ISyE 6740 – Fall 2023
Final Report

Garrett Ramos - 903848553

Finding the Next Harry Kane – Player Similarity in European Football

Problem Statement

As one of the most popular sports on Earth, soccer generates all sorts of viral, and awe-inspiring moments on the pitch for those to observe in all its glory. An estimated 1.5 billion people from all over the world watched Lionel Messi lift the World Cup trophy last year in Qatar. So, it is safe to say that the sport has the world's attention. However, the sport of soccer is more than what we see on the pitch, it is more than the amazing goals, and prized moments we see on TV. In the background, there are many people working tirelessly in order to ensure that the best players are on display for the world to watch play.

Scouts are all over the globe watching players as young as children to ensure the future of the sport is in safe hands. Technical Directors, and other analysts are working around the clock to ensure that their respective clubs' are signing the best players available, or perhaps unearthing a hidden gem. What used to be as simple as a team of scouts watching a particular player's match and using their eyes to spot the next great talent, has turned into something that clubs have invested millions of dollars every year into. Data providers are paid thousands to supply clubs with statistical profiles of players from the top leagues in Europe all the way to the lower tiers of Asia, or South America. Large teams of Analysts are employed by clubs of all sizes to evaluate this data and help supplement the eye test of talented and experienced scouts. One common task for these recruitment teams is to replace a current a player in a squad. Perhaps an important player has suffered a long term injury and will be out of action for the foreseeable future, or an aging player in the squad is in need of replacing as retirement looms near. In both situations, and others similar in nature, it is likely that the recruitment team and head coach would be interested in finding a new player that is similar in terms of playstyle and skill level to the player being replaced.

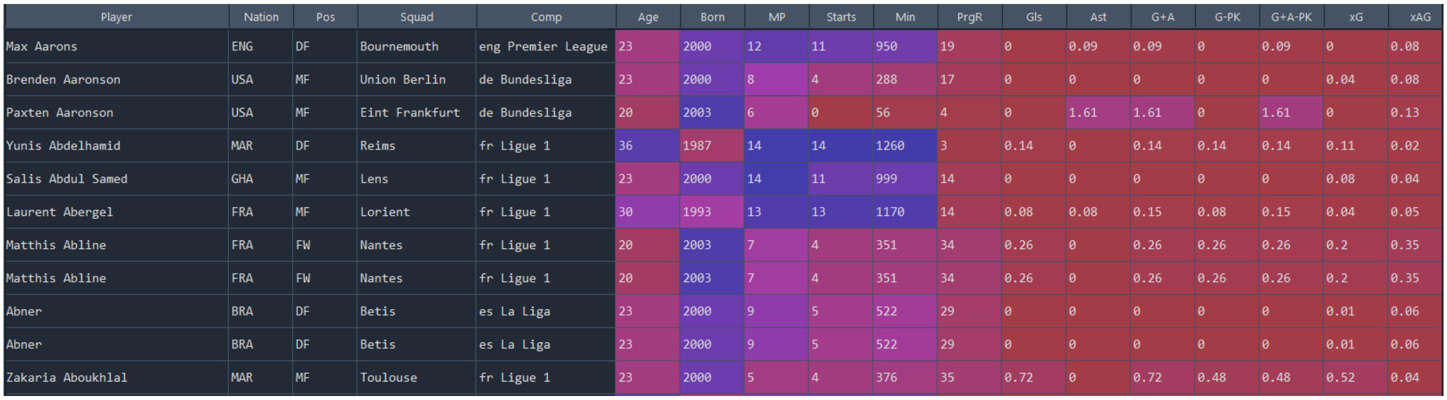
Over the summer, late into the European summer transfer window, English club Tottenham Hotspur lost arguably their greatest ever player, the club's all-time leading goal scorer Harry Kane, when he was sold to German club Bayern Munich for a reported $100 million. With so little time left in the summer window at the time of Kane's departure, Tottenham were unable to find a replacement in time before the windows' close, and were thus left with a massive void in an already thin squad heading into an always massively competitive English Premier League season.

Figure 1- Harry Kane completes his transfer to Bayern Munich

Replacing Harry Kane is not as simple as buying a player who plays the same position. There are many specific characteristics related to Kane's playstyle, and his impact on the squad which are important to quantify and use to help find a suitable replacement. As a striker, Kane's main task, as for all strikers in the sport of football is to score goals. While this is certainly the most important role Kane played in the team, what made him such a unique profile of player was his ability to not only score goals, but also create them, playing the premier roles of not just the main goal scoring threat, but also the chief playmaker. There are not many other players in the world who offer a team as much as Harry Kane provided for Tottenham, especially at the striker position. This is why a more data driven, analytical approach is needed to aid the search for the next Harry Kane at Tottenham. The goal is to use this approach to come up with a shortlist of players who could be suitable replacements in line for a potential January transfer move to Tottenham.

