CLUSTERING FOR PLAYER REPLACEMENT

Rohan Sahasrabudhe
International Institute of Sports Management
Mumbai, India
E-mail: rohansbudhe@gmail.com

ABSTRACT

Analytics is a prospering segment in the sporting sector with its application across almost all departments in a sports organisation. People are surrounded with data points anywhere they go and anytime they go and there are various means to capture these from the source. This study is an attempt to construct a model for finding out the ideal football player replacement for the football players whose careers have come to the doorstep of retirement. By performing an analysis using clustering methods of K-means Clustering and Principal Component Analysis, the desired results were obtained. The analysis is done on the data regarding skills and attributes of football players from the game FIFA 20, sorted into four categories of Goalkeepers, Defenders, Midfielders and Forwards. Other aspects such as market value of player and achievements of the player are further considered to refine the list of ideal players found and zero upon one player. The results show that a clustering model can be formed and utilised to find ideal player replacements. The model is not restricted to just football but can be applied to other team sports as well. This was concluded with taking the limitations of the study into account that further study with proper access to latest and updated data of players would prove the model in an enhanced manner.

Keywords: *sports analytics; k-means; football; replacements*

INTRODUCTION

With the coming age of technological advancement, almost everyone and everything has been upgraded themselves and the resources at their disposal. This breakthrough has also been implemented in the sports industry across the world. For example, technologies of real time decision review systems, player performance analysis, match video analysis, performance predictions and so on. In recent times, Sports Analytics is a booming field. It refers to the use of data and advanced statistics to measure performance and make informed decisions in order to gain a competitive advantage (Holman, 2018). Be it team sports such as cricket, football, basketball, rugby, baseball, kabaddi or individual sports such as badminton, table tennis, wrestling, chess, swimming, track & field events, data is present everywhere in various forms. Analysts study the data and guide the coaches, players to improve their game, and management to make wise decisions to grow as an organisation. Sports analytics helps improves players' and game performance, enhances organization's business performance, analyse player health and injury probability (Holman, 2018). Almost all areas and aspects of sports and sports organisations are leveraging analytics to gain competitive advantage.

Sports has become a business avenue for brands and the brands looks for investment opportunities in the sport, team, league or even the player. The brands have discovered these as sources of earning the return on investment. Similarly, for a club or a team, investing in a player is a big decision that involves lot of aspects to be thought over critically and not emotionally. Performance of the player and ultimately the team contributes to the return on investment made by the team in the players selected or bought or loaned for the team.

Purpose of the Study

Sean Rad, founder of Tinder, once said, "Data beats emotions", and it has been rightly proved in various dimensions in the world. Sports being a data driven activity has adopted the same data driven approach over the emotional decision making. An instance to support the Sean Rad's comment would be when Louis Van Gaal was terminated by Manchester United and Jose Mourinho was appointed as the team's manager in 2016. Van Gaal led Manchester United to FA Cup. The emotion of winning such important game in front of thousands of fans could have gone long way in helping his job security, legacy or both. But to executive at Manchester United his results over two years were underwhelming, and thus, they were concerned more with data than with emotion. Manchester United chose 'data over emotion' and brought in Jose Mourinho as their manager who statistically had the best win percentage in Premier League history (Riemer, 2016). This study is an attempt to construct a model using the clustering methods for player replacement. It would help to find ideal replacements for a player becoming vulnerable due to his/her age, injury, contract expiry, performance and so on. This model could be applied to other team sports, just the attributes or variables or parameters could vary from sport to sport.

Scope of the Study

This study uses data from the football domain. Data from the game FIFA 20 is used in this study. The data in the game is based on the performance of players in the real world. Thus, this data is used to perform the analysis as at this level, the researcher is not in a position to get the actual data from various teams. Financial worth or market value of players is not considered in the model in this study as genuine and updated data was no available in the limits of the researcher. However, market value of the player is not ignored in the research as it is included outside the model along with the required qualitative aspects.

SUPERVISED AND UNSUPERVISED LEARNING

Supervised learning is done with prior knowledge of what the output values of the sample should be. It is used to obtain a best possible relationship between input and output observable in the dataset. Whereas, unsupervised learning does not have labelled outputs and thus it infers the natural structure present in the dataset (Soni, 2018). Data scientists use these methods for machine learning problems, data mining, big data processing and neural network (Educba, n.d.). In supervised learning method input and output variables are given. The goal is to determine the function such that when new input is given, its output can be predicted using the model. It is used in image and speech recognition, forecasting, financial analysis, decision trees, training neural networks etc (Educba, n.d.). For example, classification, regression and support vector machine. Supervised learning is used to plot a trendline along the data points to predict the future outcomes based on the historical data. Whereas, in unsupervised learning method only input data is given. The goal is to model the unseen patterns or structure is given input dataset in order to learn about the data. It is used in exploratory analysis, pre-train supervised learning algorithms (Educba, n.d.). For example, clustering, association and k-means. Unsupervised learning technique shows that the observations are grouped into different clusters based on the similarities in the attributes of the characters.

CLUSTER ANALYSIS

The task of dividing the data into numerous groups such that the individual data points in a group are more similar to each other than the other data points in other groups, is called clustering (Kaushik, 2016). It simply means to sort the data into groups based on the similar traits of the data points. Cluster analysis had been

applied in wide variety of fields such as statistics, information retrieval, pattern recognition machine learning, data mining, biology, psychology and other social sciences (Kumar, n.d.).

K-means Clustering

K-means clustering is the most common and simple method for cluster analysis. It splits the data into a set of k groups (University of Cincinnati, n.d.). This classification is of data into clusters requires computation of the distance or similarity or dissimilarity between the observations. Each cluster is represented by a centroid (center) corresponding to the mean of points belonging to the cluster. The basic idea behind this method is about defining clusters in a way that the total intra-cluster or within-cluster variation is minimized (Kassambara, 2017).

The researcher is required to specify the number of clusters (K) to be created. Next, objects (k) from the dataset are to selected as initial cluster means. The observations are assigned to their closest centroid depending on the distance between the object and centroid. When k-means clustering is done in R, the software does all the calculations and iterations on its own. It calculates the new mean values of all the data points in the cluster, the total sum of squares is minimized with each iteration (University of Cincinnati, n.d.). Listed below are the advantages and disadvantages of k-means clustering (Google, n.d.).

