Demand of Football in European Leagues

DATA 450 Capstone

Matthew Wilcox

2/9/23

1 Introduction

Football is the most popular sport in the world. And no, I am not talking about American Football; Im referring to the sport in America called Soccer. Something that is incredibly important about Football is its economics. There are two economic sides of Football, the market of buying and selling players and selling the sport to the fans. This report will follow the later economic side of Football. More specifically, the demand for Football in reference to attendance.

Attendance is an important KPI for these football teams. First, they show sentiment among the supporters of the club. If attendance is consistently lower for a club, it usually indicates dissatisfaction from the supporters of the team's performance. Second, it is a key metric used for football clubs to sell sponsorships. For instance, if a company wants to sell advertising space in the stadium based on its attendance metrics, it may be able to charge more or less for these sponsorships and advertisements. The third reason it is essential is its revenue stream. For the more prominent leagues, such as the English Premier League, the revenue that comes from attendance pales to tv and other revenue streams. However, for many smaller clubs, the revenue from attendance is crucial for the success and longevity of the club.

This report aims to evaluate multiple different attributes that potentially impact attendance for a given match. The three primary attributes that will be assessed to determine their effect on attendance will be the time/day of the match, position in the table(i.e., are they winning consistently, are they fighting for promotion or relegation), and the sports odds for an individual game. Ultimately, the goal is to have a model that will predict attendance for any given match based on these attributes.

2 Dataset

Two data sets were used. The first was from Football-data.co.uk(see "Football-Data.co.uk" 2023), which contained the betting data for individual games. The second was scraped from Worldfootball.net(see "Worldfootball.net" 2023) and involved the individual match data. These will later be combined. The data is from the years 2010 to 2018. Right now, as I am still scraping and combining the data sets, the dimensions of the data are unavailable. However, the variables that will be used are stated below:

- Home team: the home team for an individual match
- Away team: the away team for an individual match
- Stadium: the stadium where the match took place
- Attendance: Attendance will be measured as a measurement of capacity. It will be a percentage. (i.e., 93% attendance)
- League: What league was the match took place in
- Betting odds: This is grouped. Currently, the data set has betting odds from multiple sources for each game. Once data has been cleaned and prepped, a better understanding of which betting odds to use will be more clear. The smaller the number for all the stats, the greater chance for that team to be the winner. For example, Home 147, Tie 250, Away 340. The home team has the highest chance of winning and then is more likely to be a tie than an away team win. For all of them, it is broken down into three different variables
  - Home team odds
  - Tie odds
  - Away team odds
- Time: This is the time that the match started
- Date: The date the match took place
- Day of the week: The day of the week a match took place. (ie. Monday or Saturday)
- Odds Goals under 2.5: the betting odds for total goals scored in a match to be under 2.5. The lower the number better the odds
- Odds goals over 2.5: the betting odds for total goals scored in a match to be over 2.5. The lower the number, the better the odds.
- Position: The numerical position in the standings a team is in.
- Secured Promotion: A true/false variable of whether a team is secured promotion from their league

- Secured Relegation: A true/false variable of whether a team is secured relegation from their league

3 Data Acquisition and Processing

The data acquisition varied. The betting data was already tabulated in CSV files and were just downloaded from Football-data.co.uk. The remaining data was scraped from Worldfootball.net. Using the beautifulsoup package, the data for match results, attendance, and stadium size were acquired.

The data processing is large. First, all the betting data is divided into individual files for each year and league. These CSVs will have to be combined together into one singular data set. Secondly, once the data set is combined into one singular data set, it will have to be merged with the match dataset. The team's data will have to be matched together to create one large dataset. Once there is a singular data set, some variables will have to be created using the data. These variables are the home and away team's current position and the related cumulative data for that season to determine the relative position.

4 Research Questions and Methodology

Time:

1. How does the day of the week affect attendance for a match? This will be viewed in a bar graph output, with attendance being the vertical axis and the average attendance for each week being the individual bars.
2. How does the start time of a match affect the attendance of that match? This will be viewed in either a histogram or scatter plot of the data, with the x-axis being the time of the game and the y-axis being the attendance level. Most likely will be a histogram.
3. How do attendance for matches change over an entire year? Similarly to the previous question, this will be viewed as a heat map of the entire calendar year, including all of the dates however.
4. How does attendance change over the season, averaging per league? This one compares the different leagues we are studying over time. This will be done with a time series line graph to see if all the leagues follow the same trends of average attendance over the entire season or if some leagues have different patterns.

Position:

6. How does the position in the table affect the attendance of a match? This one will be a bar graph where each bar is the position within the league, showing the average attendance for each position. With this visualization analysis will also be done comparing teams in relegation to the rest of the league. For this the teams within relegation and promotion to be in different colors in comparison to the rest of the leagues.

7. How does the position in the table affect the attendance of the match over a season? This will be a time series graph of the entire season. Here the teams will be grouped into thirds to show teams that are on the top and bottom of the table and how that changes over the course of a season.

