 2.7: The Less offensive team in La Liga from 2012 to 2017

From Figure 2.7, we can notice that the teams that are not scoring a lot have a small number of goals from 2012 to 2017 and they are far away in term of number of Goals with team such as Barcelona and Real Madrid.

After looking at the stats for the whole team, now we will beak it down per players by looking at the player who have the highest number of goals in the league.

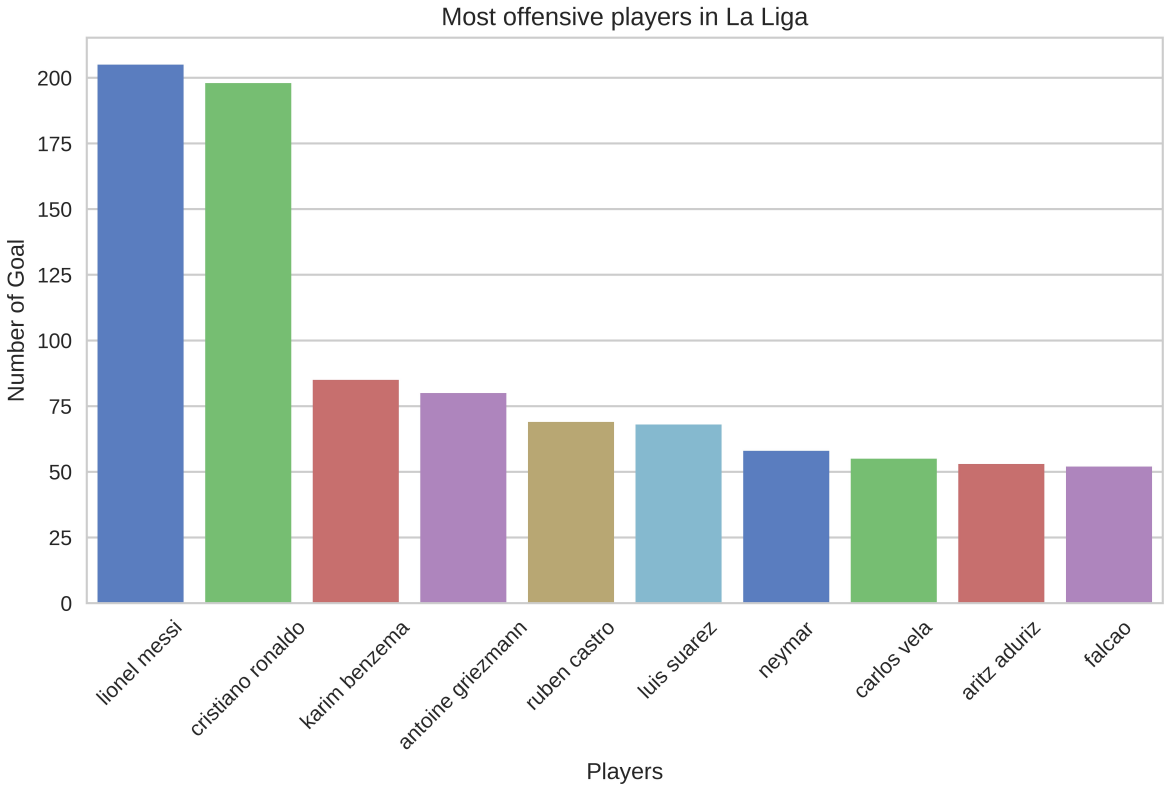


Figure 2.8: The Most offensive player in La Liga from 2012 to 2017

From Figure 2.8, we can see that the most offensive players are Lionel Messi and Cristiano Ronaldo which is in correlation with the most offensive teams that we look at before and again here the ratio with other players is too huge.

The Performance of a team can be also see by the number of fouls that they are making, In the following section we will look at the number of red cards took by the team.

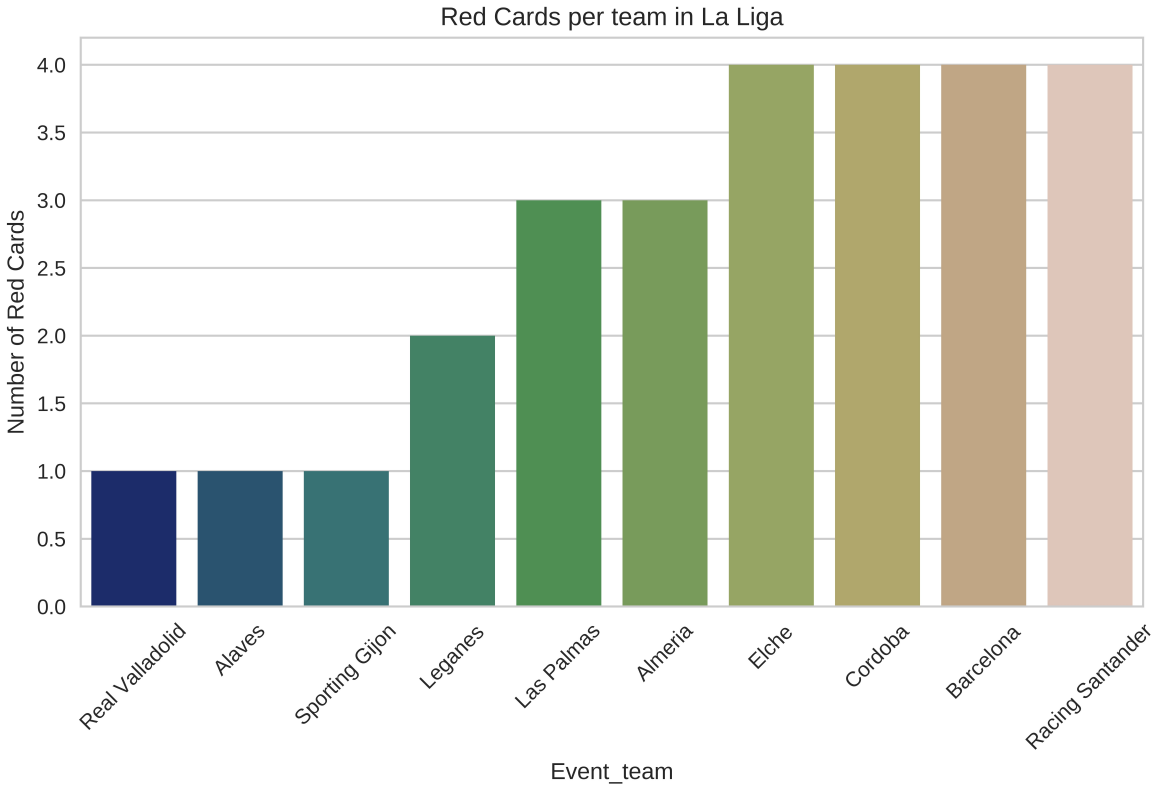


Figure 2.9: The Number of red cards per team in La Liga from 2012 to 2017

From Figure 2.9, we can remark that a part from being the most offensive team, Barcelona is also among the teams that commit less fouls in La Liga league, this can be also verify because the like to stay with the ball and by trying to get the ball their adverse tend to commit fouls on them.

2.2.3 Analysis of shots and shot places

In this part we will try to define some football terms present in our data set before doing our analysis.

A shot on target is defined as any goal attempt that [1]:

- Goes into the net regardless of intent.
- Is a clear attempt to score that would have gone into the net but for being saved by the goalkeeper or is stopped by a player who is the last-man with the goalkeeper having no

chance of preventing the goal (last line block).

Shots directly hitting the frame of the goal are not counted as shots on target, unless the ball goes in and is awarded as a goal.

Shots blocked by another player, who is not the last-man, are not counted as shots on target.

A shot off target is defined as any clear attempt to score that[1]:

- Goes over or wide of the goal without making contact with another player.
- Would have gone over or wide of the goal but for being stopped by a goalkeeper's save or by an outfield player.
- Directly hits the frame of the goal and a goal is not scored.

Blocked shots are not counted as shots off target.

A blocked shot is defined as any clear attempt to score that[1]:

- Is going on target and is blocked by an outfield player, where there are other defenders or a goalkeeper behind the blocker.

A Set piece can be define as an attempt created where the ball starts from an indirect free kick dead ball situation[1].

Shooting Accuracy : A calculation of Shots on target divided by all shots (excluding blocked attempts and own goals)[1].

2.2.4 Individual teams

Barcelona

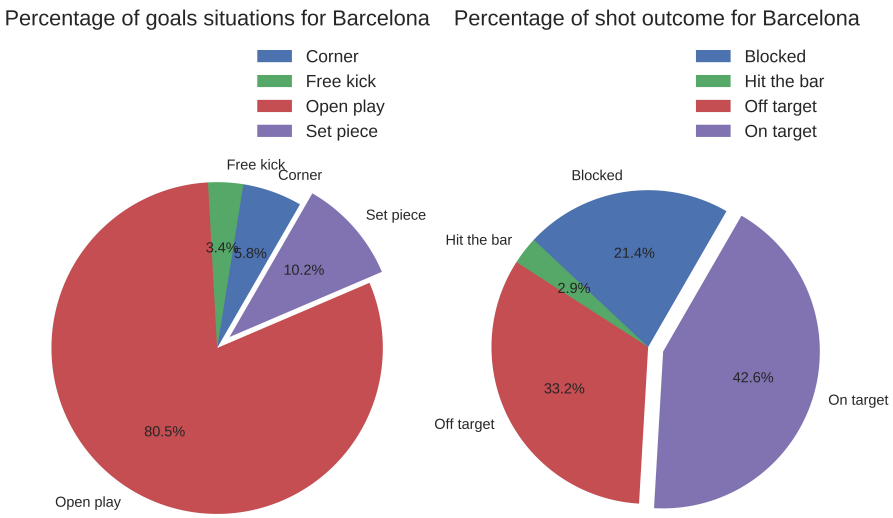


Figure 2.10: Statistics for Barcelona in La Liga from 2012 to 2017

From Figure 2.10, the plot shows that Barcelona team is using Open play and set of piece to score on most of their situation. Another remarks is 42.6 percent of the time their shot are on target which is good because it show that they are using their shot in a efficient way.

Detailed analysis of real madrid

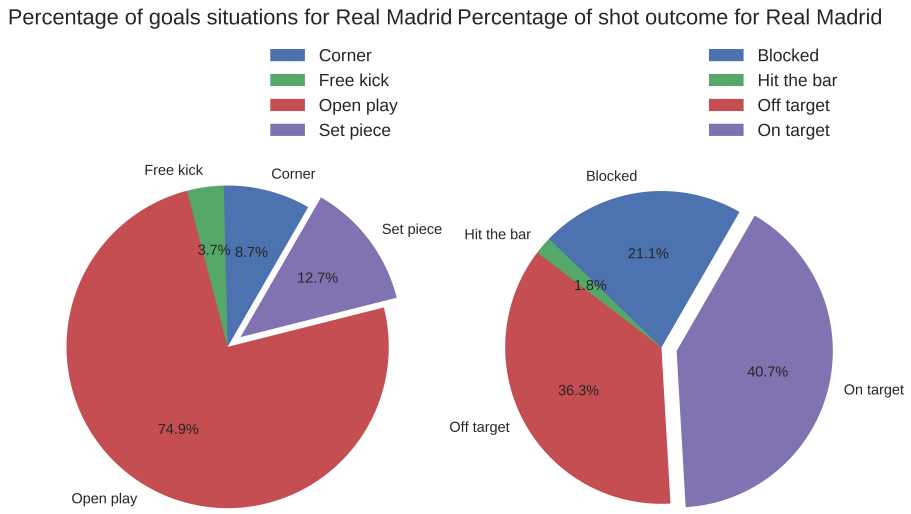


Figure 2.11: Statistics for Real Madrid in La Liga from 2012 to 2017

From Figure 2.11, the plot shows that Real Madrid team is using Open play and set of piece to score on most of their situation. Another remarks is 40.7 percent of the time their shot are on target which is good because it show that they are using their shot in a efficient way.

