Proposal: [Your project name]

DATA 450 Capstone

[Your Name]

2/2/23

1 Introduction

Football is the most popular sport in the world. And no I am not talking about American football; Im refering to the sport that in America is refered to as soccer. Something that is incredibly important with football is the economics of it. There is two economical sides of football, first being the market of buying and selling players and buying and selling the sport itself to the fans. This report will follow the later econimical side of football. More specifically the demand for football in reference to attendance.

Attendance is a importnat KPI for these football teams. First they show setiment within the supporters of the club. If attendance is consistently lower for a club it usually shows disatifaction from the supporters of the teams performance. Second it is a key metric used for football clubs to sell sponsorships. For instance a company wants to sell advertising space in the stadium, based upon their attendance metrics, they may be able to charge more or less for these sponsowrships and advertisments. The third reason it is important is its a revenue stream. For the larger leagues such as the English Premier league, the revenue that comes from attendance pales in comparison to tv revenue and other revenue streams. However for many smaller clubs the revenue from attendance is key and critical for the success and longevity of the club.

The goals of this report is to evaluate multiple different attributes that potentially impact attendance for a given match. The three primary attributes that wil be evaluated to determine their affect on attendance will be time/day of match, postion in the table(ie are the winning conistantly, are the fighting for promotion or relegation), and the sports odds for a individual game. At the end the goal is to have a model that will predict attendance for any given match based upon these attributes.

2 Dataset

[In this section, desribe the dataset(s). This includes things like where you obtained the dataset. Include a full citation, as specified here. Describe how the data was obtained by the data owner/curator, as best as you can. List the variables that you plan to use in your analysis, for example:

- weight: The patient's weight (kg)
- sex: The patient's sex, male or femalet
- age: The patient's age (months)

]

3 Data Acquisition and Processing

The data acquistion 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 aquired.

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

4 Research Questions and Methodology

[In this section, list each of the questions you will explore. Following each question, provide a detailed and specific plan for how you plan to answer the question. Include the specific steps you will take, what form the answer will take (a number? table? visualization? model? Give all the specifics), and estimate how many hours each question will take to complete.]

1. Is smoking correlated with diabetes? To answer this, I will create a filled bar plot, with the left bar representing non-smokers, the middle bar representing people who smoke moderately, and the right bar representing heavy smokers. The bars will be the same height, and each bar will be colored two colors based on the proportion of patients in the group who do or do not have diabetes.

2. Question 2? Plan for question 2.
3. Question 3? Plan for question 3.
4. etc.

5 Work plan

[Fill in the list below with a plan for what you will do each week. You should have around 7 hours worth of work each week. Writing work counts. Several tasks have already been filled in for you.]

Week 4 (2/6 - 2/12): [Just an example:

- Data tidying and recoding (4 hours)
- Question 2 (4 hours).]

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

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

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

- Presentation prep and practice (4 hours)

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

- Presentation peer review (1.5 hours)

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

- Poster prep (4 hours)

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

- Peer feedback (2.5 hours)
- Poster revisions (2 hours)

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

- Poster revisions (1 hour).

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

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 hours).

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 hours)
- Blog post revisions (2 hours)
- [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.]