Technical Report 2019-20

Clustering Footballers playing styles

Work by:
Mohamed Khanafer and Aayush Kejriwal

Table of Content

For our analysis, we used **Dataiku.**

Introduction

The scope of our analysis is to aid modern football teams in the process of individuating, selecting and acquiring the best possible fit as an addition to their current line-up.

The rationale behind our idea comes from the fact that, in the last few years, European teams have seen a lot of new players acquisitions that turned out to be a failure because a given footballer's abilities were not in line with what was expected of him.

As no such thing as an optimal cluster exists, our approach consisted of categorizing players based on a number of different attributes derived from their playing style. As we are moving towards data-backed decisions in most industries, we believe that the process of individuating the right members of a team will be ever more reliant on solutions such as ours. This is because once teams understand what is missing from their line-up through playing-data analytics, our cluster analysis will show them which players on the market have the right attributes.

To put it simply, we are running an analysis to look at features such as a player's level of aggressiveness during matches as well as more technical aspects such as his ability to score a goal with his left leg.

Variables & Data

For our analysis, we used a dataset that initially consisted of 148,640 rows and 89 columns. The data contained was about both behavioral and technical qualities of various players on the market, scored in a range from 1 to 20. All the attributes were derived from players' in-match performance and came from the football game *Football Manager*.

After the initial data understanding process, we decided that, in order for our set to be in a condition to be used to create meaningful and actionable cluster, more manipulation was needed.

Features in the original dataset

UID	AerialAbility	Corners	PenaltyTaking
Name	CommandOfArea	Crossing	Tackling
NationID	Communication	Dribbling	Technique
Born	Eccentricity	Finishing	Aggression
Age	Handling	FirstTouch	Anticipation
IntCaps	Kicking	Freekicks	Bravery
IntGoals	OneOnOnes	Heading	Composure
U21Caps	Reflexes	LongShots	Concentration
U21Goals	RushingOut	Longthrows	Vision
Height	TendencyToPunch	Marking	Decisions
Weight	Throwing	Passing	Determination

Flair	NaturalFitness	Loyalty	AttackingMidLeft
Leadership	Pace	Goalkeeper	AttackingMidRight
OffTheBall	RightFoot	Sweeper	DefenderCentral
Positioning	Stamina	Striker	DefenderLeft
Teamwork	Strength	Pressure	DefenderRight
Workrate	Consistency	Professional	DefensiveMidfielder
Acceleration	Dirtiness	Sportsmanship	MidfielderCentral
Agility	ImportantMatches	Temperament	MidfielderLeft
Balance	InjuryProness	Controversy	MidfielderRight
Jumping	Versatility	PositionsDesc	WingBackLeft
LeftFoot	Adaptability	AttackingMidCentral	WingBackRight
	Ambition		

Data Preparation Procedures

Here are the main steps taken to prepare and clean our data in order to run our model.

Since the objective of our analysis is to cluster players playing in the same position on the basis of their playing style, it is crucial to segregate players in the dataset based on their position. Running a similar cluster analysis without segregating players into their positions would likely lead to clusters representing the positions rather than the different styles of play within each position.

The columns 'Goalkeeper', 'Sweeper', 'Striker', 'AttackingMidCentral', 'AttackingMidLeft', AttackingMidRight', 'DefenderCentral', 'DefenderRight', `DefenderLeft', 'DefensiveMidfielder', 'MidfielderCentral', 'MidfielderRight', 'MidfielderLeft', 'WingBackLeft' and 'WingBackRight' represent the players competency in each of these positions. They contain values ranging from 1 to 20, with a 20 implying that the player plays best when deployed in the position represented by the column header. We noticed that for each player (row), exactly 1 of the above-mentioned columns had a value of 20, and therefore represented the players position.

For the purpose of our analysis we needed a single column representing the position, the value of which would be the column header from those mentioned above that contained a value of 20 for the respective player. To do this we deleted cell values for all non 20 values in each of the above-mentioned columns, therefore having either a NULL or a 20 value in the columns. Also, as mentioned earlier, each row had exactly one of the columns with a 20 so the rest were NULL.

