Erasmus School of Economics
Master thesis QMBA

Effectiveness of youth soccer academies

Maud de Witte (472923)
Supervisor: M. van de Velden
Second Assessor: P.H.B.F Franses

Abstract

Soccer clubs often promote their youth academies and a lot of the club's resources are put towards identifying and training young talents. However, the actual effectiveness of these academies is something that is much debated and needs to be investigated further. This study therefore aims to find whether playing at a soccer academy also actually has an influence on the player's future career. This effectiveness is measured in terms of success and multiple methods are proposed to estimate the effect. Using these methods, the conclusion can be reached that playing at an academy does increase the probability to become more successful, but not necessarily. More importantly, playing longest at an academy for 5 to 10 years that has a high budget and starting at a young age at the first academy is shown to have a large positive effect on the probability to reach a high level of success during the later stage of the player's career.

March 4, 2022

Contents

1 Introduction

It is commonly perceived advantageous to identify and develop sporting talent at an early age and it is therefore that many national sport organisations have invested increasing amounts of resources in identifying sporting talent at early ages (Vaeyens, Güllich, Warr, & Philippaerts, 2009). Similarly, most professional soccer clubs have set up high-quality talent tracks and academies in order to groom players for a club's professional team (Nesti & Sulley, 2014). Since 2012 the Premier League and its clubs have invested more than £800 million in youth development under the Elite Player Performance Plan. All clubs, including the ones in lower leagues, also applaud the success and possibilities of their own academies and glorify their part in providing world-class players for the major Leagues and renowned clubs.

The actual added value of these talent tracks is, however, not straight forward and there is no clear answer to the question of whether sport academies in general contribute to a greater athletic success. There are multiple studies that attempt to find the worth of sporting academies with conflicting results for many different types of sports. For example, Seiler (2013) suggests that youth development tracks are successful in turning young athletes competing in Olympic winter sports with potential into future medallists. On the other hand, Güllich and Emrich (2012) found that identifying talent and selecting talented youth for special elite tracks does not have a significant influence on the sporting performances of these athletes. Furthermore, negative physical and psychological effects of sport academies have been reported. For instance, Malina (2010) found a higher risk of problems such as burnouts, overdependence and injuries. On the other hand, Ivarsson et al. (2015) found that a high quality development environment can actually be beneficial for the well-being of the player. Hence, the current literature is not in agreement about the effects and the worth of youth academies.

The effect of football youth academies needs to be further investigated. When taking into account the large amount of resources that are invested into these academies by the clubs and the soccer associations as well as the possible negative effects these tracks might have on the junior athletes, it is especially important to be able to accurately estimate the added value of these academies. Furthermore, when considering the aim of developing athletes into star-players, it is also important to determine if the current tactics are contributing to that goal. If this is not the case, it might be needed to change the organisation of academies or to look into other ways to effectively stimulate and development youth potential.

This paper therefore considers the problem of determining the effect of youth soccer academies in the Netherlands and what role these academies play in the professional success of their attendants. In order to answer this main research question, several sub-questions need to be answered. First of all, it needs to be determined how to measure the effect of a youth academy. The effectiveness itself is unobserved and needs to be imputed using other observable factors or a good proxy. The next issue is how to estimate the effect and to understand which techniques are appropriate and why, and how one can accurately model and describe the effectiveness of youth football academies. This can than be used to answer the question of what the role of youth academies have in determining the success of the players and more

specifically what factors or characteristics of the academies are important for the success. Thereby the main question whether they are effective in obtaining their goal of grooming players for a professional career at top clubs can be answered.

In this thesis first the existing literature on the topic is presented and how these papers relate to this research is discussed. Secondly, in the data section the process of gathering the data is described and the used data is discussed and illustrated. Next, the used methods are introduced and discussed. Also, the general set-up is presented. The results of these models are discussed in the section after that. Furthermore, the results are interpreted and the found insights are discussed. Based on the results a conclusion is drawn and the practical implications are discussed in the conclusion. Lastly, the limitations of the research are discussed and suggestions for further research are given.

2 Literature Review

Although the literature on the athletic development is increasing, the existing literature on the effect of talent tracks on later athletic success is still limited. There is, however, quite some research on the correlation between youth competitors and their chances to continue to be successful in adult championships and leagues, both specifically for soccer and more in general for other sports. For example Barreiros, Côté, and Fonseca (2014) investigated how many Portuguese youth athletes competing in international tournaments reappear in senior competitions for soccer as well as other sports. They found that a third of the athletes that were selected at the pre-junior (under 16) age for junior teams made it to the top senior teams. Furthermore, they found that amongst the senior teams, more players did not come from the pre-junior team, but instead transitioned from non-selection teams at a later age. Gulbin, Weissensteiner, Oldenziel, and Gagné (2013) conducted a similar study, but found an even lower general conversion rate of less than 7% for 27 different sports. These studies, however, focus on the reappearance rate in general and do not specifically incorporate the influence of talent tracks on this rate.

Güllich (2007) extends this field of research to incorporate the influence of talent identification and support from an early age onwards. He investigated the ratio of youth competitors that reappeared during the Olympic games for German athletes. The main finding of his research is that most of the youth that were selected at an early age (i.e. between 8 and 12) did not become successful in the senior leagues. Athletes that entered an elite sport school at an later age, on the other hand, were found to be more successful. Similarly, the Vaeyens et al. (2009) find that in general inclusion in talent promotion programs during adolescence does not result in a higher athletic success at a later age for Olympic sports. They continue by proposing a development track based on "talent recycling", where identification and specialization of one discipline takes place at a later stage when the athlete is more mature. Other research in this area, is the one by Güllich and Emrich (2006) that looks into the relation between senior athletic performance and juvenile training volume, success and inclusion in support systems for German national squad athletes for all Olympic sports. For the inclusion in support systems they find that on the short term the correlation is generally positive, implying that if included in a support system, the athlete

will have a greater short term success. However, on the long run, thus at later stages of their athletic career, this effect is not always present and can even be the other way around. Hence, they conclude that whereas such systems can be beneficial in the short-run, they are not profitable for the long-run success of athletes. Furthermore, Emrich, Fröhlich, Klein, and Pitsch (2009) study to what extend the inclusion in a elite sport secondary school results in a higher Olympic success for German athletes. They find that for the summer Olympic games of 2004 there was no significant difference between pupils enrolled at an elite sport school and students at other schools in terms of the amount of medals won. In their study, however, the focus is also on the academic success and well-being of the students. Moreover, most of the mentioned studies focus on Olympic sports, where individual success can be measured by the medal count. For soccer, however, this is not applicable and Olympic achievement does not say much about the success of the player, as the Olympic games are not of great importance for soccer. Furthermore, many of these studies also focus other factors that might influence success than the ones considered in this paper, such as training volume, as this information is not public information.

Another field that is interesting to consider, is the research on the effect of elite private schools on academic success. Here, similar to a sport talent track, only students are admitted into the elite schools that meet certain selection criteria, which are in this case based on academic performance. As there is quite some research on this topic, the ideas and methods of these type of studies may be interesting to consider and adjust to fit this problem. For example, Lucas and Mbiti (2014) investigated the causal effect of attending a private school on the students educational progress and final exam grades using a regression discontinuity. They found little evidence that a private education results in a higher academic success. Similarly, (Wu, Wei, Zhang, & Zhou, 2019) found that there is no difference in test scores between students at elite schools and public schools in China. This research also uses a regression-discontinuity approach in order to estimate this effect. In line with these findings, Clark (2007) also concludes that students at elite schools in the UK do not perform better than others based on their test scores. However, he does find run-long positive effects of elite schools, such as a higher university attendance.

Others also find a positive effects of elite education on academic performance. For instance, Dobbie and Fryer Jr (2015) performed a similar study as the ones previously discussed, only they considered a random lottery admission, instead of admission based on certain criteria. They found that elite charter schools do have a positive influence on their attendants and their test scores. Another study that found a positive effect is the one by Cohodes (2020). However, instead of considering an elite school, they focused on talent tracks for star students at public schools. Furthermore, Berkowitz and Hoekstra (2011) find that attending an elite school causes the students to continue their studies at a more selective university. The attendance of top universities and the effect of attending an elite high school on the chance of doing so, can be related back to the research considered in this paper. One of the key performance indicators for soccer is making it into a top club. Therefore, a similar approach might be appropriate.

Although the above studies are similar to the problem considered in this paper, they are not directly related to this study. As mentioned before, many of the studies that specifically focusing on athletic talent tracks, mostly consider Olympic achievements, which is not applicable to soccer, and include

measures that are beyond the scope of this study. The ones that do focus on soccer academies only look at the conversion rate and do not investigate further what role the talent track has and which characteristics of talent tracks might be influential. Furthermore, whereas the methods used for the studies that investigate elite schools might be applicable to the problem considered in this study, the problem itself differs greatly, as these studies are focused on academic achievement rather than athletic performance. Academic performance is easily measurable, while for soccer players the reached success is not directly observable. Furthermore, many of these studies use a regression discontinuity model, which is not applicable in this case due to the lack of selection criteria and information. Therefore, the effect of youth soccer academies is an issue that should be investigated further. In order to do this, data should be gathered on players that are part of such academy and suitable methods that can be used to estimate this effect should be found. The studies discussed in this section can be used to do so, in terms determining of which factors to include and what methods to use.

3 Data

In this section the database that is used in the study is described and the data generating process is explained. First, the data scraping process is described. How this data is then used to obtain the variables that can be used for estimation is explained after that. Finally, the database with these variables is explored and some characteristics of the data are presented.

3.1 Data Scraping

The data that is used in this study is scraped from the web page www.transfermarkt.nl. This site contains information on all soccer players and clubs. The database consists of information on 1250 players and contains for each player the name, age, position and career trajectory, where the career trajectory is given by the club that the player plays at for each season between 2010/2011 and 2020/2021. These trajectories are constructed using the transfer history data available on the transfermarkt page of the player. If a player has two transfers in one year, these are combined to form a single transfer from the original club to the last club the player transfers to.

The players that are part of the database are all players that play in youth teams that compete in the Dutch A-Junioren Eredivisie, A-junioren eerste divisie or B-Junioren Eredivisie in the seasons between 15/16 and 17/18. These three considered leagues are the youth leagues where youth academy selection teams (under 19 and under 17 teams) compete. The particular seasons are chosen in order to ensure that most players are above 19 at the time of the study, and thus are no longer part of youth teams, but younger than 25. Furthermore, only players with Dutch nationality are included in the database, because this eliminates the influence of foreign youth academies and clubs. The database only includes players with no missing data, thus for which the transfer history and age are known. Furthermore, players that are still part of a youth academy at the time of the study are deleted from the database, as for these players the effect of the youth academy cannot be measured. Additionally, players that were not part of a youth academy during those years, but still made it to Eredivisie clubs are also included in the database.

For this reason, players that play in the Eredivisie during the seasons between 2019-2021, but did not play at a youth academy in the seasons between 2015 and 2018, or during earlier seasons, are selected. Only players that are younger than 25 are selected to match the characteristics of the other players in the database.

Most players did not play at a lot different youth academies. The average number of youth academies that a player has been part of during their career is 1.26. Additionally, 70% of all players only played at one academy during their career. The maximum amount of academies is 4, which is true for only 4 players. On average, there is a decreasing relation between the amount of youth academies and the number of players.

Figure 1 gives an overview of the number of years a player stayed at the academy he played at the longest. Note that only players that played at an academy are included here. It can be seen that there are two peaks, one at 2 years and one between 6 and 8 years. There are thus two types of players that can be identified. The first type plays at one or a few youth academies and is part of these academies for a long time. These are the players corresponding to the peak at 6 to 8 years in Figure 1. The other type corresponds to players that are part of more youth academies for a shorter amount of time, or are "discovered" at a later age and therefore only spend a few years in a youth academy. However, when looking at the full database, it can be seen that on average most players stay at an academy for a relatively long amount of time. Another interesting characteristic of the data is that often players are

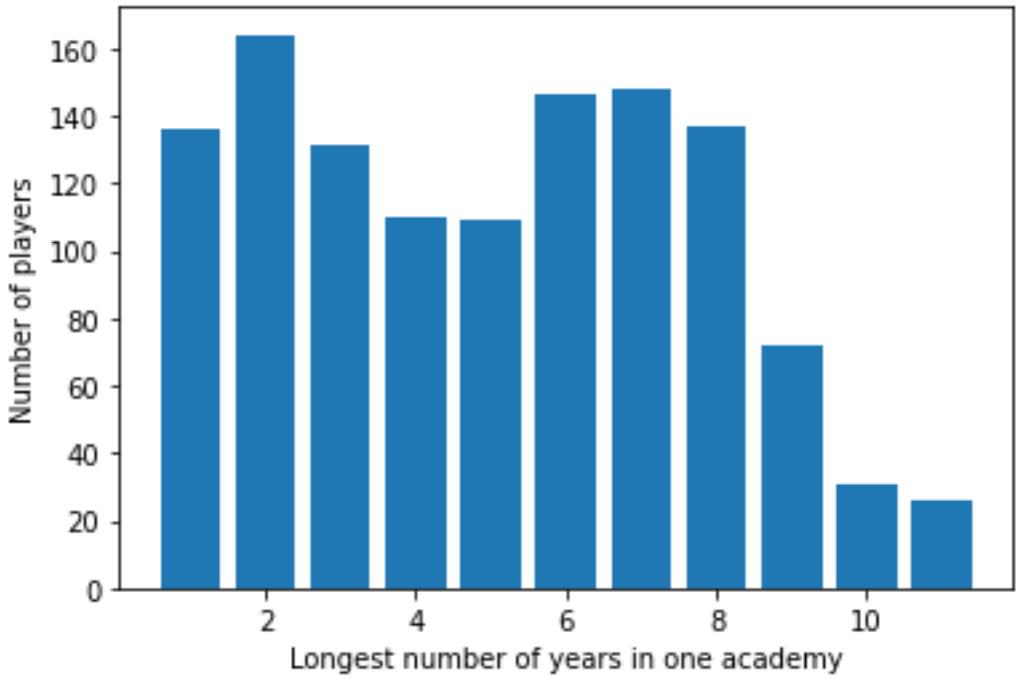


Figure 1. Count of the number of years one player played at the youth academy the played at longest

part of a youth academy for a long time during their early years and at a later age switch to another academy for one or two years. These academies are also often the most prestigious ones.

