Is player success predictable with machine-learning?

By

Shane Glean

A Master’s thesis

Submitted to

The University of Liverpool

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

07/10/2022

ABSTRACT

Is player success predictable with machine-learning?

By

Shane Glean

This thesis applies Machine-Learning (ML) to predicting the success, or lack thereof, of young (≤23) English Premier League debutantes. Only basic, publicly available, data were used from websites such as Transfermarkt, sofifa.com, and footballdatabase.eu. The only information collected was what would have been available prior to each player’s Premier League debut, leading to a restricted features space. Due to constraints on data, this study was represented as a binary classification problem, and four classification models were tested: K-nearest neighbours (KNN), Multi-layer perceptron (MLP), Random Forest (RF), and Support Vector Machines (SVM). In addition, all models apart from the MLP were tested in two forms. Once with all features and once with fourteen components explaining most of the variance from a Principal Components Analysis (PCA). Lastly, all playing positions were analysed to see what variables correlated most with success.

SVM and MLP had the strongest all-round performance relative to the other algorithms, producing accuracies of 76% and a Macro Average F1-score of 74%. The others tended to overfit, with feature reduction not making much of a difference at all. Also, whether a player had won an individual award prior to their debut came out as the variable most correlated with ‘success’ across positions, with some exceptions. On the other hand, variables least correlated with success were heterogenous across positions. The thesis’ conclusion, limitations, and opportunities for future research are discussed at the end.

DECLARATION

I hereby certify that this dissertation constitutes my own product, that where the language of others is set forth, quotation marks so indicate, and that appropriate credit is given where I have used the language, ideas, expressions or writings of another.

I declare that the dissertation describes original work that has not previously been presented for the award of any other degree of any institution.

Signed,

Your-name-here

ACKNOWLEDGEMENTS

Thank you to my supervisor, Dr. Bezerra, for your guidance throughout and for always making time for me. Thank you to Dr. McCabe for your feedback in the earlier stages of this report that helped me improve the depth of my work. Lastly, thank you to the University of Liverpool for the financial support that gave me the opportunity to do something I never thought would be possible for someone like me.

**TABLE OF CONTENTS**

Page

[LIST OF TABLES 2](#_Toc115867881)

[LIST OF FIGURES 3](#_Toc115867882)

[Chapter 1. Introduction 4](#_Toc115867883)

[Chapter 2. Data 10](#_Toc115867885)

[Chapter 3. Methodology 24](#_Toc115867887)

[Chapter 4: Results 29](#_Toc115867889)

**Chapter 5: Conclusion 33**

[Learning Points 41](#_Toc115867890)

[Appendix 43](#_Toc115867891)

LIST OF TABLES

Page

[Table 1: List of features 17](#_Toc115360339)

[Table 2: Hyperparameters Selected 28](#_Toc115360340)

[Table 3: Evaluation metrics for each model 29](#_Toc115360341)

[Table 4: Most and least correlated variables to success by position 31](#_Toc115360342)

LIST OF FIGURES

Page

[Figure 1: Players were most likely to make their debut between ages 18-22 18](#_Toc115450081)

[Figure 2: As players approach their ‘peak’ ages, their value increases. 19](#_Toc115450082)

[Figure 3: Centre-Backs have a high proportion of successes, while Central Midfielers do not. 20](#_Toc115450083)

[Figure 4: Correlation Covariance Matrix 22](#_Toc115450084)

[Figure 6: Average weights of different ages 44](#_Toc115450085)

[Figure 7: Average heights of different age groups 44](#_Toc115450086)

[Figure 8: Percentage of successes by position 45](#_Toc115450087)

[Figure 9: Covariance Matrix 45](#_Toc115450088)

[Figure 10: Covariates of success by position 46](file:///C:\Users\shani\Downloads\First%20draft.docx#_Toc115450089)

[Figure 11: Covariates of success by position 46](file:///C:\Users\shani\Downloads\First%20draft.docx#_Toc115450090)

[Figure 12: Confusion matrices for Random Forest (left) and Random forest + PCA (right) 46](file:///C:\Users\shani\Downloads\First%20draft.docx#_Toc115450091)

[Figure 13: Confusion Matrix for KNN (left) and KNN + PCA (right) 46](file:///C:\Users\shani\Downloads\First%20draft.docx#_Toc115450092)

[Figure 14: Confusion Matrix for SVM (left) and SVM + PCA 46](file:///C:\Users\shani\Downloads\First%20draft.docx#_Toc115450093)

[Figure 15: Confusion Matrix for MLP 46](file:///C:\Users\shani\Downloads\First%20draft.docx#_Toc115450094)

# Introduction

Football or soccer is by many estimates the biggest sport in the world (McCarthy, 2017; Shvili, 2020), and with a global audience of around 3.2 billion people (Premier League, 2019) and annual revenues in excess of 5 billion pounds (Deloitte, 2022), the English Premier League may well be its largest asset. The league’s international status enables it to offer massive financial incentives to all teams in the league, with even larger spoils going to the league’s champions (Sigsworth, 2022). The problem is the level of competition in the league is such that merely avoiding relegation is incredibly difficult for many teams in the league, without the right group of players that is. For instance, it is estimated that the quality of players at a club’s disposal explains somewhere between 81-90% of the variation in what league position a club finish in (Anderson and Sally, 2013; Kuper and Szymanski, 2019, pp.99,113). It is no surprise then that the wealthiest teams that can afford better players usually turn out to be champions, with Manchester City and Paris Saint-Germain being the most obvious examples of this fact in recent years.

Most clubs are not naïve not to this fact: around 18 billion pounds has collectively been spent on player transfers in the Premier League over the last decade (Statista, 2021). This has been supported by the blossoming of football analytics companies, such as StatsPerform and SmarterScout, which charge clubs to use their proprietary software and statistical models to get advanced insights on potential signings and reduce the risk of wasting money on the wrong players (Soccerment, 2020). Some clubs have taken even further steps in this direction, with teams like Brentford and Liverpool employing teams of physicists and mathematicians (Williams, 2020; Herbert, 2021) to create statistical models that give their teams an edge in finding talented players overlooked by their rivals. One innovation of Brentford in particular is their arbitrage of young players, buying players several years before “Peak ages” (Ram, 2020), which appears to be between ages 25-27 for outfield positions (Dendir, 2016; Worville, 2021), and then selling these players on at much higher prices.

The idea is that younger players can often be undervalued assets due to the uncertainty they represent to potential buyers. On the one hand, being younger means there is usually a smaller sample size of games of which to get a reliable measure of the player’s qualities. On the other hand, there is some inherent volatility to younger players due to the uncertainty about how they will evolve as they mature (Bergkamp et al., 2022), or whether severe injuries may stifle their progression (Larruskain et al., 2021). The key insight, however, is that the earlier you can accurately identify talented players the better.

**1.1 Aims**

My thesis aims to explore three issues in parallel, namely:

* Can ML be useful in classifying which young English Premier League debutantes will have ‘successful’ careers using only basic player data?
* Which algorithms perform best for this kind of classification problem?
* What player characteristics appear most relevant to player success, and how does it vary by position?

‘Young debutantes’ just refers to players that are aged ≤23 and playing for the very first time in the league. ‘Success’ here is defined as achieving ≥100 appearances in one of the top 5 European leagues (Bundesliga, La Liga, Ligue 1, English Premier League, and Serie A).

## 

## Literature Review

In the English Premier League alone, billions of pounds are spent on acquiring players each year (Lange, 2021), and there is a good reason for this: the overall quality of a team’s players appears to be a primary determinant of a team’s success (Anderson and Sally, 2013; Beiderbeck et al., 2020). As a means of quality control to ensure they are investing in the appropriate players, many teams make use of commercial sports analytics companies, such as Smarterscout and STATSPerform, that collect massive amounts of data on individual players and rank them on an array of position-specific metrics (Arastey 2021; Altman, 2022).

While football teams’ use of ML is heavily focused on talent identification, the same cannot be said of the academic literature. At the niche intersection of ML and football, there has been a lot of academic interest around the use of supervised classification methods in predicting match outcomes over the last decade (Constantinou, Fenton, and Neil, 2012; Berrar et al., 2019), though this has turned out to be a rather difficult problem (Berrar et al., 2019, pp.123-124). Other applications of ML to football have looked at predicting injury risk using some blend of psychological data, physiological data, and tracking data (Majumdar et al., 2022), as well as some recent attempts to look at improving offensive play using a blend of event and tracking data (Herold et al., 2019).