We used the IF formula to create the column labelled 'PlayingPosition':

if(isNotNull(Goalkeeper), "Goalkeeper", if(isNotNull(Sweeper), "Centerback", if(isNotNull(Striker), "Striker", if(isNotNull(AttackingMidCentral), "Att.Mid", if(isNotNull(AttackingMidLeft), "Winger", if(isNotNull(AttackingMidRight), "Winger", if(isNotNull(DefenderCentral), "Centerback", if(isNotNull(DefenderLeft), "Fullback", if(isNotNull(DefenderRight), "Fullback", if(isNotNull(DefensiveMidfielder), "Def.Mid", if(isNotNull(MidfielderCentral), "CentralMid", if(isNotNull(MidfielderLeft), "Winger", if(isNotNull(MidfielderRight), "Winger", if(isNotNull(WingBackLeft), "Fullback", if(isNotNull(WingBackRight), "Fullback", "ERROR"))))))))))))))))

The resulting column 'PlayingPosition' contained the following 8 possible values: 'Goalkeeper', 'Centerback', 'Fullback', 'Def.Mid', 'CentralMid', 'Att.Mid', 'Winger' and 'Striker'. (The 'ERROR' used in the formula was used to check for problems with the formula)

We then deleted the 15 above mentioned columns that were used to get this information.

(The column `PositionsDesc' provides us with the position as well. However, it contains multiple positions for the same players within the same column, separated by a '/'. Therefore we split the column to get multiple columns with at most one position per column. We assumed the first of the resulting set of columns represented the best position of the player, since it is impossible to individually go over each player individually and select the right position. However, upon inspecting the result, we noticed many players were not assigned to their best position. Therefore we deleted the 'PositionsDesc' column along with all the split column outputs, and used the above mentioned method to arrive at a more accurate value for player position.)

- Further deleted the following 21 columns since they do not have a bearing on the playing style of a footballer and are therefore not required in our analysis: 'NationID', 'Born', 'Age', IntCaps', 'IntGoals', 'U21Caps', 'U21Goals', 'Longthrows', `Leadreship', `Consistency', 'ImportantMatches', 'InjuryProneness', `Versatility', 'Adaptability', `Ambition', 'Loyalty', 'Pressure', 'Professional', 'Sportsmanship', 'Temperament' and 'Controversy'.

- The columns 'LeftFoot' and 'RightFoot' represented the players comfort in using the specific foot. It contained values ranging from 1 to 20, with 20 implying that the player is comfortable using that foot, or that it is his natural foot. Every player had a 20 in at least one of the 2 columns. To compute their ability with their non-dominant foot, we created a new column called 'WeakFoot', using the formula: min(LeftFoot, RightFoot).

- Since we were left with a large number of equally important player attributes, we grouped similar attributes together to form new columns, and then deleted the old individual attributes used to calculate the grouped attributes. The players are rated on each attribute on a scale of 1 to 20, with 20 being the best at that skill. Since our new grouped attributes were made of varying number of individual attributes, a simple addition of the base attributes would lead to different possible totals of the newly created attributes. This could potentially lead the clusters to give more importance to certain created attributes over others. To prevent this, we took an average of individual attributes while creating the new attributes.

Following are the newly created attributes with their corresponding constituent attributes in brackets:

1. GKHighBalls (AerealAbility, CommandOfArea, Handling, TendencyToPunch)
2. GKDistribution (Throwing, Kicking)
3. GKCommunication (Communication)
4. GKShotStopping (Eccentricity, OneOnOnes, Reflexes, RushingOut)
5. Crossing (Corners, Crossing)
6. Passing (Vision, Passing, Teamwork)
7. Dribbling (Dribbling, FirstTouch, Technique, Flair)
8. Scoring (Finishing, PenaltyTaking)
9. LongRange (FreeKicks, LongShots)
10. Movement (Acceleration, Balance, Agility, Pace)
11. Positioning (Anticipation, OffTheBall, Positioning)
12. AerealThreat (Heading, Jumping, Strength)
13. Defending (Marking, Tackling)
14. Endurance (Determination, Workrate, NaturalFitness, Stamina)
15. DecisionMaking (Composure, Concentration, Decisions)
16. Aggression (Bravery, Dirtiness, Aggression)

- These steps got us the prepared dataset, which we then split into 8 datasets based on playing positions as discussed in step 1. For this we used the SPLIT function under ACTIONS.