3.2 Dependent Variable

As mentioned before, the approach and methods used when investigating the effect of elite schools can be used and adapted to fit this problem. However, in these studies, typically test scores are used as an indicator of academic achievement. For football, there is no such numerical indicator for success. Instead, an appropriate indicator might be the indicator proposed by Grossmann and Lames (2013), namely playing at a top soccer club at the senior age. In that sense, it resembles studies that investigate the continuation of elite school students at (top ranking) universities.

When using a similar indicator for success, it is necessary to define which clubs are considered as top clubs and which are not. In order to do so, a similar benchmark as the one used by Grossmann and Lames can be applied. They define an indicator for individual success, where they label players that appeared on a professional team as "successful" and other as "unsuccessful", where professional teams are the top-3 Bundeliga teams or top-3 clubs of comparable foreign leagues. This idea can be extended by making the explanatory variable an ordinal variable, representing different levels of success. For each player, the status of their career is measured 3 years after leaving their last youth academy. This time span is chosen to measure the effect of the youth academy, rather than the possible effect other teams might have on their career before transferring to a top club. The ordinal dependent variables measures 6 levels of success. The highest level of success is defined as playing at a club that was part of the Champions League between seasons 18/19 and 20/21. One level lower corresponds to all clubs that were part of the Europa League during these seasons. The next level corresponds to clubs playing at the national level at the eredivisie. Another level lower includes the clubs that play at the Keuken Kampioen League (which is the second tier league) and one level lower to the Tweede Divisie (which is the third tier league). The lowest level of success corresponds to clubs which are part of a lower league or players that terminated their career.

Table 1 shows the distribution of the dependent variable. The table clearly shows that the fast majority of the players reach a low level of success. Furthermore, only very few players play in the highest two levels of success. As this is very little data to work with, the last two categories of the dependent variable are combined. In this case, the last category consists of 30 observations, making up only 2.4% of the total data. However, when looking at the properties of each class, it can be seen that category 5 and 6 are very similar, yet significantly different from category 4. Thus, in order to preserve as much information from the data as possible, these two categories are not further merged with category 4. This also allows for a more straightforward interpretation, namely category 5 can now be seen as playing internationally. This is obviously different from playing at the eredivisie level in the Netherlands. The properties of the categories are further highlighted in Section 3.4.

3.3 Explanatory variables

In order to get an accurate estimate of the effect of talent tracks and insight into what factors might have an influence on the effect of academies, explanatory variables are included in the estimation. Several variables are created, both academy and player specific. An overview of all explainatory variables can be

Table 1. Distribution of dependent variable

Level of succes	Number of players	Percentage
1 (Regional or terminated)	796	63.68
2 (Tweede Divisie)	144	11.52
3 (Keukenkampioen)	167	13.36
4 (Eredivisie)	113	9.04
5 (Europa League)	21	1.68
6 (Champions League)	9	0.72
Total	1250	100

found in Appendix A.

3.3.1 Variables per club

The academies that are considered are the ones of the clubs that play in either the Dutch eredivisie or the Keukenkampioen divisie. All of these clubs are part of the Betaald Voetbal Organisatie (BVO) and can be considered the major clubs in the Netherlands. For each of the considered clubs a few characteristics are used. Based on the previously discussed literature in Section 2, the included characteristics are the perceived quality and the budget of the academy.

As quality is difficult to observe, a good proxy is needed. The KNVB has a quality and performance program which focuses on the quality of the youth development of clubs. Using this program they have assigned a quality status to each participating club, which ranges from international status to local status. This status is based on talent identification, talent development and learning climate. All clubs, from large to local clubs can participate in the program by paying a fee and the screening and assessment of the academy is done by an independent company. This status is thus independent of the ranking or budget of the club. This quality status will be used as a proxy for the quality of a youth academy. Of the 5 tiers of quality distinguished by the program, the considered academies are all in the top 3 tiers.

For the budget of a youth academy a proxy is also required, as the clubs are not required to share this information and are often not keen to do so. The proxy that is used in this case is the overall budget of the club. Here the assumption is made that the larger the overall budget of a club, the larger the budget of the youth track. The budget is then transformed to an ordered categorical variable, by creating 5 different bins in which the budget can lie. An overview of the used bins can be found in Appendix A.

3.3.2 Variables per player

For each individual player several variables are created. These include the number of years in the youth academy that the player longest part of, the total amount of youth academies the player has been part of, the age the player joined the academy and whether or not the player transferred to a higher ranking academy at the end of their early career. Furthermore, a categorical variables indicating the player's position is included in the database. These variables are all created based on the players career trajectory. The focus here lies on the academy that the player was longest part of. This is based on

the previously discussed observation that players often are part of only a couple of academies and stay there relatively long. Therefore, the academy that the player stayed at longest is likely to have the most influence on the player's development. The number of years that the player stayed at this academy and the age at which they joined the academy are included as explanatory variables. Furthermore, the total amount of youth academies that a player has been part of is included to investigate whether the amount of youth academies also may have an influence. As mentioned before, some players change academies to a higher ranking academy during the last years of their youth career. Therefore a dummy whether or not a player had such a "level-up switch" is added to the dataset.

3.4 Properties of categories

Some preliminary exploration of the data can be performed in order to give more insight in the data and the categories. Figure 2 shows boxplots of each of the used categorical variables. It can be seen that for the number of youth clubs and the level up variable, the mean is equal across all categories of the dependent variable. For the other features it can be seen that there is a difference in means across categories. However, there are also categories that have the same mean, or where the means are close to each other. For example, for status and budget there are two means. Interesting here is that the mean is not increasing and the mean for category 4 is actually lower than for category 5. For the number of years, it can be seen that the first four categories have approximately the same mean, and that the mean for the highest level of success slightly increases. For the starting age, there is a clear downwards trend. These plots therefore suggest that the age at which a player starts playing at a youth academy, the status and the budget of the academy he played at longest and the number of years he spend at the academy that he played at longest are the variables that most likely will be important.

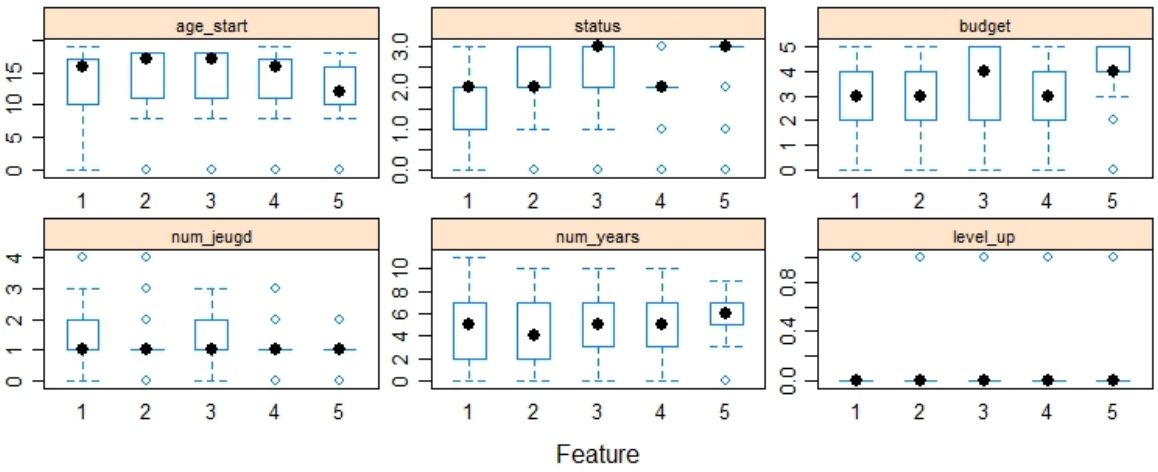


Figure 2. Boxplots of the 6 explanatory variables

4 Models and Methods

Multiple methods are proposed for estimation of the data in order to get more insight into the data and the effects. Both a parametric (OLS and ordinal logistic regression) and a non-parametric (decision and regression tree methods) approach are used for estimation. In this section the different models are introduced and the general set-up and methodology is discussed.

4.1 Proportional Odds and Non-Proportional Odds model

A model that is common to estimate an ordinal dependent variable in practice ordinal logistic regression. Often Proportional Odds models are used for this, which is based on the cumulative distribution function. There are different models that can be used, each differing in assumptions and necessary data structure. Following the framework provided by Liu and Koirala (2012) and Lu, Ma, and Xing (2021), these different models and their assumptions are discussed.

4.1.1 Proportional Odds model (POM)

The Proportional Odds model estimates the probability of either being in or below a certain category of the dependent variable given the set of predictors. This probability can be estimated by (Ananth & Kleinbaum, 1997)

$$P(Y \leq y_j|x) = \frac{exp(\alpha_j - x'\beta)}{1 + exp(\alpha_j - x'\beta)}, \qquad j = 1, ..., J-1, \tag{4.1}$$

where $x = (x_1, ..., x_p)$ is the set of predictors, α_j is the unknown cutoff point of each of the $J-1$ regions, where J is then number of categories of the dependent variable, with $\alpha_1 \leq \alpha_2 \leq ... \leq \alpha_{J-1}$ and $\beta = (\beta_1, ..., \beta_p)'$ is the vector of logit coefficients. This model can be linearized by estimating the logarithm of the cumulative odds, defined by (Ananth & Kleinbaum, 1997)

$$logit(P(Y \leq y_j|x)) = log(\frac{P(Y \leq y_j|x)}{1 - P(Y \leq y_j|x)}) = \alpha_j - x'\beta, \qquad j = 1, ..., J-1. \tag{4.2}$$

The above model is valid under the parallel lines assumption. This assumption states that the coefficient are constant across all different categories of the dependent variable. This assumption can be tested by the Brant test (Brant, 1990), which tests the equality of the parameter estimates. This test tests the equality of coefficients between each pair of $J-1$ regions as well as the overall equality. The test performs a Wald chi-square test for each of these binary logits. For an insignificant overall model test statistic the parallel assumption holds. If this is not the case, the test statistics for each independent variable indicates for which variables the assumption holds and for which it does not.

The model can be estimated using maximum likelihood. The log-likelihood function is defined as (Peterson & Harrell Jr, 1990)

$$L = \sum_{i=1}^{n} \sum_{j=0}^{J-1} I_{ij} log\{P(Y = j|x_i\} = \sum_{i=1}^{n} \sum_{j=0}^{J-1} I_{ij} log P_{ij}, \tag{4.3}$$

where i is the number of observations, I_{ij} is an indicator function that is 1 if $Y_i = j$ and 0 otherwise for

all $j = 1, ..., J-1$, and

$$P_{ij} = \begin{cases} 1 - C_{ij} & \text{if } Y_i = 0, \\ C_{ij} - C_{i,j+1} & \text{if } 0 < Y_i < k, \\ C_{ik} & \text{if } Y_i = k. \end{cases}$$

with $C_{ij} = P(Y \geq y_j|x_i) = 1 - P(Y \leq y_j|x_i)$. Note that this probability can be calculated using Equation 4.1 on the individual level, thus using the set of predictor values specific to each individual observation. Maximizing this function with respect to the regression coefficients yields the ML estimators. This maximization can be done using Newton-Raphson algorithm with line-search. The NR algorithm is a iterative method that searches for the optimal value and this model can be extended by adding a line-search in order to more efficiently find the next candidates and therefore speeding up convergence (Fujiwara, Okamoto, Kameari, & Ahagon, 2005). This method is efficient, converges quickly and and is computational cheap as the log-likelihood described above is relatively well-behaved (Christensen, 2018).