Advantages:

- Relatively simple to implement
- Scales to large data sets
- Guarantees convergence
- Can warm-start the positions of centroids
- Easily adapt to new examples
- Generalizes to clusters of different shapes and sizes

Disadvantages:

- Choosing k manually
- Being dependent on initial values
- Clustering data of varying sizes and density
- Clustering outliers
- Scaling with number of dimensions

FIFA 20 PLAYER RATING

The Fédération Internationale de Football Association (FIFA), the governing body of football in the world, along with EA Sports launch a video game called FIFA every year. The game's aesthetics are very accurate to footballing world in real life. The appearance of players, infrastructure, commentary, player transfers, graphics of the game, playing style of players, performance ratings and many more such innovative and creative aspects that match the real-life situation. Ratings given to the performance of the players is based on the players' skills, attributes and performance in real-life. How are the ratings decided? EA Sports employs a team of producers (25) and external data contributors (400) led by the Head of Data Collection & Licensing; this team's responsibility is to ensure that data of all the players is up to date. Suggestions and alterations to the database are constantly provided by another team or community of 6000 FIFA Date Reviewers got Talent Scouts (Murphy, 2019). The process is quite complicated but has vast scope. It ensures that the information and data in the game is as accurate as possible across leagues, teams, players from all levels.

METHODOLOGY

The objective here is to construct a clustering model for finding out ideal replacement for an existing player in a team sport. Existing players in a team could become vulnerable due to his/her age, injury, contract expiry, performance and many other reasons. This model would help the coaches, support staff and the management to find out ideal replacement options for that particular player.

Significance of the Study

This model if successful can prove vital for decision making for sporting teams. Selection of players is an essential task in order to build a good team. As seen in the movie (based on a true story) "Moneyball", the baseball team general manager challenged the system and defied conventional wisdom when he was forced to rebuild his small-market team on a limited budget. Despite criticism, he developed a roster of misfits with the help of a young economist using and analyzing the historical data of various players and changed the way the game was played (Miller, 2011). In the move the team does not win the league but improve their performance significantly. Thus, with the coming age of numeric data and visual content, scouting players has become easier, time-saving and inexpensive.

Data Collection

The researcher uses secondary data for this study. FIFA 20 video game data has been used for this model. This data was obtained from an online source called Kaggle (www.kaggle.com). The data was uploaded October 2019 by Stefano Leone and was accessed and retrieved by the researcher on 12th December 2019 (Leone, 2019). The uploaded files contain player datasets of FIFA video game from the year 2015 to 2020. But this study only uses the complete player dataset FIFA 20.

Selection of sample

The dataset consists details of 18,278 players registered with FIFA and are a part of the FIFA 20 video game. The dataset shows numerous attributes of these players that are given points based on their performance in the real world. These attributes are of different types: demographic attributes, financial attributes, physical attributes, skill-based and performance-based attributes. The players are divided into four major segments based on their playing position: Goalkeepers, Defenders, Midfielders and Forwards. Within these segments, further clusters are formed that are used to achieve the desired objective. The list of attributes of is given further.

Methods of data analysis

Modelling and analysis in this study is a two-step process. First step consists of cluster analysis in which the players are grouped into clusters and based on those results, a list of ideal players for replacements would be made. Second step is concerned with a qualitative aspect wherein players' background would be considered to further narrow down the list of desired layers. Figure 1 show a list of all the attributes that are used in the K-means clustering model. These attributes demographic, physical, and financial aspects related to the football players. Categorical variables (attributes) such as nationality, club, player positions, and preferred foot were used in K-means clustering model, as it is one of the limitations of the method. Market value, wage, release clause was also not included in the model because they do not affect the performance of the player, but performance of the player defines them; also, some of the values were found to be missing.

Figure 1: List of attributes used in the clustering model

short_name	overall	attacking_crossing	movement_acceleration	mentality_aggression	gk_diving
age	potential	attacking_finishing	movement_sprint_speed	mentality_interceptions	gk_handling
dob	pace	attacking_heading_accuracy	movement_agility	mentality_positioning	gk_kicking
height_cm	shooting	attacking_short_passing	movement_reactions	mentality_vision	gk_reflexes
weight_kg	passing	attacking_volleys	movement_balance	mentality_penalties	gk_speed
nationality	dribbling	skill_dribbling	power_shot_power	mentality_composure	gk_positioning
club	defending	skill_curve	power_jumping	defending_marking	goalkeeping_diving
player_positions	physic	skill_fk_accuracy	power_stamina	defending_standing_tackle	goalkeeping_handling
preferred_foot	value_eur	skill_long_passing	power_strength	defending_sliding_tackle	goalkeeping_kicking
contract_valid_until	wage_eur	skill_ball_control	power_long_shots		goalkeeping_positioning
	release_clause_eur				goalkeeping_reflexes

1) K-means Clustering

The cluster analysis is done using R. Data was scaled (standardized) and centered, before performing the K-means Clustering. Optimal number of clusters to be created would be found and then the k-means clustering would be performed. Once the clusters are obtained, a cluster plot would be made accommodating all the players in the dataset in the said number of clusters. The method requires the researcher to mention the number of clusters to be generated. The optimal number of clusters is found out using the fviz_nbclust() function in R (in factoextra package). The graph shows the variance within the clusters. As number of clusters (k) increases, the variance within the cluster decreases (Kassambara, 2017). It is represented by a bend (or elbow) in the graph which shows that the additional clusters beyond the bend point have little value.

2) Subset Analysis

Once the players are categorised into their respective clusters, a table would be made with other categorical information such as age, playing positions, preferred leg and contract validity. These will help us to further narrow down the list to get the desired result of ideal players for replacement. The newly made table was then exported from R to MS-Excel. Filters were applied in order to arrive at the desired subset in each segment of players, i.e. Goalkeepers, Defenders, Midfielders and Forwards.