- We decided to make clusters based on playing styles for players in the defensive midfield position. Therefore, we focused on the split dataset that had 'PlayingPosition' as Def.Mid.

- Since midfielders do not need goalkeeping skills, we deleted the 4 columns representing goalkeeping skills. These columns are 'GKHighBalls', 'GKDistribution', 'GKCommunication' and 'GKShotStopping'.

- Running a cluster analysis on the remaining attributes we calculated earlier revealed a flaw in our approach. Since some players were bad in some good, the clusters ended up grouping good players together and bad players together, without revealing anything about their playing styles. The problem was that we were using absolute values of their skill in each attribute. Bad players had low values across all attributes while good players had high values.

- To tackle this problem, we created a new column called 'OverallAbility' as a sum of 9 relevant attributes calculated in step 4.

  (Crossing+LongRange+Scoring+Defending+DecisionMaking+Dribbling+Positioning+Passing+AerealThreat)

  We did not consider the attribute 'Aggression' in this calculation as it represents the nature of a player and has nothing to do with their skill level. We also did not include 'Movement' and 'Endurance' as these are physical traits that are innate to a player and do not improve much over time. While they do impact the ability of a player to perform well, they are not a skill. In other words, the attribute we chose to calculate 'OverallAbility' were attributes that according to us can be improved upon as a player becomes better at the game.

- Once we had the column ‘OverallAbility’, we created a new column for each of its constituent attributes. These new columns contained the particular attribute as a percentage of ‘OverallAbility’. All the columns were names as the attribute the belonged to followed by a ‘%’ sign. For example, ‘Passing%’ was a column created, showing a players passing attribute as a percentage of his overall ability.

  The 9 columns thus created were: ‘AerealThreat%’, ‘Passing%’, ‘Positioning%’, ‘Dribbling%’, ‘DecisionMaking%’, ‘Defending%’, ‘Scoring%’, ‘LongRange%’ and ‘Crossing%’.

  By using the percentage values, we treated each player as though they were all of the same level, or equally good overall. While this is not true in reality, the objective of the clustering exercise was not to distinguish good players from bad, but to distinguish players based on the attributes that they rely on more heavily while playing and those that they are relatively weak at. This gives us a better understanding of playing styles instead of ability.

  Of course, once clusters are formed by grouping similar playing styles, an overall ability filter can be applied to select players that not only match the style a club is looking for but also has the level of ability required.

The Clustering Model

To train the clusters, we chose the default k-means clustering, and built clusters with k= 3, 4 and 5. We chose against trying higher values of k because having too many cluster would at some point start to defeat the purpose of grouping similar players together by being too exact in the matching criteria. Also, at some point having too many clusters would be useless since the size of the clusters would themselves become smaller, leaving limited options of players to choose from once a club identifies the cluster they want to search in, we decided this limit should be set at 5.

For our final cluster model, the features we chose were ‘Passing%’, ‘Positioning%’, ‘Dribbling%’, ‘DecisionMaking%’, ‘LongRange%’ and ‘Crossing%’. These attributes represent the supplementary skills a defensive midfielder would need in a football match and could therefore be a basis of useful differentiation.

Attributes such as ‘AerealThreat%’ and ‘Scoring%’ represent skills that a defensive midfielder, by the nature of their position, would almost never need to use in a match. Differentiating players on the basis of them would therefore be of no use to a club since they wouldn’t need those skills. ‘Defending%’ was also left out because it represents a core skill and the primary job of any defensive midfielder. Using it as a feature doesn’t tell us much.

It serves clubs much better to filter players on the basis of their defensive skills after finding players that match the style they need. Once again, we only used useful supplementary skills for a defensive midfielder since differentiating clusters on this basis serves the purpose of our clustering exercise better.