4.1.2 Non-Proportional Odds model (NPOM)

In case the parallel lines assumption does not hold, the POM is not valid. In this case the Non-Proportional Odds model or the Generalized Ordinal Logit model can be used, where the parameters are allowed to differ for each category of the dependent variable. In this case the model is described by (Fu et al., 1999)

$$P(Y \leq y_j|x) = \frac{exp(\alpha_j - x'\beta_j)}{1 + exp(\alpha_j - x'\beta_j)}, \qquad j = 1, ..., J-1, \tag{4.4}$$

where $\beta_j = (\beta_{j1}, ..., \beta_{jp})'$ is the vector of coefficients of each cut-off point j. Similarly as in the POM this model can be transformed to the linear form by estimating the natural logarithm of the cumulative odds, given by

$$\alpha_j - x'\beta_j, \qquad j = 1, ..., J-1. \tag{4.5}$$

This model can also be estimated using ML. This can be done by substituting Equation 4.4 in Equation 4.3.

4.1.3 Partial Proportional Odds model (PPOM)

In the NPOM it is not defined for which parameters the parallel assumption holds. However, it might be the case that for some parameters the assumption holds, while for others it does not. In this case the NPOM estimates more coefficients than needed. In general, the number of coefficients that needs to be estimated in the NPOM grows with the number of categories of the dependent variable. The Partial Proporional Odds model splits estimation into two subsets, one for the coefficients for which the assumption holds and one for which the assumption does not hold, therefore reducing the number of

parameters that needs to be estimated. The model is defined as (Peterson & Harrell Jr, 1990)

$$P(Y \leq y_j|x) = \frac{exp(-\alpha_j - x'\beta - t'\gamma_j)}{1 + exp(-\alpha_j - x'\beta - t'\gamma_j)}, \qquad j = 1, ..., J-1, \tag{4.6}$$

where x is the (p x 1) vector of the full set of covariates, $\beta = (\beta_1, ..., \beta_p)'$ is the set of corresponding parameters, t is the (q x 1) vector of covariates corresponding to the $q \leq p$ parameters that do not satisfy the assumption, γ_j is the corresponding vector with regression coefficients where each region j for which the assumption does not hold has their own (q x 1) vector of coefficients. Similarly as before, the model can be transformed to a linear form. This form is given by

$$-\alpha_j - x'\beta - t'\gamma_j, \qquad j = 1, ..., J-1. \tag{4.7}$$

Note that if all parameters do not hold the assumption, thus $q = J-1, p = 0$, the model becomes NPOM and can be estimated by Equation 4.4. Similarly, if $q = 0, p = J-1$, all parameters hold the parallel lines assumption and the model becomes POM. In this case Equation 4.6 simplifies to 4.1. This model is also estimated using ML. This is again done by maximizing Equation 4.3 where the probabilities are calculated using Equation 4.6.

4.2 Decision trees

The above models are parametric, which means that it is assumed that the data can be modelled by a collection of parameters. Therefore, the main limitation of these models is that certain assumptions have to hold in order for these proportional odds models to be valid (Weng, Meng, & Wang, 2013). The main assumption for the PO models is that the proportional odds can be estimated by a linear model. Non-parametric models are an interesting alternative, as they do not need this assumption to hold in order to be valid. These methods have some advantages over parametric models, such as handling non-linear relationships and dealing with a large set of explanatory variables (Bernard, 2015).

Classification trees partition a large dataset into smaller, more homogeneous subsets. More specifically, the aim is to find certain subsets of the data that have the same value of the dependent variable and similar values of the explanatory variables. This is achieved by searching for splits in the data that create more "pure" subsets, thus sets that are more similar to each other than the larger subset before the split. This purity is measured by a certain measure or variable, depending on the used training algorithm. This process of finding and implementing splits continues until there are no more possible splits or some stopping criteria is satisfied.

Although classification trees do not provide estimates per variable that can be explicitly interpreted, they provide splitting rules that can be visualized in a decision tree. These trees are easily interpretable based on the splitting rules the importance of the variables as well as the practical interpretation can be deducted. Furthermore, the found splitting rules gain insight into the (non-linear) interaction between different factors.

One danger of using classification by decision trees, is the risk of overfitting the data, resulting in a model that does not generalize well (Bayam, Liebowitz, & Agresti, 2005). Pruning methods can be used to solve this problem. Pruning methods remove splits that do not improve the generalizibility of the model at the lower-level nodes (Mingers, 1989), while remaining the overall accuracy of prediction. There are different pruning algorithms that can be used to avoid overfitting, differing in how the trees are built and the used splitting criteria. Three methods to built and prune decision trees are CART, CHAID and conditional inference trees. These algorithms will now be introduced and explained.

4.2.1 The Classification and Regression Tree (CART)

The Classification and Regression Tree algorithm searches for possible splits based on the impurity of the data at a given node. Whereas normally the Gini index is used to measure this impurity, CART can be used for an ordinal dependent variable by using the generalized Gini impurity function. Following the notation introduced by Archer (2010), the generalized Gini impurity function at node t is defined as

$$i_{GG}(t) = \sum_{k=1}^{J} \sum_{l=1}^{J} C(w_k|w_l)p(w_k|t)p(w_l|t), \tag{4.8}$$

where J is the number of categories of the dependent variable with categories $w_1 < w_2 < ... < w_J$, $C(w_k|w_l)$ is the misclassification cost of labeling a datapoint belonging to category w_l with w_k and $p(w_k|t)$ corresponds to the fraction of datapoints belonging to category w_k at node t (Breiman, Friedman, Olshen, & Stone, 1984). Different misclassification cost functions can be used, differing in how costly each type of mistake is considered. In this study two types will be considered, a quadratic one which is defined as the squared difference between the actual score and the assigned score, or the absolute value of the difference. Note that in both cases this means that the larger the error, the higher the resulting cost. However, due to the quadratic form, the quadratic function assigns a higher cost for each type of error.

CART is a greedy algorithm that at each stage looks for the binary split that decreases the impurity the most, thus thereby finding the two resulting subsets that are most similar. The splitting procedure continues until no further splits are possible or until the impurity is not reduced enough based on some threshold. This threshold is the cost complexity factor, which is the factor by which the split must increase the fit.

The cost complexity factor can be treated as a hyper-parameter. This means that the factor can be tuned during the cross-validation to find the value for which the highest prediction accuracy is obtained. Using cross-validation the training data is split into a fold used for training and one used for testing. For each possible value of the parameter, a tree is built on the training fold. This tree is then used make a prediction of the testing sample and the prediction accuracy is measured. This tuning process is discussed in more detail in Section 4.4. The value that results in the highest accuracy on the hold-out sample is selected as the optimal value. The accuracy can be based on either the total misclassification rate, thus the rate of datapoints that are not correctly classified by the model or the misclassification cost, where $C(w_k|w_l)$ from Equation 4.8 is minimized. Through tuning this hyper-parameter, the tree

itself is automatically pruned. Note that the higher the cost complexity factor, the smaller the tree and vica versa. As the value that performs best on the hold-out sample is chosen, it follows that this is the value of the factor that results in the tree that generalizes best, thus solving the overfitting issue.

4.2.2 Chi-square Automatic Interaction Dedication (CHAID)

Chi-square Automatic Interaction Dedication (CHAID) introduced by Kass (1980) splits the data based on the Chi-squared statistic. The algorithm works with a categorical or ordinal dependent variable and predictors. Hence, nominal predictors need to be transformed to a categorical variable first, which can be done by treating each value of a discrete variable as a separate category or by making used-specified intervals for continuous variables. The algorithm then works as follows:

[1] For each predictor take the cross-table with the categories of the dependent and the predictor and repeat:

  (a) For each possible combination of the categories of the predictor calculate the χ^2 statistic of the merging of the two categories. Choose the combination that yields the lowest statistic and merge the categories if the statistic is below a certain critical value. Else, do not merge and move to step [2].

  (b) If after the merge the new category consists of 3 or more of the original categories, determine the best binary split based on the χ^2 statistic for independence of two subsets. If this split is significant based on a certain pre-defined critical value, implement the split. Go back to (a).

[2] For each predictor calculate the significance based χ^2 statistic for the independence of two subsets that result from splitting based on that predictor, using a Bonferroni correction. Select the most significant predictor. If based on a certain pre-defined critical value the statistic is significant, split the data based on the (merged) categories of this predictor.

[3] Consider the new options to partition the data resulting from step [4] and return to [1].

The algorithm takes into account the ordinal structure of the predictors by only allowing adjacent categories to be merged. For an ordinal dependent variable, ordinality can be taken into account by assigning scores to each ordinal category, where the highest category of the ordinal variable has the highest score and vica versa (Magidson & Vermunt, 2005).

As only significant mergers of categories of the independent variables in step 1a and only significant splits in step 2 are allowed, the growth of the tree is limited. Therefore, the algorithm automatically prunes the tree. Differently from most other algorithms, CHAID allows multiple splits at a single level instead of only binary splits. Therefore, it might result in a better fit. However, as the number of categories of the variables and the number of predictor grows, the computational time of the algorithm grows. This can be reduced by only considering a certain subset of mergers and splits in step [1] and [2] or only further investigating groups that are larger than a certain threshold in step [3] of the algorithm. This, however, means that the implemented splits are not necessarily the ones that fit the data the best. Nevertheless, in practice, CHAID is effective and results in satisfactory results (Wilkinson, 1992).

4.2.3 Conditional Inference Trees (CIT)

One problem of the CART method is that the variable selection is biased towards predictors with a large number of categories (Hothorn, Hornik, & Zeileis, 2006). This is due to the fact that the distribution of the test statistic or splitting criteria differs according to the number of categories of the predictor, as pointed out by White and Liu (1994). CHAID is one of the first methods to try to solve this problem by splitting the variable selection and the splitting. Hothorn et al. (2006) also present a solution to the bias problem by introducing the conditional inference framework. This method also splits the variable selection and the splitting decision, but works with recursive binary splitting and differs in the used selection criteria. The methods works by repeating the following steps:

[1] Test the global hypothesis of independence between any of the covariates and the dependent variable. If this hypothesis cannot be rejected stop, else:

   (a) Consider all possible combinations of predictors and the dependent variable and for each combination test the null hypothesis of independence.

   (b) Select the variable with the highest statistical significance and thus the strongest relation with the dependent variable.

[2] Select the optimal split based on the chosen variable by searching over all possible splits and selecting the most significant one.

As illustrated in step 1a and 1b, the selection is based on a permutation test. Step 2 is also based on a permutation test. The idea of a permutation test is to consider each permutation of the set of predictors (step 1) or each possible split (step 2) and to individually test the null hypothesis of independence. Testing the independence of each permutation or split can be done using a certain linear test statistic measuring the dependence between two sets. This gives a complicated test statistic, which is explained in more detail by Hothorn et al. (2006). This test statistic can be altered to explicitly handle both ordinal predictors and an ordinal response variable, where the ordinal variables are captured in score vectors. Again, the highest score corresponds to the highest ordinal category and the other way around. The statistic has a χ^2 distribution and therefore for each permutation a p-value can be calculated. For the testing of the global hypothesis in step 1, a Bonferroni correction is used and the minimum of the adjusted p-values is chosen.

This method also automatically prunes the tree by using a statistical stopping criteria in step 1. This ensures that the procedure stops as soon as the covariate and the dependent variable are not significantly independent, hence preventing overfitting.

4.3 Regression Trees

Another tree building method that can be considered is using regression trees, which lies somewhere between logistic regression and decision tree. Using regression trees methods for an ordinal dependent variable, is something that is not covered extensively in existing literature on ordinal problems. Kramer,

Widmer, Pfahringer, and De Groeve (2001) are one of the first to look into this method and propose an altered Structural Classification and Regression Trees (S-CART) algorithm to solve it. They show that the method obtains good results in practice and is adequate in minimizing both the prediction accuracy and distance-based error. Furthermore, their propose method is adaptable and easily interpretable.

The proposed algorithm by Kramer et al. is an altered type of S-CART algorithm. The original S-CART algorithm is similar to the CART algorithm. It is a greedy algorithm that searches for the best binary split at each node. However, in this case the dependent variable is treated as numerical. At each node the split is based on the minimization of the mean squared error, where:

$$MSE = \frac{1}{n_1 + n_2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (y_{ij} - \hat{y_i})^2. \tag{4.9}$$

Here n_i is the number of observations assigned to branch i, y_{ij} is the dependent variable and $\hat{y_i}$ is a certain target value, which is predicted by the model.

As mentioned before, the S-CART algorithm treats the dependent variable as numerical. This means that the ordering is explicitly taken into account. On the other hand, the categorical nature of the dependent variable is neglected. This categorical structure can be taking into account by altering the target value $\hat{y_i}$. The target variable in the original S-CART algorithm is the mean of the observations covered by the node. However, as this is in general not a class of the dependent variable, this target variable can made to ensure that during the building of the tree only valid values are allowed. Following Kramer et al. different target values can be used, the median, the rounded mean and the mode, where each of these chooses a single class value to be target value. Consider the set of dependent variable values that the observations in node i have. The median target value then selects the median of the class labels. The rounded mean first calculates the mean of the observations and then rounds this real value to the closest dependent variable class. Finally, the mode is the most prevalent value.

