The Alan Turing Institute

Data Study Group Final Report: PlayerLens

10 – 14 December 2018

Player Pathways: Understanding career paths that deliver success for professional football players and clubs

https://doi.org/10.5281/zenodo.3558253

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1

Contents

1 Executive summary **2**
- 1.1 Challenge overview . . . 2
- 1.2 Data overview . . . 3
- 1.3 Main objectives . . . 3
- 1.4 Approach . . . 4
- 1.5 Main conclusions . . . 5
- 1.6 Limitations, recommendations and future work . . . 6

2 Data overview **6**
- 2.1 Dataset description . . . 6
- 2.2 Data quality issues . . . 7

3 Defining 'Success' **8**
- 3.1 Potential proxy measures of individual player success . . . 8
- 3.2 Potential proxy measures of club success . . . 13
- 3.3 Potential novel models for measuring individual player success . . . 13

4 Experiments **14**
- 4.1 Identifying determinants and indicators of individual player success . . . 14
- 4.2 Identifying determinants and indicators of club success . . . 35

5 Future work and possible research avenues **46**

6 Team members **46**

7 References **48**

1 Executive summary

1.1 Challenge overview

In football, as in any profession, it is understood that careful career decisions could be critical in dictating one's potential for professional success. Similarly, the well-considered formation of a cohesive, complimentary team is a critical ingredient in determining club performance. But what are the journeys that deliver such individual success, and how might one navigate the optimal path to maximise their career trajectory and achieve their desired outcomes? Moreover, is there a model combination of career histories and player backgrounds that can help distinguish between teams that have experienced different levels of success?

This data challenge considers whether it is possible to measure different definitions of success and identify key determinants for achieving each outcome. The study aims to derive preliminary insights for advising players and clubs in crucial decision making processes, primarily in relation to those made in the arena in which football players are available for transfer to clubs - the so called 'transfer market'.

The challenge focuses on the career paths and other key attributes of players, and seeks to identify the journeys and characteristics which deliver most success to both individual players and clubs. This is of particular interest to PlayerLens, the challenge owner, who work closely with both players and teams at the point of considering transfer opportunities.

Methods for understanding indicators of performance outcome are presented. These could be used to gain insight into questions such as, 'Is there an archetypal career path to the highest level of football?', 'Is there a recommended age at which players should change club?', 'Is it beneficial to experience lower league football?' and 'Is there an optimal composition of players that a club should aim for, with regards to player pathways?'.

1.2 Data overview

A unique database of detailed career records relating to more than 40,000 professional football players was provided by Opta, the world leading football data provider. This contained demographic information, performance statistics and career history (historic transfer details) for each player, where the term “appearances” below identifies occasions in which a player had playing time during a game. Data includes:

- player date of birth
- player position
- current career status (active/inactive)
- number of first team appearances for each club in career history
- number of appearances for national team
- duration of stay at each club in career history
- transfer type, for each transfer in career (free transfer, transfer, loan, back from loan and free agent)
- a rich set of game stats and in-game data (e.g. count of goals, assists, cards received, appearances and minutes played)

The transfer is a ‘free transfer’ when the player transferring to another team does not include a fee. A loan is when a player temporarily play for a club other than the one he is currently contracted to.

1.3 Main objectives

The overarching aim of the challenge was to develop methods for gaining insights that could serve as a basis for informing more successful transfer decisions, both from the perspectives of individual players and clubs.

The main objectives of this study were therefore identified as:

1. From a player perspective: To gain a better understanding of how transfer and loan decisions might have an effect on a player’s performance outcomes and on their overall career trajectory.

2. From a club perspective: To identify whether particular compositions of individual player pathways act as a determinant or indicator of a team's success, and consider how this information might be used to advise different clubs regarding future transfer decisions.

Alongside these core transfer related objectives, the data provided also offered scope to analyse other physical attributes and key performance characteristics of a player to assess whether there might be additional, non-transfer related determinants of success. This avenue of exploration therefore prompted a third core objective of the study:

3. Identify key player attributes which might contribute to performance/career success.

1.4 Approach

Several distinct streams of proof-of-concept work were undertaken during the week to achieve the objectives outlined in Section 1.3 and better understand key determinants and indicators of player and club success. This work has been summarised in this document into two main sections. The first in Section 4.1 is on the key features of successful player careers. The second in Section 4.2 is an analysis of team composition leading to success looking at team as an aggregate of player careers.

Section 4.1 contains three distinct player-level analyses; first a random forest method, next an application of decision trees and finally an exploration of elastic nets, event trees and chain event graphs (CEGs). Section 4.2 contains a novel application of sequence state analysis for gaining insight into club-level player composition and a discussion of the model composition of 'successful' clubs. These analyses will be frequently referred to as 'the methods' throughout this document.

However, before any of this work could be undertaken, it was necessary to first define the meaning of the term 'success' (in this case, the dependent variable for all of the analysis) in the context of this study, to enable the analysis of its determinants, as desired. Since no single, agreed upon measure for success exists, defining the terminology from

both an individual and player perspective constituted the first challenge.

For the purpose of this study (and due to time limitations), a range of simple but informative measures were employed in the development of the above methods, whilst a more sophisticated approach was developed in parallel. These considerations, and the development of this more sophisticated 'success' measure, are also outlined in this document as an additional third core section (Section 3). This measure can serve as a started point for methods presented in any future iterations of the work.

1.5 Main conclusions

In summary, three data-driven approaches for assessing player and club attributes relative to a variety of performance measures have been developed and preliminary observations were made. In each case, transfer decisions have been identified as important determinants of success for both players and clubs, as expected, with loans having been highlighted as potentially detrimental to the success of a player's career. Running the decision trees explicitly found that players either returning from loan, or who are involved in a free transfer, are less likely to be successful based on the measures of success analysed (Section 4.1.2). At the team level, it was discovered that the top-flight squad (Manchester City) comprised of fewer players with loan experience than the mid-table squad (West Ham United), but more experience at other European clubs (Section 4.2). Taken together, these findings suggest that, under the assumptions made in our analysis (e.g., our quantification of 'success'), players with foreign football experience, but minimal loan experience, may be most likely to become elite footballers, and thus form successful teams.

It is important to note that all results are presented under the caveat that these outputs have been developed based on simple and potentially biased measures of success, and further tests with more appropriate measures should be carried out as a check that the insights derived are not an artefact of the choice of success measure (or of another

assumption).

1.6 Limitations, recommendations and future work

As mentioned previously, the tight time constraints of the challenge led to the pragmatic adoption of a number of success measures which were readily available and easy to employ, regardless of their potential limitations (outlined in Section 3). The development and adoption of increasingly sophisticated model of success, such as that detailed in Section 3.1.2, would be recommended in any extension of this work to increase the utility of the final results derived. Possible future work includes investigating such an improved success measure when applying the methodologies proposed in this document to evaluate player success.

Second, extending the analysis of the club level exploration would also be recommended as the short timeframe also limited the potential for more detailed analysis of the optimal composition of player pathways exhibited by a 'successful' club. For example, the club pathway study (Section 4.2) does not, at present, contain a formalised analysis of the optimal ratio of homogeneity: diversity across players. Such an extension might greatly enhance the insights gained. Further work might also see an implementation of the distance matrix and/or clustering of successful teams by common pathways to construct an optimal framework.

2 Data overview

2.1 Dataset description

A database of records relating to 44,142 football players was provided. This detailed players who are currently playing, or have played, for any professional club or national team world-wide. The data was split into three related datasets, which when combined, provided a rich set of information relating to each player's performance and career history:

- **Player Information:** Contained player characteristics and other personal information, such as the player's name, date of birth, prominent foot, height, weight, nationality, and country of birth.
- **Statistics Information:** Contained aggregated match statistics for each player, including number of assists, yellow cards, and time played, grouped by tournament (both cup and league) and level of play (both club and international).
- **Membership Information:** Contained information regarding the clubs at which a player has been a member, the type of transfer via which they had moved to the club, and the time spent at the given club (if not still a current member).

NB: The terms *variables*, *attributes* and *features* are used interchangeably throughout this document to refer to the information contained in the datasets relating to the players.

2.2 Data quality issues

Some records were incomplete, with some demographic information not provided for a number of players. Though this was not considered enough of an issue to halt the analysis for the proof-of-concept model, this should be investigated and rectified, or accounted for (if possible), if the work is to be replicated or extended in the future.[1]

[1]There was also a delay in the delivery of the data. Checks of the of the data initially provided highlighted some missing players. Subsequently, an update of the data, which contained a complete set of information for the final count of 44,142 players, was provided on day 3 of the challenge. The original data structure was retained in the update, which allowed the methods presented in this document to be developed on the original datasets in the meantime, but all results presented were derived from the final, complete data transfer.

3 Defining ‘Success’

In order to identify determinants and indicators of successful players and clubs, the overarching aim of the challenge, it was first necessary to define ‘success’. Since no single, agreed upon, objective measure is already available, this presented a key challenge.

Several potential measures were suggested and discussed within the group. The practical and theoretical pros and cons of each measure suggested were evaluated against the technical ease and speed with which it could be employed within the challenge timeframe. Some suggestions were based on pre-existing proxy measures already available for use, while others considered the potential for developing new, novel success measures. The main considerations are summarised in Section 3.1.

We reiterate that quantifying success is an open question in and of itself. Several of the measures suggested in this report were pragmatically adopted and tested to enable the development of the methods, but there is ample room for further improvement.

3.1 Potential proxy measures of individual player success

3.1.1 Basic proxies proposed

Several potential success measures were discussed, each with different benefits and limitations. The benefits were largely limited to ease, simplicity and ability to calculate from the available data, while limitations surrounded inherent bias in the calculation or the input data selected. These are discussed in the table below, including potential solutions for reducing such bias or considering additional real-world complexity into the calculation of the model to “improve” the outcome.