All that being said, there has been some interesting work presented at the MIT Sloan Sports Analytics Conference in 2021 providing some proof of concept that ML can be useful for predicting player performance. One paper from Patton et al (2021) examined if it was possible to forecast the probability of college NCAA basketball players becoming professionals in the NBA by combining tracking data with event data (p.1). They tested two ensemble methods, one predicting the probability of a player making it to the NBA (pp.6-8) and another for predicting their draft pick (pp.9-11), with the latter often being a key indicator of which players are most talented (p.11). The problem, however, is that while the results reported for both models suggest strong performance and provide proof of concept for ML’s capabilities in predicting successful young talent, it is not possible to apply to their methodology to the present study. Firstly, the authors work for the analytics company StatsPerform, hence they have access to richer features like tracking data which is generally unavailable to the public. Secondly, perhaps for proprietary reasons, the authors leave the details of their ensemble methods rather opaque. For example, they mention that they utilise Microsoft’s LightGBM classifier, a multi-layer perceptron, and the Random Forest algorithm for the ensemble (Patton et al., 2021, pp.4,7,10-11), but do not disclose their hyperparameters or offer a detailed explanation of how the models fit together.

The other relevant paper from the conference, presented by Wiliams, Clarke, and Brugler (2021), is much more promising, since it designs a model to predict player development in the National Football League using only publicly available data. The results of their study are represented as a direct comparison to an earlier published model used as a benchmark, and unlike the previous paper the authors are highly transparent about their methods. The authors’ preferred algorithm is based on the K-nearest neighbours (KNN) algorithm for regression (p.8), chosen for its simplicity and versatility, though the model is not exactly a nearest neighbour algorithm since some nearest neighbours unlikely to be of interest are excluded. This is done with Locality-Sensitive Hashing: a dimensionality reduction technique frequently used to lessen the computational complexity of nearest neighbour search problems. As for the hyperparameters of their model, the similarity metric chosen was Mahalanobis-Distance, which was preferred to Euclidean distance due to its ability to “simultaneously standardize the scale of each axis and adjusts for the presence of correlation between dimensions of the data” (Williams, Clarke, and Brugler, 2021, p.11). The selection of the K value was far more straightforward, simply being chosen by testing 10 values between 5 and 50 in increments of 5. Their model’s results show it to be superior to its benchmark in terms of its loss, as measured by Root Mean Squared Error.

The last paper of interest, authored by Berrar, Lopes, and Dubitzky (2019), offers the largest source of inspiration for my project despite its slightly different subject matter. Using a dataset containing only the essential information for over 216000 games in 52 different leagues, the authors train a KNN classifier to predict the outcomes of future results. Like the study before, KNN is preferred for its simplicity and versatility. Their chosen evaluation metrics are Ranked Probability Score (RPS) and accuracy, choosing the former over the more frequently chosen ‘Brier Score’ due to its ability to discriminate against less similar outcomes. For example, if their model predicted a team to lose and the team won, this would be penalized more harshly than if the model had predicted a draw. They compare their model to models used by other researchers using the same dataset, finding a modest improvement, but the central insight of this paper is that adding contextual information can improve performance.

To show this, the authors integrated data across seasons and merged all teams in the dataset into one league to get an overall sense of each team’s relative strength based on their recent performances. By complimenting this with a rating system for each team that compared them to other teams in their league, the authors could develop metrics to represent how teams perform in attack and defence depending on whether they are playing at home or away- which is important since it considers the well-known ‘home team advantage’ in football (Pollard, 2008). It is this integration of context and background football knowledge that allows this model to outperform other models (Berrar et al., 2019) on the same dataset, and this has influenced the inclusion of metrics that in this dataset that may otherwise have been ignored.

**1.3 Structure of this report**

The rest of this report will first detail how the data was collected, what data was collected, and how it was preprocessed. After this the methodology of the study will be discussed, which will explain both the design choices and how they were implemented. The final chapter will reveal what the results of this study were, alongside a conclusion that will lay out the key points, the limitations, as well as directions for future study.

1. **Ethical Considerations**

All research was conducted in line with British Computer Society’s (BCS) Code of Conduct. No private data was used, there are no conflicts of interest to report, and no information has been withheld since all data and code is available to the reader in the Appendix. Finally, any mistakes or misrepresentations are unintended and my responsibility alone. I encourage any feedback pointing out such flaws to achieve greater accuracy.

# Data

Data Collection

The data was collected from 4 sources:

* The Website *Transfermarkt* which contains a list of all Premier League debutantes and their ages at the time of their debut. It also contains a range of basic player information on the player such as height, goals, assists, appearances, estimated market value, and if the player has won any individual awards.
* The website *footballdatabase.eu,* which produces an annual score for each player based on how their actions contribute to their team’s performance. It also had a detailed record of appearances, goals, and assists in case any data was missing from *Transfermarkt.*
* *Sofifa.com* has all the information on the attributes given to each player on the video game FIFA and maintains a record of how those attributes change over the course of each season.
* *Fifa.com* and *Uefa.com* were used for their records of the youth tournaments they govern, like the UEFA Euro U19 Championships or the FIFA U20 World Cup. These tournaments will usually have a player of the tournament, a team of the tournament, and a top scorer- all which would count as individual awards.
* The website *premierleague.com* was used for its records of youth leagues, namely youth players of the season and top scorers.

The entire dataset is basically comprised of a player’s basic physical data, details of their playing history, and their FIFA attributes. Only players that made their debut between 2010 and 2020 were used, and after players with insufficient data were removed 680 players remained. Finally, the goal of this thesis is to make a more genuine prediction about what players would go on to have successful careers or not, consequently the only data used for each player is what would have been available *prior* to their debut. For instance, the number of goals scored for each player in the dataset are only the goals they scored before making their Premier League debut. The idea being that, if any model using this data was successful, it would have relevance to real world applications in predicting player outcomes.

Given the constraint that all features must be what was available prior to the player’s debut, the following 24 features were selected.

