MŰEGYETEM 1782

Budapest University of Technology and Economics
Faculty of Electrical Engineering and Informatics
Department of Telecommunications and Media Informatics

Career handbook for professional soccer players

Balázs Ács, Roland Kovács

CONSULTANT
Dr. László Toka

BUDAPEST, 2021

Content

1 Introduction

In the early 2000s sports analysts had a hard time, because of the limited data available for this purpose. Fortunately, with the growing importance of the statistical approach to achieve competitive advantage, the quantity and quality of sport-related data has also increased. First baseball, basketball, hockey, and American football enjoyed the benefits of this approach, but now with the explosive growth of available soccer data, there are plenty of areas where a solution or tool can be developed, which can lead to a competitive advantage for both teams and players.

During this research our goal was to create the basics of a tool, a so-called handbook, which can identify the optimal career and development plan of a player to achieve their desired goals. For example, which is the optimal career path to play in one of the top leagues of the world, or what path should the player follow to be one of the most expensive players in the world? To recognize these patterns, the usage of time series data is essential, so we decided to collect 15 years of data from different players and teams from all over the world. We would like to generalize this decision supporter solution, so we did not just consider cutting-edge teams and leagues. The data we were able to find and collect provided us insights of teams and players not only from the top leagues, but from the less developed countries and lower tier leagues as well. In these countries the international appearance of the clubs is negligible or minimal so they can provide us a higher view during the modeling and evaluation phase.

In this part of the project, we focused on creating the basis of this solution. After we gathered and cleaned the required data, we dealt with the soccer market value inflation, which we will detail later. Then, we set up an acquaintance network among the footballers based on if they have ever been teammates. Network science has become widespread in recent years, allowing us to explore more and more networks. By examining complex networks, we can obtain information that would not otherwise be possible and that can have a massive effect on the examined theories. Using the collected data and the data excluded from the networks we analyzed different scenarios, that can be the target of many professional footballers. After developing this tool, we want to answer questions like what skills should the young players develop in order to play in certain leagues, or what are the most vital attributes of the players that managed to get into the team of the season or achieve a desired market value?

During this research, we examined professional footballers' different properties and looked for the main differences between the stand-out players and the rest of the players. Finding these differences could highlight those attributes, skills that should be developed with greater emphasis to become world class players. Identifying successful professional career path patterns is important for soccer players, their managers, and clubs, in order to make optimal career decisions, club selection and lucrative player transfers, respectively. Specifically for players who aspire for excellence during their careers ahead of them, it is of paramount importance to make reasonable choices in decision situations, e.g., focusing on improving specific skills, selecting their next club. In this paper, we seek the knowledge of what made the greatest soccer players in terms of those decisions. We implemented and tested our model on four selected career goals: what attributes are necessary to improve to achieve a market value over €100M, to play in the English Premier League, to get into the Team of the Season (TOTS) and to increase the most important skills over 80 (on a 1-99 scale).

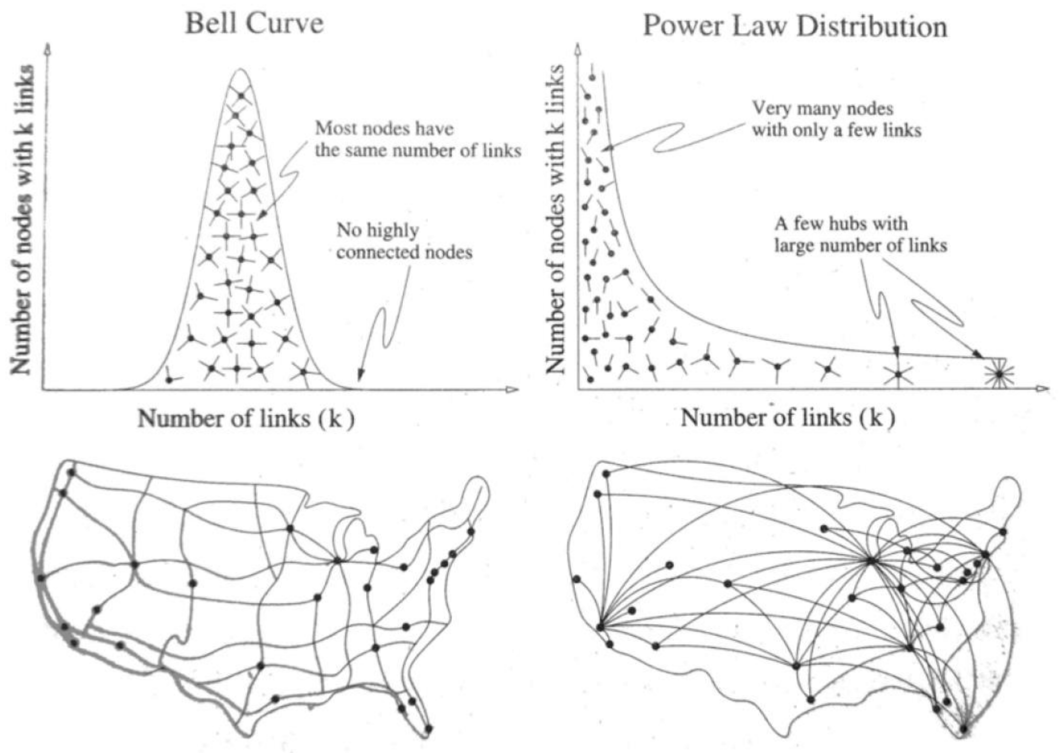
The paper is organized as follows. In the next section we present the results of the relevant research from network science and sports analytics. In Section 3 we present our data collection process, including data cleaning, feature elimination, and dealing with the inflation of the players' transfer market value. Then, we describe the basic properties of our constructed network, with great emphasis on scale-free networks and the small-world property and present the excluded feature of this network. Then, in Section 5, we specify the steps we took to develop different machine learning models to analyse the importance of the features in different scenarios. Using these important metrics, we created different Key Performance Indicators (KPI), that can represent the most vital attributes of the footballers for each of the examined years. In this way, we got time series for each of the examined footballers. Analysing these KPIs could help us a lot, to identify the main differences between different groups of players, and to better understand what the key is of the success among the world class players. In this dissertation we did not cover the examination of time series, we only validated their authenticity, and analysed the features that were used to create the KPIs. Finally, we summarize the most important conclusions in Section 6.

2 Related work

2.1 Networks

One of the biggest breakthroughs in the world of network science was brought by Pál Erdős and Alfréd Rényi who discovered random networks [1]. The above-mentioned scientists both wanted to examine complex graphs, but at that time it was not that obvious how to model networks. They thought most of the networks that occur in real life are unpredictable, asymmetric in structure, and rather appear random. Due to this assumption, the formation of graphs was characterized by the principle of randomness, which means that the best way to build a graph is to add the edges completely randomly between the nodes. Accordingly, each vertex has the same probability to collect edges, so most of the nodes have approximately the same degree. This means that if we want to draw a histogram of the degrees, we get a curve with a Poisson distribution. This was proven by Erdős's student, Béla Bollobás in 1982 [2]. The degree distribution, in this case, follows a bell curve, it has a maximum point, and the other vertices do not deviate much from this. We do not find vertices with very extreme degrees that differ from the average to a large extent. This suggests that, if we look at a social network, all people have nearly the same number of acquaintances. Or if we look at the World Wide Web and measure the connectivity of websites, pretty much each page points to the same number of other pages. Although most networks today are known to be non-random networks, their discovery has greatly contributed to the development of network science.

As stated above, numerous networks have been proven to be small-world networks. These networks have unique characteristics. According to Granovetter's studies, small-world networks have higher clustering coefficients thanks to the many complete subgraphs [3, 4]. Clustering coefficient is a metric in network science, which measures the probability if two neighbours of a vertex are also adjacent to each other. Real-world networks usually have a clustering coefficient between 0.1 and 0.5 [5]. Another vital feature of small-world networks is that the length of the average distance grows only logarithmically with the number of vertices [6]. Consequently, the average distance is relatively small. According to Amaral and his co-authors, the diameter, which is the largest shortest distance between nodes, is also small in the small-world networks [7]. Additionally, the degree distribution of such networks follows a power-law distribution. This means most of the nodes have a small degree while only a few have greater degrees. These nodes are responsible for the weak connections, and we usually call those hubs. Such scale-free networks are all small-world networks [7] and inherently differ from random networks.


