European Football Database (Spanish La liga 2021-22 season)

Soham Sengupta

Roll No.- MDS202241

Email- sohams@cmi.ac.in

October 28,2022

Abstract

In this project we have taken a European football database data on Spanish football top flight league (la liga) for 2021-22 season. We use visualization techniques in R to evaluate the unpredictability of the game and performance of teams , along with the variation in odds provided throughout the match by various betting providers.

Introduction

Sports betting is a rapidly growing industry that obtained a worldwide market size of over 200 billion United States (US) dollars in 2019 (Ibisworld, 2020). In total, there are over 30,000 sports-betting-related businesses globally (Ibisworld, 2020). Prior to the COVID-19 pandemic, the sports-betting industry in the regions of Asia, the Middle East, and South America had grown at above-average rates (Ibisworld, 2020), while in 2021 weekly sports betting in the United States doubled (Morning Consult, 2022).

Football is the most popular sport in the world , evidently betting in football also is quite common. In view of unpredictability of the game , betting providers use much statistical approach in estimating the odds. The birth and subsequent explosion of the internet at the end of the 1990s truly revolutionised the way we bet on football. The new school bookies, such as Betfair and bet365, began to pop up in abundance, offering an entirely online operation. Football data became more easily available and betters at individual level started to try comeup with their own predictions.

Here we take the data on Spanish top division (La Liga) data for 2021-22 season and see the changes on betting odds provided by different betters with course of match and how they relate to various match results. Also we see the unpredictability of a match by seeing the extent of change in result at full time as compared to that in halftime.

Dataset Description

The data is obtained from the European Football Database (https://www.football-data.co.uk/). This is a database that have detailed match data of different tiers of europe's

top 11 football leagues from 1993 onwards. For our purpose , Spanish top division (La Liga) data of 2021-22 season has been taken. The data set provides detailed data of every matcg including result, shots taken , shot on target, fouls committed, disciplinary statistics and more. It also provides betting odds data by different betting providers.

Key variables-

Div = League Division

Date = Match Date (dd/mm/yy)

Time = Time of match kick off

HomeTeam = Home Team

AwayTeam = Away Team

FTHG and HG = Full Time Home Team Goals

FTAG and AG = Full Time Away Team Goals

FTR and Res = Full Time Result (H=Home Win, D=Draw, A=Away Win)

HTHG = Half Time Home Team Goals

HTAG = Half Time Away Team Goals

HTR = Half Time Result (H=Home Win, D=Draw, A=Away Win)

HS = Home Team Shots

AS = Away Team Shots

HST = Home Team Shots on Target

AST = Away Team Shots on Target

HF = Home Team Fouls Committed

AF = Away Team Fouls Committed

HY = Home Team Yellow Cards

AY = Away Team Yellow Cards

HR = Home Team Red Cards

AR = Away Team Red Cards

And also betting odds of different providers

B365H = Bet365 home win odds

BWH = Bet&Win home win odds

IWH = Interwetten home win odds

PSH and PH = Pinnacle home win odds

VCH = VC Bet home win odds

WHH = William Hill home win odds

AvgH = Average home win odds

AvgD = Average draw odds

AvgA = Average away win odds

Closing odds (last odds before match starts)

As above but with an additional "C" character following the bookmaker abbreviation/Max/Avg

Graphical Presentation and Summary of analysis

Summary analysis shows that full time home goal for the season to be 1.421 whereas that of away team being 1.082, which signifies home team being more successful in the season. Also each match averages 2.5 goals per match which reveals the competitive nature of the tournament. It also reveals that average half time goals to be just above 1 , which signifies that majority of the goals were scored in the first half in the season.