This splitting continues until some stopping criterion is met. This can for example be when the tree reaches a certain depth, or when the cardinally of the resulting groups is below a certain threshold. After growing the tree, it needs to be pruned in order to deal with the issue of overfitting. Similarly as in the CART algorithm described before, this can be done by cross-validation. The pruning in this case it then is based on the overall MSE and the tree size that results in the lowest MSE across the hold-out samples is selected.

4.4 Setup and methodology

As mentioned before, the data consists of $n = 1250$ observations, where each observation corresponds to a single player. The dependent variable is an ordinal variable with $J = 5$ different categories. The data is randomly split into a training set and a test set using a 80/20 ratio. The model is trained and built using 80% of the dataset, while the other 20% is reserved for testing the final model's performance. The splitting is performed inside each class, thereby maintaining the class distribution of the full data. This ensures that the test set as well as the training set are representable samples of the full data.

An issue with the training of the different models is that due to the imbalance in the dataset, a model that is built using this data will likely favour the majority class. As 60% of the data falls into category 1, the training is likely to result in a model that will almost always predict class 1, as pointed out by Chawla, Bowyer, Hall, and Kegelmeyer (2002). Such model will thus give limited information about the other categories of the dependent variable, as it only focuses on the majority class. One solution that can be applied for all used models is to use sampling techniques in order to balance the data used for training. The used sampling technique is called up-sampling. This technique over-samples the minority classes. This is achieved by randomly sampling from the minority classes with replacement until the minority classes have the same size as the majority class. The resulting training sample will thus have a larger dimension than before the balancing, giving the models more training data in all classes to train the model with. This sampling is only applied to the training data and the test set is left untouched. This is due to the fact that the test set is used to evaluate the model performance on previously unseen real data. Using sampled data for this evaluation will bias the evaluation, as this data then does not correspond with reality.

The POM is fitted on the training set and the parallel lines assumption is tested using the Brant test. If this assumption does not hold for all parameters, the NPOM and the PPOM are fitted. The models are then compared using likelihood ratios tests and the Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC). Furthermore, the predictive performance is evaluated using the test set. Moreover, OLS is fitted on the dataset as a benchmark model. This model is simple, easy to estimate and the interpretation is uncomplicated. Therefore, it is worthwhile to consider and see how this model performs relative to the more complicated proportional odds models. Based on the indicators of fit, performance and validity of the underlying assumptions, the best fitting model can be selected.

Next the decision and regression trees are built. As mentioned before, during the training of the model, different model parameters need to be tuned. To do this a repeated k fold cross validation scheme is used. This means that the data is split into k folds. At each iteration of the cross validation, one of the k folds is chosen as the hold-out sample, while the other $k-1$ folds are used to train the model. After training the model, its performance is determined using the hold-out sample. This is repeated such that each fold is used as the hold-out sample, thus k times. The way in which the data is divided, thus the different folds that are created, may have an influence on the outcome of the cross-validation. Therefore, the entire cross-validation is repeated a fixed number of times and the average performance is determined.

The actual tuning of the parameters is then based on this average performance across the different repeats and hold-out samples. The tuning is based on a grid search. All possible values of the parameters are considered and for each value the average performance is established using the repeated cross validation. The parameter settings that are chosen are the ones that obtain the best performance. Using these optimal setting the final tree is built on the full training data.

The balancing of the training data is done inside the cross-validation. This means that after determining the $k-1$ folds that are used as the training set, this training set is balanced. This balanced training set is then used to train the model. The hold-out sample remains untouched. The motivation for this is the

same as before, namely that the testing of the model's performance to predict unseen data should be based on data that is close to reality. Therefore, the balancing should be performed inside the cross-validation and only on the training set.

Algorithm 1 gives an overview of the full re-sampling scheme that is used when training the tree methods.

Algorithm 1 Full re-sampling scheme of training trees

```
Determine model and corresponding parameter sets to tune
for all parameters do
    for all re-sampling iterations do
        for all r repeats of the CV do
            Split the data in k folds
            for each iteration of the CV do
                Select one of the k folds as hold-out sample
                Using a sampling method balance the remaining k − 1 folds
                Fit the model using this balances training set
                Predict the hold-out sample
            end for
            Determine the average performance across all k hold-out predictions
        end for
        Determine the average performance across all r repeats of the CV
    end for
    Determine the parameter value that results in the best performance
end for
Fit the model using the full training data and the found optimal parameter set
```

After training the trees and fitting the final models based on the training set, the model performance can be evaluated using the test set consisting of the remaining 20% of the full data.

As both parametric and non-parametric methods are used and the types of methods different greatly, it is difficult to directly compare them to each other. However, the main goal is to gain insight into the relationships between the variables and thus the focus is on descriptive analysis instead of predictive power. Therefore, the relations and insights found in all different models, combined with the validity of the models, can be investigated to reach a conclusion about the driving factors of a player's success.

5 Results

Each of the discussed methods is implemented in R studio 3.6.3. The results of the methods are presented and discussed per (type of) model in this section. The models and their findings are then compared in order to establish which models are preferred for estimation.

5.1 Ordinary Least Squares model

First of all, OLS is used to fit the data. The first assumption that this model makes is linearity, meaning that the relation between the dependent and the independent variable can be modeled by a linear relationship. The other assumptions of linear regression are normality of the residuals, independence of the errors and equal variances of the error terms. These can all be validated after fitting the model.

Fitting the model results in the estimates given in Table 2. As the status and the budget of the academy are categorical variables, dummies are created for each of the levels of these variables. The coefficient estimates for these dummies are then relative to the base of the variable. For example, for the budget of the academy, the base is having played longest at an academy with a budget in the first category. It can be seen that playing at an academy with a larger budget has a positive effect on the level of success. The category corresponding to having played at no academy at all is omitted from the regression due to high co-linearity. For the status of the academy a player has been part of longest, the base is also the first level. Here, there is also a positive effect of having played at a youth academy with a higher status. The level with value of 0 corresponds to having not played at an academy at all. This surprisingly also has a significant positive effect on the level of success. Furthermore, the last coefficient that is significant is the one for the level-up switch. As this is a binary variable, the effect is also relative to the base, which is having made no such switch. Thus, making a level-up switch has on average a positive effect on the level of success that a player reaches after leaving the youth academy.

Table 2. Results of OLS

Variable	Estimate
Intercept	1.926*** (0.189)
budget2	0.419* (0.167)
budget3	0.356 (0.187)
budget4	1.028*** (0.196)
budget5	1.134*** (0.214)
level_up1	0.601*** (0.175)
age_start	-0.004 (0.007)
num_jeugd	-0.120 (0.054)
num_years	0.062 (0.012)
status0	1.217*** (0.235)
status2	0.199* (0.092)
status3	0.335 (0.136)

Note. Standard errors are in parentheses;
*$p < 0.05$,**$p < 0.01$, ***$p < 0.001$

When making predictions using the test set, the numerical predictions can be transformed to class predictions by rounding to the nearest integer. This gives an overall accuracy of 0.0968. This is particularly low because the model fails to detect the lowest category of the dependent variable, as can be seen in the confusion matrix in Table 3. Instead, the model predicts these data points often as class 2 or 3. However, looking at the confusion matrix, it can be seen that for the other classes the model also performs poorly.

Table 3. Confusion matrix for OLS

		Actual Class				
		1	2	3	4	5
Predicted Class	1	0	0	0	0	0
	2	53	8	6	6	1
	3	89	17	16	17	3
	4	17	3	11	0	1
	5	0	0	0	0	0

The assumptions made about the residuals can be visually checked using the plots given in Figure 3.

Figure 3a and 3c can be used to validate the assumption of homoscedasticity of the error terms, where the fitted values correspond to the found outcome of the dependent variable. It can be seen that for the scale location plot there is a clear upwards trend, thus for higher fitted values, the variances of the residuals are also higher. Looking at the fitted value plot, it can also be seen that for higher fitted values, higher residuals are obtained, also indicating heteroscedasticity. Transformation of the dependent variable did not improve the heteroscedasticity, where the log and the square root of the dependent variable were considered. The red line in Figure 3a can be used to validate the linearity assumption, where this line should be approximately horizontal around zero. Here, it can be seen that overall there is a deviation from zero, thus suggesting that this assumption is also violated. Finally, Figure 3b can be used to validate the assumption of normality of the OLS residuals. Clearly, the residuals here are not positioned along the line, thus the residuals do not follow a normal distribution. The violation of the normality and the homoskedacity of the error terms can result in the OLS standard errors to be biased. This therefore means that the confidence intervals and the hypothesis tests might not be valid and the significance of the parameter estimates cannot be assumed to be accurate. The violation of the linearity can result in the parameter estimates to be invalid, therefore leading to poor predictions. Thus, combining these issues, the OLS model clearly has issues, as the estimates as well as the found p-values cannot be assumed to be true.

Furthermore, the model treats the dependent variable as numerical, meaning that although the natural ordering is taken into account, the categorical nature of the dependent variable is ignored. However, the model can be used as a benchmark model and it can be compared to the Proportional Odds model that does explicitly take into account the structure of the dependent variable.

5.2 Proportional Odds models

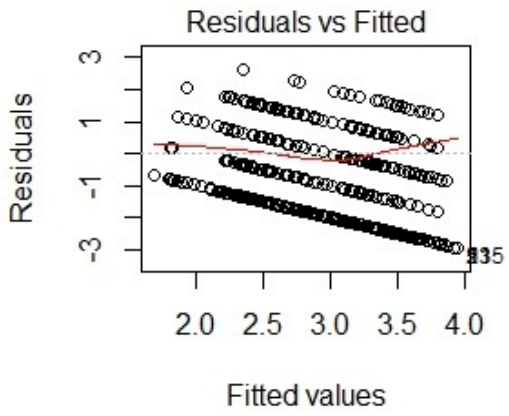
5.2.1 POM

Following the presented framework in section 4.4, first the proportional odds model is fitted. The results of this model can be found in Table 4. The first part of the table gives the estimates per explanatory variable and the second part gives the threshold estimates, which correspond to the α_j's and β_i's in Equation 4.1 respectively.

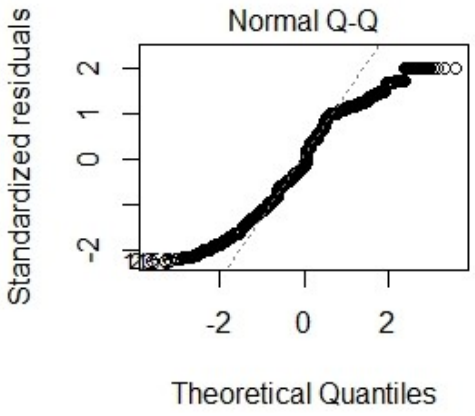
This model finds the coefficients of all variables to be significant. The positive coefficients for the budget, status, level-up dummy and the number of years in the academy imply that a higher value of these variables leads to an increase of the log odds of moving into a higher category of the ordinal dependent variable. A negative effect corresponds to a decrease of this log odds. In order to give a more direct interpretation, the log-odds ratio should be considered. Note that the found coefficients correspond to an increase of one unit of the explanatory variable, or:

$$logit(P(Y \leq y_j | x_i = a) - logit(P(Y \leq y_j | x_i = a - 1) = \beta_i,$$

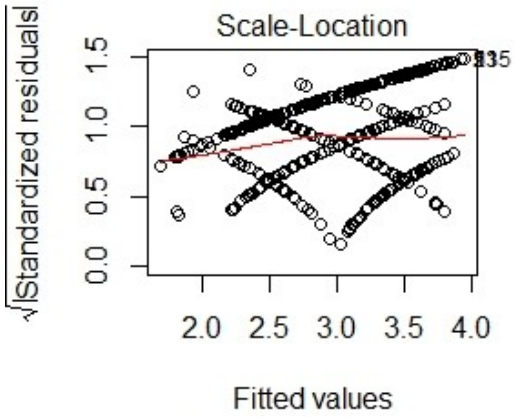
for each $j = 1, ..., J-1$, $i = 1, ..., p$ and where a and $a-1$ correspond to (valid) values of the explanatory



(a) Residuals versus fitted values



(b) QQ Normality plot of Residuals



(c) Scale location plot of Residuals

Figure 3. Visual diagnostic plots of model residuals

variable. Taking the exponent on both sides gives

$$\frac{P(Y \leq y_j | x_i = a)}{P(Y > y_j | x_i = a)} / \frac{P(Y \leq y_j | x_i = a-1)}{P(Y > y_j | x_i = a-1)} = exp(\beta_i),$$

which is the log odds ratio. Here the property $log(a) - log(b) = log(\frac{a}{b})$ is used. This relation can now be used to interpret the coefficients.

The odds of a player that a played longest at an academy with a higher budget to fall into a higher category of the dependent variable is $exp(0.33) = 1.39$ times the odds of a player that played at an academy with a lower budget to fall into a higher category. Hence, the higher the budget of the academy that a player was longest part of, the higher the probability of success for this player. Using the same reasoning, being part of an academy with a high status, playing at an academy for a longer amount of years and switching to an academy with a higher status during the end of the youth trajectory, all give a higher probability to reach a higher level of success during the later career.

