Box-To-Box: An Analysis and Clustering of European Soccer Players

BOSTON UNIVERSITY

By Chanan Sukangium, Tachapan Kongboonma, Zhenjie Cao

Introduction

What differentiates a striker from a defender? Soccer is undoubtedly a tactical game and every position is unique in its purpose and function. Our team is curious to see what exactly makes a player more suitable to be a striker than a defender whether it's their height, dribbling skill, or even their weight.

Goals

We aim to figure out:

1. Which attributes best influence the overall rating of a player through correlation?
2. Which skill attributes are prominent and significant among specific positions (striker, midfielder, defender)?

The Data

Relevant data for our project have been retrieved through downloading a European Soccer Player Kaggle Dataset as an SQLite then converting relevant tables in database shell to .csv to be used in data frame and duplicate data and those with 'null' fields are removed. All of the numerical and subjective ratings (eg. dribbling, ball control, strength) are from a game EA Sports Fifa that uses real professional scouts to assign numerical ratings to players' attributes within the range 1 to 100; we feel that this is the least biased method to rate a soccer player. Some of the key features of the data include: soccer player's height, weight, overall rating, dribbling, agility, long shot.

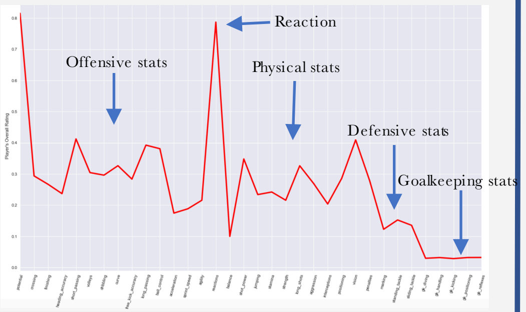
Reference

https://www.kaggle.com/hugomathien/soccer

Analysis 1

1. Compute correlation coefficient between the overall rating of player against every single attributes of a player (acceleration, stamina, dribbling etc.)
2. Compile all the correlation value and plot in a single line graph to determine which attribute has the highest influence on the overall rating of a player

Correlation Coefficient Between Overall Ratings and Individual Attribute

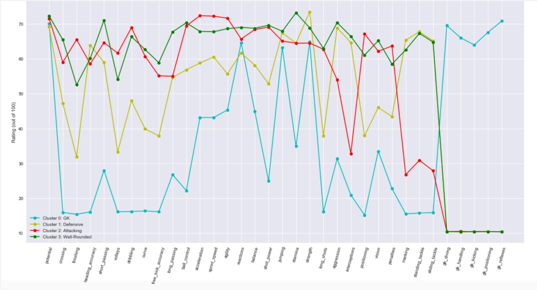


- Our result shows that offensive and physical attributes tend to have higher influence than defensive and goalkeeping attributes. Comparing this finding to the Ballon d'Or 2017 awards (annual shortlist of world's top 30 soccer players), over 73% of the players in the list are offensive players, 16.7% defensive players and 10% are goalkeepers; this indicates the dominance of offensive attributes in influencing overall football ability.
- "Reaction" is the sole outlier in this graph with correlation value (~0.79) significantly higher than the second highest being "short passing" (~0.42). One explanation for this is the fact that reaction is a vital trait for all positions in soccer, from attackers to goalkeepers.

Analysis 2

1. Perform K-means clustering on data frame with 34 attributes.
2. Merge the result back into the original data frames
3. Create 4 data frames based on the cluster assignment (0-3)
4. Find average of the 34 attributes for each data frames and plot to compare with other clusters

K-means clustering (K = 4) based on 34 player attributes



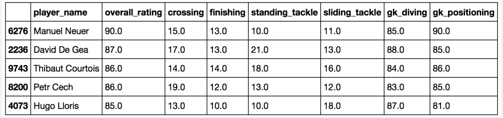
- Each cluster shows varying average value on different attributes which can be interpreted into different type of soccer position.
- Based on the result, we found that cluster 0 represents goalkeepers due to high goalkeeping and physical attributes.
- Cluster 1 represents defensive players due to high physical and defensive attributes (strength and tackling)
- Cluster 2 represents offensive players due to high offensive attributes (finishing, dribbling)
- Cluster 3 represents well-rounded, versatile players with relatively high rating across majority of the attributes except for goalkeeping stats.

