Chanan Suksangium
Tachapan Kongboonma
Zhenjie Cao

CS506 Final Project - Analysis and Clustering of European League Soccer Players

Abstract

The goal and motivation of this data analysis project is to cluster European league soccer players based on all of their 34 skill attributes. We expect the players to be grouped by their positions and play style in real life; the skill attributes used in this project include features such as ‘overall rating’, ‘dribbling’, ‘tackling’, and ‘goalkeeper diving’ from the range 0 to 100 (more information about this dataset may be found below under ‘Data’ section). We feel that these 34 attributes are enough to distinguish players by type (eg. defensive vs offensive) except for perhaps the versatile jack-of-all-trades type of players. For data analysis, this project uses K-Means clustering to group similar players according to attributes and correlation coefficients to determine if any particular skills are more influential a player’s overall rating.

Introduction and Hypothesis

As soccer enthusiasts, our team strived to understand which attributes of a professional soccer player define their play style and position on the field. We originally wanted to analyze players based on height and weight as explained in the original Project Proposal, but found those data to be quite dull and not insightful after performing various forms of clustering. Thus we switched to work with data containing 34 columns of skill attributes to draw more interesting results.

This forces us to form a new hypothesis: we expect that individual skill attributes define a professional soccer player’s role on the field, so this study analyzes 34 different attributes, and performs clustering; we predict to see players grouped by positions: forwards, defenders, goalkeepers, and midfielders since the skill attribute ratings for each position should vary from each other eg. forwards have high offensive stats but low defensive stats. One possible application of this project is to identify players playing out of position if their assigned cluster conflicts with real life roles because their attributes are more suitable elsewhere on the field.

Data

The dataset used in this project is the European Soccer Database from Kaggle, which is a 300mb SQLite file that contains 8 different tables, consisting of data about players, teams, and matches. For the purpose of this project, we only needed ‘Player’ and ‘Player_Attributes’ tables since our goal is to analyze players.

- ‘Player’ table contains concrete information and objective data eg. name, height, weight.
- ‘Player_Attributes’ table contains subjective data, which is overall rating of a player and 34 other individual attributes that describe a player such as ‘Finishing’, ‘Short Passing’, and ‘Standing Tackle’. These attributes are in the range 1 to 100 and are directly from a highly popular video game EA Sports FIFA (best selling console game in 2016), who employs real professional soccer scouts and analysts to rate players and each of their attributes. We feel that this is the most objective way of rating players numerically.

Player table:

id	player_api_id	player_name	player_fifa_api_id	birthday	height	weight
1	505942	Aaron Appindangoye	218353	1992-02-29 00:00:00	182	187
2	155782	Aaron Cresswell	189615	1989-12-15 00:00:00	170	146
3	162549	Aaron Doran	186170	1991-05-13 00:00:00	170	163

Player_Attributes table:

id	player_fifa_api_id	player_api_id	date	overall_rating	potential	preferred_foot	attacking_work_rate
1	218353	505942	2016-02-18 00:00:00	67	71	right	medium
2	218353	505942	2015-11-19 00:00:00	67	71	right	medium

Data Extraction and Cleaning

1. The first step involved using command line shell for SQLite to query relevant tables and extract them as .csv file, so the format is friendly for Pandas library.
2. After reading in 2 .csv files into notebook as Pandas dataframe, irrelevant and duplicate columns (birthday, player_api_id) are dropped, duplicate rows are removed based on 'player_fifa_api_id' since data contains some duplicate ratings from older versions of FIFA (latest one is kept), and players with Null in any attributes are removed.
3. Two data frames are merged on 'fifa_api_id' so each player corresponds to their attributes for further analysis and clustering.

Analysis 1: Correlation Coefficients for Skill Attributes

Explanation:

The first data analysis step taken is to find correlation coefficients between overall rating of a player and every single skill attributes; this will give a further understanding of the dataset regarding influence of each attributes on overall rating, and may help give more explanation in the main analysis after clustering players based on attributes.

