Interdisciplinary Project in Data Science Proposal

Recommending Similar Players Using Technical Performance Indicators

Alexander Lorenz
e12239877@student.tuwien.ac.at

Main supervisor: PD Dr. Jakob Müllner, Wirtschaftsuniversität Wien
Co-supervisor: Univ.Prof. Dr. Allan Hanbury, Technische Universität Wien

Domain-specific lecture: 4694 Foundations of International Business (WU Wien)

1 Abstract

The aim of this project to is to develop a data-driven method that supports coaching and scouting staff to identify one-to-one replacements for key football players that leave the club or suffer from a long term injuries, by recommending the most similar players given a query player based on technical performance indicators, including dribbling, passing, shooting, and other dimensions. Therefore, a feature engineering pipeline is designed to derive advanced features based on game-event data for each technical performance dimension to construct player vectors, which capture the playing style characteristics of football players. Unsupervised learning algorithms and similarity measurements are employed for providing player recommendations.

2 Motivation

Modern football has become increasingly data-driven, especially in game analysis, physical health metrics, and player scouting. In the domain of player scouting, scouting staffs use platforms such as wyscout [1] to explore rich tabular data, visualizations, and short video clips that reveal information across various dimensions of a player's characteristics (e.g. dribbling, shooting, passing attributes). Certainly, to effectively compare the player and their statistics, the scouts have to appropriately account for contextual factors and biases across league levels, teams, and attribute ranges to understand the order of magnitude of each player statistic.

Eventually, scouting new players among thousands of players from scratch based on a desired set of player attributes may become a challenging task. Furthermore, instantly identifying patterns in the data that capture nuances of the player's playing style that the

club is seeking is ambitious. Subsequently, players may be left unseen or incorrectly labeled as the wrong player even if they share similar characteristics.

Simultaneously, football players that have a significantly positive impact on the team performance receive offers and eventually leave the club. At the same time, impactful players can get complicated injuries that can force them to miss matches for months.
In these cases, coaches seek players with similar characteristics that can seamlessly integrate into the team and serve as a one-to-one replacement.

2.1 Research Objective

The objective of this project is to develop a data-driven scouting method that derives advanced player statistics from raw game-event data, which capture player characteristics. This shall be used to recommend the top-k most similar players given a query player that can serve as replacement.

2.2 Research Question

To what extend can high-dimensional game-related player statistics effectively capture and identify player characteristics? Can a one-to-one replacement player be identified for a given query player, that share the same characteristics, using a feature-engineered dataset?

3 Methodology

In general, this project will follow along with the CRISP-DM process model. As data source, we will use the StatsBomb Python API [2], which provides rich game-event data for the entire season of 2015/16 of the Premier League, Ligue 1, Bundesliga, Seria A and Ligue 1.

In the next step of data preparation, we will perform feature engineering to transform the raw game-event data into advanced features that reflect the technical performance of a player [3], aiming to encapsulate the player characteristic. Moreover, adding complexity to a generic variable (e.g passing accuracy) by accounting for a given event its location on the pitch, whether it was under pressure, and whether it had a positive outcome (e.g passing accuracy under pressured to final 1/3) can provide highly informative insights on purposeful actions taken by a player [4]. The feature engineering part is expected to be the most time-demanding part of this project.

In the modeling part, we use both supervised and unsupervised learning techniques to train and evaluate to what extend the feature-engineered dataset can effectively separate players by their positions. The player position variable is highly informative and provides valuable insights into the player's role, behavior and the playing style on the pitch [4]. Finally, we experiment with suitable unsupervised learning algorithms and similarity measurements to recommend the top-k most similar players based on a given query player.

At last, a shallow web application is deployed to demonstrate the proof of work. It shall provide the ability the to select a player and display the top-k recommended players based on the selected player.

4 Expected Results

In general, a feature engineering pipeline shall be implemented, which generates a data set that captures different player characteristics. The corresponding success criteria are multi-class classifiers, trained to predict player position, that obtain sufficient predictions scores for accuracy, precision, and recall, while providing explainability and interpretability.

For the recommendation part, the metrics recall@k and precision@k are used to determine whether the recommendation list actually contains relevant player in terms of equivalent positions.

More practically speaking, when selecting a defensive midfielder as query object, who is known for aggressive pressing, ball-winning ability and an active engagement for tackles, duels and recoveries, while obtaining offensive skills in contributing the ball, dribbling and goal-scorer potential, it is expected to get in return a list of players that also share these characteristics. Hence, surprising and unexpected recommended players that actually do not have the same position in common and share the same characteristics are encouraged.

References

[1] Wyscout, https://www.hudl.com/en_gb/products/wyscout

[2] Statsbomb API, https://github.com/statsbomb/statsbombpy

[3] Yang, G., Leicht, A. S., Lago, C., & Gómez, M. Á. (2018). Key team physical and technical performance indicators indicative of team quality in the soccer Chinese super league. Research in Sports Medicine, 26(2), 158–167. https://doi.org/10.1080/15438627.2018.1431539

[4] Bloomfield, J., Polman, R., & O'Donoghue, P. (2007). Physical Demands of Different Positions in FA Premier League Soccer. Journal of sports science & medicine, 6(1), 63–70.