In order to get a higher silhouette and thus more defined clusters, we tried to group 'Attribute%' values and reached a silhouette of up to 4.29. However, the clusters did not make sense anymore, and therefore discarded it and continued with the model here defined.

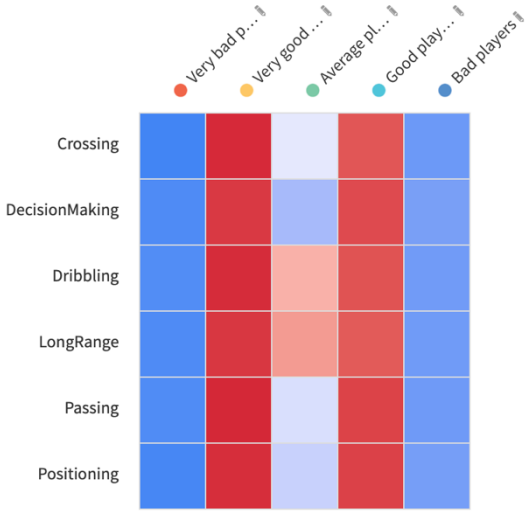
The Results

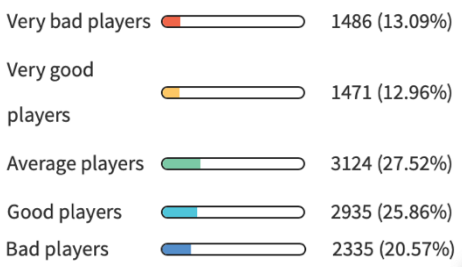
Hypothesis

In the last decades, defensive midfielders mainly had one style of play, that being of "game-breakers". They were required to be good at breaking up the opponent's plays and at passing the ball around to build up the plays of their team.
That rigid role has evolved in modern football and defensive midfielders have now re-invented their playing styles, deriving other styles as well.
Before our cluster analysis, we thus hypothesized that clusters of playing styles could be well defined and this would help clubs better make decisions in their recruitment processes.

First Problematic clusters

The first problematic cluster analysis (as mentioned in the Data Preparation section) run using absolute values of players' skill in each attribute yielded the following heatmap:



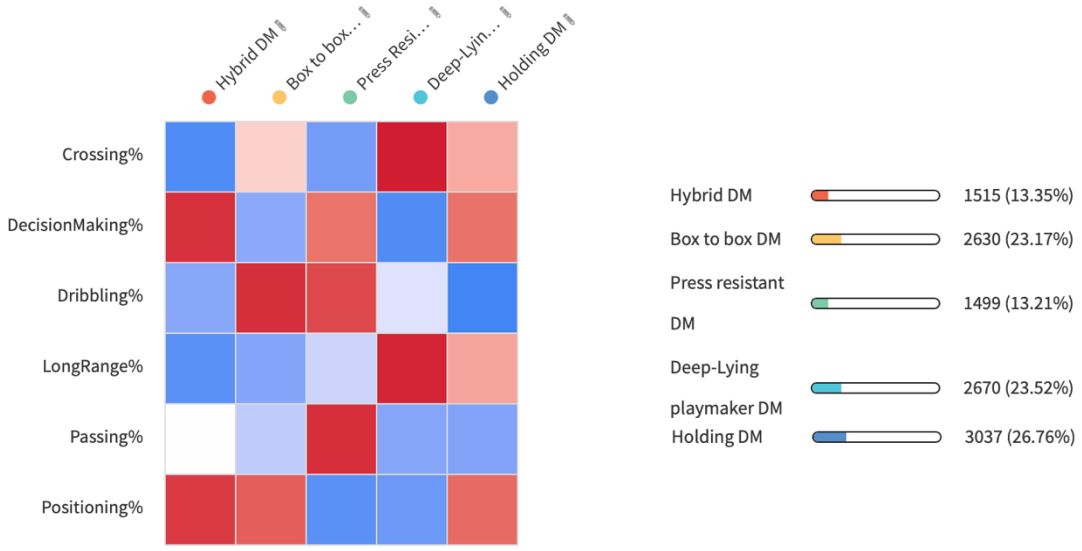


No specific playing styles could be derived from these cluster because the bad players had low values across all attributes while good players had high values.