3) Qualitative Analysis

After finding the ideal players, a background check could be done to find out the past achievements, controversies (if any), social conduct, goodwill gestures of the players in order to develop personal image along with the professional aspects. The league in which the player presently plays also matter a lot as it determines the competition level each player is exposed to. Number of years played and experience gained by a player impacts the profile of a player in a big way.

ANALYSIS OF FINDINGS

K-means Clustering was done on R. The attributes for all four segments of players and the R codes used for the analysis are available in the appendices. Finding out the list of ideal players for replacement was done on MS-Excel with use of filters. Other qualitative data was obtained from various other sources cited in this study.

Goalkeepers

Figure 2 indicates that the optimal number of clusters to be formed for Goalkeepers is five (5). So, the researcher ran the k-means clustering function with k=5 on the Goalkeepers dataset. Figure 3 represents the cluster plot of Goalkeepers. The cluster numbers were then paired with the individual players along with other categorical factors, as one can see for the example in in Table 1.

There are 2,036 Goalkeepers in the dataset that are divided in 5 cluster as shown above in Figure 3. There are 380 goalkeepers belonging to cluster 1, 343 goalkeepers in cluster 2, 389 goalkeepers in cluster 3, 572 goalkeepers in cluster 4 and 352 goalkeepers in cluster 5. Example: Miguel Angel Moya, a Spanish right-footed goalkeeper who plays for Real Sociedad, is 35 years old and with a contract expiration in 2020. Ideal replacements for him from the k-means clustering model would be the following players listed in Table 1.