Bell Curve
Power Law Distribution
Number of nodes with k links
Most nodes have the same number of links
No highly connected nodes
Number of links (k)
Number of nodes with k links
Very many nodes with only a few links
A few hubs with large number of links
Number of links (k)

1. Figure: The difference between random networks and scale-free networks [4]

Research has already been done in the world of sports related to the analysis of networks. Yuji Yamamoto and Keiko Yokoyama examined the networks that emerged in a football game, representing the players and the passes between them. They concluded that the degree distribution follows a power law and that the exponent values are very similar to real world networks. They also managed to identify the key players who play a big role in the team's performance [8]. Javier López Pena and Hugo Touchette also used information about passes to create networks and to describe football strategy [9], just like Raffaele Trequattrini et al. did, who analysed an UEFA Champions League match [10]. They visualized the line-up of the teams and determined the importance of the players. Pablo Medina et al. used social network analysis to determine match results. They not only developed and analysed networks but also studied their relevance to the results [11]. Paolo Cintia et al. measured team performance with networks as well. They not only used passes to determine the edges between players, but many other actions as well, like tackles, fouls, clearances, etc. They observed that the network indicators correlate with the success of the teams, and then used it to predict the outcome of the matches [12]. Filipe Manuel Clemente et al. suggested that defenders and midfielders have the most connectivity in the team [13]. Also, Clemente et al. got similar results when they analysed the Switzerland national football team in the 2014 FIFA World Cup [14]. E. Arriaza-Ardiles et al. used graph theory and complex networks to understand the play structure of the team. They used clustering and centrality metrics to describe the offensive play [15]. Opposed to the listed works that are all analysing in-game relationships, our intention is to create a graph theoretical model on player level in order to access relationship information between footballers.

2.2 Football careers

Many related works have appeared on this topic. Here we discuss what are the key differences compared to our work, and what conclusions can be drawn.

Bettina Schroepf and Martin Lames searched for career patterns in German football youth national teams [16]. They managed to find 8 typical career types. It has been found that the careers of youth players last up to 1 or 2 years in Germany and only a few players can achieve long-lasting careers. It is interesting how big is the churn rate by young players in major European football nations, so it is already a privilege here for someone to be among the pros. In our paper we have broaden the search and tried to find career paths of adult players around the globe. A recent study revealed that in Portugal the length of a career as a football player decreased, but the years of youth career increased: it follows that the career path started earlier in the last 3 decades [17]. Remaining in Portugal, Monteiro et al. identified that the best result the players performed was at the age of 27 and they ended their career at around 33 [18]. Other researchers revealed that the peak performance by female football players is around the age of 25 [19] or between 25 and 27 [20] in case of male players. In our study we preserved the players between the ages of 15 and 47 and the majority of the examined players were between the ages of 22 and 27.

Identifying the career potential of athletes is a very difficult task. Coaches play a big role in this aspect, because they have the closest relationship with the players. However, they also encounter difficulties. In the article of Cripps, A. J. et al., the authors found that coaches can predict the career outcome more accurately for late maturing athletes, but they are less accurate for early maturing players [21]. It is important to keep in mind that the way of maturation can easily affect the career path both positively and negatively.

It is also interesting how it is possible to draw a parallel between psychology and performance in terms of success. Schmid, M. J. et al., found patterns in rowing by connecting these variables. They measured the proactivity, ambition, and commitment of the athletes for the past year and 30 months later. As a result, if a highly motivated rower had poor performance in the past year, it was more likely that he or she performed at a very high level in the future.

Athletes with low morale and motivation were more likely to drop out or perform weaker [22]. Good motivation is a key to perform better. Highly achievement-oriented players have a better chance to accomplish an outstanding career [23]. We only used skill and performance variables but based on these statements, there is potential in the introduction of certain psychological features into our model.

Vroonen, R. et al. [24] predicted the potential score (available on Sofifa.com) of professional football players. They selected a player and searched for similar players from the same age. Based on the evolution of the latter, they predicted if he or she had great potential score in the future. In contrast, in our paper we recognized continuous market value growth patterns and we have shown which skill development is required to achieve this goal. We did not predict the potential scores of Sofifa, rather we paralleled the available information from Sofifa with the development of real market prices and we successfully explained the real-life career path patterns we were looking for with them.

3 Data collection and preparation

Our goal was to collect data diversified by nationality, club, league, or international popularity. We used two sources for this type of data: Sofifa.com and Transfermarkt.com. From Sofifa we were able to collect 15 years of data about players included in the FIFA database. The data from this site plays a big role in this analysis, because it contains 21 different skill scores (e.g., dribbling, short passing, finishing), market values, wages, and other personal information (e.g., age, weight) for each player. All skill variables are stored in a range from 1 to 99, the higher is the better. It is important to note that the different positions (e.g., ST, GK, CAM) require different skills. For example, for a goalkeeper the goalkeeper handling skill is more important than finishing. Besides the skill and personal features, the market value is a key element.

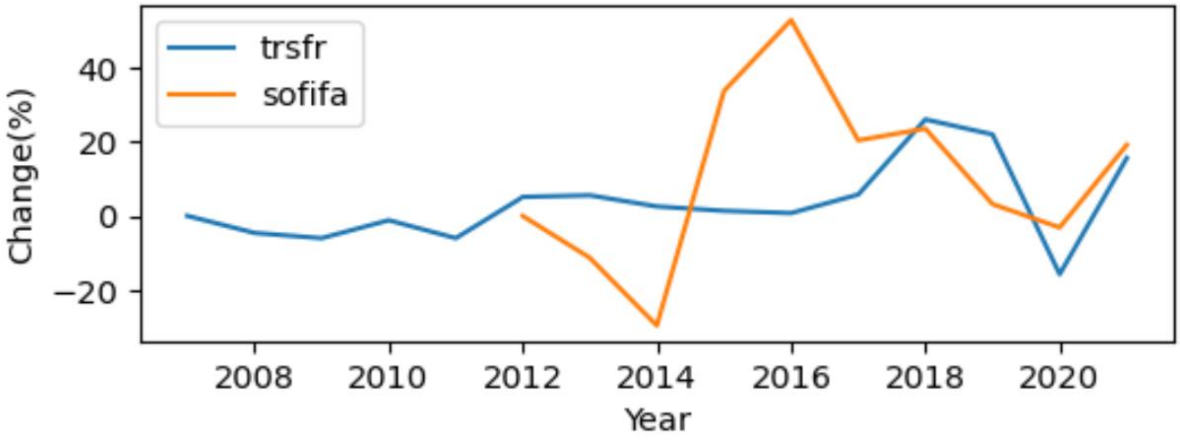
Collecting the data from Sofifa was not enough because of two reasons. First, between 2007 and 2011, the market value of players is missing from the Sofifa database. As market values are essential in our analysis, we had to collect them from another site, i.e., Transfermarkt.com. Furthermore, within the Sofifa player pricing data there is a significant difference in the year-over-year change of the mean market values compared to the Transfermarkt values, and there is a great fluctuation in the values between 2012 and 2016, raising suspicion about the quality of data.

The range of the collected data is 15 years (from 2007 to 2021) from both sources. In order to deal with the distance between the high-end and low player values, we performed a logarithmic scaling transformation on the prices.

3.1 Market value differences between Transfermarkt and Sofifa

We found a significant difference between the player values we collected from the two sources. First of all, we wanted to figure out if the two datasets are from the same population. For this purpose, we used the Mann-Whitney U test and Kruskal-Wallis H test. We tested the market values every year, from 2012 to 2021 (before 2012 the Sofifa dataset had missing pricing values). We denote the Transfermarkt dataset as TR and the Sofifa dataset as SO throughout this dissertation. Except for 2012, the TR and SO datasets were different, we had to reject H0 at 0.05 significance level: we can clearly state that the market values reflected in TR and in SO are different.

Moreover, the mean year-over-year changes are also significantly different: in Figure 2 it is clear that the price evolution is far from being the same in the two datasets. Since the market values before 2012 were not available in the SO dataset and the fluctuation has been much greater over the years than in TR, we decided to use only the TR prices during the analysis.



2. Figure: Transfermarkt vs. Sofifa mean year-over-year market value change

3.2 The 2014-16 market value boom

In Figure 2 it is noticeable that something happened between 2014 and 2016. The SO dataset shows a strong market value incrementation in this period, and the TR values are also increasing during and after this phenomenon. After the decrease from 2012 to 2014, in 2015 there was a 33% growth and from 2015 to 2016 this growth increased to 53% in SO. This exceptionally large jump between 2014 and 2016 is due in part to affluent Arab and Russian investors. In these years United Arab Emirates, Qatar and Saudi Arabia suddenly appeared in the list of the top 20 spenders in the football market across the world. In 2014 Al Arabi (Qatar Stars League) spent over 50 million dollars on the transfer market and with this Al Arabi was the eighth placed club inside the top 10 spenders. In the 2016 transfer season, Arab football teams spent over 200 million dollars on transfers [25].