Similarly, the odds of a player that played at more youth clubs to fall into a higher category of the dependent variable is $exp(-0.270) = 0.76$ times the odds of a player that played at fewer youth academies to fall into a higher category. In other words, the more youth academies a player has been part of, the lower the likelihood to reach a higher level of success. Furthermore, starting at a later age also decreases the probability of a player to fall into a higher category of success.

Table 4. Results of POM

Variable	Estimate
budget	0.33*** (0.054)
level_up	0.854*** (0.247)
age_start	-0.039*** (0.009)
num_jeugd	-0.270*** (0.070)
num_years	0.029* (0.014)
status	0.198* (0.087)
Threshold	Estimate
1—2	-0.675*** (0.154)
2—3	0.359** (0.154)
3—4	1.219*** (0.154)
4—5	2.269*** (0.158)

Note. Standard errors are in parentheses; $*p < 0.05$,$***p < 0.001$

This model has an accuracy of 0.3387 when making predictions using the test set. Table 5 shows the confusion matrix. It can be seen that the model performs especially well for class 1 and class 5. However, both these classes are predicted too often by the model, causing the detection rate of the other classes to be low.

Table 5. Confusion matrix for POM

		Actual Class				
		1	2	3	4	5
Predicted Class	1	72	14	12	10	1
	2	23	1	1	3	0
	3	6	0	4	4	0
	4	31	2	0	3	0
	5	27	11	16	3	4

Next, in order to validate the parallel lines assumption of the model, a Brant test is performed. The results of this test can be found in Table 6. The overall test statistic is significant, thus suggesting that the parallel lines assumption is violated. When further investigating the test statistics for each variable separately, it can be seen that in fact all variables violate the assumption. It should also be noted that for each variable the evidence that this is the case is strong, as the probability is 0. This means that all variables should be treated as non-proportional and the NPOM from Section 4.1.2 should be used.

5.2.2 NPOM

Therefore, next the NPOM is fitted, where the parallel lines assumption is relaxed for all explanatory variables. The results are shown in Table 7. Note that all coefficients are now threshold coefficients, thus representing the effects of each variable on the different category switches. However, this is again

Table 6. Results of Brant test for parallel lines assumption

	χ^2	df	probability
Overall	294.17	18	0
num_jeugd	12.90	3	0
num_years	14.21	3	0
level_up	23.27	3	0
age_start	82.26	3	0
status	38.28	3	0
budget	17.59	3	0

the effect on the log odds. It can be seen that these coefficients are indeed very different for different thresholds. However, not all thresholds have a significant effect.

This model has some interesting results. For the budget it can be seen that the higher the category, the larger the effect is. The same goes for the level-up dummy, where the effect is especially strong for the transition between category 4 and 5. This means that the effect of such a "level-up switch" has a much higher positive effect on the probability to reach the highest level of success. For both these variables, the transition from category 1 to 2 does not have a significant effect. The effect of the starting age, on the other hand, is significant for each transition. Interesting here is that for the transition from 1 to 2, the effect is positive, while for the others it is negative. This indicates that while the first transition has a negative influence on the probability to reach a higher category of success, the others have a positive influence and actually increase the probability of becoming more successful. For the status of the academy a similar effect can be seen, where a negative effect is found for the transition between category 3 and 4. This is actually in line with Figure 2, where the mean of category 3 is higher than category 4. For the number of academies a player has played at, a negative effect is found. Thus, the higher the number of academies, the lower the likelihood to belong to a higher category of success. Lastly, for the number of years a player has played at the academy he was part of longest, only the last two transitions are significant. This indicates that playing longer at the academy only has a positive effect on the likelihood to fall into the highest 2 categories of the dependent variable.

Due to the threshold structure, the interpretation becomes even more difficult for the NPOM. However, in order to give a more clear interpretation and to visualize the found and described effects, the probabilities of belonging to a class can be computed and plotted for different levels of each explanatory variable. This results in the plots shown in Figure 4. Note that for each plot only the concerning explanatory variable is changed, while the others are kept constant. The different lines correspond to the different levels of success. For the budget of the academy, it can be seen that playing at an academy with a higher budget strongly increases the probability to fall into the highest category of success. For the other levels of budget, the probability to belong to lower levels of success is generally higher. For the variable indicating whether a player has made a "level-up" switch at the end of their youth career, the probability to fall into a lower class is higher for players that did not make a level-up switch, while for players that did make such a switch, the probability to fall into a higher category is greater. For the starting age, it can be seen that especially for the second level of success there is a strong effect. Starting at a late age greatly increases the probability to fall into this category. For all other levels of success, there is a negative trend.

Table 7. Results of PPOM

Variable	Estimate
1—2.budget	0.037 (0.081)
2—3.budget	0.286*** (0.064)
3—4.budget	0.361*** (0.063)
4—5.budget	0.554*** (0.081)
1—2.level_up	0.411 (0.324)
2—3.level_up	0.456* (0.267)
3—4.level_up	0.784** (0.265)
4—5.level_up	2.045*** (0.302)
1—2.age_start	0.045*** (0.012)
2—3.age_start	-0.059*** (0.010)
3—4.age_start	-0.057*** (0.010)
4—5.age_start	-0.062*** (0.014)
1—2.num_jeugd	-0.432*** (0.088)
2—3.num_jeugd	-0.113 (0.078)
3—4.num_jeugd	-0.251** (0.083)
4—5.num_jeugd	-0.458*** (0.112)
1—2.num_years	0.006 (0.017)
2—3.num_years	0.012 (0.016)
3—4.num_years	0.046** (0.017)
4—5.num_years	0.045* (0.024)
1—2.status	0.462*** (0.126)
2—3.status	0.245** (0.102)
3—4.status	-0.092 (0.102)
4—5.status	0.458*** (0.136)
Threshold	Estimate
1—2	-0.183 (0.174)
2—3	0.176 (0.167)
3—4	0.580*** (0.175)
4—5	3.355*** (0.310)

Note. Standard errors are in parentheses
$^*p < 0.05$,$^{**}p < 0.01$, $^{***}p < 0.001$

Furthermore, playing at more youth academies increases the probability to fall into the lowest category, while decreasing the probability to fall into the highest. Additionally, playing at the academy that a player has been part of longest for a longer amount of years increases the probability to fall into the highest 2 categories of the dependent variable, while decreasing the probability for the other 3 categories. Lastly, playing at an academy with a higher status causes a higher probability for either level 2, 3 or 5. For level 1 and 4, it can be seen that there is a clear negative relation. Although this model works with threshold estimates, the overall trends when looking at the class probabilities are on average not very different from the found trends when using the POM.

Furthermore, also using Figure 4, a comparison can be made between players that have been part of an academy and players that have not. Note that the players that have not been part of an academy have a value of 0 on each of the explanatory variables. Looking at the plots, the highest probability for this category is either for level 1 or level 2, thus the lowest two categories of success. Furthermore, these two categories for most explanatory variables have a negative trend, while the higher levels of success often have a line with a positive slope. This suggest that on average, players that have been part of an academy have a higher probability to reach a higher level of success.

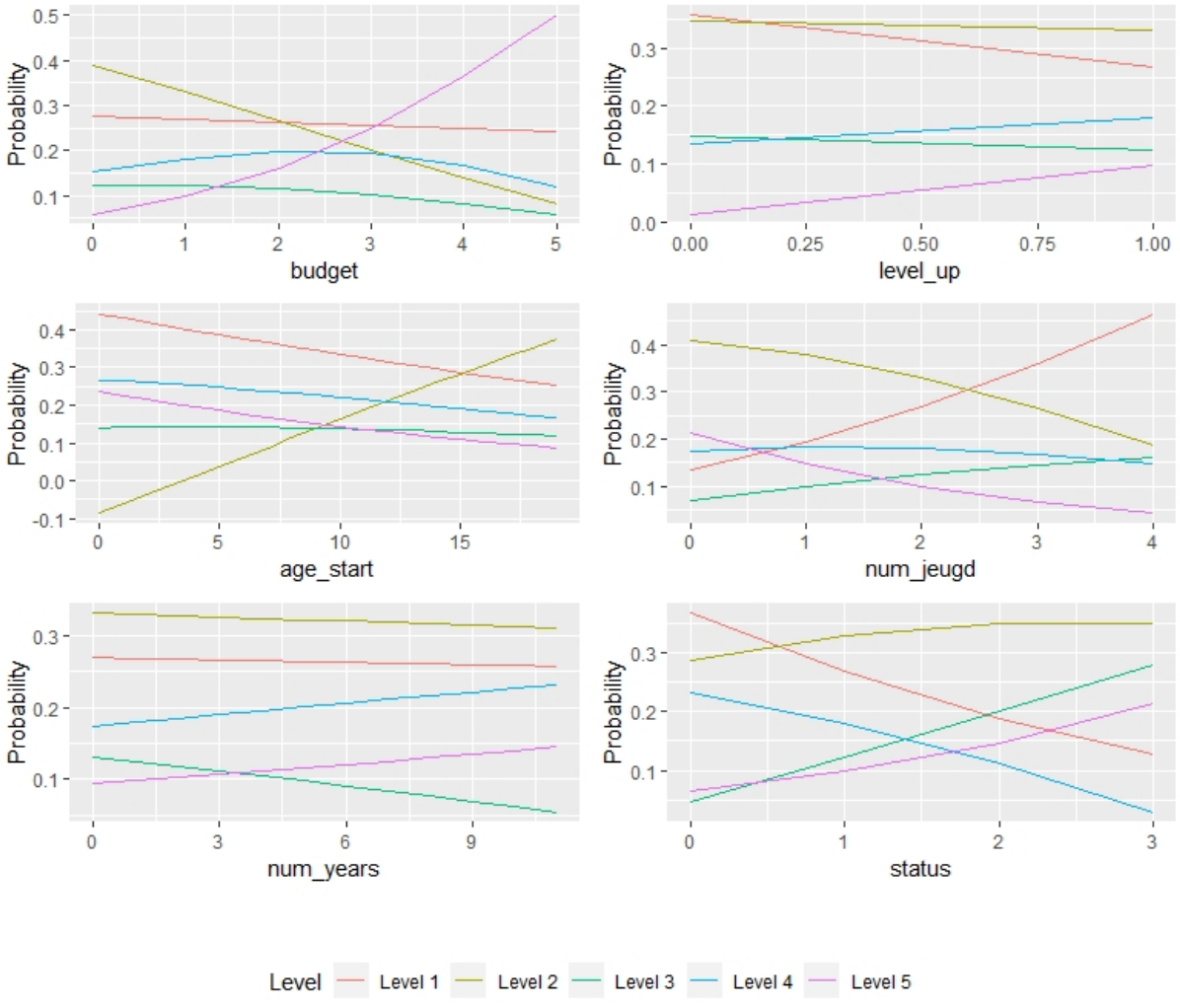


Figure 4. Probability of belonging to given class for each level of the concerning explanatory variable

When using the test set for making predictions, the model has an accuracy of 0.2177. As can be seen in the confusion matrix in Table 8, this model performs well for the highest class. For the other classes, the detection rate is much lower. Especially for class 1 and 4, where for class 1 often a higher class is predicted and for class 4 a lower class is predicted.

Table 8. Confusion matrix for NPOM

		Actual Class				
		1	2	3	4	5
Predicted Class	1	31	3	3	4	1
	2	53	7	8	6	0
	3	12	5	4	4	0
	4	38	6	3	8	0
	5	25	7	15	1	4

5.2.3 Comparison of models

When comparing the two models in terms of fit, several measures of goodness of fit can be considered, such as the AIC and the BIC, which can be used across models if the two are nested. These are summarized

in Table 9. It can be seen that both in terms of AIC and BIC, the NPOM scores lower than the POM, eventhough this model is more complex. This model is thus favoured over the more restricted POM model. Looking at the values for the OLS model, it can be seen that both measures are quite high. Keeping in mind the simplicity of the model, this means that the fit is most probably not satisfactory.

Table 9. Goodness of fit measures for POM and NPOM

Model	AIC	BIC
OLS	10873.8	10952.66
POM	9953.749	10014.41
NPOM	9646.714	9816.568

As shown in section 4.1, the NPOM and the POM are nested models, where the POM is a special case of the more general NPOM. Therefore, a likelihood ratio test can be performed to test the null hypothesis that the POM and the NPOM perform similarly in terms of fit versus the alternative that the NPOM performs better. This gives a p-value of $2.2e-16$, which is significant at 0.05%. Therefore, the null can be rejected and hence there is evidence that the NPOM model has a better fit.

In terms of accuracy the POM model performs best, as this model has the best detection rate for class 1. However, the NPOM has a higher detection rate for the other classes. The OLS model clearly performs poorly in terms of accuracy. Thus, the NPOM is preferred amongst the three when looking at predictions and performance.