In summary, we can see that both Barcelona and Real Madrid have similar style play, and this can be adopt by other team if they want to perform well.

Barcelona vs Real Madrid

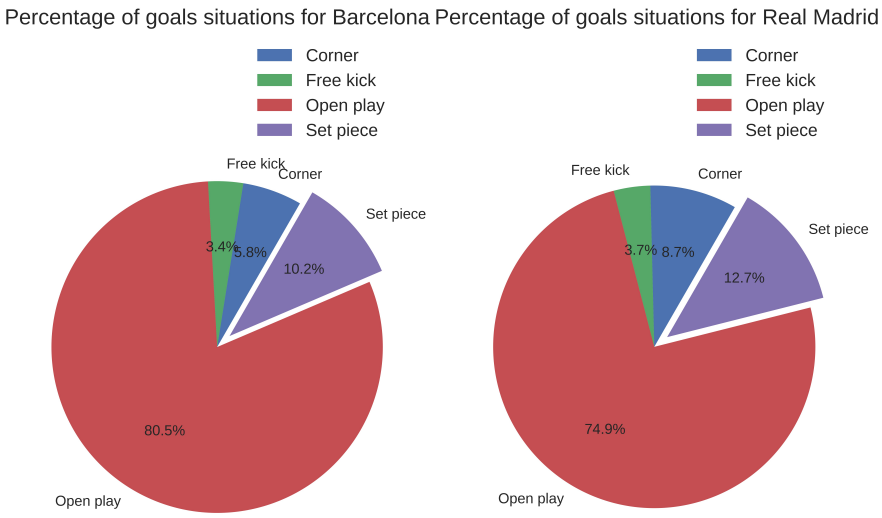


Figure 2.12: Statistics for Barcelona vs Real Madrid in La Liga from 2012 to 2017 in terms of goals situations

The plots show that about 75 percent of the goals scored are from open play and the fact that Read Madrid is the team using most an set of piece game.

2.2.5 Shooting Accuracy

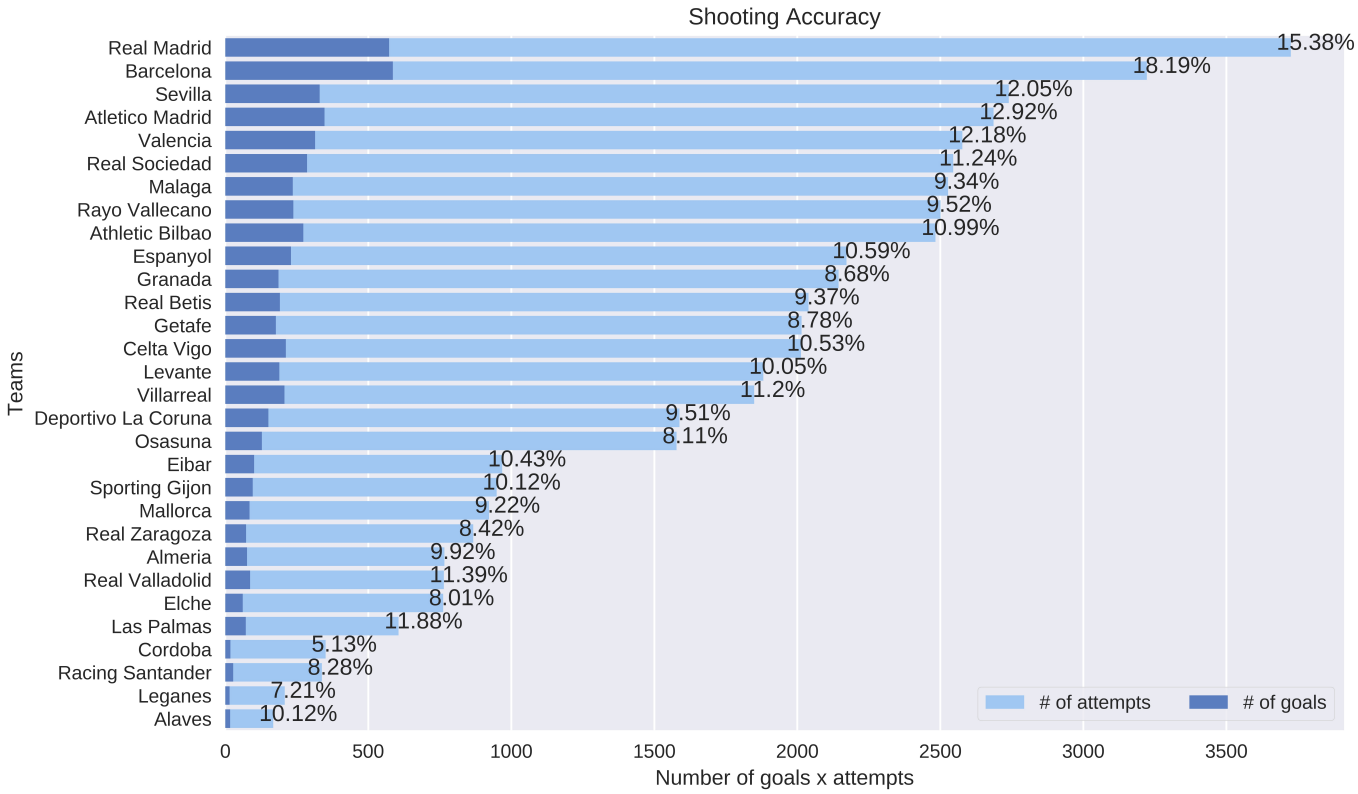


Figure 2.13: Shooting Accuracy for team in La Liga from 2012 to 2017

Figure 2.13, show the shoot accuracy for all the team in La Liga, and from this figure we can understand that team like Barcelona or Real Madrid are team that when you are playing with them, you have to be careful the defense because they need a small numbers of attempts to score.

2.2.6 Last-minute winners in Europe's top 5 leagues

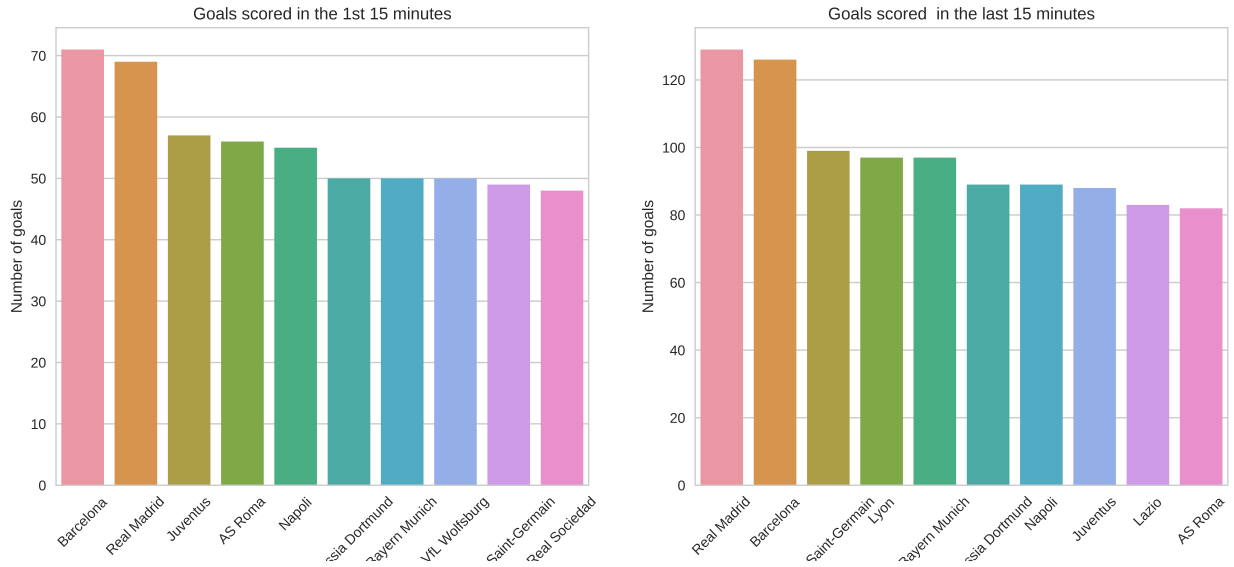


Figure 2.14: Last-minute winners in Europe's top 5 leagues from 2012 to 2017

We can notice that almost all the best teams of the 5 leagues appeared in both lists: the teams that scored the most goals in the beginning of the match and in the end of the match. Taht show that we need to be 100 percent focus from the beginning until the end facing this kind of team.

2.3 Detailed Analysis of Cards

2.3.1 Analysis of the time cards are most likely to be served

Figure 2.9 and 2.16 we notice that players are most likely to be served towards the end of first half (45 minutes) and towards the end of second half (90 minutes) because usually teams are trying to score goals and this increases aggressiveness among the players. In order to improve the team performance, players should balance out the game to avoid panic towards the end. Early research into football analysis [6] shows that a consistent game with nice strategies like passes increases the team's performance.

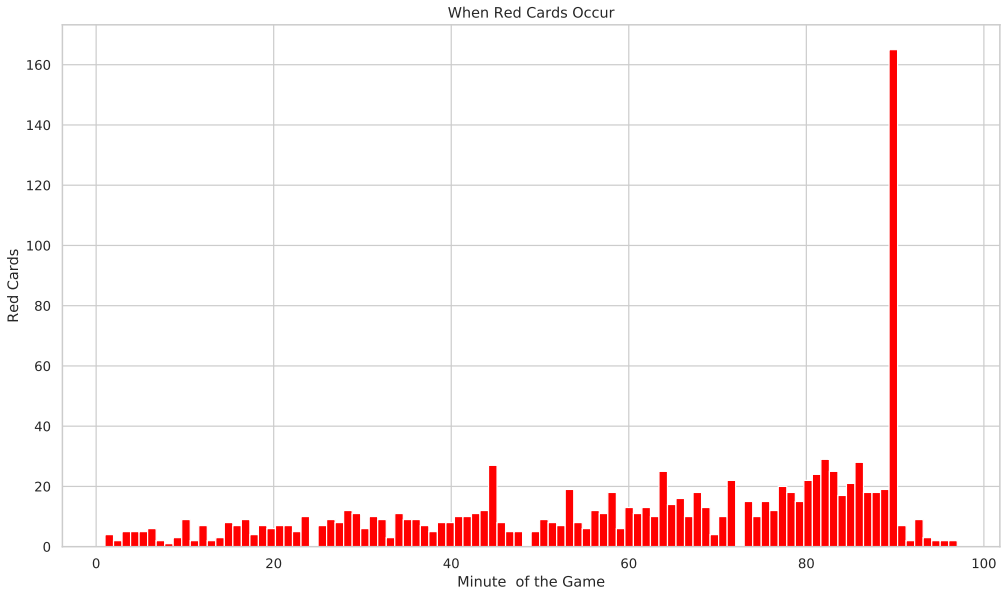


Figure 2.15: Time red cards are served looks into the details of the game over the 100 minutes interval and assesses the time when the cards are served