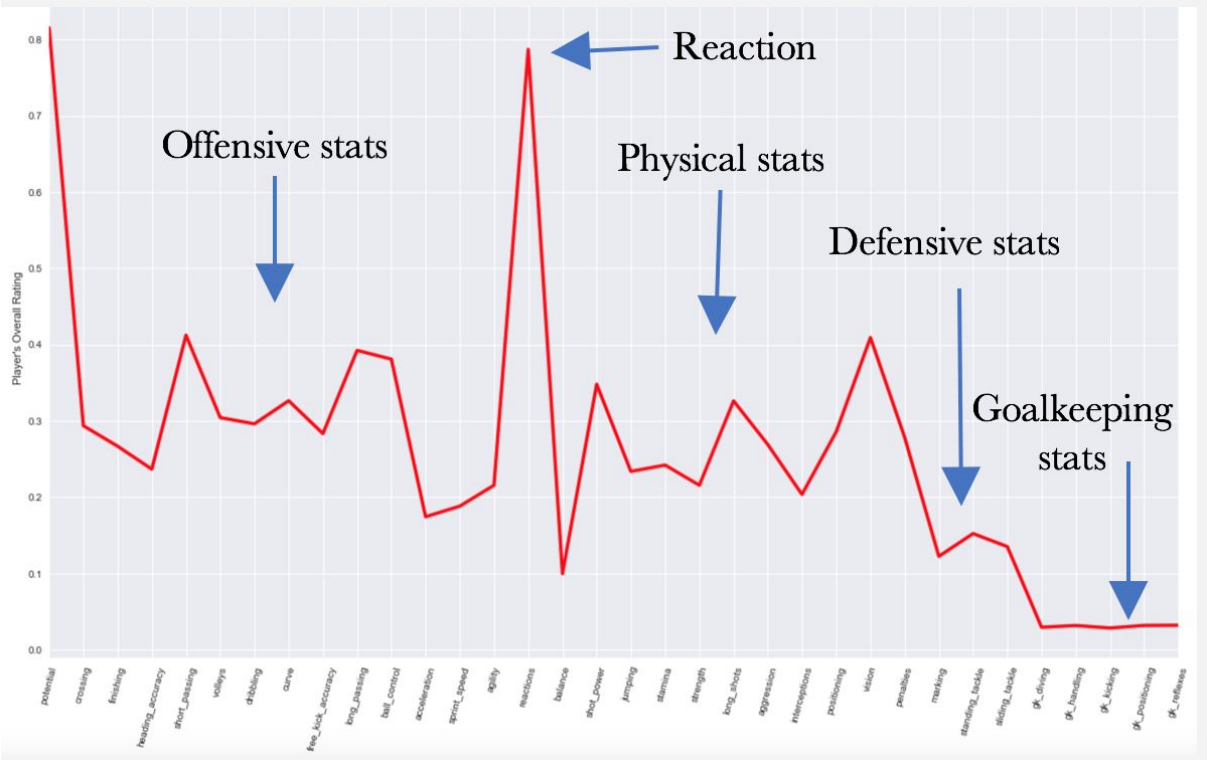
Steps:

1. A list of length 34 is created through list comprehension that calls correlation (.corr()) between 'overall_rating' column with every single skill attribute
2. Each correlation coefficient is plotted on a line graph with range 0 to 1 on Y-axis as correlation coefficient value and each skill attribute along the X-axis

Result:

Looking at the plotted correlation coefficient graph below, it could be seen that the offensive (dribbling, volley, finishing) and physical (stamina, strength, vision) attributes are significantly higher in correlation to overall rating compared to defensive (sliding tackle) and goalkeeping attributes. It is difficult to make a concrete statement about the accuracy of this in real life since the ratings are subjectively given after all, but linking this to the most recent Ballon d'Or Awards (shortlist of world's top 30 soccer players), 22 (~73%) players have attacking play styles, 5 (16.7%) players are defensive, and 3 (10%) players are goalkeepers. This shortlist definitely supports our findings, but soccer skill is too complex to be put into numbers, so perhaps we could conclude that offensive attributes are more influential in our perception of a player's skill due to the culture of the sport eg. highlights and score sheet only show goals and assists, not

defensive higlights. Aside from ‘potential’ correlation coefficient that could be ignored, there is a sole outlier attribute which is ‘reaction’ at ~0.79 because the second highest coefficient for attribute is ‘short_passing’ at ~0.42. One explanation we came up for this outlier is that reaction is the only skill trait that is prevalent across all positions in soccer from striker to goalkeeper.



The line graph above shows correlation coefficient (0-1) for each of the 34 skill attributes

Analysis 2: K-Means Clustering of Soccer Players by Skill Attributes

Explanation:

To answer our original hypothesis on determining a player’s position based on attributes, we decided to perform k-means clustering (k = 4: 1 cluster for each position in soccer) to see how attributes predict position of a player. K-Means clustering was used here for hard assigned clusters because we feel that players belong to a specific position on field, so soft assignment clustering will not define player type as accurately; our attempt with hierarchical showed that cluster averages are less distinct than K-Means due to the myriad average players in dataset.

Steps:

1. K-Means clustering(k = 4) is called on dataframe of all players with 34 attributes columns
2. List of cluster assignment (0-3) is appended back to original data frame to match players with their respective cluster assignment
3. Players are sorted by clusters and separated in 4 data frames and average values for all attributes of each data frame is calculated for comparison and visualization
4. Averages of all attributes are plotted on graph to compare differences between clusters