3.3 Handling the football market value inflation

We also tackled the issue of price inflation. The inflation in the world of football seems to supersede the regular monetary inflation. For example, 1 British pound in 1990 was worth 2.27 in 2019, but in football 1 pound of 1990 would be worth about 40 now. As a stellar illustration: while in 1989 the whole squad of Manchester United was worth 20 million pounds [26], today the most valuable player in the world is Kylian Mbappé with 144 million pounds. With this observation we calculated the inflation rate of market values over the collected 15 years. We tried different approaches to handle inflation: we adjusted the market values of each year to match the mean, median and the third quartile statistics with those of the latest year values. By doing so our intention was to set the values of each year to their present value as close as possible.

We found that the best statistics was the third quartile (Q3): we set the Q3 in 2021 as the base value, being 1.8M Euros, and linearly scaled each year's values to bring their Q3 to this value. After this transformation every year had 1.8M Euros as Q3: the YOY became uniform, the sudden changes in pricing and even the effects of the 2014-2016 market value boom disappeared.

3.4 Feature elimination

During our work we required a strong feature elimination because of the properties of the soccer skills. It is easy to understand that playing in different positions requires different skills. For example, a goalkeeper does not need to be a good free kick shooter, and an attacker does not need to have a good diving skill. In our dataset we had a total of 33 independent skill variables like dribbling, finishing, marking, heading, and goalkeeper skills. We decided to find the most important skills for each position based on the market value of the players. We used the Random Forest Regressor as a model for feature selection. In this case we wanted to predict from X independent variables (skills) an Y continuous dependent variable (market value). First, we omitted and merged certain positions, because in some cases very little data was available from the scraped dataset (e.g. count of RWB players was 88). For example, we merged the position CF (central forward) with ST (striker), because the selected features for each position were similar or the same. Table 3 shows the final positions. We reduced the number of positions from 14 to 10. It contains the positions that we used with full names and the short names we will use in the next sections.

CAM	Central Attacking Midfielder
CB	Centre Back
CDM	Central Defensive Midfielder
CM	Centre Midfielder
GK	Goalkeeper
LB	Left Back
LM	Left Midfielder
RM	Right Back
RM	Right Midfielder
ST	Striker

Table 1: Player positions

During feature selection, we collected the five most important features for each position, and we used only those during the modelling phase. The feature selection was implemented with 3 different methods. In the following sections we show these methods and results, then our final feature selection decision. We illustrate the feature selection process with the Centre Back (CB) position.

3.4.1 Recursive feature elimination

The first method we used is the recursive feature elimination. It is basically a backward selection of features. In the beginning it uses all the features to build the model and computes an importance score for each variable. Next step is to remove the least important feature and rebuild the model. This iterates until the number of selected features reaches the predefined value (in our case it is 5) [27]. It performs a greedy search, and it may consume a lot of time to calculate the results. In the worst case, if the dataset has N features, it will create 2N combination of features [28]. Table 2 shows the top 5 features of the recursive feature elimination selected from the CB player list.

Recursive feature elimination	
1	Heading accuracy
2	Acceleration
3	Sprint_speed
4	Reactions
5	Strength

Table 2: Top 5 CB skills with recursive feature elimination

3.4.2 Backward elimination

The next method is the backward elimination. In a backward elimination process, there are six important steps. The first step is to choose the significance level (known as the P value). Here we used 5%. The second step is to fit the selected model with all independent features, 33 in this case, then to find the highest P value. If it is greater than the significance level, then remove this predictor and fit the model again. This iteration goes until the highest P value is less than the significance level [29]. In Table 3 we can see the top 5 features of CB players selected with this method. Besides Acceleration and Sprint_speed there are new selected features here.

Backward elimination	
1	Finishing
2	Volleys
3	Acceleration
4	Sprint_speed
5	Agility

Table 3: Top 5 CB skills with backward elimination

3.4.3 Extra tree classifier

Last, but not least we did feature selection with an Extra Tree Classifier. It is very similar to Random Forest, except it is creating the decision trees in the forest differently. This ensemble learning technique aggregates the results of multiple de-correlated decision trees in a forest. To construct the forest, it is needed to select the best features in each tree. At the beginning, every tree is provided with n random features and the best features will be selected by a mathematical formula, which is in this case the Gini index. This leads to multiple de-correlated decision trees in the forest. Now it is possible to perform the feature selection, by the Gini importance. This is computed by the normalized total reduction for each feature [31]. We ordered these features in descending order by the Gini importance and selected the top 5 most important. After looking at Table 4 we can see that every feature selection gave us different results. The extra tree classifier method is closer to the recursive feature elimination results, but in order to make a clear decision we had to test these features in a regression model.

Extra tree classifier	
1	Standing_tackle
2	Interceptions
3	Sliding_tackle
4	Heading_accuracy
5	Reactions

Table 4: Top 5 CB skills with extra tree classifier

3.4.4 Final feature selection and dimension reduction

To find out which five features are the best, we created three XGBoost models based on the market value of the players to test the three feature selection method. Table 5 shows the results of the models on the CB players (the other positions had almost the same results, so we represent here only the CB players)

XGBoost model results for market value prediction (CB)				
Method	MAE	R^2	RMSE	Mean cross-validation score
Recursive feature elimination	0.18	0.81	0.24	0.77
Backward elimination	0.41	0.27	0.54	0.14
Extra tree classifier	0.16	0.83	0.23	0.80

Table 5: Feature elimination performances

As we can see the backward elimination performed poorly, but the other two were almost the same. In the end we decided to continue our analysis with the features from the extra tree classifier elimination method.

Finally, we needed dimension reduction for two reasons. First, the importance of the skills and in some cases the required skills are different by position. Second, one of our targets is to find the career patterns of the players with outstanding skill improvement during their professional time. Therefore, we created five universal skill variables (Skill 1, Skill 2, Skill 3, Skill 4, and Skill 5) from the 18 different skills that remained after the first feature elimination. From now on, Skill 1 means the most important skill characteristic of a given player, Skill 2 is the second one, and so on. With this we were able to deal with the problem of different positions. For example, we can now compare how the goalkeepers' most important ability against strikers has evolved over the examined years. Table 6 shows all the selected features broken down into positions ordered by its importance.

Position	**Skill 1**	**Skill 2**	**Skill 3**	**Skill 4**	**Skill 5**
CAM	Ball_control	Dribbling	Vision	Short_passing	Finishing
CB	Standing_tackle	Interceptions	Sliding_tackle	Heading_accuracy	Reactions
CDM	Standing_tackle	Interceptions	Short_passing	Reactions	Ball_control
CM	Short_passing	Ball_control	Dribbling	Vision	Long_passing
GK	Gk_reflexes	Gk_diving	Gk_positioning	Gk_handling	Reactions
LB	Standing_tackle	Sliding_tackle	Interceptions	Ball_control	Reactions
LM	Ball_control	Dribbling	Positioning	Short_passing	Reactions
RB	Standing_tackle	Sliding_tackle	Ball_control	Interceptions	Crossing
RM	Dribbling	Ball_control	Reactions	Short_passing	Positioning
ST	Finishing	Positioning	Shot_power	Ball_control	Reactions

Table 6: Features selected by positions

3.5 Other targets collected

To better differentiate the top players from the others, we scraped two other features as well, what we used as a target in our models. One of these features is the information about the footballers playing in their national teams. We created an IsNational binary attribute, which is 1 in case the given player played at least one match in the respective national team in the given year (according to Transfermarkt data), otherwise 0. Only the best players can play in their country's national team, so using this as a feature could help the separation of the exceptional players. Similarly, we named the second such feature IsTOTS, that represents the Team of the Season. If a player got into the Team of the Season in the given year, this attribute is 1, otherwise 0. We pulled this information from the Futhead website, which deals with FIFA computer game information about the footballers [31]. Every year Electronic Arts, the developer of FIFA, creates the Team of the Season for many leagues, based on the real performance of the footballers. Getting into this team is a great honour, meaning the given player had an outstanding season. Unfortunately, this information is only available between 2012 and 2019, so we had to narrow our dataset down during the career book research process.