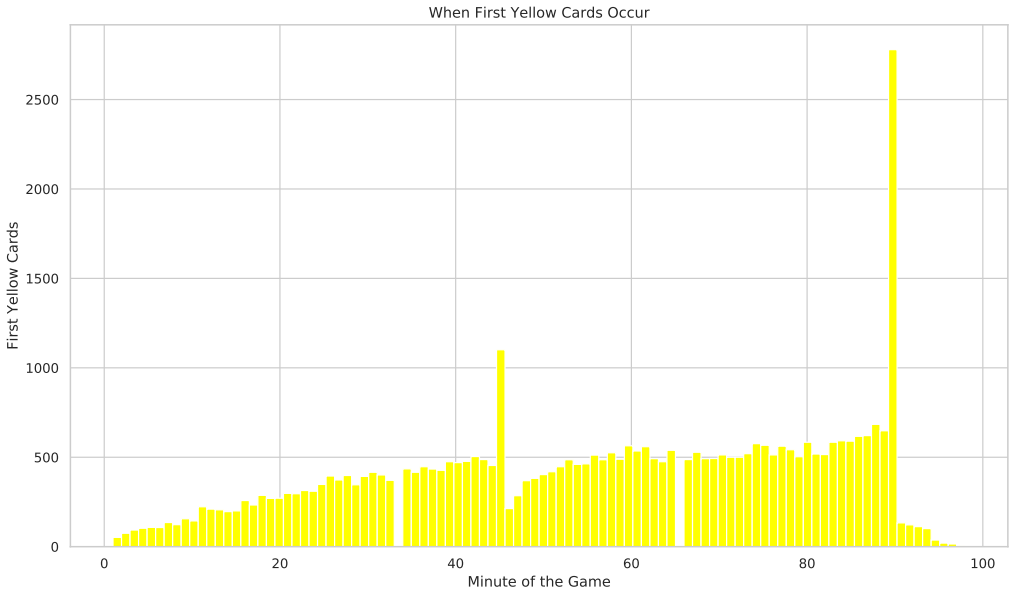


Figure 2.16: Time yellow cards are served looks into the details of the game over the 100 minutes interval and assesses the time when the yellow cards are served. This explains the intensity and panic among players to score towards the end of the first and second half of the game

2.3.2 Correlation between getting served and the team's performance

Figures 2.17, 2.18 and 2.19 give a detailed correlation between the teams getting served and their performance in the game. From 2.17, we that one of the members of the team getting a red card reduces the morale of the players. There is a slight increase in attempts, fouls, corners, and hand balls. However from 2.18, we see that a player getting a yellow cards highly affects the players' performance. There is a big increase in: fouls with a correlation of 0.9, free kicks with a correlation of 0.7, and hand ball with a correlation of 0.6.

Figure 2.19 shows a comparison between the correlation of yellow cards, red cards and second yellow cards with other events. We see that when players get the first yellow card, they are driven to improve the team performance by increasing the number of attempts and so on, while when they get the second yellow card, their motivation decreases, the activities don't increase that much and with a red card, the motivation is slight increased.

Figure 2.19

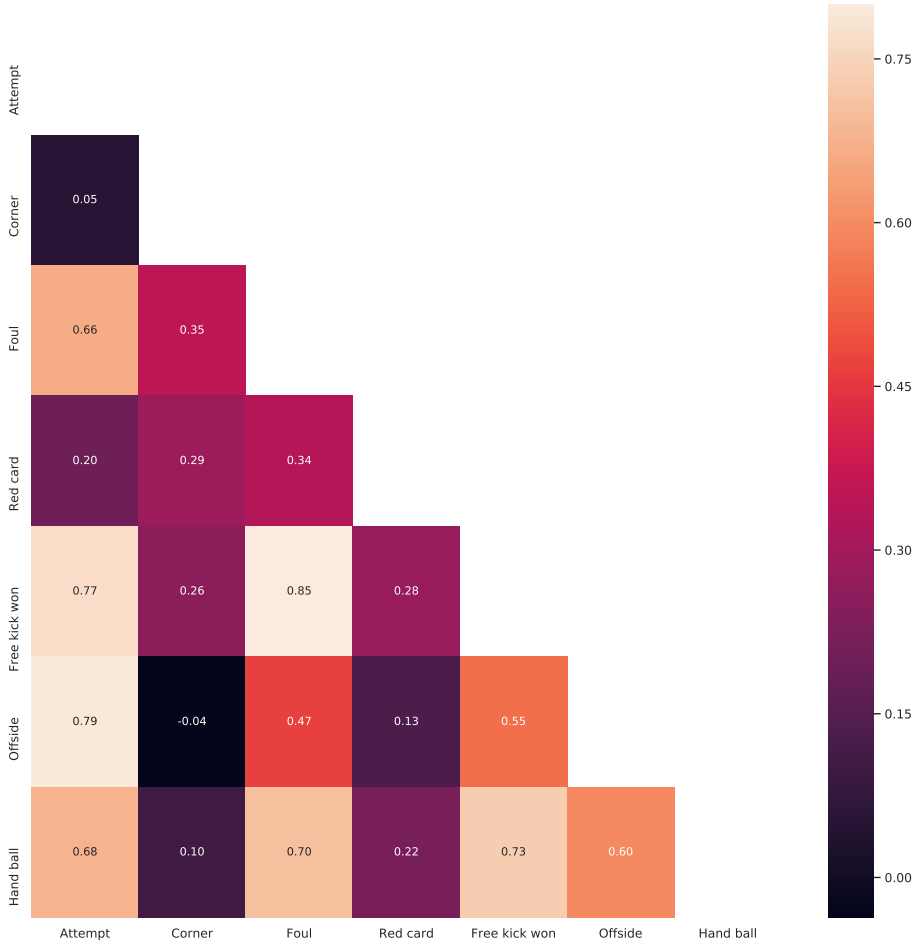


Figure 2.17: Correlation between red cards and team performance shows how players behave after one of their teammates is given a red card. We expound on the relationship between getting served a red card and other actions the team takes

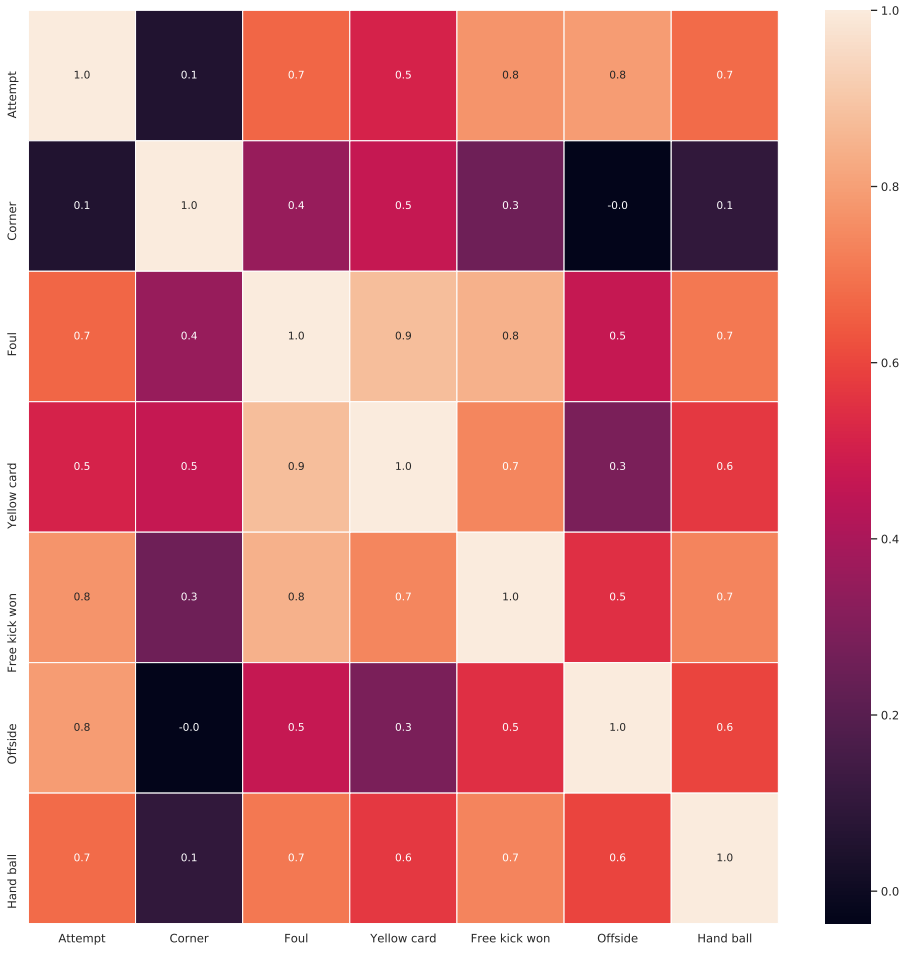


Figure 2.18: Correlation between first yellow card and team performance shows how players try improve performance through increased attempts, corners and reduced fouls in order to improve performance

