Identifying and Recommending Players of Similar Skill Level Through Clustering

Jordan Miranda

September 27th, 2021

BrainStation

Introduction

This report has been prepared to provide a technical summary of this capstone project that's been built over the course of seven weeks. The overall objective of this capstone project is to identify and recommend soccer players of similar skill levels through clustering. With clustering the goal is to create – based on cluster labels – unique profiles of player-types. The goal with these labels/unique profiles is to serve as a reference for what type of player any given player (chosen by the user) is in the top five European leagues.

Business Use Case

The business use case for creating a reference book that contains players of similar skill levels is that it will allow football clubs to expedite their player scouting process when beginning their initial search efforts. If the club is in search of replacing a player whom they know may leave their club soon, being provided with a list of players that are of similar skill level and/or playstyle compared to their departing player would be an incredible convenience. It would allow the club to immediately begin in-depth player analysis and recruiting. Bypassing the need to send out player-scouts to source a list of potential candidates. This would save the club money by reducing the expense of paying the cost of their scouts travelling to identify players. It would also save the club time as they would not have to wait on the logistics of sending out scouts to games across the globe and waiting for their reports. In summary, receiving an instant list containing players of similar skill level would allow clubs to begin evaluation immediately after they decide they're in need of bringing in new talent, saving the club time, money, and effort.

Background

The players in our dataset consist of players only within the top five European soccer leagues. These leagues and their respective countries are listed as follows: Premier League (England), Ligue 1 (France), Bundesliga (Germany), Serie A (Italy), and lastly, La Liga (Spain). As these leagues are considered the best leagues worldwide, they are the most recognized leagues commercially as well. This means data on these leagues is credible, complete, available from multiple sources, and unlikely to be tampered with or containing missing values.

Only in recent years has soccer seen a data analytics boom. For its incredibly lengthy history, soccer was long considered a sport too proud to consider statistics and analytics. It was a sport in which you "watched the game with your eyes, not with a stat sheet" and where a player's skill was often judged by what's called "the eye-test" – forming an opinion on a player after seeing them with your own two eyes. In recent years there's been an acceleration in the acceptance of data analytics within soccer. More and more clubs are using data to make decisions on how they operate. Much of the use of data in soccer thus far has been on optimizing players and their play on the pitch. With this project, the hope is to take a different approach on data usage in soccer. The goal with this project is to streamline and optimize a football club's operations when it comes to player recruitment.

The Data

The dataset that was used for this project was every player's stats who played in the top five European leagues from the 2010-2011 season to the 2020-2021 season. The data format was originally in tables hosted on fbref.com, with the stats provided courtesy of Statsbomb. Each season was presented as an individual table. Exporting the tables was done by navigating through a dropdown menu converting each season's respective table to a csv file.

Regarding the shape of these dataframes, they were often varying in rows of approximately 100, as some seasons teams would register more players. Most commonly the number of rows in each season's dataset was between 2600 to 2750. For the seasons of 2010 to 2016, the number of features was 23. However, beginning in 2017 the introduction of predictive metrics was included to the tables and thus from 2017-2020 the number of features increased to 32 features. Regardless of which season the dataset was from, the number of object datatype features stayed the same – 5. The remaining 18 (for seasons pre-2017) or 27 (for seasons post-2017) consisted of floating point or integer datatype features. These features consisted of metrics measuring the player's season. After being familiarized with the source of the dataset and its features, data cleaning and feature engineering began.

Data Cleaning, EDA, and Feature Engineering

Before merging all the datasets together to form a collection of the past 11 seasons there were some features that needed necessary editing and some features that were to be removed. Using a for loop, all datasets had a "Year" feature added that tracked which season the player played in. Each dataset had 3 features removed due to HTML-encodings and a singular redundant feature. Due to emojis in the tables we had to perform string splitting on 3 features to extract the plain text values. Lastly, renaming certain sets of metrics was needed as the website tables had given them a category heading to which they belonged to but was not included in the csv files, causing duplicate feature names.

Once each dataset had their identified features modified, they were all concatenated to create one dataset representing the past 11 seasons played in Europe's top five leagues. At this point, the focus shifted to dealing with missing values.