Potential proxy	Comments
A player's total playing time (in any specified competition type).	Could be improved by appropriately accounting for length of career, or age of player and/or level/type of fixture. ***Basic measure used in Section 4.1.1.***
A player's total goals scored (in any specified competition type).	Doesn't normalise by number of games played, or account for likelihood of scoring in different positions. ***Basic measure used in Section 4.1.1.***
A player's total appearances (in any specified competition type).	Could be improved by appropriately accounting for length of career, or age of player and/or level/type of fixture. ***Basic measure used in Section 4.1.1.***
A score based on the rank of each player's current club.	Requires data on club rankings, issues detailed below.
Appearance in the Guardian's ranking of the top 100 footballers.	***This measure was used in the analysis carried out in Section 4.1.2., where it will be discussed in more detail.***

3.1.2 Developing new complex success measures

Motivation

While the methods outlined in Section 4.1 and Section 4.2 have been developed based on a range of measures for success (outlined in Section 3.1.1, above), better, more appropriate measures of success could improve all of the methodologies presented. Though no measure can conclusively capture all aspects that define player quality, a first attempt at a potential composite measure which aims to capture multiple dimensions of player success is proposed here. Unfortunately, obtaining all of the data we needed was very time consuming, so the measure

presented has to be considered as a proposal for future work, and as a way to check the obtained results.

Approach

1. Defining success

Feedback from PlayerLENS indicated that the total minutes played by a player was a useful indicator of player success. Nonetheless, we propose that this should also consider the team in which the individual plays; a minute played at Real Madrid (a top tier club), for example, is more indicative of success than a minute played at West Ham United. Given this, we propose a success measure which captures these factors.

Additionally, our proposed method incorporates the player's career length to date, to ensure that young players are not penalised for their career age. Specifically, we consider the last five years for each player, or the total number of seasons played, whichever is lower (i.e., *N*=*min*{5, number of seasons played}). The proposed measure is:

$$y_i = \frac{\sum_{j=1}^{N} w_j x_{i,j} u_{i,j}}{\sum_{j=1}^{N} w_j} \tag{1}$$

where:

- y_i is the success measure of a given player, i.
- $x_{i,j}$ is the normalised number of minutes played by player i during season j. This value is derived by dividing the number of minutes played by player i during season j by the total number of minutes played by his teammates.
- $u_{i,j}$ is a measure of club success. This is derived from the number of points a club receives using the UEFA club coefficients table. To this value, we assign the number of points earned at the end of season j. If a club is omitted from the UEFA rankings (i.e. the clubs that have not played in a UEFA tournament during season j), the values are homogeneously distributed according to the club's final position in it's national league. The team that is top of the league will receive a score of 1 and the team at the bottom of the league will receive a score of 0, all other scores will be scaled between 0 and 1.

- w_j is a weighting parameter.

The measure can account for several characteristics:

- Greater importance is placed on recent performances of the player. As such, the weights $w_1, \ldots, w_N$ should be a decreasing sequence.
- Playing time is re-scaled according to the number of matches played by a given club. This is to distinguish between the number of minutes that *could* have been played by a given player, relative to the number of minutes that *were* played. This is important in controlling for differing season lengths across leagues and cup tournaments.
- Minutes played at an elite club are deemed more relevant to player success than minutes played at average teams, as such, our methodology adjusts minutes according to UEFA rankings.
- Our method also allows for comparison of players with varying career lengths, ensuring that players with less than 5 years of career experience are not penalised for their inexperience.

2. Defining and adding weights

The weights can be defined on the basis of three different approaches, each of them with some advantages and some limitations:

1. In the present report we arbitrarily set the weights. This method does not require a training set and is somewhat subjective.
2. An alternative approach could involve fitting a linear model on a training set. Here, the response variable is a proxy of the success measure. The explanatory variables are defined as $z_j = x_j u_j$ and obtaining the estimates of the parameters $w_1, \ldots, w_N$. This approach is data-driven, but requires a large array of variables that can be considered relevant proxy for the value y. One possibility would be to consider the Guardian's list of the top 100 footballers as a proxy for career success. From these player rankings, we could infer player quality and refine our weights.
3. A second alternative could involve estimating the parameters $w_1, \ldots, w_N$ under a constraint that allows the estimation of just one

parameter ξ, i.e. considering $w_j = e^{-\xi(N-j)}$. From a technical point of view, it means to find

$$\hat{\xi} : min_{\xi} \sum_{i=1}^{n} \left(y_i - \frac{\sum_{j=1}^{N_i} e^{-\xi(N_i-j)} x_{i,j} u_{i,j}}{\sum_{j=1}^{N_i} e^{-\xi(N_i-j)}} \right)^2 ,$$

with n sample size and N_i the value of N for player i. As with approach (2), this method requires a training set. Figure 16 shows the weights according to different values of ξ.

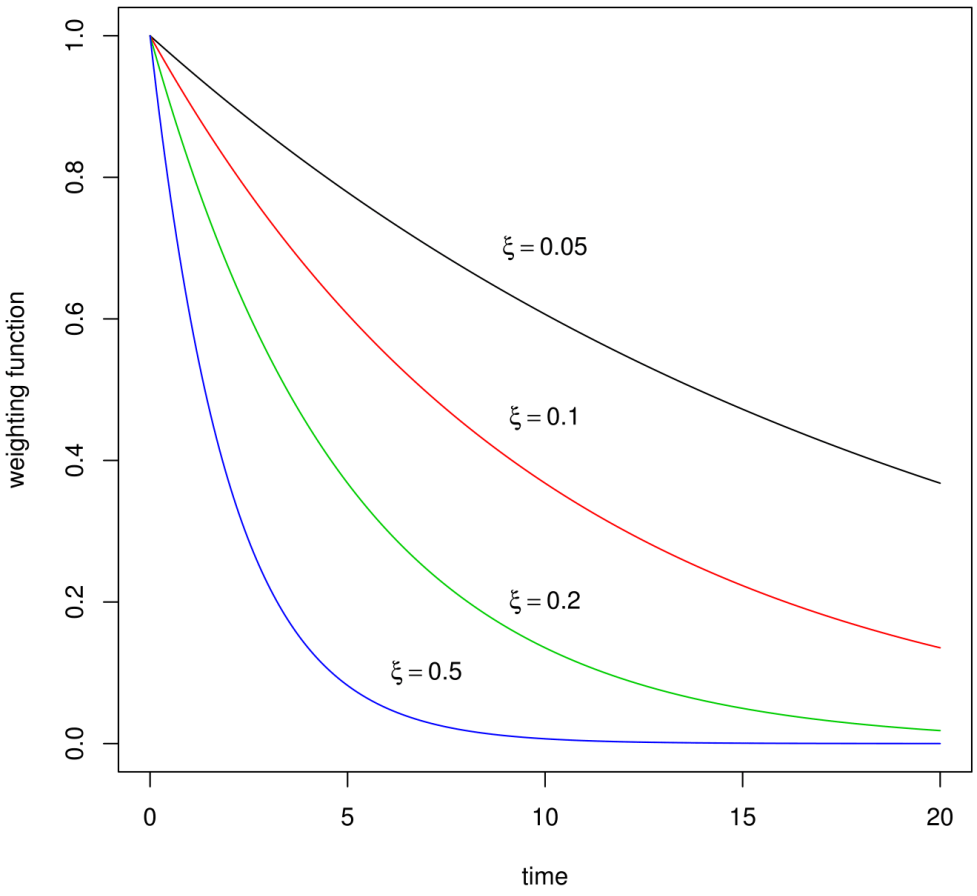

Figure 1: Weighting function for the different values of ξ.

Discussion and future work

Future work might also consider assigning weights to different competitions. For example, a minute played in the UEFA Champions League could be viewed as more valuable than a minute played in the FA cup. We might consider multiplying the number of minutes played in the UEFA Champions League by a factor of 2, multiplying the number of minutes played in Europa League by 1.5 etc., on the basis of tournament prestige.

A further extension might also consider different values for u_j that allow for a comparison across continents or account for player position. For example, a goalkeeper is less likely to be substituted than a midfielder.

3.2 Potential proxy measures of club success

Scraping and using existing rankings available online for each football club was considered, however, this was deemed to be infeasible given the time constraints. Additionally, only country specific rankings were found, which would present difficulty when comparing teams across different leagues in different regions.

3.3 Potential novel models for measuring individual player success

Three models for deriving new success measures from the data provided were proposed:

Model 1: A scaled measure of the total minutes played by the player (in any club match, including domestic and international games) adjusted according to the rank of the club to which the player belongs.

This would require data on club rankings, the issues of which are outlined above, so was therefore not progressed.

Model 2: A normalised total of the number of minutes played by the individual player for a given team, calculated as the number of minutes played within a team as a proportion of the number of *possible* minutes that the player could have played for the team.

Because this measure is derived from data drawn directly from our dataset, it is a data-driven measure of player success which does not discriminate against players who are earlier in their career.

This measure was employed in the analysis carried out in Section 4.1.3., where the calculation and application is outlined in more detail.

Model 3: A complex composite measure combining the concepts in both Method 1 and Method 2 (above), based on total minutes played, a rank of the club at which the player belongs and the current length of the player's career.

Though this model was developed throughout the week, and was thus not ready to be applied in any of the analysis which was run simultaneously, it presents a potential improvement to the less developed methods which were more readily available and therefore used. In the future, this model could be adopted as an alternative measure of success, however, there is still scope to extend the model further.

Details of the development of the measure, and the scope for improvements, are presented in Section 3.1.2.