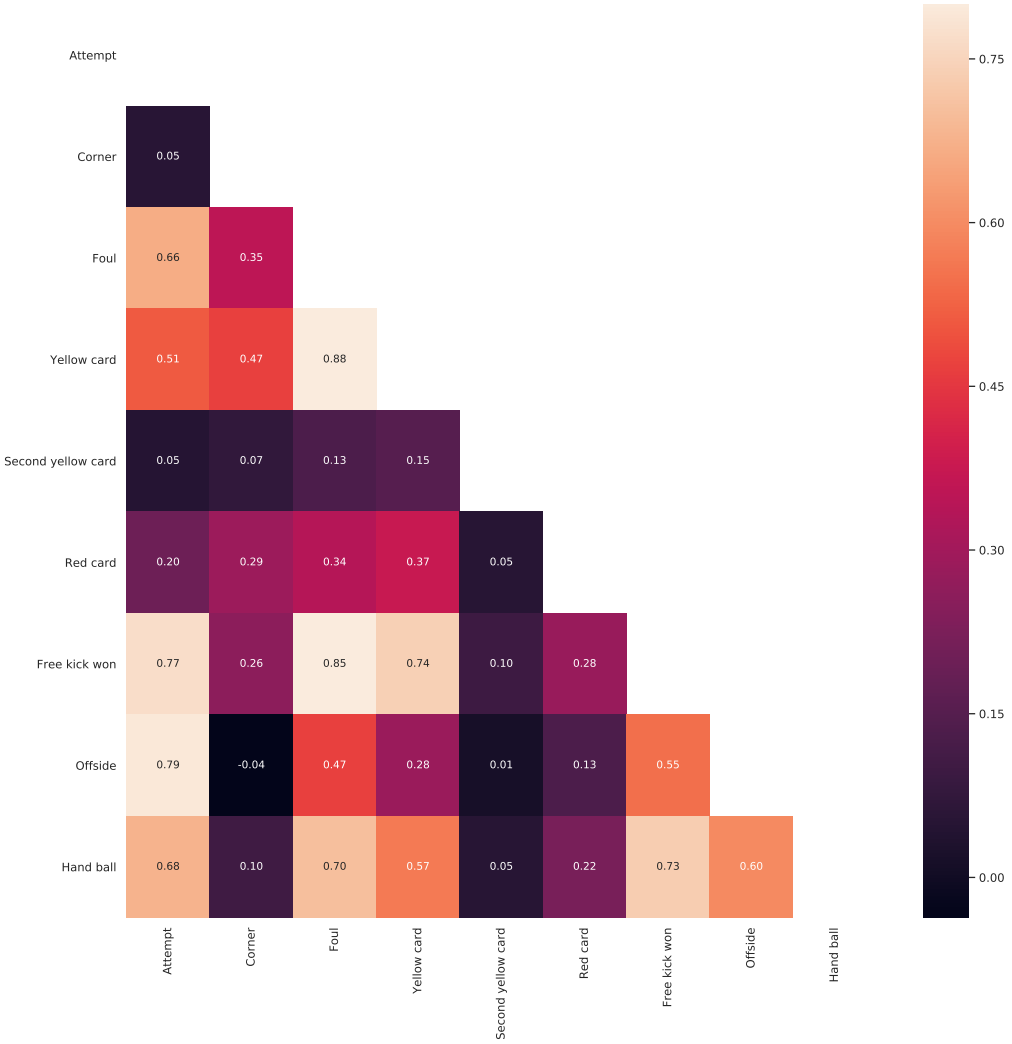


Figure 2.19: Correlation between getting a card and the team's performance

2.4 Evaluation of Characteristics of Teams in a League

2.4.1 Evaluation of whether the league is balanced

In this section, we seek to ensure that the teams in the leagues are all capable and competent. Although, the final ranking of teams at the end of the season indicates that teams are certainly not equal in competence and capability, it's not always the case. Some teams are more strategic than others in the league. We study the points and goals of each team to create a performance index. The index shows the ratio of the number of goals received and scored by each team which indicates the similarity or difference in their competence. If there is no noticeable difference in the ratios, then the teams are similar and if there is a big difference then the teams are different. The following figures show an example of a balanced and unbalanced league.

```
#####################################################
## the Informations about the Premier League 2015 ######
#####################################################
             ATeam ButsEncaisses ButsMarques RatioE RatioM RatioME
19  Manchester City           38          81    1.0   2.13    2.13
15          Chelsea           31          66   0.82   1.74    2.13
5           Arsenal           34          69   0.89   1.82    2.03
4    Manchester Utd           35          61   0.92   1.61    1.74
13      Southampton           30          52   0.79   1.37    1.73
3        Stoke City           43          47   1.13   1.24    1.09
17        Tottenham           53          56   1.39   1.47    1.06
8         Liverpool           47          47   1.24   1.24     1.0
14   Crystal Palace           47          46   1.24   1.21    0.98
1          West Ham           46          44   1.21   1.16    0.96
11          Everton           47          43   1.24   1.13    0.91
10          Swansea           47          43   1.24   1.13    0.91
6    Leicester City           52          44   1.37   1.16    0.85
0         West Brom           48          36   1.26   0.95    0.75
18            Hull           47          32   1.24   0.84    0.68
7         Newcastle           60          40   1.58   1.05    0.67
2               QPR           72          42   1.89   1.11    0.58
9           Burnley           49          27   1.29   0.71    0.55
16       Sunderland           50          27   1.32   0.71    0.54
12      Aston Villa           57          30    1.5   0.79    0.53
#####################################################
## the Informations about the Premier League 2016 ######
#####################################################
             ATeam ButsEncaisses ButsMarques RatioE RatioM RatioME
12        Tottenham           33          68   0.87   1.79    2.06
8           Arsenal           34          65   0.89   1.71    1.91
3    Leicester City           34          63   0.89   1.66    1.85
18  Manchester City           40          67   1.05   1.76    1.68
14      Southampton           39          56   1.05   1.51    1.44
5    Manchester Utd           35          48   0.92   1.26    1.37
19        Liverpool           49          61   1.29   1.61    1.24
11         West Ham           51          61   1.34   1.61     1.2
2           Chelsea           51          58   1.34   1.53    1.14
1           Everton           52          57   1.37    1.5     1.1
15          Swansea           49          40   1.29   1.05    0.82
16       Sunderland           57          46   1.54   1.24    0.81
13          Watford           48          38   1.26    1.0    0.79
7        Stoke City           54          41   1.42   1.08    0.76
9         West Brom           46          34   1.21   0.89    0.74
17   Crystal Palace           50          36   1.32   0.95    0.72
0       Bournemouth           65          44   1.71   1.16    0.68
6         Newcastle           63          42   1.66   1.11    0.67
4      Norwich City           65          38   1.71    1.0    0.58
10      Aston Villa           74          26   1.95   0.68    0.35
```

Figure 2.20: Premier League, an example of balanced league from 2015 to 2016

In figure 2.20, RatioME column represents the ratio between the number of goals scored and the number of goals in the net for each team. We notice that,for the season 2015,the gap between the first (Manchester City) and the last (Aston villa) is not too high, in comparison to what is shown in 2.21 which represents an unbalanced league.

```
#######################################################
## the Informations about the League One 2016 ######
#######################################################
                ATeam ButsEncaisses ButsMarques RatioE RatioM RatioME
17  Paris Saint-Germain           19          95   0.51   2.57     5.0
8                 Lyon           38          65   1.03   1.76    1.71
4                 Nice           35          54   0.95   1.46    1.54
0                Lille           26          38    0.7   1.03    1.46
6            Marseille           39          47   1.05   1.27    1.21
12           St Etienne           34          39   0.94   1.08    1.15
1           Montpellier           43          49   1.16   1.32    1.14
10            AS Monaco           48          53    1.3   1.43     1.1
14               Angers           35          36   0.95   0.97    1.03
11         Stade Rennes           50          49   1.35   1.32    0.98
7              Bordeaux           53          49   1.43   1.32    0.92
13             Guingamp           52          44   1.41   1.19    0.85
2                Bastia           40          32   1.08   0.86     0.8
16              Lorient           57          45   1.54   1.22    0.79
5                Nantes           40          31   1.08   0.84    0.78
9               Toulouse           52          40   1.41   1.08    0.77
15                 Caen           52          37   1.41    1.0    0.71
18       Stade de Reims           55          38   1.49   1.03    0.69
19          GFC Ajaccio           55          36   1.53    1.0    0.65
3                 Troyes           80          26   2.16    0.7    0.33
#######################################################
## the Informations about the League One 2017 ######
#######################################################
                ATeam ButsEncaisses ButsMarques RatioE RatioM RatioME
1             AS Monaco           19          61    0.9    2.9    3.21
19  Paris Saint-Germain           15          39   0.71   1.86     2.6
8                 Nice           14          34   0.67   1.62    2.43
10                Lyon           23          35   1.15   1.75    1.52
18           St Etienne           15          19   0.71    0.9    1.27
17             Guingamp           21          24    1.0   1.14    1.14
16             Toulouse           22          22   1.05   1.05     1.0
9             Marseille           25          24   1.19   1.14    0.96
6             Dijon FCO           29          26   1.38   1.24     0.9
5              Bordeaux           22          19   1.05    0.9    0.86
11         Stade Rennes           24          20   1.14   0.95    0.83
3           Montpellier           33          27   1.57   1.29    0.82
7     AS Nancy Lorraine           20          16    1.0    0.8     0.8
0                Bastia           24          18   1.14   0.86    0.75
15                Lille           23          17    1.1   0.81    0.74
13               Angers           26          17   1.24   0.81    0.65
2                  Caen           32          20    1.6    1.0    0.62
14              Lorient           40          21    1.9    1.0    0.53
4                  Metz           34          17    1.7   0.85     0.5
12               Nantes           28          13   1.33   0.62    0.46
```

Figure 2.21: League One, an example of unbalanced league from 2015 to 2016

In figure 2.21, we notice that for the season 2016, the gap between the first (Paris Saint-Germain) and the last (Troyes) is too high. this an example of unbalanced league.

2.4.2 Evaluation of dominant teams

Here, we focus on the five teams with the highest points at the end of the league. Knowledge of the dominant teams is essential in determining the teams that represent the league in the biggest world league called the Champions League. It is very competitive for teams to enter the champions league. The number of points obtained by a team depends on the scores of their matches during the season. A win counts for 3 points, a loss counts for 0 and a draw counts for 1 point. Using the data set, we determined the dominant teams using a function that calculates every teams' accumulated points by the end of the season. Figure 2.22 shows the the teams that dominated the Bundesliga league.

```
#####################################################
## the Informations about the Bundesliga 2012 ######
#####################################################
                      ATeam ATot ATotH BWin BWinH CDefeat CDefeatH Null Points
0          Borussia Dortmund   33    16   24    13       3        2    6     78
8              Bayern Munich   32    16   21    13       7        5    4     67
9                 Schalke 04   33    17   19    13      10        7    4     61
14  Borussia Monchengladbach   34    17   17     9       8        7    9     60
16          Bayer Leverkusen   33    17   15     8      10        5    8     53
5              VfB Stuttgart   33    16   15    10      11        7    7     52
6                Hannover 96   33    16   12    10      10       10   11     47
15             VfL Wolfsburg   32    17   13    10      14        9    5     44
11                  Nurnberg   32    16   12     6      14        8    6     42
10            TSG Hoffenheim   33    17   10     4      12        8   11     41
2              Werder Bremen   32    16   10     8      13        8    9     39
12               SC Freiburg   33    16   10     6      14        9    9     39
7                      Mainz   34    17    8     6      14        7   12     36
1                FC Augsburg   31    15    7     5      12        8   12     33
13                Hamburg SV   32    16    7     3      13        7   12     33
3              Hertha Berlin   33    16    7     4      16        7   10     31
4                 FC Cologne   32    16    8     5      19       12    5     29
17            Kaiserslautern   33    17    4     2      19        9   10     22
#####################################################
## the Informations about the Bundesliga 2013 ######
#####################################################
                      ATeam ATot ATotH BWin BWinH CDefeat CDefeatH Null Points
16             Bayern Munich   34    17   29    14       2        1    3     90
0          Borussia Dortmund   34    17   19    10       5        1   10     67
12          Bayer Leverkusen   34    17   20    13       7        5    7     67
13                Schalke 04   33    17   16    10      11        7    6     54
5                SC Freiburg   34    17   14     8      11        6    9     51
7        Eintracht Frankfurt   34    17   14     9      12        7    8     50
2                 Hamburg SV   34    17   13     7      14        7    7     46
6   Borussia Monchengladbach   34    17   12     8      12        8   10     46
14                  Nurnberg   34    17   12     8      12        8   10     46
15             VfL Wolfsburg   34    17   10     3      10        4   14     44
17               Hannover 96   33    16   12     8      15       12    6     42
1              VfB Stuttgart   34    17   11     5      15        7    8     41
8                      Mainz   34    17   10     7      13        7   11     41
9              Werder Bremen   34    17    9     5      16        9    9     36
4                FC Augsburg   34    17    8     5      17       10    9     33
11            TSG Hoffenheim   34    17    8     5      18       11    8     32
10        Fortuna Dusseldorf   34    17    7     5      18       12    9     30
3       SpVgg Greuther Furth   34    17    5     0      21        8    8     23
```

Figure 2.22: Teams that dominated Bundesliga from 2012 to 2013

According to 2.22,we see that Bayern Munich, Borussia Dortmund,Bayer Leverkusen and Schalke

teams are consistently in the top five.

2.4.3 Characteristics of the top teams in a league

From the previous section, we found that Bayern Munich, Borussia Dortmund,Bayer Leverkusen and Schalke were the top teams in the Bundesliga league. Knowledge of the dominant teams and their characteristics would help a coach improve their team's performance. The major characteristics we looked at in this project is the performance index which teams have in common locally (in the same league) and generally (all leagues). To correctly evaluate the performance index, we answered the questions below :

- What is the total number of victories, especially at home?
- What is the total number of goals, especially at home?
- What is the average number of goals per game?
- What is the total number of losses, especially at home?

We focus on home performances because they are very important in the leagues. Indeed, a team is at home when the match is played in their country or for the specific case of the leagues on their premises. A home match is special in the sense that in case of a perfect tie between two teams, the performances of each team at home and away from home can be used to decide between them. Therefore, when a team is at home, it must dominate the match to not give the chance to the opponent to take advantage of being away.