4 Network research approach and findings

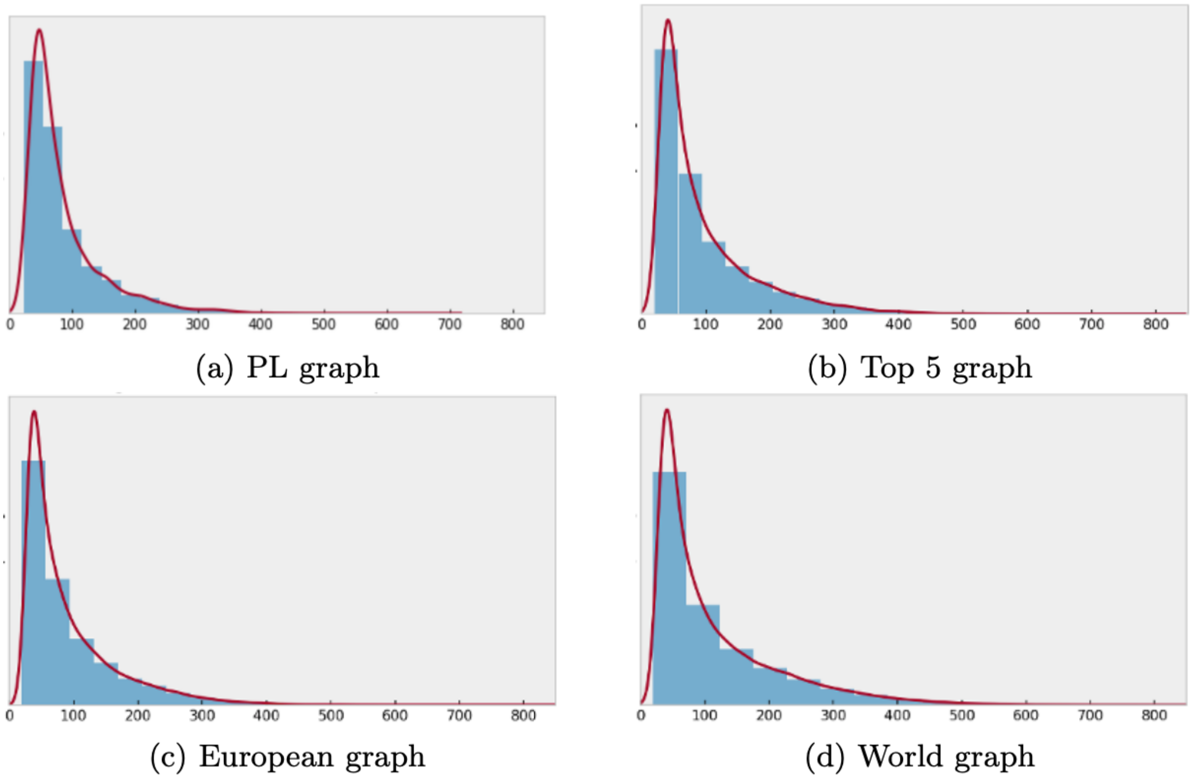
4.1 Analysing the acquaintance network between players

In this section we present the graphs we created for modelling the relationship among players. We created four different graphs with more and more leagues involved. In this way, we could examine how the network metrics have changed with the increasing number of nodes. We started the analysis with one of the most competitive leagues in the world, the English Premier League. For this graph, we used the information available on the official website of the league [32]. We studied the teams from the 1992/93 season to the 2020/21 one, so this part of our research covered the entire history of the Premier League. We created three more graphs: we named the first one as Top 5 as it includes the best 5 European football leagues; the second graph, named as European, contains 24 first division leagues from Europe; the third graph, named as World, contains 58 leagues from all over the world. Both first, second, third and in some cases even fourth divisions are included. These latter 3 graphs are based on Sofifa computer game data, which is published every year with updated squads. We used information about players since the 2007 release. We created familiarity graphs, in which the players became the nodes and two players considered to be adjacent if they were teammates for at least a season. After defining the edge list, the developed graphs have the following metrics, summarized in Table 7.

Metrics	Premier League	Top 5	European	World
Nodes	6,407	22,509	42,827	92,969
Edges	271,083	1,070,595	1,964,483	5,299,404
Average degree	84.62	95.13	91.74	114.00
Max degree	486	570	583	801
Min degree	23	20	19	19

Table 7: The basic metrics of the player graphs.

First, we examined the degree distribution of the graphs, which can be decisive in answering our question. Small-world networks have power function distribution, which means that most players have only a few connections, but some players have a lot. Indeed, a power function is followed by the distribution of the degrees in the four graphs we created as Figure 3 shows.



(a) PL graph

(b) Top 5 graph

(c) European graph

(d) World graph

Figure 3: Degree distribution of the graphs follow power law

The average shortest distance and the diameter are both relatively small in small-world networks. Quite precisely, these numbers are roughly equal to the logarithm of the number of nodes. The average distance is 2.63, while the diameter is 5 of the Premier League graph. As 6,407 base 10 logarithm is 3.8, the average distance is even smaller than what is required for a small-world graph. This means that in the Premier League, the distance between players is very short, averaging less than 3 players and even between the farthest players, the distance is only 5. As shown in Table 7, these metrics are also small for the other graphs and increase at a slower rate than the logarithm of the number of vertices. It is utterly amazing that in the examined leagues, which cover the entire world, players are just over three steps apart on average. What is more, it only takes a maximum of six steps to connect any two players, even if they play in the farthest parts of the world.

The third characteristic which we examined is the clustering coefficient metric. It seems logical that this should be comparatively big, since teammates form a complete subgraph within the graph. The whole network is made up of such subgraphs connected by players who have turned up in several teams. When examining the clustering coefficient, we are curious about the extent to which there are triangles in the graph, so this property is also commonly referred to as triadic closure [33]. We will use these two words as synonyms hereafter. This metric for the Premier League players' graph is 0.41 which is closer to the up- per limit of the standard value described earlier. Table 8 suggests that with the increasing number of vertices the clustering coefficient becomes smaller. But as it is relatively still far from zero, the last condition is also met, so in the world of football, players make small-world networks.

Metrics	Premier League	Top 5	European	World
Nodes	6,407	22,509	42,827	92,969
Log of nodes	3.81	4.35	4.63	4.97
Average distance	2.63	3.00	3.17	3.24
Diameter	5	6	6	6
Clustering coefficient	0.41	0.31	0.31	0.25

Table 8: The distances and clustering coefficient of the graphs

As we can see, the world of the football players is small, no matter how many championships we take into account, since all the properties that apply to graphs with small-world properties are fulfilled in them.

4.2 Network feature extraction

To get information about the players' relationship with each other, we also excluded features from the networks. We used the players' graph to get information about the players' acquaintances. For this purpose, we built graphs for every year to avoid the scenario when a player gets acquainted with a player with whom he played later. For example, if the given year is 2017, we used the graph that has the data from 2012 to 2016 to avoid a player's score being influenced by a later teammate. We extracted the degree number, which represent the number of different teammates the given player had; the eigenvector, which is a centrality metric to measure the node's importance in the graph (it takes into account the degree of the vertices and the degree of its neighbours when determining the importance value); the number of links each player has with other players who played in their national team, and the number of teammates who got into the Team of the Season. With these features we wanted to include information about the structure of the footballers' world, and our aim is to improve the performance of the models with the extra information.

5 Key Performance Indicators modelling and validation

Our aim is to create a tool that helps to understand the differences between world class players and the other players. To be aware of the differences could mean a competitive advantage, as the player would know exactly what is needed to be improved to reach his or her goals, what career path should be taken, and what decisions should be made. Also, with analysing the differences between the two groups we can identify the optimal strategic choices towards multiple potential aims a soccer player can have.

We analysed four such potential aims, created machine learning models to predict the potential to reach these goals, and with the models' feature importance, we created Key Performance Indicators (KPI) for each year for the players. In this way we got different time series for each of the examined players, that included the different features of the models with various weights. Analysing this time series could give us a picture about the improvement of the players and comparing the two groups (those who reached the given goal, and those who did not) could highlight the necessary steps the players have to take in order to reach the world class level. We validated the correctness of our KPIs with Dynamic Time Warping. In the following sections, we present the creation of the KPI time series.

5.1 Classifier models to create the Key Performance Indicators

We defined four aims a player could have, in order to reach his or her full potential. The first one is to get into the Team of the Season, the second is to play in one of the Top 5 league (in our case the Premier League), the third is to reach a relatively high transfer market value, and the fourth is to greatly develop the main skill of the player's position, making him or her unique. These goals could be real goals for many young players who want to become world class players one day.

