Turner Schwiebert

Seminar in Data Analytics

Dr. Sarah Supp

October 14, 2022

Research Design

As previously stated, this analytical research project will identify the number of modern positional roles in professional soccer (and which players fit which roles) This form of player profiling of professional soccer players requires relevant data as well as advanced analytical methods. This is evident in previous research and projects with similar goals. In fact, there have been many studies conducted in similar sports contexts that have used k-means clustering to profile team playing styles, basketball players, and college-level football players respectively (Spagnolo, 2007; Shelly, 2020; Gyarmati, 2014). Furthermore, k-means clustering preserves the variance in the raw data when assigning observations to clusters, making it a very good choice for answering this project's research question (Shelly, 2020). Thus, the use of k-means clustering is common when profiling players in this context, so it is a good choice for this project as well. Another clustering method that will be used for raw data exploration is hierarchical clustering. Hierarchical clustering is a good method for determining how many clusters are present at specific levels within data sets, and this method has been used in previously published research projects as well (Gyarmati, 2014; Duman, 2021). Both k-means clustering and hierarchical clustering are reproducible clustering methods available through the "caret" package in R along with data manipulation and visualization conducted with "dplyr" and "ggplot2" packages, respectively (R Core Team, 2022; Kuhn, 2022, Wickham, 2022; Wickham, 2016). It is important to note that k-means clustering uses randomly generated initial centers for clustering, so results

will vary based on the initial center used. However, initial centers can be reproduced by using the same "seed" in R, so there are still no reproducibility issues associated with this method.

Before running any sort of clustering analysis, the first requirement for answering this research question on player profiling is acquiring data on the players themselves. This was done through FBRef, a free, open-source platform for professional soccer data. FBRef is part of the Sports Reference family of sports data websites. They are widely renowned for providing availability of this form of data. In fact, previous published research projects have also used data from Sports Reference websites (Terner, 2021). In order to collect this data, the website has been web-scraped for "per 90" player performance data covering the past 365 days. It is important to note that "Per 90" refers to a professional player's production per 90 in-game minutes. Using this unit makes the most sense because a soccer game lasts 90 minutes. Nevertheless, the web-scraping process started by collecting a list of player names as well as unique FBRef player IDs. These lists were then used to create URLs for each professional player's "Complete Scouting Report" for all players in Europe's top 5 leagues: Premier League, Bundesliga, La Liga, Serie A, and Ligue 1. The variables collected include traditional position, age, height, weight, footedness, as well as the aforementioned "per 90" performance data covering shooting, passing/pass types, goal/shot creation, defense, possession, and miscellaneous aspects of soccer. Thus, all variables other than position and footedness are numerical and continuous, while position and footedness are categorical. Finally, this web-scraping process was conducted in python using the packages "os" for directory manipulation, "selenium" for opening the league web pages automatically, "io", "request", "lxml" for requesting, receiving, and parsing raw responses, "pandas" for post-collection data manipulation, and "tqdm" for tracking web-scraping progress at run time (Van Rossum, 2009; os; Muthukadan, 2018; io; Chandra, 2015; Behnel,

2005; McKinney, 2010). It is also important to note that there may be exclusion criteria for players based on how many minutes they have played in the past 365 days.

The validity of analytical results is paramount in any research project. However, a project consisting only of advanced "black box" clustering methods requires complex model validation solutions. This problem will be solved in various ways. First, I have already conducted domain research that includes past success in using these methods in this domain. This makes me confident that results, once tuned, will be correct and insightful. Second, extensive exploratory analysis using hierarchical clustering will be used to ensure that the correct "k" has been chosen for the k-means clustering algorithm. Third, the composition of clusters will be vetted in detail using the prevalence of traditional positions, ensuring that results make sense given personal as well as domain knowledge of famous players and their roles in the past 365 days. Finally, the clustering analysis will follow a "two-pronged" approach with the first phase pooling all players in one data set and running it through a clustering algorithm. The second phase will be conducted using the same data separated into data frames by traditional player position and then run through k-means clustering. This second phase will reveal insights into player roles within each position. Regardless of results, actionable and interesting insights will be generated at each phase.

Works/Tools Cited:

Behnel, S., Faassen, M., & Bicking, I. (2005). *lxml: XML and HTML with Python*. Lxml.

Chandra, R. V., & Varanasi, B. S. (2015). *Python requests essentials*. Packt Publishing Ltd.

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

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

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

Kuhn, Max (2022). caret: Classification and Regression Training. R package version 6.0-91. https://CRAN.R-project.org/package=caret

McKinney, W., & others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56).

Muthukadan, B. (2018) *Selenium with Python*. https://selenium-python.readthedocs.io

os - *Miscellaneous operating system interfaces*. https://docs.python.org/3/library/os

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

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.

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.

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

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Wickham, H., François, R., Henry, L. and Müller, K. (2022). dplyr: A Grammar of Data Manipulation. R package version 1.0.8. https://CRAN.R-project.org/package=dplyr

Wickham, H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

https://cran.r-project.org/web/packages/ggplot2