Here is when we added the new columns that contained the particular attribute as a percentage of 'OverallAbility' to be able not to distinguish good players from bad, but to distinguish players based on the attributes that they rely on more heavily while playing and those that they are relatively weak at. The goal again being to better understand the styles of playing instead of ability.

Final clusters

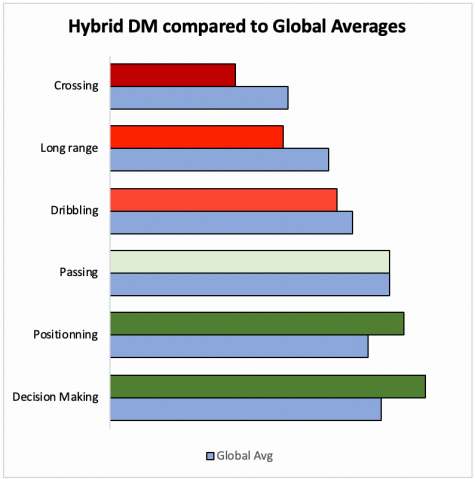
Running the cluster analysis again with this time the proper measures gave us more defined style of plays as shown in the heatmap below:



As a first impression, the clusters made more sense with more defined and different clusters on the heatmap. The silhouette value of this analysis was 0.1101. Running the same analysis with a detection of outliers gave us a slightly lower silhouette and thus decided not to detect outliers in our analysis.

Description of clusters

- **Hybrid DM**: With good decision making and positioning, this category of midfielders represent 13.35% of the defensive midfielders in the dataset. They can be seen as players that not only defend but can easily contribute to the attacking parts of the play from where their naming as "hybrid".

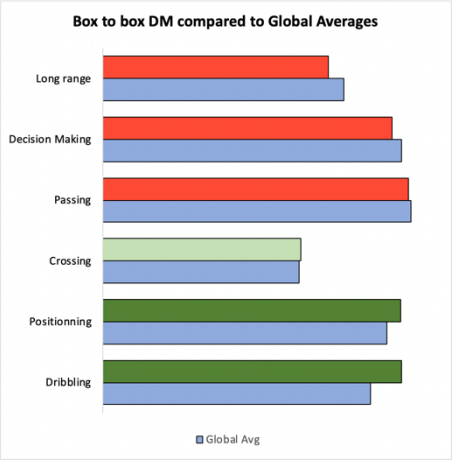


DecisionMaking% is in average 16.53% greater: mean of 14.62 against 12.55 globally

Crossing% is in average 29.70% smaller: mean of 5.80 against 8.25 globally

LongRange% is in average 20.84% smaller: mean of 8.03 against 10.14 globally

Box to box DM: with above the average dribbling and positioning abilities, these players can be seen as a good link between the two opposite sides of the fields. They represent 23.1% of the defensive midfielders in the dataset.

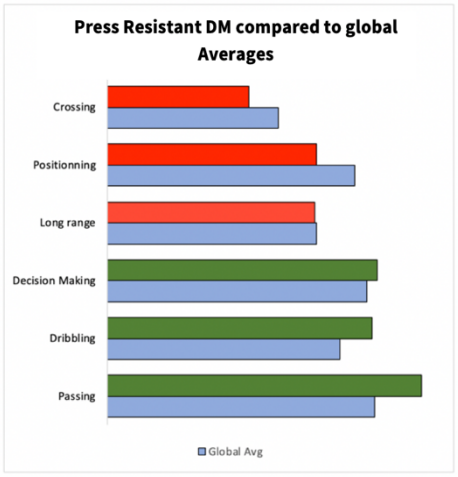


Dribbling% is in average 11.62% greater: mean of 12.55 against 11.24 globally

Positioning% is in average 4.71% greater: mean of 12.51 against 11.95 globally

LongRange% is in average 6.72% smaller: mean of 9.46 against 10.14 globally

Press Resistant DM: this cluster is the lowest represented in the dataset of defensive midfielders with 13.21%. Their set of technical skills differentiate them from the other clusters with higher passing, dribbling and decision-making abilities. This profile of players with such a playing style are more and more demanded in modern football.