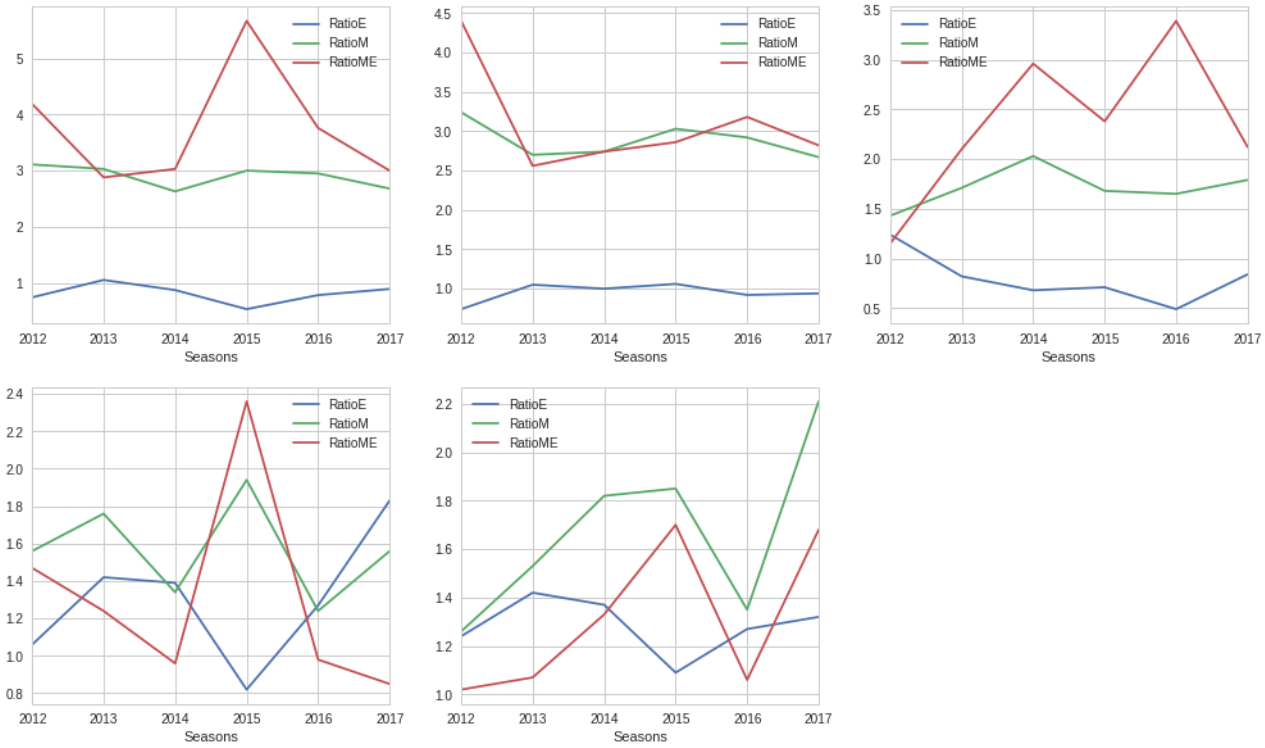


Figure 2.23: Top 5 Teams that dominated La liga from 2012 to 2017 in terms of Ratio

Figure 2.22 indicates that dominant teams score more goals their opponents every year.This performance index is verified not only for this league but for other leagues in the data set as well. Therefore, we can conclude that for a team to dominate the league, they have to score more goals than their opponents throughout the year.

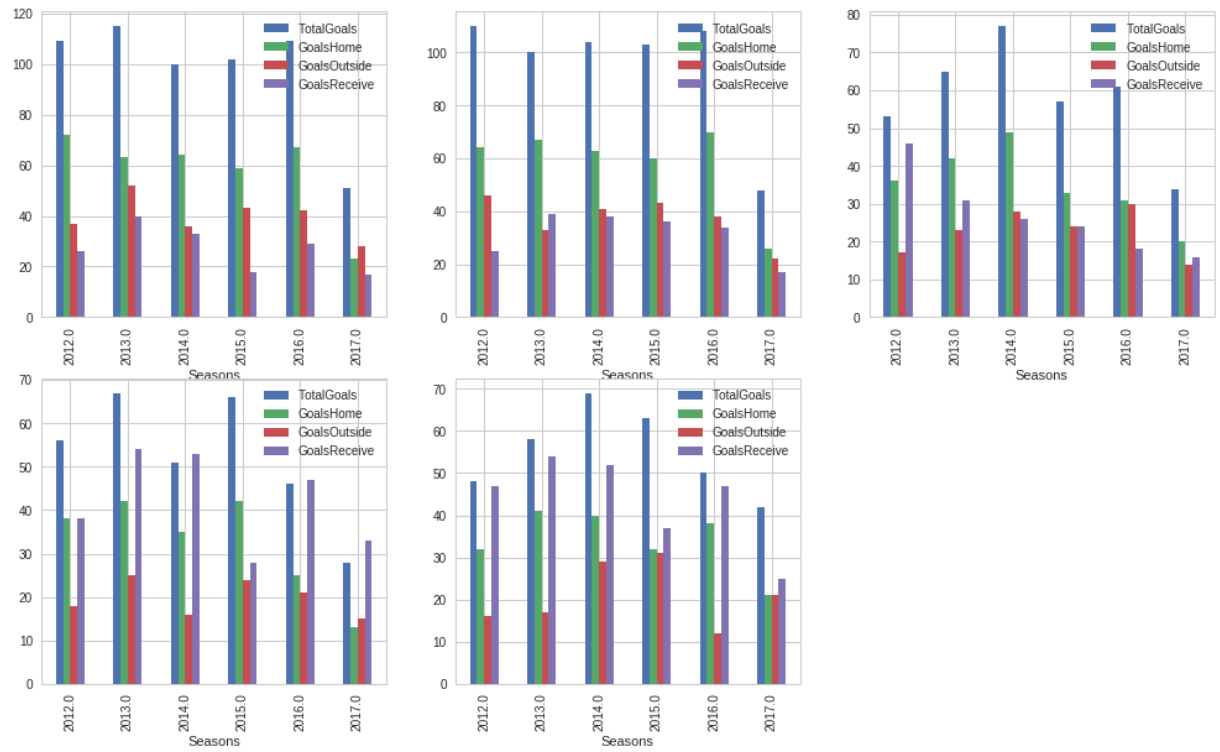


Figure 2.24: Top 5 Teams dominated La liga from 2012 to 2017 in terms of Number of Goals

this second figure gives more information about the performance of team at home. we can notice that the dominant team scores more goals at home than outside. this can be explain that, at home the needs to keep advantage in his side as it is already said . Therefore it's likely to score at home than outside.

```
#####################################################
## the Informations about the La Liga 2012 ######
#####################################################
              ATeam ATot ATotH BWin BWinH CDefeat CDefeatH Null Points
9       Real Madrid   34    17   29    15       1        1    4     91
8         Barcelona   35    18   26    17       2        2    7     85
1          Valencia   36    18   16    10       9        6   11     59
6   Atletico Madrid   37    19   15    11      12        9   10     55
17          Levante   35    18   16    11      12        8    7     55
#####################################################
## the Informations about the La Liga 2013 ######
#####################################################
              ATeam ATot ATotH BWin BWinH CDefeat CDefeatH Null Points
5         Barcelona   38    19   31    18       2        2    5     98
4       Real Madrid   37    19   25    16       5        5    7     82
19  Atletico Madrid   38    19   23    14       8        5    7     76
13    Real Sociedad   37    18   18    10       7        5   12     66
17         Valencia   38    19   19    13      11        8    8     65
#####################################################
## the Informations about the La Liga 2014 ######
#####################################################
              ATeam ATot ATotH BWin BWinH CDefeat CDefeatH Null Points
18  Atletico Madrid   38    19   28    15       5        5    5     89
3         Barcelona   38    19   27    16       5        4    6     87
5       Real Madrid   38    19   27    16       5        3    6     87
11  Athletic Bilbao   38    19   20    13       8        6   10     70
4           Sevilla   38    19   18    11      12        8    8     62
#####################################################
## the Informations about the La Liga 2015 ######
#####################################################
              ATeam ATot ATotH BWin BWinH CDefeat CDefeatH Null Points
7         Barcelona   34    17   27    15       3        2    4     85
9       Real Madrid   34    18   26    15       6        5    2     80
10         Valencia   34    17   21    15       3        3   10     73
0           Sevilla   34    17   21    12       6        5    7     70
15  Atletico Madrid   34    16   20    11       5        3    9     69
#####################################################
## the Informations about the La Liga 2016 ######
#####################################################
              ATeam ATot ATotH BWin BWinH CDefeat CDefeatH Null Points
14        Barcelona   37    19   28    16       5        3    4     88
12      Real Madrid   37    19   27    16       4        2    6     87
2   Atletico Madrid   37    18   25    13       6        5    6     81
10       Villarreal   37    19   18    12       9        6   10     64
13       Celta Vigo   37    19   17     9      11        7    9     60
```

Figure 2.25: Top 5 Teams dominated La liga from 2012 to 2017 in terms of Number of victories

We all know that to be on top, you must win as much as possible of your matches. this figure highlights some relevant information. you can easily notice that the dominant teams win more matches at home. In other words, they don't loss easily at home. for example, if we take Real Madrid in 2012, the figure tells us that the team plays 34 matches , 17 matches were hosted by them,they win 29 matches this season with 15 victories at home and they only loss 1 match at home. The same remark can be done for all the others teams in the top 5 of the others leagues.

2.4.4 Best teams

In the world of football, each league ranks the participating teams in descending order of the number of points obtained at the end of the season. This Local Ranking by Country, gives a global information on the best teams in the world. In other words, this ranking allows us to know the most offensive teams, the most defensive teams, the championships where we score the most goals and so on. At the beginning of each season, each team must adopt a strategy to maintain or to have a good place in this ranking to maintain the top of the world rankings and keep some advantages such as advertising and a high odds for the bets. Note also that a football team is a company , means that it makes investments early in the season and expect gains at the end. The strategy of a team gathers at the same time, the line of conduct of the players during the matches and the decisions that must take the administrative staff for the smooth running of the team. The ultimate goal is the fulfillment of everyone, the heavy task is given to the Coach, who must both present the weaknesses and strengths of his team to allow the financial staff to make investments in the direction of bringing solutions to the problems highlighted. To do this, the coach must at the same time,make a study of the performances of his team compared to the other teams of the league. Hence the objective of this part of our investigation will generally consist of presenting the best teams of each league during the seasons and then give performance indices that characterize them.

To achieve this goal, we will successively answer certain questions that will not guide us in assessing the performance of a team. The main questions are:

- What type of league is it?
- Is the league balanced?
- Which teams have dominated the championship since time?
- What are the characteristics of these teams?

2.4.5 What type of league is it?

The answer to this question gives a general idea about the type of league. Is it offensive? Is it aggressive? Etc ... As part of this work, we will focus mainly on the following criteria: the number of goals per game, the number of cards per game. The first test demonstrates the offensive character of the league. In this type of championship, the priority is to score as many goals as possible and there are usually high scores of matches. In other words, the championship is rich in goals. The second character shows the aggressive rating and defensive league. In this type of league, the number of cards (yellow / red) that is proportional to the number of faults committed is generally very high.

