A Rainy Day in Stoke: Comparing Soccer Players

Justin Chen, Morissa Chen, Mihir Dhoot, Ashwin Nallan

Intro

Soccer is the world's most popular sport, with a fanbase of over 3.5 billion people. And one question has always divided them: "Can Lionel Messi, the best player in the world, dominate on a cold, rainy night in Stoke"? We wanted to try and answer that quesion, and in the process bring fans both casual and intense a new way to look at soccer.

The aim of this project is to build an analysis of all soccer players regardless of league, season and even gender. We will take each player's statistical profile, and build and visualize player comparisons. Through our tool, a fan could find lesser-known players similar to their favorites -- and even try to answer if Messi's playing style would be successful in England.

Data

We scraped data from **fbref.com**, a website that provides soccer statistics for most competitions. Statistics were scraped at the **player level per season** to compare individual players on a season-by-season basis. Additionally, statistics were scraped at the team level for additional context on individual performance.

Our team-level data was fairly high-level, focused on **wins/losses** and **basic stats** like goals scored and allowed. Our player-level data presented many issues with incomplete or spotty data. To accomodate this, we filtered the data to include only attributes with over 50% of data available, preserving **102,000 player-seasons** of information across **33 attributes**.

	index	M	W	null_avg
21	('Performance', 'Ast')	0.002057	0.000602	0.001329
22	('Performance', 'CrdR')	0.000051	0.000000	0.000026
23	('Performance', 'CrdY')	0.000000	0.000000	0.000000
24	('Performance', 'Fls')	0.244371	0.618045	0.431208
25	('Performance', 'Gls')	0.000010	0.000000	0.000005
26	('Performance', 'OG')	0.761358	0.168421	0.464889
27	('Performance', 'PK')	0.000432	0.000602	0.000517

These attributes include goals, shots on target, assists, tackles+interceptions, fouls, and yellow and red cards.

League Comparator

Soccer fanatics have long debated which league is the "toughest" to perform in. Our project title itself is derived from the famous question "Can Messi dominate on a cold, rainy night in Stoke", which describes the "merciless, physical" playing style in England. Thus, when comparing player performance across all leagues, we first wanted to **account for a league's style**.

We attempted to **build profiles** for each league based on its attacking and defensive statistic distributions to **normalize player performances** accordingly. The league comparator tool allows a user to see the differences across leagues with respect to each statistic, from Goals and Assists to Fouls and Yellow Cards. We were also able to generate clusters of similar leagues. Our analysis expanded past prior literature by including leagues beyond the top 5 in Europe.

Clustering

We conducted a **k-means clustering** on player data to further analyze and find similarites between players. Using the Elbow Method, we chose to have 6 distinct clusters. Although the **sihlouette coefficient** of our clustering model was not high, we found **key distinctions** between the clusters.

Cluster 2, for example, clearly clustered **high-quality defensive players**, with high tackle+interception rates. Cluster 4 identified **good strikers with frequent playing time**, with the highest "% of Starts" and "Goals+Assists per 90 minutes".

cluster	Count	Age	ATT	MID	DEF	%Min	%MP	%Starts	SoT/90	G+A/90	Tkl+Int	Fls	CrdY	CrdR
0	39024	24.79	0.22	0.52	0.49	0.21	0.31	0.20	0.31	0.07	1.07	9.87	1.48	0.10
1	11261	25.42	0.68	0.63	0.06	0.64	0.79	0.66	0.97	0.49	1.53	34.56	3.63	0.19
2	3235	26.01	0.14	0.51	0.58	0.65	0.74	0.66	0.25	0.14	70.43	28.53	4.84	0.25
3	13858	24.68	0.69	0.52	0.09	0.22	0.38	0.21	0.99	0.53	0.94	11.63	1.30	0.08
4	3052	27.10	0.74	0.48	0.07	0.70	0.80	0.72	1.23	0.64	2.69	38.05	4.14	0.23
5	30120	26.32	0.10	0.52	0.62	0.65	0.73	0.66	0.26	0.11	0.28	33.60	4.90	0.32

Player Comparator Tool

The player comparator tool allows users to **enter** a player name and **find** similar players. This tool can, among other things, help users **find new players** based on interesting **players they already know**. This is a more user-friendly approach than past attempts to compare soccer players, which focused on one or two stats and presented the data in static form to a user.

The **nodes depict players** while the **edges indicate similarity** between players. The **dark green** nodes depict the queried player as well as its directly similar counterparts **S**, while the **light green** nodes depict the directly similar players to **S**. Similarity was calculated using pairwise distances as informed by the clustering. Users can **hover over nodes to view player information** including name, position, and seasons played. If curious, users can then double click a node to search for similar players to that node. The tool also supports zooming, panning, and pinning nodes.

Conclusion

Through our interactions with the statistical profiles of soccer players on a season-by-season basis, we were able to draw **clear comparisons** between players not currently thought of as similar. Our emphasis on pairwise-distance similarity allowed us to compare a women's soccer star striker like Alex Morgan to a men's soccer star winger like Neymar. or players in lesser-know leagues like Korea to famous stars in Europe.

There were some **areas of improvement** for future analyses, however:

1) We were unable to use many of the attributes we scraped, such as Expected Goals, as these "newer" statistics have just recently begun to be tracked.
2) Data on passing was also sparse, and likely affected our clustering model and similarity graph's ability to accurately compare midfielder "passing maestros".
3) We noticed that our similarity graph still largely paired women together, despite womens' seasons representing only 4% of the data; it would be interesting to investigate why that was.

Ultimately, we think we were **largely successful** in creating a way for soccer fans to **invest in new leagues and players**. We hope to build on this work to raise interest in women's soccer, and make soccer in general **more accessible** to all.