Combining this with the fact that the POM violates the parallel lines assumption and for the OLS model the underlying assumptions are also violated and that the NPOM performs best in terms of fit, the NPOM is the favoured model amongst the parametric models.

5.3 Decision Trees

Next, the different decision trees are fitted. This is done with the cross-validation settings of $k = 10$ folds and this is repeated 5 times. The decision trees and their results are discussed for each type of used algorithm.

5.3.1 CART

In order to validate the importance of using the re-sampling algorithms to balance the data, the model is run without using this sampling technique and the performance is compared to the model with using the technique. In the case of not using sampling, the model always predicts class 1 of the dependent variable and the tree only consists of a root node. This gives an accuracy of 0.6368 on the full data, which is the same as the number of times class 1 occurs in the data. As all other data points are incorrectly classified, a model that can also classify other categories of the dependent variable is preferred, keeping in mind that the aim is to interpret the data, rather than reaching the best predictive model. Such model reflects more information about the data and thus sampling to balance to data should be used.

Therefore, CART is fitted in combination with up-sampling. As explained in Section 4.2.1 two different misclassification cost functions are considered, differing in the way the misclassification is penalized.

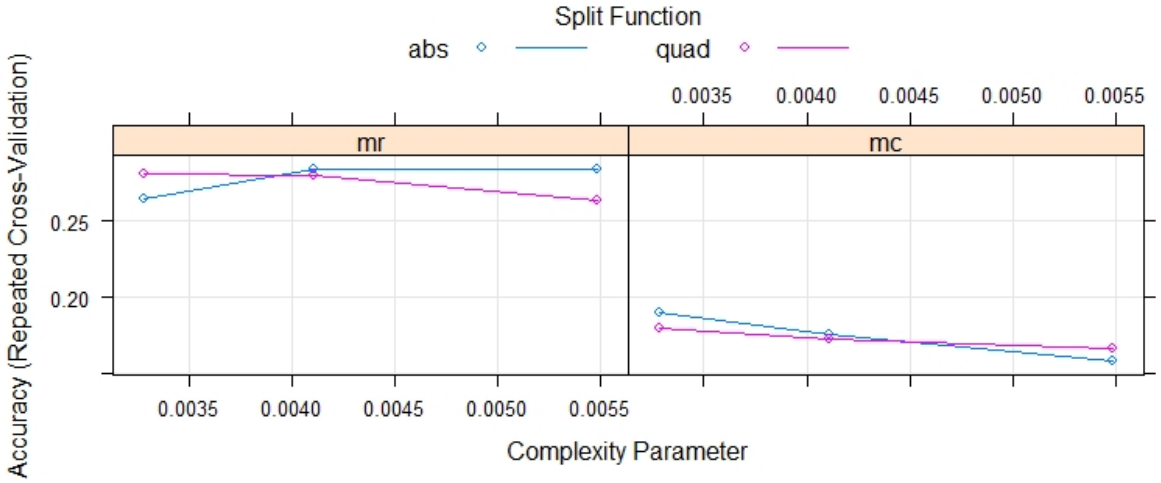


Figure 5. Accuracy for different hyper-parameter settings based on repeated cross-validation for CART with up-sampling. Here *mr* corresponds to the misclassification rate and *mc* to the misclassification cost.

Furthermore, two criteria for the tuning are considered, the misclassification rate and the cost of misclassification. During the training of the model the optimal combination of these two is searched for. Figure 5 shows the accuracy based on repeated cross-validation for the different settings and different values of the cost complexity factor. Clearly, tuning based on the misclassification rate results in a better model, especially in combination with an absolute value cost function. It can be seen that the accuracy stays constant after a complexity parameter of approximately 0.004. As a higher complexity parameter results in a smaller model, which generalizes better and is easier to interpret, a further customized grid search is performed to explore the effect of higher cost complexity factors. This grid search finds that for higher values than 0.008, the accuracy starts to decrease rapidly, while up until that point the accuracy does not differ greatly. Therefore, for the final model a cost complexity factor of $cp = 0.008$ is chosen, which results in a much smaller model, yet with an average accuracy close to the one with 0.004.

Fitting this final model on the test set, gives an accuracy of 0.1734. In order to gain more insight into what causes this low accuracy, it is useful to inspect the confusion matrix given in Table 10. The table indicates that the low accuracy is mainly caused by the low ability to detect class 1. This is also reflected in the sensitivity of this class, which is the faction of the correctly predicted values of the concerning class, which is only 0.06289 for the lowest class. However, it can be seen that often the model predicts class 2 for these wrongly predicted observations. As this class is adjacent to class 1 in terms of ordinality and both correspond to low success, this misclassification is not very harmful in terms of interpretation. For classes 4 and 5, the sensitivity is quite low, 0.17391 and 0.400000 respectively, thus suggesting that the model struggles with identifying a high level of success. Note that in this case the focus is on identifying the right class, rather than identifying which observations do not belong in a certain class, thus the sensitivity of each class gives the most relevant information.

Figure 6 shows the final decision tree. The first thing to notice here that clearly there are some non-linear effects found. Furthermore, the most important variables are the budget of the academy, the starting age and the number of years a player was part of the academy he played at longest, as these are the

Table 10. Confusion matrix for CART

		Actual Class				
		1	2	3	4	5
Predicted Class	1	10	0	0	1	0
	2	103	18	15	14	2
	3	22	5	9	4	1
	4	9	2	1	4	0
	5	15	3	8	0	2

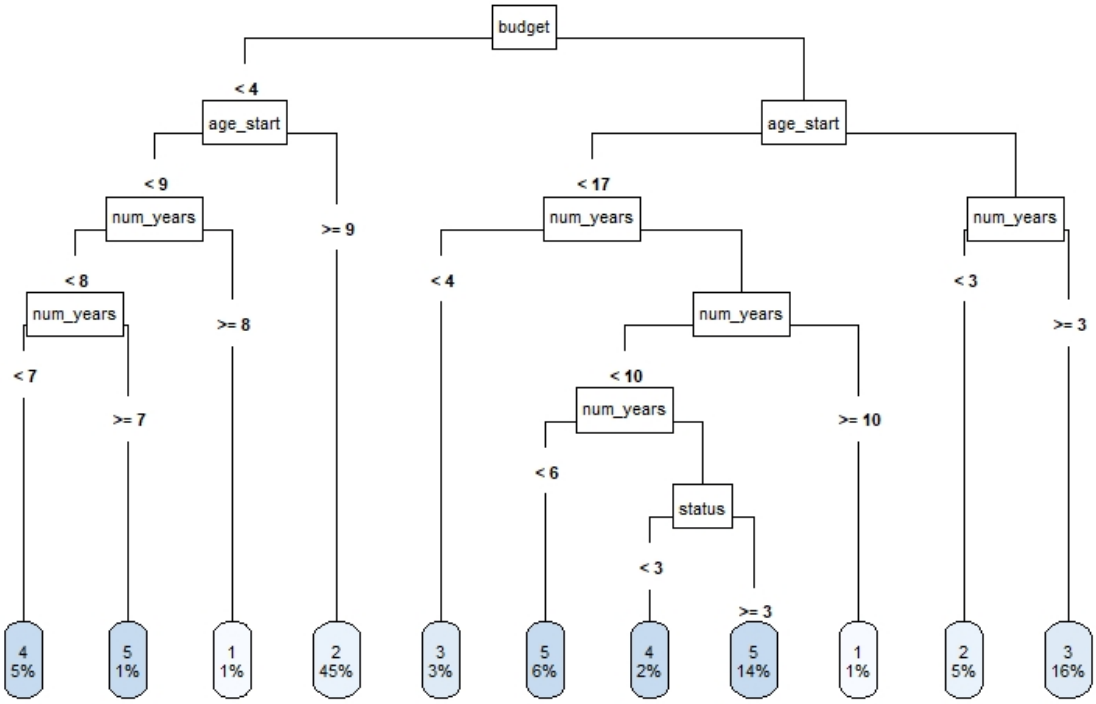


Figure 6. Decision tree using CART

variables the splits are based upon. Considering players that are labeled with category 4 or 5 as successful players, it can be seen that these players can be found in two parts of the tree. First of all, these are the players that are part of a youth academy that falls into the highest budget category, started playing at an academy before their teenage years (before 17) and played for less that 10 years at that academy. Note that there is an additional split on status for players that played between 6 and 10 years at the academy, where if the status of the academy is high, the player is labeled as part of the highest category of success. The other type of players with high success were part of an academy longest with a lower budget, but started at very young age (before 8). Another interesting finding of the tree is that some players at academies with a high budget fall into a low category of success, which is either class 1 or 2. However, as can be seen in the leave nodes this is only a small portion of the data. These are players that started at a late age or played at the academy for a long time. However, most players with low success are players that played longest at an academy with a lower budget and started playing at a later age than 8. Lastly, players that are not part of any academy in this case are labeled as category 4. However, as most of these players actually belong to the lowest 2 categories, this classification is not very accurate.

5.4 CHAID

As discussed in Section 4.2.2, all variables must be categorical for CHAID. In this case, all variables are transformed to ordinal variables with a natural ordering. This ensures that the algorithm only merges adjacent categories, thereby improving the interpretability. The significance boundaries are chosen to be 0.05. This corresponds to the significance levels used in step 1a and 2 of the CHAID algorithm described in Section 4.2.2. However, in this case the tree needs some additional pruning. The resulting tree with a 5% significance level has 201 terminal nodes and thus is not easily interpretable. In order to enhance this interpretability, the depth of the tree is limited to 3. Furthermore, step 1b is ignored in order to limit the number of splits and to reduce the computing time.

The model has an overall accuracy of 0.2984. Table 11 explains why the accuracy of the model is higher than for the CART model. This is mainly because the model is better in predicting class 1, for which it has a detection rate of 0.7727. However, it can also be seen that the model often incorrectly predicts class 1. For all other classes, class 1 is predicted more often than the actual class, resulting in low detection rates for all other classes. This is a problem especially for the highest two classes, as the model does not label them as having a high success correctly, but rather as belonging to low a success category.

Table 11. Confusion matrix for CHAID

		Actual Class				
		1	2	3	4	5
Predicted Class	1	51	32	10	32	34
	2	6	8	5	5	4
	3	4	6	4	3	16
	4	5	6	3	7	2
	5	0	0	1	0	4

Figure 7 shows the final tree. First of all, it can be seen that the tree indeed has multi-way splits, which is one of the key characteristics of CHAID. The split at the root is based on status, where all categories are considered separately. Players that did not play at any academy are reflected in the category with value 0 and are thus always labeled as belonging to the lowest class of success. Having played at an academy therefore can have a positive influence on the level of success, as it increases the chance to be labeled as belonging to a more successful category. Most players that played longest at an academy with a low status, are also labeled as having low success. Only the ones that play at the academy for 5 or 6 years are found to have a high success. For players at an academy with mediocre status, already more players are labeled as successful. More specifically, only players that player for less than 4 or more than 10 years are not labeled as successful. The last split with value 3 of the status variable has a similar structure, where players that played between 2 and 4 years or more than 10 years are assigned to the low classes. Thus, playing at an academy longest with a higher status only results in a higher success if the player is part of the academy for not too short and not too long.

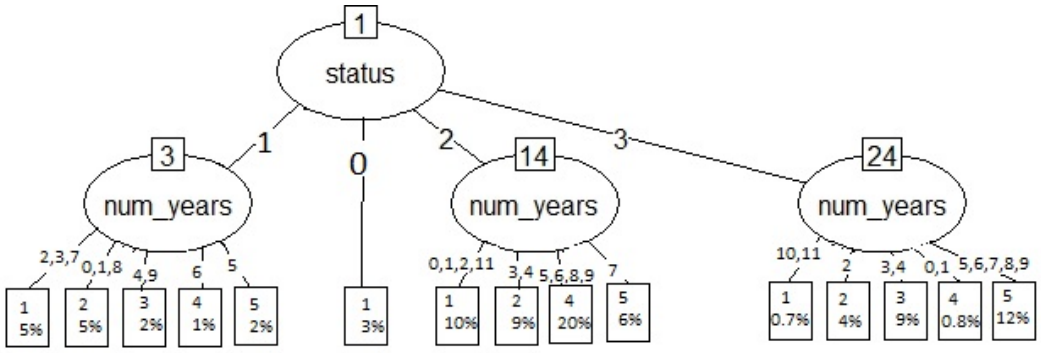


Figure 7. Decision tree using CHAID

5.4.1 CIT

The last decision tree that is fitted is the conditional inference tree. Initially, a significance level of 5% is chosen. However, as this results in a very large model, which is difficult to interpret, the significance level is reduced to 0.02. This more restrictive significance bound forces the model to built a smaller tree, hence enhancing the interpretability.

This results in a model with an accuracy of 0.1169 when using the test sample for prediction. This is again very low. However, similarly as before, this is mainly because of the model's inability to detect class 1. This can be seen in Table 12, which shows that class 1 is most often labeled as class 2 by the model. Furthermore, class 3 and 4 also have a low sensitivity. Both are also often falsely predicted as class 2. For class 3, this is not a very large problem, but in terms of interpretation the difference between class 4 and 2 is very different. Thus, this is not desirable. However, class 5 is often correctly predicted by the model.