4 Experiments

4.1 Identifying determinants and indicators of individual player success

4.1.1 Method 1: Random Forest

This initial method is employed to identify the features which have the greatest influence on player performance. In ranking the features by

importance, we can use these to predict player success based on his feature values.

Approach

Random forest will be applied to rank and compare the importance of various player attributes in determining ‘success’, where success is defined by a given set of possible performance indicators; ‘number of appearances’, ‘number of goals’ and ‘minutes played’.

In the random forest, each node in the decision trees is a value of a single feature that is designed to split its descendent nodes so that eventually similar response variables (the performance indicator) end up in the same pool.

One of the important measurement is Mean Decrease Accuracy (MDA). MDA measures the loss of accuracy. By randomly permuting the value of a single feature that matches the distribution of the samples, the accuracy of the tree given the response variable is computed. By repeating this process for each feature, the mean loss of accuracy of each feature to the whole forest can be obtained, which is used to rank the features accordingly.

Data and data preparation

The data is first separated by the type of competitions the player has played, ranging from the domestic league, domestic cup, domestic super cup, international cup and international super cup. Within each subset, some players may play multiple games at the same level of competition.

Performance indicators such as goals, appearances and minutes the players played were aggregated for all the matches the players played in each competition type. These were further combined with the player’s demographic information, to produce a record for each player including their basic information and match performances. An additional feature of age is also inferred. This leads to 5 datasets with the following number of players in each:

Tournament	*Player count*
Domestic league	32,441
Domestic cup	28,121
Domestic super cup	3,918
International cup	18,705
International super cup	771

Analysis, results and conclusion

Considering initially just the minutes played in all super cup competitions as the performance indicator (or 'success measure'), 6 features are identified as important. These are highlighted in green in Figure 1, where the whiskers of the box plot indicate the 5% and 95% percentiles.

The variables coloured in yellow and red are rejected due to their relatively low values. This has been decided by z-scores as an additional measurement to take the accuracy loss fluctuation into account. Features with significantly high z-scores are tagged as acceptable, important features. Unsurprisingly, age appears to be the most influential factor, meaning older players tend to play more time in these competitions.

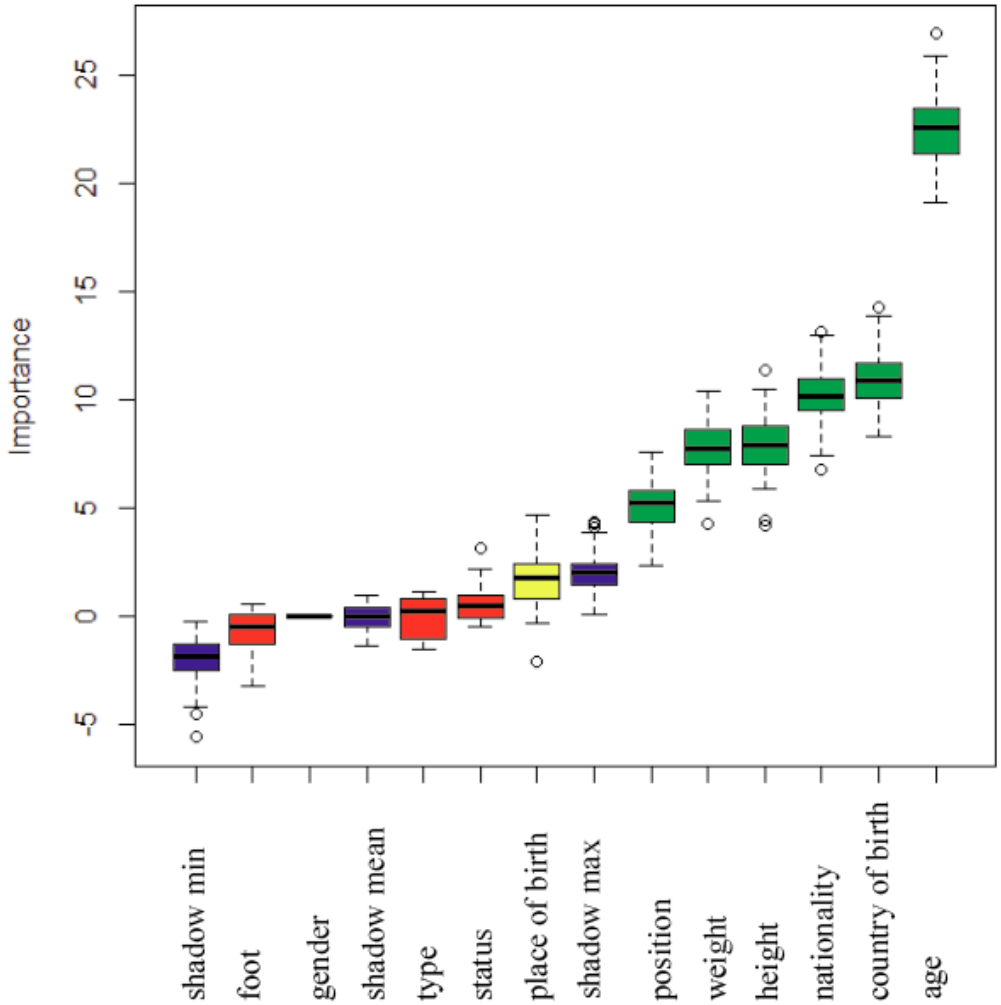

Figure 2: Exploring feature importance with Random Forest; from right to left, the most to least important feature when considering the minutes played in all super cup competitions as the success measure.

Comparing feature rankings based on different performance indicators within the domestic league, it appears that most features rank of similar importance for ‘minutes played’ and ‘number of appearances’. But taking ‘number of goals’ as the measure of player success, the player’s position becomes the most important variable (Figure 2).

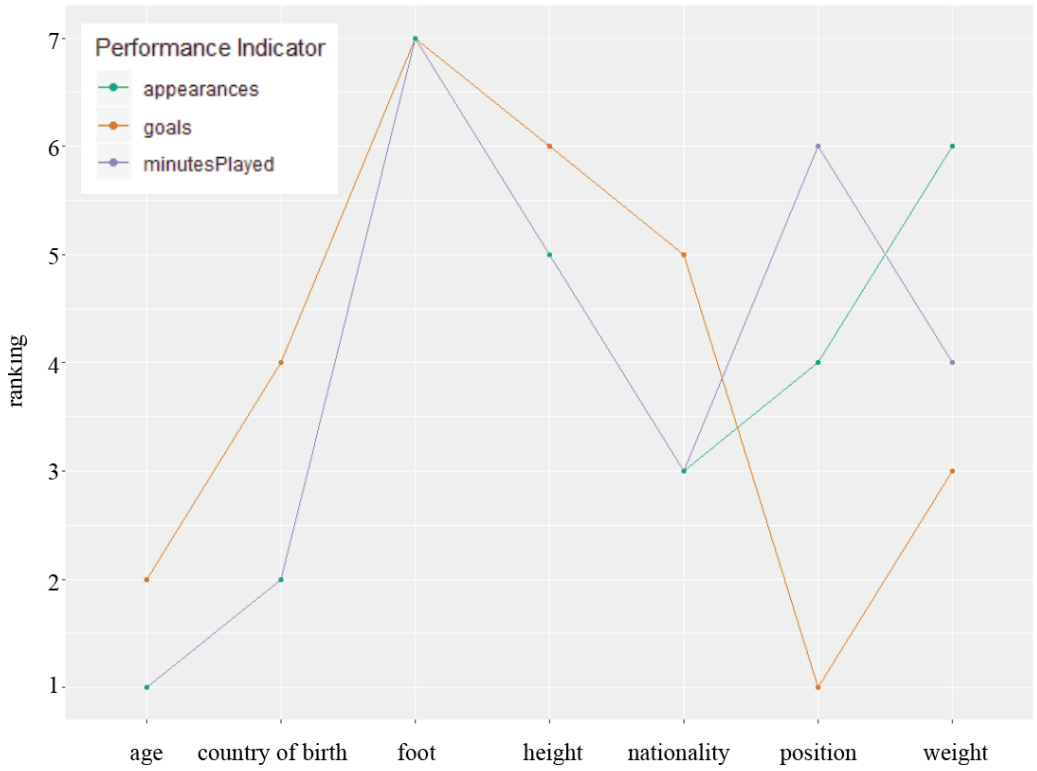

Figure 3: Comparison of feature rankings with different performance indicators (success measures); from bottom to top, the most to the least important feature.

Similarly, a single performance indicator such as ‘minutes played’ can be considered, and the importance of features can be compared for minutes played across different competition types (Figure 3). Now, the trends of the rankings are all fairly similar, with age appearing as the most important variable and predominant foot as the least.