Figure 2: Optimal number of clusters for Goalkeepers

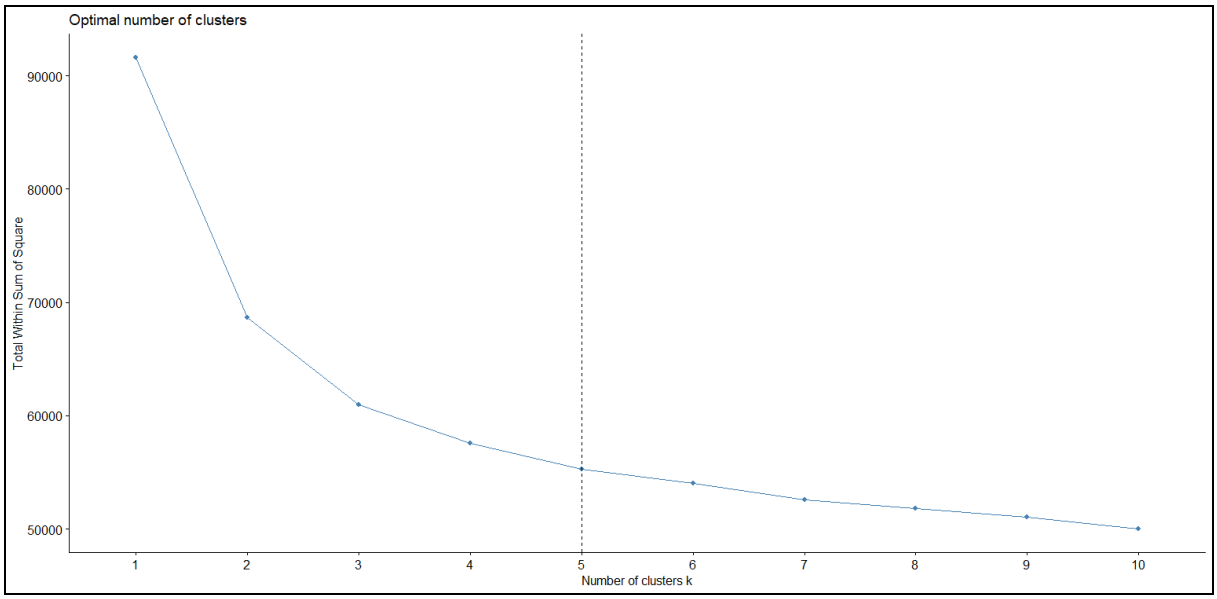


Figure 3: Cluster plot of Goalkeepers

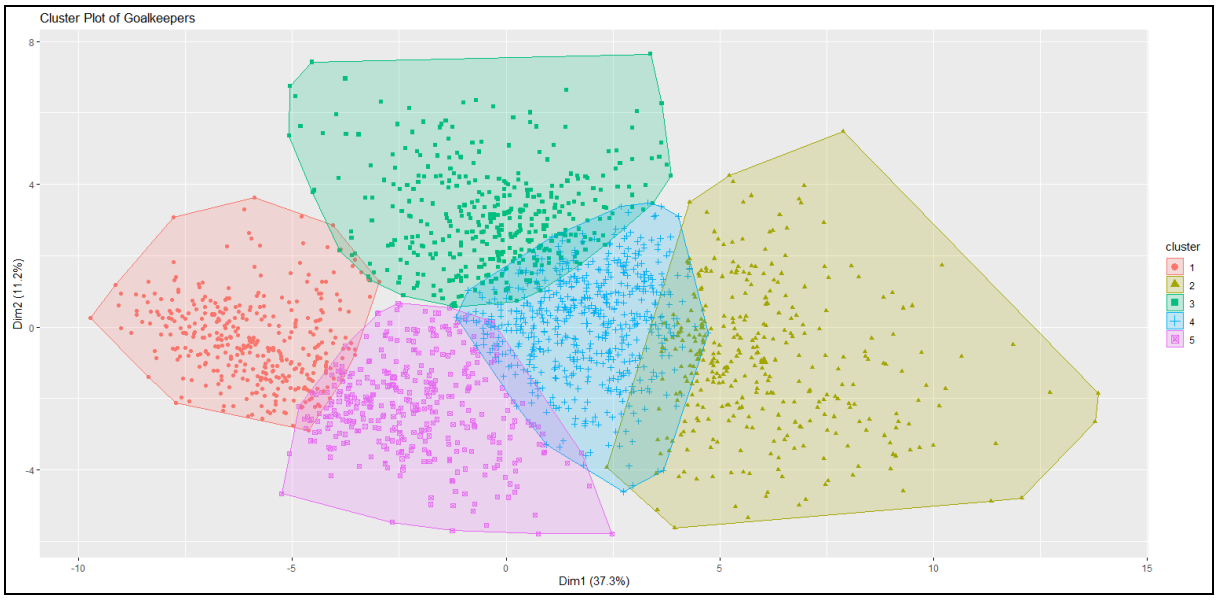


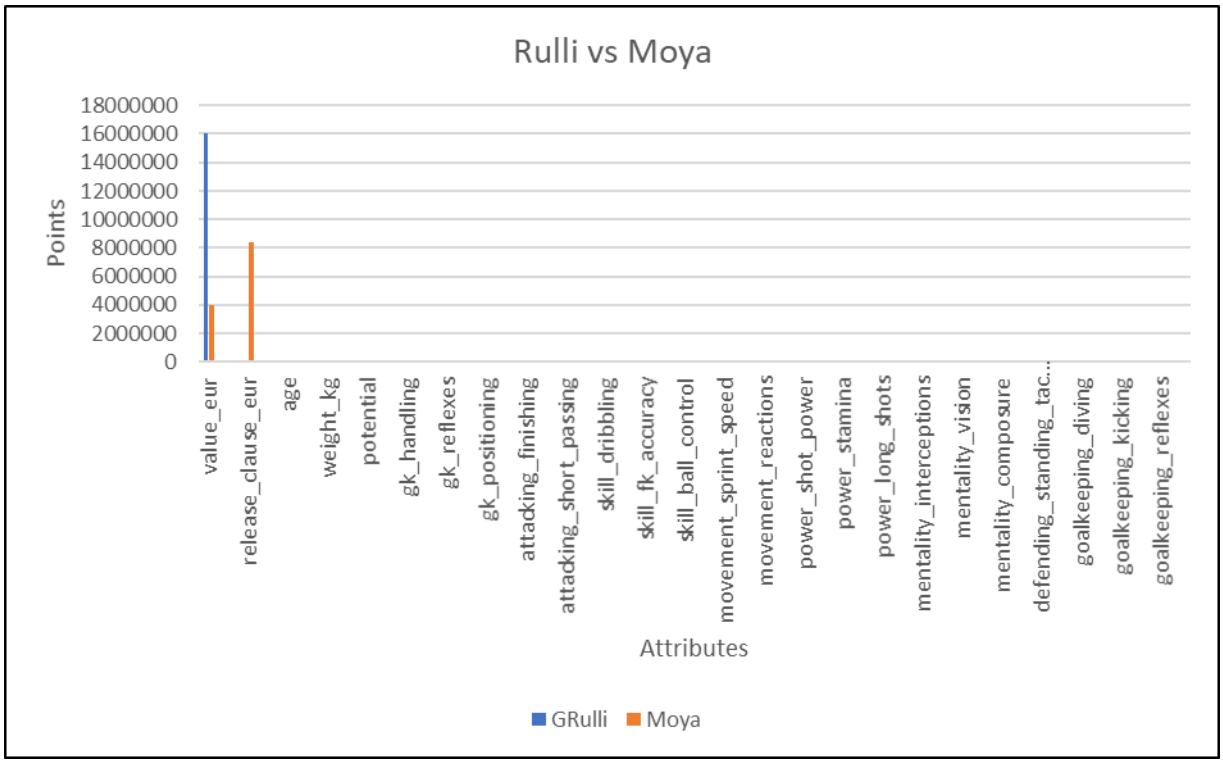
Table 1: List of ideal player replacements for Miguel Angel Moya present in Goalkeepers Cluster 2

Name	*Nationality*	*Club*	*Age*	*Position*	*Preferred foot*	*Contract validity*	*Cluster*
Moya	Spain	Real Sociedad	35	GK	Right	2020	2
DeGea	Spain	Manchester United	28	GK	Right	2020	2

GRulli	Argentina	Montpellier HSC	27	GK	Right	2020	2
WBenÃtez	Argentina	OGC Nice	26	GK	Right	2020	2
LKarius	Germany	BeÅŸiktaÅŸ JK	26	GK	Right	2020	2
ASchwolow	Germany	SC Freiburg	27	GK	Right	2020	2
ClÃ¡udioRamos	Portugal	CD Tondela	27	GK	Right	2020	2
MSportiello	Italy	Atalanta	27	GK	Right	2020	2
ALunin	Ukraine	Real Valladolid CF	20	GK	Right	2020	2

From the above-mentioned players, the background for these players would be checked for After finding the ideal players, a background check could be done to find out the past achievements, controversies (if any), social conduct, goodwill gestures of the players in order to develop personal image along with the professional aspects. For example, Geronimo Rulli, an Argentinian right-footed goalkeeper who plays for Montpellier (on loan from Real Sociedad) in French Ligue 1, is 27 years old and with a contract expiration in 2020 could be an ideal replacement for Miguel Angel Moya. He has played for Spanish club Real Sociedad previously and the club can call him back from Montpellier to play in La Liga.