|  |  |  |
| --- | --- | --- |
| Column name | Description | Data Type |
| Position | This is the player’s main playing position. The eight positions in the dataset were Goalkeeper (GK), Centre-back (CB), Fullback (FB), Centre-defensive midfield (CDM), Centre midfield (CM), Centre-attacking midfield (CAM), Winger, and Striker (STR) | String |
| Club Status | A binary variable denoting whether this player come from one of the “Big 5” clubs of Arsenal, Chelsea, Liverpool, Manchester City, or Manchester United. These are selected as standouts because of their prestigious youth systems (as evidenced by the amount of youth leagues and tournaments they have one within the last 10 years) and their consistently strong performances across many (as evidenced by the trophies these teams have won in the last decade.). | Integer |
| Foreign | Another binary variable documenting whether the player was homegrown (born in the UK) or has come from overseas. | Integer |
| Age at debut | The age, in years and months, each player made their debut. | Float |
| Birth quarter | Using the methodology of Hansel, Venckel, and Williams (2007), this variable documents what quarter of months a player was born in between the dates of September 1st and August 31st. These dates are chosen because they dictate what age group a player is allocated to in the UK. For instance, a player born in November would be given the number 1, because he would be born in one of the first three months from September. | Integer |
| Height | The player’s height in cm. | Integer |
| Weight | The player’s weight in kilograms | Integer |
| International appearances | Simply the total amount of youth or senior national team appearances a player has made prior to their debut. | Integer |
| Nation Status | A binary variable denoting whether the national team the player has appeared is one of the more successful national teams or not. For instance, Spain and France would be labelled “1” because their youth, and men’s, teams have a strong record in international competitions, mostly due to the massive competition for selection in these countries. On the other hand, a country like San Marino would be labelled “0”, due to their poor record at all levels of competition and relatively little selective pressure for each position. | Integer |
| Market value on debut. | This is an estimate of a player’s market value in millions (sterling) based on either crowdsourced estimates from the website *Transfermarkt,* or their most recent transfer value. The value selected depended on how long after their transfer their debut was made. For instance, if a player was signed in July and made their debut in August their most recent transfer value would be used. On the other hand, for a player who has never been transferred before, the crowdsourced estimate would be used. | Float |
| Pro debut. | This is a binary variable that tracks whether a player is making their professional debut or just their Premier League debut. This is intended to act as a proxy for the experience level of the player. A player labelled “1” is playing in their first ever professional game, whereas a player labelled “0” has played in at least one professional game before. |  |
| Assists | This is the number of assists the player has made in all competitions there is publicly available data on prior to their debut. | Integer |
| Goals | The number of goals a player has scored in all competitions prior to their debut | Integer |
| FDB rank | This is the score given to the player on the FBDB index on the footballdatabase.eu website. This score tracks how a players statistical performances have contributed to their team’s performance. As these scores are only made annually, to ensure the data being used is only what could have been available at the time, the score selected was the one for the year to previous to that in which they made their debut. | Integer |
| Overall | This refers to overall rating they were given on the video game FIFA, which is a composite score of other rated attributes given to them in the game. | Integer |
| Potential | This is the maximum overall rating the player could reach in FIFA. This is intended to mimic real life, where young players start off not as good as the established players, but they improve over time. | Integer |
| Weak foot | A FIFA rating between 1-5 stating how strong a player’s less dominant foot is, with a rating of 5 meaning the player’s weak foot is just as strong as their dominant foot. This is intended to help replicate the real-life ambidexterity of some players. | Integer |
| Skill moves | This FIFA rating, between 1-5, determines the extent of the skill moves the player can use in game, with a rating of 5 meaning the player can use the most advanced skill moves in the game. | Integer |
| Pace/diving | This is a composite score of a player’s ‘sprint speed’ and ‘acceleration’ attributes on FIFA. It controls how fast the player can run with and without the ball in the game.  Diving is solely for goalkeepers, which dictates how well goalkeepers can dive to save shots. Both are scored between 1-99. | Integer |
| Dribbling/reflexes | Dribbling is a composite score of a player’s ‘dribbling’, ‘ball control’, ‘balance’, ‘agility’, ‘reactions’, and ‘composure’ ratings on FIFA. It dictates how well a player can dribble with the ball in the game.  Reflexes is an attribute only relevant to goalkeepers, and it controls how quickly goalkeepers can react to shots on goal. | Integer |
| Passing/kicking | Passing is a composite score of a player’s ‘short passing’, ‘long passing’, ‘crossing’, ‘curve’, ‘free kick accuracy’, and ‘vision’. It controls how accurately a player can pass the ball in the game.  Kicking, a rating only relevant to goalkeepers, controls the accuracy and length of goal kicks. | Integer |
| Shooting/handling | Shhoting is a composite score of a player’s ‘finishing’, ‘shot power’, ‘penalties’, ‘long shots’, ‘positioning’, and ‘volleys’. It controls how well a player shoots in the game.  Handling is only relevant to goalkeepers, and it controls how good they are at catching the ball in the game. | Integer |
| Physical/positioning | Physical is a composite score of a player’s ‘stamina’, ‘strength’, ‘jumping’, and ‘aggression’. This rating controls most of the player’s athletic qualities.  Positioning is a goalkeeper-specific attribute that controls how well the goalkeeper positions himself to stop shots and intercept crosses. | Integer |
| Defending/speed | Defending is a composite score of a player’s ‘marking’, ‘standing tackle’, ‘sliding tackle’, ‘interceptions’, and ‘heading accuracy’. It dictates how well the player defends in the game.  Speed is another goalkeeper-specific attribute, and it controls how quickly the goalkeeper can come off their line. | Integer |

Table 1: List of features

2.2 Data Preparation

As a dataset containing the relevant information was not available at the start of this project, it had to be created by the author. It contains 680 samples of Premier League debutantes between the ages of 16 and 23, alongside all the public information the author could find that was available prior to each player’s debut. This was the product of merging data from several different sources which was all compiled in a .csv file, with a link available to the reader in the Appendix. Before any exploration of the data occurred, the data was checked to ensure that there were no null values, and that any outliers or players who had values that were likely wrong were removed.

## Data Analysis

After all the data was collected and quality checked, it was explored to get a better sense of what it consists of. This exploration is accompanied by some relevant visualizations to aid the reader’s understanding.

2.31 Most popular debut ages

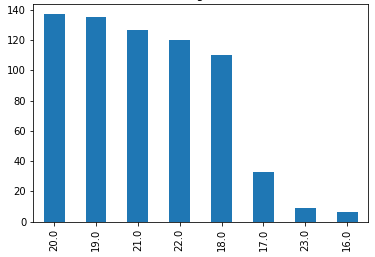


Figure 1: Players were most likely to make their debut between ages 18-22

As shown by Figure 1, the most popular ages for young players making their Premier League debut were between 18 and 22. Relatively few players make their debut before age 18, and only 6 players made their debut before age 17. No player made their debut before age 16 between the 2010-2020 Premier League seasons. The reason the number of 23 year olds in the dataset was so low was due to defining ‘young’ as any player ≤23, and that all ages collected were in years and months. For example, anyone aged 20 would have their age in years and months stored as a float all the way until they reached aged 21, whereas for age group 23 only players that were exactly 23 years and 0 months old were stored.

2.32 Average estimated Market Value on debut by age

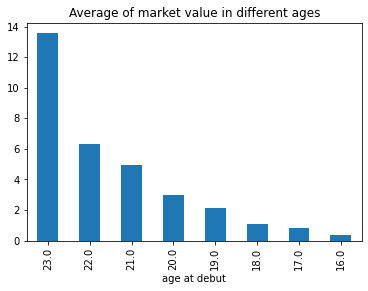


Figure 2: As players approach their ‘peak’ ages, their value increases.

Estimates of market value (in millions) are taken from crowdsourced estimates from the website *Transfermarkt*, and, where possible, the player’s most recent transfer value. As shown by Figure 2, when these values are averaged for each age group there appears to be a linear increase in estimated value with age. This makes some sense, as players mature they tend to gain more experience, leading to improved performances corresponding to higher estimated values. It is potentially confounded by the fact that in the Premier League older debutantes are more likely to have been transferred from another team, the reason being is that transfer values often exceeded estimated market value. Also, and perhaps most importantly, there were only nine 23-year-olds in the dataset, so a wider selection of similarly aged players may yield different results.

2.33 Positions and success

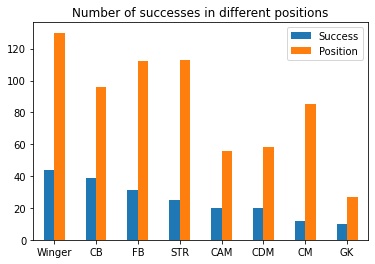


Figure 3: Centre-Backs have a high proportion of successes, while Central Midfielers do not.

In the total dataset, only around 30% of the players in the dataset were deemed to be successes but this varied tremendously by position. Centre-Backs were by far the highest-achieving position, with 41% of debutantes becoming successes, with Central Midfielders being the lowest achievers since only 14% of them beng deemed successes. Interestingly, the success rate for positions closely related to Central Midfield, namely Central-Attacking Midfield and Central-Defensive Midfield, had a much higher success rate. Perhaps indicative of higher returns to specialisation in football.

The other intriguing aspect of this chart is the high number of Wingers, Full-backs, and Strikers making their debuts relative to other positions. This may be a product of sampling bias, but it is possibly more likely to be a product of a real desire for younger players in certain positions. One idea is that wider positions, such as wingers and fullbacks, are far more physically demanding (Martin-Garcia et al., 2018) and require players that are closer to their athletic peak (Worville, 2021). For most athletes explosive sports like football this peak will likely arrive in their mid-to-early 20s (Allen and Hopkins, 2015), which goes some way to explaining a preference for young players in these positions. Some speculative research has also suggested that executive functioning and problem-solving ability may be more important for offensive players (Vestberg et al., 2012; Nakisa and Rahbardar, 2021), and given that these abilities appear to peak in a person’s early-to-mid 20s (Hartshorne and Germine, 2015), this may help to explain why younger players are more prevalent in offensive positions.

**2.34 Feature analysis**

The correlation covariance matrix is a useful tool to learn about your data’s collinearity, as well as providing a means of investigating data quality. To illustrate the latter, if the data is valid then we should expect to see a relationship between features that are related to each other. The strong relationships between height and weight, along with goals and assists, suggest this requirement is met.

Two of the strongest covariates are between the FIFA attributes ‘pace’ and ‘dribbling’, and ‘Overall’ and ‘potential’, which make sense within the context of the game. Players with a high pace rating are often wingers, which is the position that often selects for better dribblers. Similarly with a player’s ‘Overall’ rating, since players that start with a higher rating in the game normally have a correspondingly higher ‘potential’ rating.

This matrix also reveals that ‘awards’, meaning whether a player as won an individual award prior to their debut or not, appears to be the strongest covariate of success in the dataset. On the one hand, it is not surprising that better players are able distinguish themselves from their peers at earlier ages. On the other hand, it only covaries at 0.44, suggesting it may not be that reliable of a signal. Also, prior to the study, there was a worry that a player’s pre-debut FIFA attributes would bear no relationship to a player’s success, but this proved not to be true as three of the strongest covariates of success came from the video game.