8. How does securing promotion or the league and securing relegation affect attendance compared to before securing their position? This will be a grouped bar graph with one group being promotion and the other being relegation. They will have two bars, one being the average attendance before securing, and the second would be the average attendance after securing their position.

9. How does =betting odds relate to the individual teams positions? This will be a scatterplot with the difference between position and the difference between bettign odds for a team.

Betting:

10. How does attendance change when there is a heavy favorite vs. a closely betting team? This will be viewed in a histogram with the attendance being measured by the betting odds for a game.

11. How do the betting odds realte to the actural result of a match? The focus of this question is to compare to see how often the betting odds successfully predicted the outcome of a game. This will be viewed with a scatterplot where the x axis will consist of the hometeams odds of winning and the y axis be the away team odds of winning. Each individual data point will be colored if the match predicted whom would win correctly or not. Will most likely use just 2 team moneylines for this information as it will be more difficult to visualize the 3 dimension space that would occur when introducing ties.

12. How do the predicted goals being greater than 2.5 vs. lower than 2.5 have an on the attendance of a game? This will be two bars. One will be for games where the odds being over 2.5 is greater, and the other is for games where lower than 2.5 is higher. These bars will be the average attendance for these two games. Also may go and break this up by leagues to see if the different cultures and fanbases may be more accepting of lower vs. higher-scoring games.

13. How do the odds of the home team winning vs. tieing vs. losing impact attendance? This will be three scatter plots, all with the same axis scaling. The first will be the home team's odds of winning compared to attendance. The second will be tieing, and the third

will be losing. The goal is to see if a team is expected to lose a game if its attendance is lower vs. a team that is supposed to win the attendance is higher.

Final Model:

14. How can all these factors be used to predict the attendance outcome of these games? Here I will be making models with two different outcomes one being the exact attendance figure and the other focusing on percentage of the stadium capicity filled. Three different modeling methodolgoy will be used. The first attempted will be a linear model, the second will be a polnomial regressiojn mode and the third will be a decision tree regression model. The variables looking to be used are the league, the home team, the away team, the postions(difference between the postion of the home team and away team as well as the position of just the home team), the betting odds for that individual game, odds for number of goals scored, and the time/date of the match.

5 Work plan

Week 4 (2/6 - 2/12):

- Data cleaning and prepreation (5 hours)
- Question 1 (1 hr)
- Question 2 (1 hr)

Week 5 (2/13 - 2/19):

- Question 3 (1 hr)
- Question 4 (1 hr)
- Question 5 (2 hrs)
- Questjion 6 (2 hrs)

Week 6 (2/20 - 2/26):

- Question 7 (2.5 hrs)
- Question 8 (2 hrs)
- Question 9 (2.5 hrs)

Week 7 (2/27 - 3/5):

- Presentation prep and practice (3 hrs)
- Question 10 (1 hr)

- Question 11 (1.5 hrs)
- Question 12 (1.5 hrs)

Week 8 (3/6 - 3/12): *Presentations in class on Thurs 3/9.*

- Question 13 (1.5 hrs)
- Question 14 (4 hrs)
- Presentation peer review (1.5 hrs)

Week 9 (3/20 - 3/26):

- Question 14 (2 hrs)
- Poster prep (4 hrs)

Week 10 (3/27 - 4/2): *Poster Draft 1 due Monday 3/27. Peer feedback due Thursday 3/30.*

- Peer feedback (2 hrs)
- Poster revisions (2 hrs)
- Revision of graphs (3 hrs)

Week 11 (4/3 - 4/9): *Poster Draft 2 due Monday 4/3. Final Poster due Saturday 4/8.*

- Exploratory map/scatter of all games for blog post (4 hrs)
- Poster revisions (1 hr)

Week 12 (4/10 - 4/16):

- Draft Blog post (5 hrs)

Week 13 (4/17 - 4/23): [All project work should be done by the end of this week. The remaining time will be used for writing up and presenting your results.]

- Draft blog post (5 hrs)

Week 14 (4/24 - 4/30): *Blog post draft 1 due Monday 4/24. Peer feedback due Thursday 4/27. Blog post draft 2 due Sunday 4/30.*

- Peer feedback (2.5 hrs)
- Blog post revisions (2 hrs)
- [Do not schedule any other tasks for this week.]

Week 15 (5/1 - 5/7): *Final blog post due Tuesday 5/2.*

- Final presentation prep and practice.
- [Do not schedule any other tasks for this week.]

Final Exam Week (5/8): *Final Presentations during final exam slot, Monday May 9th 3:20-6:40pm.* [Do not schedule any other tasks for this week.]

References

"Football-Data.co.uk." 2023. 2023. https://www.football-data.co.uk/data.php.
"Worldfootball.net." 2023. 2023. https://www.worldfootball.net/.