Using the above-mentioned data, we created target variables that serve the purpose of differentiating the selected players and the other players. We trained the model on our dataset that contains Sofifa data between 2012 and 2019 with the grouped main skills, the features extracted from the graph, and the IsNational and IsTOTS features. This means 92,120 records of 29,231 different footballers. We developed four different models for the four different aims, choosing the target variable in the following way:

- In the first case, the target variable was the IsTOTS feature.
- In the second case, we created an IsPL variable, and filled in with ones, if the player plays in the Premier League in the given year.
- In the third case, we used the inflated market value to decide which players are in the world class category. We created a binary variable called IsValue, and filled with 1, if the player's inflated market value is at least 7. We choose this value after analysing the histogram of that field, which can be seen in the Figure 4 below. In 2019 this value means exactly 10 million euros.
- In the fourth case, we created a target variable based on the Skill 1 variable, which is the most essential skill of the players according to their position. We chose 80 as the turning point, so we marked the players with Skill1 feature who reached this limit with ones.

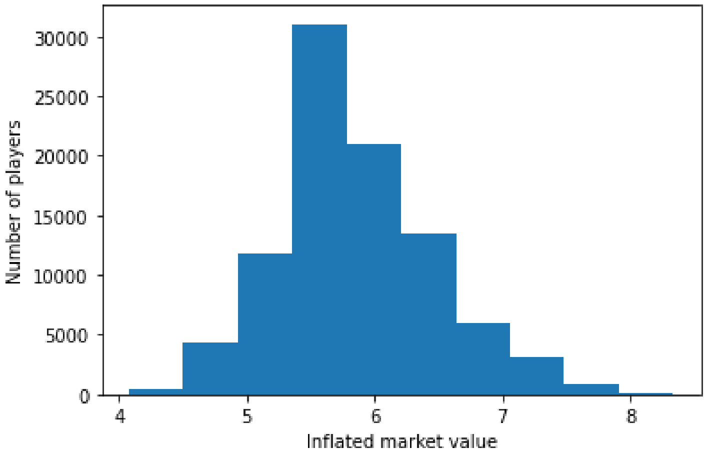


Figure 4: Histogram of the Inflated market value

During modelling, we used the same process for all of the models to define the one with the best performance. First, we checked the distribution of the target variable in all four cases. The most imbalanced target variable was in the case of the IsTOTS model, having only 1,733 positive cases out of the possible 92,120. This was followed by the IsPL model with only 3,932 positives, and by the Value model with 4,465 positives. The Skill model's target variable was more balanced, with 7,072 positive cases.

In the next step, we tried out numerous different models and compared its results. We used cross validation to train the models and choose the best performing one according to the AUC (Area Under Curve) metric on the test set. In case many models had similar AUC, we also considered the Accuracy and the F1 metric. In three cases, the Random Forest Classifier worked the best, but for the IsPL model the XGBoost Classifier achieved slightly better results.

As the dataset contained some categorical variables (like the Country, League, and Position), we had to encode those variables into numerical ones, to be able to use those as well. We checked the performance of the chosen models with different encoding techniques, and similarly to the model selection, we chose the one with the best AUC. CatBoostEncoder worked the best for the IsTOTS and the Value models, JamesSteinEncoder had the best AUC for the IsPL model, and for the Skill 1 model we chose the MEstimateEncoder.

Due to the imbalanced data set, we used SMOTE (Synthetic Minority Oversampling Technique) to reproduce the favourable data, if there were too few positive cases and it led to performance degradation. The SMOTE technique uses oversampling with the k-nearest algorithm to smooth the data set. After trying several techniques, oversampling with BorderlineSMOTE worked the best. Using this method the AUC metric increased greatly, but creating too many records resulted in over-fitting, so we only added that many cases that was necessary to reach a 30% rate of positive cases. Only the Skill 1 model did not require this oversampling technique.

Finally, we used grid search to find the best parameters. We chose the set of hyperparameters with the highest AUC metric. The best set of parameters are 500 estimators, 0.9 maximum samples, and 11 maximum features. The final AUC of the different models can be seen in below Table 9.

Skill1 model	**IsPL model**	**IsTOTS model**	**Value model**
87.65%	83.15%	85.20%	86.88%

Table 9: AUC metric of the models

As mentioned before, we used these models to identify the most vital features in each case. To do that, we used the feature importance function of the models, that weights every feature according to its predictive power. The 10 most vital features of the models can be seen in the following Table 10.

Skill1 model	**IsPL model**	**IsTOTS model**	**Value model**
Skill 2	Value	TOTS teammates	Skill 1
Value	Country	Value	Skill 2
Skill 4	National teammates	Skill 1	Skill 4
Skill 3	Division	Skill 2	League
Skill 5	Eigenvalue	Skill 4	Skill 3
Eigenvalue	Degree	Skill 5	Skill 5
Position	Skill 1	Skill 3	TOTS teammates
National teammates	Position	National teammates	National teammates
Degree number	TOTS teammates	League	Country
Country	Skill 3	Eigenvalue	Eigenvalue

Table 10: The most vital features of the models

Different models have different order in the most important features. Every feature excluded from the network made it to the Top 10, that means playing with better teammates can have a positive effect on the players' career. We can also see that, if a player wants to improve one of his or her skills, other skills also need to be improved as well. The Position attribute is also present in the Skill model, which suggests different positions require different skills. The most important features of the PL model are the Value and Country, but after this, the number of national teammates is the third. Playing with players who made it to their national team can be an advantage, if somebody wants to play in the Premier League. The best way to get into the Team of the Season, is to play with many other players, who already did it. This could represent a good team, whose players usually made it to the TOTS. Transfer market value and skills also play a great part in that. The best way to increase your market value is to increase your skills and play in better leagues, but also having teammates who play in the national team and get into the TOTS has a good effect on that.

Using the feature importance, we identified what are the most essential attributes that need to be developed or need to be focused on in order to reach different goals. Using the weights of these feature importance metrics, we created Key Performance Indicators for every year of the players. In this way we got time series that consist of these KPIs. Analysing these time series could not just highlight what attributes need to be improved, but also the way and rate of the improvement. Therefore, grouping the players into two groups (selected and others), and comparing the KPIs should be the next step. We grouped the players differently in each model:

- In the TOTS model, the selected players are the ones, who got into the TOTS at least once during the examined period.
- In the PL model, the selected players are the ones, who played in the Premier League at least once during the examined period.
- In the Value model, the selected players are the ones, whose average inflated market value during the examined period was at least 7.
- In the Skill 1 model, the selected players are the ones, whose average Skill 1 attribute during the examined period was at least 80.

In Table 11 we show an example of the grouping where the goal is to reach a desired market value. Here, the aim is to reach the inflated market value of 7, which means about 10 million Euros in real life. The feature importances in Table 11 indicates that the skill scores are vital for this objective. This table shows the mean skill scores of each group.

Group	Skill 1	Skill 2	Skill 3	Skill 4	Skill 5
Selected	82.2	81.7	79.3	79.1	76.5
Other	68.9	68.1	66.2	66.4	63.8

Table 11: The average skill scores in the groups.

In the group of selected players, the skill scores are significantly higher than in the other group. It seems that our grouping is correct, but before analysing the evolution of these skills or the KPI time series we validated those correctness with comparing the result within the groups, and between the two groups using Dynamic Time Warping. In the following section we present the result of it.

5.2 Dynamic Time Warping for selection validation

In the previous section we found the most important features for each goal with the models we created. From the feature weights we created KPIs for each player in each year so in the end we had a KPI time series for every player in our dataset as well. Based on our hypothesis of what the difference might be between the outstanding players and other players in each career goal, we divided the players into two groups: selected and other players. The next step is to validate these assumptions with the help of the newly created KPIs.

Since our dataset contains players with different length of career time (someone has three years, someone has five, etc.) and we did not limit the number of examined years (e.g. for the first five years), we had data from 2012 to 2019 with different length of individual time series. In order to compare time series with different lengths and find out if our grouping was correct, we used Dynamic Time Warping (DTW). We measured three distances here:

- The distance between the KPI of selected players (the grouping was successful if the distance is relatively small).
- The distance between the KPI of the other players (the grouping was successful if the distance is relatively small).
- The distance between the KPI of the two groups (the grouping was successful if the distance is relatively higher than the measurement within the groups).