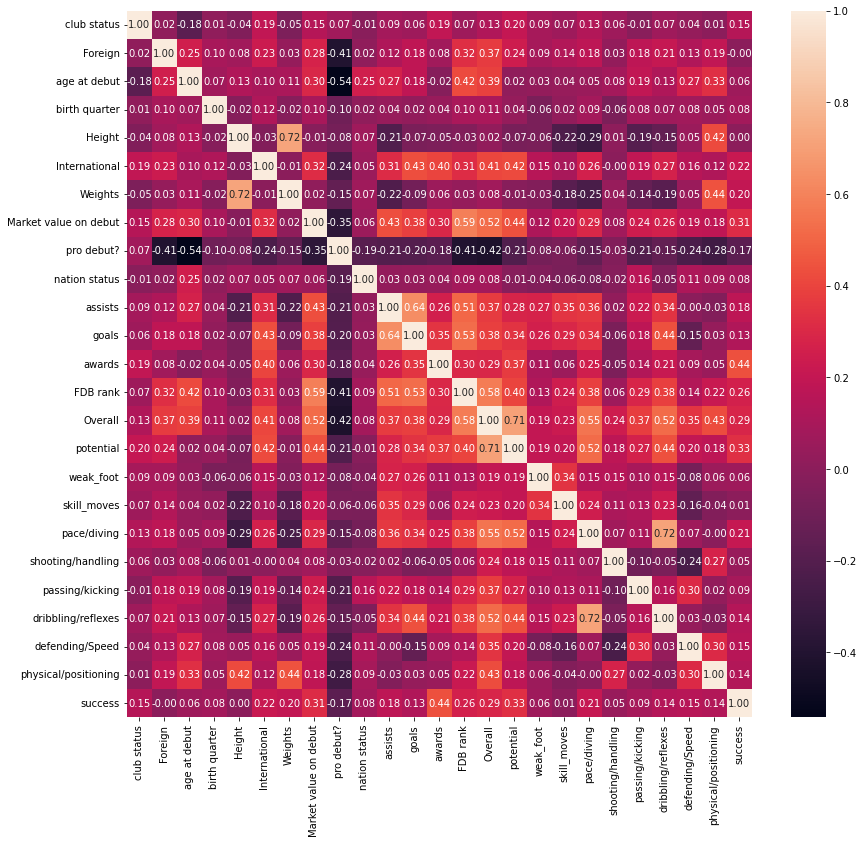


Figure 4: Correlation Covariance Matrix

The last interesting thing about the covariates of ‘success’ was the lack of a relationship with ‘birth quarter’. Several highly cited studies (Jimenez and Pain, 2008; Mujika et al., 2009) have suggested that players that are born in the later months of their annual age-groups will be disadvantaged as they progress in elite sport, in a phenomenon known as the “Relative Age Effect”. The idea being that age gaps on the order of months can be a massive difference at younger ages, as children born just a few months earlier than their peers often gain significant physical advantages (Votteler and Honer, 2013). Though no such effect was found to be relevant to success in this analysis, a closer look at the data revealed that this was not entirely true. For all UK-based players, whose administrative cutoff for each age group is September 1st, most players were born between September and February indicating most players were born in the 6 months after the cutoff. This was similar for foreign players, whose administrative cutoff for each age group is January 1st, as most players were born between January and June. Therefore, while the Relative Age Effect may not directly influence which players become successful, it may well indirectly influence it through producing a selection bias in favour of relatively older players.

# Methodology

Design

In accordance with the primary aims of this project, four ML algorithms were chosen to be tested on this dataset. The first was the K-Nearest Neighbours (KNN) algorithm, which was selected for its simplicity and versatility (Guo et al., 2003), but also for the promise it has shown in similar studies (Berrar, Lopes, and Dubitzky, 2019; Williams, Clarke, and Brugler, 2021). The second algorithm used was Random Forest (RF) for its proven versatility, but also for its robustness against overfitting (ref) which was somewhat of a concern given the imbalanced nature of the dataset. The next algorithm chosen was Support Vector Machines (SVM), for its effectiveness in dealing highly sparse datasets (Li et al., 2015) like the one used for this study. The final algorithm chosen was the Multi-layer Perceptron (MLP), since there were no studies that could be found applying this algorithm to this kind of problem and because of its unique propensity for customization in terms of hyperparameters.

Only the classifier variant of these models were imported because this was made to be a binary classification problem. This was not originally intended but, due to a lack of samples and a lack of readily available data of good quality for each sample, binary classification was the best way to proceed to meet the aims of the study given the constraints. The study could still test how applicable ML is to this problem, as well as how ML models perform relative to each other, even if multiclass classification and regression are not available options. The basis of the classification was labelling each sample with “1” or “0”, with the former representing ‘successful’ and the latter representing ‘unsuccessful’.

I chose to define ‘success’ in this study as any player that has made ≥100 appearances in one of the top 5 European football leagues: the Bundesliga, La Liga, Ligue1, the English Premier League, and Serie A. There were two general exceptions that allowed players who had not met the threshold to nonetheless be judged as successes:

1. A player who had made their debut in 2019 or 2020 and had made over half of the available appearances in one of the top 5 Leagues since then,
2. A player who was very close to meeting this threshold before a serious injury cut their season short.

An example of exception (1) and (2) would be someone like Gabriel Martinelli of Arsenal. Not only did he not have much time to meet the appearance threshold having only made his Premier League debut in the 2019/2020 season, but also sustained a series of injuries that have prevented him from playing for over 6 months.

The author appreciates that any definition of ‘success’ would have necessarily been somewhat arbitrary and inevitably miss some of the nuance of player outcomes, but this threshold definition was chosen because it mostly captured what the study cared about while avoiding many of the pitfalls of other definitions. For instance, if one were to define success as having won a major trophy, this might favour players who happened to be part of good teams despite rarely ever playing for them. Likewise, if one were to define ‘success’ as having won an individual award, this would not only mean that we must conclude almost all professional footballers are ‘unsuccessful’, but it would also disproportionately penalize players who play in defensive positions. The most prestigious individual award in football, the Ballon d’Or (Anderson et al., 2019), illustrates this perfectly. Only one defender has won the award since 2000, and over half of the awards during that time have been won by the same two players.

Lastly, on top of the ML aims of this study, I also wanted to see what the strongest covariate of success by position were. The idea behind this is that different positions require different skillsets from players, so a characteristic that is highly important for Centre Back success may have little importance to Wingers.

## Implementation

Everything was coded in Python 3.8 in a Jupyter Notebook environment, using the sklearn and Keras ML libraries. These libraries enabled me to import all relevant ML classifiers alongside their relevant evaluation metrics. Numpy, Pandas, and matplotlib helped me structure the data and visualize findings. Before the training process began all features were transformed into the same scale with sklearn’s MinMaxScaler, and the training split ratio was 80:20. GridSearchCV was used to tune the hyperparameters for each model, which works by testing various combinations of pre-determined hyperparameters to see which combination delivers the best results. Only the number of layers and nodes were tuned for the MLP as it was assumed the other hyperparameter values would not change much.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithm | | Hyperparameter | Values tested | Value selected |
| KNN | K\_value | | [5,7,9,11,13,15] | 15 |
| Distance metric | | ['minkowski','euclidean','manhattan'] | Minkowski (p=2) |
| Weights | | ['uniform','distance'] | Uniform |
| MLP | Hidden layers | | [1, 2, 3, 4] | 2 hidden layers |
| Nodes | | 'layer1':[64,32]  'layer2':[32, 16, 8] | Hidden\_layer1 = 32 Hidden\_layer2 = 8 |
| Optimizer | | N/A | Adam |
| Loss function | | N/A | binary\_crossentropy |
| Activation function for hidden layers | | N/A | ReLu |
| Activation function for output layer | | N/A | Sigmoid |
| epochs | | N/A | 20 |
| Batch size | | N/A | 2 |
| Random Forest | Criterion | | ['gini', 'entropy'] | gini |
| Max\_features | | [4,5,6,7,8] | 6 |
| Max\_depth | | ['auto', 'sqrt', 'log2'] | Auto |
| n\_estimators | | [50, 100, 200, 500] | 100 |
| SVM | gamma | | [1,0.1,0.01,0.001] | .01 |
| C | | [0.1,1, 10, 100] | 100 |
| kernel | | ['rbf', 'poly', 'sigmoid'] | rbf |

Table 2: Hyperparameters Selected

To evaluate each model’s best performance, sklearn’s classification metrics were imported. Each model’s precision, recall, F1-Score, and macro average was measured for each class and both classes together. In addition, each model’s accuracy was recorded along with a confusion matrix to provide a visual representation of model performance. Weighted average was also collected, but given that only 30% of players in the dataset were deemed to be successes this metric along with accuracy can be misleading.

