Project Proposal

October 30, 2020

Home field advantage in the NFL - is it the crowd or the stadium?

by Chau Nguyen

Project Github repository: https://github.com/chav-ngvyen/data_science_1_project

I. Statement

In this project, I want to study data from the National Football League - and learn how much home field advantage matter to the outcome in football.

Before starting this project, my prior belief is that what is called "home field advantage" is actually a combination of how much a team has to travel, how much rest they have inbetween games, and crowd energy. The 2020 NFL season so far has presented an interesting natural experiment in this sense: teams still have to travel, but many stadiums are not packed. How does this affect a team's performance?

My plan is build my own dataset from webscraping football analytics websites, then build a simple model on data for seasons prior to 2020, and see how the model does with games in the 2020 season.

II. Data sources

I intend to build a dataset from scratch. My preliminary model is

$$Point\,spread = \beta_0 + \beta_1 Home\,field + \beta_2 Fans\,Attendance$$
$$+ \beta_3 Distance\,traveled + \beta_4 Time\,since\,last\,game + \beta_5 Opponent\,quality$$

I expect that most of the data needed can be found on Wikipedia and https://www.pro-football-reference.com/ - I will go into details on my plans on how to obtain each data variable in the next section.

III. Plan to obtain data

I will obtain all my data from scraping tables from the above websites. More specificially:

a) *Point spread* and *Home field* can be scraped from https://www.pro-football-reference.com/years/2018/games.htm (example for 2018).

b) *Fans attendance* can be scraped from https://www.pro-football-reference.com/years/2018/attendance.htm.

(Pro-football-reference only has attendance data from 1992 onwards, so 1992 will be the first year in my dataset.)

c) *Time since last game* can be calculated for each team with the table obtained from part a.

d) *Distance traveled* will be trickier to put together. This is because 1) Some teams will have moved cities & stadiums since the dataset starts in 1992, so I can't use a static dataset of distance between each current city/ stadium; 2) I will need to account for games played in London & Mexico City.

- The data for a stadium's city will be scraped from this Wikipedia table.
- I will use links for each stadium in the table to scrape the stadium's coordinates, then use the Haversine formula to calculate the distance between 2 stadiums (instead of airports).

e) Opponent quality will likely be a set of variables - for example, the overall ranking of the opponent's defense https://www.pro-football-reference.com/years/1992/opp.htm.

IV. Methods I aim to employ

- Web scraping: I need to write an efficient scraper for the data listed above. I will likely write 2 scrapers for each variable - one for data from seasons 1992 - 2019, which will only need to be run once, and one for the 2020 season data, which I can re-run as least once a week for new updates.
- Data wrangling: Because I'd like to be able to feed new 2020 data into the model This also means that I'll need to pay good attention into the cleaning & merging of the data so that I can add new weekly data with ease.
- Data visualization: Since each game is played by 2 teams, I expect the final dataset will be a dyadic dataset. I will need to think more about how I can visualize sports/ dyadic data.
- Machine learning component: I want to first build a model based on data from the 1992 - 2019 seasons, then try to apply it to the 2020 season.

V. My definition of success

The challenges I set for myself with this project are to be able to:

- Ask interesting questions to study a completely new domain to me - sports, and be able to conceptualize what kind of data I would need to answer these questions.
- Scrape data from the web myself, merge scraped data together to build my own dataset.
- Build at least one simple machine learning model on data in pre-covid seasons and see how it does in the 2020 season.
- I will not base my success on being able to get statistical significant results my first try. Results would be a nice thing to have, but it's not something I will be overly concerned with at this stage.