Data and Preprocessing

The data necessary for this endeavor was sourced from two places. Primarily, the feature data utilized was scraped from Football Reference using the Python library 'Requests'. In addition, data related to the playing position of each player in the dataset was taken from Transfermarkt. Overall, the dataset consists of 2,015 outfield players from Europe's top 5 domestic leagues including the English Premier League, France's Ligue 1, Italy's Serie A, Germany's Bundesliga, and Spain's La Liga. There are 98 unique features in the dataset, which are statistics are taken from the 2023-24 season, which began back in August. These features include all facets of the game including goal scoring, chance creation, defending, and possession.



Player	Nation	Pos	Squad	Comp	Age	Born	MP	Starts	Min	PrgR	Gls	Ast	G+A	G-PK	G+A-PK	xG	xAG
Max Aarons	ENG	DF	Bournemouth	eng Premier League	23	2000	12	11	950	19	0	0.09	0.09	0	0.09	0	0.08
Brenden Aaronson	USA	MF	Union Berlin	de Bundesliga	23	2000	8	4	288	17	0	0	0	0	0	0.04	0.08
Paxten Aaronson	USA	MF	Eint Frankfurt	de Bundesliga	20	2003	6	0	56	4	0	1.61	1.61	0	1.61	0	0.13
Yunis Abdelhamid	MAR	DF	Reims	fr Ligue 1	36	1987	14	14	1260	3	0.14	0	0.14	0.14	0.14	0.11	0.02
Salis Abdul Samed	GHA	MF	Lens	fr Ligue 1	23	2000	14	11	999	14	0	0	0	0	0	0.08	0.04
Laurent Abergel	FRA	MF	Lorient	fr Ligue 1	30	1993	13	13	1170	14	0.08	0.08	0.15	0.08	0.15	0.04	0.05
Matthis Abline	FRA	FW	Nantes	fr Ligue 1	20	2003	7	4	351	34	0.26	0	0.26	0.26	0.26	0.2	0.35
Matthis Abline	FRA	FW	Nantes	fr Ligue 1	20	2003	7	4	351	34	0.26	0	0.26	0.26	0.26	0.2	0.35
Abner	BRA	DF	Betis	es La Liga	23	2000	9	5	522	29	0	0	0	0	0	0.01	0.06
Abner	BRA	DF	Betis	es La Liga	23	2000	9	5	522	29	0	0	0	0	0	0.01	0.06
Zakaria Aboukhlal	MAR	MF	Toulouse	fr Ligue 1	23	2000	5	4	376	35	0.72	0	0.72	0.48	0.48	0.52	0.04

Figure 2- Snapshot of the dataset used including just a few of the features used

In terms of data preprocessing, significant cleaning was required in order to make the dataset suitable for model building, as well as any further analysis. The scraping process involved certain tables from Football Reference, these tables often included some of the same columns including many of the biographical or descriptive, and non-stat related columns which required removal. Additionally, many of the features were in the counting stat form, and even though all features were eventually scaled and standardized, it was preferable to have most of the stats in the per 90 minutes format as is customary to most football related analysis. Features such as tackle success, or pass completion which were in percentage form were left as such. Many of the players' names included accent marks, which needed to be dealt with in order to make it easy to search for and analyze particular players by name. For this purpose, all characters with accent marks were converted to the closest base character resemblance.

Methodology

The methodology employed involved a twofold process centered around an unsupervised learning approach. Primarily, K-Means clustering was conducted on the dataset in order to find clusters of players which were statistically similar. Second, cosine similarity was implemented on the players which were clustered together, so that an additional measure was taken to identify players most similar to one another. Cosine similarity also provided a means of which to quantifiably compare certain players to one another and detect just how similar certain players are to one another.

An important aspect of the K-Means implementation involved a great deal of data manipulation. With nearly 100 different features in the dataset, dimensionality reduction was always going to prove necessary. Depending on what position the player plays there are going to be many features which are potentially irrelevant to the predominant playstyle of said player. For example, in the case of Harry Kane, who plays as a striker, defending attributes are not going to reasonably define Kane's playstyle in any reasonable way. Features such as tackles, interceptions, or blocks are simply not important enough to certain groups of players to be included in the modeling process. In order to navigate this important aspect of the modeling process, PCA was used to represent the dataset. To do this, all 98 features were standardized and scaled, before 25 components were chosen for the PCA process. So, the dataset including 98 features was reduced to only the 25 PCA components to use in the K-Means algorithm.