To see how much redundancy there was in the feature space, a Principal Components Analysis (PCA) was performed. Two versions of three of the ML algorithms were implemented: one running all features, and one running only the top 14 components from the PCA. This was to see how well each algorithms performance could be replicated, or perhaps improved upon, by getting rid of some of the noisier features. The MLP was only tested using all features, as it was felt that reducing the feature might negate most of the benefits of the algorithm. The total number of samples for each unique position was iterated through to see what the strongest covariates of success were by position, and a corresponding heatmap of results was produced which can be found in the Appendix.

# Chapter 4: Results

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | KNN | KNN+PCA | MLP | RF | RF+PCA | SVM | SVM+PCA |
| Precision (successful) | 43% | 43% | 61% | 49% | 47% | 63% | 65% |
| Recall (unsuccessful) | 75% | 75% | 7% | 73% | 74% | 69% | 64% |
| F1-score (successful) | 55% | 55% | 65% | 59% | 58% | 66% | 65% |
| Precision (unsuccessful) | 92% | 92% | 85% | 90% | 91% | 84% | 79% |
| Recall (unsuccessful) | 74% | 74% | 80% | 76% | 75% | 80% | 80% |
| F1-Score (unsuccessful) | 82% | 82% | 82% | 82% | 82% | 82% | 80% |
| Macro Average Precision | 67% | 67% | 73% | 69% | 69% | 74% | 72% |
| Macro Average Recall | 75% | 75% | 75% | 74% | 75% | 75% | 72% |
| Macro Average F1-Score | 68% | 68% | 74% | 70% | 70% | 74% | 72% |
| Weighted Average Precision | 82% | 82% | 78% | 80% | 81% | 77% | 74% |
| Weighted Average Recall | 74% | 74% | 76% | 75% | 75% | 76% | 74% |
| Weighted Average F1-Score | 76% | 76% | 77% | 76% | 77% | 77% | 74% |
| Accuracy | 74% | 74% | 76% | 75% | 75% | 76% | 74% |

1. **Model Results**

These are the results of all models after hyperparameter tuning and training. As can be seen from the table below, SVM had the strongest all-round performance, with MLP coming in as a close second. A closer look at the data gives us some indication as to why this was the case, namely that SVM and MLP were best suited to the task due to their robustness and

Table 3: Evaluation metrics for each model

resistance to overfitting. For example, both models were less precise for the unsuccessful class and had lower recall for the successful classes than the other algorithms. This suggests that the other algorithms were much more prone to overfitting, which produced much higher precision for the unsuccessful samples because of their homogeneity, but this strategy was less optimal for dealing heterogeneity of the successful samples.

KNN with all features and KNN with 14 components produced the exact same performance on all metrics. The KNN algorithms were also very precise in judging which players were unsuccessful, but given the imbalanced nature of the dataset it is not too surprising. Relative to the other models, the KNN algorithms produced the highest recall score for the successful class, though this seems to have come at the cost of precision since it had the worst precision score for the successful class.

Both RF models scored similarly to both KNN models, though they were slightly better on most metrics. Despite RF being an algorithm that ought to be less inclined to overfit it does show some evidence of overfitting in the results. This is probably less to do with overfitting and more related to the dataset not lending itself well to the kind of logical operations that comprise the RF algorithm. To be precise, the succesful class in particular has a large amount of overlap in terms of feature values with the unsuccesful class, leading to a high amount of impurity for many nodes.

Another thing to note is that when each algorithm was tested using only 14 Principal Components, the results did not differ very much. On the contrary, on some metrics the algorithms improved. However, much to my surprise, getting rid of some of the noisier features did not produce much of a benefit.

1. **Correlation covariates for each position**

Just like the correlation covariance matrix from chapter 2, ‘awards’ comes out as the variable most correlated with success for most positions. One of the exceptions to this trend are Attacking Midfielders, whose strongest covariate is the number of assists they have made prior to their debut. This makes sense given that one of the primary functions of this position is to create goalscoring opportunities for other members of the team. It should be noted however that there were relatively few Attacking Midfielders in the dataset used, likely because there are a lack of attacking midfielders generally (Worville, 2021), in any case one should not overinterpret these results.

|  |  |  |
| --- | --- | --- |
|  | **Strongest covariate** | **Weakest Covariate** |
| **Goalkeepers** | | Potential (.52) | Birth quarter (.04) |
| **Centre-backs** | | Awards (.43) | Birth quarter and Shooting (-.01) |
| **Fullbacks** | | FDB rank (.42) | Foreign (.01) |
| **Defensive midfielders** | | Awards (.57) | Club status (-0.02) |
| **Central midfielders** | | Awards (0.53) | Skill moves (-.03) |
| **Attacking midfielders** | | Assists (.49) | Passing and foreign (-.03) |
| **Wingers** | | Awards (.45) | Shooting (.01) |
| **Strikers** | | Awards (.49) | Skill moves (-.04) |

Table 4: Most and least correlated variables to success by position

Fullbacks were another exception to the ‘awards’ trend. I initially thought this was due to it being a defensive position because they are often not selected for individual awards, but this seems unlikely given that ‘awards’ is the strongest covariate for Defensive Midfielder’s and Centre Backs. A closer look at the data revealed that ‘FDB rank’, which is a measure of a player’s performance the year prior to their debut, was a strong covariate for most positions in the dataset, so perhaps it is not surprising it came out as the strongest for at least one position.

The weakest correlates of each position’s success were far more diverse and delivered some unexpected results. The Relative Age Effect that was measured through the ‘birth quarter’ variable appeared to have little influence on the success of Goalkeepers and Centre Backs, which is somewhat contrary to the findings of other authors (Salinero et al., 2013). Though, as mentioned in chapter three, this is possibly a result of selection effects. Whether a player was born or raised in the UK, as measured by the variable ‘foreign’, was among the weakest covariates for most positions analysed, but especially in the case of Attacking Midfielder’s and Fullbacks. In fact, whether a player was foreign or not did not appear to matter for most positions, possibly suggesting that signing foreign players may not be clearly better than sticking with homegrown players, at least at younger ages.

The two results that were most surprising were the weakest covariates for Wingers and Attacking midfielders. Shooting was found to be least correlated with success for Wingers and passing for Attacking Midfielders: attributes that are highly relevant for their positions. This is even more perplexing given that the strongest correlate of success for Attacking Midfielders is a trait that should be closely related to one’s passing ability. It is hard to know for certain the reason behind these somewhat contradictory findings, but the most likely explanation relates to a lack of information on behalf of FIFA. Oftentimes young players who have made few or no appearances yet are underrated by FIFA and are given arbitrary ratings for all attributes apart from ‘potential’.

**Chapter 5: Conclusions**

1. **Conclusions**

This study showed proof of concept that it is possible to predict which players will be successful significantly better than chance using only basic, publicly available, data. Furthermore, the study demonstrates that simple ML algorithms, such as SVM and MLP, seem to show promise for this kind of task. It also shows that across most positions, winning an individual award prior to debut is the strongest correlate of whether a player will be successful. Additionally, the study also shows that some aspects of FIFA representations of a player can be useful in making these kinds of predictions, particularly a player’s “Overall” and “Potential” rating.

1. **Limitations**

One of the major difficulties with this study was finding good quality data that fit the criteria of what I needed. For instance, there are some good sources of more detailed player data on websites such as *theanalyst.com* or *sofascore,* but these websites have only recently come into existence and do not have much information on players for the years prior to 2016. The original plan was to train a regression model to predict how many appearances the player would make, but with only 26 features there simply was not enough data for any model to produce granular enough predictions. Furthermore, with a sample size of only 680, there would likely not be enough training data to produce strong enough results for any multiclass classification problem. In the end, I believed the best way to proceed was by transforming the problem into one of binary classification, as I felt this would still achieve most of the project’s aims. It should be noted however, that due to constraints on data it is hard to draw strong conclusions from the study.

Another issue was that in adopting this binary approach had to find a quantifiable way to distinguish better players from the rest. As discussed in the design section of chapter 3, I eventually chose the appearance threshold of ≥100 to define which players are successful or not. While I stand by this decision, I acknowledge that it is imperfect. One troubling aspect of this definition is that it places extraordinary players in the same category as merely good players and, depending on how much time or money a team is going to invest in a player, it would be helpful to know which of these categories the player is most likely to be in.