Table 12. Confusion matrix for CIT

		Actual Class				
		1	2	3	4	5
Predicted Class	1	3	0	0	0	0
	2	116	17	15	18	1
	3	11	5	3	3	0
	4	6	1	0	2	0
	5	23	5	15	0	4

To gain further insight into the model using the CIT method, the decision tree is shown in Figure 8. Again, the first split is based on budget, indicating that this is the most important variable. Furthermore, this tree is larger than the other models, with more splitting rules and also smaller resulting subgroups, which is reflected by the smaller percentages in the leave nodes. Again, different types of players can be identified. Starting with players that played longest at an academy with a high budget, many different player profiles can be deducted. Players that only played for a short amount of years, more specifically 2 years or less, are not labeled as successful. However, players that played for 3 years can be successful, but only if the academy has a low status (below 2) and they started at a late age (after 17), or if the

academy has a high status and falls into the highest budget category. Furthermore, players that played for longer than 3 years at the academy they played at for the longest period and only played at the single academy are also found to have a high success. Players that played at more academies are also successful, except when they played for 5 or 6 years at the academy. For players that played at an academy with a low budget longest, there is also a possibility to be labeled as successful. This is the case when the player played at the academy he was part of longest for a long amount of years, where the budget of this club is in the middle category and more importantly in the case that the player made a level-up switch at the end of his youth career. Lastly, players that have not played at any academy are labeled as category 2, hence having reached a low success.

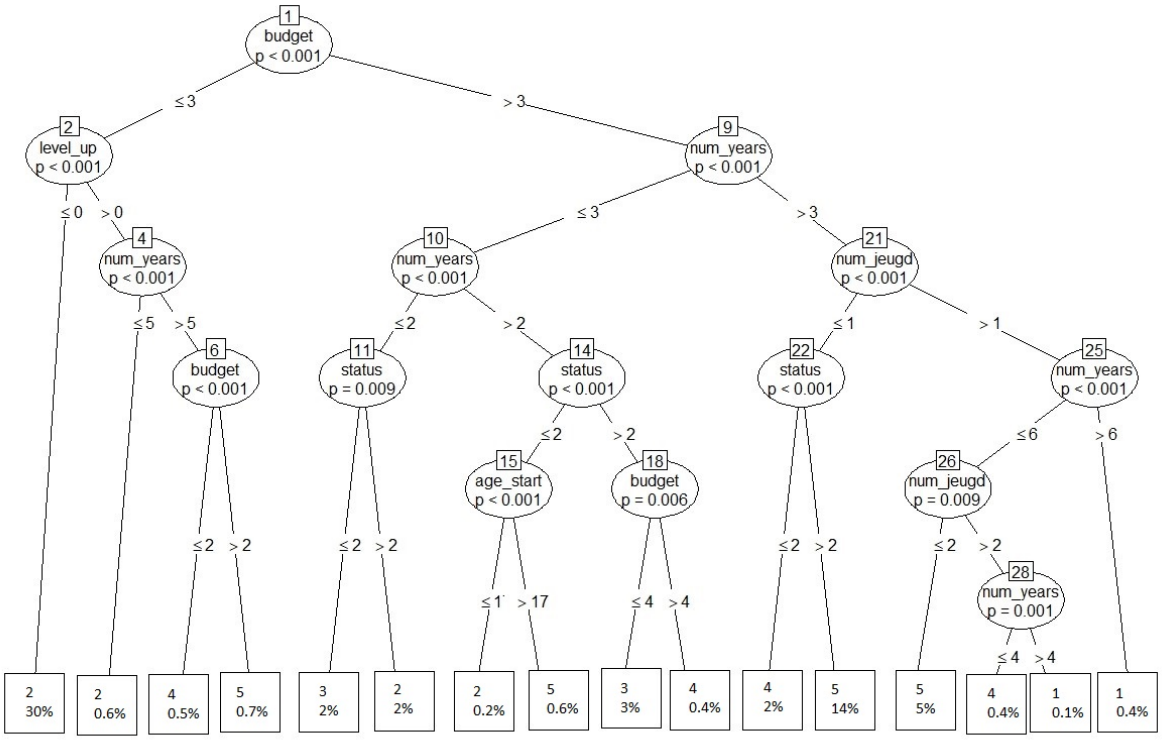


Figure 8. Decision tree using CIT

5.5 Regression Trees

The regression trees are built using the S-CART algorithm with the three different settings to incorporate the ordinal structure. The accuracies of the models with the median, rounded mean and mode are 0.1331, 0.129 and 0.129 respectively. Again, to gain insight into the performance, the confusion matrices can be considered, which are shown in Table 13. For all three settings it applies that the model does not or barely predicts class 1. It can be seen that these players are labeled as class 2 in most cases. This class is predicted too often, however, in all three models. For the model with the median and the rounded mean, the performance is especially good for classes 2 and 3, where the prediction is correct most of the times.

For the one with the mode, the prediction is less accurate for class 3. Additionally, the model does not perform very well for the highest two classes in all three cases. These players are often labeled with a lower class than the actual class they belong to.

Table 13. Confusion matrices for regression trees with S-CART

13.1: S-CART with median

		Actual Class				
		1	2	3	4	5
Predicted Class	1	0	0	0	0	0
	2	79	15	9	11	0
	3	59	9	15	10	2
	4	16	4	8	2	2
	5	5	0	1	0	1

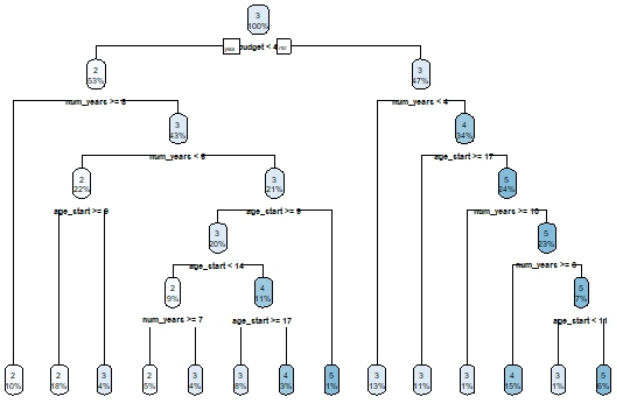
13.2: S-CART with rounded mean

		Actual Class				
		1	2	3	4	5
Predicted Class	1	0	0	0	0	0
	2	80	15	9	11	0
	3	58	9	15	10	2
	4	20	4	9	2	3
	5	1	0	0	0	0

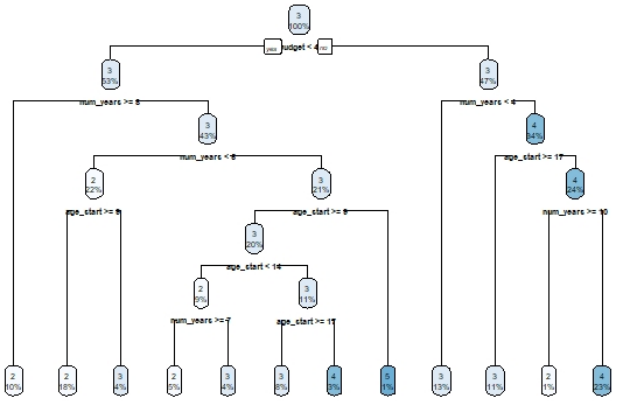
13.3: S-CART with mode

		Actual Class				
		1	2	3	4	5
Predicted Class	1	1	0	0	0	0
	2	128	22	18	21	2
	3	12	2	7	1	1
	4	0	0	0	0	0
	5	18	4	8	1	2

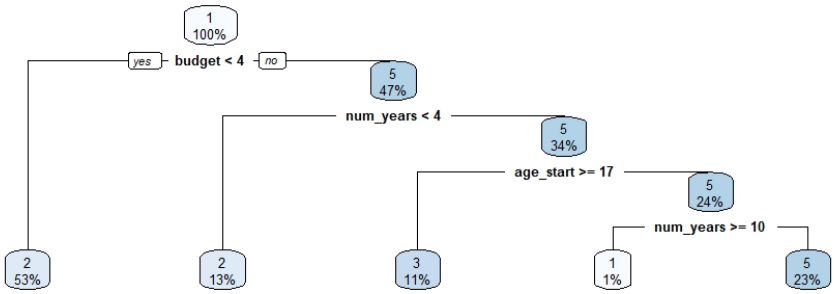
Figure 9 shows the resulting trees. First of all, the model with the median and rounded mean do not differ much in terms of structure and splits. It can be seen that only the right part of the tree is different, where the median has a few additional splits. Keeping in mind that the model does not predict the lowest class, the players that are labeled as having a low success are thus captured in class 2 and mostly appear in the left side of the tree. This corresponds to players that played longest at an academy with a low budget. However, players that played at such academy between 5 and 8 years and started playing before the age of 17 are labeled as category 4 or 5, thus having reach a high level of success. On the right side of the tree, more players are labeled with these higher classes of success. These are players that played at an academy longest with a high budget. More specifically, players that played between 6 and 10 years at the academy and started at an age between 11 and 17 are found to have a high success according to the model with the median. For the model with the rounded mean the same applies, only this model is less restrictive and players that played for 4 to 6 years at the academy also end up in the highest classes of the dependent variable. The model with the mode is much smaller. Here while only players that played at an academy with a high budget longest are labeled as successful, all other players are instantly labeled as not successful. Furthermore, only players that additionally to playing at such academy with a high budget, played between 4 and 10 years at the academy and started before the age of 17 are labeled as successful. In this sense, the model is similar to the other two in terms of interpretation, only having



(a) Model with median



(b) Model with rounded mean



(c) Model with mode

Figure 9. Trees built with S-CART

less splits.

Combining the structure, size and the performance of the trees, the preferred model can be chosen. First, it should be noticed that in terms of interpretation, there is no large difference between the trees. However, the tree with the mode is much smaller and thus more straight-forward to interpret. Furthermore, the models with the median and rounded mean never predict class 1 and the mode model never predicts class 4. However, the mode model performs better for class 2 and class 5 than the other two models. Therefore, overall the model with the mode is the preferred model amongst the three regression tree models.

5.6 Comparing the different models

Different type of models are used to model the reached level success of players and to find the impact of youth academies. Once again keeping in mind that this study is descriptive, rather than predictive, choosing which models are preferred is not based on only on the performance and accuracy, but also on the interpretability and complexity. However, the performance measures of the models can be used in order to determine how relevant the model is and the model's ability to detect different classes.

Although many different models are used, some obviously are not to be preferred. More specifically, when looking at the parametric models, the OLS and the POM model can be ignored, as they do not perform well in terms accuracy and validity of model assumptions. Rather, the focus should be on the found results of the NPOM model, as also mentioned in section 5.2.3. For the regression trees, the preferred model is the one using the mode, as discussed in the previous section. For the decision trees, many different models are used, all differing in methods, results and performance. In order to make a comparison across all different models a good criteria is needed. Such comparison can be based on accuracy. However, the detection rate per class is gives more information, as this reflects the ability of the model to identify each class separately. These detection rates are shown in Table 14. When looking at the overall accuracy, the first 3 methods obtain the highest values. Furthermore, clearly CIT and S-CART do not perform very well. When inspecting the first 3 methods more carefully, it can be seen that CHAID especially performs well for class 1, while the model has a lower detection rate for all the other classes than the other two models. On the other hand, NPOM seems to perform on average the best for all the classes. This therefore is the model that is most preferable. This model also has the second highest overall accuracy. When looking at the tree methods, the CART model is chosen to be the preferred model, as although the overall accuracy is lower than for CHAID, this methods performs better on average for each individual class.

	Class 1	Class 2	Class 3	Class 4	Class 5	Overall Accuracy
NPOM	0.1950	0.2500	0.1212	0.3478	0.8000	0.2177
CART	0.0629	0.6429	0.2727	0.1739	0.4000	0.1734
CHAID	0.7727	0.1539	0.1739	0.1489	0.0667	0.2984
CIT	0.0189	0.6071	0.0909	0.0870	0.8000	0.1169
S-CART (mode)	0.0063	0.7857	0.2121	0.0000	0.4000	0.1290

Table 14. Detection rate per class and overall accuracy of all models

The main advantage of the NPOM model is the ability to compute class probabilities. These are more

informative that a single class label. However, the model is quite large and has a threshold structure. This, in combination with the odds structure, leads to a complex model that is not easily interpretable. Furthermore, the model works under a linearity assumption, which might not be valid. Therefore, the CART model might be preferred, due to the less complex structure and ability to include non-linear effects. The tree structure allows for easy interpretation of these effects and investigating how the different variable interact with each other.