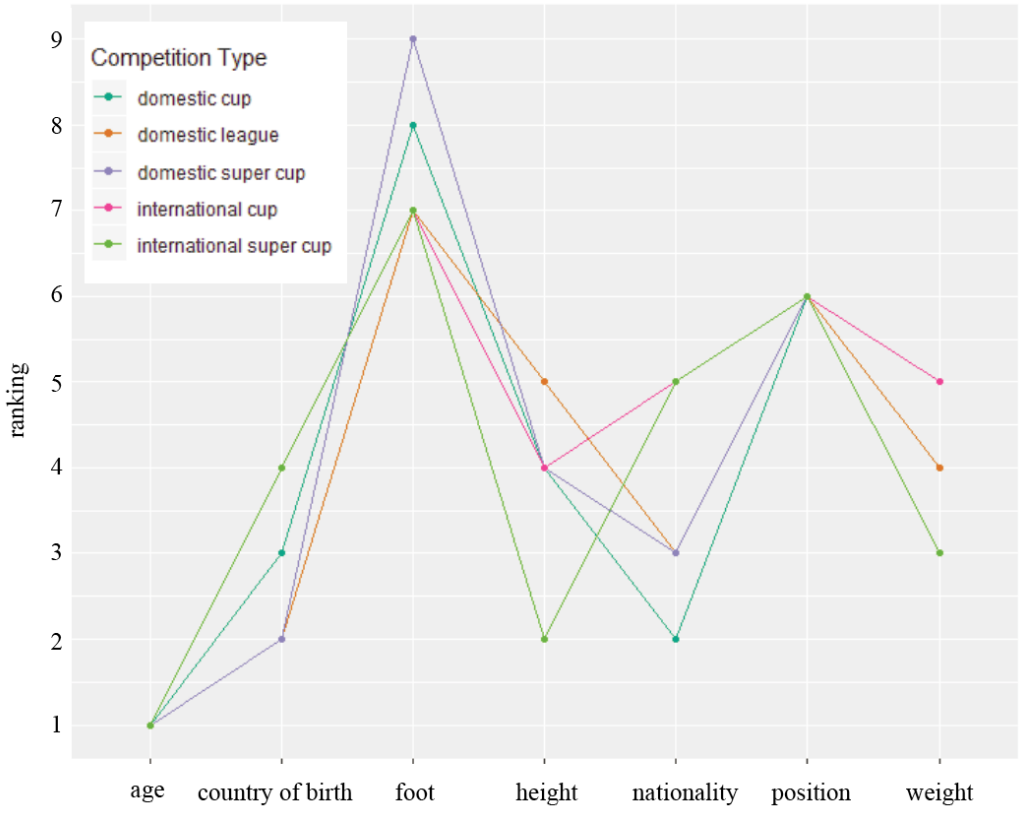

Figure 4: Feature rankings with different competition types; from bottom to top, the most to the least important feature.

Discussion: Limitations and future exploration

Just considering this limited set of attributes and performance indicators, the results seem fairly obvious and reasonable. The utility of these results is limited based on the crude nature of the performance indicators measured. As previously highlighted, measures of success such as 'minutes played' do not account for length of career, so it seems intuitive that older players will have played longer, and as such, age would be identified as the most significant feature to having played more.

However, the methodology presented here simply demonstrates a proof-of-concept. The same process could be re-applied to both a wider set of attributes, such as which club the player belongs to and the tier of the

club, or more importantly, to more appropriate or complex performance indicators (measures of success). In doing so, more useful insights may be derived.

4.1.2 Method 2: Decision Trees

Approach

This second analysis is interested in exploring the behaviour of Positive Deviants (PDs), players who outperform their peers in a given domain to identify the features which differentiate these successful players.

Here, the measure of success is derived from the Guardian's ranking of the 100 best footballers, decided by a panel of 169 experts from 63 nations who were asked to identify the top male footballers in 2017 (https://www.theguardian.com/football/ng-interactive/2017/dec/19/the-100-best-footballers-in-the-world-2017-interactive). The players included in this list are classified as elite players (PDs), with all others classified as non-elite players (non-PDs).

An exploratory comparison of player attributes for elite and non-elite players suggests that there is some relationship between these attributes and success (Figure 4). For example, the results show left- or both-footed players to have a higher likelihood of being an elite player, while players with free transfers and those returning from loans are less likely to be elite.

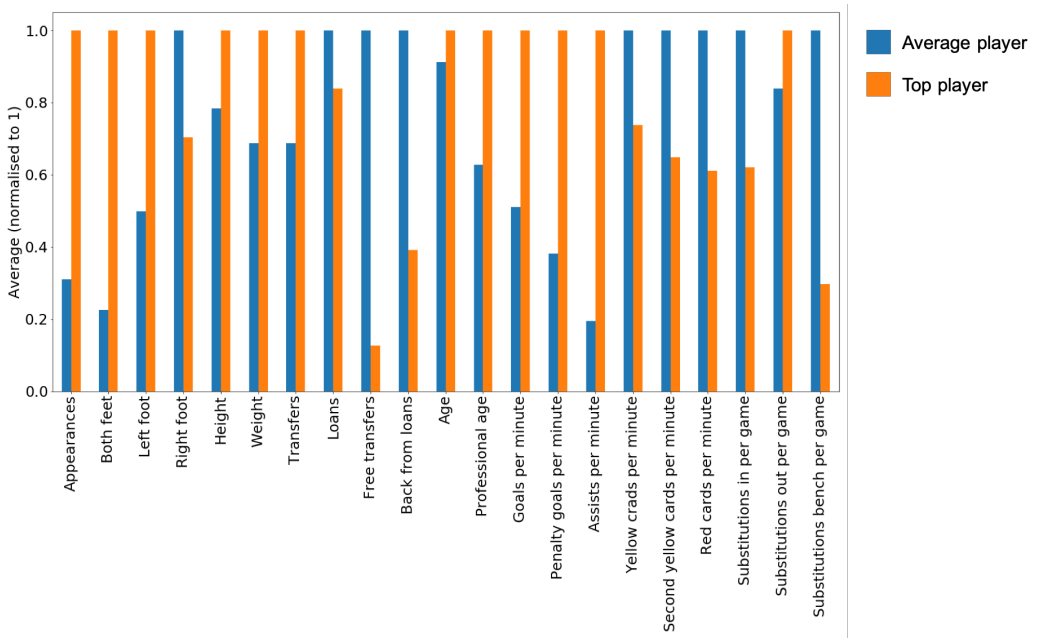

Figure 5: Comparison of the career statistics of top players vs. average players, using the Guardians ranking of the 100 best footballers as the success measure.

Decision trees were used to explore these relationships further and better understand the attributes of elite players.

A decision tree is a tree-like structure in which:

- each *node* represents a variable
- each *branch* represents the values of the variable that lead to the leaf, and
- each *leaf* represents a class label.

A segment is a combination of nodes and branches leading to the desired leaf (i.e., PD class). For example, if we have a segment **s1** having situational variables **v1** (variables you can't control like your age, height and weight) and non situational variables **v2** (variables you can control like transfers and clubs to join) the situational variables could be used to identify potential target segments and the non situational variables could

be used as segment specific recommendations.

Decision trees benefit from being intuitive and relatively easy to interpret. The outcomes could be a categorical variable (e.g., 1 = elite player; 0 = non-elite player), but similarly, it can be used for continuous variables such as the number of goals scored per minute, or the number of minutes played during a season.

Here, we explore different decision trees for a range of feature selections and target variables.

Data and data preparation

The following career information and physical attributes were employed in the application of this method:

Career Characteristics:

- Count of transfers
- Count of loans
- Count of free transfers
- Count of returns from loans
- Count of substitutions in
- Count of substitutions out
- Count of games spent on subs bench

Physical Characteristics

- Prominent foot: left, right or both
- Weight (kg)
- Height (cm)

Further variables were also derived from the data available.

- Age
- Professional age (years played)
- Goals per minute
- Assists per minute
- Penalty goals per minute
- Yellow cards per minute
- Second yellow cards per minute
- Red cards per minute

In order to create the success measure to be used, the Guardian list of

player rankings was also imported. This was used to derive a PD value of:

- 1 = Present in elite list ($N = 100$)
- 0 = Absent from the elite list ($N > 40,000$)

Analysis, results and conclusion

1. All players

In figure 5, we use all available features for all players to identify the characteristics that best predict whether a player is absent or present in the Guardian's list of elite players.

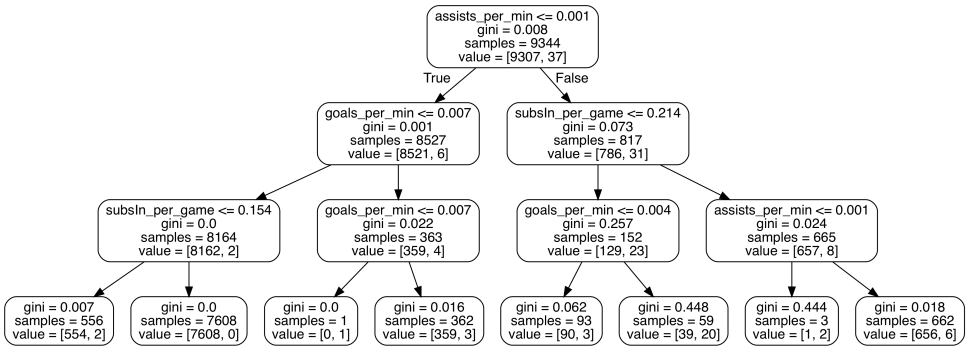

Figure 6: Decision tree to predict inclusion or exclusion in the Guardian's list of elite players.

Here, different criteria are automatically applied to split the full dataset and identify combinations of features that best predict elite or non-elite status. The *value* parameter describes the composition of each of the subsets.

We can see that a player at a later career stage (at least 9.5 years of professional football) with a high rate of assist is most likely to fall into the category of highly successful players.

2. Attacking Players

This methodology can also be applied to specific groups within the data. Here we focus exclusively on attacking players. The results of this analysis can be seen in Figure 6.

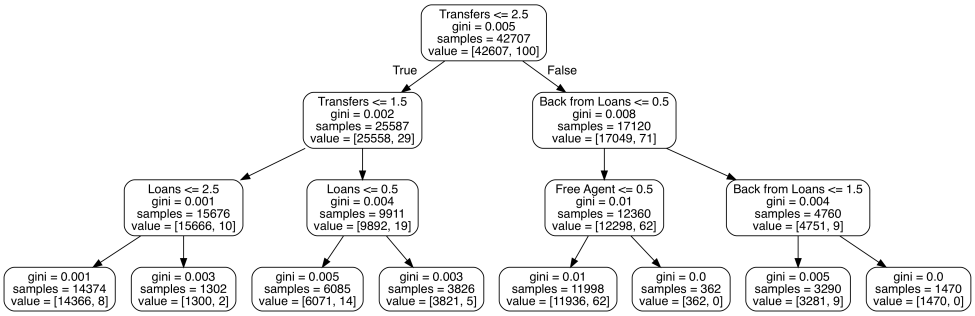

Figure 7: Decision tree to predict an attacking player's inclusion or exclusion in the Guardian's list of elite players.