Figure 4: Geronimo Rulli vs Miguel Angel Moya - Attributes Comparison

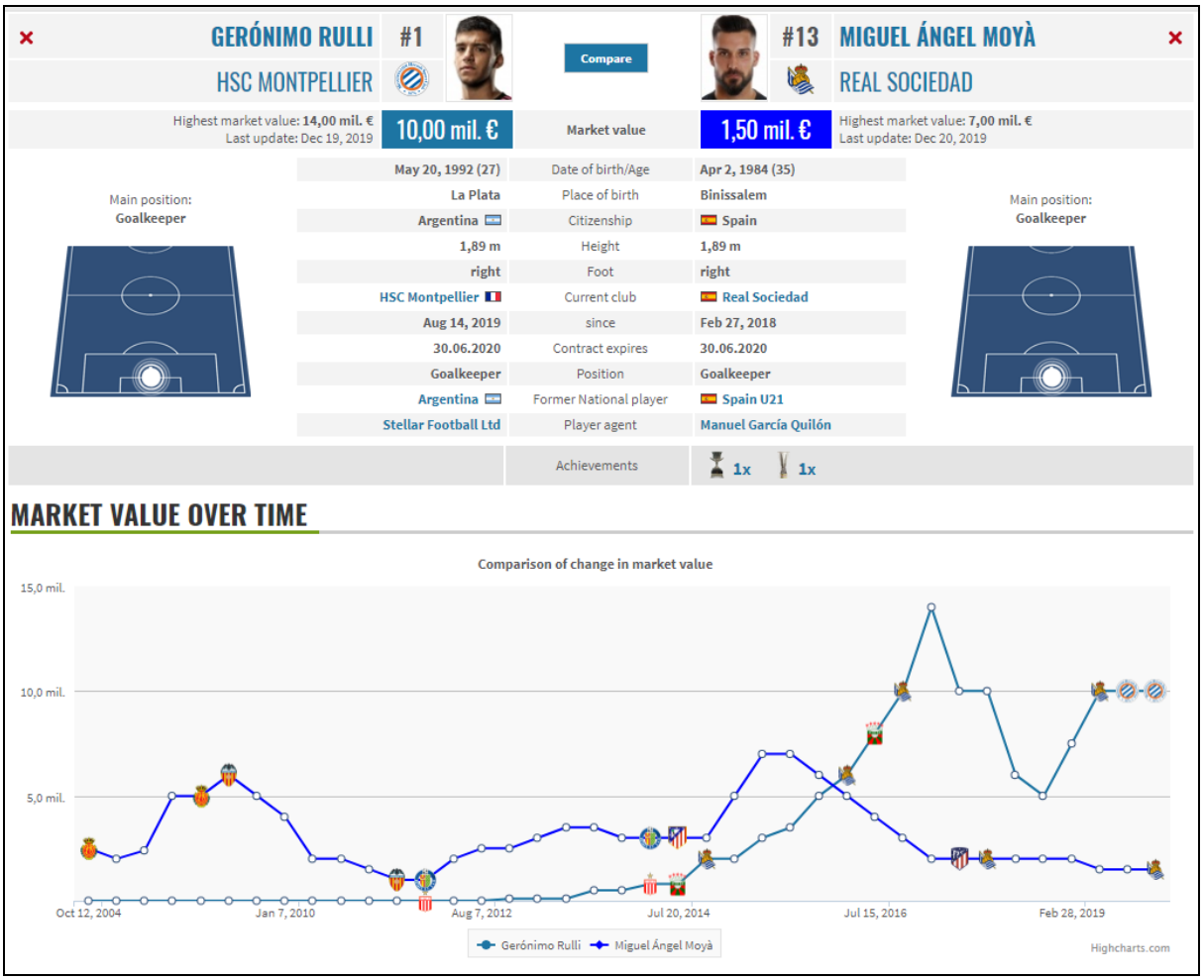


As one can see from Figure 4, data shows that Rulli and Moya are quite similar based on the points of their attributes. There is hardly a difference of one-two points in their attributes as a Goalkeeper. The difference that exists is due to the experience at the level where they played and the number of years played by the players. In some attributes, Rulli is better than Moya. So as per this data, one can say that Rulli can be considered as an ideal replacement for Moya.

Selection of a player is not done just on the basis of the points of his or her attributes. There are several other factors considered before making a final choice. There is a financial aspect to the selection which is

represented by the player's market value. It can be seen in Figure 5 that Geronimo Rulli has higher market value than Miguel Angel Moya. Having a high market value signifies that the player has been performing well in the matches. Rulli hasn't won any trophies in his career, whereas Moya has won 2. Even though Rulli has not been able to win any trophies yet, his performance (indicated by market value in Figure 5) has improved over the years and has become better than Moya.

Figure 5: Geronimo Rulli vs Miguel Angel Moya (Transfer Market, 2019)



Moya has played in the Spanish La Liga for all his career years. Rulli also has an experience of 5 seasons in the Spanish league, but currently plays in French Ligue 1 league which, in comparison with Spanish league, is similar in competitiveness. Considering his decent performance in Spanish and French leagues, Rulli could be considered a replacement for Moya.

Defenders

There are 5938 Defenders in the dataset that are divided in 5 cluster. There are 1394 defenders belonging to cluster 1, 981 defenders in cluster 2, 1603 defenders in cluster 3, 805 defenders in cluster 4 and 1155 defenders in cluster 5. Example: Thiago Silva, a Brazilian defender (right-footed centre back) who plays for Paris Saint-Germain, is 35 years old and with a contract expiration in 2020. Ideal replacements for him from the k-means clustering model would be the following players listed in Table 2.

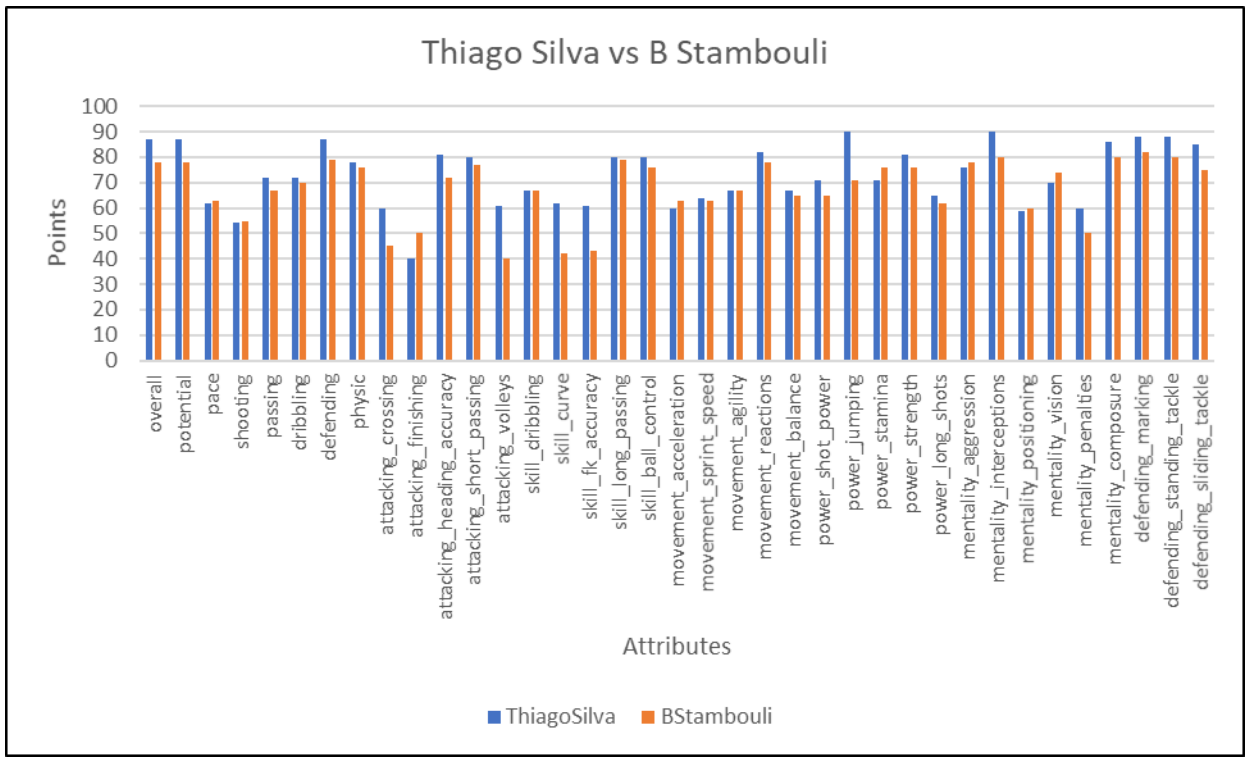
For the below-mentioned players, subset and qualitative analysis was done. For example, Benjamin Stambouli, a French defender (right-footed centre back) who plays for FC Schalke 04 in Bundesliga, is 29