When comparing and combining the results of all the above models, and keeping in mind that the NPOM and the CART are the preferred models, some interesting overall observations can be made. First of all, players that have not played at any academy are found to be likely to have a low success. Having played at an academy does not necessarily increase the probability for success and this depends on the characteristics of the player and the academy. First of all, it is clear that the budget of the academy plays an important role. Clearly, playing at an academy with a high budget increases the probability to fall into a higher category of success. As the NPOM showed, the probability to reach a higher success strongly increases as the budget increases. However, this seems to apply especially for players that started playing at an academy at a young age and did not play at the academy for longer than 9 or 10 years. Furthermore, players that played at an academy longest with a low budget can also be successful, if they started at a very young age and play for approximately 5 to 10 years at the academy. Thus, combining these two observations with the results of the NPOM, it can be concluded that a high budget, a young starting age and playing at an academy for about 5 to 10 years all have a positive influence on the probability to fall into a higher level of success.

6 Conclusion

This study aims to explore the effectiveness of Dutch soccer youth academies on the future career of their attendants. This is done by exploring the relation between being part of a youth academy and the level of success of their players during their early career and investigating what factors and characteristics of the youth academies and their players play an important role in determining the level of success that a player reaches after leaving the youth academy.

In order to do so, first the issue of how to measure the effectiveness is addressed. This effectiveness cannot be directly measured, as it is unobservable. Hence, it should be captured in an observable proxy. As stated before, in this paper the effectiveness is measured by the level of success a player reaches in the three years after leaving the youth academy. Obviously, a higher level of success means that the youth academy is more effective in boosting the athlete's career. The level of success of a player is captured by the leagues that the clubs that the player plays at during these three years compete in. This information is scraped using the online player transfer database.

The next part of the problem is measuring the effect and finding the appropriate method or methods to do so. In order to measure the effect and to investigate which factors are important, first a database is created. This database consists of different players with different characteristics of the players and the

academies they played at. Using the transfer history that was scraped, the database is built and different variables are created that give information about the players and the academy that the players played at longest.

Different models are used to describe the level of success and to find the relation with the different explanatory variables. As the level of success is an ordinal variable, the models should be able to incorporate this ordinal structure. In terms of parametric models, therefore the Proportional Odds and the Non-Proportional Models are used. These are compared to OLS, which is a simple technique and therefore functions as a benchmark model. Using the POM, class probabilities can be computed and used to determine what the impact of the explanatory variables are on these class probabilities. Furthermore, in order to incorporate non-linear effects and avoiding being limited to making linearity assumptions, different tree algorithms are used. For decision trees CART, CHAID and Conditional Inference trees are used and for regression trees an altered S-CART model is used. To prevent overfitting issues, these models are tuned and optimal trees are found and interpreted. All these models, both the parametric and non-parametric, are combined with up-sampling in order to deal with the large unbalance in the data.

All models differ in terms of fit, results and structure. However, comparing the different models in terms of accuracy and more importantly the detection rate of each individual class, two models are preferred. This first is the NPOM, which overall has the most satisfactory performance in terms of class detection rate. One main advantage of this model is the ability to compute class probabilities. However, due to the odds structure, the interpretation is not straight-forward and as mentioned before the model relies on a linearity assumption. Therefore, the CART model is also interesting to consider. Of all tree methods, this is the preferred model in terms of class detection rate.

After establishing the two preferred models for estimation, the results of the models can be combined in order to answer the main research question on whether youth academies enhance the future success of their attendants and what factors play a role. When looking at the findings of the models, several observations can be made. First of all, players that do not play at any academy have a low probability of success. Having played at an academy generally does increase the probability to become more successful, but not necessarily. The most important driving factor of the probability of success is the budget of the academy that the player played at longest. Players that were part of an academy with a high budget have a much higher probability to become successful in their early career, especially when combined with a young starting age and having played at the academy for less than 9 or 10 years. Nevertheless, also players that played at an academy with a lower budget can become successful. This is again for players that started playing at an academy at a young age and played for 5 to 10 years at the academy with the low budget. Thus, playing at an academy longest with a high budget, playing there for 5 to 10 years and start playing at a youth academy at a young age all have a positive effect on becoming more successful in the future.

One of the most important limitations of this research lies in the data generating process. Here, multiple problems can be detected. First of all, probably the most important issue is the large unbalance in the

data. This unbalance results in issues with fit and measuring the accuracy of the different models. One solution is used here, which is up-sampling the data, however more research can be done on different solutions for this problem, where different methods might result in different and possibly more accurate results. However, another solution might be using a different data generation process, where a more balanced dataset is created. Thus, more players that are successful or less players with low success are included in the database. Additionally, in general having a larger database would be beneficial. More players can be included to give a more complete database and more accurate results.

Another limitation is the number and type of explanatory variables that are included in the analysis. In order to give a more complete overview of all the possible influencing factors, thus giving more accurate results, more explanatory variables should be included in the dataset. This also allows for including different type of variables, such as ones focusing on the psychological information of the players, more detailed information on the academies and other factors that might have an influence. In this study only public information is used, but it would be interesting to also include factors such as training volume, the size of the academy or the talent or performance of the player. This means extending the scope of the study, which can lead to interesting results and give a more complete overview of what drives success for players.

Lastly, this study proposes different models for estimation. However, some other models that incorporate the ordinal structure might also be interesting to consider. For example, using random forests or a latent structure might be an option. As such models are not much covered in the existing literature on ordinal estimation, this might be interesting to look into. Furthermore, it might be interesting to use selection criteria in order to be able to use discontinuity models. If this data is available, these methods can be used to more directly measure the effect and compare selected players to players to did not get scouted for a youth academy.

References

Ananth, C. V., & Kleinbaum, D. G. (1997). Regression models for ordinal responses: a review of methods and applications. *International journal of epidemiology*, *26*(6), 1323–1333.

Archer, K. J. (2010). rpartordinal: An r package for deriving a classification tree for predicting an ordinal response. *Journal of Statistical Software*, *34*, 7.

Barreiros, A., Côté, J., & Fonseca, A. M. (2014). From early to adult sport success: Analysing athletes' progression in national squads. *European journal of sport science*, *14*(sup1), S178–S182.

Bayam, E., Liebowitz, J., & Agresti, W. (2005). Older drivers and accidents: A meta analysis and data mining application on traffic accident data. *Expert Systems with Applications*, *29*(3), 598–629.

Berkowitz, D., & Hoekstra, M. (2011). Does high school quality matter? evidence from admissions data. *Economics of Education review*, *30*(2), 280–288.

Bernard, J. M. (2015). *An application of data analytics to outcomes of missouri motor vehicle crashes*. University of Missouri-Saint Louis.

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, 1171–1178.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees (wadsworth, belmont, ca). *ISBN-13*, 978–0412048418.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Christensen, R. H. B. (2018). Cumulative link models for ordinal regression with the r package ordinal. *Submitted in J. Stat. Software*.

Clark, D. (2007). Selective schools and academic achievement.

Cohodes, S. R. (2020). The long-run impacts of specialized programming for high-achieving students. *American Economic Journal: Economic Policy*, *12*(1), 127–66.

Dobbie, W., & Fryer Jr, R. G. (2015). The medium-term impacts of high-achieving charter schools. *Journal of Political Economy*, *123*(5), 985–1037.

Emrich, E., Fröhlich, M., Klein, M., & Pitsch, W. (2009). Evaluation of the elite schools of sport: empirical findings from an individual and collective point of view. *International Review for the Sociology of Sport*, *44*(2-3), 151–171.

Fu, V. K., et al. (1999). Estimating generalized ordered logit models. *Stata technical bulletin*, *8*(44).

Fujiwara, K., Okamoto, Y., Kameari, A., & Ahagon, A. (2005). The newton-raphson method accelerated by using a line search-comparison between energy functional and residual minimization. *IEEE transactions on magnetics*, *41*(5), 1724–1727.

Grossmann, B., & Lames, M. (2013). Relative age effect (rae) in football talents–the role of youth academies in transition to professional status in germany. *International Journal of Performance Analysis in Sport*, *13*(1), 120–134.

Gulbin, J., Weissensteiner, J., Oldenziel, K., & Gagné, F. (2013). Patterns of performance development in elite athletes. *European journal of sport science*, *13*(6), 605–614.

Güllich, A. (2007). *Training – support – success: Control-related assumptions and empirical findings.* Saarbrücken: University of the Saarland [in German].

Güllich, A., & Emrich, E. (2006). Evaluation of the support of young athletes in the elite sports system. *European Journal for Sport and Society*, *3*(2), 85–108.

Güllich, A., & Emrich, E. (2012). Individualistic and collectivistic approach in athlete support programmes in the german high-performance sport system. *European Journal for Sport and Society*, *9*(4), 243–268.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, *15*(3), 651–674.

Ivarsson, A., Stenling, A., Fallby, J., Johnson, U., Borg, E., & Johansson, G. (2015). The predictive ability of the talent development environment on youth elite football players' well-being: A person-centered approach. *Psychology of Sport and Exercise*, *16*, 15–23.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *29*(2), 119–127.

Kramer, S., Widmer, G., Pfahringer, B., & De Groeve, M. (2001). Prediction of ordinal classes using regression trees. *Fundamenta Informaticae*, *47*(1-2), 1–13.

Liu, X., & Koirala, H. (2012). Ordinal regression analysis: Using generalized ordinal logistic regression models to estimate educational data. *Journal of modern Applied Statistical methods*, *11*(1), 21.

Lu, J., Ma, X., & Xing, Y. (2021). Risk factors affecting the severity of disruptions in metro operation in shanghai, 2013-2016. *Journal of Transportation Safety & Security*, *13*(1), 69–92.

Lucas, A. M., & Mbiti, I. M. (2014). Effects of school quality on student achievement: Discontinuity evidence from kenya. *American Economic Journal: Applied Economics*, *6*(3), 234–63.

Magidson, J., & Vermunt, J. K. (2005). An extension of the chaid tree-based segmentation algorithm to multiple dependent variables. In *Classification—the ubiquitous challenge* (pp. 176–183). Springer.

Malina, R. M. (2010). Early sport specialization: roots, effectiveness, risks. *Current sports medicine reports*, *9*(6), 364–371.

Mingers, J. (1989). An empirical comparison of pruning methods for decision tree induction. *Machine learning*, *4*(2), 227–243.

Nesti, M., & Sulley, C. (2014). *Youth development in football: Lessons from the world's best academies.* Routledge.

Peterson, B., & Harrell Jr, F. E. (1990). Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *39*(2), 205–217.

Seiler, S. (2013). Evaluating the (your country here) olympic medal count. *International journal of sports physiology and performance*, *8*(2), 203–210.

Vaeyens, R., Güllich, A., Warr, C. R., & Philippaerts, R. (2009). Talent identification and promotion programmes of olympic athletes. *Journal of sports sciences*, *27*(13), 1367–1380.

Weng, J., Meng, Q., & Wang, D. Z. (2013). Tree-based logistic regression approach for work zone casualty risk assessment. *Risk analysis*, *33*(3), 493–504.

White, A. P., & Liu, W. Z. (1994). Bias in information-based measures in decision tree induction.

Machine Learning, *15*(3), 321–329.

Wilkinson, L. (1992). Tree structured data analysis: Aid, chaid and cart. *Retrieved February*, *1*, 2008.

Wu, J., Wei, X., Zhang, H., & Zhou, X. (2019). Elite schools, magnet classes, and academic performances: Regression-discontinuity evidence from china. *China Economic Review*, *55*, 143–167.

Appendices

A Overview of variables

Variable	Type	Category	Description
Success	Ordinal	1	Regional or terminated
		2	Tweede Divisie
		3	Keukenkampioen
		4	Eredivisie
		5	Europa League or Champions League
Status	Ordinal	1	Regional Quality status
		2	National Quality status
		3	International Quality status
Budget	Ordinal	1	[0, 6000000)
		2	[6000000, 12000000)
		3	[12000000, 20000000)
		4	[20000000, 26000000)
		5	[26000000, ∞)
level_up	Binary	1	Player has transferred to a higher ranking academy during the last 2 years of his youth career
		0	otherwise
num_jeugd	Continuous		Number of youth academies that the player has been part of
num_years	Continuous		Number of years the player played at the academy that he has been part of longest
age_start	Continuous		Age at which player started played at a youth academy

B Short description of code

Separate files are used the data scraping, building the dataframe for estimation and the actual estimation. Each will be shortly described.

The *scraping* file is a Phyton file that is used for scraping the player data and transfer history(see Section 3.1). The code uses the HTML structure of the site to scrape the data using the beautiful soup package. First all players that need to be included are found, then the data of these players are scraped from their personal page. Which players should be included is described in Section 3.1.

The *createvariables* file is a Phyton file that creates the database that is used for estimation using the scraped data. This file furthermore uses a list of all considered youth academies and their budget category as input data. The code creates the dependent variable (see Section 3.2), gets the academy specific variables (see Section 3.3.1) and the player specific variables (see Section 3.3.2) and aggregates them in a dataframe.

The *newfit* file is a R file that is used for the estimation. This file uses the dataframe created in the *createvariables* file as input data. Here the models of Section 4 are estimated and the results are given and visualized, using appropriate statistical packages. The procedure of estimation is described in Section 4.4 more accurately.