3. Goals scored

If we wish to predict a particular outcome or another measure of success, like the number of goals scored by a player, the decision tree methodology can be used with a continuous target variable, such as assists or goals per hour. Figure 7 shows a decision tree for a given player's rate of scoring (goals per hour).

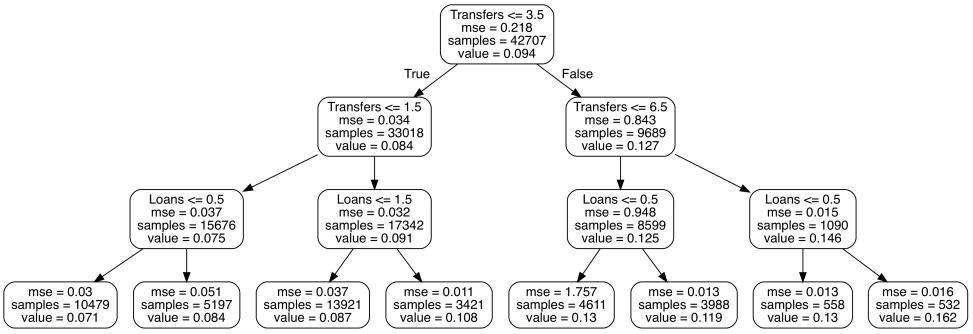

Figure 8: Decision tree to predict the number of goals scored per hour by a given player.

When the outcome variable is the average number of goals scored per minute by the players in each subgroup, we find that a high number of transfers is a strong indicator of the number of goals scored per hour.

4. Transfer history

Finally, in other use cases we might only care about a subset of the variables, such as the transfer history of a player. In Figure 8 we predict elite/non-elite status as a function of a player's transfer history, to demonstrate the effects of particular (types of) career paths on predicting the success measure.

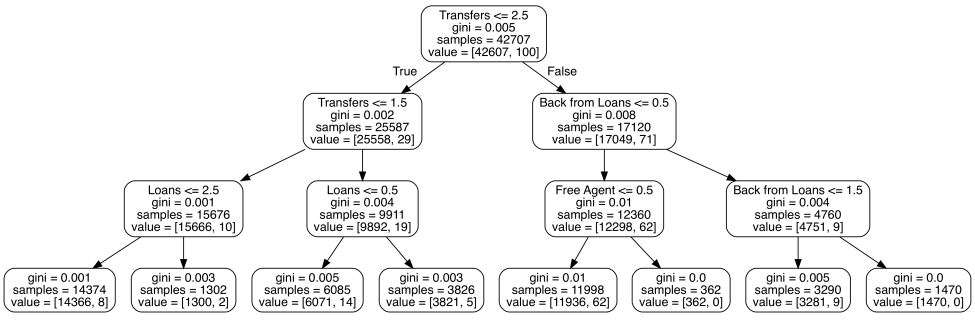

Figure 9: Decision tree to predict inclusion or exclusion in the Guardian's list of elite players, based on transfer history.

This shows that elite players typically meet one of the following criteria:

- Has a minimum of 3 transfers.
- Has few returns from loan.
- Has no spells as a free agent.

Discussion: Limitations and future exploration

Here we provide a proof-of-concept to identify the characteristics of PDs and non-PDs, using the Guardian's list of elite football players. Future research should use a more objective measure, which can be generalised to players outside of the top 100 players. The measure developed in Section 3.1.2, for example, would be recommended. From this, players that are 2SD from the mean or median could be identified as "above average" or "below average" players.

It is also possible to base the non-average cut off using the PlayerLens dataset. For example, we might classify players with more than 450 minutes of experience in the UEFA Champions League as elite, for a simpler alternative.

Additionally, in the future, one might consider re-categorising some variables. For example, grouping nationalities into fewer bins to allow for simplified comparison.

4.1.3 Method 3: Elastic Nets, Event Trees and Chain Event Graphs (CEGs)

The analysis outlined in this section seeks to identify the features and career pathways of successful football players by considering the wide range of attribute variables available. The regularised regression method Elastic Net was employed in this section of the analysis, alongside the probabilistic graphical models Event Trees and Chain Event Graphs.

Approach

Since using all of the possible variables would be computationally expensive, the first step involves an identification of the features which

are most relevant for defining player or team success. A common approach is to use a regression method to estimates the relationship between variables to identify the variables which affect the success measure, and to what extent. In this case, the regularised regression method elastic net was used. Regularisation prevents overfitting to the data by adding a penalty for increased complexity of the model. This method linearly combines the penalties of the lasso and ridge methods. Like other regression methods, the elastic net also provides coefficients associated with each variable which tell us how influential each variable is in determining the success measure. If the influence is non-zero, the result indicates whether the influence is positive or negative.

Once the important features are extracted from the elastic net, the next step is to use an event tree and a chain event graph (CEG) (Smith & Anderson, 2008) to explore the pathways that lead to success based on the chosen success measure. The event tree is a graphical representation of a process. This helps us to identify the potential pathways for a player. This is an efficient and quick way to visualise the data and all of the pathways contained in the data. A CEG is subsequently built from the event tree, clustering entire or partial pathways. The result of this process reveals the different pathways along which the same success measure can be reached. The CEG is fit under a Bayesian setting and a weakly informative is set prior. For clustering the nodes of the event tree (and thereby clustering entire or partial pathways), a Bayesian hierarchical clustering algorithm called Agglomerative Hierarchical Clustering (AHC) is used. This merges nodes at each step that give the highest improvement to the log marginal score.

This work can be extended for a more detailed analysis of the pathways, for information see the future works section.

The success measure detailed in Model 2 in Section 3.3 is employed in this analysis. This measure considers the proportion of minutes that the player was active for out of the total *possible* minutes that they could have played at any club. Time spent playing in any game in the ‘UEFA Champions League’, ‘UEFA Europa League’, or ‘UEFA Europa League Play-offs’ was inflated to weight these games as more important. Weightings of 2, 1.5, and 1.5 were used, respectively. This success measure will be referred to as the ‘club success measure’.

Data selection and preparation

The following variables from the 'Player Output' and 'Stat Output' datasets were merged to be employed in the application of this method:

Player Output:

- Nationality
- Country of birth
- Date of birth
- Height (cm)
- Weight (kg)
- Type
- Status (active/inactive)
- Prominent foot
- Player position

Stat Output: (counts)

- Goals
- Assists
- Penalty goals
- Appearances
- Yellow cards
- Second yellow cards
- Red cards
- Substitute in
- Substitute out
- Subs On bench
- Minutes played

Membership Output:

- Club name

These variables were selected to facilitate analysis aimed at answering specific research questions, following checks for multicollinearity.

A 'country of birth' variable was considered, however, comparisons with the 'nationality' variable revealed discrepancies in 5% of the records, so the 'country of birth' variable was removed based on domain expert recommendations.

Finally, players who have no associated statistics included in the data have been removed.

Once the variables had been selected, the data was cleaned and prepared for analysis.

1. The date of birth was converted to current age.
2. Where demographic information was missing (discussed in Section 2.2), averages were used (using the mode for categorical variables and the mean for numerical variables).
3. The Chi Merge algorithm was used to create bins (intervals) for age, height and weight variables.

Analysis, results and conclusion

Under normal circumstances, a leave-one-out cross-validation would be adopted when carrying out this analysis. Cross-validation involves running the model multiple times times with a proportion of data out of the sample and testing on the removed data to measure accuracy. Leave-one-out cross-validation involves removing one data point for the dataset and training the model on the remaining data points before obtaining a prediction for the removed data point, using the newly trained model. This process enables the training of the model just once on the whole dataset whilst simultaneously obtaining immediate predictions. Due to time constraints and large quantity of data involved with this study, this step was omitted on this occasion. Instead, the only evaluation metric is the R^2, obtained from a simple linear regression of the observed and predicted outputs.

Table 1 below shows the variables which have been identified by the elastic net as relevant to player success. The magnitude of positive or negative influence is also noted in the *Coefficient* column. The actual size of the coefficients are only important relative to each other, it is therefore not necessarily a weakness to have small coefficients. All other variables, which are not included in the table, made a trivial contribution to our model.

Variable	Coefficient
Goals	8.44×10^{-5}
Appearances	8.33×10^{-5}
Yellow cards	5.08×10^{-5}
Substitute out	2.35×10^{-5}
Substitute in	-1.56×10^{-5}
Subs on bench	-7.55×10^{-5}
Assists	-7.78×10^{-4}

Table 1: Variables influencing player success.

The variables listed in Table 1 were therefore used as inputs into the elastic net model to predict success. A simple linear regression on the observed and predicted success measures was run to gain an understanding of how well the model performed. The R^2 value for this

model was about 0 .707 which is around 70.7%, indicating that 70.7% of the variability in the success measure within the model was explained by these influential variables. This is a good R^2 value, however, an inspection of the players which it ranks as top players, we empirically conclude that this success measure is perhaps not very strong. The teams that are said to be most successful are unknown and play with unsuccessful teams.