The first features that carried the largest number of missing values were the predictive metrics which was to be expected as we had accumulated seven seasons worth of missing values. Ultimately it was decided that these features would be dropped as the focus of our analysis was to focus on players as they are now, not what a third-party predicts they could be. Beyond these features, there was only a handful of missing values across a variety of features such as age or assists. These values were filled by using a trusted outside source – Transfermarkt.com. After locating which players were missing values and where, the values from Transfermarkt were used to fill them in. Lastly, we decided to remove all goalkeepers from our dataset due to the analysis's focus being on outfield players. At this point, exploration of our data began.

In our exploratory data analysis, the goal was attempting to uncover features that could end up being an issue due to outliers. As many of the features were absolute value metrics this wasn't a big problem within these features. Where it became an issue was in the "Per 90" features. Due to low minutes played, players who managed to get an assist or goal in their small amount of time on the field ended up with massively inflated "per 90" metrics due to the numbers extrapolating from the small sample size. Luckily, these players were caught out in our modelling and were removed (more on this under the Modelling section). After exploratory data analysis concluded, feature engineering became a large focus in this project.

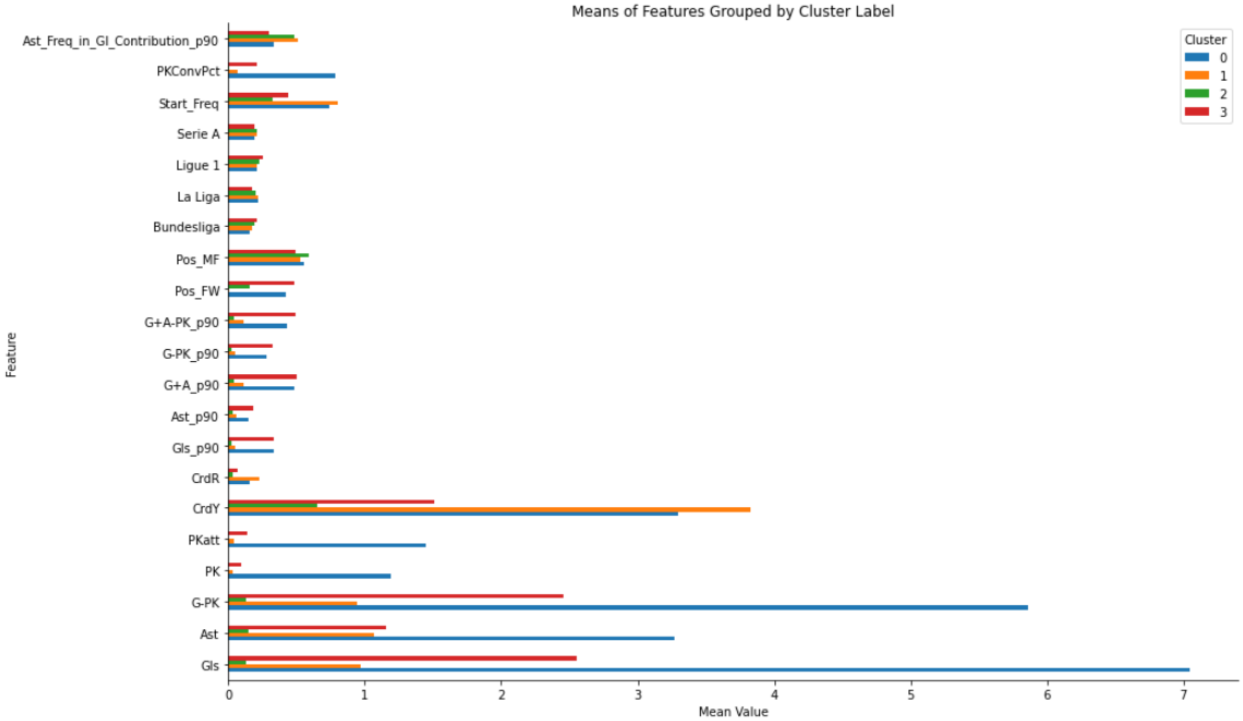
In feature engineering the goal was to create as many relative value metrics as possible. As many of the players in our dataset weren't playing as much as the best were, it was no wonder they weren't scoring as many goals. We needed to establish a way to compare both types of players equally. There were already plenty of "per 90" metrics created for us which was helpful, but we decided to create five additional features that measured players on a relative scale. One hot encoding was performed on two categorical values – the league a player played in and their position. Prior to one hot encoding the player's position we consolidated this feature down to three main values – forward, midfield, and defense. Lastly, we performed a group-by function on the player's names to get the averages of all their metrics over their careers. Overall, this feature engineering process increased our feature count to 29 features while the group-by function reduced our rows down to 7968 unique players over 11 seasons.