Finally, because the data only concerns male players from the English Premier League, it is hard to know whether the findings would replicate to other leagues. For example, whether a player has won individual awards may be more, or less, important for other top European leagues.

1. **Future Directions**

Firstly, in showing a proof of concept, it is my hope that this can serve as a basis for others in the academic world to do similar work with even more rigour and richer features. The dearth of academic literature concerning this topic is surprising given how important talent identification is to the most popular sport in the world, though this may at least in part be a result of the same difficulties in acquiring data that hampered this study.

Finally, the results here potentially show promise for developing a tool that might be able to make a reasonably accurate prediction on which players will be successful or not. With even more data, it may be possible to perform regressions to predict how many appearances a player might make, or how many goals they will go on to score, and so on. On the one hand, if such predictions can be made using only public data by an algorithm with minimal loss this presents a huge opportunity for clubs who cannot afford to consult expensive analytics companies like StatsPerform. And it is worth being optimistic about this scenario, especially as richer metrics like expected goals and optical tracking are becoming more widely available every year, likely leading to even better results than this study.

References

Allen, S.V. and Hopkins, W.G., 2015. Age of peak competitive performance of elite athletes: a systematic review. *Sports Medicine*, *45*(10), pp.1431-1441.

Altman, D., 2022. “How we spot smarterscout young prospects”. Smarterscout., URL: [How we spot smarterscout young prospects](https://smarterscout.com/articles/liverpool-harvey-elliott-sepp-van-den-berg-tottenham-troy-parrott-premier-league-efl-championship-talent-spotting-scouting-recruiting) [Accessed 07-07-2022]

Anderson, C., and Sally, D., 2013. *The Numbers Game.* Chapter 10., Random House.

Anderson, C., Arrondel, L., Blais, A., Daoust, J., Laslier, J., & Van der Straeten, K. (2020). Messi, Ronaldo, and the Politics of Celebrity Elections: Voting for the Best Soccer Player in the World. *Perspectives on Politics,* *18*(1), 91-110. doi:10.1017/S1537592719002391

Arastey, G., 2021. “ARTIFICIAL INTELLIGENCE IN SPORTS”. Sports Performance Analysis. URL: [Artificial Intelligence (AI) in Sports | Sport Performance Analysis](https://www.sportperformanceanalysis.com/article/artificial-intelligence-ai-in-sports). [Accessed 07-07-2022]

Baboota, R. and Kaur, H., 2019. Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting*, *35*(2), pp.741-755.

Beiderbeck, D., Frevel, N., Krüger, H., [Küpper](https://www.mckinsey.com/our-people/jorn-kupper), J., and [Tacke](https://www.mckinsey.com/our-people/tilman-tacke), T., 2020: “The value pitch: The importance of team value management”. Mckinsey & Company., URL: [The value pitch: The importance of team value management | McKinsey](https://www.mckinsey.com/industries/consumer-packaged-goods/our-insights/the-value-pitch-the-importance-of-team-value-management). [Accessed 06-07-2022]

Bergkamp, T.L., Frencken, W.G., Niessen, A.S.M., Meijer, R.R. and den Hartigh, R.J., 2022. How soccer scouts identify talented players. European Journal of Sport Science, 22(7), pp.994-1004.

Berrar, D., Lopes, P. and Dubitzky, W., 2019. Incorporating domain knowledge in machine learning for soccer outcome prediction. *Machine learning*, *108*(1), pp.97-126.

Berrar, D., Lopes, P., Davis, J. and Dubitzky, W., 2019. Guest editorial: special issue on machine learning for soccer. Machine Learning, 108(1), pp.1-7.

Christou, L., 2018., “How Sheikh Mansour bought Manchester City Premier League success”., The Verdict., URL: [How Sheikh Mansour bought Manchester City Premier League trophies - Verdict](https://www.verdict.co.uk/manchester-city-trophies-sheikh-mansour/#:~:text=Manchester%20City%20has%20spent%20over%20%C2%A31.4%20billion%20to,transfer%20spending%20been%20below%20the%20Premier%20League%20average.). [Accessed 04-07-2022]

Constantinou, A.C., Fenton, N.E. and Neil, M., 2012. pi-football: A Bayesian network model for forecasting Association Football match outcomes. *Knowledge-Based Systems*, *36*, pp.322-339.

Deloitte. 2022. [Annual Review of Football Finance 2022 | Deloitte UK](https://www2.deloitte.com/uk/en/pages/sports-business-group/articles/annual-review-of-football-finance.html). [Accessed 23-09-2022]

Dendir, S., 2016. When do soccer players peak? A note. *Journal of Sports Analytics*, *2*(2), pp.89-105.

Dubitzky, W., Lopes, P., Davis, J. and Berrar, D., 2019. The open international soccer database for machine learning. *Machine Learning*, *108*(1), pp.9-28.

Guo, G., Wang, H., Bell, D., Bi, Y. and Greer, K., 2003, November. KNN model-based approach in classification. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"* (pp. 986-996). Springer, Berlin, Heidelberg.

Harper, J., 2021. “Data experts are becoming football’s best signings”. *BBC*. URL: [Data experts are becoming football's best signings - BBC News](https://www.bbc.co.uk/news/business-56164159). [Accessed 14-09-2022]

Hartshorne, J.K. and Germine, L.T., 2015. When does cognitive functioning peak? The asynchronous rise and fall of different cognitive abilities across the life span. *Psychological science*, *26*(4), pp.433-443.

Herbert, J., 2021 “Meet Matthew Benham, a sports gambler who turned $700k into over $300 million”. TheSportsRoom. URL: <https://www.thesportsroom.org/matthew-benham-turned-700k-into-300m/>. [Accessed 04-10-2022]

Herold, M., Goes, F., Nopp, S., Bauer, P., Thompson, C. and Meyer, T., 2019. Machine learning in men’s professional football: Current applications and future directions for improving attacking play. *International Journal of Sports Science & Coaching*, *14*(6), pp.798-817.

Jiménez, I.P. and Pain, M.T., 2008. Relative age effect in Spanish association football: Its extent and implications for wasted potential. *Journal of sports sciences*, *26*(10), pp.995-1003.

Joseph, A., Fenton, N.E. and Neil, M., 2006. Predicting football results using Bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, *19*(7), pp.544-553.

Kuper, S. and Szymanski, S., 2018. Soccernomics: Why England loses, why Germany and Brazil win, and why the US, Japan, Australia, Turkey--and even Iraq--are destined to become the kings of the world's most popular sport. Hachette UK. pp. 99,113

Larruskain, J., Lekue, J.A., Martin-Garetxana, I., Barrio, I., McCall, A. and Gil, S.M., 2021. Injuries are negatively associated with player progression in an elite football academy. Science and Medicine in Football, pp.1-10.

Lange, D., 2021. “Premier League soccer clubs spending on transfer fees 2012-2021”. Statista. URL: [• Premier League transfer fee spending 2021 | Statista](https://www.statista.com/statistics/746589/premier-league-transfer-fee-spending/). [Accessed 06-07-2022]

Li, X., Wang, H., Gu, B. and Ling, C.X., 2015, June. Data sparseness in linear SVM. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Majumdar, A., Bakirov, R., Hodges, D., Scott, S. and Rees, T., 2022. Machine Learning for Understanding and Predicting Injuries in Football. *Sports Medicine-Open*, *8*(1), pp.1-10.

Martín-García, A., Casamichana, D., Díaz, A.G., Cos, F. and Gabbett, T.J., 2018. Positional differences in the most demanding passages of play in football competition. *Journal of sports science & medicine*, *17*(4), p.563.

McCarthy, N., 2017. “Most popular spectator sports worldwide”. Statista. URL: <https://www.statista.com/chart/10042/the-most-popular-spectator-sports-worldwide/>. [Accessed 04-10-2022]

Mujika\*, I., Vaeyens\*, R., Matthys, S.P., Santisteban, J., Goiriena, J. and Philippaerts, R., 2009. The relative age effect in a professional football club setting. *Journal of sports sciences*, *27*(11), pp.1153-1158.

Nakisa, N. and Ghasemzadeh Rahbardar, M., 2021. Comparison of IQ, EI, Sports Performance, and Psychological Characteristics of Young Male Soccer Players in Different Playing Positions. *Annals of Applied Sport Science*, *9*(1), pp.0-0.

Oliver, J.L., Ayala, F., Croix, M.B.D.S., Lloyd, R.S., Myer, G.D. and Read, P.J., 2020. Using machine learning to improve our understanding of injury risk and prediction in elite male youth football players. *Journal of science and medicine in sport*, *23*(11), pp.1044-1048.