In terms of the steps required to conduct the K-Means algorithm, it was important to select a value for 'k', the number of clusters to be used. The elbow plot below helps indicate which range of k-values are most optimal.

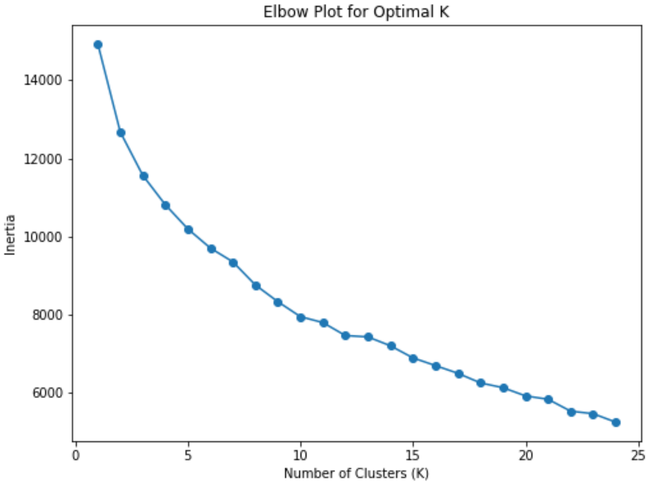


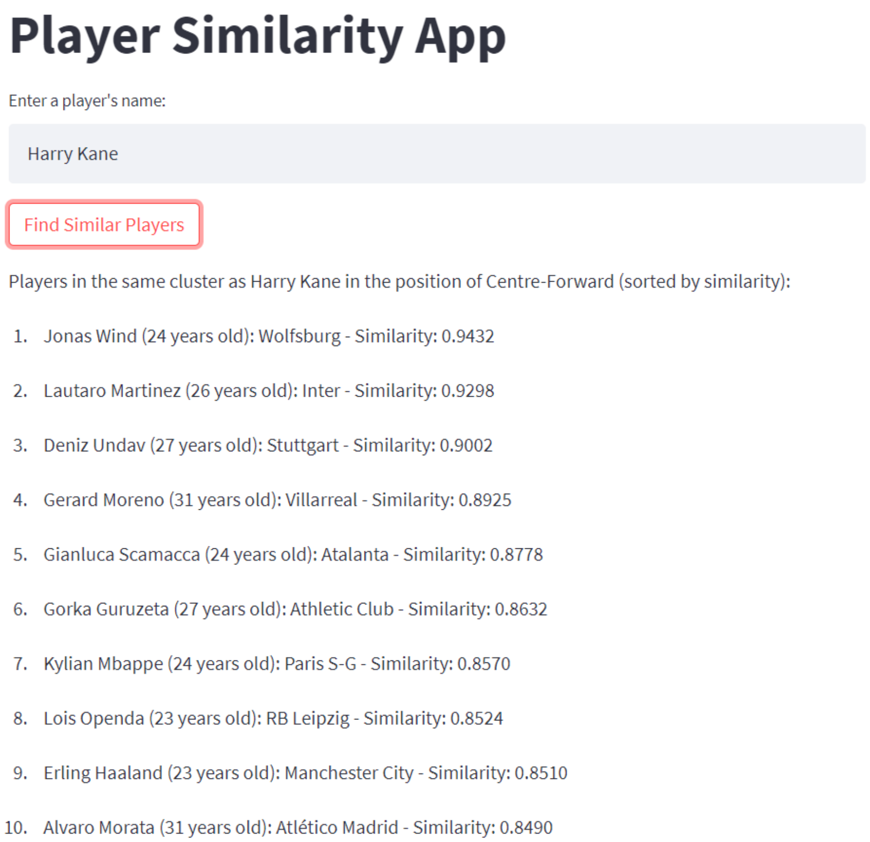
Figure 3- Elbow plot used to find optimal 'K' value

In the plot above, we are looking for the 'elbow', or in other words, the part in the graph when the sequential k-value is no longer reducing the inertia by a great deal. Based on this understanding, it was deemed that k-values from 10-15 would provide the most satisfactory results. For the analysis related to this report and specifically that of the case study involving the player Harry Kane and his potential replacements, a k-value of 15 was used.

Once clusters were formed, we wanted to focus solely on the players within the same cluster as the player of interest Harry Kane. Cosine similarity was then performed on this cluster of players, where each player in the cluster was compared to Harry Kane's statistical profile. In this case, the PCA components were once again used in place of the regular features, after observing more optimal performance. After calculating similarity measures for each of the players who shared a cluster with Harry Kane, a shortlist of ideal replacement candidates was ready to be made.

