Turner Schwiebert

Seminar in Data Analytics

Dr. Sarah Supp

September 30, 2022

Skill Outline

Programming Tools:

1. Web-scraping FBRef (Fbref, 2022; Terner, 2021) in Python using
    a. “selenium” package to automate the scrolling process on league websites.
    b. “requests” package to execute get requests and receive website elements.
    c. “lxml” package to parse response.
    d. “pandas” package to create the final data set.
2. Data Exploration (Clustering) in R using
    a. “caret” package to experiment with different clustering algorithms
3. Statistical Approaches
    a. K-means clustering, used in previous studies with similar goals (Spagnolo, 2007; Shelly, 2020; Gyarmati, 2014)
    b. Hierarchical clustering (Gyarmati, 2014; Duman, 2021)
4. Cluster Validation
    a. I will use my personal domain knowledge as well as the prevalence of “traditional positions” to decide whether or not the finalized clusters provide any valuable insights.

Works Cited:

Anıl Duman, E., Sennaroğlu, B., & Tuzkaya, G. (2021). A cluster analysis of basketball players for each of the five traditionally defined positions. *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology*, 17543371211062064. https://doi.org/10.1177/17543371211062064

This study used hierarchical clustering to better understand the different basketball players in the traditional 5 basketball positions. This is relevant to my project because it is doing something very similar but for basketball. Thus, I am more confident in potentially using hierarchical clustering in my project.

FBRef. (2022). Retrieved From https://fbref.com/en/

This is the website that will web scrape for performance data. It is an open source platform for soccer data, and it is part of the Sports Reference family of websites. For an example of the web page I will scrape data from, see the following link: https://fbref.com/en/players/d70ce98e/scout/365_euro/Lionel-Messi-Scouting-Report

There is metadata under the player's name and the "per 90" values under the "per 90" column in the data table.

Gyarmati, L., Kwak, H., & Rodriguez, P. (2014). Searching for a unique style in soccer. *arXiv preprint arXiv:1409.0308*. https://doi.org/10.48550/arXiv.1409.0308

This study uses PCA, hierarchical clustering and k-means clustering to analyze the different playing styles of professional Spanish soccer teams. This study makes me confident in the use of both k-means and hierarchical clustering in this context.

Shelly, Z., Burch, R. F., Tian, W., Strawderman, L., Piroli, A., & Bichey, C. (2020). Using K-means clustering to create training groups for elite American football student-athletes based on game demands. *International Journal of Kinesiology and Sports Science*, *8*(2), 47-63.
This study used k-means clustering to cluster similar American football student athletes into training groups using Catapult data. The researchers chose k-means because it captures the most detail out of the data set compared to other methods, so it makes me confident that k-means is the right method for my project.

Spagnolo, P., Mosca, N., Nitti, M., & Distante, A. (2007, September). An unsupervised approach for segmentation and clustering of soccer players. In *International Machine Vision and Image Processing Conference (IMVIP 2007)* (pp. 133-142). IEEE.
This study is very similar to the study that I plan on doing this semester. In fact, I was initially concerned that this filled the knowledge gap that I plan to fill. However, the researchers clustered players based on video game data, not real performance data. Regardless, their exploration of clustering methods like k-means in this context gives me confidence that real performance data will be able to group players into positional roles.

Terner, Z., & Franks, A. (2021). Modeling player and team performance in basketball. *Annual Review of Statistics and Its Application*, *8*, 1-23.

This study uses basketball data from Sports Reference, the parent company of the website I will be web scraping for my data. Seeing a published study use this open-source data gives me confidence that the data is correct, up-to-date, and usable for this formal context.