If all three prerequisites were met, then we considered our grouping successful and we could declare that we found significant difference between the careers of the players. In Table 12 we collected the results of Dynamic Time Warping for each career goal, where we calculated the mean distances for each group observation.

Measurement	PL	Value	TOTS	Skill 1
Selected players	0.33	2.42	0.57	0.35
Other players	0.34	3.59	0.20	0.43
Between groups	0.53	9.35	0.72	1.02

Table 12: Mean DTW distances of career goals

As we can see in each case the distance between the groups is higher than the distance within the groups. The most outstanding differences can be identified by the skill and value goals and the PL goal is also acceptable. By the TOTS examination we can see that the inner distance of the selected players is between the two other measurements. This is partly due to the low number of players ever achieving this title, but as the distance between the groups is larger, we also accepted this result and considered the separation to be satisfying.

In Figure 5, 6 and 7 we present the concrete result of Dynamic Time Warping by continuing our example with the value goal. In Figure 5 we can see the visualized DTW distances between the first 2 players within the selected group.

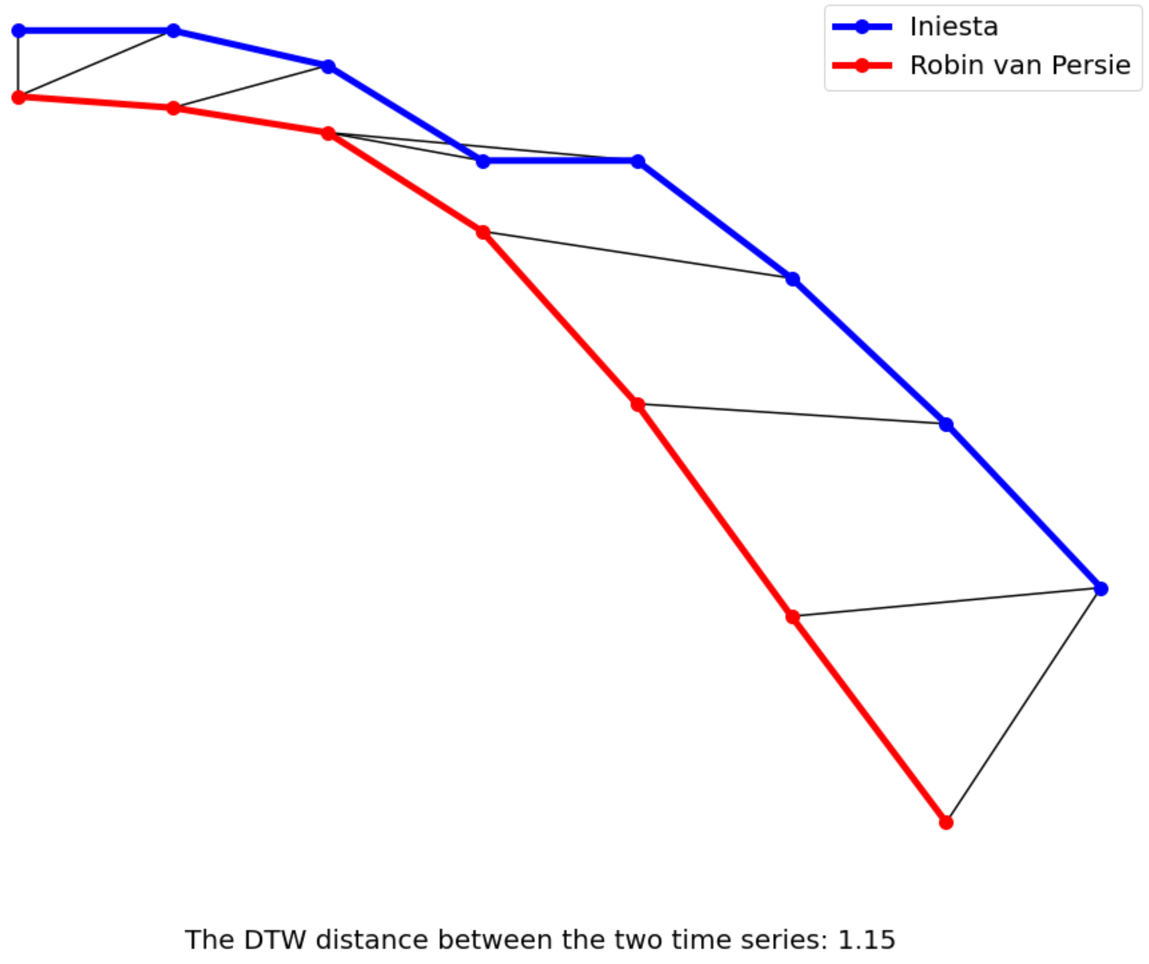


Figure 5: Distance within the selected players

We can see that the distance is relatively small: 1.15. In this example, this means that if the other two prerequisites are met, then our selection is correct, the two players belong to the same group. Figure 6 shows the results within the other players group.

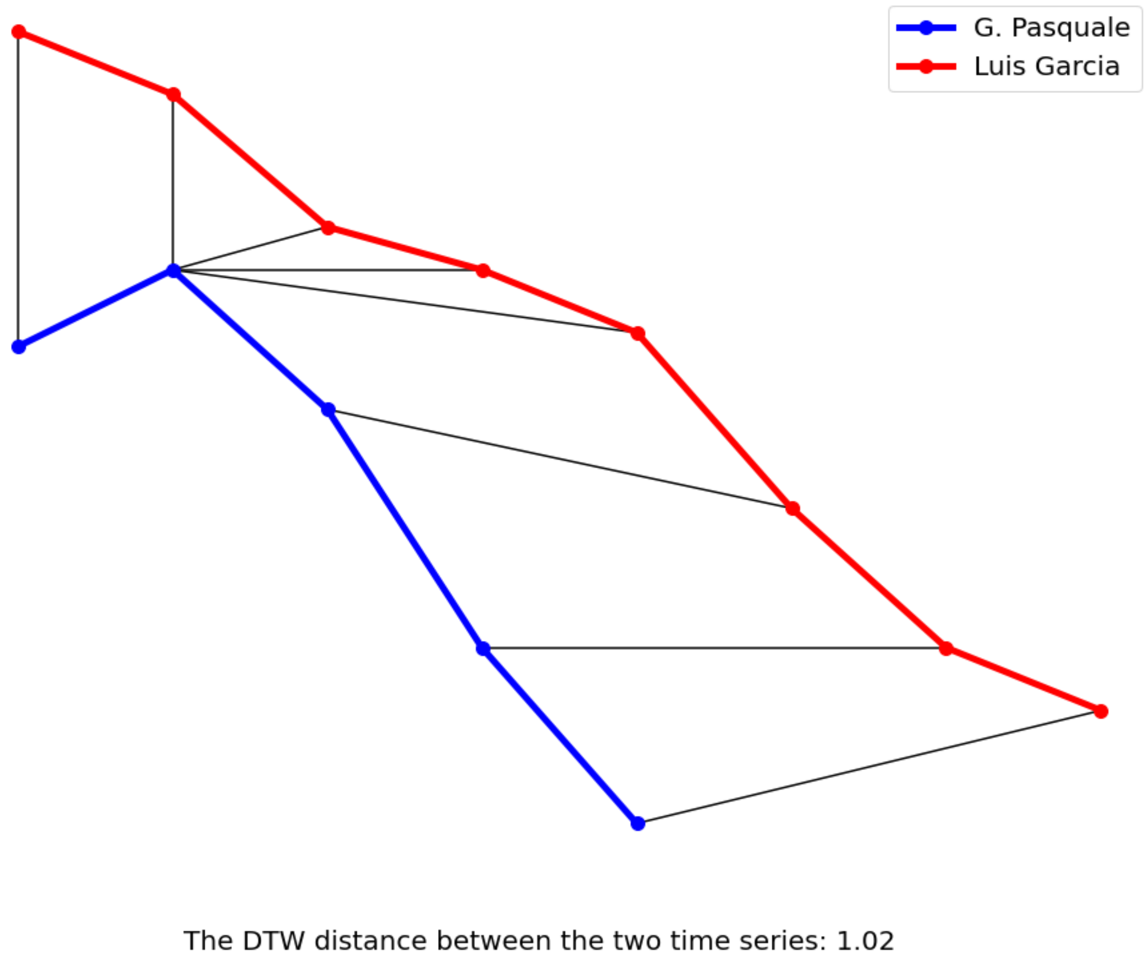


Figure 6: Distance within the other players

This comparison gives almost the same result; the distance is: 1.02. Now, if our estimations were correct the distance between the two groups should be relatively higher.