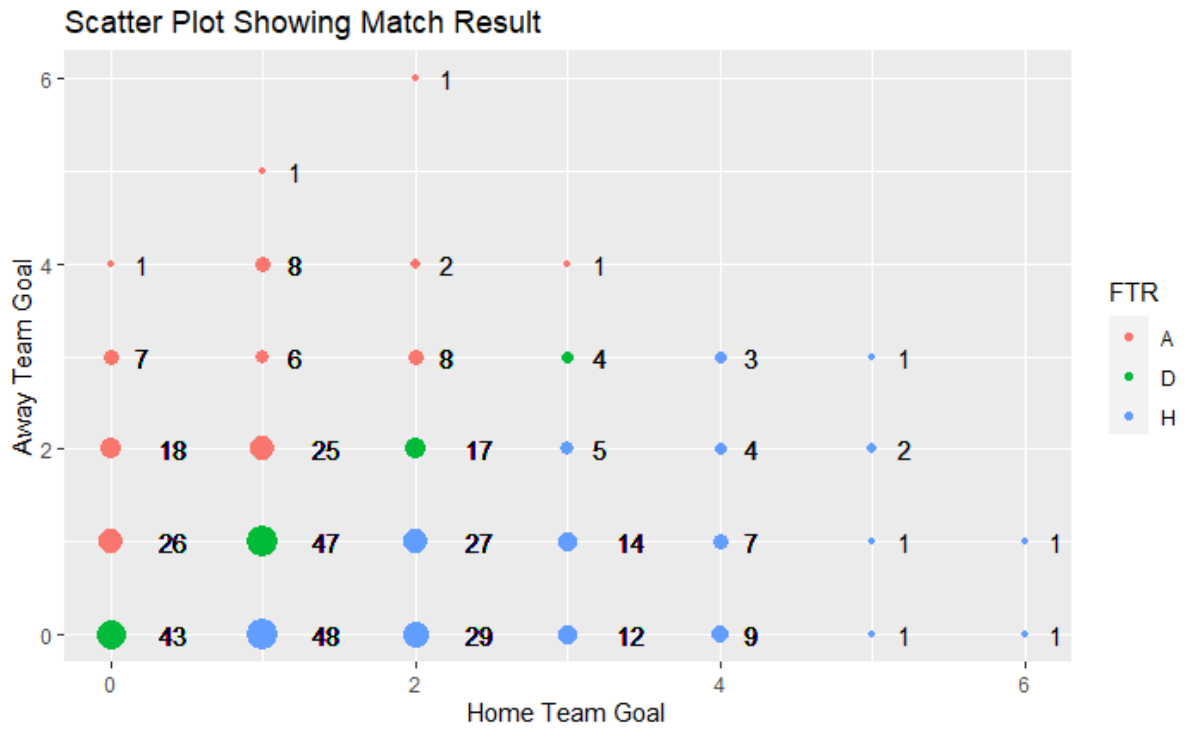

Fig 1- Scatter Plot Showing Match Result

First plot is a scatter plot showing the match results throughout the season, The scatter plot have home team goal on the x axis and away team goal on the y axis , size of the point

corresponds to no. of time that result occurs and the color of the point corresponds to the result of the match.

The plot reveals that most of the match result were concentrated at low scoring games (<3 goals) with 1-0 being the most common score line , followed by 2-0 win for home. We also see that highest goal scored in a match was 2-6 win for away team. Also highest no. of goals scored by a team in a single match being 6, which occurred 3 times. Also , we can clearly see that more matches have been won by the home team.