Conclusion

To check the accuracy of our labelling of clusters from graph in Analysis 2, we listed the top 5 players from each cluster to check with real data and see whether these players actually play for these positions or not.

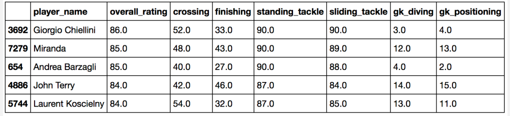
Data frame showing top 5 players from each cluster

Cluster 0: Goalkeepers



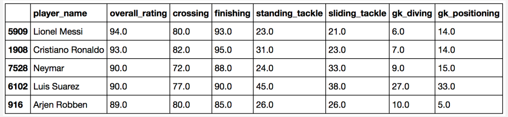
	player_name	overall_rating	crossing	finishing	standing_tackle	sliding_tackle	gk_diving	gk_positioning
6276	Manuel Neuer	90.0	15.0	13.0	10.0	11.0	85.0	90.0
2236	David De Gea	87.0	17.0	13.0	21.0	13.0	88.0	85.0
9743	Thibaut Courtois	86.0	14.0	14.0	18.0	16.0	84.0	86.0
8200	Petr Cech	86.0	19.0	12.0	13.0	12.0	83.0	85.0
4073	Hugo Lloris	85.0	13.0	10.0	10.0	18.0	87.0	81.0

Cluster 1: Defensive Players



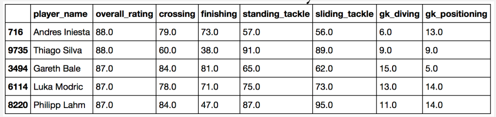
	player_name	overall_rating	crossing	finishing	standing_tackle	sliding_tackle	gk_diving	gk_positioning
3692	Giorgio Chiellini	86.0	52.0	33.0	90.0	90.0	3.0	4.0
7279	Miranda	85.0	48.0	43.0	90.0	89.0	12.0	13.0
654	Andrea Barzagli	85.0	40.0	27.0	90.0	88.0	4.0	2.0
4886	John Terry	84.0	42.0	46.0	87.0	84.0	14.0	15.0
5744	Laurent Koscielny	84.0	54.0	32.0	87.0	85.0	13.0	11.0

Cluster 2: Offensive Players



	player_name	overall_rating	crossing	finishing	standing_tackle	sliding_tackle	gk_diving	gk_positioning
5909	Lionel Messi	94.0	80.0	93.0	23.0	21.0	6.0	14.0
1908	Cristiano Ronaldo	93.0	82.0	95.0	31.0	23.0	7.0	14.0
7528	Neymar	90.0	72.0	88.0	24.0	33.0	9.0	15.0
6102	Luis Suarez	90.0	77.0	90.0	45.0	38.0	27.0	33.0
916	Arjen Robben	89.0	80.0	85.0	26.0	26.0	10.0	5.0

Cluster 3: Well-Rounded Players



	player_name	overall_rating	crossing	finishing	standing_tackle	sliding_tackle	gk_diving	gk_positioning
716	Andres Iniesta	88.0	79.0	73.0	57.0	56.0	6.0	13.0
9735	Thiago Silva	88.0	60.0	38.0	91.0	89.0	9.0	9.0
3494	Gareth Bale	87.0	84.0	81.0	65.0	62.0	15.0	5.0
6114	Luka Modric	87.0	78.0	71.0	75.0	73.0	13.0	14.0
8220	Philipp Lahm	87.0	84.0	47.0	87.0	95.0	11.0	14.0

Our findings are fairly accurate where top 5 players of Cluster 0, 1, and 2 are belong in the predicted cluster. We found in Cluster 3 that 4 players are midfielder and the last player is a defender, so well-rounded cluster could be interpreted as midfielders.

Future Work

Given more time, we will consider other factors in determining a players position including dominant foot, height, and weight to see if positions can be broken down further more eg. Striker vs Winger

With more data from outside European leagues, we can build a model to predict a player's ideal position based on their attributes.