One possible explanation for this weakness could be the inclusion of too many dummy variables into the model (for instance, for the nationality variable there were approximately 180 different nationalities and a dummy column for each nationality). Additionally, many of the input variables used were directly involved in the computation of the success measure, which could have influenced their ability to explain the variability of the measure. The lack of success could also be attributed to the failure to cross-validate in the execution of the elastic net. Finally, it is also possible that the success measure itself is not objectively poor, but the other variables in the dataset do not influence it enough to account for majority of its variability. Since a success measure which is able to pick out the important features from those contained in the dataset is desireable, a measure such as this one, which is built using these same variables, is not suitable to be used in this approach.

Exploration reveals that it is the success measure itself which is not representative and is identifying unknown players as the most successful. Therefore, a step back will be taken to consider just the raw, non-normalised, count of minutes played as a measure of success instead. In doing so, more well-known players are once again identified as 'successful', suggesting this to be a more appropriate success measure to be employed in the chain event graph after all.

This analysis is now re-run based on this new measure of success. The results now identify more well-known players as successful, which is more in line with the expectations of the domain experts, and thus is deemed to be a better model.

This method explained 98% of the variance (R^2 of .98). Since the aim is to identify players with high success scores, only the positively influential variables, which are 'Appearances', 'Yellow Cards' and 'Assists', are considered to be of interest. The remaining variables were either

non-influential or negatively influential.

A Event Tree and a Chain Event Graph (CEG) was employed at this stage, based upon these variables, in the order listed above, along with ‘minutes played’ (the success measure). If the process unfolds over time, the order of the variables is important, since a different ordering of the variables could lead to different interpretations. As these variables don’t describe a process that unfolds over time, the variable ordering is trivial, though there is nonetheless a partial order; number of appearances directly influences the potential number of yellow cards and assists, as such, this must be the first variable in the ordering. The outcome variable, time played in minutes, is our final variable. There is no clear order between yellow cards and assists, so an ordering was fixed at random.

While the visualisation of an event tree is intuitive for a temporally evolving process, it is also powerful for exploring pathways for scenarios that do not represent an evolving process, such as in this case. Here we are interested in how different realisations, bins, or categories of variables combine to lead to changes in the outcome variable, our player success measure. Essentially, the event tree and its resulting CEG enable us to explore the pathway patterns of successful players. To ensure a sufficient number of edge counts along tree branches, we binned each of the first three variables into four bins. These bins were labelled as ‘variable name g(1/2/3/4)’, where g1 was the bin with lowest values and g4 was the group with highest values. The success measure of total time played was binned into three groups as ‘low’, ‘medium’, ‘high’.

Figure 9 shows the event tree with the variables in the order listed above. Here we see that most of the players follow one of a few possible pathways (root to leaf paths) and most of the other pathways are sparsely populated.

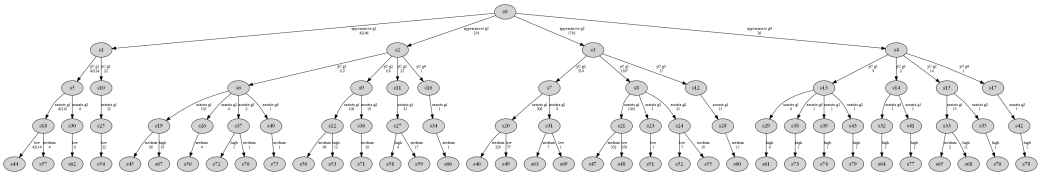

Figure 10: The event tree generated for the variables revealed by the elastic net as positively influencing success.

We can also see that the event tree has sampling zeros, that is, some paths are not a possibility at all (at least, within our dataset). For example, a player cannot have low appearances, high yellow cards, high assists, and high scores. Some of these zeroes could be structural, for instance, it may actually be impossible by the game rules for a player to have low appearances, high yellow cards, high assists, and high scores.

A CEG is now employed for merging pathways of this event tree. Figure 10 shows the nodes that the algorithm considered merging at step 1.

Figure 11: A Chain Event Graph showing the nodes that the Agglomerative Hierarchical Clustering (AHC) algorithm considers merging in the first step. Edges between two nodes indicates that the log marginal score is improved by merging the nodes in the first round of the AHC.

The final step of the algorithm merges the following nodes s6, s13, s5, s18, s9, s14, s15. The last two sets merge nodes that do not correspond to the same variable, and in the second set, both nodes belong to the same path. As such, we can ignore these last two sets. The interesting merge is the first set. Here, the model reveals that the probability of success follows a similar pathway for players with fairly high appearances (level g3) and low yellow cards as it does for players with very high appearances (level g4) and low yellow cards, which makes some intuitive sense.

Moreover, the observation that appearances as a substitute on the bench positively predicts the number of minutes played (from the elastic net output) now also makes sense. A player receiving 5 yellow cards will be suspended, so a team may wish to keep a player with a high number of yellow cards on the bench until an important match, to reduce the risk of suspension. Given that most of the individuals following these two partial pathways have medium or high score based on their level of assists, it would be interesting to map out the club membership and time at each club for a group of players who follow one of these pathways. This could show some interesting links in their careers which might be a determinant of success (at least based on our success measure).

In theory, this could also be applied to the success measure adopted in Section 4.1.2, a measure based on the Guardian's top 100 players ranking list. Again, this was used to create a binary variable, where a player appearing in the top 100 is given a label of 1, and those not appearing are labelled 0. However, when running the previous elastic net with the same input variables as before (including minutes played), but with this new success measure as the output variable, the net failed to return any results. This could be due to the large volume of features with different types (e.g., some dummy variables, some continuous).

Finally, new input variables were created to represent transfer types. The number of times a player was transferred was counted, broken down by transfer type, and likelihood of appearing in the top 100 list was modelled. Transfers were coded as:

- Transfer
- Loan
- Free agent
- Player swap
- Trial
- Free transfer
- Return back from loan

Re-running the analysis produces the following results (Table 2). Again, the coefficient represents the magnitude of the affect of the feature relative to other features:

Variable	Coefficient
Transfer	7.34×10^{-4}
Loan	2.94×10^{-4}
Free transfer	-7.49×10^{-4}
Return back from loan	-1.09×10^{-3}

Table 1: Variables influencing player success.

Visualising these results (Figure 11), we can see that Transfer and Loan are positively associated with the success measure, and Free Transfer and

Back from Loan are negatively associated. The negative effect of returning from loan is the largest, relatively speaking. Additionally, the number of Transfers and Free transfers are similar with respect to magnitude, albeit in different directions.

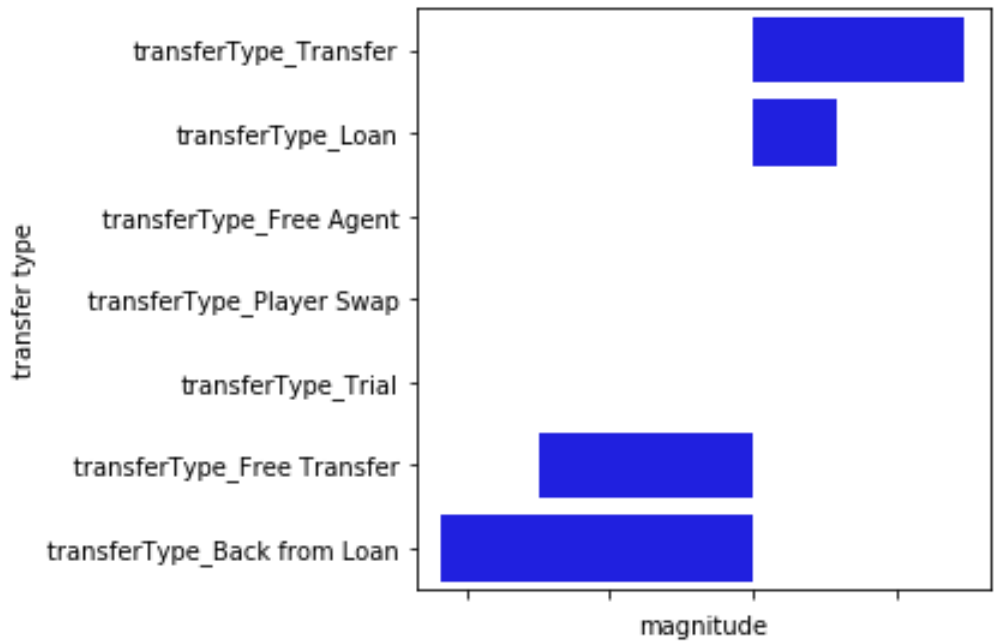

Figure 12: The magnitudes of the effects of the features from the elastic net.

If we had more time, we could use these features in the CEG to find pathways.

We evaluated the model using a simple linear regression of the observed and predicted success measure. Our model fit was low, R^2 = .0016 (0.16% of the variance). The model is poor at predicting whether a player will be in the top 100 as a regression, but is nonetheless successful identifying variables that are intuitively important for career success.

Discussion: Limitations and future exploration

Caution should be applied when interpreting any of these analyses as causal. The employed methods analysed the data to pick up patterns in the characteristics and career pathways of players but do not imply existence of a causal mechanisms. Additionally, given that we are interested in determining the features or patterns of player success, our ideal dataset would have only players who have finished their football careers, since players can peak at different points in the career and so the career stages of players aren't directly comparable. The current dataset has players at different stages of their careers. This does not mean that the results of our analyses are not useful, rather, that we need to exercise caution when interpreting these results.

Going forwards, the pathways identified by the CEG can be used to extract a list of players who have followed these pathways, the temporal career pathways of successful players can be plotted (as done in Figure 13, Section 4.2) and, if patterns do emerge from this, validity of these findings could be tested empirically.

Moreover, instead of analysing whether nationality is influential in determining the success of a player, we could instead ask questions like 'are Brazilian players more likely to be successful than players of other nationalities?' and 'does the success of a player differ on the basis of the continent of birth?'.