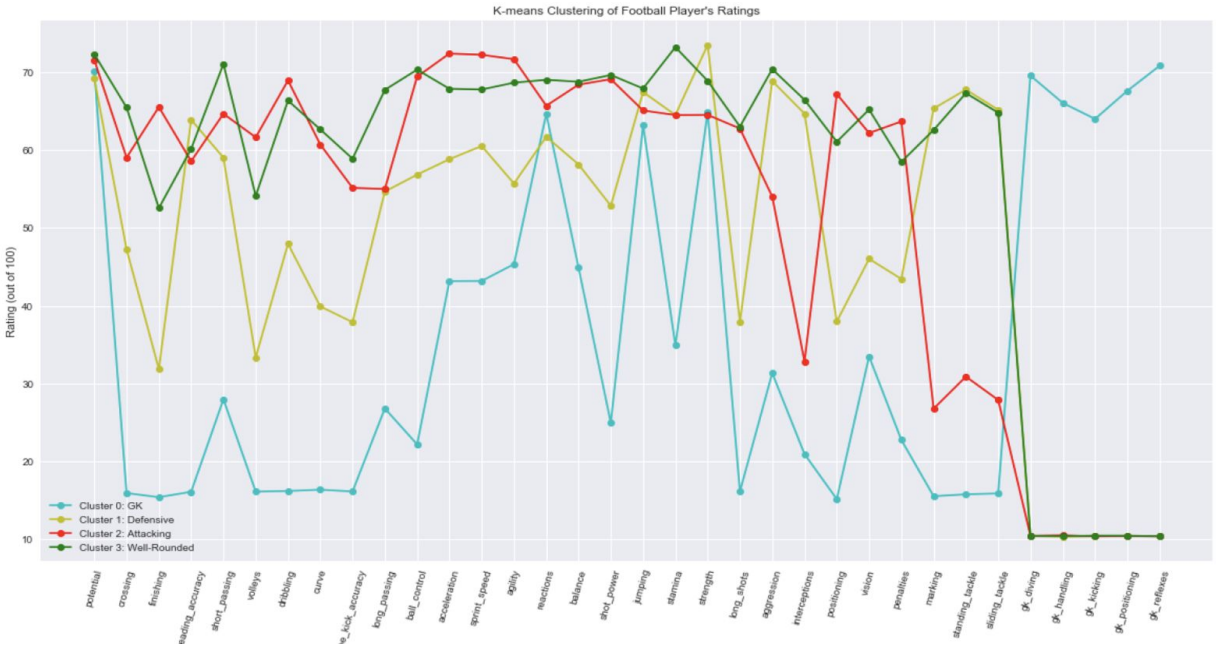
Results:

Looking at the graph below, we interpreted that:

- Cluster 0 (Cyan) contains goalkeepers due to high goalkeeping and physical stats

- Cluster 1 (Yellow) contains defensive players due to high tackling stats
- Cluster 2 (Red) contains attacking players due to high attacking stats and low defensive
- Cluster 3 (Green) contains well-rounded players with no low stats aside from gk

We can note that physical attributes (stamina, strength) are generally high across each cluster because they are still essential for professional soccer players to be fit despite position on field.



The line graph above shows correlation average of cluster for each of the 34 skill attributes

Additionally, 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 belong to the appropriate clusters. The 4 tables below represent the top 5 players from each cluster along with the accuracy percentage on predicting the position of these players.

Cluster 0: Goalkeepers (100%)

	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 (100%)

	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 (100%)

	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 (80%)

	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

Conclusion and Future Work: Beyond Europe

Looking back at our 2 original goals, the aim of this project is to :

1. Figure out which attributes most influence the overall rating of a player through correlation coefficient to gain insight of skill rating system of this dataset.
2. Identify skill attributes are most prominent for a specific position on the field.

The results obtained from analysis 1 & 2 are just as we hypothesized: skill attributes in professional soccer do define a player's position on the field as the clustering accurately grouped similar-role players together.

Analysis 1 provides insightful correlation coefficients between a player's overall rating and individual stats because soccer tends to highlight the performance of offensive players more than defensive players. This claim is supported by the skewed distribution of offensive players in Ballon d'Or Awards (world's top 30 soccer players). Our findings from analysis 1 reflects this with higher correlation between overall rating and offensive compared to defensive attributes.

Analysis 2 shows that skill attributes in soccer are quite specific to a position on the field. This allows the clusters to be fairly distinct from each other since each position in soccer have their strengths and weaknesses; a smaller, fast-paced winger would not be suitable as a defender. The 4 clusters generated from the K-Means algorithm are distinct enough for us to identify which position these clusters represent by looking at the cluster averages for each attributes. Our decision of using K-Means algorithm is based on the fact that the algorithm works well with a high dimension set of data; in our case, we have 34 attributes. Our attempt of trying hierarchical clustering resulted in less distinct clusters based on averages, most likely due to large amount of average players that dominate the dataset with less insightful ratings. Our project shows an accurate prediction because we were able to correctly match soccer positions with clusters based on average numbers on the line graph.

Despite our findings being rather accurate, the result of this project should not be taken too heavily into consideration because after all, the skill attributes of these soccer players cannot be assigned a discrete, numerical value. EA Sports FIFA is, however, the most comprehensive source in numerically rating professional soccer players, and their accuracy continues to improve as they release new versions of the game annually.

Given more time, we want to explore other physical factors that may additionally help determine a player's position including dominant foot, height, and weight; it will be interesting to see how these other physical stats may affect our clustering, and perhaps the positions can be broken down even further more, for example to a striker or a winger rather than just an offensive player. We also want to expand our data beyond European Professional Soccer Leagues to the rest of the world to also test accuracy there. With enough data, we wish to build a prediction model to help predict a player's position based on limited data, which can potentially have its application in predicting play styles of new or unknown soccer players.