Football Analytics: Harnessing big data to identify players with similar technical attributes

Project Proposal

Student Name: Kelvin Lim Wan
Student Number: 929715
Subject: Research Project (COMP90055)
Period: Semester 2, 2022

Supervisor: Tilman Dingler
Institution: University of Melbourne
Industry Mentor: Daniel Pelchen
Industry Organisation: Traits Insights

Motivation

In recent years, clubs at every level of the football pyramid have become smarter and more efficient. How? Through the use of data and analytics (Anderson and Sally, 2013). Analysts are now recording data from thousands of actions during games and training sessions to help shape pre-match preparations and post-game debriefs, pinpoint transfer targets and develop young talents (Thakkar and Shah, 2021). This is revolutionary for the transfer market, mainly because the financial gap in football's top leagues is constantly increasing (Thakkar and Shah, 2021). To compete against the affluent clubs requires creative thinking from less wealthy and successful clubs. For instance, Southampton F.C. has created the *black box*: a live database collecting player metrics from every major league. This has enabled them to acquire players of undervalued talent and sell them for a profit. Virgil van Dijk, Sadio Mane and Luke Shaw are prime examples (Anderson and Sally, 2013).

Football is an extremely complex sport: it is low-scoring, continuous, time-varying and free-flowing (Thakkar and Shah, 2021). As a result, data points collected during matches are not as rich and informative as in other sports, which poses a challenge for football data analysts. To illustrate, basketball and Australian football are high-scoring, while tennis and baseball are time-segmented; this allows experts to derive accurate and informative insights to aid executives in their decision-making. Besides, there are aspects of a player such as their professionalism and football IQ that cannot be aggregated into data, and still requires watching live performances or talking to the player (Anderson and Sally, 2013).

One of the main challenges for a football team is upon the unforeseen departure of a player, when there is no evident replacement in the current squad. After all, teams are trained to play in a specific system, where each player carries out a specific role (Anderson and Sally, 2013). The most sensible solution is to find a like-to-like replacement, that is, a new player with the same attributes as the departed player. Naturally, no two players are identical, but by leveraging data and analytics, it is possible to assess players similarity based on their technical profiles. An old-school approach is to send scouts to watch players in live matches across several locations (Thakkar and Shah, 2021). This process is very time-consuming and not scalable, especially in the rapidly-growing football industry. In addition, there is a risk of unconscious bias in one's opinion of a player, due to cultural differences or a preferred style of play. With big data, clubs are able to analyse a plethora of players all over the world in just a few clicks, while also eliminating the cognitive bias from human scouts.

Further, the football transfer market has seen some remarkable transfers in recent years, with the current transfer fee record being €222 million (A$333 million) for the transfer of Neymar from FC Barcelona to Paris Saint-Germain F.C. in August 2017 (Anderson and Sally, 2013). One of the reasons for those excessive prices is that people get obsessed with one player; "they think 'this is the guy, we need to have him, and we are willing to pay over the odds'" explains Anderson (2013). What data can do is help generate options: to find players that are similar to another player, or who would fit into the team in a slightly different way. It allows clubs to walk away from a bad deal.

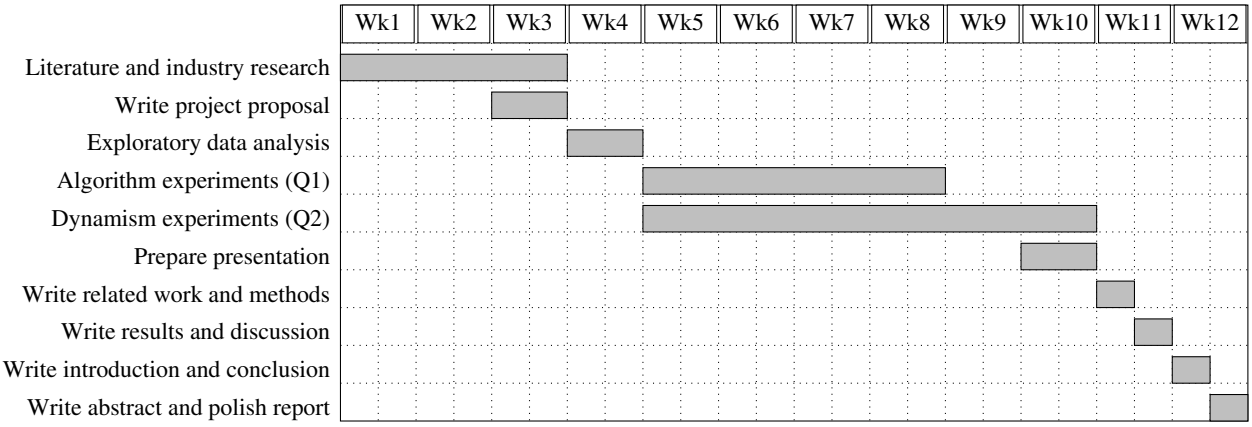
Objectives

This research project aims to leverage data analytics tools and machine learning techniques to answer the following research questions:

1. Which algorithm works best to group football players in terms of their technical attributes using a framework of composite variables?
2. How can a similarity index be automated for dynamic samples of composite variables and differing frameworks?

Proposed Methodology

The proposed plan and methods for the delivery of the research project is outlined in the Gantt chart below. The project spans across the 12 university teaching weeks between 25 July 2022 and 23 October 2022, with a presentation in week 11 (exact date to be confirmed) and the thesis submission on 31 October 2022.



For question 1, some ideas include using PCA (Principal Component Analysis) for dimensionality reduction and K-means clustering or DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to group similar players. For question 2, a potential tool is Microsoft's Power BI as it allows for data dynamicity. A more concise plan and more tailored methods will be devised upon further research and data exploration.

Implications

In answering the two research questions, we aim to provide football clubs with a tool that allows them to extract a list of players who are the next closest fit to a certain queried player according to their technical profiles. The tool would also automate such queries around any number of tiered composite variables, specific to each club, and update as the samples evolve. Football clubs would then be able to make data-driven decisions regarding transfers which would result in cutting cost on in-person scouts and a more scalable and unbiased outlook on player recruitment.

Reference List

Anderson, C. and Sally, D., 2013. The Numbers Game: Why Everything You Know About Football is Wrong. Penguin Books.

Thakkar, P. and Shah, M., 2021. An Assessment of Football Through the Lens of Data Science. Annals of Data Science, 8(4), pp.823-836.