To detect the league type of our data set, we use two functions that each enter a league and give results for each season, the average number of goals per game and the number of cards per game. The results are given by the following figure:

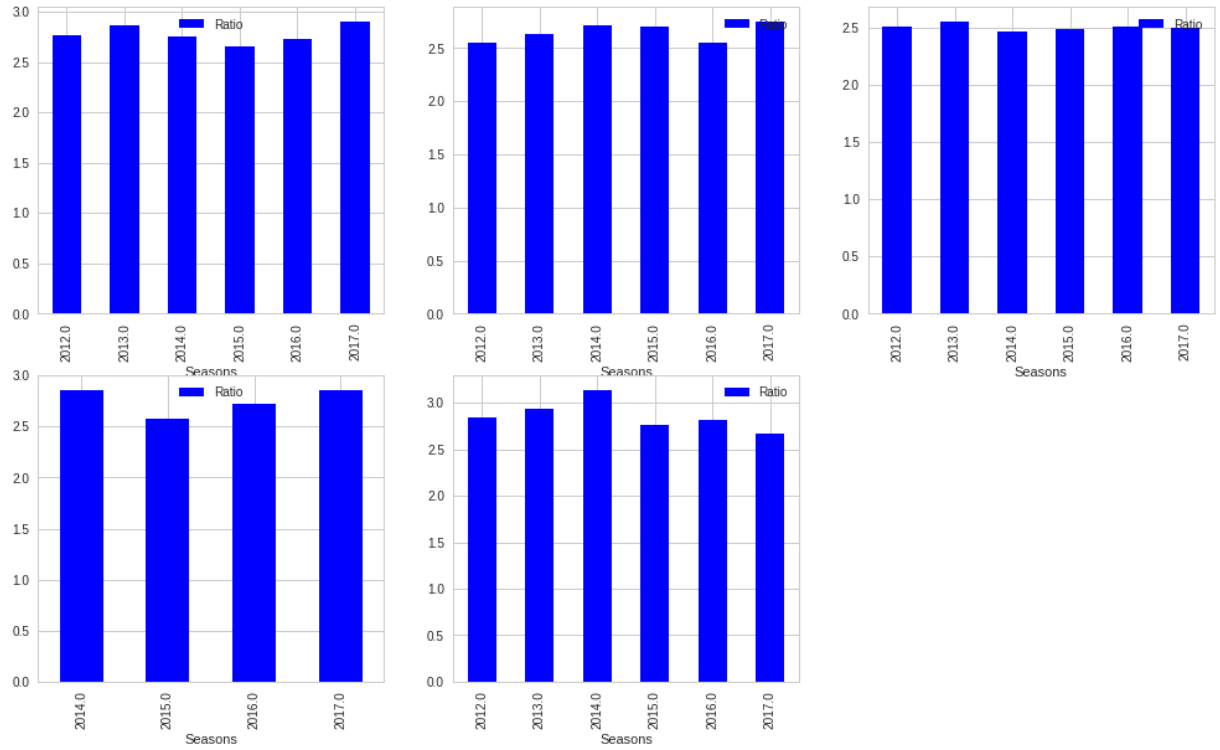


Figure 2.26: Average of goals per match of the leagues from 2012 to 2017

We note that we have on average more than two goals per game since 2012 to 2017 of all the leagues is offensive.we realized that all these leagues are offensive although the density of number of goals per season is not the same. This is because not all leagues have the same number of teams and therefore not the same number of matches. As a result, teams with more matches are more likely to score than teams with few matches.

2.4.6 Weakest teams

In this part we will do the opposite of what was done in the previous part. Here we will give the characteristics of weak teams. the objective of this part is to highlight the weaknesses of these teams. These characteristics are points on which these teams must improve to have better rank in the final ranking. Just like in the previous part, we will observe the performances in general and in particular the performances at home.

```
## the Informations about the League One 2012 ######
#####################################################
         ATeam ATot ATotH BWin BWinH CDefeat CDefeatH Null Points
13     Lorient   38    19    9     8      17       12   12     39
4         Nice   35    16    9     6      16       12   10     37
1         Caen   36    17    8     5      17       10   11     35
9    Dijon FCO   37    19    8     6      19       11   10     34
18  AJ Auxerre   38    19    7     6      18       11   13     34
#####################################################
## the Informations about the League One 2013 ######
#####################################################
                   ATeam ATot ATotH BWin BWinH CDefeat CDefeatH Null Points
3   Evian Thonon Gaillard   38    19   10     6      18       12   10     40
13             AC Ajaccio   37    19    8     7      15        8   14     38
16                 Troyes   37    18    8     6      15       11   14     38
18      AS Nancy Lorraine   36    17    8     4      17        9   11     35
8                   Brest   37    19    8     5      24       13    5     29
#####################################################
## the Informations about the League One 2014 ######
#####################################################
           ATeam ATot ATotH BWin BWinH CDefeat CDefeatH Null Points
15          Nice   38    19   12    10      20       13    6     42
19        Nantes   37    18   10     5      15        8   12     42
9        Sochaux   37    18   10     8      18       13    9     39
2   Valenciennes   38    19    7     4      23       13    8     29
8     AC Ajaccio   38    19    4     3      23       13   11     23
#####################################################
## the Informations about the League One 2015 ######
#####################################################
                   ATeam ATot ATotH BWin BWinH CDefeat CDefeatH Null Points
14               Toulouse   38    19   12     8      20       15    6     42
0          Stade de Reims   38    19   11     7      18       10    9     42
4   Evian Thonon Gaillard   38    19   11     7      23       13    4     37
12                  Metz   38    19    7     6      22       13    9     30
11                  Lens   38    19    7     5      23       13    8     29
#####################################################
## the Informations about the League One 2016 ######
#####################################################
             ATeam ATot ATotH BWin BWinH CDefeat CDefeatH Null Points
16         Lorient   37    18   10     6      14        9   13     43
9         Toulouse   37    19    8     6      16       10   13     37
19     GFC Ajaccio   36    19    8     5      15        8   13     37
18  Stade de Reims   37    18    9     6      19       12    9     36
3           Troyes   37    18    3     1      26       15    8     17
```

Figure 2.27: 5 weakest Teams for League One from 2012 to 2017 in terms of Number of victories

In general, we can notice that these teams are the opposite of the Dominant Teams . they loss more matches at home than they win.That means that they give advantage to the opponents even when they are at home.

3. Predictive Models

The goal of this report is to extract as much information from the football events dataset to achieve the main objective which is to provide football decision makers with useful insights by spotting weaknesses and strengths in the teams/players in order to take decisions aiming to improve the team performance. We came up with two predictive models that can give insights from the dataset. One is to predict whether an event will become a goal or not and the second is to predict the match final result based on the odds of the match. The pipeline used in the modeling tasks is:

- Data cleaning.
- Exploratory data analysis. (performed comprehensibly in the previous chapter)
- Feature selection.
- Try several machine learning models on a performance metric.
- Evaluate the models on the testing set.
- Interpret the model results.
- Draw conclusions.

3.1 Predicting a Goal from Events

3.1.1 Data cleaning

Data cleaning goal is to transform the data into a format that is appropriate for any machine learning predictive model that will be used. Usually in this step we handle missing values, correct the data types of the features and remove outliers. In the events dataset we have many of the features that have missing values as illustrated in Table 3.1. We tried two solutions for handling missing values: one is to fill them with a new category 'UNK' which means (Unknown) and the second solution is to remove the null values and we will report the prediction results with both methods.

Table 3.1: Proportions of missing values in the dataset

Feature name	Number of Missing Values	% of Total Values
player_in	889294	94.5
player_out	889271	94.5
odd_over	842329	89.5
odd_under	842329	89.5
odd_bts	842329	89.5
odd_bts_n	842329	89.5
assist_method	773104	82.2
event_type2	726716	77.2
shot_place	713550	75.8
shot_outcome	712511	75.7
situation	711872	75.6
bodypart	711824	75.6
player2	649699	69.0
location	473942	50.4
player	61000	6.5

For the data types of the features, all the categorical features have the 'object' data type and the numerical features have appropriate data types whether float for odds features or int for 'time', 'fthg' and 'ftag' features but they had big unneeded size like 'int64' or 'float64'. We converted the 'object' type to 'category' type for the categorical features and assigned similar but smaller size data type for the numerical features.

3.1.2 Feature selection

As it can be seen from the dataset description section, most of our variables/features are categorical features and our target variable is categorical. We relied on our domain knowledge to choose the features. We decided to manually choose 6 categorical features, 3 numerical features, and one engineered boolean feature 'first_half' used in the prediction tasks which indicates whether the event happened in the first half of the match or in the second half. The chosen features and their data types can be shown in Table3.2.

Table 3.2: Manually chosen features

Feature name	Information
odd_h	941009 non-null float16
odd_d	941009 non-null float16
odd_a	941009 non-null float16
assist_method	941009 non-null category
location	941009 non-null category
side	941009 non-null category
shot_place	941009 non-null category
situation	941009 non-null category
bodypart	941009 non-null category
first_half	941009 non-null bool

3.1.3 Machine learning models evaluation

Our problem is a binary classification which can be addressed with many machine learning models. In order to apply machine learning models we transformed the categorical features with one-hot encoding (for features with more than 2 categories) and a binary mapping (for features with exactly 2 categories). In this section we are going to train our data with the following models:

- Logistic Regression (LR): A widely used predictive model, in the binary logistic regression form (our case), it uses the logistic function to model the target variable by assigning probability to each outcome by maximizing the likelihood function.[9]
- Random Forest (RF): It is one of the ensemble methods that combines many decision trees to have more accurate prediction. One of the reasons to use RF model is that it can show us the relative importance of each feature on the prediction.[3]
- Gradient Boosting (GB): It is one of the ensemble techniques in which the estimators/predictor models are not independent. In this technique it is sequential training; each predictor is learning from the mistakes of the previous one. [5]

After substituting the null values in the chosen features, we applied the previous machine learning models and measured the performance by two metrics: accuracy and balanced accuracy; because accuracy is not capturing the class imbalance problem if exist and it measures how many correctly classified data points without taking into consideration the ratio of true negatives and true positives which is the difference between accuracy and the balanced accuracy.

Table 3.3: Machine learning models evaluation [after substituting null values with "UNK"]

	Imbalanced classes		Balanced Classes with Random Under Sampling	
Model name	Accuracy	Balanced Accuracy	Accuracy	Balanced Accuracy
LR	98.42 %	80.37 %	94.51 %	97.18 %
RF	98.2 %	78.08 %	95.14 %	97.01 %
GB	98.39 %	75.77 %	94.48 %	97.14 %

In our task we figured that we have class imbalance problem which is shown in figure 3.1. We applied three techniques to handle the imbalance in classes. The first is random under sampling, which randomly removes data points from the majority class. The second is random over sampling, and this performs the opposite of under sampling by generating new samples by randomly sampling with replacement from the minority class samples. The last technique is Synthetic Minority Oversampling Technique (SMOTE) for numerical and categorical features SMOTENC. SMOTE technique over sample the minority but with a different technique than random sampling. It interpolates the values of the newly generated data points after choosing K-nearest neighbours from a minority data point. In table 3.3 we show the two metrics (accuracy and balanced accuracy) with the 4 models and with applying the first method of handling missing values (substituting null values with 'UNK' category) without handling the class imbalance and with handling it with (random under sampling) method.

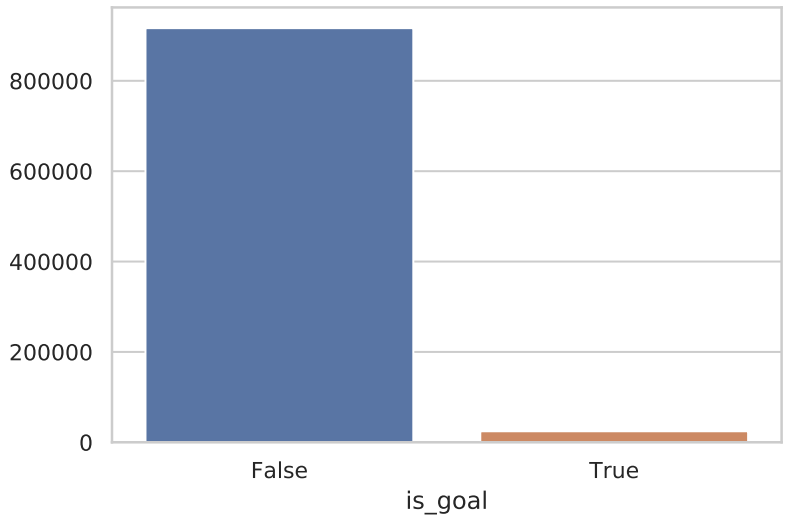


Figure 3.1: Imbalanced Classes

Table 3.4: Machine learning models evaluation [after removing null values]

	Balanced Classes with SMOTENC		Balanced Classes with Random Over Sampling	
Model name	Accuracy	Balanced Accuracy	Accuracy	Balanced Accuracy
LR	83.68%	88.87%	83.11%	89.28%
RF	89.21%	80.07%	90.91%	75.25%
GB	80.96%	88.60%	81.36 %	88.82%

After exploring how each feature contributed to the number of goals, it is obvious that filling the null values by 'UNK' was not a good solution. We thought of removing the null values and have less number of data points in order to achieve more reliable results. We applied two techniques for over sampling that we discussed before: random over sampling and SMOTENC. The results for the accuracy and balanced accuracy is shown in table 3.4. As we can see both metrics dropped from above 90s to range from 80 to 90, and this is understandable because most of the null values contributed heavily towards 'no_goal' as shown in figures 3.3, 3.2 and 3.4.