years old and with a contract expiration in 2020 could be an ideal replacement for Thiago Silva. He has played for Paris Saint-Germain previously and the club can take him back from FC Schalke 04 to play in French Ligue 1.

Table 2: List of ideal player replacements for Thiago Silva present in Defenders Cluster 5

Name	*Nationality*	*Club*	*Age*	*Position*	*Preferred foot*	*Contract validity*	*Cluster*
ThiagoSilva	Brazil	Paris Saint-Germain	35	CB	Right	2020	5
BStambouli	France	FC Schalke 04	29	CB, CDM	Right	2020	5
KThÃ©ophile-Catherine	France	Dinamo Zagreb	29	CB, RB	Right	2020	5
BKamara	France	Olympique de Marseille	19	CB	Right	2020	5
OLewicki	Sweden	MalmÃ¶ FF	26	CB, CDM, CM	Right	2020	5
Nolaskoain	Spain	Deportivo de La CoruÃ±a	20	CB, CDM, CAM	Right	2020	5
PFranco	Colombia	AmÃ©rica de Cali	28	CB, CDM	Right	2020	5

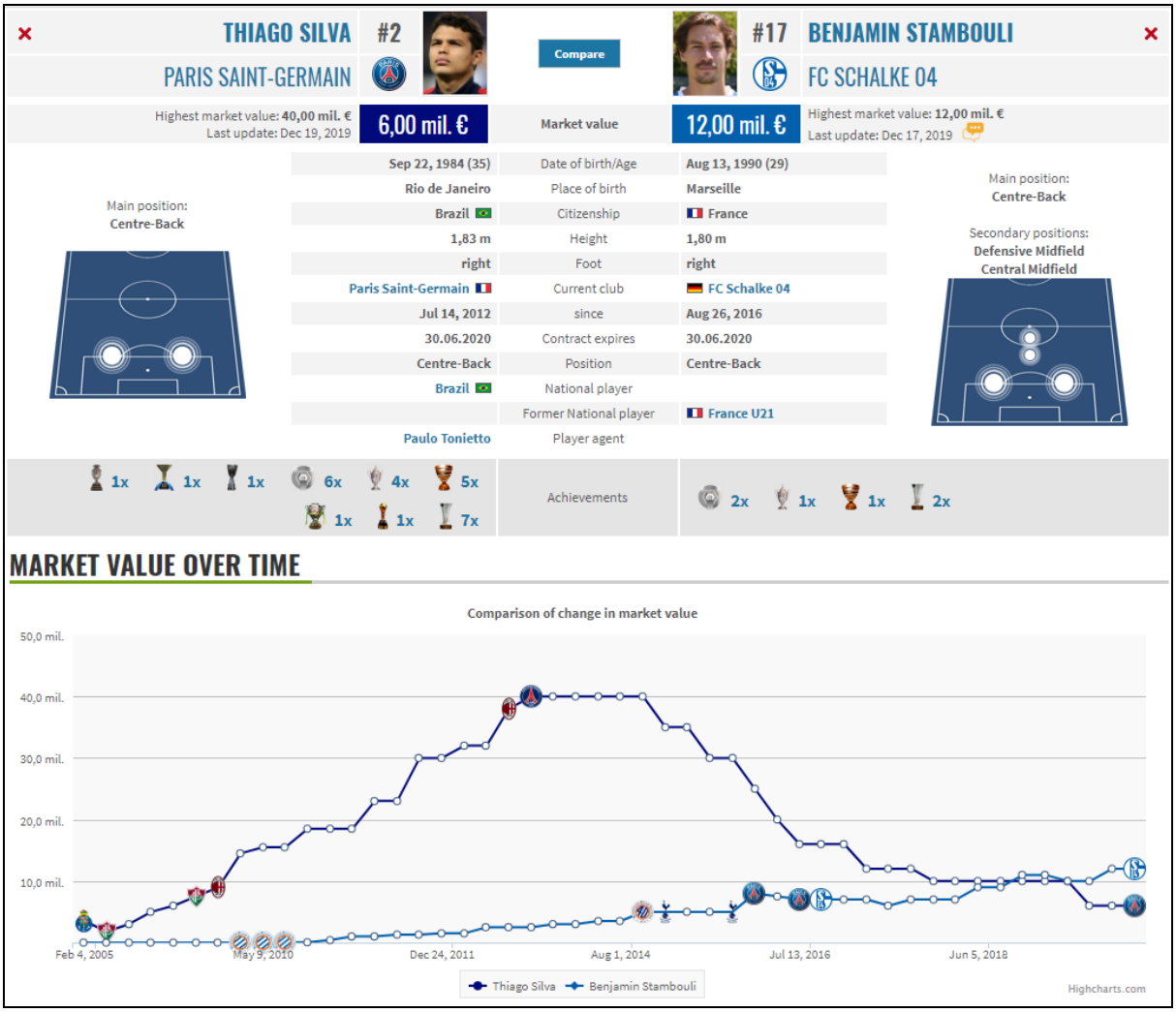
Figure 6: Thiago Silva vs Benjamin Stambouli - Attributes Comparison