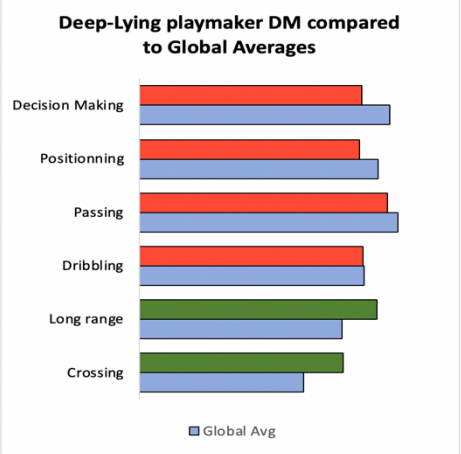


Passing% is in average 17.63% greater: mean of 15.22 against 12.94 globally

Positioning% is in average 15.13% smaller: mean of 10.14 against 11.95 globally

Dribbling% is in average 14.06% greater: mean of 12.82 against 11.24 globally

Deep-Lying playmaker DM: with greater crossing and long-range abilities than other midfielders, this cluster of players can be seen as playmakers in the forefront, a role normally given to more offensive players. This group of players represent 23.52% of the defensive midfielders in the dataset.

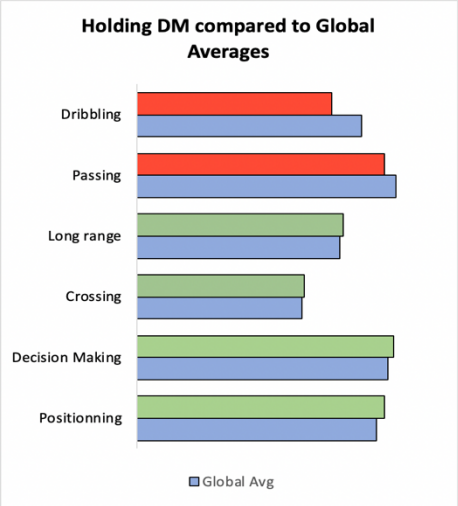


Crossing% is in average 24.10% greater: mean of 10.24 against 8.25 globally

LongRange% is in average 17.43% greater: mean of 11.91 against 10.14 globally

DecisionMaking% is in average 11.28% smaller: mean of 11.13 against 12.55 globally

- **Holding DM**: except for dribbling, this cluster of players are the most balanced ones in terms of overall attributes. This is the group that resembles the most this “typical” playing style long associated with this position. And they represent the biggest part of the dataset with 26.76%.



Dribbling% is in average 13.46% smaller: mean of 9.73 against 11.24 globally

Passing% is in average 4.33% smaller: mean of 12.38 against 12.94 globally

Positioning% is in average 3.26% greater: mean of 12.34 against 11.95 globally

Validation of the Results

Given our knowledge in football, we were able to validate this model. What we did is that we looked at the most known players in each cluster and analyzed if the cluster the player was put in by the model made sense.
And for each of the 5 clusters, we chose 3 players that we know and confirmed they effectively had similar playing style. Here are the players we chose and validated the model on:
For **Hybrid DM**: Sami Khedira, Sergio Busquets, and John Obi Mikel
For **Box to Box DM**: Emre Can, Daniele De Rossi, and Fabinho
For **Technically better DM**: Nemanja Matic, Maxime Gonalons, and Marcos Llorente
For **Deep-lying playmaker DM**: Luiz Gustavo, Andrea Pirlo, and Xabi Alonso
For **Holding DM**: Fernandinho, Casemiro, and Gareth Barry

In our analysis, we chose Defensive Midfielders to run our model and make sure it was accurate. However, this is just a pilot project and a similar analysis can be run on all other positions as well.