Turner Schwiebert

Seminar in Data Analytics

Dr. Sarah Supp

September 16, 2022

Project Proposal: Analyzing Roles in Modern Professional Soccer

Soccer analytics is a very crowded space. Much research has already been done on the financial, physical, and performance aspects of the sport (Bush, 2015; Di Salvo, 2007). Professional teams also conduct their own analyses in an attempt to improve their team's performance. In fact, teams in Europe's top leagues constantly improve both physically and tactically year on year (Bradley, 2016). Part of this tactical improvement is due to soccer's top coaches constantly coming up with innovative ways to use players. For example, one revelation at the professional level in the past few seasons has been the use of full backs. Some teams employ their fullbacks as box-to-box defenders who attack and defend on the wing. Other teams "invert" their full backs, and these players who technically play a wide position end up occupying space in the middle. Thus, there is much variability even within specific positions, and full backs are just one example of this (Konefal, 2015). More and more players are being asked to do things for their team that do not fit their expected positional requirements (Konefal, 2019). My aim for my research project is to better understand these complex requirements, a current gap in soccer research. I plan on analyzing data measuring player performance, physical attributes, and potentially spatial data in some capacity. This analysis will result in clusters of players who perform similarly, or, in other words, fill the same positional role. These clusters will form the basis of insights into the current landscape of positional roles in modern soccer beyond conventional positions.

To accomplish this, I will first redevelop my FBRef web scraping infrastructure that I created in an independent project a few years ago. It no longer works, so I plan on completely redesigning it to fit my needs for this project. FBRef is an open-source platform with a huge amount of publicly available soccer data (FBRef). I will be collecting data on all players in Europe's top five leagues: the English Premier League, German Bundesliga, Italian Serie A, French Ligue 1, and Spain's La Liga. More specifically, I will be collecting data from the players' "Complete Scouting Report" page. It contains advanced performance data that measures a players offensive and defensive output per 90 minutes (length of a game) as well as physical variables like age, height, weight, primary foot, and their dependence on their primary foot. I am also open to using spatial data or some proxy that measures what spaces players are occupying on the field on average.

The methods I will need to use to complete my project include, but are not limited to, web scraping, data wrangling/preparation, and clustering. First, as previously mentioned, I plan on web scraping FBRef. Thus, I will need to refresh myself on my FBRef web scraping process that I used in my old project. Beyond web scraping techniques, I will also need to become familiar with the FBRef's website and how it denotes unique players. This will be integral in creating the necessary URLs for each player, since each scout report is on a different web page. Next, I feel confident that I will be able to prepare the web scraped data once collected. I have extensive experience with wrangling and cleaning data through the DA curriculum and my internship, so I am not worried about catching up at that stage of my project. Lastly, I will likely need to learn more about advanced, unsupervised clustering methods to capture the positional roles through the performance data. I have some experience with these techniques through DA350, but I feel that I will need to broaden my knowledge and learn new methods. Regardless

of my previous experience with the techniques, I am confident that clustering will provide meaningful insights due to similar techniques that have been used in the same context (Spagnolo, 2007).

My semester-long project will be conducted in two phases: collection and analysis. First, I plan on finishing my data collection phase as quickly as possible, likely by the end of September. If needed, I will revisit this phase later in the semester if I need more data to explain the positional roles and, thus, generate insights. The analysis phase will likely take much longer and consist of choosing which clustering algorithm to use and then improving the model over time. I foresee the analysis phase finishing at the project deadline. For a more detailed explanation of my semester-long timeline, please see the table below.

Steps	Date of Completion
Domain Review	Sep 23, 2022
FBRef Data Collection	Oct 1, 2022
Choose Clustering Method	Oct 12, 2022
Finish Clustering	Nov 5, 2022

Works Cited

Bradley, P. S., Archer, D. T., Hogg, B., Schuth, G., Bush, M., Carling, C., & Barnes, C. (2016). Tier-specific evolution of match performance characteristics in the English Premier League: it's getting tougher at the top. *Journal of sports sciences*, *34*(10), 980–987. https://doi.org/10.1080/02640414.2015.1082614

In this short excerpt, researchers discuss the increasing performance levels among leagues in England. Their findings show that performance consistently improves year after year, showing that the game is constantly evolving. Part of this evolution is strategies and tactics, such as new and innovative use of players (and their positional requirements).

Bush, M., Barnes, C., Archer, D. T., Hogg, B., & Bradley, P. S. (2015). Evolution of match performance parameters for various playing positions in the English Premier League. *Human movement science*, *39*, 1–11. https://doi.org/10.1016/j.humov.2014.10.003

In this study, researchers conducted advanced research into physical exertion of players in England's top league, the English Premier League. They stratified their research by position and found some small differences in physical output. This small difference in physical output is very interesting, and it is something I need to keep in mind while researching performance differences in positions.

Di Salvo, V., Baron, R., Tschan, H., Montero, F. C., Bachl, N., & Pigozzi, F. (2007). Performance characteristics according to playing position in elite soccer. *International journal of sports medicine*, *28*(03), 222-227.

This study looks at the differences in players' physical exertion between the conventional soccer positions. This is relevant to my project because it studies different aspects of the game across the positions, which is similar to what I will be doing.

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.

Konefal, M., Chmura, P., Andrzejewski, M., Pukszta, D., & Chmura, J. (2015). Analysis of match performance of full-backs from selected European soccer leagues. *Central European Journal of Sport Sciences and Medicine*, *3*(3).

In this study, researchers analyzed how full backs are used differently across Europe's top leagues. The researchers focused on league-specific differences in how full backs are used, which is an interesting perspective that I should keep in mind while conducting my research.

Konefał, M., Chmura, P., Zając, T., Chmura, J., Kowalczuk, E. & Andrzejewski, M. (2019). A New Approach to the Analysis of Pitch-Positions in Professional Soccer. Journal of Human Kinetics, 66 (1) 143-153. https://doi.org/10.2478/hukin-2018-0067

In this study, researchers examined how players' positions affected various performance statistics. This work is relevant to my research because the authors asked the opposite of my question. They analyzed the difference in performance between concrete positions, but I am planning to research positional roles through performance data. Thus, I can learn a lot from how they went about their research.

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 in this context gives me confidence that real performance data will be able to group players into positional roles.