4.2 Identifying determinants and indicators of club success

Motivation

So far, we have considered success from a player perspective. There is still much to learn, however, from the club perspective, which could be particularly valuable in helping to inform player recruitment.

This section looks to answer questions such as: 'what are the common features of successful football clubs?', 'can team success be predicted by looking at the combined career paths of first team players?'. Understanding how career pathways combine to form a successful team

would create a potential for clubs to identify the 'ideal' recruitment mix and ultimately better inform their future recruitment decisions.

4.2.1 Preliminary Approach

In the membership data, it is possible to identify current players at specific clubs and to pull out their past career progression, to date. Initial exploratory analysis employed network diagrams to illustrate the paths of players towards their current club. This is demonstrated in the below example for AC Milan, where AC Milan is the centre node to which all of the surrounding player paths lead.

Figure 13: Example network graph illustrating the path that AC Milan players have taken to reach their current club. Each node represents a club; the central node is AC Milan. Each path represents a player (or players) who have moved from one node (club) to another

This visualisation benefits from providing a clear depiction of the players' paths towards the club of interest. In addition, it is effective in highlighting if a given team has key 'feeder' clubs. For example, at the bottom-right in Figure 12, we can see six current AC Milan players pass through Atalanta BC before moving either directly or indirectly on to their current destination.

The graph could be extended to incorporate weighted edges to demonstrate the volume of players who share a given path, or directional edges to show the direction of movement (i.e., incoming vs. outgoing players). Finally, loans and transfers might be coloured differently to

indicate whether a player is on loan at, or has been purchased by, a given club.

Despite these benefits, network graphs are not able to capture all of the variables that might be valuable for recruiters. First, there is no temporal element. From the current analysis, we cannot tell the length of a player's spell at a club or when they moved to another club. Second, it is unclear whether current teammates already played together at another club in the past, or whether they simply passed through the same club at different times.

Extended Approach

It is possible to consider the career path of a player as a sequence of states over the time-span of their career to date. These states could be based on a variety of know attributes. For instance, this could be the club membership or other features which might be of interest to the club. These could be whether the players have previously played together, whether the player is coming from a 'known' feeder club or whether they are from a team at which no other team member has ever played. The states could also represent more personal player information including how active the player was in any given season, i.e. how many appearances the player made, and whether this was for their club, or for a club which they were being loaned to.

In developing sequences of states for individual players, it is then possible to aggregate these individual sequences, effectively developing a picture of the composition of the team based on the features of each player's previous career experience.

This extended approach makes possible to visualise and apply some sequence analysis measures to the bundle of career pathways forming a team. Results can inform the recruitment strategy of a club to appreciate how a new recruit will fit into the new team, also looking at how their career path help reach the mix of seniority and youth, international and domestic experience that made competing clubs to achieve success.

The state sequence analysis carried out in this study has been done through an implementation of the 'TraMineR' package in R which takes sequential data as an input and enables its visualisation and mining

(Gabadinho et al., 2011).This approach has been previously employed when exploring patterns of life trajectories (Gauthier et al., 2010), as well as to explore career paths of IT professionals (Campagnolo et al., 2018). To our knowledge, sequence analysis has not yet been applied to the career paths of footballers.

Data selection and preparation

Although the theoretical underpinnings should be transferable, once the player's career paths have been transformed into state sequences, there are some unique features that require consideration.

Here we focus on one element of a career path - career progression - where the sequence of states reflects a player's career progression through different clubs. Each player could play at any number of different clubs within their careers. The number of potential states introduces noise that could obscure the identification of patterns.

To overcome this, we have assigned each club to a category that broadly captures their status, thus reducing the number of potential states. Teams were categorised as follows:

- **Category 1:** Non-English clubs who have played in the final of the UEFA Champions League in the past 5 seasons.
- **Category 2:** English Premier League (EPL) clubs who have played in the final of the UEFA Champions League in the past 5 seasons.
- **Category 3:** EPL clubs who have have qualified for UEFA competitions in the past 5 seasons.
- **Category 4:** Non-English clubs who have have qualified for UEFA competitions in the past 5 seasons.
- **Category 5:** EPL clubs who have have not qualified for UEFA competitions in the past 5 seasons.
- **Category 6:** English, Non-EPL clubs.
- **Category 7:** Non-English clubs who have have not qualified for UEFA competitions in the past 5 seasons.

In addition to club categories, each state will also take into account the player's membership at that club. That is, whether the player is / was on

loan at a given club **(L)** or whether the move was permanent **(P)**. These rules result in 14 categories (potential states):

- Category 1P
- Category 1L
- Category 2P
- Category 2L
- ...
- Category 7P
- Category 7L

Due to time constraints, the categories were generated simply by looking at a team's standing in their domestic league and their international activity. Some empirical assumptions have been made in the development of the categories. For example, we focus exclusively on the UEFA Champions League and UEFA Europa League to measure a club's international activity. Additionally, we assume that European teams have a higher standing than internationally-active EPL teams, as just one EPL team (Liverpool) has achieved a Champions League final in the past 5 years.

Based on these working assumptions, the categories can be considered as an ordinal list from Category 1 to Category 7, with teams in Category 1 as the most successful, and Category 7 the least.

As mentioned previously, categories could have been developed based on other features. For example, if we wanted to know whether players have played together at other clubs in the past, We could have used the following categories:

- **Category A:** Playing for a team that contains another future player from a given team (i.e., shared experience with overlap).
- **Category B:** Playing for a team that another future player from a given team has previously played at (i.e., shared experience with no overlap).
- **Category C:** No overlap with future teammates.

Based on the conjecture that players who have played together previously might subsequently work well together, this type of categorisation could have some relevance.

Moreover, in previous applications of these approaches the states remain static throughout the time period analysed. Yet in our case if the state is defined as the standing of the club, this too may differ from season to season. For pragmatic application, in this study each club is given a fixed status, or category, which remains consistent throughout the time period analysed, though this could be adapted in the future to consider fluctuations in team ranking.

Analysis, results and conclusion

1. Basic application As a proof of concept, Manchester City, a 'top-flight' EPL club, was chosen for analysis. Manchester City is an interesting case study because their rise to the top is relatively recent. As such, it would be interesting to evaluate changes in their squad composition during their ascent.

In Figure 13 below we isolated the current 26 squad members for Manchester City and plotted their career paths. For each player, we assigned a state to each of their former clubs using the categorisation method outlined above. Additionally, we specify whether the player was on loan or the transfer was permanent. Where the player had not yet begun their professional career, we assigned a 15th category (**PC**, or pre-career) to this time period.

The mix of player career pathways at Manchester City looks as follows:

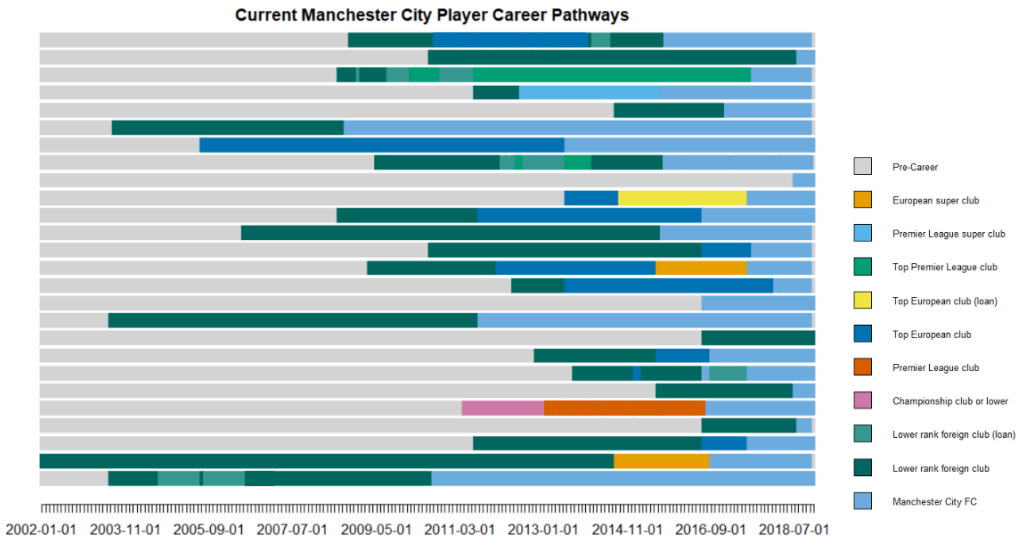

Figure 14: Visualisation of the career pathways of the 26 individuals who currently play at Manchester City.

Each row in Figure 13 represents the career pathway of one player. The colours capture the different status of each club they have played for. The grey is used to represent time before the beginning of their professional career. The longer the coloured bar, the more experienced the player is. The final block (sky blue) represents the transfer of the player to Manchester City.

Re-visualising this information as a distribution plot demonstrates changes in club recruitment over time (Figure 14).