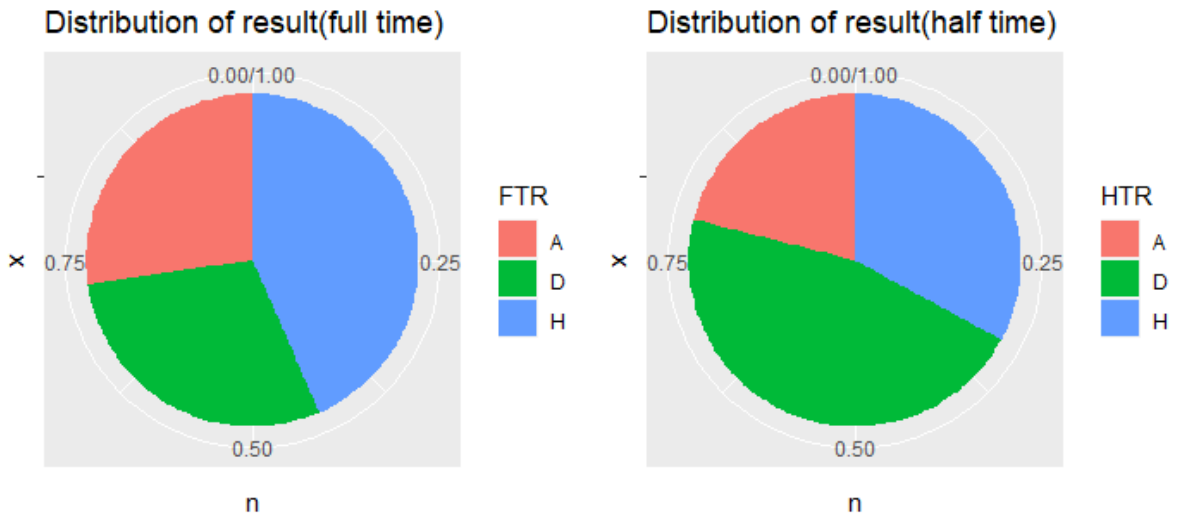

Fig 2- Pie Chart Showing distribution of result at half time and full time

Second Plot, which is a pie chart showing the result at half time and full time. We can see as concluded from the previous plot that most matches have been won by the won team which is just below 50%, followed by drawn matches and then the matches in which away team wins. At half time most of the game remained in draw , but quite evidently home team taking the win in the second half . The graph reveals a good amount of variation between half time and full time results showing the unpredictability of the game.

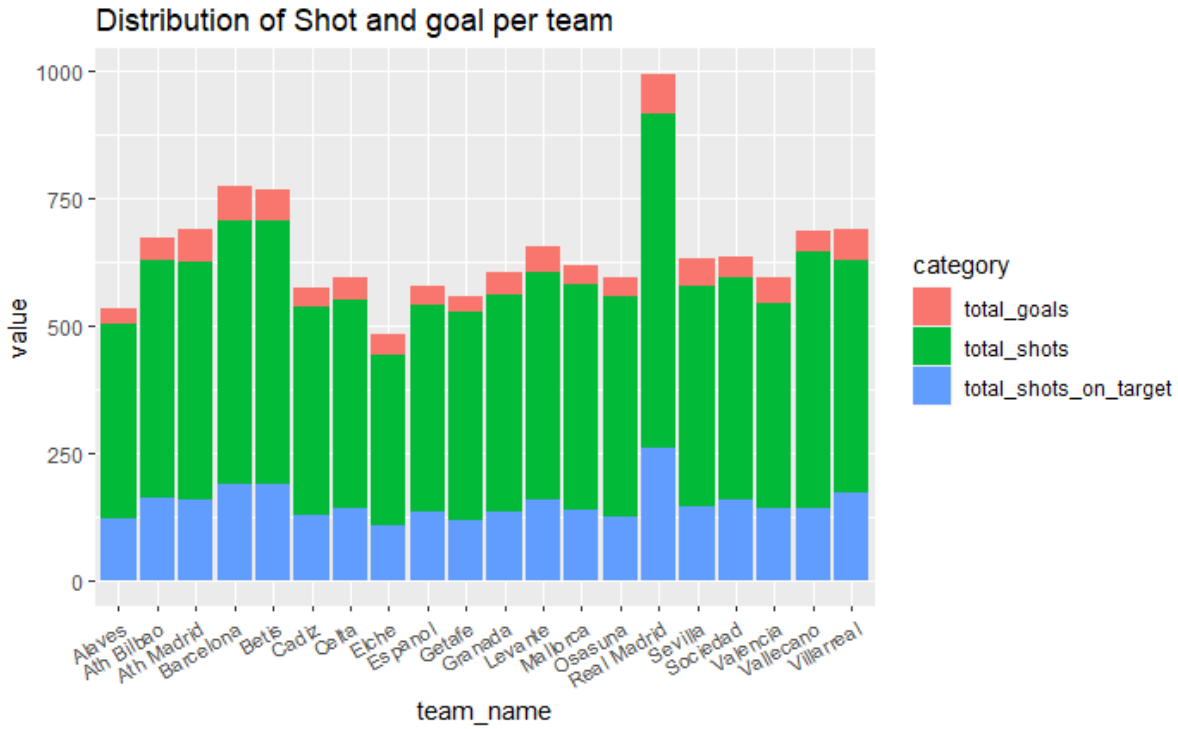

Fig 3.- Subdivided Bar plot showing no. of shot taken and goal scored by each team in the league

Third Plot gives us an indepth analysis of no. of shot taken , shot on target and goal scored by each team in the competition . It is quite evident that the maximum no. of shot taken , as well as shot on target and goal scored was by the eventual league winner Real Madrid, which is not surprising. Followed by Barcelona who finished second. Lowest no. of shots were taken by Elche who finished 13th , whereas lowest goals were scored by Alaves who finished last.