As one can see from Figure 6, data shows that Silva and Stambouli are quite similar based on the points of their attributes. There are some differences among few attributes as a Defender. The difference that exists is due to the experience at the level where they played and the number of years played by the players. In some attributes, Stambouli is better than Silva, but one can say that it could vary because of the age of the player. So as per this data, one can say that Stambouli can be considered as an ideal replacement for Silva. Selection

of a player is not done just on the basis of the points of his or her attributes. There are several other factors considered before making a final choice. There is a financial aspect to the selection which is represented by the player's market value. It can be seen in Figure 7 that Benjamin Stambouli has higher market value than Thiago Silva. Having a high market value signifies that the player has been performing well in the matches. The trend line in Figure 7 does not show much gap between the market value of both the players. Performance wise they both are quite at the same level. Stambouli has a little higher market value than Silva, and he also brings an additional advantage that he can also play as a centre defensive midfielder if need be.

Figure 7: Thiago Silva vs Benjamin Stambouli (Transfer Market, 2019)



Silva has been playing in the French Ligue 1 for almost 8 seasons now. Prior to his he has played in Italian and Brazilian League. Stambouli has been playing in the German Bundesliga for about 4 seasons now. Thiago Silva has won 27 trophies throughout his career, which signifies the quality and value he adds to the team. Stambouli, on the other hand, has won 6 trophies so far. Number of years Silva has played for and that too at very top competitive level, and yet winning so many trophies, this experience brings a for the team is unmatched. This is a massive difference among the two players and stands as an obstacle for selecting Stambouli as a replacement for Silva.

Midfielders

There are 6862 Midfielders in the dataset that are divided in 5 clusters. There are 944 midfielders belonging to cluster 1, 1719 midfielders in cluster 2, 1340 midfielders in cluster 3, 1682 midfielders in cluster 4 and 1177 midfielders in cluster 5. Example: Fernandinho, a Brazillian midfielder (right-footed centre defensive midfielder) who plays for Manchester City, is 34 years old and with a contract expiration in 2020. Ideal replacements for him from the k-means clustering model were listed and analysed. Sebastian Rudy, a German midfielder (right-footed centre defensive midfielder) who plays for TSG 1899 Hoffenheim (on loan from FC Schalke 04) in Bundesliga, is 29 years old and with a contract expiration in 2020 could be an ideal replacement for Fernandinho. He has played for Bayern München and FC Schalke 04 previously and Manchester City can take him from TSG 1899 Hoffenheim 04 to play in Premier League.

Data showed that Fernandinho and Rudy are quite similar based on the points of their attributes. There are some differences among few attributes as a Midfielder. The difference that exists is due to the experience at the level where they played and the number of years played by the players. In some attributes, Stambouli is better than Silva, but one can say that it could vary because of the age of the player. So as per this data, one can say that Rudy can be considered as an ideal replacement for Fernandinho. Selection of a player is not done just on the basis of the points of his or her attributes. There are several other factors considered before making a final choice. Fernandinho has higher market value than Sebastian Rudy, although the difference is not very much. Since 2015, Fernandinho's market value has been falling and during that time Rudy's market value was increasing. At a point in 2018, they both were almost at the same level. Presently Rudy has a little lower market value, which means that he would cost less to Manchester City in terms of wages. Rudy can also play as a right back if need arises.

Fernandinho has been playing in the Premier League for about 7 seasons now. Rudy has been playing in the Bundesliga since 2017, but in different teams. English Premier League is considered to be the most competitive football league across the world. Fernandinho has been playing in a top English club as a starting-11 player and has contributed to achieve many titles during these years. Performance-wise Fernandinho has been exceptional and has won 28 trophies in his career, which signifies the quality and value he adds to the team. Rudy has won 4 trophies so far. Number of years Fernandinho has played for and that too at very top competitive level, and yet winning so many trophies, this experience brings a for the team is unmatched. This difference among the two players stands as an obstacle for selecting Rudy as a replacement for Fernandinho.

Forwards

There are 3442 Forwards in the dataset that are divided in 5 clusters. There are 767 forwards belonging to cluster 1, 664 forwards in cluster 2, 428 forwards in cluster 3, 686 forwards in cluster 4 and 897 forwards in cluster 5. Example: Aritz Aduriz, a Spanish forward (right-footed striker) who plays for Athletic Club de Bilbao, is 38 years old and with a contract expiration in 2020. Ideal replacements for him from the k-means clustering model were listed and analysed. Enes Ünal, a Turkish Forward (right-footed striker) who plays for Real Valladolid CF (on loan from Villarreal) in La Liga, is 22 years old and with a contract expiration in 2020 could be an ideal replacement for Aritz Aduriz. He has played for Levante and Villarreal previously and been a member of Manchester City Reserves and Athletic Club de Bilbao can take him from Real Valladolid CF to continue playing in LA Liga.

Data showed that Aduriz and Ünal are quite similar based on the points of their attributes. There are some differences among few attributes as a Forward, in some cases Aduriz is better and in some Ünal is better. So as per this data, one can say that Unal can be considered as an ideal replacement for Aduriz. Selection of a

player is not done just on the basis of the points of his or her attributes. There are several other factors considered before making a final choice. There is a financial aspect to the selection which is represented by the player's market value. Enes Ünal is 17 years younger than Aritz Aduriz and has a higher market value. Aduriz has won just 3 trophies in his long career of 14 years. Ünal being a young player, has better performance as compared to Aduriz and has so much amount of time in front of him to perform and win trophies. Moreover, Ünal is already playing in La Liga (the Spanish League) and also has some experience of training with Premier League giants Manchester City, thus, possessing a mix of skills from English and Spanish football.

Aduriz has been playing in the Spanish La Liga for all his life. Ünal has played majority of football of his career in Spain and still plays in the Spanish league. They both are exposed to same competition level. Considering his decent performance in the league, Ünal could be considered a replacement for Aduriz.