Modelling

In our modelling process we used three unique clustering models – K-Means, Agglomerative Clustering, and DBSCAN. The metrics we used to evaluate our models were inertia (where applicable) and silhouette scores. We also evaluated whether the number of clusters made sense in the context of our goal – if we received high silhouette scores after 30 clusters, the model failed as we weren't looking to create 30 unique player profiles. Each model was put through a for loop to uncover what was the optimal K-value (or epsilon value for DBSCAN). Evaluation of the optimal K-value was done using inertia and silhouette scores. The K-Means model was iterated upon twice due to the first model flagging outliers with extremely inflated metrics (as discussed in our EDA). These players were removed due to their small sample size of minutes played. Once the model was run again, we received our best model as it had the lowest inertia possible with four clusters (which made intuitive sense) and a relatively high silhouette score when compared to the three other models. The cluster labels were extracted from this model and attached to our dataset to begin our findings process.

Findings

After performing a group-by function on our cluster labels, we created a horizontal bar chart of relevant key metrics to uncover the differences between our cluster labels (pictured below).



With this visual we were able to gather insights and create a player profile for each cluster label.

Cluster 0 players were the overachievers. These players had a high impact on the games they played in, with extremely high average goals and assists. They typically played in the attacking roles and when they contributed to their teams goals it was typically as the goal scorer.

Cluster 1 players were the reliable performers. Players in this column were typically in defense-oriented roles. It was identified these players are defenders by nature due to the high number of yellow and red cards these players received, which is typical of a defender. They were considered reliable due to their high number of matches played, high average minutes per game, and their high starting frequency. These players had clearly earned their spot in their teams starting eleven and continued to show their value every game.

Cluster 2 players were classified as the unproven youth. Players in this cluster were much younger on average than the other clusters. They had very low matches played, started very little and were receiving very little game time. The lack of stats did tell us something about the players in this cluster – these players have yet to prove to their coaching staff they deserve to be on the field.

Cluster 3 players were seen as the upcoming young talent. The average age of players in this cluster was 24, right between cluster 0 and 2. They were also middle of the pack for matches and minutes played, as well as starting frequency. Players in this cluster were heavily attacking oriented. Given their age and situation, players in this cluster were extremely proficient, impacting the game by scoring goals and assisting just as frequently as the overachievers in cluster 0, but at a younger age!

These findings were not what we initially set out for when we began this project. Initially we wanted to see if we could identify players of similar playstyle. With the findings we uncovered in our modelling it seems rather we identified players of similar skill level. Nonetheless, the findings in the cluster labels make sense in context and are in a similar line to what we initially sought out when beginning this project. Let's now examine how these findings can be used in a business context.

Business Application

In professional soccer, there are many clubs with many different motivations. Some want to challenge for trophies, others want to maintain the position they're in, and others want to focus on developing players in hopes of turning a profit. Using these cluster labels will better help clubs of all motivations identify which players they should look at when performing player scouting. For example, a club that invests and develops young players would likely want to aim for players with the cluster labels 2 and 3, the unproven youth and the upcoming young talent as it would fit their motivations. Restating our initial business use earlier, by providing the club with this reference list it would enable them to expedite their searching and recruitment process.

Next Steps

There are two main things that can be implemented to improve upon the work completed thus far. The first is an actual player recommender system. With the current clustering label system in place, it only gives broad recommendations. By creating a recommender system, we could provide 3-5 specific players given a user-inputted player name that a club may be seeking to replace. The second improvement is to include more specific stats in both offensive and defensive contexts. Having only the basic stats did not allow us to differentiate upon playstyles. By introducing these more granular metrics I believe we would see better results when looking for specific playstyles.