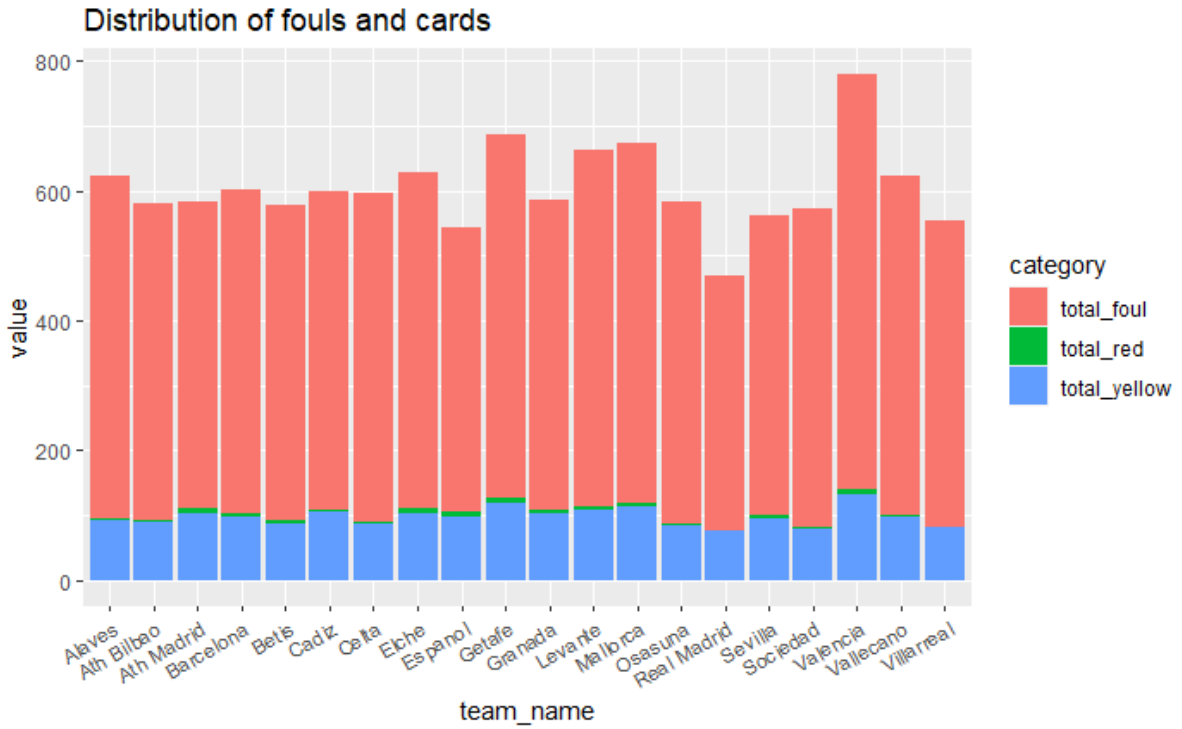

Fig 4.- Subdivided Bar plot showing no. of disciplinary statistics of each team in the league

Fourth Plot gives us an indepth analysis of disciplinary statistics of each team throughout the season which consists of no. of fouls committed , yellow card received and red card received. It is quite evident from the plot that highest no. of fouls commited along with highest yellow and red card was received by Valencia who finished 9th. On contrary the most disciplined team of the tournament being the champion Real Madrid. This plot also shows that high correlation between fouls commited and no. of cards received.