Evaluation and Final Results

K-Means clustering was conducted on the dataset limited only to the players who play in the same position as Harry Kane, which reduced the dataset, and potential replacements for Kane to 268. At the conclusion of the clustering algorithm, 14 players were found to be in the same cluster as Harry Kane. The shortlist was reduced to only the 10 most similar players sorted by their cosine similarity score. The final shortlist is as follows:



Player Similarity App

Enter a player's name:

Harry Kane

Find Similar Players

Players in the same cluster as Harry Kane in the position of Centre-Forward (sorted by similarity):

1. Jonas Wind (24 years old): Wolfsburg - Similarity: 0.9432
2. Lautaro Martinez (26 years old): Inter - Similarity: 0.9298
3. Deniz Undav (27 years old): Stuttgart - Similarity: 0.9002
4. Gerard Moreno (31 years old): Villarreal - Similarity: 0.8925
5. Gianluca Scamacca (24 years old): Atalanta - Similarity: 0.8778
6. Gorka Guruzeta (27 years old): Athletic Club - Similarity: 0.8632
7. Kylian Mbappe (24 years old): Paris S-G - Similarity: 0.8570
8. Lois Openda (23 years old): RB Leipzig - Similarity: 0.8524
9. Erling Haaland (23 years old): Manchester City - Similarity: 0.8510
10. Alvaro Morata (31 years old): Atlético Madrid - Similarity: 0.8490

Figure 4- Players most similar to Harry Kane

24 year old Danish striker Jonas Wind emerges as the most similar player profile to Harry Kane with a similarity score of 0.9432. Closely behind Wind is Argentine striker Lautaro Martinez, who interestingly enough was rumored to be moving to Spurs back in 2021, when many at the time thought Harry Kane was pushing for a move of his own to Manchester City. We can observe more closely how some of the shortlisted players compare to Harry Kane on a statistical level with the radar plots below:

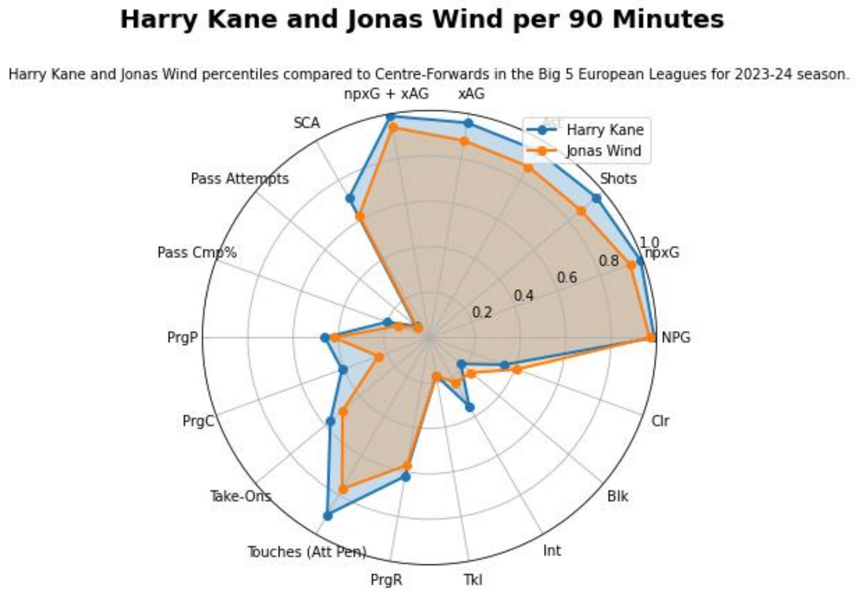


Figure 5- Harry Kane compared to Jonas Wind

Looking at the plot above, we can see how the two statistical profiles of Harry Kane and Jonas Wind closely resemble each other. Kane is slightly superior in just about every attribute highlighted in the graphic above, but Wind is not too far behind, and at 24 years old Wind has the time to further develop, and improve on his ability.

The second player in the shortlist was Lautaro Martinez. As previously mentioned, news outlets had reported just a couple years earlier of rumors involving a potential move to Tottenham Hotspur for Martinez. The analysis conducted and outline in this report may lend some insight into why the club had previously expressed interest in the player. If we look at the plot below comparing Kane and Martinez we can observe some of the key similarities. Specifically, when it comes to shooting, goal scoring, and touches in the penalty box, we can acknowledge that Martinez offers many of the same attacking qualities. However, when it comes to Kane's playmaking and chance creation, Martinez lacks in comparison.