Patton, A.N., Scott, M., Walker, N., Ottenwess, A., Power, P., Cherukumudi, A. and Lucey, P., 2021. Predicting NBA Talent from Enormous Amounts of College Basketball Tracking Data. Boston: MIT Sloan Sports Analytics Conference.

Pollard, R., 2008. Home advantage in football: A current review of an unsolved puzzle. *The open sports sciences journal*, *1*(1).

Premier League., 2019. “Premier League global audience on the rise.”. premierleague.com. URL: [Premier League global audience on the rise](https://www.premierleague.com/news/1280062). [Accessed 23-09-2022]

Ram, S. P., 2020. “Moneyball in Football: Analysing Brentford's Recruitment Strategy (eflanalysis.com)”. eflanalysis.com. [Accessed 05-09-2022]

Rommers, N., Rössler, R., Verhagen, E., Vandecasteele, F., Verstockt, S., Vaeyens, R., Lenoir, M., D’Hondt, E. and Witvrouw, E., 2020. A machine learning approach to assess injury risk in elite youth football players. *Medicine and science in sports and exercise*, *52*(8), pp.1745-1751.

Salinero, J.J., Pérez, B., Burillo, P. and Lesma, M.L., 2013. Relative age effect in european professional football. Analysis by position. *Journal of Human Sport and Exercise*, *8*(4), pp.966-973.

Shvili, J., 2020. “Most Popular Sports in the World”. WorldAtlas. URL: <https://www.worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html>. [Accessed 04-10-2022]

Sigsworth, T., 2022. “Premier League Prize Money 2021-2022”. Inews. URL: [Premier League prize money 2021-22: How much every team will earn after final day confirms positions (inews.co.uk)](https://inews.co.uk/sport/football/premier-league-prize-money-2021-22-how-much-every-team-earn-1644024). [Accessed 23-09-2022]

Soccerment. 2020. “The Growing Importance of Football Analytics”. URL: https://soccerment.com/the-importance-of-football-analytics/ . [Accessed 04-10-2022]

Transfermarkt., 2022., “Income and Expenditures”., Transfermarkt., URL: [Premier League - Transfer income and expenditures | Transfermarkt](https://www.transfermarkt.co.uk/premier-league/einnahmenausgaben/wettbewerb/GB1/plus/0?ids=a&sa=&saison_id=2009&saison_id_bis=2022&nat=&pos=&altersklasse=&w_s=&leihe=false&intern=0). [Accessed 26-06-2022]

Vestberg, T., Gustafson, R., Maurex, L., Ingvar, M. and Petrovic, P., 2012. Executive functions predict the success of top-soccer players. *PloS one*, *7*(4), p.e34731.

Votteler, A. and Höner, O., 2014. The relative age effect in the German Football TID Programme: Biases in motor performance diagnostics and effects on single motor abilities and skills in groups of selected players. *European journal of sport science*, *14*(5), pp.433-442.

Williams, A., Brugler, S., and Clarke, B., 2021. MAYFIELD: Machine Learning Algorithm for Yearly Forecasting Indicators and Estimation of Long-Run Player Development. MIT Sloan Sports Analytics Conference.

Williams, S., 2020. “Behind the badge: The physicist who leads Liverpool’s data department”. Liverpoollfc.com. URL: https://www.liverpoolfc.com/news/behind-the-badge/398645-ian-graham-liverpool-fc-behind-the-badge . [Accessed 04-10-2022]

Worville, T., 2021., [What age do players in different positions peak? - The Athletic](https://theathletic.com/2935360/2021/11/15/what-age-do-players-in-different-positions-peak/) ., The Athletic., [Accessed 26-06-2022]

# Learning Points

1. **Skills acquired or improved**

This thesis forced me to lean on skillsets I either was not proficient in or did not have altogether. Firstly, prior to the project my coding was not a strength. Even within the limited remit of machine-learning-based coding there were significant gaps in my knowledge that I had to correct quickly. For instance, prior to the study I had never created a classification Multilayer Perceptron, had never created a correlation covariance matrix, and never performed a Principal Components Analysis. Consequently, in order to complete this project I had to both improve my overall coding ability and learn how to use some new techniques.

1. **Knowledge gained**

Though I had used most of the ML algorithms on this project before, I did not have a deep understanding of how they worked or under what circumstances they would be the preferred options. The thesis forced me to improve my knowledge in this space, particularly when analysing the performance of different algorithms. The thesis also helped me understand how disproportionately important the data collection and pre-processing steps are for studies of this kind. This was demonstrated in a first attempt at this project, where due to a lack of features and some incorrectly labelled samples, the results were not promising at all. After gathering more relevant data and correcting certain inaccuracies performance improved significantly.

I also learned how hard it can be to both uncover the influence of certain variables, and the difficulty in making any sort of causal inference due to either confounding factors or unknown directionality. A good example in this study was the Relative Age Effect (RAE), which did not appear to have any relevance to whether a player was successful or not, but a closer look revealed it that it was likely having some effect on the composition of the sample. Therefore, it is hard to say it was irrelevant even if it had no *direct* influence on success.

1. **What I would do differently**

Firstly, I would perhaps choose a more workable problem to explore. The major difficulty in this study was the lack of relevant and available data. I thought the problem was important enough to persist anyway, and the results make me happy I did. However, in hindsight I was risking far too much in pursuing a thesis that may not have even been possible due to data constraints.

Secondly, I would take time to learn certain scraping techniques to make the data gathering process more time efficient. As mentioned previously, my programming skill is limited, and I did not feel comfortable learning that sort of thing in the fear that it may have taken too long. Again, in hindsight the tradeoff was not nearly as steep as I feared relative to the alternative, as I literally spent weeks in terms of hours finding and entering data by hand into a spreadsheet for 19040 cells. On the one hand, collecting the data by hand allowed greater precision, but this was not worth the excessive time spent on it.

Lastly, I would think through the problem more carefully and ensure everything is clearly defined. One of the things pointed about by both my supervisor and he second reviewer was that some terms were not clearly defined. In particular, the ‘success’ definition that my whole project rested on. While I eventually arrived at a suitable definition for it, this should have better thought through from the onset to avoid confusion to both the readers and myself.

APPENDIX

###### Libraries and code

The following libraries were used for this study:

* matplotlib.pyplot
* pandas
* numpy
* seaborn
* sklearn.model\_selection
* sklearn.preprocessing
* sklearn.metrics
* sklearn.pipeline
* sklearn.decomposition
* sklearn.ensemble
* sklearn.model\_selection
* sklearn.neighbors
* sklearn.svm
* keras.wrappers.scikit\_learn
* keras.models
* keras.layers
* tensorflow

An editable version of the code can be found at [MLfootballProject.ipynb - Colaboratory (google.com)](https://colab.research.google.com/drive/1weSM8HYU-qFtPqxmk31wbTEPIL1b0yJz#scrollTo=sjFkFMf1ubzi).

The dataset used and the code can be found at GitHub at: [GitHub - H4CK3RM4N3/Thesis: Contains player dataset and code used to analyse it](https://github.com/H4CK3RM4N3/Thesis).