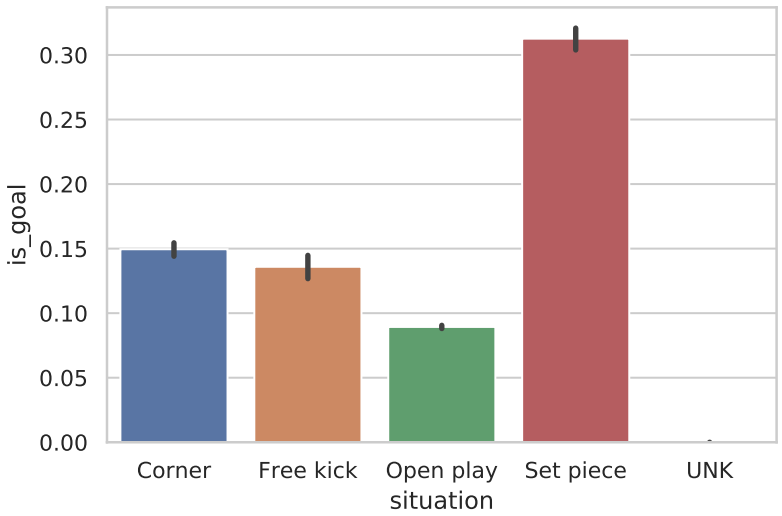


Figure 3.2: Situation contribution to 'is_goal' variable

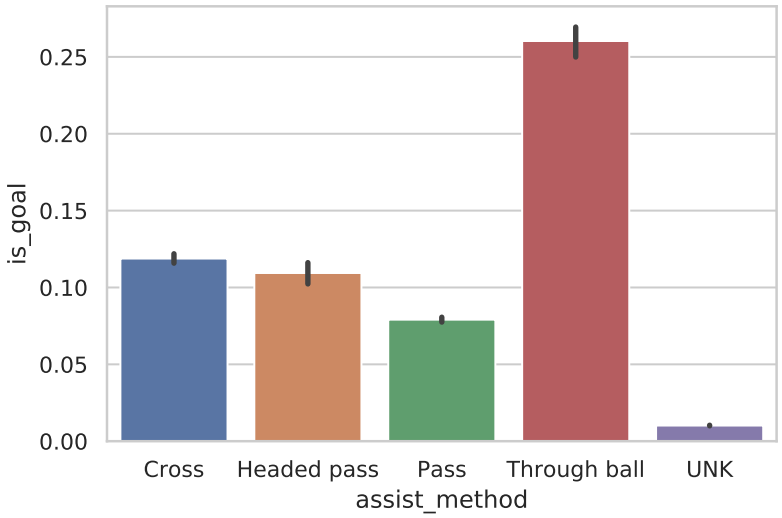


Figure 3.3: Assist method contribution to 'is_goal' variable

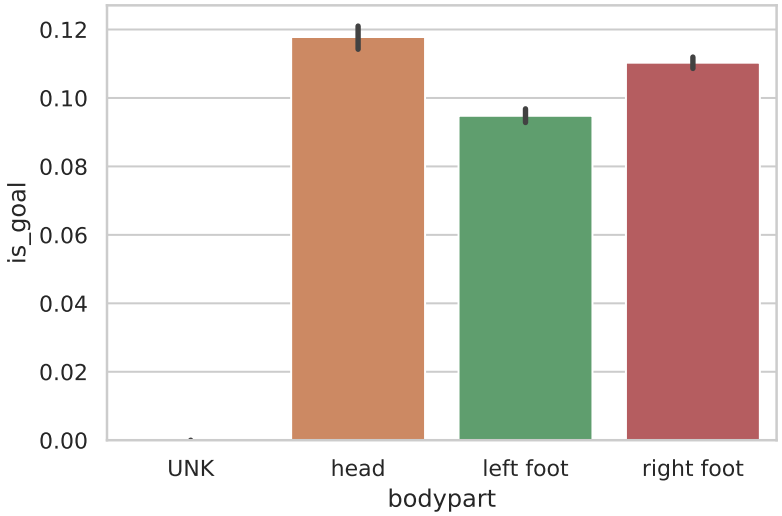


Figure 3.4: Body part contribution to 'is_goal' variable

3.1.4 Interpreting model results

This task is important not for achieving the highest accuracy only, but also to interpret what features might be important towards being goal. Such features can give insights to coaches to build better training strategies in the future. As a general overview of what features are important we might need to look at the relative weights trained during the random forest model as shown in figure 3.5.

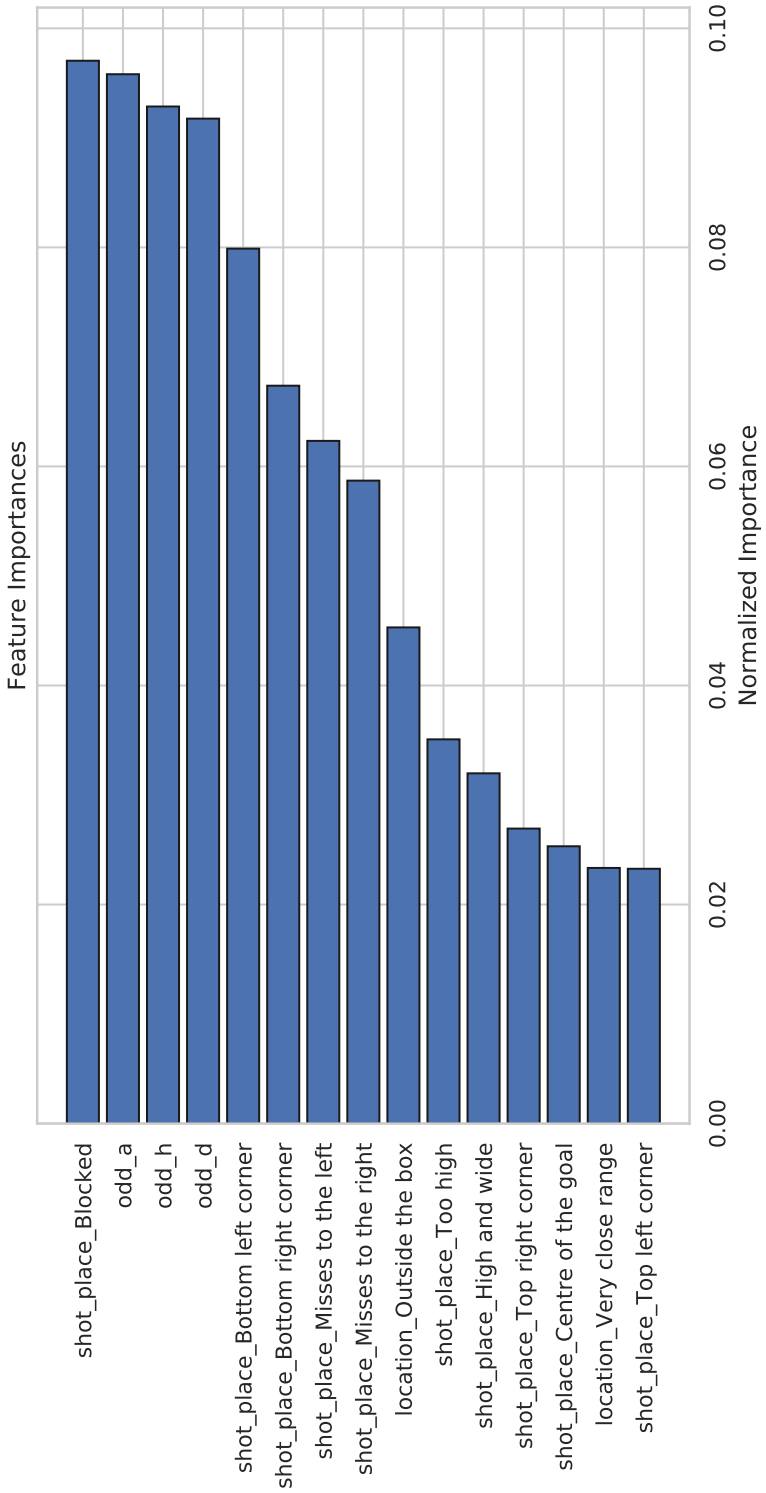


Figure 3.5: Features importance from RF model [after removing null values + SMOTE]

What features importance that random forest models give us is an overview of the most important

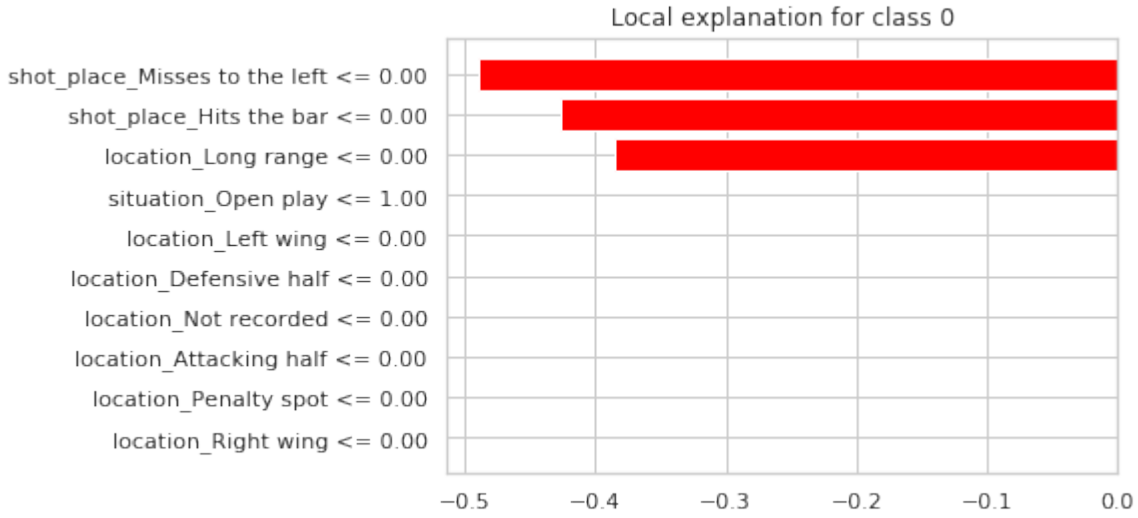


Figure 3.6: LIME explanation for wrong prediction with Logistic Regression

features in the whole model. Instead, if someone is using this model and she wants to inspect for a given event, what were the most important features that force specific prediction, features importance will not be beneficial. For this reason we are using one of the recently published machine learning black-box model interpretation algorithms which is called LIME. LIME is short for (local interpretable model-agnostic explanations). It aims to explain individual predictions of any machine learning model that provide probability predictions for the target classes. For example, figure 3.6 shows the explanation for class 0 (i.e. not a goal class) of this prediction that was wrongly classified as a 'goal'.