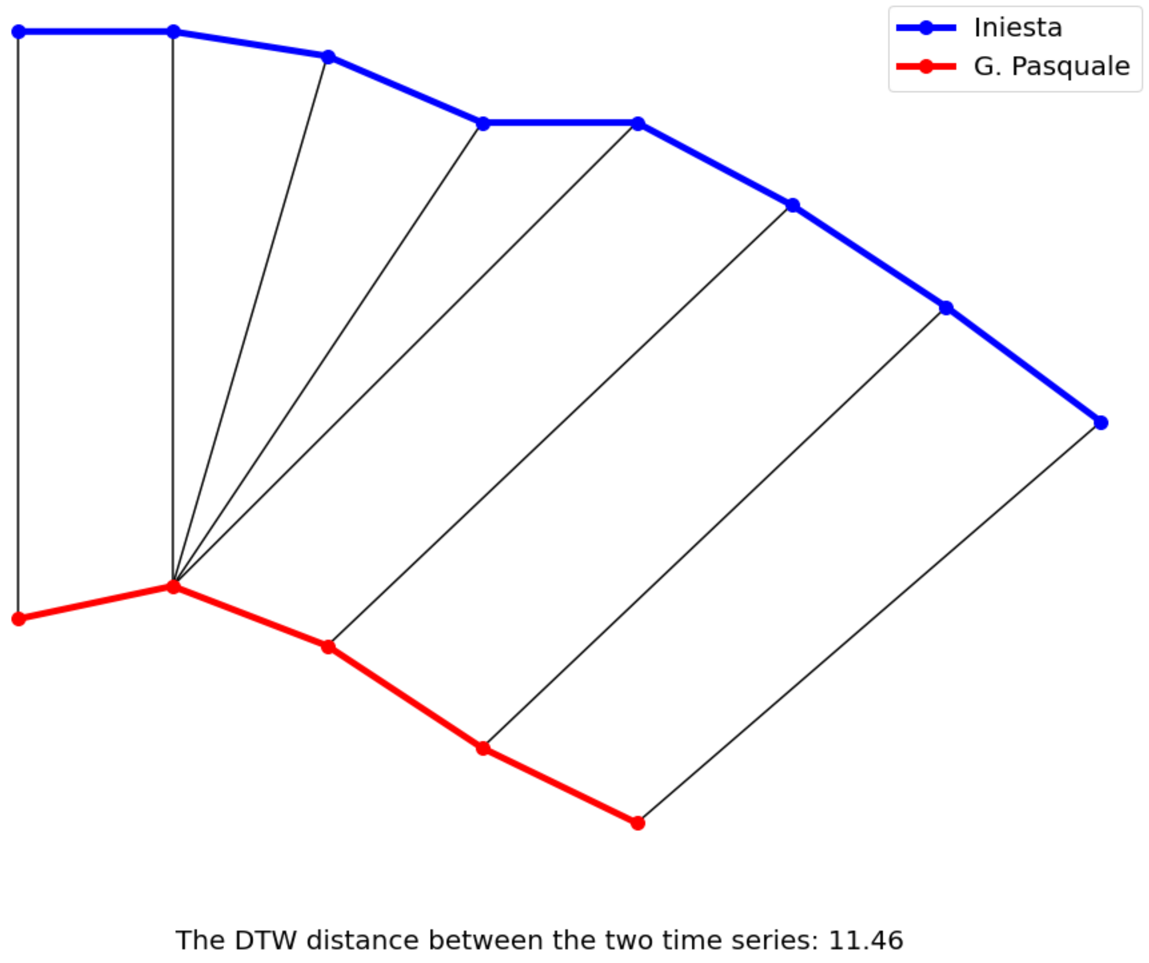


Figure 7: Distance between the two groups

In Figure 7 we compared Iniesta (selected) with Pasquale (other). It is easy to interpret that our estimations were correct. The distance is: 11.46. This means that the distance between the groups is much greater than within the groups, and the classification of the players was successful. This example almost represents the exact values we calculated in Table 12 (Value).

Finally, we conducted some tests to see how our results perform in practice. In this example, we tested the evolution of the five skill variables by the value goal as these variables were the most important by creating the KPI for this objective. How should the players develop their skills to increase the probability of achieving this goal?

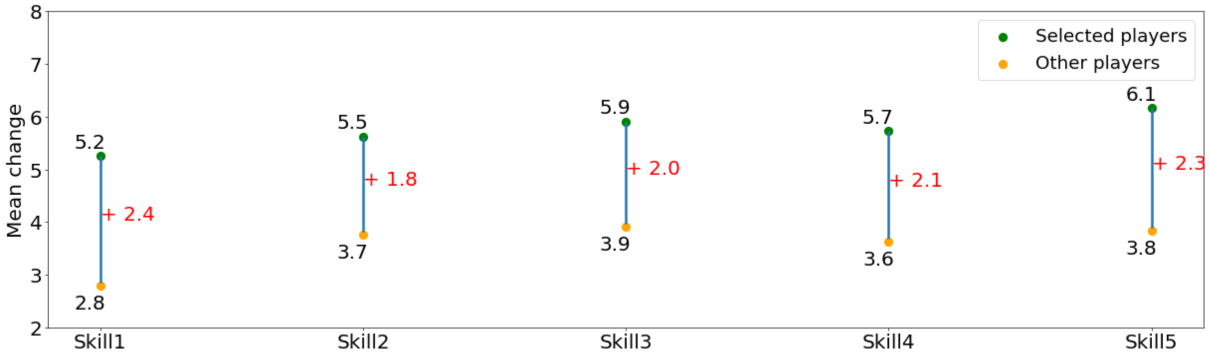


Figure 8: The evolution of the skills

Figure 8 shows how exactly these skills should improve during the examined years. We can notice that in each group the skill scores are increasing over the years, but the skills of the selected players have an increased growth rate. We could say that, if someone is already a high-end player, then it is more difficult to improve his or her skills, but it seems that it is a must, because it is one of the key factors to achieve an outstanding career and become successful. In Figure 9 we finalize our value goal example with the Skill 1 evolution comparison between a selected player (G. Chiellini) and a not selected player (J. Jonsson).