CONCLUSION

From this study one is able to find replacements for players using the statistical methods of cluster analysis. The two main methods used in the study were Principal Component Analysis and K-means Clustering that showed the required results. The researcher was able to give a list of ideal replacements for a particular player based on the clusters formed. From the examples shown in previous sections, one can understand that just cluster analysis is not enough for the player selection; there are certain qualitative aspects that are needed to be considered.

In the examples above, it was seen that among Goalkeepers, Rulli could be selected as a replacement for Moya. Among Defenders, Stambouli was shown ideal as per the personal attributes but Thiago Silva has an upper hand over Stambouli in terms of experience, hence, Stambouli could not replace Silva. Among Midfielders, again Fernandinho has an advantage over Rudy due to experience, hence, Rudy not fit to replace Fernandinho. Among Forwards, Ünal could be given a chance to replace Aduriz as there is not much difference between them on qualitative aspect as well.

This study covers personal attributes and skills of the players, their market value, and some qualitative aspects such as achievements of the player and competitiveness of the league. But another significant aspect from the financial point of view remains, which is not covered in this study. Expenditures to replace a player are bound by a particular budget. The budgetary constraints could not be obtained for this research otherwise it would have made the model better.

Limitations

The researcher did not have access to the latest data of players, as the ratings keep on changing continuously, depending on the performance of the players. In this research, market value is used as an indicator of performance of the players, which is true in a sense. But having higher market value means higher outlays for procuring the player with high market value. This study and the examples used within do not touch upon the spending capacities of the teams or clubs, since the data was not in limits of the researcher. As the data set contained over 18,000 players, some players were plotted at the intersection of two clusters, suggesting the players could belong to either of the clusters.

Recommendations

League specific clustering model can be formed to refine the results. This the number of players would reduce from over 18,000 to somewhere around 800 to 1,000. This will help in identifying the potential replacement within the league in a much better way. Further, some specific attributes could be identified that have the most impact on player performance and then a model could be developed using only those attributes. This could help to overcome the limitation players lying at the intersection of two clusters, suggesting the players could belong to either of the clusters.

Scope for future research

The previously mentioned cases of players and their replacement are just examples of what could be done. With much more accurate data about budget allocation of teams and clubs, market values of players, past wages and other financial data, one can analyse and find out ideal replacement for a particular that matches the required skill set and well within the spending limits of the team. Transfer spending of clubs and teams will make an important factor for determining the results of such a model for player replacement. A similar model could be developed for coaches and managers in order to get the best performing coach or manager in your club or team. A comparative structure could be developed using the model to show the success rate of the coach or manager with other points or scores for qualities, skills and other attributes. Such models can be designed for other team sports as well. These models are widely used in Baseball, American Football, Basketball in the USA and some European countries. In India, such model could be implemented to sorts such as Football, Hockey, Kabaddi, Volleyball, Cricket etc. These models could be used by teams in player auctions for Indian Premier League, Pro Kabaddi League, Premier Badminton League etc. in India along with budgetary allocations and spending limits.

ACKNOWLEDGMENT

I would like to express my gratitude and sincere thanks to my Professor Nicole D'Silva, International Institute of Sports Management, for instilling confidence in me to carry out this study and extending valuable guidance and encouragement from time to time, without which it would not have been possible to undertake and complete this project. I would like to thank my colleagues, friends and parents for their valuable comments and suggestions for making this a cherishable experience for me.

REFERENCES

1. Educba. (n.d.). Supervised Learning vs Unsupervised Learning. Retrieved February 09, 2020, from EDUCBA: https://www.educba.com/supervised-learning-vs-unsupervised-learning/
2. Google. (n.d.). k-Means Advantages and Disadvantages. Retrieved February 23, 2020, from Google Developers: https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages
3. Holman, V. (2018, November 15). What is Sports Analytics? Retrieved January 28, 2020, from Agile Sports Analytics: https://www.agilesportsanalytics.com/what-is-sports-analytics/
4. Kassambara, A. (2017). Practical Guide to Cluster Analysis in R - Unsupervised Machine Learning (1 ed.). Statistical Tools for High-throughput Data Analysis (STHDA). Retrieved January 10, 2020, from http://www.sthda.com/english/articles/25-clusteranalysis-in-r-practical-guide/

5. Kaushik, S. (2016, November 3). An Introduction to Clustering and different methods of clustering. Retrieved January 13, 2020, from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/
6. Kumar. (n.d.). Cluster Analysis: Basic Concepts and Algorithms. In Kumar. Retrieved January 13, 2020
7. Leone, S. (2019, September 27). FIFA 20 complete player dataset. Retrieved December 12, 2019, from Kaggle: https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset
8. Miller, B. (Director). (2011). Moneyball [Motion Picture]. USA: Columbia Pictures. Retrieved December 17, 2019, from https://www.sonypictures.com/movies/moneyball
9. Murphy, R. (2019, September 12). FIFA player ratings explained: How are the card number & stats decided? Retrieved February 22, 2020, from Goal: https://www.goal.com/en-ae/news/fifa-player-ratings-explained-how-are-the-card-number-stats/1hszd2fgr7wgf1n2b2yjdpgynu
10. Riemer, Y. (2016, May 22). Congrats, You're Fired. Retrieved February 20, 2020, from Yair Riemer On Startups, Entrepreneurship, and Marketing: http://yairriemer.com/tag/data-beats-emotion/
11. Soni, D. (2018, March 22). Supervised vs. Unsupervised Learning. Retrieved February 09, 2020, from Towards Data Science: https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d
12. Transfer Market. (2019, December). Player Comparison. Retrieved February 23, 2020, from Transfer Market: https://www.transfermarkt.co.in/vergleich/spielervergleich/statistik
13. University of Cincinnati. (n.d.). K-means Cluster Analysis. Retrieved January 08, 2020, from UC Business Analytics R Programming Guide: https://uc-r.github.io/kmeans_clustering
14. https://www.indiatoday.in/sports/football/story/manchester-united-sack-louis-van-gaal-jose-mourinho-to-take-over-325058-2016-05-23
15. https://blockgeni.com/machine-learning-creating-a-similarity-measure/