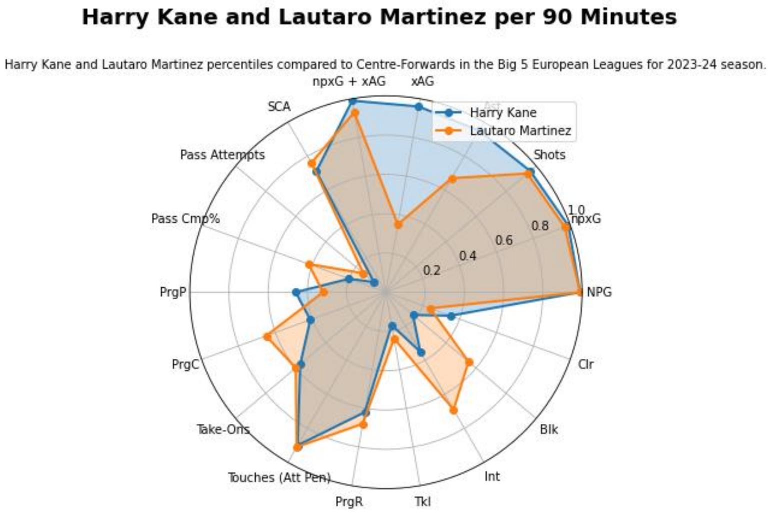


Figure 6 - Harry Kane compared to Lautaro Martinez

The last player from the shortlist of which will be observed more closely is Belgian striker Lois Openda. At 23 years old, Openda is the youngest player on the shortlist, and would fit the current trajectory of the sporting project at Tottenham who are looking to rebuild with a younger squad. Similar to Wind, and Martinez above, we can see a great deal of similarities shared between Kane and Openda, specifically in the shooting and goal scoring, and goal creating departments.

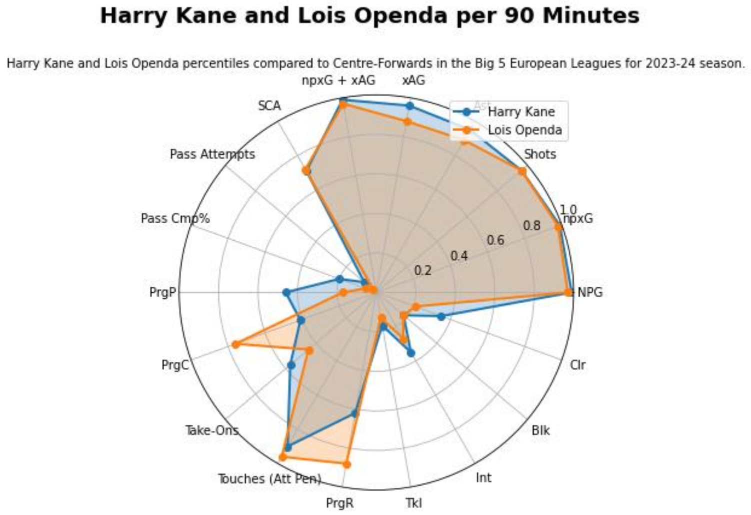


Figure 7- Harry Kane compared to Lois Openda

Overall, the results were more than satisfactory. Jonas Wind was a near identical statistical profile to Harry Kane, while Lautaro Martinez was a player that the club actually attempted to purchase less than 2 years ago. That being said, there are a few things that could have been implanted to either improve the analysis or at the very least provide more insight. Often times when searching for a new player, the most important aspect of completing the transfer is not just limited to the player's ability, but the financial logistics. A player being within the budget of the club is practically just as important as whether or not the player is good enough. Finding and implementing data associated with a players' transfer value, or current wages could have added a more realistic component to the analysis conducted in this report. Players such as Erling Haaland or Kylian Mbappe who were on the shortlist of players to replace Harry Kane, are two of the most expensive players in the world and would be simply unattainable for a club like Tottenham.

Lastly, the player pool in the dataset was limited only to players who are playing in the top 5 European leagues. There were thousands of professional players, many of which likely could have been suitable options for the shortlist, playing smaller leagues that were not represented in the dataset. Tottenham have recently signed players from Portugal, and the Netherlands, and so it would have made a lot of sense to include those leagues as well.

Citations

Football Reference Data - https://fbref.com/en/

TransferMarkt Position Data - https://github.com/griffisben/Soccer-Analyses/blob/main/TransfermarktPositions-Jase_Ziv83.csv