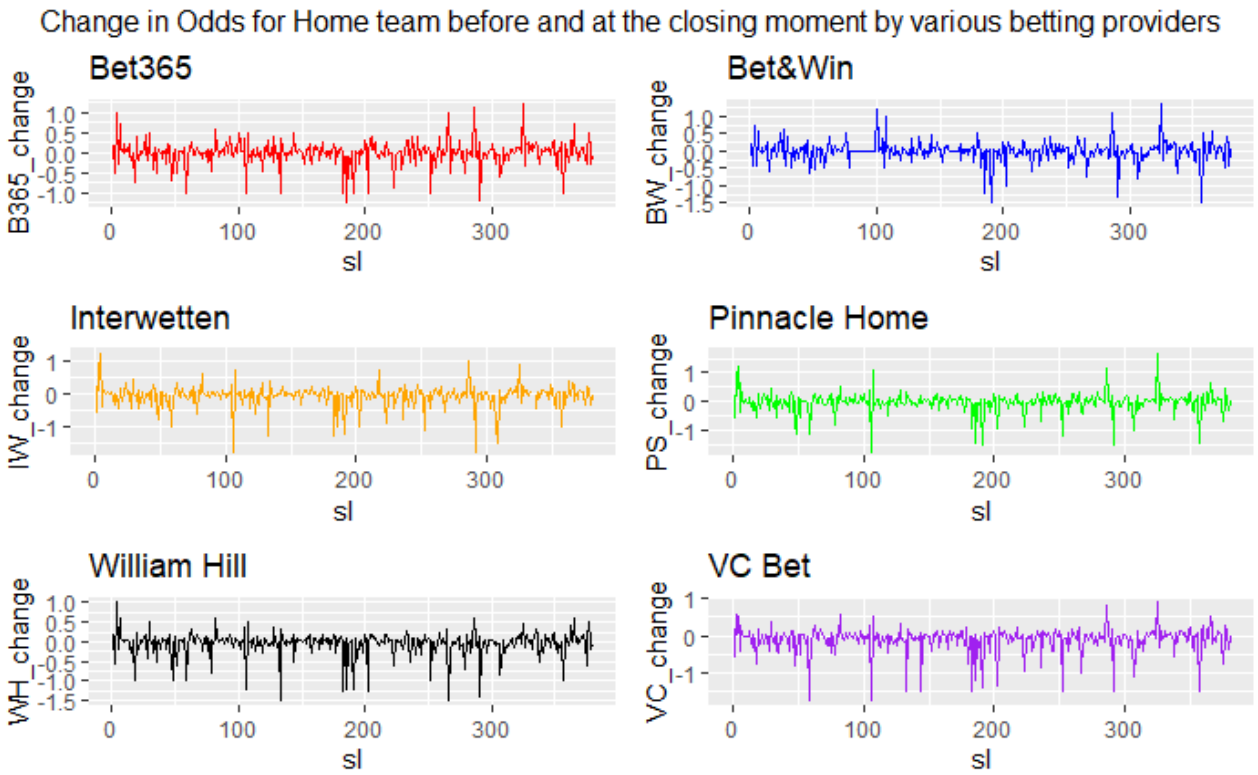

Fig 5.- Plot showing change in odds for home win before the match starts and at the closing moment by various betting providers

Our fifth plot shows the change in odds for home win before the match starts and at the closing moment by various betting providers throughout the course of the season. Here a negative change in odds refers to the match being more in favour of the home team than predicted , whereas increase in odds means the match away from the home team. Odds refer to the amount of money return to expect for $1 of bet. From the above plot we see that VC Bet have a tendency of underestimating the home team as compared to the other. Whereas Pinnacle Home being the most stable odds provider . There is high variation in William Hill and Bet365 as well. This clearly reveals the unpredictable nature of the matches.

Fig6.- Change in Average Odds for Home , Away team and draw by various betting providers

Sixth plot shows the change in odds for the win of home away team and draw taking average odds provided by all the betting providers. We can clearly see that the change in odds of home team has been somewhat more stable than that of draw and away team . There has been a huge underestimation of away team in one of the game by the providers which shows a huge drop in change of odds.

Conclusion

From the above analysis it is clear that the chance of home team winning is higher hence betting providers tend to give a higher probability (1/odds) to the home team. Also we see that the competition is highly competitive with change in half and full time result being very frequent. For season 2021-22 it has been seen the team with most shot and best disciplinary record throughout the tournament has won the league.

It has been shown that there are much fluctuation in the odds provided by the betting provider at the start of the match to what a actual match may be. It is highly risky to bet given the scenario and betters have to have their own knowledge. Finally , we see that sometimes the betting odds totally gets very eratical especially for away team win or draw , which provides a higher opportunity of upset and hence a greater return.