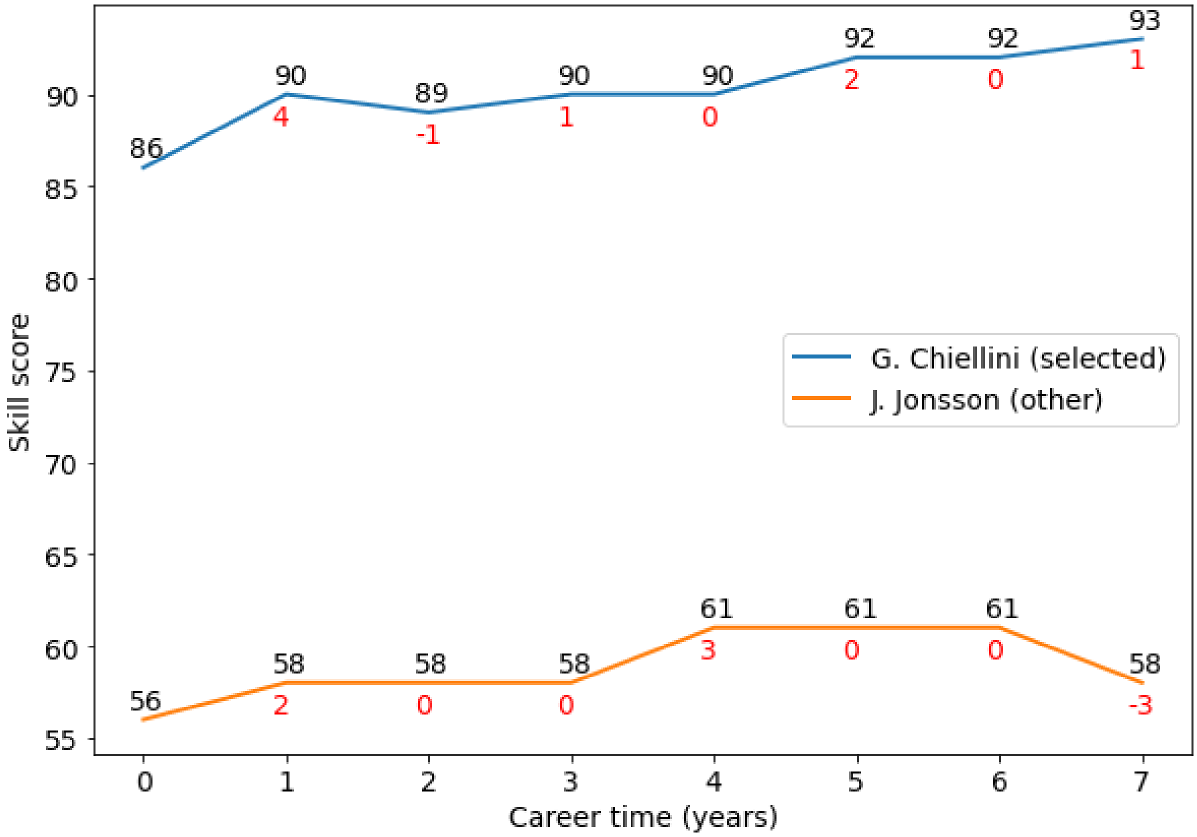


Figure 9: Skill 1 evolution comparison

The former player achieved the desired market value goal (10 million Euros) during his career and we can see a mostly positive increase in his most important skill (CB – Standing_tackle) with a mean change of +7. The latter player's position is also CB, however his most important skill mostly stagnates and after a strong (+3) increase we can see a strong (-3) decrease with an overall mean change of +2. We can conclude that even if the player is at a high level, it is possible and at the same time necessary to develop his or her most important skills for excellence. If J. Jonsson had been able to develop his Skill 1 to around 63 and maintain it throughout his career (we should not forget about his other skills and features, where the same way of thinking is applied) he could have greatly increased the likelihood of achieving the goal and becoming an outstanding player.

6 Conclusion

The success of a soccer player is not entirely pre-destined by their physical ability, talent, and motivation. There are certain decisions along the way that greatly affect the arc of their career: which skills to develop, which club to sign a contract with. In this paper we identify the optimal strategic choices towards multiple potential aims a soccer player can have. Players, managers, and clubs may use the lessons learned in their strategic decisions. We successfully handled the football market inflation, found the most important skill features for each position, and dealt with the unique position skill requirements. We created new features from network and graph analysis and also took into account national team members and Team of the Season players. We built classification models to find all the features that were essential to build our KPIs and divide our dataset into selected and other players based on the desired career goal. Then we validated our player separation with Dynamic Time Warping where we compared the KPI time series, and we could declare that this methodology is successful. Finally, we did a practical test of our findings and showed in an example (market value goal) how the players should evolve their skills in order to achieve this goal and what are the exact differences.

In the future we plan to improve our models and methodology with a KPI analysis where we compare not just the skill, network, graph and other variables, but the KPIs we created for each career goal. With those time series comparisons, we will be able to dig deeper and find more hidden connections and unique properties of the professional soccer career paths.

References

1. Erdos, P., Renyi, A.: On the evolution of random graphs. Publl. Math. Inst. Hung. Acad., vol. 5., pp. 17–61 (1960)
2. Barabasi, A.-L.: Linked: How Everything is Connected to Everything Else and What It Means for Business, Science and Everyday Life. Plume Books, New York (2003)
3. Amaral, L. A. N, Scala, A. , Barthelemy, M., Stanley, H. E. : Classes of small-world network (2000)
4. Granovetter MS: The strength of weak ties. In: Am J Sociol 78., pp. 1360–1380. (1973)
5. Javier, M. H., Piet, V. M.: Classification of graph metrics (2011)
6. Newman, M., Barabasi, A.-L., Watts, D. J. (Eds.): The Structure and dynamics o networks. Princeton University Press (2006)
7. Amaral, L. A. N, Scala, A., Barthelemy, M., Stanley, H. E.: Classes of small-world network (2000)
8. Yamamoto Y, Yokoyama K: Common and Unique Network Dynamics in Football Games. In: PLoS ONE 6(12). (2011)
9. Pena, J. L., Touchette, H..: A network theory analysis of football strategies, in Sports Physics. Proc. 2012 Euromech Physics of Sports Conference, ed C. Clanet., pp. 517–528. (2012)
10. Trequattrini, R., Lombardi, R., Battista, M.: Network analysis and football team performance: a first application (2015)
11. Medina, P., Carrasco, S., Rogan, J., Montes, F., Meisel, J. D., Lemoine, P., Penas, C. L., Valdivia, J. A.: Is a social network approach relevant to football results? In: Chaos, Solitons and Fractals, vol. 142 (2021)
12. Cintia, P., Rinzivillo, S., Pappalardo, L.: A network-based approach to evaluate the performance of football teams. Workshop on Machine Learning and Data Mining for Sports Analytics, pp. 46–54., Porto, Portugal (2015)
13. Clemente, F. M., Couceiro, M. S., Martins, F. M. L., Mendes, R. S.: Using network metrics to investigate football team players' connections: A pilot study, Motriz, Rio Claro, vol. 20 n.3, pp. 262–271. (2014)
14. Clemente, F. M., Martins F. M. L., Kalamaras, D., Oliveira, J., Oliveira, P., Mendes, R. S.: The social network of Switzerland football team on FIFA World

Cup 2014, In Acta Kinesiologica 9, pp. 25–30. (2015)

15. Arriaza-Ardilesa, E., Martín-González, J.M, Zunigac, M. D., Sánchez-Floresd, J., de Saae, Y., García-Mansoe, J.M.: Applying graphs and complex networks to football metric interpretation. In: Human Movement Science 57, pp. 236–243. (2018)
16. Schroepf, B., Lames, M.: Career patterns in German football youth national teams – A longitudinal study. International Journal of Sports Science & Coaching 13(3), 05–414 (2018)
17. Carapinheira, A. et al.: Career Termination of Portuguese Elite Football Players: Comparison between the Last Three Decades. Sports 6(4), 155 (2018)
18. Monteiro, R. et al.: Identification of key career indicators in Portuguese football players. International Journal of Sports Science & Coaching 15(4), 533–541 (2020)
19. Barreira, J.: Age of Peak Performance of Elite Women's Soccer Players. International Journal of Sports Science 6(3), 121–124 (2016)
20. Dendir, S.: When Do Soccer Players Peak? A Note. Journal of Sports Analytics 2(2), 89–105 (2016)
21. Cripps, A. J., Hopper, L. S., Joyce, C.: Can coaches predict long-term career attainment outcomes in adolescent athletes? International Journal of Sports Science & Coaching 14(3), 324–328 (2019)
22. Schmid, M. J., Conzelmann, A., Zuber, C.: Patterns of achievement-motivated behavior and performance as predictors for future success in rowing: A person- oriented study. International Journal of Sports Science & Coaching 16(1), 101–109 (2021)
23. Zuber, C., Zibung, M., Conzelmann, A.: Motivational patterns as an instrument for predicting success in promising young football players. International Journal of Sports Science 33(2), 160–168 (2015)
24. Vroonen, R. et al.: Predicting the potential of professional soccer players. In: Davis, J., Kaytoue, M., Zimmermann, A. (eds.) ECML PKDD 2017, Proceedings of the 4th Workshop on Machine Learning and Data Mining for Sports Analytics, vol. 1971, pp. 1–10. Springer, Skopje (2017)
25. Raisi, O. A.: The Economics of Middle East's Football Transfers, https://www.sportsjournal.ae/the-economics-of-middle-easts-football-transfers/. Last accessed 21 June 2021
26. Christou, L.: The true extent of spiralling inflation in football's transfer market, https://www.verdict.co.uk/football-transfer-market-inflation/. Last accessed 21 June 2021

27. Recursive Feature Elimination, https://bookdown.org/max/FES/recursive-feature-elimination.html. Last accessed 15 Aug 2021

28 Beginner's Guide to Feature Selection in Python, https://www.datacamp.com/community/tutorials/feature-selection-python. Last accessed 15 Aug 2021

29 Backward Elimination for Feature Selection in Machine Learning, https://towardsdatascience.com/backward-elimination-for-feature-selection-in-machine-learning-c6a3a8f8cef4. Last Accessed 20 Aug 2021

30 ML Extra Tree Classifier for Feature Selection, https://www.geeksforgeeks.org/ml-extra-tree-classifier-for-feature-selection. Last Accessed: 21 Aug 2021

31. Futhead website, https://www.futhead.com. Last accessed 25 Oct 2021

32 Premier League official website, https://www.premierleague.com/players. Last accessed 20 Aug 2021

33. David, E., Jon, K.: Networks, Crowds, and Markets: Reasoning about a Highly Connected World. In: Cambridge University Press, pp. 48–50. (2010)