3.1.5 Deep learning model

In machine learning models in order to use any of the predictive models, we needed to encode the categorical features as one-hot encoding. This technique can also be used if we trained a deep neural network. But instead, we used embeddings for the categorical features which constitute a unique representation for each feature. Using the embeddings and the other numerical features we trained a fully connected neural network with 2 layers (200 and 100 units in each respectively) with both techniques of handling missing values (substituting the null values with 'UNK' class and removing the null values then applying one of the over sampling techniques). The deep learning model outperformed in all cases reaching an accuracy over 98% with the ability to add more features that we could not add in a normal machine learning settings (e.g. event_team); because the number of columns created with the one-hot encoding for the event_team feature exceeded the capacity of the computation resources we have.

3.2 Predicting Match Results from the Odds

Odds are the amount of money that a better will receive for every 1 $ he bets on that result (because we have odds on winning at home, away or the drawing). For example, if the odds for a draw is 5, you will get 5$ for every 1$ you bet on a draw. Bookmakers tend to put high odds on the winning of the weak team and vice versa (i.e. they tend to put lower odd ratio for the most probable result to maintain a margin of profit in any case).

We used the odds features for the three cases (home win, away win and draw) and the difference in the odds as features. We tried to use the odds to predict the match final results (win, lose, or draw) but the accuracy was very low. We used logistic regression, gradient boosting, random forest and a fully connected network with embeddings for categorical features. The accuracy score of the used models are summarized in table 3.5.

Table 3.5: Prediction models evaluation for game result from the odds

Model name	Accuracy
LR	54.8 %
RF	48 %
GB	54.2 %
Deep learning	52 %

3.3 Predicting Number of Goals in a Match

3.3.1 Motivation

In this section, we are trying to build a model predict the number of goals in a match according to match events.

3.3.2 Data Understanding

To build this model we select 6 features.

1. features To build this model we select 6 features
   (a) Event type: Corner, Foul, Substitution, Red card, Yellow card, Hand ball, Offside, etc.
   (b) Location: Centre of the box, Outside the box , Left side of the six yard box, Long range, etc.
   (c) Shoot place: Too high, Bit Too high, Bottom left corner, Top centre of the goal , etc.
   (d) shoot outcome: On target, OFF target, Blocked or Hit the bar.

3.3.3 Data pre-processing

1. Home and away goals: First we separate each event into home or away match according to data-set, Also we run othor experiment by adding home and away matches which they have the same match ID and we summation the number of goal in home and away matches.
2. One hot encoding: We run the experiments into the original data-set (integers values), also we convert the data-set to binaries to accelerate the training.
3. Missing data: we put any missing data equal -1, because we have integers values and zeroes.
4. labels: Our labels will be number of goals per match, so we make it like our features (integers and binary-encoded for two separated experiments).
5. Vectored: We select 5 features (over all we select 6 features but we use side as anther ID to separate matches have the same ID match), In every match we have 180 events (maximum number of events per match) so we have 2-d array size (5 X 180), we convert it into 1-d array (1 X 900).

3.3.4 Methodology

We use Recurrent Neural Network (RNN) as a classifier,

1. The inputs will 2d-array has size (number of matches X Vectored features events).
2. Labels: numbers of goal in a match.
3. Output: prediction number of goals in this match (Float number we round it).

3.3.5 Network-Model

1. Training data: Events of 2000 matches.
2. Test data: Events of Events100 matches.
3. accuracy: round(prediction)== label.
4. Loss: Mean Square Error (MSE).
5. Optimizer: Adam.
6. Device: CPU.

3.3.6 Hyper-parameters

1. Batch size: 100
2. Learning rate: 1e-3
3. Weight decay: 1e-3

3.3.7 Experiments

Combined Matches:Home and Away matches in one match

In this experiment, we summation all events in home and away matches which are sharing the same match ID. Also we summation number of goals in home and away matches.

Separated Matches: Home Match and Away Match

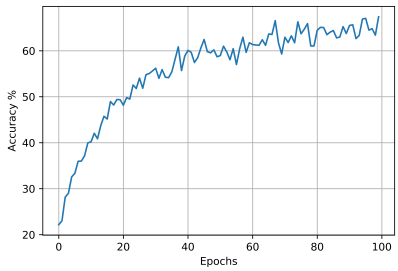
In this experiment, we separate home and away matches, Also we separate number of goals in home and away matches.

1. Home matches without one-hot encoding.
2. Away matches without one-hot encoding.
3. Home matches with one-hot encoding.
4. Away matches with one-hot encoding.

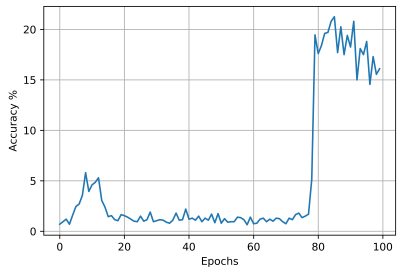
3.3.8 Results

We focus on measuring the accuracy and training time in each experiment.

Separated Matches:



(a) Accuracy Home matches without one-hot encoding



(b) Accuracy Away matches without one-hot encoding

Figure 3.7: Accuracy of Separated Matches

The average accuracy shown in table below:

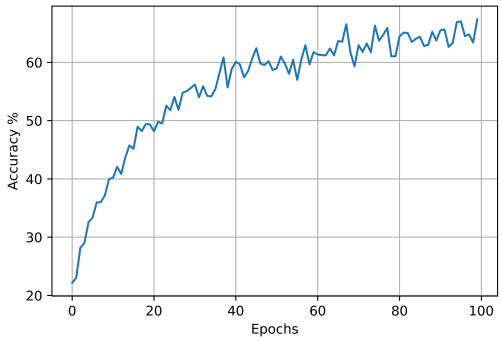
Home without encoding	Away without encoding
56.017%	5.098%

The training time (in minutes) shown in table below:

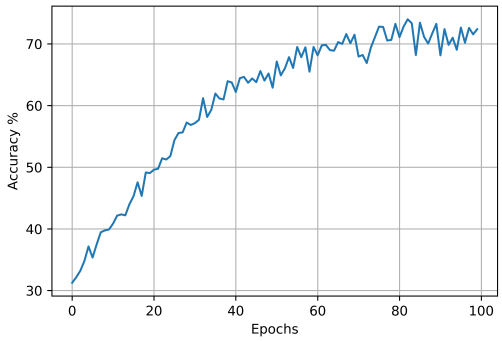
Home without encoding	Away without encoding
83.967	108.25

Combined Matches:

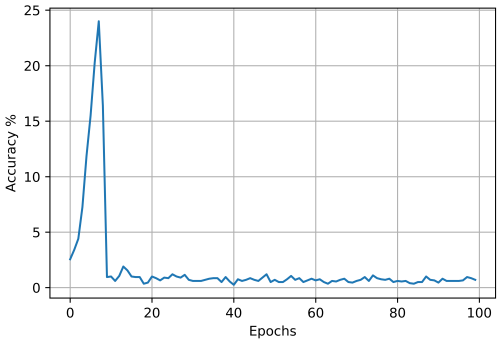
We run this model for 4 model components. in figure 3.8



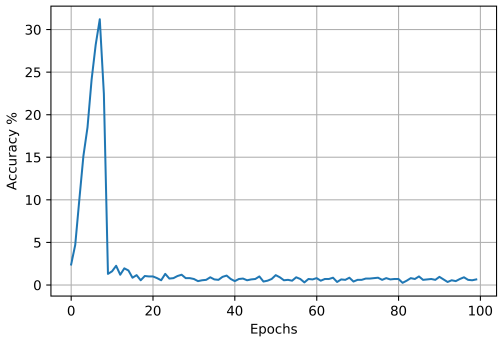
(a) Accuracy Home matches without one-hot encoding



(b) Accuracy Away matches without one-hot encoding



(c) Accuracy Home matches with one-hot encoding



(d) Accuracy Away matches with one-hot encoding

Figure 3.8: Accuracy of Combined Matches

The average accuracy shown in table below:

Home without encoding	Away without encoding	Home with encoding	Away with encoding
56.017%	60.937%	1.726%	2.27%

The training time (in minutes) shown in table below:

Home without encoding	Away without encoding	Home with encoding	Away with encoding
83.067	80.467	84.05	83.033

4. Challenges and Lessons Learned

In this chapter, we highlight the challenges faced by the team during the project execution and how they were resolved. Additionally, we talk about the lessons we learned during this experience.

4.1 Challenges Encountered

4.1.1 Technical challenges

Data engineering

- Difficulty in ascertaining the relevance of different features in modeling team performance.
- Large number of missing values in the dataset, which forced us to use smaller subset of the whole dataset, and this affected the accuracy of the predictive models.
- Absence of data that is to say missing values making it extremely hard to make predictions.
- Unclassified data like for example values given for unclassified body parts making it hard to know which specific feature contributed to the team performance.

4.1.2 Non-technical issues

- Lack of domain knowledge for over 50% of the team members.

Team collaboration

We started with two long meetings to come up with the main objective since we are given a dataset without clear problem formulation. We had to come up with meaningful questions that we can answer through data analysis and predictive models. We also performed a daily (maximum every two days) stand-up meeting to check each one status.

4.2 Steps Taken to Solve the Challenges

- We read papers and blog articles to get a deep understanding of the topic.
- We created GitHub, Google and overleaf documents to make collaboration easier to coordinate.
- We created a WhatsApp group to facilitate consistent communication in the amongst team members.

4.3 Lessons Learned

- Most of us learned how to use git and GitHub in a collaborative setting.
- We learned how to analyze, model and make predictions given a Kaggle data set. We all got a chance to experience to experiment with the platform and learned how to make meaningful contributions.
- We had a knowledge about how machine learning pipeline works for prediction tasks.
- Gained knowledge about dealing with categorical features practically, since most of our features are categorical.

5. Future Works and Conclusions

Football data analysis is wide area of research that affect many people in many fields: starting from betters, lovers till the entities that invest heavily in the football industry. The dataset we had did not have some of other useful information that would make the models more intuitive, but we could do data analysis to gain useful insights from the events data. Here are some of the future work:

- In the future work , we will try to use this dataset to predict the outcome of each league and then with that predict which team will be able to attend the champion league and which team will win the champion league.
- If we have the player prices among the years we would predict how the price will become according to his performance.
- Go deeper, add more layers and batch normalization.

Bibliography

[1] Dictionary of football terms. https://www.optasports.com/news/optas-event-definitions/. Accessed: 2018-12-7.

[2] Histoire du football. https://fr.wikipedia.org/wiki/Histoire_du_football. Accessed: 2018-12-6.

[3] Niklas Donges. Random forest model. https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd. Accessed: 2018-12-7.

[4] Stefano Dóttavio, Mario Esposito, Antonio Lombardo, Laura Pantanella, Bruno Ruscello, and Tommaso Valente. Socialsoccer.

[5] Prince Grover. Gradient boosting model. https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d. Accessed: 2018-12-7.

[6] Mike Hughes and Ian Franks. Analysis of passing sequences, shots and goals in soccer. *Journal of Sports Sciences*, 2005.

[7] Belle Selene Xia and Peng Gong. Review of business intelligence through data analysis. *Benchmarking: An International Journal*, 21(2):300–311, 2014.

[8] Manuel Stein, Halldór Janetzko, Daniel Seebacher, Alexander Jäger, Manuel Nagel, Jürgen Hölsch, Sven Kosub, Tobias Schreck, Daniel A Keim, and Michael Grossniklaus. How to make sense of team sport data: From acquisition to data modeling and research aspects. *Data*, 2(1):2, 2017.

[9] Wikipedia. Logistic regression definition. https://en.wikipedia.org/wiki/Logistic_regression. Accessed: 2018-12-7.