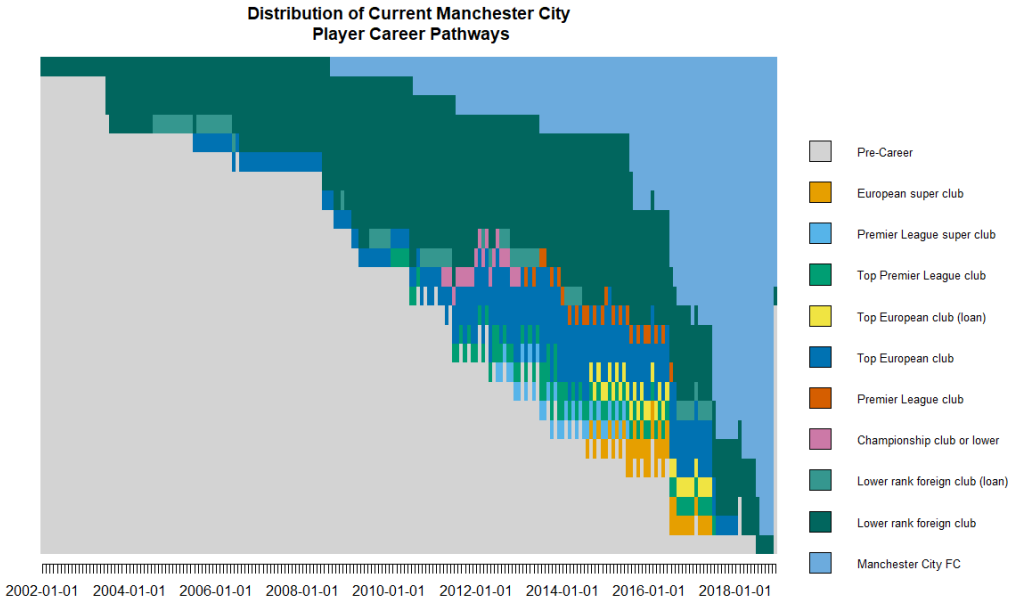

Figure 15: Visualisation of the distribution of team membership for individuals who currently play at Manchester City.

From Figure 13 a number of interesting observations can be made:

- The current team is relatively young, with approximately 50% of the players beginning their career in the past 6 years.
- Any player who has been previously loaned appears to have only been loaned to non-English clubs.
- Just one player has made their way from the Championship to Manchester City. This player made this move via a Premier League club of lower status.
- Few players have moved from other top Premier League clubs.
- Many players are having long European careers (without international game experience) before moving to Manchester City.

1. Extension: Club comparisons

When visually compared with Figure 15 below representing career paths of West Ham players, the sequence plot shows that Manchester City has

a relatively younger team. Also, by only looking at the transfers, West Ham recruits much more from the domestic league (5) than Manchester City (2). For Manchester City in particular, the number of players recruited from lower rank foreign clubs (11) is nearly the same than those recruited from Top level or Super clubs (10).

By observing the full career pathways however, we learn that West Ham recruits players that went through loans much more than Manchester City. And when West Ham recruits from Top Premier League clubs, it is in most cases mature players, arguably close to the end of their career. Other differences can be visually identified regarding recruitment from lower rank foreign clubs (dark green). This is the dominant strategy in both clubs and arguably in the entire Premier League. However, City recruited less players directly from lower ranked foreign clubs (9) than West Ham (15). Furthermore, from the 2016 summer transfer window the trend drastically reduced at Manchester city (only 4 players) while it remained pretty unchanged at West Ham (11 players). If the existence of differences in the recruitment strategies of Manchester City and West Ham might not come across a surprise, the ability of this method to capture differences in recruitment strategies can prove more relevant when comparing clubs of similar standing e.g Manchester City and Chelsea.

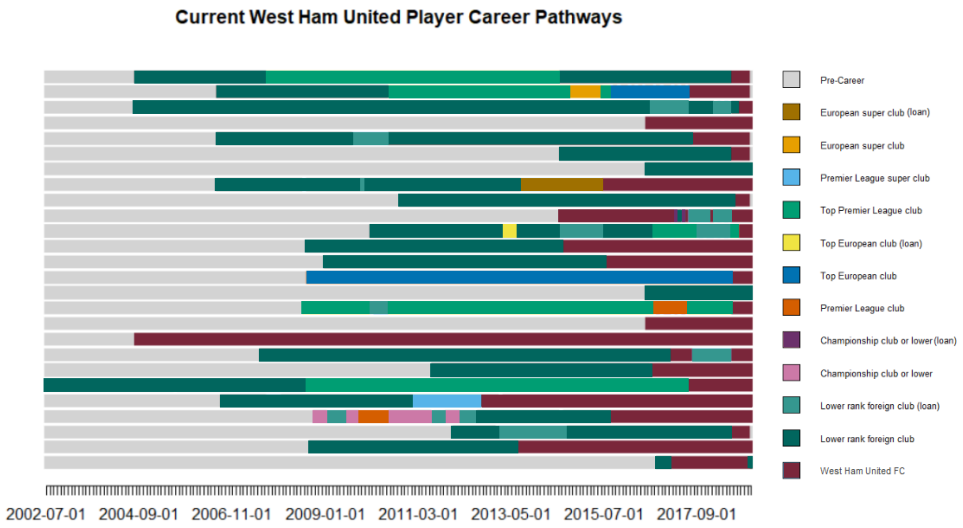

Figure 16: Visualisation of the career pathways of the 26 individuals who currently play at West Ham United.

Discussion: Limitations and future exploration

In this work, broad assumptions have been made to facilitate sequence analysis. For example, the aggregation of individual clubs to club types.

Additionally, many of the insights presented have been derived from descriptive analysis. However, the literature does outline scope for more a systematic approach through the employment of sequential analysis methods to develop measures of difference, variation and similarity. Measures such as multi-channel sequence analysis can enable a more objective comparison of the different club compositions,.

Sequence analysis could be adopted to compare sets of status sequences for a single team, where each set is based on a different set of features, i.e. career pathways (as demonstrated in this study) compared alongside information relating to whether players have shared previous clubs, or any other features of player pathways that might be of interest. However, the time available did not allow for such investigation on this occasion.

Further research is required to understand the ‘typical’ recruitment mix

for different clubs. This can be then be considered as a benchmark by clubs seeking similar success and used to inform their own recruitment decisions.

5 Future work and possible research avenues

Due to the rich nature of this field of study, there are several lines of potential future research. Many of these have been outlined throughout the document, but the main recommendations are summarised below:

Primarily, we suggest developing and implementing the more sophisticated model of player success that is outlined in Section 3.1.2. Second, we recommend developing time sequencing data into the Chain Event Graph to focus on new factors and improve the player attribute analysis (Section 4.1.3). Additionally, the decision tree analysis would benefit from the inclusion of cross validation and testing measures.

Moreover, broad assumptions formed the basis for the club-level analysis. This work would benefit from some testing of these assumptions. This club-level analysis could also be enhanced by developing the cross-club comparison to provide a better understanding of the optimal recruitment mix.

Finally, we recommend combining the work on player-level and club-level trends, to provide a richer understanding of individual- and group-level interactions.

6 Team members

Aditi Shenvi is a second year PhD student in the Mathematics of Systems CDT at the University of Warwick. She works with developing bespoke Bayesian graphical models which have applications in public

health interventions containing asymmetries in the form of context-specific information.

Amanda Otley is a second year PhD student at the University of Leeds in the Institute for Data Analytics. Her current research focuses on shifting geodemographic classification systems from a national to a place-specific perspective in partnership with Leeds City Council and TransUnion.

Basma Albanna is a Development Informatics PhD student at the Global Development Institute, University of Manchester. Her research explores the possibility of leveraging big data sources to identify individuals or communities who are able to achieve better outcomes than their peers - despite having the same socioeconomic constraints - in order to analyse and disseminate their underlying behaviours and practices.

Bhavan Chahal is a second year PhD student in the Mathematics of Systems CDT at the University of Warwick. Her work involves using deep learning tools to infer house prices, and she is currently working on inferring house prices from Google Street View and Zoopla images.

Daniel Justus is currently working as data science engineer at Digital Catapult where he supports start ups and scale ups solving their data science and machine learning challenges. He earned a PhD in computational neuroscience investigating brain circuits that underlie the ability of spatial navigation.

Haoyuan Zhang is a final year PhD student at Queen Mary University of London. He works on safety and reliability problems. His current research involves Bayesian parameter learning from data and knowledge for deterioration learning and Bayesian Networks for maintenance decision support.

Jacopo Diquigiovanni is a first year Ph.D. student in Statistical Sciences at the University of Padua (Italy). His research focuses on algorithms and models for network data, methods for clustering and the improvement of predictive algorithms in the field of sports betting.

Naomi Muggleton is currently working as a postdoctoral researcher on an industry-funded project at Warwick Business School, to implement behavioural insight at Lloyds Banking Group. Her doctoral research examined the evolution of social norms. More recently, she completed a

summer collaboration with the Alan Turing Institute and The British Museum.

Principal Investigator - Gian Marco Campagnolo is a Faculty Fellow at the Turing Institute and Lecturer in Science, Technology and Innovation Studies at the University of Edinburgh. In previous research, he applied sequential analysis methods to study career patterns of IT professionals. His recent work addresses the application of data science to performance analysis in football. His work has been presented at different venues including Football Associations and the Opta Pro Analytics Forum. He is also a qualified football coach and tactical analyst.

7 References

Campagnolo, G., Williams, R., Alex, B., Acerbi, A., Chapple, D. (2017). *Sensitizing social data science: Combining empirical social research with computational approaches to the analysis of career data, Working paper*, 1-49. DOI: 20.500.11820/5d00df5f-3359-4188-afc2-4ebadf8d826e.

Elzinga, C. H. and Liefbroer, A. C. (2007). De-standardization of family-life trajectories of young adults: A cross-national comparison using sequence analysis. *European Journal of Population/Revue europenne de Dmographie*, **23**(3-4), 225-250.

Gabadinho, A., Ritschard, G., Mueller, N. S. Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, **40**(4), 1-37.

Gauthier, J.A., Widmer, E.D., Bucher, P., Notredame, C. (2010) Multi-channel Sequence Analysis Applied to Social Science Data, *Sociological Methodology*, **40**, 1-38.

Smith, J. Q. and Anderson, P. E. (2008). Conditional independence and chain event graphs. *Artificial Intelligence*, **172**(1), 42-68.

turing.ac.uk
@turinginst