1. **Charts and figures**

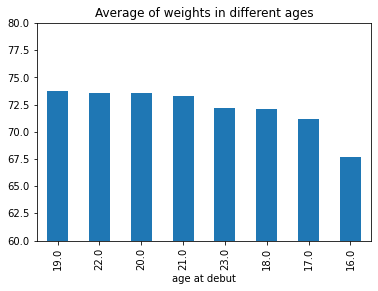


Figure 5: Average weights of different ages

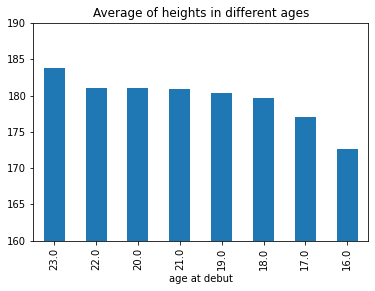


Figure 6: Average heights of different age groups

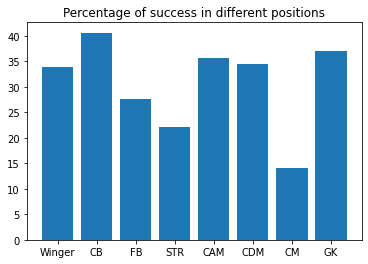


Figure 7: Percentage of successes by position



Figure 8: Covariance Matrix

Figure 9: Covariates of success by position

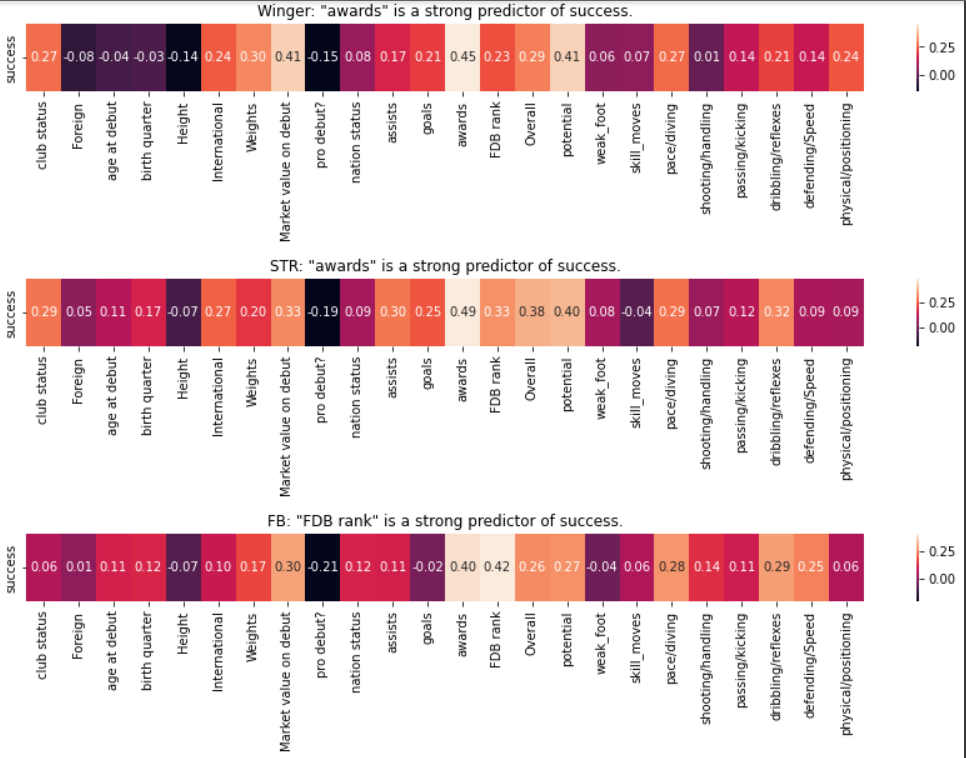
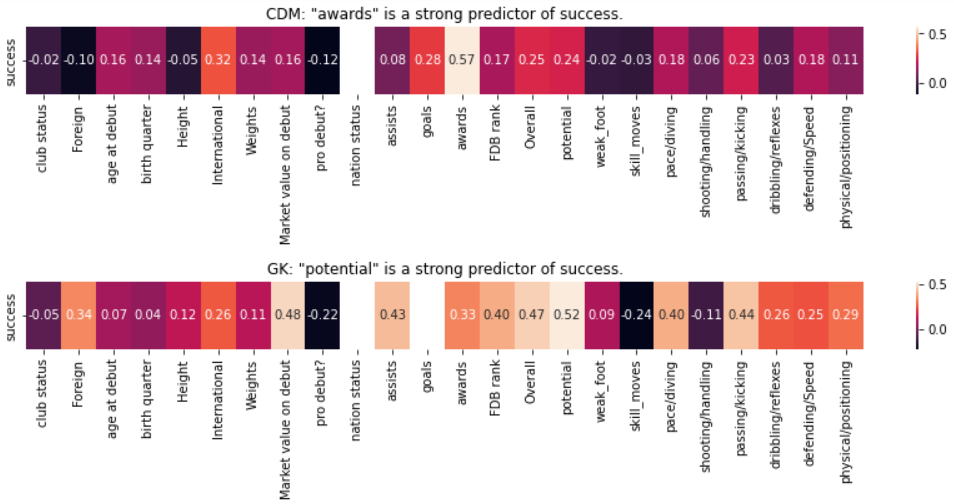
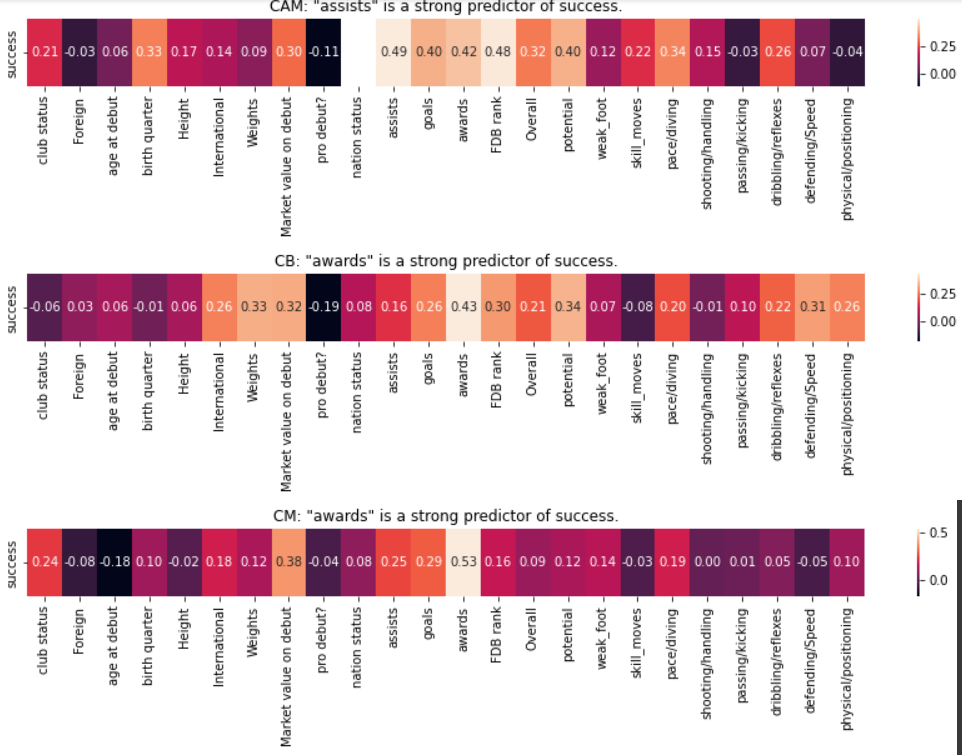


Figure 10: Covariates of success by position

Figure 11: Confusion matrices for Random Forest (left) and Random forest + PCA (right)

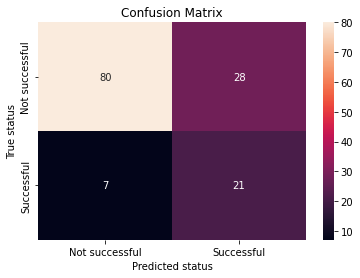
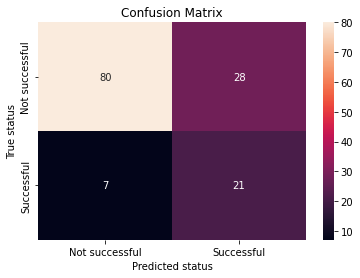
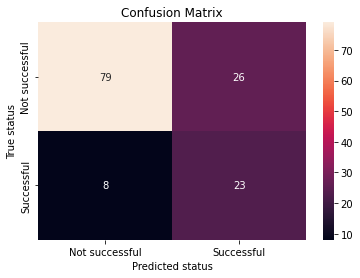
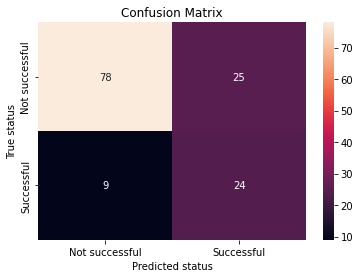


Figure 12: Confusion Matrix for KNN (left) and KNN + PCA (right)

Figure 13: Confusion Matrix for SVM (left) and SVM + PCA

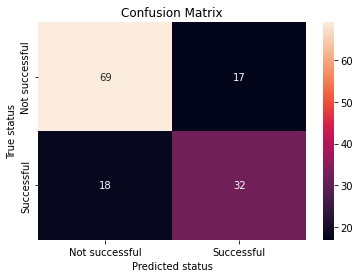
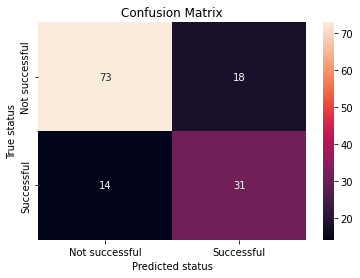


Figure 14: Confusion Matrix for MLP

